

Lecture Notes in Electrical Engineering 315

Hamzah Asyrani Sulaiman

Mohd Azlishah Othman

Mohd Fairuz Iskandar Othman

Yahaya Abd Rahim

Naim Che Pee

Editors

Advanced Computer and Communication Engineering Technology

Proceedings of the 1st International
Conference on Communication and
Computer Engineering

 Springer

Lecture Notes in Electrical Engineering

Volume 315

Board of Series editors

Leopoldo Angrisani, Napoli, Italy
Marco Arteaga, Coyoacán, México
Samarjit Chakraborty, München, Germany
Jiming Chen, Hangzhou, P.R. China
Tan Kay Chen, Singapore, Singapore
Rüdiger Dillmann, Karlsruhe, Germany
Haibin Duan, Beijing, China
Gianluigi Ferrari, Parma, Italy
Manuel Ferre, Madrid, Spain
Sandra Hirche, München, Germany
Faryar Jabbari, Irvine, USA
Janusz Kacprzyk, Warsaw, Poland
Alaa Khamis, New Cairo City, Egypt
Torsten Kroeger, Stanford, USA
Tan Cher Ming, Singapore, Singapore
Wolfgang Minker, Ulm, Germany
Pradeep Misra, Dayton, USA
Sebastian Möller, Berlin, Germany
Subhas Mukhopadhyay, Palmerston, New Zealand
Cun-Zheng Ning, Tempe, USA
Toyoaki Nishida, Sakyo-ku, Japan
Federica Pascucci, Roma, Italy
Tariq Samad, Minneapolis, USA
Gan Woon Seng, Nanyang Avenue, Singapore
Germano Veiga, Porto, Portugal
Haitao Wu, Beijing, China
Junjie James Zhang, Charlotte, USA

About this Series

“Lecture Notes in Electrical Engineering (LNEE)” is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering

LNEE publishes authored monographs and contributed volumes which present cutting edge research information as well as new perspectives on classical fields, while maintaining Springer’s high standards of academic excellence. Also considered for publication are lecture materials, proceedings, and other related materials of exceptionally high quality and interest. The subject matter should be original and timely, reporting the latest research and developments in all areas of electrical engineering.

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer’s other Lecture Notes series, LNEE will be distributed through Springer’s print and electronic publishing channels.

More information about this series at <http://www.springer.com/series/7818>

Hamzah Asyrani Sulaiman
Mohd Azlishah Othman
Mohd Fairuz Iskandar Othman
Yahaya Abd Rahim · Naim Che Pee
Editors

Advanced Computer and Communication Engineering Technology

Proceedings of the 1st International
Conference on Communication
and Computer Engineering

Editors

Hamzah Asyrani Sulaiman
Mohd Azlishah Othman
Mohd Fairuz Iskandar Othman
Yahaya Abd Rahim
Naim Che Pee
Universiti Teknikal Malaysia Melaka
Melaka
Malaysia

ISSN 1876-1100

ISSN 1876-1119 (electronic)

ISBN 978-3-319-07673-7

ISBN 978-3-319-07674-4 (eBook)

DOI 10.1007/978-3-319-07674-4

Library of Congress Control Number: 2014947654

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law. The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The first International Conference on Communication and Computer Engineering (ICOCOE 2014) was held in the historical city of Malacca, Malaysia, between May 20 and 21, 2014. The conference focuses on industrial and manufacturing theory and applications of electronics, communications, computing, and information technology. The objectives of the conference were to provide an avenue for researchers to present high-quality research and be involved in professional interactions for the advancement of science, technology, and fellowship. More than 300 participants from 15 countries registered for the conference. ICOCOE 2014 provided a platform for them to present their latest invention and advanced technology and research, share their expertise, establish new contacts, and discuss a range of communication and computer engineering topics for the betterment and future advancement of these fields.

ICOCOE 2014 is the first ever event handled by Prime Systems, an educational consulting sector provider. It was successful in getting overwhelming response from public and private sectors, local and foreign universities, research institutions, stakeholders, and various industries from all over the region. Nearly 200 papers were received with only around 102 papers accepted. Associate Professor Dr. Muhammad Ramlee from Universiti Teknologi Malaysia gave a keynote address on RF and microwave communication system in advanced systems while Prof. Eryk Dutkiewicz from Macquarie University presented his latest work on wireless communications. Several technical sessions were arranged for the conference.

This book contains a selection of revised manuscripts presented at ICOCOE 2014 within the two main conference themes, which are communications and computer engineering research areas. This book covers various aspects of advanced computer and communication engineering, specifically on industrial and manufacturing theory and applications of electronics, communications, computing, and information technology. The development of latest technologies has been

highlighted, aimed at various computer and communication-related professionals such as telecommunication engineers, computer engineers and scientists, researchers, academicians, and students. Furthermore, several applications involving cutting edge communication and computer systems are also discussed.

The editors wish to acknowledge Prime Systems for successfully organizing and sponsoring this conference.

Hamzah Asyrani Sulaiman
Mohd Azlishah Othman
Mohd Fairuz Iskandar Othman
Yahaya Abd Rahim
Naim Che Pee

General Committee and Advisors

General Committee and Advisor/Editorial Board

Mohd Azlishah Bin Othman	Universiti Teknikal Malaysia Melaka, Malaysia
Ahmad Naim Bin Che Pee	Universiti Teknikal Malaysia Melaka, Malaysia
Hamzah Asyrani Bin Sulaiman	Universiti Teknikal Malaysia Melaka, Malaysia
Yahaya Bin Abdul Rahim	Universiti Teknikal Malaysia Melaka, Malaysia
Mohd Fairuz Iskandar Bin Othman	Universiti Teknikal Malaysia Melaka, Malaysia

International Advisor/Technical Program Committee

Prof. Hj. Kamaruzaman Jusoff
Dr. Abdel Rahman Mohammad Said Al-Tawaha
Prof. Dr. R.J. Godwin
Prof. Dr. Erik Valdemar Cuevas Jimenez
Prof. Dr. Arun Kumar Gupta
Prof. Robert Morelos-Zaragoza

Running Committee

Chairman

Alishamuddin Enjom

Co-Chairman

Shahrizan Jamaludin

Secretary

Nurulhuda Abu Talib

Treasurer

Zamani Marjom

Technical Program Chair

Siti Zawiyah Iskandar

Contents

Part I Communication

1	RWA: Comparison of Genetic Algorithms and Simulated Annealing in Dynamic Traffic	3
	Arturo Rodriguez, Antonio Gutierrez, Luis Rivera and Leonardo Ramirez	
2	Route Optimization in Proxy Mobile IPv6 Test-Bed via RSSI APPs	15
	Nur Haliza Binti Abdul Wahab, L.A. Latif, S.H.S. Ariffin, N. Fisal and N. Effiyana Ghazali	
3	Polytetrafluoroethylene Glass Microfiber Reinforced Slotted Patch Antenna for Satellite Band Applications	29
	M. Samsuzzaman, T. Islam and M.R.I. Faruque	
4	2.4 GHz Circularly Polarized Microstrip Antenna for RFID Application	37
	Rudy Yuwono and Ronanobelta Syakura	
5	On Understanding Centrality in Directed Citation Graph	43
	Ismael A. Jannoud and Mohammad Z. Masoud	
6	Channel Capacity of Indoor MIMO Systems in the Presence of Spatial Diversity	53
	M. Senon, M.N. Husain, A.R. Othman, M.Z.A. Aziz, K.A.A. Rashid, M.M. Saad, M.T. Ahmad and J.S. Hamidon	

7	Design of Multi-band Antenna for Wireless MIMO Communication Systems	63
	M.M. Saad, M.N. Husain, M.Z.A. Aziz, A.R. Othman, K.A.A. Rashid and M. Senon	
8	Design of Linear Polarization Antenna for Wireless MIMO Application	73
	K.A.A. Rashid, M.N. Husain, A.R. Othman, M.Z.A. Aziz, M.M. Saad, M. Senon, M.T. Ahmad and J.S. Hamidon	
9	The Effect of IV Characteristics on Optical Control of SDR Si IMPATT Diode	85
	T.S.M. Arshad, M.A. Othman, M.N. Hussain and Y.A. Rahim	
10	Variable Intrinsic Region in CMOS PIN Photodiode for I–V Characteristic Analysis	95
	M.A. Othman, N.Y.M. Yasin, T.S.M. Arshad, Z.A.F.M. Napiyah, M.M. Ismail, H.A. Sulaiman, M.H. Misran, M.A. Meor Said and R.A. Ramlee	
11	Variable Depletion Region in CMOS PN Photodiode for I–V Characteristic Analysis	103
	M.A. Othman, T.S.M. Arshad, Z.A.F.M. Napiyah, M.M. Ismail, N.Y.M. Yasin, H.A. Sulaiman, M.H. Misran, M.A. Meor Said and R.A. Ramlee	
12	U-Slot Rectangular Patch Antenna for Dual Band Application	111
	Mohammad Shawkat Habib, I.M. Rafiqul, Khaizuran Abdullah and M. Jamil Jakpar	
13	Analysis of Synthetic Storm Technique Based on Ku-Band Satellite Beacon Measurements in Malaysia	121
	Ali K. Lwas, I.M. Rafiqul, Mohamed Hadi Habaebi, Ahmad F. Ismail, Mandeep Singh, Jalel Chebil, Al-Hareth Zyoud and Hassan Dao	
14	The Evolution of Double Weight Codes Family in Spectral Amplitude Coding OCDMA	129
	N. Din Keraf, S.A. Aljunid, A.R. Arief and P. Ehkan	

15 Performance Evaluation of LTE Scheduling Techniques for Heterogeneous Traffic and Different Mobility Scenarios 141
 Lukmanhakim Sukeran, Mohamed Hadi Habaebi,
 Al-Hareth Zyoud, Musse Mohamud Ahmad, Shihab Hameed,
 Amelia Wong and I.M. Rafiqul

16 Design for Energy-Aware IP Over WDM Networks with Hibernation Mode and Group-Node Techniques 151
 M.N.M. Warip, Ivan Andonovic, Ivan Glesk,
 R. Badlishah Ahmad, P. Ehkan,
 Mohamed Elshaikh Elobaid Said Ahmed,
 Shamsul Jamel Elias and Fazrul Faiz Zakaria

17 Content Based Image Retrieval Using Color Layout Descriptor and Generic Fourier Descriptor 163
 Muhammad Imran, Rathiah Hashim and Noor Elaiza

18 Pilot Based Pre FFT Signal to Noise Ratio Estimation for OFDM Systems in Rayleigh-Fading Channel 171
 A.M. Khan, Varun Jeoti and M. Azman Zakariya

19 The Use of Convolutional Code for Narrowband Interference Suppression in OFDM-DVBT System. 183
 Aizura Abdullah, Muhammad Sobrun Jamil Jamal,
 Khaizuran Abdullah, Ahmad Fadzil Ismail and Ani Liza Asnawi

20 Two-Elements Crescent Shaped Printed Antenna for Wireless Applications 195
 Wan Noor Najwa Wan Marzudi, Zuhairiah Zainal Abidin,
 Ma Yue and Raed A. Abd-Alhameed

21 Wideband Linearly Polarized Printed Monopole Antenna for C-Band 205
 Touhidul Alam, Mohammad Rashed Iqbal Faruque
 and Mohammad Tariqul Islam

22 A Novel Anti-collision Protocol for Optimization of Remote Sensing in Dense Reader Network. 213
 Faiza Nawaz and Varun Jeoti

23 Double Square Loop Frequency Selective Surface (FSS) for GSM Shielding 223
 Nur Khalida Binti Abdul Khalid
 and Fauziahanim Binti Che Seman

24 Analysis of the Active Region of Archimedean Spiral Antenna 231
 Abdirahman Mohamoud Shire and Fauziahanim Che Seman

25 Optimization of BER Performance in the MIMO-OFDMA System for Mobile WiMAX System Using Different Equalization Algorithm. 241
 Azlina Idris, Norhayati Abdullah, Nor Azlizan Hussein
 and D.M. Ali

26 Performance Analysis of Polling Delay in Transparent and Non-transparent Multi-hop Relay WiMAX Network 251
 Mohd Daud A. Hassan, Habibah Hashim
 and D.M. Ali

27 Blind Source Computer Device Identification from Recorded Calls 263
 Mehdi Jahani-rad, Ainuddin Wahid Abdul Wahab
 and Nor Badrul Anuar

28 Visibility for Network Security Enhancement in Internet Protocol Over Ethernet Networks 277
 W.K. Alzubaidi, Longzheng Cai,
 Shaymaa A. Alyawer and Erika Siebert-Cole

29 Comparative Analysis of Different Single Cell Metamaterial 289
 Pankaj Rameshchandra Katiyar
 and Wan Nor Liza Binti Wan Mahadi

30 Distributed Video Coding with Frame Estimation at Decoder 299
 Kin Honn Chiam and Mohd Fadzli Mohd Salleh

31 Performance Analysis of an OCDMA System Based on SPD Detection Utilizing Different Type of Optical Filters for Access Networks 309
 Sarah G. Abdulqader, Hilal A. Fadhil, S.A. Aljunid
 and Anuar Mat Safar

32	Deployment of Optimized Algorithm for MPEG-4 Data Over Wireless Multimedia Sensor Network	321
	Norlezah Hashim, Sharifah Hafizah Syed Ariffin, Farizah Yunus, Fakrulradzi Idris and Norsheila Fisal	
33	Partially Compensated Power Control Technique for LTE-A Macro-Femto Networks	331
	Sawsan Ali Saad, Mahamod Ismail and Rosdiadee Nordin	
34	Design and Development of the Visible Light Communication System.	343
	Anuar Musa, Mazlaini Yahya, Nazaruddin Omar, Mohd Kamarulzamin Salleh and Noor Aisyah Mohd Akib	
35	The Embroidered Antenna on Bending Performances for UWB Application	349
	M.S. Shakhirul, A. Sahadah, M. Jusoh, A.H. Ismail and Hasliza A. Rahim	
36	The Embroidered Wearable Antenna for UWB Application.	357
	M.S. Shakhirul, A. Sahadah, M. Jusoh, A.H. Ismail, C.M. Nor and F.S. Munirah	
37	Bowtie Shaped Substrate Integrated Waveguide Bandpass Filter	365
	Z. Baharudin, M.Z.U. Rehman, M.A. Zakariya, M.H.M. Khir, M.T. Khan and J.J. Adz	
38	Logical Topology Design with Low Power Consumption and Reconfiguration Overhead in IP-over-WDM Networks	375
	Bingbing Li and Young-Chon Kim	
Part II Computer		
39	Systematic Analysis on Mobile Botnet Detection Techniques Using Genetic Algorithm.	389
	M.Z.A. Rahman and Madihah Mohd Saudi	

40	An Empirical Study of the Evolution of PHP MVC Framework	399
	Rashidah F. Olanrewaju, Thouhedul Islam and N. Ali	
41	Evolutionary Approach of General System Theory Applied on Web Applications Analysis	411
	Aneta Bartuskova, Ondrej Krejcar and Kamil Kuca	
42	A Novel Distributed Image Steganography Method Based on Block-DCT	423
	Rosemary Koikara, Dip Jyoti Deka, Mitali Gogoi and Rig Das	
43	An Improved History-Based Test Prioritization Technique Using Code Coverage	437
	Avinash Gupta, Nayneesh Mishra, Aprna Tripathi, Manu Vardhan and Dharmender Singh Kushwaha	
44	Local Pricewatch Information Solicitation and Sharing Model Using Mobile Crowdsourcing	449
	Hazleen Aris	
45	Enhancement of Nurse Scheduling Steps Using Particle Swarm Optimization	459
	Norhayati Mohd Rasip, A.S.H. Basari, Nuzulha Khilwani Ibrahim and Burairah Hussin	
46	Hardware Implementation of MFCC-Based Feature Extraction for Speaker Recognition	471
	P. Ehkan, F.F. Zakaria, M.N.M. Warip, Z. Sauli and M. Elshaikh	
47	Parallel ASIP Based Design of Turbo Decoder	481
	F.F. Zakaria, P. Ehkan, M.N.M. Warip and M. Elshaikh	
48	A Comparative Study of Web Application Testing and Mobile Application Testing	491
	Maryam Ahmed and Rosziati Ibrahim	
49	Multi-objective Functions in Grid Scheduling	501
	Zafril Rizal M. Azmi, M.A. Ameen and Imran Edzereiq Kamarudin	

50	Experimental Analysis on Available Bandwidth Estimation Tools for Wireless Mesh Network	525
	Imran Edzereiq Kamarudin, M.A. Ameen and Zafril Rizal M. Azmi	
51	A Survey of Petri Net Tools	537
	Weng Jie Thong and M.A. Ameen	
52	Towards a Exceptional Distributed Database Model for Multi DBMS	553
	Mohammad Hasan Ali and Mohd Azlishah Othman	
53	Semantic Search Engine Using Natural Language Processing	561
	Sudhakar Pandiarajan, V.M. Yazhmozhi and P. Praveen kumar	
54	Integration of Mobile Based Learning Model Through Augmented Reality Book by Incorporating Students Attention Elements	573
	Zarwina Yusoff, Halina Mohamed Dahlan and Norris Syed Abdullah	
55	Law Reckoner for Indian Judiciary: An Android Application for Retrieving Law Information Using Data Mining Methods	585
	S. Poonkuzhali, R. Kishore Kumar and Ciddarth Viswanathan	
56	Enhancing the Efficiency of Software Reliability by Detection and Elimination of Software Failures Through Univariate Outlier Mining	595
	S. Poonkuzhali, R. Kishore Kumar and R. Kumar	
57	A Survey on the Application of Robotic Teacher in Malaysia	605
	Noraidah Blar and Fairul Azni Jafar	
58	A Novel Method for Distributed Image Steganography	615
	Bismita Choudhury, Rig Das and Themrichon Tuithung	
59	An Efficient Beam Scanning Algorithm for Hidden Node Collision Avoidance in Wireless Sensor Networks	627
	Moorthy Sujatha and Raghuvel Subramaniam Bhuvaneshwaran	

60	Evaluation of Stereo Matching Algorithms and Dynamic Programming for 3D Triangulation	641
	Teo Chee Huat and N.A. Manap	
61	Image Enhancement Filter Evaluation on Corrosion Visual Inspection	651
	Syahril Anuar Idris and Fairul Azni Jafar	
62	A Framework for Sharing Communication Media in Supporting Creative Task in Collaborative Workspace	661
	Norzilah Musa, Siti Z.Z. Abidin and Nasiroh Omar	
63	Joint Torque Estimation Model of sEMG Signal for Arm Rehabilitation Device Using Artificial Neural Network Techniques	671
	M.H. Jali, T.A. Izzuddin, Z.H. Bohari, H. Sarkawi, M.F. Sulaima, M.F. Baharom and W.M. Bukhari	
64	Enhancement of RSA Key Generation Using Identity	683
	Norhidayah Muhammad, Jasni Mohamad Zain, M.Y.M. Saman and Mohd Fadhil Ramle	
65	Rules Mining Based on Clustering of Inbound Tourists in Thailand	693
	Wirot Yotsawat and Anongnart Srivihok	
66	Designing a New Model for Worm Response Using Security Metrics	707
	Madiah Mohd Saudi and Bachok M. Taib	
67	Neural Network Training Algorithm for Carbon Dioxide Emissions Forecast: A Performance Comparison	717
	Herrini Mohd Pauzi and Lazim Abdullah	
68	Theorem Prover Based Static Analyzer: Comparison Analysis Between ESC/Java2 and KeY	727
	Aneesa Saeed and S.H.A. Hamid	
69	Designing a New Model for Trojan Horse Detection Using Sequential Minimal Optimization	739
	Madiah Mohd Saudi, Areej Mustafa Abuzaid, Bachok M. Taib and Zul Hilmi Abdullah	

70	An Access Control Framework in an Ad Hoc Network Infrastructure.	747
	Tanya Koohpayeh Araghi, Mazdak Zamani, A.A. Manaf and Sagheb Kohpayeh Araghi	
71	Enhancement of Medical Image Compression by Using Threshold Predicting Wavelet-Based Algorithm.	755
	N.S.A.M. Taujuddin and Rosziati Ibrahim	
72	Detection and Revocation of Misbehaving Vehicles from VANET	767
	Atanu Mondal and Sulata Mitra	
73	A Novel Steganalysis Method Based on Histogram Analysis	779
	Bismita Choudhury, Rig Das and Arup Baruah	
74	Pattern Recognition Techniques: Studies on Appropriate Classifications.	791
	Sasan Karamizadeh, Shahidan M. Abdullah, Mazdak Zamani and Atabak Kherikhah	
75	Environmental Noise Analysis for Robust Automatic Speech Recognition.	801
	N. Sai Bala Kishore, M. Rao Venkata and M. Nagamani	
76	Performance Comparison of Selected Classification Algorithms Based on Fuzzy Soft Set for Medical Data	813
	Saima Anwar Lashari and Rosziati Ibrahim	
77	A Hybrid Selection Method Based on HCELFS and SVM for the Diagnosis of Oral Cancer Staging.	821
	Fatihah Mohd, Zainab Abu Bakar, Noor Maizura Mohamad Noor, Zainul Ahmad Rajion and Norkhafizah Saddki	
78	A Linear Assignment Method of Simple Additive Weighting System in Linear Programming Approach Under Interval Type-2 Fuzzy Set Concepts for MCDM Problem	833
	Nurnadiah Zamri and Lazim Abdullah	

79 Hybridization Denoising Method for Digital Image in Low-Light Condition 843
 Suhaila Sari, Sharifah Zahidah Hasan Al Fakkri,
 Hazli Roslan and Zarina Tukiran

80 The Improved Models of Internet Pricing Scheme of Multi Service Multi Link Networks with Various Capacity Links . . . 851
 Fitri Maya Puspita, Kamaruzzaman Seman and Bachok M. Taib

81 Improving the Models of Internet Charging in Single Link Multiple Class QoS Networks 863
 Irmeilyana Saidi Ahmad, Indrawati, Fitri Maya Puspita
 and Lisma Herdayana

82 A New Aggregating Phase for Interval Type-2 Fuzzy TOPSIS Using the ELECTRE I Method 873
 Nurnadiah Zamri and Lazim Abdullah

83 The Role of Green IT and IT for Green Within Green Supply Chain Management: A Preliminary Finding from ISO14001 Companies in Malaysia 883
 K.S. Savita, P.D.D. Dominic and Kalai Anand Ratnam

84 Integrating e-Learning with Radio Frequency Identification (RFID) for Learning Disabilities: A Preliminary Study 895
 Wan Fatin Fatihah Yahya, Noor Maizura Mohamad Noor,
 Mohd Pouzi Hamzah, Mohamad Nor Hassan,
 Nur Fadila Akma Mamat and Mohd Arizal Shamsil Mat Rifin

85 Palmprint Identification Using Invariant Moments Algorithm Based on Wavelet Transform 905
 Inass Shahadha Hussein and M.J. Nordin

86 Auto Mobile Ad Hoc Mechanism in Delay Tolerant Network 915
 Muhammad Affandy Azman, Sharifah Hafizah Syed Ariffin,
 Norsheila Fisal, Mazlan Abbas, Mohd Husaini Mohd Fauzi
 and Sharifah K. Syed-Yusof

87 An Exploratory Study on Blind Users’ Mental Model in Computer Accessibility 925
 Manoranjitham Muniandy and Suziah Sulaiman

88	Case-Based Reasoning and Profiling System for Learning Mathematics (CBR-PROMATH)	939
	Nur Azlina Mohamed Mokmin and Mona Masood	
89	What Is the Influence of Users' Characteristics on Their Ability to Detect Phishing Emails?	949
	Ibrahim Alseadon, M.F.I. Othman and Taizan Chan	
90	Adaptive and Dynamic Service Composition for Cloud-Based Mobile Application	963
	R. Kanesaraj Ramasamy, Fang-Fang Chua and Su-Cheng Haw	
91	Web Service Composition Using Windows Workflow for Cloud-Based Mobile Application	975
	R. Kanesaraj Ramasamy, Fang-Fang Chua and Su-Cheng Haw	
92	An Effective Image Retrieval Method Based on Fractal Dimension Using Kernel Density Estimation	987
	Zhang Qin, Huang Xiaoqing and Liu Wenbo	
93	Bio Terapi Solat: 3D Integration in Solat Technique for Therapeutic Means	1001
	Arifah Fasha Rosmani, Noor Azura Zainuddin, Siti Zulaiha Ahmad and Siti Zubaida Ramli	
94	Enhanced Interactive Mathematical Learning Courseware Using Mental Arithmetic for Preschool Children	1013
	Siti Zulaiha Ahmad, Noor Asmaliyana Ahmad, Arifah Fasha Rosmani, Umi Hanim Mazlan and Mohammad Hafiz Ismail	
95	Comparative Evaluation of Ensemble Learning and Supervised Learning in Android Malwares Using Network-Based Analysis	1025
	Ali Feizollah, Nor Badrul Anuar, Rosli Salleh and Fairuz Amalina	
96	Tailored MFCCs for Sound Environment Classification in Hearing Aids	1037
	Roberto Gil-Pita, Beatriz López-Garrido and Manuel Rosa-Zurera	
97	Metamodelling Architecture for Modelling Domains with Different Mathematical Structure	1049
	Vitaliy Mezhujev	

98	Use Case Based Approach to Analyze Software Change Impact and Its Regression Test Effort Estimation	1057
	Avinash Gupta, Aprna Tripathi and Dharmendra Singh Kuswaha	
99	A Review of Image Segmentation Methodologies in Medical Image	1069
	Lay Khoon Lee, Siau Chuin Liew and Weng Jie Thong	
100	The Utilization of Template Matching Method for License Plate Recognition: A Case Study in Malaysia.	1081
	Norazira A. Jalil, A.S.H. Basari, Sazilah Salam, Nuzulha Khilwani Ibrahim and Mohd Adili Norasikin	

Part I
Communication

Chapter 1

RWA: Comparison of Genetic Algorithms and Simulated Annealing in Dynamic Traffic

Arturo Rodriguez, Antonio Gutierrez, Luis Rivera and Leonardo Ramirez

Abstract Modern telecommunications are supporting every day a progressive demand for services, which in turn generates greater requirements from the attention capacity in photonic transport networks. This phenomenon forces us to improve the routing systems, to minimize the blocking probability and minimize the use of the network, among other indicators, in order to attend current demand and to have the capacity to attend future demand. This paper compares four studies on routing and wavelength assignment with the aim of supporting the improvement of the already mentioned indicators. A comparison is made between optimizing algorithms and heuristic simulated annealing and genetic algorithms, using comparative indicators such as blocking probability and the use of the network. The results show that the heuristic algorithms are potentially better for a high load dynamic demand (greater than 120 erlangs) that would function much better under stress. GINT proposes genetic algorithms as a solution to the coming future demand of data transport.

Keywords Simulated annealing · Genetic algorithm · NSFNET · Wavelength

A. Rodriguez (✉) · A. Gutierrez · L. Rivera
Department of Industrial Technology, Research Group New Technologies (GINT),
Universidad Santiago de Chile, Santiago, Chile
e-mail: arturo.rodriquez@usach.cl

A. Gutierrez
e-mail: antonio.gutierrez@usach.cl

L. Rivera
e-mail: luis.rivera@usach.cl

L. Ramirez
Division of Technology Development and Innovation Research Group on Telemedicine
(TIGUM), Universidad Militar de Nueva Granada, Bogotá, Colombia
e-mail: leonardo.ramirez@unimilitar.edu.co

1.1 Introduction

Currently, in the field of fiber optics, the discussion is centered on commutation, which is preferable because of its high response speed to routing. The closeness of the upper layers of OSI to the linkage layer has triggered the reflection of the permanence of the IP layer, but not of the IP address. In any case, the search to establish routes that determine a fast and secure connection should be solved through different algorithms found in the literature or innovative algorithms proposed by researchers, such as those of Dijkstra and Floyd-Warshall among others, however those optimizing algorithms are not efficient with respect to optical demand, which requires good routes rather than optimum routes, due to the need to decrease the probability of blocking the network and improving its use.

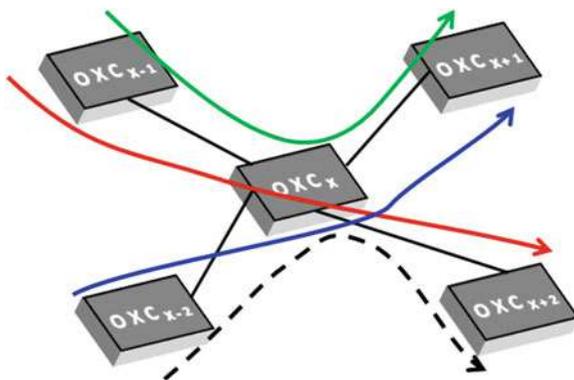
The traffic supported by a network can be classified as static and dynamic [1]. In the optical transport networks the traffic was of the static type because the demand was not sufficient to congest the system; but the static characteristics of traffic have turned toward dynamic traffic, and the RWA (Routing Wavelength Assignment) problem has become important, requiring the selection of a route and a wavelength to establish the connection, under the restriction that there should be no change in wavelength along the chosen route, which is known as CCW (Continuity Constraint Wavelength). This restriction helps to improve the delay on the route of present demand [2], but the use of the network increases, increasing the probability of blocking the future demand, because the more routes are established, the scarcer the roads become.

In general, this problem has been solved by subdividing it into two parts: The first one solves the route to be followed based on minimizing some pre-established condition [3, 4], while the second part solves the problem of assigning the wavelengths. There are other proposals that solve the problem from an integral perspective, i.e., without subdividing it [5]. The present paper makes a comparison of the two methods from the heuristic standpoint [6, 3].

1.2 Description of the RWA Problem

In All Optical Networks (AON) the transmitters are connected with each receiver at all wavelengths, and therefore as many signals are required as there are wavelengths. When many sessions use the same wavelength, these signals cannot be sent simultaneously along the same fiber to avoid collisions within the network, consequently blocking the session. It must be ensured that if a receiver is awaiting data at a given wavelength in a given time interval, only one signal at that wavelength must get to the receiver; if two or more signals arrive under those conditions, we face a phenomenon called contention, which can be avoided by isolating the signals in space and/or in time, i.e., by sending the signal along another road and/or at another time. Temporary isolation is in general what we are trying to avoid, because it causes latency.

Fig. 1.1 Example of contention in optical networks



An optical network is formed by OXC (Optical Cross Connect) or optical commutators linked by optical fibers. Depending on the technology of these commutators, they can commute fibers, wavelengths, wavelength bands, time division multiplexing (TDM). The optical routes are formed by optical links that direct the light beam (data) through the same wavelength according to the wavelength continuity restriction, but this possibility is exhausted when the demand exceeds the number of possible sessions; when this restriction cannot be satisfied, the request is blocked. Routing systems reusing the wavelength, where it is allowed to be changed, have been developed, however the studies do not show significant changes in routing performance [7]. Figure 1.1 shows 5 OXC that are part of a larger system. The established routes with their corresponding wavelengths (colors) are $(x-1, x, x+2/\text{Red})$; $(x-1, x, x+1/\text{Green})$ and $(x-2, x, x+1/\text{Light blue})$; when a service request arrives whose solutions goes through the $(x-2, x, x+2)$ links as shown in blue color, the specified route cannot be assigned the red wavelength because it is being used in the $(x, x+2)$ link, so either another wavelength should be used or a new route should be found. This problem is called contention.

We must define some variables to formulate the problem. We know that in the network there is current traffic and arriving or requested traffic. Traffic can be static or dynamic, depending on the existing relations between the current and arriving traffic. It is dynamic when the average connection time of current traffic is much less than the average time between services request arrivals. This scenario is not optimizable, so heuristic algorithms are used, most of which offer greater solutions that are not necessarily optimum, but this richness provides much support when the contentions problem arises, avoiding a new execution of the algorithmic process. The problem is to establish the routes and the wavelength assignments in the network for various criteria, with minimum probability of blocking the requests, low transport latency along the route, low level of jitter, etc. The problem is known as RWA, Routing Wavelength Assignment (routing and wavelength assignment). Different strategies have been tested to satisfy the demand of an optical network, and optimization criteria and algorithms have been used [8]. The optimization criteria have been based on:

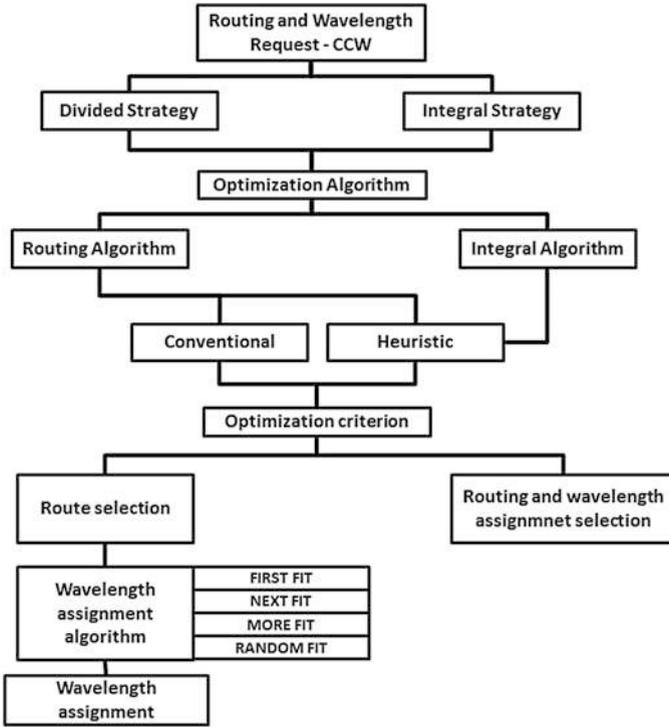


Fig. 1.2 RWA-CCW solution strategies

- Minimizing the probability of blocking new requests [5].
- Minimizing the use of the network's wavelengths [9, 10, 11].
- Minimizing the use of wavelengths per link [12, 13].
- Minimizing dispersion on the route [14].
- Minimizing the ASE (Amplified Spontaneous Emission) [15]

The optimizations algorithms can be classified into conventional, which are algorithms used in electronic routing or only as an initial population value for heuristic algorithms, such as minimum cost roads, dijkstra; incremental cost roads, bellman fulkerson; minimum delay roads, bellman ford; and Linear Programming [16], and heuristics, which are widely used algorithms in the search for good routes and not necessarily optimum routes, such as Ant [17, 18], Genetic [19, 6], Tabu Search [20], and Simulated Annealing [3, 4]. The diagram in Fig. 1.2 shows the processes located in the literature for solving the RWA problem. It is seen that they go from using problem division strategies until it is completely solved, making use of different criteria and algorithms that allow getting a route and a wavelength that allow the information to be transported. It should be stressed that in optical

networks the aim is not necessarily to optimize, but what is sought is rather a route that will work without generating a substantial increase of the probability of blocking and without enlarging the network.

1.3 Model Description

In this demand model the nodes N participate in the routing request from the network with a probability p_l (for the l th access node connected with the l th optical node X_l). This is the routing request probability that will be used as a simulation parameter. The demand M_l of the l th access node follows a Poisson distribution with an average rate μ_e .

$$P_{(M_l=m)} = \frac{\mu_e^m}{m!} e^{-\mu_e} \quad (1.1)$$

With this model we can have a varied range of requests around the whole network and we can vary the load intensity at all the nodes, and the demand can be changed from static to dynamic according to the study that it is desired to make. The comparative work was done under dynamic demand.

1.4 Description of the Algorithms

The selected strategies are to divide the problem into two parts or to use it wholly. Simulated annealing was used in the first case, and genetic Algorithms in the second case, both of them carried out previously by the Grupo de Investigacion de Nuevas Tecnologias (GINT) Industrial Technology Department, Facultad Tecnol6gica of the Universidad Santiago de Chile.

1.4.1 Network Demand

Requests arrive at the boundary nodes (with Poisson behavior, but different probabilistic scenarios could be studied). These requests bring along four parameters that must be satisfied, otherwise the request will be blocked.

$$v_i^s = (r_o, r_D, n_c, t_c) \quad (1.2)$$

Where:

v_i^s is the vector that represents the i th request that arrives at the s th boundary node (Edge Router)

r_O is the identification number of the node of origin
 r_D is the identification number of the destination node of the arrival request
 n_C is the number of connections requested for the (rO,rD) pair
 t_C is the time of connections requested for the (rO,rD) pair

Three matrices have been established: The first linking matrix, C, that will always be setting the available capacity of the links, so it can be used to monitor when a link is not available; the second wavelength matrix, λ , which determines the wavelength in use; and the third, the time matrix T, is a matrix that will have the function of keeping count of the connection time of each wavelength in each link, and must be updated dynamically. Its structure is similar to that of matrix λ , but its elements will have the value 1 when the wavelengths are available in the given link, and the time will be negative when it is in use, and it will decrease as time goes by, and in this way it can be detected in the aptitude functions FA as not to use it. The definitions of the matrices are given below [21, 19, 6].

$$c = \left\{ \begin{array}{l} c_{ij} / (c_{ij} = 0 \ \forall \ i = j \ \wedge \ c_{ij} = G \ NEE \\ \wedge \ c_{ij} = g \ \forall \ i \neq j \ SEE \ \wedge \ j \in [0, N - 1] \\ \wedge \ k \in [0, n_w - 1] \end{array} \right\} \quad (1.3)$$

$G =$ Very large number
 $g =$ Instantaneous cost of the link
 $NEE =$ There is no link
 $SEE =$ There is a link

$$\lambda = \left\{ \begin{array}{l} \lambda_{ijk} / (\lambda_{ijk} = 0 \ LOU \ \vee \ NEE) \ \vee \\ (\lambda_{ijk} = k + 1 \ LOD) \ \forall \ i \in [0, N - 1] \\ \wedge \ j \in [0, N - 1] \ \wedge \ k \in [0, n_w - 1] \end{array} \right\} \quad (1.4)$$

$n_w =$ Number of wavelengths in the network

$$T = \left\{ \begin{array}{l} t_{ijk} / (t_{ijk} = -t_c \ LOU \ t_{ijk} = 0 \ NEE) \ \vee \\ (t_{ijk} = 1 \ LOD) \ \forall \ i \in [0, N - 1] \\ \wedge \ j \in [0, N - 1] \ \wedge \ k \in [0, n_w - 1] \end{array} \right\} \quad (1.5)$$

$LOU =$ Wavelength in the link is being used
 $LOD =$ Wavelength is not being used

Fig. 1.3 Initial population matrix

$$S_0 = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{matrix} & \begin{bmatrix} 1 & 3 & 9 & 3 & 2 & 2 & 5 & 6 & 9 & 4 \\ 1 & 4 & 9 & 2 & 6 & 3 & 4 & 4 & 2 & 4 \\ 2 & 1 & 1 & 1 & 3 & 4 & 5 & 3 & 7 & 1 & 4 \\ 3 & 1 & 4 & 2 & 7 & 3 & 2 & 4 & 5 & 8 & 4 \\ 4 & 1 & 5 & 9 & 7 & 5 & 1 & 2 & 1 & 9 & 4 \\ 5 & 1 & 6 & 8 & 5 & 7 & 4 & 8 & 6 & 1 & 4 \\ 6 & 1 & 7 & 8 & 2 & 3 & 6 & 9 & 7 & 6 & 4 \\ 7 & 1 & 3 & 2 & 8 & 5 & 3 & 2 & 4 & 7 & 4 \end{bmatrix} \end{matrix}$$

1.4.2 Simulated Annealing Algorithm

Let S_0 be the two-dimensional matrix of order $m \times N$, where m is the number of rows of the matrix, which indicates the population sample used, in addition to being a simulations variable, while N indicates the number of nodes.

$$S_i = \left\{ \begin{matrix} s_{xy}/s_{xy} = a \wedge a \in [0, N - 1] \\ \wedge x \in [0, m - 1] \wedge y \in [0, N - 1] \\ \forall N, m = 2k \wedge k \in \mathbb{Z}^+ \end{matrix} \right\} \tag{1.6}$$

Futhermore:

$$s_{0y} = r_O \wedge s_{xN-1} = r_D \quad \forall x, y \tag{1.7}$$

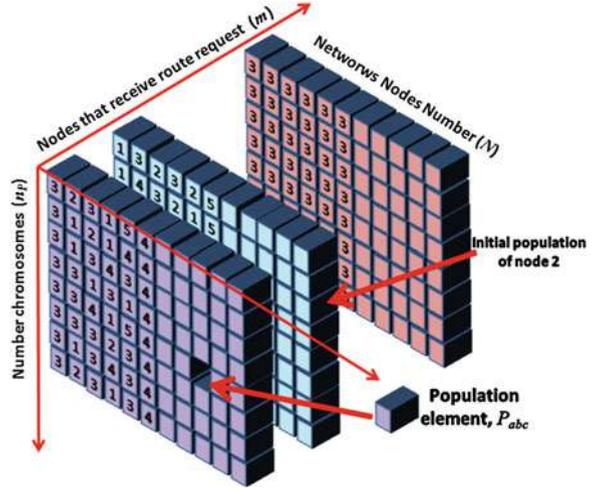
When a request arrives at node 1, matrix S_0 is established; if it arrives at node 2, matrix S_1 is established, and so on. When $v_{15}^0 = (1, 4, 3, 128)$ arrives, we read: 15° service request entering node 1, which requests route from node 1 to node 4 with three 128 ms. connections. Therefore, according to (1.8), the first column has the origin node and the last column has the destination node. The rest of the elements of the initial population are filled randomly. For example, (Shown Fig. 1.3), for $m = 8$ and $N = 10$, and the request for connection from node 1 to node 4.

Then rotation of the internal layers at different speeds must take place to find the routes. Also, the wavelength assignments were made with First Fit, to be able to compare them under the same conditions in terms of stopping criteria, aptitude functions, and other parameters of the algorithm. They can be seen in detail in [3, 4].

1.4.3 Genetic Algorithms

Let P_0 be the initial population of the genetic algorithm to be used, and n_p the number of chromosomes that constitute these initial populations. There will also be “ m ” genetic algorithms for the “ m ” nodes that are requesting transport service.

Fig. 1.4 3D matrix of the initial population



Within the genetic algorithms there are various implementations that allow different behaviors of n_P (regeneration policies); in this research this value remains constant and it is a simulation parameter. In this way, the three-dimensional matrix P_0 of $n_P \times (N + 3n_W + 2) \times N$ elements is defined.

$$P_0 = \left\{ \begin{array}{l} p_{abc}/p_{abc} = y \wedge a \in [0, n_p - 1] \\ \wedge b \in [0, N + 3n_W + 2] \wedge c \in [0, m - 1] \\ \vee y \in [0, N - 1] \end{array} \right\} \quad (1.8)$$

$$P_{0bc} = r_O \wedge P_{a0c} = r_D \quad \forall a, b, c \quad (1.9)$$

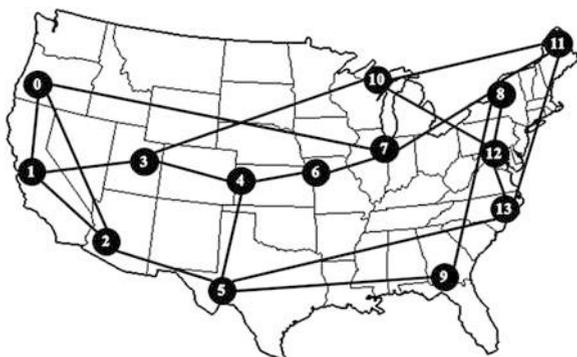
This matrix is filled randomly, the additional columns are for saving the calculations of the aptitude functions referring to each available wavelength. La Fig. 1.4 shows the 3D matrix of the initial populations that will serve for the genetic process. For example, Fig. 1.4 shows the initial population of a network with 6 ($m = 6$) nodes and 8 ($n_p = 8$) chromosomes without intergenerational numerical change, and in Fig. 1.5 each commuter is a gene that constitutes the chromosome, whose information is the route that is sought. For example, in row 4 the chromosome is 2-3-3-2-1-4, where the first and the last genes are the origin and the destination. The rest of the genes are filled randomly in the population matrix. Assuming that the network has two wavelengths ($n_W = 2$), in Fig. 1.5, we would have one of six genetic algorithms with “a” from 0 to 9, “b” from 0 to 5 and “c” from 0 to 5.

If the request that arrives at node 2 is $v_{15}^0 = (2, 4, 3, 128)$, then the matrix would be:

Fig. 1.5 2D matrix of the initial population of node 2, belonging to P0

		0	1	2	3	4	5	6	7	8	9	10	
	0	2	1	2	2	3	4						COMPUTATIONAL USE ZONE
	1	2	3	1	1	2	4						
	2	2	4	3	3	1	4						
	3	2	2	4	4	2	4						
	4	2	3	3	2	1	4						
	5	2	2	4	4	3	4						
	6	2	1	5	5	4	4						
	7	2	2	3	3	1	4						
	8	2	5	2	2	2	4						
	9	2	2	3	1	1	4						

Fig. 1.6 NSFNET network used with the four methods



Then the algorithm is executed to find the routes; as to the stopping criteria, aptitude functions and other parameters of the algorithm, they can be seen in detail in [21, 19, 6].

1.5 Comparison Scenario

Once the simulation had been made, it was compared with two similar reports under the same simulation conditions.

The network used for the test was the NSFNET (National Science Foundation NETWORK), which has 16 nodes and 25 fiber optics links (Shown Fig. 1.6). The parameters used were similar to those presented in [1]. Comparisons were made of probability of blocking and use of the network, varying the load in the [0,180] interval, with 10-erlang increments. The number of connections made during the simulation in both scenarios, simulated annealing (SA) and genetic algorithms (GA), was 10^8 connection requests.

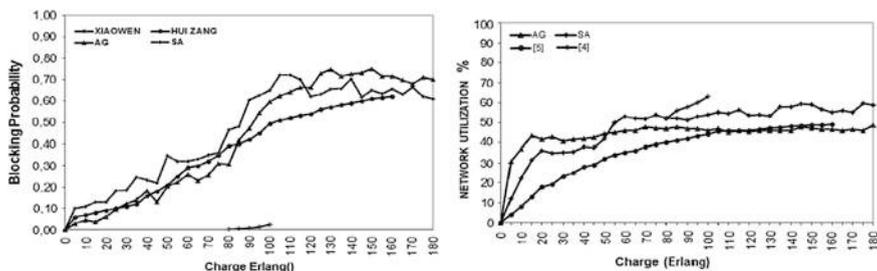


Fig. 1.7 Comparison of blocking probability and network use in the NSFNET

1.6 Results

Comparisons were made of the simulated annealing [3], genetic algorithms [19], of Dr. Xiao Wen [2], and of Dr. Hui Zang [1] simulations with the blocking probability indicator, which measures the probability of blocking the service requests under a given load, and the network use indicator, which measures network occupation under a given load. Figure 1.7 graphs the blocking probability, showing that up to 120 Erlangs the GA has better performance than SA, but above 120 Erlangs the SA is better than GA, but in no case are they better than that of Hui Zang; the excellent performance of the Xiao Wen algorithm is worth noting, but it is because 40 wavelengths are used, compared to eight wavelengths that were used in the other three studies. Figure 1.7, graphs the percentage use of the network, comparing the network use with the previously mentioned reports. The SA algorithm improves the network use for loads smaller than 50 Erlangs, but for greater loads the GA algorithm uses fewer network resources. But, under 110 Erlangs the SA and GA algorithms are not better than that of [1], but the GA algorithm is better for loads greater than 110 Erlangs. Note that work with SA and GA reaches 180 Erlangs, while the other reports involve simulations with lower load intervals. On the other hand, the performance of the algorithm reported in [2] involves high network consumption, even with many wavelengths, generating a problem to satisfy future load demand.

1.7 Conclusions

The work reported in [2, 1] is always aimed at optimum solutions, while the work based on SA and GA looks for solutions without the need to go over the whole possible universe, thereby saving a large amount of operating time. From the results it can be concluded that heuristic algorithms have great potential for future demand because they function much better at high loads and small dynamic traffic, and the latter should become very important because it allows the use of the

network with an optimum sense. This research team agrees that it will be necessary to develop the re-use of wavelengths to be able to remove the restrictions given in CCW and improve the results while better proposals are developed.

References

1. Zang, H., Jue, J.: Dynamic lightpath establishment in wavelength-routed WDM networks. *IEEE Commun. Mag.* **39**, 100–108 (2001)
2. Chu, X., Li B., Zhang, Z.: A dynamic RWA algorithm in a wavelength-routed all-optical network with wavelength converters, INFOCOM 2003. Twenty-second annual joint conference of the IEEE computer and communications, vol. 3, pp. 1795–1804. IEEE Societies (2003)
3. Rodríguez, A., Saavedra, F., Ramírez, L.: Simulated Annealing una Propuesta de Solución al Problema RWA en Redes Fotónicas. IEEE Intercon UNI, Lima-Perú (2011)
4. Guenduez, H., Kadir, H.: A well-arranged simulated annealing approach for the location-routing problem with time windows. In: 46th Hawaii International Conference on System Sciences (HICSS), pp. 1144–1153 (2013)
5. Liu, L., Xue, Q., Zhang, J., Wang, F., Wu, J., Tong, Lin, J., Bin, L.: Investigation of uniform models for analysing network blocking probability in ASON, communications, networking in China, 2006. ChinaCom '06. First International Conference on, pp. 1–4 (2006)
6. Rodríguez, A., Saavedra, F., Ramírez, L.: Solución Simultánea Del Enrutamiento Y Asignación De Longitud De Onda en redes WDM Con Algoritmos Genéticos. IEEE COLCOM, Bogota, Colombia (2008)
7. Ramaswami, R.: Optical networking technologies: what worked, what didn't. *IEEE Commun. Mag.* **44**, 132–139 (2006)
8. Balasis, F., Wang, X., Xu, S., Tanaka, Y.: A dynamic physical impairment-aware routing, wavelength assignment scheme for 10/40/100 Gbps mixed line rate wavelength switched optical networks. In: 15th International Conference on Advanced Communication Technology (ICACT), 2013, pp. 116–121 (2013)
9. Zang, H., Jue, J., Mukherjee, B.: A review of routing, wavelength assignment approaches for wavelength-routed optical WDM networks. *Opt Netw. Mag.* **1**, 47–60 (2000)
10. Ramaswami, R., Sivarajam, K.: Design of logical topologies for wavelength-routed optical networks. *IEEE J. Sel. Areas Commun.* **14**, 840–851 (1996)
11. Wang, J., Qi, X., Chen, B.: Wavelength assignment for multicast in all optical WDM networks with splitting constraints. *IEEE/ACM Trans. Networking* **14**:169–182 (2006)
12. Kuri, J., Puech, N., Gagnaire, M., Dotaro, E., Douville, R.: Routing, wavelength assignment of scheduled lightpath demands. *IEEE J. Sel. Areas Commun.* **21**, 1231–1240 (2003)
13. Saengudomlert, P., Modiano, E., Gallager, R.: Dynamic wavelength assignment for WDM all-optic tree networks. *IEEE/ACM Trans.* **13**, 895–905 (2005)
14. Gurzi, P., Steenhaut, K., Nowe, A.: Minimum cost flow based R&WA algorithm for dispersion, OSNR limited all-optical networks. In: 15th International Conference on Optical Network Design, Modeling (ONDM), 2011, pp. 1–6 (2011)
15. Feres, M., Trevelin, L.: RWA algorithm aware of PMD, ASE for all-optical networks. In: 11th International Conference on Transparent Optical Networks, 2009. ICTON '09, pp. 1–4 (2009)
16. Ozdaglar, E., Bertsekas, D.: Routing, wavelength assignment in optical networks. *IEEE/ACM Trans. Networking* **11**, 259–272 (2003)
17. Zeng, P., Yu, H.: An ant-based routing algorithm to achieve the lifetime bound for target tracking sensor networks. In: IEEE International Conference on Communications, ICC '06, vol. 8, pp. 3444–3449 (2006)

18. Tan, S.: Ant-based Physical Attack, Amplifier Spontaneous Emission-aware routing. *Communication Technology (ICCT)*, 2012 IEEE 14th International Conference on, pp. 650–653 (2012)
19. Rodríguez, A., Saavedra, F.: Enrutamiento y asignación de longitudes de onda en redes WDM: Solución simultánea basada en algoritmos genéticos. *IEEE Intercon UNSA, Arequipa-Perú* (2009)
20. Assis, K., Ferreira, D., Giozza, W.: Hybrid algorithms for routing assignment wavelengths in optical networks. *Lat. Am. Trans.* **8**, 214–220 (2010)
21. Barpanda, R., Turuk, A., Sahoo, B., Majhi, B.: Genetic algorithm techniques to solve routing, wavelength assignment problem in wavelength division multiplexing all-optical networks. In: *Third International Conference on Communication Systems, Networks (COMSNETS)*, pp. 1–8 (2011)
22. Zang, H., Sahasrabudde, L., Jue, J., Ramamurthy, S., Mukherjee, B.: Connection management for wavelength-routed WDM networks, global telecommunications conference, 1999. *GLOBECOM '99*, vol. 2, pp. 1428–1432 (1999)

Chapter 2

Route Optimization in Proxy Mobile IPv6 Test-Bed via RSSI APPs

Nur Haliza Binti Abdul Wahab, L.A. Latif, S.H.S. Ariffin,
N. Fisal and N. Effiyana Ghazali

Abstract Proxy Mobile IPv6 is the new protocol, but there is a problem on Localized Routing (LR) algorithm which it is not ready build in PMIPv6 protocol. As for this in PMIPv6 in RFC5213 address the needed to enable Localized Routing (LR) but it not specify a complete procedure to establish Route Optimization (RO). RFC6279 state the problem statement on LR issue with several scenarios to tackle down. LR is important especially to minimize data delay and decrease handover latency due to the un-optimized data route. Data packets on PMIPv6 protocol without LR always need to travel to Local Mobility Anchor (LMA) with result in end-to-end delay. Therefore, this paper propose the new LR algorithm to optimized data route for selected scenarios to reduce handover and data packet delay. This paper also will illustrate the setting up of the PMIPv6 and test-bed performance.

2.1 Introduction

Users demand on mobility dedicate Internet Engineering Task Force (IETF) to designed a standard communications protocol named as Mobile IP or IP mobility to allow Mobile Node (MN) to maintain their permanent IP address while moving around from one network to another network.

Mobile IP for Internet Protocol version 4 (IPv4) addresses first described in IETF and documented in [1] while Mobile IPv6 (MIPv6) protocol was designed by IETF to support mobility implementation for the next IP generation of IPv6 [2].

N.H.B.A. Wahab · L.A. Latif (✉) · N. Fisal · N.E. Ghazali
UTM Razak School, Universiti Teknologi Malaysia, UTM, Kuala Lumpur, Malaysia
e-mail: liza@ic.utm.my

N.H.B.A. Wahab
e-mail: nurhaliza197@gmail.com

S.H.S. Ariffin
UTM MIMOS Center of Excellence, Universiti Teknologi Malaysia, UTM, Skudai,
Malaysia

Table 2.1 Differences between MIPv6 and Pmipv6

Protocol criteria	MIPv6	PMIPv6
Mobility scope	Global	Local
Location management	Yes	Yes
Required infrastructure	Home agent	Local mobility anchor (LMA), mobile access gateway (MAG)
MN modification	Yes	No
Handover latency	Bad	Good
Localized routing	Yes	No

Although MIPv6 enable a MN to move but this ability is still not sufficient for true mobility as MIPv6 requires the MN to modify its client functionality in the IPv6 stack. Plus, the need of MN to involve in IP related to signaling and enabling efficient handover is an additional and critical requirement thus IETF again introduces new protocol to tackle the issues occur in MIPv6. The new protocol known as Proxy Mobile IPv6 (PMIPv6). Table 2.1 shows the differentiation between MIPv6 and PMIPv6.

PMIPv6 provide new approach named as Network-Based Localized Mobility Management (NetLMM) which has been developed to relocate mobility procedures from mobile devices to network components to support in IP networks.

There are three main advantages [1, 2] of PMIPv6 and they are:-

- Handover performance optimization.
PMIPv6 can reduce latency in IP handovers by limiting the mobility management within the PMIPv6 domain. Therefore, it can largely avoid remote service which not only cause long service delays but consume more network resources.
- Reduction in handover-related signaling overhead.
The handover-related signaling overhead can be reduced in PMIPv6 since it avoids tunneling overhead over the air as well as the remote Binding Updates either to the HA or to the Correspondent Node (CN).
- Location privacy.
Keeping the mobile node's Home Address (MN-HoA) unchanged over the PMIPv6 domain dramatically reduces the chance that the attacker can deduce the precise location of the mobile node.

Although there are many advantages on PMIPv6 rather than MIPv6, there is a lack on PMIPv6 protocol. As state in Table 2.1, there are no Localized Routing (LR) protocol build-in in PMIPv6. Therefore [3] address the need to enable this LR features.

There are two main entities in PMIPv6. Local Mobility Anchor (LMA) and Mobile Access Gateway (MAG). The function of LMA is to act as the persistent HA for MN and as the topological anchor point for MN prefix assignments. LMA also

function as default multicast upstream for the corresponding MAG which manages the links to maintain the state of MN.

LMA is responsible to detect MN movements and change of attachment. Packets sent to and received from MN are routed via tunnels between LMA and the corresponding MAG. MAG performs the mobility-related signaling on behalf of MN which is attached to it. MAG acts as an Access Router for MN, which is the first-hop router in the Localized Mobility Management (LMM).

Figure 2.1 shows the basic operation of PMIPv6 [3] as the signaling information transverse when a Mobile Node connects to the PMIPv6 network. The signaling flow starts when a MN enters a PMIPv6 domain and MN attaches to an access link provided by a MAG.

MAG on that access link performs the access authentication process which it identifies MN, the LMA’s address and other configuration parameters. To update the LMA about the current location of MN, MAG sends Proxy Binding Update (PBU) to the LMA which also contain MAG address with identity of the MN. Upon receiving this request, the LMA allocates MN’s home network prefix and sends back to the MAG as Proxy Binding Acknowledgement. It also creates the Binding Cache entry and sets up its endpoint of the bi-directional tunnel to the MAG.

While MAG sets up its endpoint of bi-directional tunnel to the LMA, it also sets up the forwarding channel for MN’s traffic. MAG then sends Router Advertisement messages to MN on access link including the prefix allocates to MN. Upon receiving the Router Advertisement messages on the access link, MN attempts to

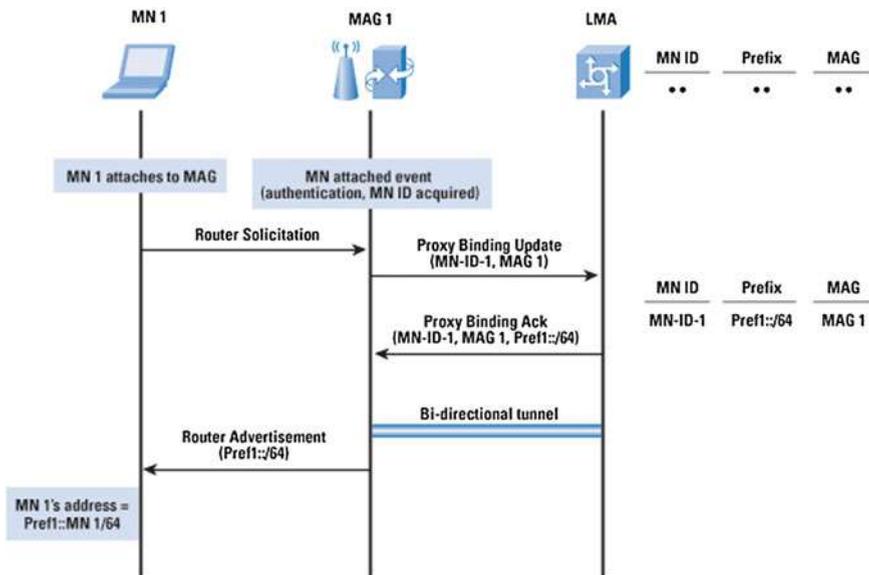
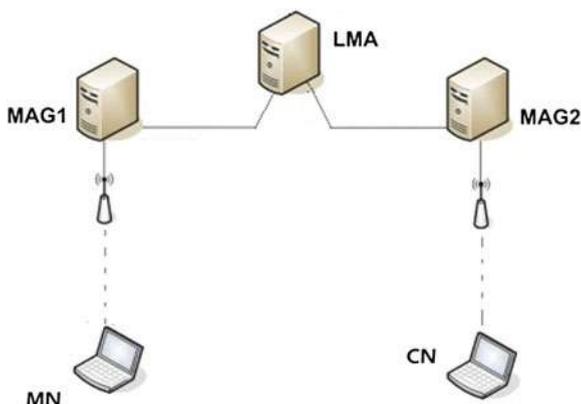


Fig. 2.1 Signaling when Mobile node connects to PMIPv6 network [3]

Fig. 2.2 Overview of scenario A21



configure an address using interface either stateful or stateless address configuration modes.

After obtaining the initial address configuration in PMIPv6 domain, if MN moves, the MAG updates the location of the MN and will signal the LMA and advertises the same prefix to MN. MN keeps the address configured when it first enter the LMA domain. This means that MAG configured the same link local address for specific MN.

PMIPv6 cannot directly apply the LR protocol in MIPv6 as MN is not involved in mobility signaling thus it cannot perform signaling to set up a RO [3, 4]. In PMIPv6, data packets have to always go through the LMA even if the CN and MN sit on the same network.

In [3], PMIPv6 protocol does not fully specify the procedure to establish such localization routing state. The problem specified in [4] which have 5 scenarios to be considered.

All scenarios refer to “A” cases as assumption that both the MN and the CN are registered with an LMA according to the PMIPv6 protocol [3, 94]. The scenarios represent as ‘A = (number of MAGs)(number of LMAs)’ which scenario A11, A21, A12, and A22 [4]. Scenario that has similar topology as in the office or indoor environment is scenario A21. Scenario A21 is scenario where there are two MAGs and one LMA in the network. Figure 2.2 shows the overview of scenario A21.

LR for scenario A21, needs the routing information at both MAGs so that data packet can directly forwarded between MN’s MAG and CN’s MAG. As LMA is the common anchor for the MN and the CN, LMA should maintains location information for both nodes.

The remainder of this paper is organized as follows: Section II will discuss related research work done in this field; Section III will describe the setup of the test-bed; Section IV will present the localized routing algorithm; Section V will present the results and discussion; and lastly Section V will conclude the paper.

2.2 Related Research

Over the past few years, there has been a huge interest in PMIPv6 research area after it was proposed to overcome the long registration delay problem. With PMIPv6, it can avoid tunneling overhead over the air and support the hosts without any involvement in the mobility management side.

Several related research has been proposed in this field such as in [9]. Authors in [9] proposed some enhancements for PMIPv6 as their studies show that PMIPv6 may cause high handover latency is the LMA is located far from the current MAG through simple mathematical model.

Paper [10] presented a new novel localized routing scheme called LRP. Jun et al. [10] claim that their proposed design has shorten the response time of LR compared with a classical protocol. An experimental evaluation on PMIPv6 under different implementation configurations and they evaluated the impact on the performances on PMIPv6 real test-bed.

Research journal in [11] also presented the real PMIPv6 test-bed implementation but it focus on Network Mobility (NEMO). Minoli [11] propoed a NEMO supporting scheme, which supports MN's mobility between PMIPv6 networks and mobile networks as well as the basic NEMO.

Iapichino and Bonnet [12] focused only hardware or real test-bed. Their test-bed OS used Ubuntu 9.04 and the kernel is 2.6.31. Software that used in [12] is OpenWRT Kamikaze 7.09. Linksys WRT54GL v1.1 is used as the Access Point in [12]. Iapichino and Bonnet [12] only compare their system performance result with previous work. Another related research is [13] where looking at improvement localized routing handover. Lee et al. [13] is a project which run the simulation without test-bed development.

2.3 Test-Bed Setup

The PMIPv6 test-bed architecture for this work is shown in Fig. 2.3. In this architecture, we can see that the PMIPv6 test-bed consists of several entities which three (3) computers, two (2) notebooks, two (2) access points (AP) and one (1) hub or router.

The three computers are needed to be setup as LMA, MAG1 and MAG2 while notebooks are represent and acted as MN and CN. The hub or router is needed to work as interfaces that connect LMA, MAG1 and MAG2. AP will acted as AP for MAG1 and MAG2 as well.

d_{APs} is the distance between AP1 and AP2. Communication or handover process are different and vary when d_{APs} change. An experimental are done with different d_{APs} to see the correlation between distance and handover latency.

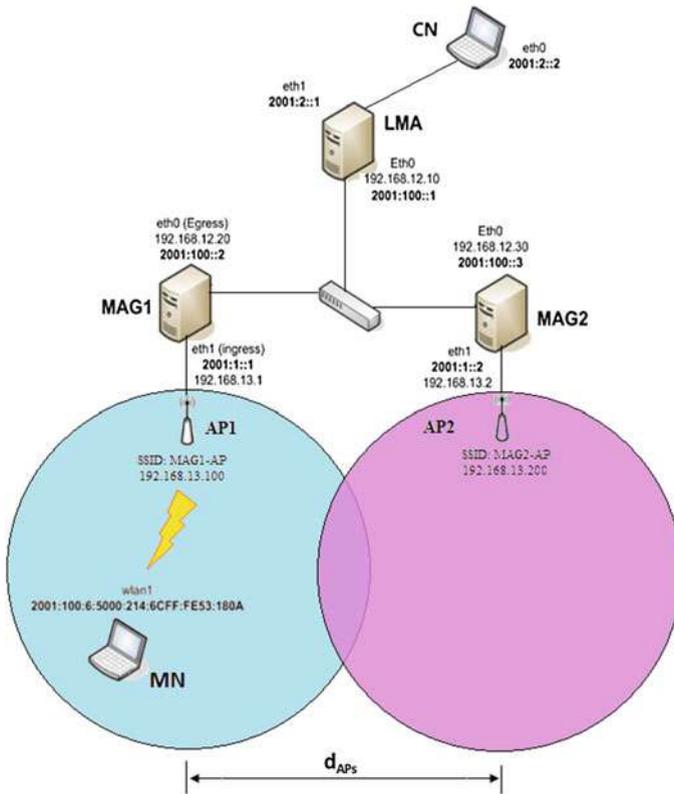


Fig. 2.3 Proxy mobile IPv6 architecture

2.3.1 Software and Hardware Requirement

The hardware specification for this PMIPv6 test-bed project are listed in Table 2.2. Specification shows all the IP address for every components in the PMIPv6 test-bed and their hardware details.

Hardware needed for LMA, MAG, MN and CN did not require specific type of processor, RAM, type of devices (computer or notebook) and more. But, for LMA and MAGs, the hardware require should use computer as it need more than two Ethernet card to be used to connect to AP, MN or CN.

The type of AP that used should be an AP that can be configure or modified their setting and program because the AP should be work on masquering mode to fulfill the PMIPv6 test-bed. In this project, Cisco Aironet 1200 Series AP are used which can be configure and modified to be used in masquering mode.

Table 2.2 Hardware specification

Device	Hardware specification	IP address
LMA	Processor :Intel [®] Core [™] i5-2310	eth0 : 2001:100::1
	Memory : 3.2 GiB	eth1 : 2001:2::1
	CPU :2.90 Ghz	
MAG1	Processor :Intel [®] Pentium [®]	
	DualCore [™] i5-2310	eth0 : 2001:100::2
	Memory : 3.2 GiB	eth1 : 2001:1::1
	CPU :2.90 Ghz	
MAG2	Processor :Intel [®] Pentium [®]	
	DualCore [™] i5-2310	eth0 : 2001:100::3
	Memory : 3.2 GiB	eth1 : 2001:1::2
	CPU :2.90 Ghz	
MN	Processor : Inter [®] Core [™] i3-M350	
	Memory : 1.8 GiB	wlan0 : 2001:100:6:5000:
	CPU : 2.27 GHz	214:6CFF:FE53:180A
CN	Processor :Intel [®] Core [™] 2 Duo Memory :1.9 GiB CPU :T6400@2.00 GHz	eth0 : 2001:2::2
AP	Cisco Aironet 1200 Series 2.4 GHz@5 GHz 54 Mbps	–

Several software used to build up this PMIPv6 test-bed project. The main software needed is Linux Open Source (OS) for the base system. The Linux type used in this work is Ubuntu 10.04 LTS that chosen to be used in all components (LMA, MAG1, MAG2, MN and CN) in this PMIPv6 test-bed.

As Linux was chosen to be as the platform system, the next important thing to look down is the kernel version for the Linux. In this project, the kernel version used is 2.6.32.59 + drm33.24 which a modification kernel from kernel version 2.6.32.

2.3.2 Kernel Setup and Installation

As mention previously, kernel version used in this PMIPv6 test-bed is 2.6.32.15 + drm.33.24. This kernel is modified version from kernel version 2.6.32. The kernel needed to enable several features to support PMIPv6 environment. After modified the kernel, several packages needed to be installed to fully run this PMIPv6 test-bed project.

2.3.3 Access Point Cisco Aironet 1200

Access Point (AP) that used in this project is Cisco Aironet 1200 Series, as mention in previous section. This AP needed to be configured to enable their SYSLOG client and acted in masquering mode. This is because, the AP should be a separator for the MAG function that connected to it.

2.4 Localized Routing Algorithm

The Route Optimization (RO) is Using Multicasting Binding Cache approach to MAGs method, add on RSSI APPs and Prediction method. RSSI APPs is the APPs that been installed to the MN so that MN will keep sending Signal Strength of every APs of MAGs detected and sent to the LMA.

LMA is a Dynamic LMA which it will calculate and compare all the Signal Strength given by MN via RSSI APPs and keep update the MAGs with all the RSSI using Multicasting Binding Cache approach. Since the LMA is in the local domain, it can always multicast its binding cache to all the MAGs to keep update cache ahead before any operation (Handover, data communication, etc.) take happen.

The Prediction method will change the data path route automatically when MAGs detect $AP-MAG1 = AP-MAG2$ or $AP-MAG1 > AP-MAG2$ for the first time follow to the topology. It's mean, before handover happen, the path route of data communication already change to the shorter path or known as Route Optimization (RO), without need to wait for the handover process to end to change the data path route.

The data travel to maximum path if there is no RO implementation. Data from CN will be sent to MAG1 to LMA, than back to MAG1 and lastly to MN ($CN \rightarrow MAG1 \rightarrow LMA \rightarrow MAG1 \rightarrow MN$), rather than with LR via RSSI APPs the data will be travel by shorter path which data from CN to MAG1 than direct to MN because they attach to the same MAG ($CN \rightarrow MAG1 \rightarrow MN$).

While MN moving, it keep sending the signal strength of both MAGs (MAG1 and MAG2) to LMA so that LMA are up to date of the MN location. As LMA which always sending BU to MAGs using multiple binding cache approach to MAGs will keep update the caches to MAGs before any operation (handover) happen.

The prediction method will change the data path when MAGs detect the signal strength of AP-MAG1 equal to signal strength of AP-MAG2 or when AP-MAG1 higher than signal strength AP-MAG2 ($AP-MAG1 \geq AP-MAG2$). This mean, before handover happen, the data path already change to the shorter path, without need to wait for the handover process to end to change the path flow. The data flow with LR via RSSI APPs and without LR shows in Fig. 2.4. In Fig. 2.4, we can see that data paths flow is the black arrows and the red arrow is the deleted or reduction path. The data flow with LR via RSSI APPs and without LR shows in Fig. 2.5.

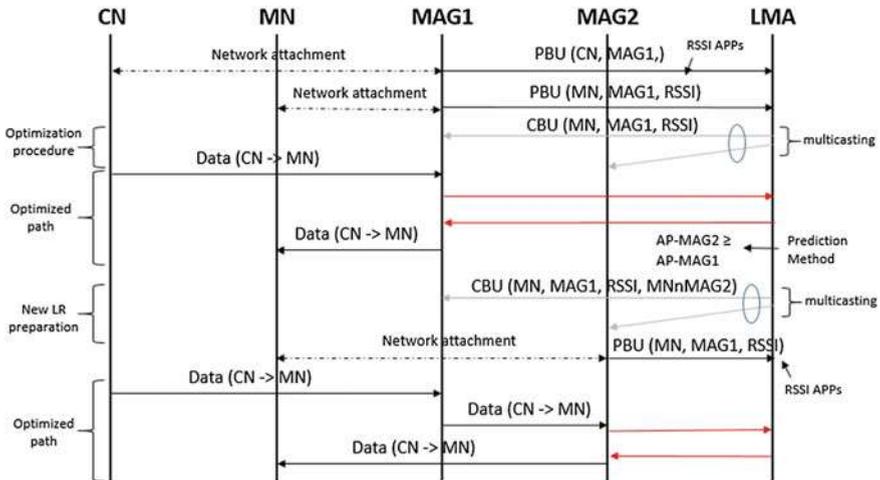


Fig. 2.4 Data flow with RO via RSSI APPs and without RO

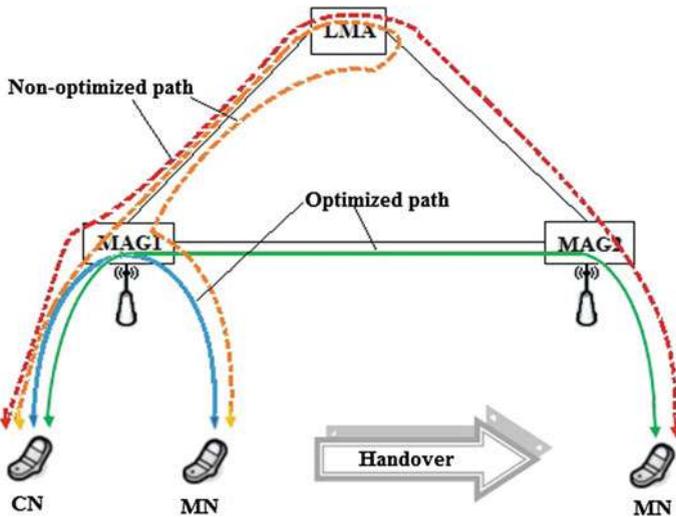


Fig. 2.5 Data flow with and without route optimization in PMIPv6

2.5 Result and Discussion

The aims of this project is to develop PMIPv6 test-bed that can be used for experimental work. By having this PMIPv6 test-bed, other work can be implement in this test-bed for enhancement especially as next work for this project is to look forward for handover and localized routing problem in PMIPv6.

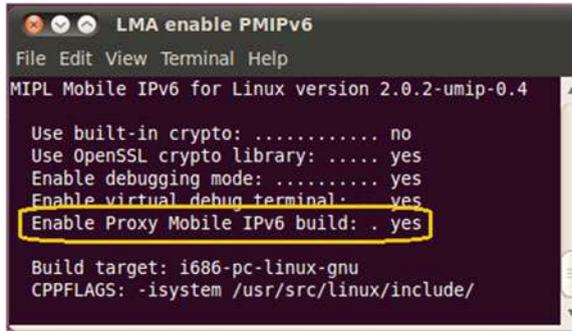


Fig. 2.6 Successfully enable PMIPv6 features

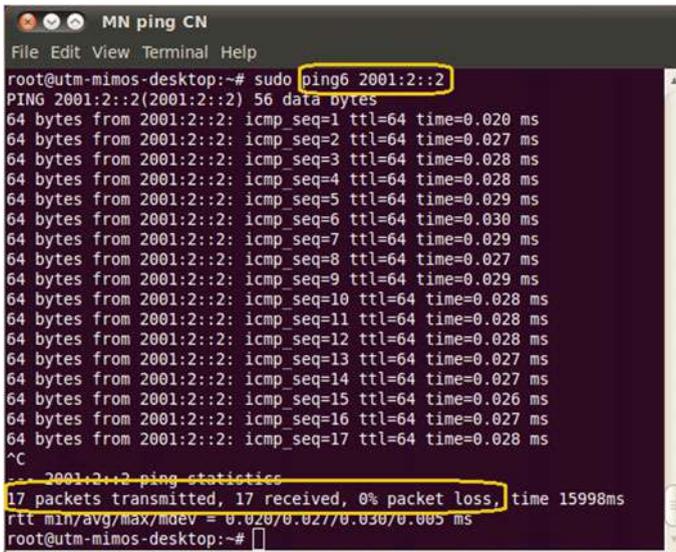


Fig. 2.7 MN successfully ping CN

To allow IPv6 addresses to be used in the system, the kernel setting should enable the IPv6. As mention in previous section, the kernel 2.6.32 should be setup to meet IPv6 forwarding. Figure 2.6 shows that the LMA system has been successfully enabled in the PMIPv6. To see this result, the LMA need to be run and the script in Fig. 2.6 will be seen.

Final result to prove that this PMIPv6 test-bed project is successfully developed is by making communication between MN and CN. Figure 2.7 shows that MN can communicate with CN with ‘ping’ procedure.

Experiment has done for topology shown in Fig. 2.5 with two scenarios which d_{APs} 5 and 10 m. The experiment has done for handover without LR and handover with LR to check and compare the handover latency. As to mention again, to

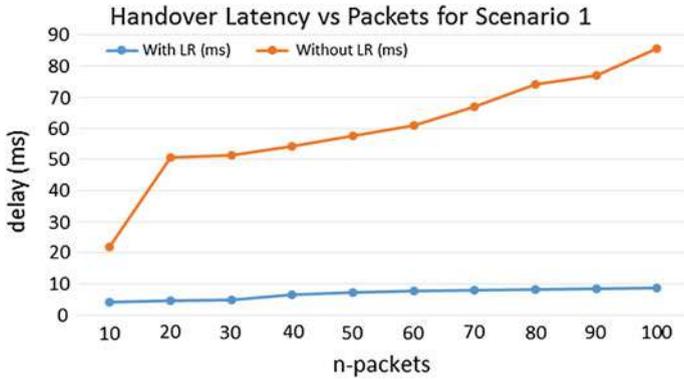


Fig. 2.8 Handover latency versus packets send for scenario 1

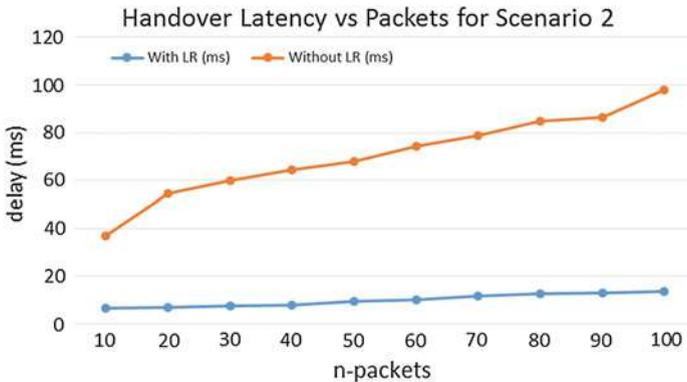


Fig. 2.9 Handover latency versus packets send for scenario 2

enable multimedia communication (seamless multimedia handover) while hand-over happen, the handover latency should be less than 50 ms.

Figure 2.8 shows the handover latency versus packets send for scenario 1. Scenario 1 is the scenario where distance between AP-MAGs is 5 m ($d_{APs} = 5$ meter). We can see that the handover latency without LR is more than 50 ms while data packets send more than 30 packets (64 bytes every packets). As for this, it cannot be used for seamless multimedia handover. When the LR algorithm (RSSI APPs LR with Dynamic LMA for PMIPv6) used, the handover latency will be minimize until less than 10 ms for 10–100 packets send. This conclude that the LR algorithm can be used for seamless multimedia handover as it reduce the handover delay.

While Fig. 2.9 shows the handover latency versus packets send for scenario 2. Scenario 2 is the scenario where distance between AP-MAGs is 10 m ($d_{APs} = 10$ m).

We can see that the handover latency without LR is more than 50 ms while data packets send more than 20 packets (64 bytes every packets). As for this, it cannot

be used for seamless multimedia handover. When the LR algorithm (RSSI APPs LR with Dynamic LMA for PMIPv6) used, the handover latency will be minimize until less than 20 ms for 10–100 packets send. This conclude that the LR algorithm can be used for seamless multimedia handover as it reduce the handover delay. The delay is higher than scenario 1 because when the d_{APs} is large, the delay time will increase.

2.6 Conclusion

This PMIPv6 test-bed was successfully run without errors and handover experiment had been done to see the handover latency for different distance between AP. This proposed work has successfully develop real test-bed for PMIPv6. The optimized routing problem in [8] has completely solved with reducing the handover delay.

2.7 Future Work

For future work, the algorithm will be run on the simulation also by using NS3.

References

1. Perkins, C. (ed.): IP Mobility Support for IPv4, Revised, RFC5944
2. Johnson, D.: Rice University, Mobility Support In IPv6, RFC3775
3. Kato, T., Takechi, R., Ono, H.: A study on mobile IPv6 based mobility management architecture, FUJITSU Science Technology Journal 37, 1, June 2001. M. Clerc, the swarm and the queen: towards a deterministic and adaptive particle swarm optimization. In: Proceedings of the IEEE Congress on Evolutionary Computation (CEC), pp. 1951–1957 (1999)
4. Gundavelli, S., Leung, K., Devarapalli, V., Chowdhury, K., Patil, B.: Proxy mobile IPv6, RFC 5213, Aug 2008. RFC5213-3 H.H. Crockell, specialization and international competitiveness. In: Etemad, H., Sulude, L.S. (eds.) Managing the Multinational Subsidiary, Croom-Helm, London (1986)
5. Liebsch, M.: PMIPv6 localized routing problem statement, draft-ietf-netext-pmip6-lr-ps-02, RFC6279, Jan 2010. In: Deb, K., Agrawal, S., Pratab, A., Meyarivan, T. (eds.) A Fast Elitist Non-dominated Sorting Genetic Algorithms for Multiobjective Optimization: NSGA II, KanGAL report 200001, Indian Institute of Technology, Kanpur, India (2000)
6. Jeong, S., Shin, M.-K., Kim, H.-J.: Performance comparison of route optimization schemes in proxy mobile IPv6”, ICACT 2010. In: 12th International Conference on Advanced Communication Technology, Gangwon-Do, Korea (2010)
7. Lee, K.-W., Seo, W.-K., Choi, J.-I., Cho, Y.-Z.: A simple route optimization detection scheme for multiple LMAs in PMIPv6 domain. In: Asia-Pacific Conference on Wearable Computing Systems, APWCS 2010, Shenzhen, China, IEEE Computer Society, 17–18 Apr 2010

8. Rasem, A., St-Hilaire, M., Makaya, C.: A comparative analysis of predictive and reactive mode of optimized PMIPv6. *IWCMC*, 722–727 (2012)
9. Lei, J., Fu, X.: Evaluating the benefits of introducing PMIPv6 for localized mobility management. In: *The International Wireless Communications and Mobile Computing Conference 2008 (IWCMC 2008)*, China, Greece, 6–8 Aug (2008)
10. Jun, T., Changxing, P., Bo, S.: *Research on the localized routing mechanism for the PMIPv6*. J. Xidian University, 2012-01
11. Minoli, D.: *Mobile Video with Mobile IPv6*. Wiley, New York (2012)
12. Iapichino, G., Bonnet, C.: Experimental evaluation of proxy mobile IPv6: an implementation perspective. In: *IEEE Wireless Communications and Networking Conference (WCNC)*, Sydney, Australia, 18–21 Apr 2010
13. Lee, H.-B., Han, Y.-H., Min, S.-G.: Network mobility support scheme on PMIPv6 networks. In: *International Journal of Computer Networks and Communication (IJCNC)*, vol. 2, Sept 2010
14. Melia, T., Bernardos, C.J., de la Oliva, A., Giust, F., Calderon, M.: IP flow mobility in PMIPv6 based networks: solution design and experimental evaluation. In: *Wireless Personal Communications Journal*. Springer, doi: [10.1007/s1277-011-0423-3](https://doi.org/10.1007/s1277-011-0423-3), Oct 2011
15. Kempf, J., Leung, K., Roberts, P., Nishida, K., Giaretta, G., Liebsch, M.: *Problem Statement for Network-Based Localized Mobility Management*, RFC 4830, IETF (2007)
16. Wahab, N.H.A.: *Location-Assisted Session Transfer for Real Time Applications*, Master Thesis. Universiti Teknologi Malaysia, Skudai (2010)
17. Nagamalai, D., Renault, E., Dhanuskodi, M.: Advances in parallel, distributed computing. In: *First International Conference on Parallel, Distributed Computing Technologies and Applications, PDCTA 2011*, Tirunelveli, Tamil Nadu, India, Proceedings Volume 203 of *Communications in Computer and Information Science*, Published by Springer, 23–25 Sept 2011
18. Udugama, A., Iqbal, M.U., Toseef, U., Goerg, C., Fan, C., Schlaeger, M.: Evaluation of a network based mobility management protocol: PMIPv6. In: *IEEE 69th Vehicular Technology Conference, VTC Spring (2009)*
19. Lee, J.-H., Lim, H.-J., Chung, T.-M.: Preventing out-of-sequence packets on the route optimization procedure in proxy mobile IPv6. In: *22nd International Conference on Advanced Information Networking and Applications (aina 2008)*
20. Chadchan, S.M., Akki, C.B.: 3GPP LTE/SAE: an overview. *Int. J. Comput. Electr. Eng.* **2**(5) (2010)
21. Nishida, K., Tanaka, I., Koshimiz, T.: Basic SAE management technology for realizing All-IP network. *NTT DOCOMO Tech. J.* **11**(3)
22. Ali-Yahiya, T.: *Understanding Lte and Its Performance*. Springer, Berlin (2011)
23. Abed, G.A., Ismail, M., Jumari, K.: Modeling and performance evaluation of LTE networks with different TCP variants. *Int. J. Inf. Commun. Eng.* **6**, 4 (2010)
24. Kim, J.H., Haw, R., Hong, C.S.: Development of PMIPv6 based 6LoWPAN sensor node mobility scheme. In: *The third AsiaFI Winter School*, Seoul National University, Seoul, Korea, 27 Feb 2010
25. Idserda, K.: *Simultaneous binding proxy mobile IPv6*. Master's Thesis, University of Twente, Enschede, The Netherlands (2008)
26. PMIPv6 Open Source, <http://www.openairinterface.org>
27. Soto, I.: Universidad Politécnica de Madrid; Carlos J. Bernardos, and María Calderón, Universidad Carlos III de Madrid; and Telemaco Melia, Alcatel Lucent Bell Labs, PMIPv6: A network-based localized mobility management solution. *Internet Protoc. J.* **13**(3)
28. Wahab, N.H.A., Latiff, L.A., Ariffin, S.H.S., Faisal, N.: Development of proxy mobile IPv6 using flat domain model test-bed. In: *7th International Conference on Computing and Convergence Technology ICCCT2012 (ICCIT, ICEI and ICACT)*, Seoul, Korea, 3–5 Dec 2012
29. Zhang, X., Zhou, X.: *Lte-Advanced Air Interface Technology*. CRC Press, Bosa Roca (2012)
30. Lucent, A.: *Strategic White Paper*

Chapter 3

Polytetrafluoroethylene Glass Microfiber Reinforced Slotted Patch Antenna for Satellite Band Applications

M. Samsuzzaman, T. Islam and M.R.I. Faruque

Abstract In this paper, a new configuration of triangular and diamond slotted patch shape microstrip fed low-profile antenna is introduced on Polytetrafluoroethylene glass microfiber reinforced material substrate for satellite applications. The antenna is composed of a rectangular shape patch slotted with eight triangles and two diamonds and elliptical slotted ground plane. The rectangular shape patch is obtained by cutting two diamond slots in the middle of the rectangular patch and six triangular slots on the left and right side of the patch, and two triangular slots in the up and down side of the patch. The slotted radiating patch, the elliptical slot ground and microstrip fed allow widening the matching bandwidth. Using a finite element based simulator, a parametric investigation was performed for the optimization. The proposed antenna offers a -10 dB impedance bandwidth of 1,020 and 1,350 MHz, respectively, and relatively stable and omnidirectional radiation patterns in the matching band.

3.1 Introduction

In modern wireless communication system, wide and multiband antennas demand is increasing due to support multi user and to provide more information with higher data transmitting and receiving rates. Among different kinds of antennas,

M. Samsuzzaman (✉) · T. Islam
Department of Electrical Electronic and Systems Engineering, Universiti Kebangsaan, Kebangsaan, Malaysia
e-mail: sobuzcse@eng.ukm.my

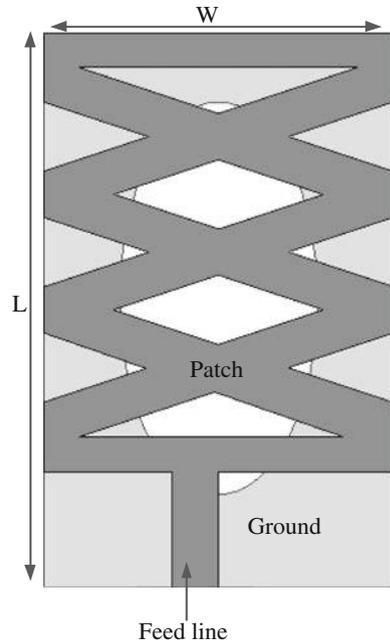
T. Islam
e-mail: titareq@yahoo.com

M.R.I. Faruque
Centre of Space Science, Universiti Kebangsaan, Kebangsaan, Malaysia
e-mail: rashedgen@yahoo.com

microstrip antennas are one of the most prominent structures due to their light weight, compatibility, low-profile, ease of fabrication, multifrequency capability and low cost. Compared to conventional three different types of antennas planar microstrip antennas printed on small pieces of printed circuit board (PCB) becomes familiar with recent wireless communication. Because it can be easily embedded into wireless devices or integrated with other radio frequency (RF) circuitry. Generally, a planar structure can be used to minimize the volumetric dimension of a wide band antenna by replacing three dimensional radiation elements with their planar design [1].

Various types of antennas have already been designed for wide and multi-band applications. A variety of dielectric materials have been used for designing and prototyping of these antennas. Basically, a dielectric material chosen for the design of wideband antennas is preferably needed to feature a higher permittivity and lower dissipation factor [2]. Materials with higher dielectric constant have higher capability to store charge and produce larger electromagnetic fields but limited isolation between conductors. On the other hand, material with lower permittivity are good insulators for lower frequency signals requiring high isolation in densely packed circuits such as mobile and satellite communications [3, 4]. The search for latest improved materials that could exchange the aluminum or other conductors is an important task for many applications in the wireless industry. Size and weight reduction, more tolerance to fatigue and easiness of manufacturing for complex structures are some advantages of using such materials. Such kinds of characteristics can lead to the operational cost reduction and performance intensification. The woven glass fabric with epoxy resin (FR4) composite is a successful and popular example of such materials in many applications. Several studies have been made in recent years to explore this material performance for antenna development. Moreover, by using a material with higher permittivity a compact antenna can be designed that is capable of achieving very wide or multiple operating band [4–8]. Ceramic-polytetrafluoroethylene composite material-based miniaturized split-ring multiband patch antenna was designed [5]. The proposed antenna obtained operating bandwidths (reflection coefficient < -10 dB) range from 5.0 to 6.5 GHz, 9.1 to 9.6 GHz, and 10.7 to 11 GHz. However, the antenna was designed on high dielectric and costly substrate. A miniaturized modified circular patch antenna was designed on ceramic-polytetrafluoroethylene (PTFE) composite material with dimension $0.22\lambda \times 0.29\lambda \times 0.23\lambda$; the proposed antenna achieved multi-band characteristics. However, the antenna failed to fulfil the requirement of Wi-Fi-WiMAX applications [6]. The dual-band operation of a microstrip patch antenna (20 mm \times 20 mm) for Ku- and K-bands has been presented [9] on glass microfiber reinforced PTFE. A compact square loop multiband patch antenna design on high dielectric ceramic composite material was proposed [10]. Though the reported antenna achieved multiband but impedance bandwidth was low and substrate cost is high compare to epoxy resin fiber. A printed wide-slot antenna designed and prototyping on available low-cost polymer resin composite material fed by a microstrip line with a rotated square slot for bandwidth enhancement and defected ground structure for gain enhancement [11]. A wideband pentagon shape

Fig. 3.1 Proposed antenna geometry layout



microstrip slot antenna was designed on epoxy resin composite material [12]. The proposed design antenna obtained 124 % impedance bandwidth, but its use in portable communication devices was limited due to large ground plane.

This article presents a simple modified rectangular shape printed antenna that exhibits dual band characteristics. An elliptical slot is etched on the ground plane for more coupling with slotted patch radiator. The simulated reflection coefficients and far field radiation pattern are presented. The simulation and analysis of the presented antenna are performed by using Finite Element Based high frequency structural simulator HFSS.

3.2 Antenna Design, Architecture and Optimization

The proposed, modified, rectangular-shaped, dual-band antenna is illustrated in Fig. 3.1. For the design studied here, the radiating element and feeding line are printed on the same side of the microstrip patch substrate, which has a thickness of 1.575 mm, a dielectric constant of $\epsilon_r = 2.33$ and $\tan\delta = 0.002$, while the other side is the ground plane of the antenna. The microstrip transmission line is used to feed the antenna. The basic antenna structure begins as a rectangular-shaped patch. The radiating patch is created by cutting two diamond-shaped slots and two equatorial triangular slots in the middle as well as three triangular slots on both sides of the

rectangular patch. An elliptical slot is etched on the ground plane of the proposed antenna. Different arm lengths of different slots and widths have been optimized to obtain optimum outputs. The dual-frequency, wide-impedance matching capability is enabled via the electromagnetic coupling effect of the ground plane to the feed line and the radiating patch. The design parameters are $L = 38$ mm, $W = 30$ mm.

3.3 Properties and Performance Analysis of Different Dielectrics Substrate

Four different substrate materials, the properties of which are summarized below are used in the simulation studies presented herein.

The FR4 substrate material consists of an epoxy matrix reinforced by woven glass. This composition of epoxy resin and fiber glass varies in thickness and is direction dependent. One of the attractive properties of polymer resin composites is that they can be shaped and reshaped without losing their material properties. The composition ratio of the material is 60 % fiber glass and 40 % epoxy resin.

Taconic TLC laminates are engineered to provide a cost effective substrate suitable for a wide range of microwave applications. TLC laminates offer superior electrical performance compared to thermoset laminates (e.g. FR-4, PPO, BT, and Polyimide and cyanate ester). TLC's construction also provides exceptional mechanical stability.

NeltecNX 9240 has superior mechanical and electrical performance, making it the material of choice for low-loss, high-frequency applications, such as wireless communications. This material is specially designed for very low-loss antenna applications. The enhanced N9000 materials reduce passive intermodulation issues in antenna and high-power designs.

RT/Duroid 5870-filled PTFE composites are designed for exacting strip-line and microstrip circuit applications. The unique filler results in a low density, lightweight material that is beneficial for high-performance, weight-sensitive applications. The very low dielectric constants of such RT/Duroid 5870 laminates are uniform from panel to panel and are constant over a wide frequency range. This substrate material is a high-frequency laminate and PTFE (Polytetrafluoroethylene) composite amplified using glass microfibers. Table 3.1 represents the different substrate dielectric properties.

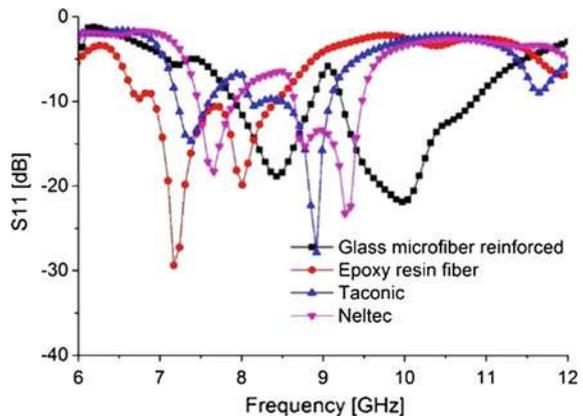
The effect of the different substrate material on the reflection coefficient of the proposed antenna is shown in Fig. 3.2. It can be clearly observed that the proposed antenna offers a wider bandwidth and an adequate return loss value compared with other reported materials. Although the antenna with epoxy resin fiber material substrate provides a lower return loss value due to its higher dielectric constant, but bandwidth is narrower, and the desired resonance is shifted. Moreover, the epoxy resin fiber loss tangent is higher compared to others. Others two material

Table 3.1 Properties of different substrate^a

Parameters	FR4_epoxy	Rogers RT/duroid 5870	Taconic	Neltec NY
Dielectric constant	4.6	2.33	2.4	2.4
Loss tangent	0.02	0.0012	0.0016	0.0016
Water absorption (%)	<0.25	0.02	<0.02	<0.02
Tensile strength	<310 MPa	450 MPa	–	–
Volume resistivity (M Ω .cm)	8 \times 10 ⁷	2 \times 10 ⁷	1 \times 10 ⁷	1 \times 10 ⁷
Surface Resistivity (M Ω)	2 \times 10 ⁵	3 \times 10 ⁷	1 \times 10 ⁷	1 \times 10 ⁷
Breakdown voltage (kV)	55	>60	–	>60
Peel Strength (N/mm)	9	5.5	12	12

^a All information collected from the different substrate data sheet from their website

Fig. 3.2 Effect of reflection coefficient of different substrate material



substrate material have achieved resonance, but low performance. The dielectric properties of the substrate material are tabulated in Table 3.1.

The far-field radiation patterns for the proposed wideband dual-frequency slotted antenna are also examined. Figure 3.3 shows the measured co— and cross—polarized radiation patterns including the horizontal (E plane) and vertical (H plane) polarization pattern for the antenna at lower band of 8.25 GHz and upper band 9.95 GHz. These results represent that the patterns are stable across the operating matching band. One can note that there is small asymmetry in the both E and H plane patterns which is due to the radiation of the microstrip line and the patch. Omnidirectional radiation pattern is observed in both E and H plane.

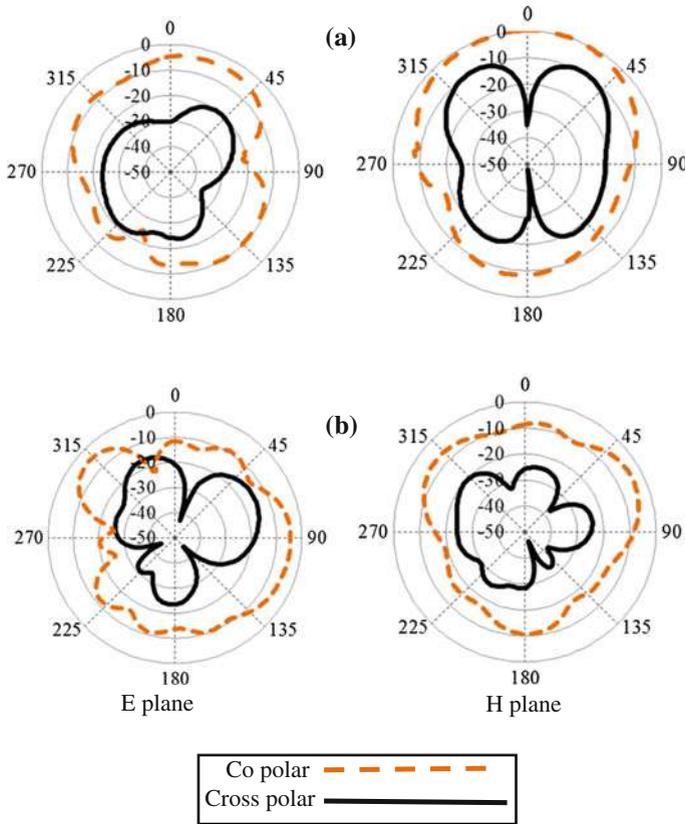


Fig. 3.3 E- and H-plane radiation patterns of the proposed antenna at **a** 8.25 GHz **b** 9.95 GHz

3.4 Conclusion

A novel dual band configuration of triangular and diamond slotted patch with slotted ground plane antenna is proposed. The antenna is designed on Polytetrafluoroethylene glass microfibre reinforced substrate. Compared to the other reported material, the proposed material antenna exhibits better performance in terms of impedance bandwidth, and stable radiation pattern. In this design a large bandwidth has been obtained by adding different slot in the patch and the ground plane. In addition, a parametric investigation was carried out to optimize the proposed design. Simulated results show that the proposed antenna can offer an impedance bandwidth of 1,020 and 1,350 MHz, respectively. The proposed shape dual band antenna has a very simple structure, which makes the design simpler and fabrications easier, and is very suitable for applications in the access points of satellite application.

References

1. Balanis, C.A.: Antenna theory: analysis and design. Wiley-Interscience, Hoboken (2012)
2. OOHIRA, K.: Development of an antenna material based on rubber that has flexibility and high impact resistance. NTN Tech. Rev. **76**(Special Issue), 58–63 (2008)
3. Art Aguayo, S.: Analyzing advances in antenna Materials. Antenna Syst. Technol. 14–15 (2010)
4. Samsuzzaman, M., Islam, M.T., Mandeep, J.S.: Parametric analysis of a glass-micro fibre-reinforced PTFE material, multiband, patch-structure antenna for satellite applications. Optoelectron. Adv. Mater. Rapid Commun. **7**(9), 760–769 (2013)
5. Ullah, M.H., Islam, M.T., Mandeep, J., Misran, N.: Ceramic-polytetrafluoroethylene composite material-based miniaturized split-ring patch antenna. Sci. Eng. Compos. Mater. **21**(3) 405–410 (2014)
6. Ullah, M.H., Islam, M.T.: Miniaturized modified circular patch monopole antenna on ceramic-polytetrafluoroethylene composite material substrate. J. Comput. Electron. 1–6 (2013)
7. Assis, R.R.d., Bianchi, I.: Analysis of microstrip antennas on carbon fiber composite material. J. Microw. Optoelectron. Electromagn. Appl. **11**, 154–161 (2012)
8. Samsuzzaman, M., Islam, T., Abd Rahman, N.H., Faruque, M.R.I., Mandeep, J.S.: Coplanar waveguide fed compact wide circular-slotted antenna for Wi-Fi/WiMAX applications. Int. J. Antennas Propag. **2014**, 10 (2014)
9. Islam, M., Islam, M., Faruque, M.: Dual-band operation of a microstrip patch antenna on a duroid 5870 substrate for ku-and k-bands. Sci. World J. **2013**, 378420 (2013)
10. Ullah, M.H., Islam, M.T.: A compact square loop patch antenna on high dielectric ceramic-PTFE composite material. Appl. Phys. A **113**(1), 185–193 (2013)
11. Samsuzzaman, M., Islam, M.T., Mandeep, J.S., Misran, N.: Printed wide-slot antenna design with bandwidth and gain enhancement on low-cost substrate. Sci. World J. **2014**, 10 (2014)
12. Rajgopal, S.K., Sharma, S.K.: Investigations on ultrawideband pentagon shape microstrip slot antenna for wireless communications. IEEE Trans. Antennas Propag **57**(5), 1353–1359 (2009)

Chapter 4

2.4 GHz Circularly Polarized Microstrip Antenna for RFID Application

Rudy Yuwono and Ronanobelta Syakura

Abstract Due to deployment of the wireless technology, the microstrip antenna can be used for RFID application especially for tag antenna. The existed RFID tag antenna usually use dipole model to get the maximum performance. However, the dipole antenna which has linear polarization performance has disadvantages on alignment that makes data detection on RFID can not be accurate. To reduce error during data detection caused of misalignment between reader and transponder, the tag antenna should have circularly polarized performance. The antenna was made with microstrip-line fed on FR-4 Epoxy substrate and works covers 2.4 GHz of its frequency range of 2 up to 3.3 GHz which S-parameter level below -9.54 dB and the direction of the transmitted power omnidirectionally and has circular polarization at frequency range of 2.39 until 2.46 GHz which has axial ratio below 3 dB.

Keywords Circularly polarized · RFID tag · Microstrip antenna

4.1 Introduction

The deployment of the wireless technologies is going rapidly especially in the antenna that supports the wireless system to receive and transmit the information. The antenna should be easy to fabricate, has various models and has proportional size with the wireless system. The antennas type which has these criteria is the microstrip antenna [1, 2, 7].

R. Yuwono (✉) · R. Syakura
Electrical Engineering Department, University of Brawijaya MT, Haryono 167, Malang,
Jawa Timur, Indonesia
e-mail: rudy_yuwono@ub.ac.id

R. Syakura
e-mail: syakour@gmail.com

Microstrip antenna can be applied in many wireless technologies. One of the wireless technologies which applied the microstrip antenna is in Radio Frequency Identification System (RFID), as the tag antenna that transmits the saved information in the transponder into reader to detect the information [8–10]. Existed tag antenna mostly used printed dipole as the model to get the maximum performance during data detection [3, 12].

Usage of the printed dipole has disadvantages especially in alignment of propagation. Printed dipole which has linear or elliptical polarization can cause problem if the tag and reader get the misalignment [4, 13, 14]. The reader can not detect the information accurately because of that misalignment. To solve the problem, the tag antenna should be circularly polarized.

4.2 Antenna Design

The tag antenna is designed with the dimension of substrate 30×40 mm with the microstrip line-fed. Microstrip-line fed method is chosen because it is easy to fabricate and match impedance. For the shape of the microstrip-line chosen the L shape to get orthogonal phase magnitude which can occur circular polarization [5, 6].

The applied substrate material is FR-4 Epoxy that has dielectric constant (ϵ_r) 4.4 with thickness 0.8 mm. Antenna models is shown on Fig. 4.1.

4.3 Result

The fabricated antenna was shown on Fig. 4.2.

The Observed Parameter antenna covers result of S-parameter (S_{11}), directivity pattern and axial ratio.

4.3.1 S_{11} Result

S_{11} which obtained the frequency of the antenna that can work shown on Fig. 4.3

Figure 4.3 Shows that the antenna can work covers frequency of 2.4 GHz in the work frequency range of 2.2 until 3.3 GHz. The frequency range of the antenna obtained from the S_{11} level below 9.54 dB which is the maximum tolerance level of the antenna that can work [11].

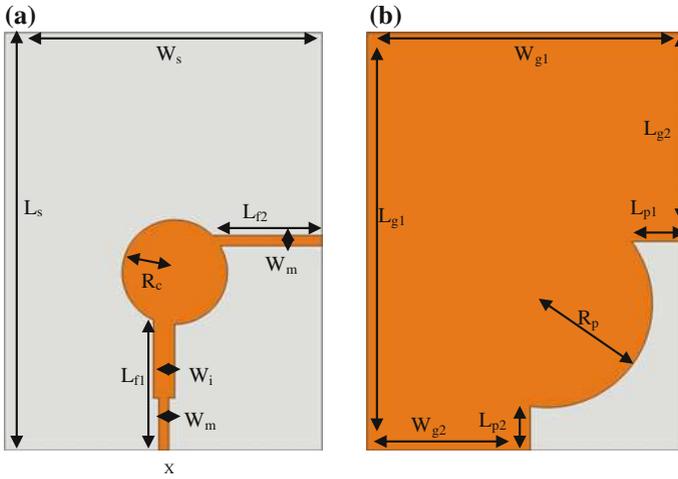


Fig. 4.1 The antenna design; X = the excitation port; **a** front view **b** rear view, with 0.8 mm FR-4 substrate ($\epsilon_r = 4.4$) $L_s = 40$ mm; $W_s = 30$ mm; microstrip-line fed with the dimension of $W_m = 1$ mm; $W_i = 2$ mm; $R_c = 5$ mm; $L_{f1} = 12.5$ mm; $L_{f2} = 10.2$ mm; ground plane with the dimension of $L_{g1} = 40$ mm; $W_{g1} = 30$ mm; $L_{g2} = 20$ mm; $W_{g2} = 15.5$ mm; $L_{p1} = 5$ mm; $L_{p2} = 4.1$ mm; $R_p = 10$ mm

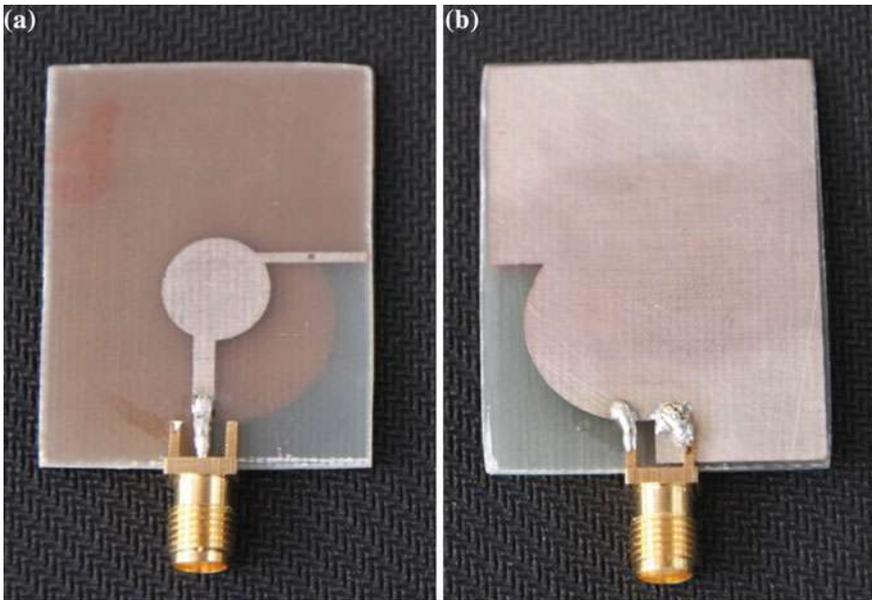


Fig. 4.2 The fabricated antenna; **a** front view; **b** rear view

Fig. 4.3 S_{11} parameter versus frequency

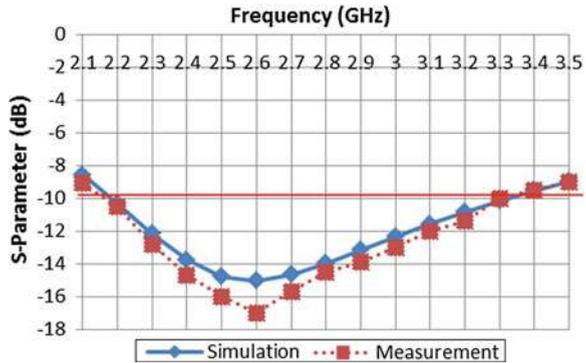
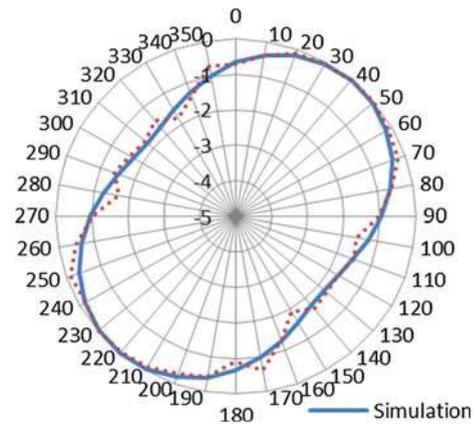


Fig. 4.4 Polar diagram of directivity pattern of the antenna at frequency of 2.4 GHz



4.3.2 Directivity Result

The directivity result of the antenna at frequency of 2.4 GHz shown on polar diagram as directivity pattern which is shown at Fig. 4.4

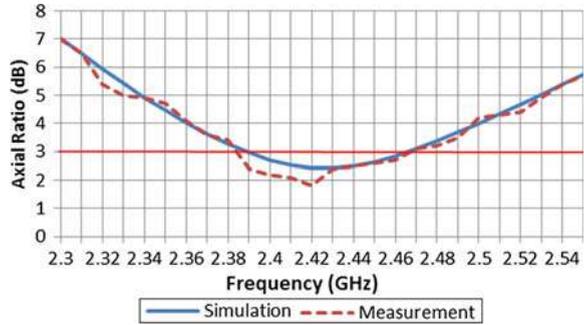
Figure 4.4 shows which the antenna has directivity pattern of omnidirectional. It is obtained from the directivity pattern that can transmit the power in all direction.

4.3.3 Axial Ratio Result

The result of the axial ratio shown on Fig. 4.5

Figure 4.5 show that the antenna has circular polarization at frequency range of 2.39 up to 2.46 GHz obtained from the axial ratio level that has value below 3 dB. With that performance the antenna can be used for RFID tag with circular polarization.

Fig. 4.5 Axial ratio with the function of frequency



4.4 Conclusion

The RFID tag antenna which made with microstrip-line fed method on 0.8 mm FR-4 epoxy with the dimension of the substrate 30×40 mm has performance can work covers frequency of 2.4 GHz in the range of frequency 2 up to 3.3 GHz with the directivity pattern omnidirectional. The antenna also has circular polarization which is obtained from the axial ratio at frequency of 2.39 up to 2.45 GHz which is used for RFID application.

Acknowledgments The Authors would like to express sincere gratitude to University of Brawijaya for their support under contract number SK DEKAN No 26/UN10.6/PG/2013, June 3, 2013

References

1. Balanis, C.A: Antenna Theory Analysis and Design. Wiley, New York (2005)
2. Bhartia, P., Bahl, Inder, Garg, R.: Ittipiboon: Microstrip Antenna Design Handbook. Artech House, Boston (2001)
3. Deavours, D.: UHF RFID Antennas. In: Bolic, M., Simplot-Ryl, D., Stojmenovic, I. (eds.) RFID Systems Research Trends and Challenges, p. 68. Wiley, New York (2010)
4. Finkenzeller, K.: Klaus: RFID Handbook, 2nd edn. Wiley, London (2003)
5. Joseph, R., Fukusako, T.: Circularly polarized broadband antenna with circular slot on circular ground plane. Prog. Electromagnet. Res. C **26**, 205–217 (2012)
6. Kumar, G., Ray, K.P.: Broadband Microstrip Antennas. Artech House, Boston (2003)
7. Preradovic, S., Karmakar, N.C.: Design of fully printable planar chipless RFID transponder with 35-bit data capacity. In: Proceeding of the 39th European Microwave Conference (2009)
8. Preradovic, S., Karmakar, N.C.: Design of chipless RFID tag for operation on flexible laminates. In: IEEE Antennas and Propagation Letters, vol. 9 (2010)
9. Preradovic, S., Karmakar, N.C.: Multiresonator-Based Chipless RFID-Barcode of the Future. Springer, New York (2012)
10. Nakar, P.S.: Design of a compact microstrip patch antenna for use in wireless/cellular devices. Electronic Theses, Treatises and Dissertations Paper 2790, <http://diginole.lib.fsu.edu/etd/2790>, (2004)

11. Yuwono, R., Baskoro, A., Erfan, A.D.: Design of circular patch microstrip antenna for 2.4 GHz RFID applications. In: *Lecture Notes in Electrical Engineering*, vol 235, pp. 21–28. Springer, Heidelberg (2013)
12. Yuwono, R., Wahyu, D.A., Muhammad, F.E.P.: Design of circular patch microstrip antenna with egg slot for 2.4 GHz ultra-wideband radio frequency identification (UWB RFID) tag applications. *Applied Mechanics and Materials*. vol 513–517 Trans Tech Publications Zurich-Durten Switzerland 3414–3418 (2014)
13. Ruengwaree, A.; Yuwono, R.; Kompa, G.: A noble rugby-ball antenna for pulse radiation. In: *The European Conference on Wireless Technology*, pp. 455–458. IEEE Conference Publications (2005)
14. Yuwono, R., Wendaria, P.: Design of circular patch microstrip antenna with rugby ball slot for ultra wideband applications. In: *The 2nd International Conference on Radar, Antenna, Microwave, Electronic and Telecommunication (ICRAMET)* pp. 130–132 (2013)

Chapter 5

On Understanding Centrality in Directed Citation Graph

Ismael A. Jannoud and Mohammad Z. Masoud

Abstract Modeling complex networks as directed/undirected graphs is considered one of the most common methods in network science. Citation graph is a directed graph of scientific published papers. This graph has been studied massively in the past decade. Citation graph can be utilized to study relationships between authors and papers. It can be used to study the characteristics of citation network to demonstrate the growth model, graph type and to predicted hot new topics. In this paper, we attempt to study the relationship between popularity of a paper and the publication date. The purpose of this study is to demonstrate the relation between paper quality and hot topics. Betweenness metric has been used to measure the popularity of a published paper. Moreover, a comparison between betweenness and citation count (node degree) has been conducted to show that papers may have a small citation count, however, they may have a great impact in research field. We have generated a directed citation graph by crawling paper information from CiteSeerx. Our study shows that date of publication is important to write a popular paper. However, high quality papers get opportunity to be popular regardless the date of publication.

Keywords Node degree · Citation graph · Graph metrics · Cluster coefficient · Betweenness

5.1 Introduction

Citation is the process of quotation information, results and conclusions from book, paper, or websites in a research work. Scientific research papers (SRP) contain from few references to few hundreds. Citation is one of the metrics that

I.A. Jannoud (✉) · M.Z. Masoud
Al-Zaytoonah University of Jordan, Amman, Jordan
e-mail: ismael.jannoud@zuj.edu.jo; ismaelj@yahoo.com

M.Z. Masoud
e-mail: eng.mohammad.masoud@gmail.com

used to measure the popularity and the importance of SRP. Moreover, citation may be used to find the roots of science fields. Citation process generates a complex directed graph, which is called citation graph or citation network. In this graph, nodes (papers) are connected with directed edges that begin from newly published papers and point to old papers.

Citation in these networks are the edges, papers are the nodes. This definition converted these networks into complex directed graphs. The emerged of network science, its applications and tools increased the possibilities in studying complex networks. Citation network is one of the complex networks that have been studied heavily in the past decade. Paper popularity, author ranking, author relations and community studies are some examples of the vast range of studies that have been conducted in this field.

Mapping graph parameters and properties into physical meanings that reflects useful meaning is the niche of network science. This mapping helped in predicting special phenomenon of citation networks, such as, power-law distribution, graph type and the expanding models of these networks. However, many graph parameters have not been studied and mapped in citation networks, such as, betweenness and closeness.

In this work, paper importance will be measured. Paper centrality in the citation network will be extracted. Timing of papers and their centrality will be examined. The purpose of this work is to demonstrate the important of hot topics time and the position of papers in citation networks. Graph betweenness will be utilized to measure the centrality of papers. To conduct our experiment massive amount of papers in a certain field must be harvested. To facilitate our experiment, a web crawler has been implemented to collect paper information from CiteSeerx website. To this end, we have generated a direct graph from the harvested data. We seek to answer the following questions using graph analysis:

- What is the meaning of centrality in citation network?
- What the relationship is between hot topics timing and paper centrality?
- Can researchers write a paper with high citation value without considering the hot topic?

The rest of this paper is organized as follows; this section ends with the related works that have been conducted in this area. Section 5.3 introduces the performance metrics. Section 5.4 provides a description of the experiment that has been implemented. This section ends with the results and discussion. Finally, Sect. 5.6 concludes this work.

5.2 Related Work

Studying real network as graphs inspired researchers over the years. They heavily studied social, information, technological, and biological graphs [1]. These studies gain researchers more insights of how these networks may evolve and how bugs, errors and diseases may separate.

Many types and kinds of networks have been studied as graphs. For example, in [2, 3] it has been reported that the World Wide Web (WWW) follows an exponential degree distribution with more than 269 thousands nodes and around 1 and half million edges. This study provided that the WWW is a small world graph. In the measurement works [4], authors attempted to generate Internet graph to study its properties. They have produced an undirected graph with 10 thousands nodes and 31 thousands edges. A 0.035 global cluster coefficient value has been computed. These values have been computed over the years again to show the development of the Internet [5].

A third type of networks have been studied in [6], the author of this study generated a software-classes directed-graph. With 1,377 nodes and 2,213 edges, the author studied the properties of this small graph. The author reported a mean node-node distance with 1.5 hops. In addition, global clustering coefficient and degree correlation coefficient were computed. This study showed that software classes don't follow the small world phenomenon. Furthermore, in the works of [7, 8], authors generated a small undirected graph based on harvested data to simulate P2P network. A graph of 880 nodes and 1,296 edges has been implemented and studied. The author found that the average shortest path in this graph is 4.2 hops and the global clustering coefficient is 0.012. Unfortunately, their implemented graph was too small to mimic a swarm of P2P networks. In [9], authors attempted to study human's neural networks as a directed graph. They have constructed a directed graph with 307 neural and 2,359 edges between them. Subsequently, they studied the properties of this graph. They reported a global clustering coefficient of 0.18 and an average shortest path value of 3.97. These values demonstrated that human's neural network is a small world graph.

Finally, Citation networks have not been forgotten by researchers. In [5], the author studied the citation network as a directed-graph. The author computed node number and vertexes. However, the cluster coefficient and degree correlation were not computed. These results made it hard to predict node distribution in citation network. In [10] a citation graph has been generated to study the impact of co-authors on the popularity of papers. In addition, co-authors are connected if they have shared work. Our work in the citation network differs from these work in two points. First, we utilized a new graph parameter to study the centrality and popularity of a paper in a certain field. Second, time has been utilized to demonstrate the relation between time and popularity.

5.3 Performance Metric

This section introduces the metrics that have been utilized from graph theory. Graph centrality through betweenness and node degree will be utilized. Graph centrality is a metric to evaluate the importance of any node in a graph. Many methods are utilized to measure the centrality of a node. Node degree, eigenvector, closeness and betweenness are the most common methods to calculate the

centrality of nodes. In citation graph, centrality metric may be used to extract the core nodes of the graph. These core nodes may be introduced as the root of a scientific field. Any research in a scientific area starts from an idea in the past. However, there is a time when this field emerges as a hot topic. Core nodes are the nodes that convert a topic into a hot topic.

In this work, we have utilized the betweenness and node degree of the nodes in the citation graph to measure graph centrality. In the next section node degree and betweenness will be defined.

5.3.1 Betweenness and Node Degree Centrality

Betweenness of a node is defined as total number of paths that flow through this node [11]. Equation 5.1 shows the mathematical method of calculating betweenness of a node. In a directed graph such as citation graph, betweenness may be used to demonstrate the importance of a scientific paper.

$$G(v) = \sum Q_{si}(v)/Q_{si} \quad (5.1)$$

Node degree is defined as the total number of incoming and outgoing connections that connects a node with other nodes.

Node degree may be used as centrality metric. However, some nodes may be important and they may have less degree than other nodes. This fact is emerged in betweenness centrality more than node degree centrality. We believe that betweenness centrality results are more accurate for two reasons. First, in scientific research timing is an important factor. The important of a paper depends on the time. Old popular paper will not get more citation after a long time. Rich will be richer if they are not too old unlike other graphs, such as, the AS graph. Second, an important paper may be cited by little number of papers that may become more and more popular. These two facts can be found crystal clear in the betweenness centrality. Figure 5.1 shows an example of the differences between node degree and node betweenness. Nodes 1–7 have node degree more than node 8. However, node 8 is more important since it connects between this node and nodes: 9, 10.

5.4 The Experiment

A web-crawler has been implemented to harvest scientific paper information from CiteSeerx website. CiteSeerx is a scientific literature digital library. It focus primary on scientific research papers in the fields of computer and information science. It has many functionalities and methods that allow it to index postscript and PDF articles.

Our web-crawler starts by retrieving the search results of a specific topic. We have used ad hoc as our main topic to generate our graph. The maximum allowed retrieved results are 500 papers. Subsequently, the crawler uses the retrieved 500 paper names,

Fig. 5.1 Example of the differences between node degree and node betweenness centrality

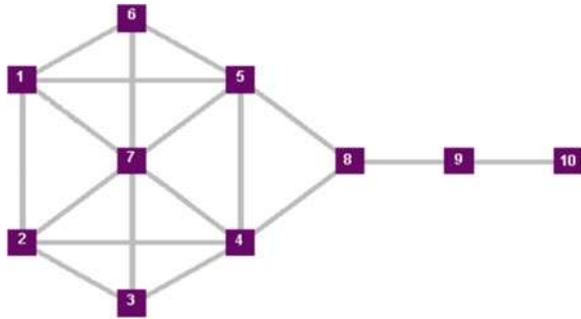
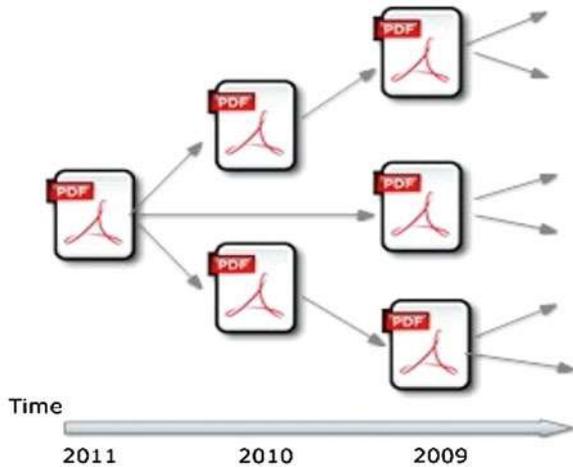


Fig. 5.2 Example of the generated direct citation graph



which are the seed list of the crawler, to retrieve their cited paper. Each cited retrieved paper is added into the list to retrieve its citations and so on. The harvesting process started in 15th of January and continued for 10 days. The crawling process stopped when encountered papers that have no information in the CiteSeerx web-site.

The harvested data of the conducted experiment has been utilized to generate a directed citation graph. Figure 5.2 shows an example of the generated direct citation graph.

Table 5.1 lists the properties of this graph. We can notes from the graph that the clustering coefficient of this graph is too small comparing with other graphs, such as, the AS graph. In addition, the average path length is longer since it's a directed graph.

5.5 Results

Figure 5.3 shows the CDF of node betweenness value. From this figure we can observe that more than 88 % of the nodes have a betweenness value less than 1. In addition we can observe that less than 4 % of the nodes have a betweenness value

Table 5.1 The constructed citation graphs properties

Property	Citation graph
Number of nodes	91,211
Number of links	221,677
Average node degree	2.43
Cluster coefficient	0.067
Average path length	7.4
Graph diameter	24

Fig. 5.3 The CDF of node betweenness value

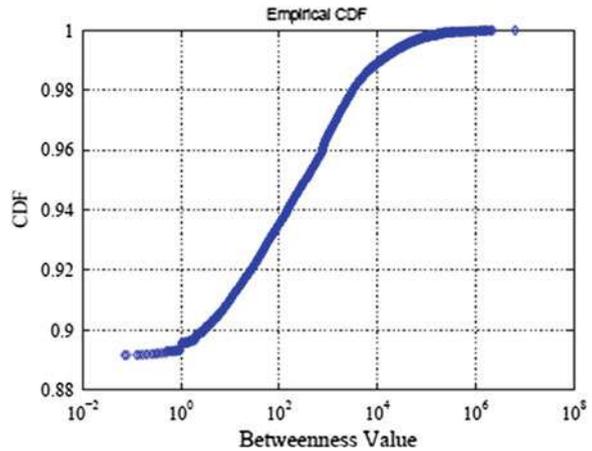
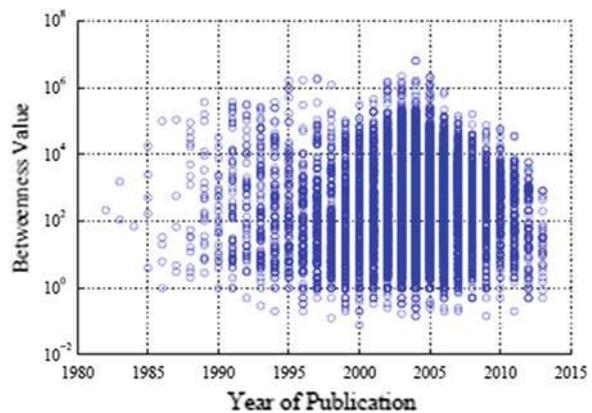


Fig. 5.4 The relation between publication year and the betweenness value



less than 100. More over we can observe that a tiny number of nodes (0.02 %) have a vast betweenness value. These nodes are the central of our citation graph.

Figure 5.4 shows a relation between publication year and the betweenness value. To compute this figure, 10 K papers have been excluded from the result. These papers have no publication year information in CiteSeerx web-site. The rest

Fig. 5.5 The relationship between betweenness value and node degree

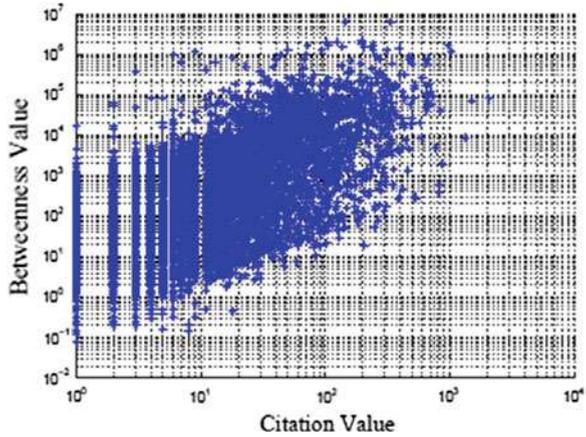
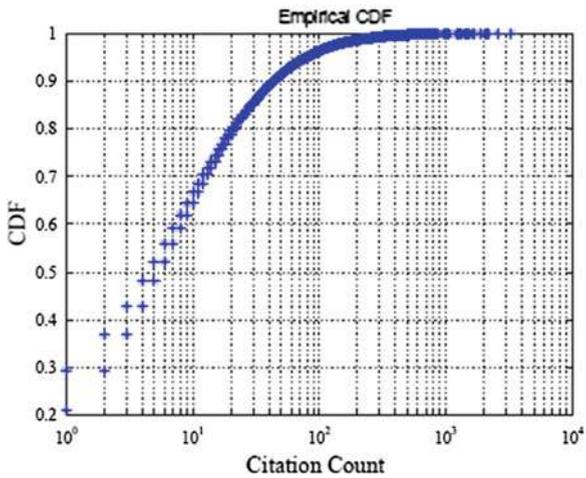


Fig. 5.6 The CDF of node citation count



of the papers (more than 8 K) have been used to generate this relation. We can observe that the figure is a random figure. We notice that for each year of publication, the betweenness value varies. In addition, we observed that the highest betweenness value was in 2005. However, the papers in ad hoc started before this year. This figure proves that high quality papers we be in the core of any field regardless of the year of publication.

Figure 5.5 shows the relation between betweenness value and node degree or citation count. This figure shows that the relation between betweenness and citation count is fuzzy. With the increased number of citation count, betweenness value increases.

However, some node with high citation value has small betweenness. In addition, we can observe that the node with highest betweenness value is not the

node with the highest citation count. The figure has a fuzzy slop to the right. This figure shows the differences between betweenness centrality and node degree centrality. This figure demonstrates that some nodes with less citation count may be more important than others. Betweenness may be used to show this fact.

Finally, Fig. 5.6 shows the CDF of node citation count (node degree). We can observe the similarity between the CDF of betweenness value in figure and this figure. However, unlike the betweenness CDF that has less than 10 % of the nodes with significant betweenness value, we can observe that more than 33 % of the nodes have significant node degree (more than 10). This fact demonstrate how betweenness centrality can be used to filter a massive number of nodes to obtain the most central and important once in a graph.

5.6 Conclusion

Researches utilized the citation graph to study and introduce the publication networks for various reasons. The last decade witnessed a heavy study of the properties of this graph. In this work, we have studied the impact of the publication date on the popularity of published papers. Three main graph properties have been utilized to study the relationship between popularity and hot topic (date of publication). Betweenness, global and local cluster coefficients and node degree distributions has been computed. To this end we have generated a citation graph by crawling paper information of ad hoc (research field) from CiteSeerx web-site. Our results demonstrated three main points. First, the probability of getting high citation count increases with the time or the date of publication. Second, there is a probability that a high quality paper will get a good citation count even if it is late in the field. Finally, the data shows a fuzzy slope in the graphs. This uncertainty requires more experiments.

References

1. Newman, M.E.J.: The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003)
2. Barabasi, A.-L., Albert, R., Jeong, H.: Scale-free characteristics of random networks: the topology of the world-wide web. *Phys. A* **281**, 69–77 (2000)
3. Albert, R., Jeong, H., Barabasi, A.-L.: Diameter of the world-wide web. *Nature* **40**, 130–131 (1999)
4. Dorogovtsev, S.N., Mendes, J.F.F.: Language as an evolving word web. *Proc. R. Soc. Lond. Ser. B* **268**, 2603–2606 (1999)
5. Masoud, M., Hei, X., Cheng, W.: A graph-theoretic study of the flattening internet as topology. In: *IEEE ICON*, December 2013 (2013)
6. Valverde, S., Cancho, R.F., Sole, R.V.: Scale-free networks from optimal design. *Europhys Lett.* **60**, 512–517 (2002)

7. Ripeanu, M., Foster, I., Iamnitchi, A.: Mapping the gnutella network: properties of large-scale peer to-peer systems and implications for system design. *IEEE Internet Comput.* **6**(1) (2002)
8. Adamic, L.A., Lukose, R.M., Puniyani, A.R., A. Huberman, B.: Search in power-law networks. *Phys. Rev. E* **64**, 046135 (2001)
9. White, J.G., Southgate, J.N.T.E., Brenner, S.: The structure of the nervous system of the nematode in *C. Elegans*. *Philos. Trans. R. Soc. Lond.* **314**, 1–340 (1986)
10. Healy, K.: A co-citation network for philosophy, June 2013 (2013)
11. Lewis, T.G.: *Network science theory and practice*. Wiley, Hoboken (2008)

Chapter 6

Channel Capacity of Indoor MIMO Systems in the Presence of Spatial Diversity

M. Senon, M.N. Husain, A.R. Othman, M.Z.A. Aziz,
K.A.A. Rashid, M.M. Saad, M.T. Ahmad and J.S. Hamidon

Abstract The capacity of Multiple Input Multiple Output (MIMO) systems has received much attention in recent years. This paper analyze the capacities of MIMO channel model for indoor propagation with increasing the distance between transmit and receive antenna through simulation and measurement. A spatial diversity method is employed during measurement and simulation process. The investigation on the channel capacity for various distance and spacing of both transmitter and receiver antenna have been done. The investigations of channel capacity are included with difference distance between transmitter and receiver sides and different in element antenna spacing. For the simulation, the path loss for the free space and physical effect are been considered. The 2×2 rectangular microstrip patch array antenna is used in order to characterize channel parameter at 2.4 GHz operating frequency. The system measurement was been conducted in UTeM Microwave Laboratory, according to the real situation in indoor environment.

6.1 Introduction

Today, demand for high data rate and channel bandwidth is increasing due to the modern and future application requirement in wireless communication systems. The Multiple-input-multiple-output (MIMO) system was invented to make the multipath propagation mechanism as an advantage in order to increase the channel capacity. This system was characterized of multiple antennas that used at the transmitter and receiver sides and can increase the channel capacity (b/s/Hz) without increased the bandwidth and transmit power [1, 2]. The used of diversity was to increase the probability at the receiver end where at least one of the signals

M. Senon (✉) · M.N. Husain · A.R. Othman · M.Z.A. Aziz · K.A.A. Rashid
M.M. Saad · M.T. Ahmad · J.S. Hamidon
Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka,
Melaka, Malaysia
e-mail: P021110007@utem.edu.my; misida.senon@gmail.com

were received correctly [3]. The spatial diversity was one of the diversity techniques [4–7]. It can be done by space apart between the antennas but when used at the limited volume or space.

In WLAN technology, the signal will propagate through numerous paths which effected from the indoor environment. This phenomenon is call multipath propagation. So the MIMO channel system is introduced to solve this problem by exploiting the richness of multipath propagation. Additionally, the use of multiple antennas at both the transmitter and receiver provides significant increase in wireless channel capacity.

The MIMO channel models can be divided into the non-physical and physical models [8–10], in order to characterize the channel performance, the statistical model is used to find the channel capacity for every scenario. The MIMO channel model can be characterized by modelling the channel with consideration of physical parameter such as distance and scattering for every scenario.

This paper will discuss and analyzed the simulation of capacity effect to the wireless MIMO communication channel model for different spatial diversity is applied at the both sides of transmitter and receiver. Then the measured channel capacity is compared to channel capacity obtain from Kronecker and Weischselberger model.

6.2 Channel Model and Channel Capacity

6.2.1 MIMO Channel Model for LOS Scenario

Wireless propagation is dominated by the daily changes by environment features such as reflection, diffraction and transmission depending on the location of transmitter and receiver.

In this project, reflection is avoided because LOS is considered as a major signal. Since no reflection between n -th transmitter and m -th receiver, the channel coefficient h , only consists a direct path between these antennas. Because of the operation in free-space condition, the power receive of Friis free-space equation is considered [11].

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 r_{mm}^2 L} \quad (6.1)$$

Where P_t is the transmitted power, G_t and G_r are transmitter and receiver antenna gain and λ is the wavelength in meters. The system loss, L , such as cable loss or antenna efficiency are not related to free-space propagation.

6.2.2 System Configuration

In this paper, parameters of MIMO channel model identified with construct a different configuration in indoor environment. Figure 6.1 shows the antenna configuration and measurement setup. The development of algorithm is starting by calculating pathloss of the system. Then the correlation coefficient and eigen analysis are calculated sequentially to find the channel capacity of MIMO wireless communication system.

The Matlab simulation tool is use to compare MIMO channel models based on the theoretical data and measured data. The measurement campaign is done to collect the data required to verify the MIMO channel model (Kronecker model and Weichelberger model) and to estimate the model parameters that charaterize different configuration.

For this project, fixed wireless system is more to line-of-sight environment while the distance (d) is assumed much higher than antenna spacing (l). From geometrical arrangement, the different path length ($r_{n,m}$) for LOS arrangement between transmitter n th and receiver m th is calculate such as Eq. 6.2 [12, 13].

$$r_{n,m} = \sqrt{d^2 + (l(n - m))^2} \quad (6.2)$$

For the second scenario, both transmitter and receiver antenna is placed with different antenna spacing (l). Thus, the path length ($r_{n,m}$) is depending on the shifting of the antenna spacing. The paths length ($r_{n,m}$) for different antenna spacing at transmitter (l_m) and receiver (l_n) is given by:

$$r_{n,m} = \sqrt{d^2 + (l_m - l_n)^2} \quad (6.3)$$

The third configuration is a multiple altitude antenna configuration which is the uniform antenna spacing is placed at both transmitters and receiver antenna front end but with different altitude.

6.2.3 MIMO Channel Capacity

The data collected from the measurement are analyzed for every configuration. Equation (6.4) show the MIMO channel matrix \mathbf{H} where \mathbf{N} and \mathbf{M} representing the number of transmitting and receiving antennas [14].

$$\mathbf{H} = \begin{bmatrix} \rho_{11} & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & \rho_{22} & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & \rho_{33} & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & \rho_{44} \end{bmatrix}_{M \times N} \quad (6.4)$$

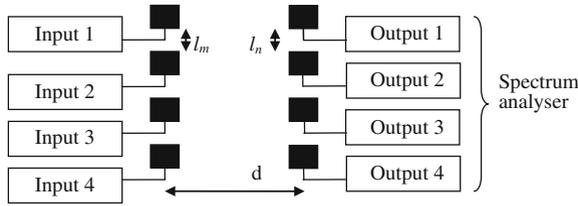


Fig. 6.1 Antenna configuration

The channel capacity is calculate by using Eq. (6.5) [15].

$$C_{MIMO} = \sum_{i=1}^i \log_2 \left(1 + \frac{SNR}{N} \lambda_i \right) b/s/Hz \quad (6.5)$$

6.3 Results and Discussion

Experimental work has been done by collecting data with configurations of typical and spatial diversity technique. The different measurement tools are helping in characterizing MIMO channel parameter.

6.3.1 Typical and Spatial Diversity

The measurement work was started by collecting the data from the typical setup. Table 6.1 show thesimulation and measurement for typical and spatial diversity setup with l_m and l_n is antenna spacing at transmitter ($l_m = l_n = \lambda$).

From the table, the Kronecker and Weichselberger model shows huge different compared to the measured data. However the Weichselberger model is closer to the measured data because this model uses joint correlation at both ends. It also prove the [16] work that Kronecker model is fail to predict the channel capacity for the system used more than 2×2 antenna.

6.3.2 MIMO System with Spatial Diversity Technique

Measurement for spatial diversity setup has been done by varying the spacing (l) between antennas at transmitter by fixing antenna spacing at receiver and vice versa. From the observation, by changing the spacing between the antennas, it will change the value of eigenvalue and capacity.

Table 6.2 shows the MIMO channel capacity for configuration by changing l at receiver with fixed antenna spacing at transmitter.

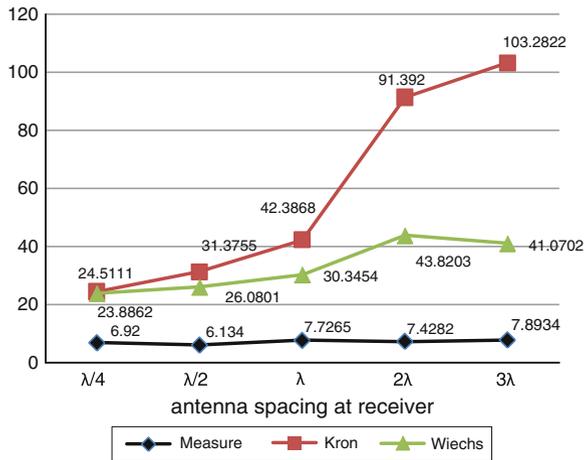
Table 6.1 Comparison of MIMO channel capacity for simulation and measured data of typical configuration at $l = \lambda$

Simulation		Measurement
Kronecker	Weichselberger	
42.3868 b/s/Hz	30.3454 b/s/Hz	7.86 b/s/Hz

Table 6.2 Result for MIMO channel capacity by changing l at receiver with fixed antenna spacing at transmitter

Antenna spacing at transmitter $l_m = \lambda$	Antenna spacing at receiver, l_n				
	$\lambda/4$	$\lambda/2$	λ	2λ	3λ
Channel capacity (b/s/Hz)	6.92	6.134	7.7265	7.4282	7.8934

Fig. 6.2 Comparison between measurement result and simulation result for spatial diversity at receiver



The result from Table 6.2 is a comparison of measurement result for MIMO Channel capacity by changing l at receiver with fixed antenna spacing at transmitter. From the table, the largest MIMO channel capacity is 7.8934 b/s/Hz at 3λ antenna spacing configuration. Due to the previous literature on spatial diversity, the capacity increase by expansion of spacing between the antennas.

From the Fig. 6.2, the result shows the channel capacity of Weichselberger model is closer to the measured data compare to Kronecker model.

The observation result of MIMO channel of that graph also shown capacity is increase by increasing the number of antennas. However, the capacity from measured data does not increase linearly by increasing antennas spacing.

Table 6.3 shows the result for antenna spacing at transmitter is changing with fixed l at receiver. Compared to typical configuration, the MIMO channel capacity is lower when applying spatial diversity at the transmitter which is the inter

Table 6.3 Result for antenna spacing at transmitter is changing with fixed l at receiver

Antenna spacing at receiver $l_n = \lambda$	Antenna spacing at transmitter (I_m)		
	$\lambda/4$	$\lambda/2$	λ
Channel capacity (b/s/Hz)	7.3274	6.3301	7.7265

Fig. 6.3 MIMO channel capacities with spatial diversity at transmitter

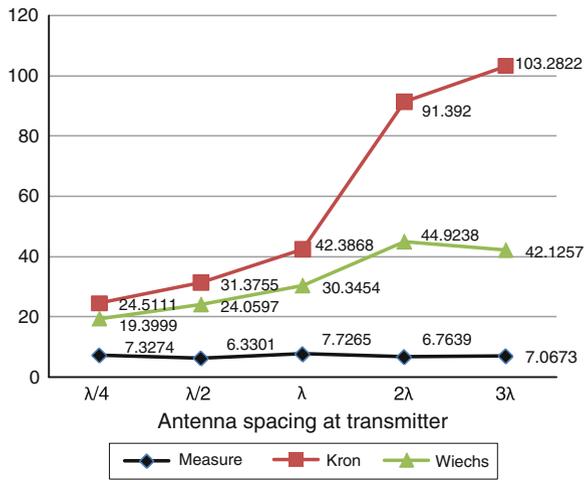
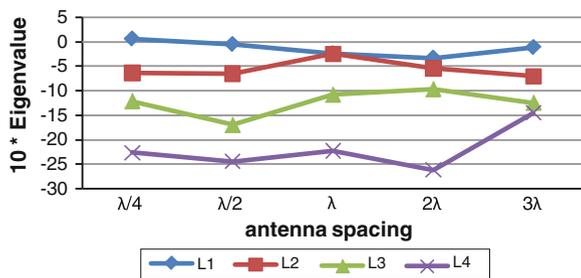


Fig. 6.4 Eigenvalue for measure data with spatial diversity at transmitter



element antenna spacing is shifted from $l = \lambda/4$ until 3λ . From the observation, the highest MIMO channel capacity is 7.7265 b/s/Hz while the lowest value of the system is 6.3301 b/s/Hz when applying $\lambda/2$ inter element antenna spacing. The differentiation between highest and lowest value is 18.07 %.

Figure 6.3 show the comparison between simulation and measurement result of MIMO channel capacity with spatial diversity at transmitter. The result has shown similar to the system with spatial diversity at transmitter where the simulation result is much differs from the measurement result. The MIMO channel capacity increase by increasing the spacing between the antennas for all models.

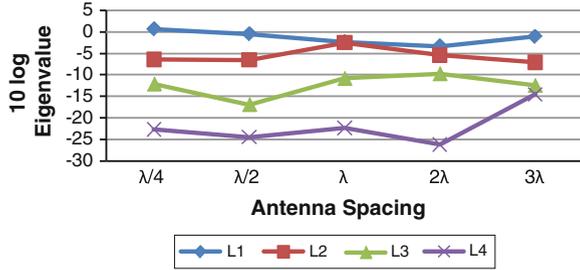
Table 6.4 Correlation coefficient at $l_m = 0.5$

	Receiver ($l_n = \lambda/4$)				Receiver ($l_n = \lambda$)			
-0.08116	-	0.07135	0.278584	0.184097	-0.22019	-0.05203	-0.06486	
	0.25826							
-0.36418	0.073011	0.320501	-0.02726	0.14998	0.53665	-0.30272	0.12837	
0.013381	-0.1958	0.069882	-0.51134	-0.20811	-0.23324	0.401047	-0.2107	
0.168741	0.105638	0.428591	0.447742	0.331321	0.037825	-0.36922	0.280644	
-0.08116	-	0.07135	0.278584	0.184097	-0.22019	-0.05203	-0.06486	
	0.25826							
-0.36418	0.073011	0.320501	-0.02726	0.14998	0.53665	-0.30272	0.12837	

Table 6.5 Result average correlation coefficient

	$l_n = \lambda/4$	$l_n = \lambda$
Average correlation	0.033713	0.024304

Fig. 6.5 Eigenvalue for MIMO system with spatial diversity at receiver



6.3.3 Eigenvalue Analysis and MIMO Channel Correlation Coefficient

Figure 6.4 shows the result of Eigenvalue analysis for measurement result of spatial diversity at transmitter.

From the graph, the highest eigenvalue of the system is L1 especially when the antennas are placed with $\lambda/2$ inter element spacing but the L2 and L3 is lower than the system with λ antenna spacing.

Correlation coefficient is used in order to calculate MIMO channel capacity. Table 6.4 summarize the correlation coefficient of MIMO system with spatial diversity at the receiver. The Table 6.5 had shown the average of correlation coefficient from the previous table. The average correlation is decrease by increasing the spacing (l_n). Due to the result of average correlation, the capacity of the system is better when the correlation coefficient is higher.

Figure 6.5 show the eigen value of the MIMO system when applying the spatial diversity at receiver. The channel capacity increase when the eigenvalue of the system close within each other such as at $l_n = 3\lambda$. Other than the L1, the eigenvalue of L4 also give a significant contribution in enhancing the capacity at $l_n = 3\lambda$ because the value is highest at that configuration compared to others.

6.4 Conclusion

The paper presented a brief analysis of the MIMO channel capacity by using spatial diversity technique at 2.4 GHz operating frequency. The analysis on the experiment data shows the MIMO channel capacity increase by increasing antenna spacing.

The capacity of Weichelberger model is more suitable to the measured channel capacity compared to Kronecker model when physical effect were considered.

References

1. Jensen, M.A., Wallace, J.W.: A review of antennas and propagation for MIMO wireless communications (invited paper). *IEE Trans. Antennas Propagat.* **52**, 2810–2824 (2004)
2. Foschini, G.F., Gans, M.J.: On limits of wireless communication in a fading environment when using multiple antennas. *Wirel. Pers. Commun.* **6**(3), 311–335 (1998)
3. Duman, T.M., Ghrayeb, A.: *Coding for MIMO Communication Systems*. John Wiley & Sons Ltd, England (2007)
4. Poon, A.S.Y.: *A Spatial Channel Model For Multiple—Antenna Systems*. 2004
5. Anreddy, V. R & Ingram, M. A. (2006). Capacity of measured Ricean and Rayleigh indoor MIMO channels at 2.4 GHz with polarization and spatial diversity. In: Anreddy, V.R., Ingram, M.A. (eds.) *IEEE Wireless Communication and Networking Conference*, pp. 946–951 (2006)
6. Kermaol, J.P., et al.: Experimental investigation of multipath richness for multi-element transmit and receive antenna arrays. In: Kermaol, J.P., Mogensen, P.E., Jensen, S.H., Andersen, J.B., Frederiksen, F., Sorensen, T.B., Pedersen, K.I. (eds.) *IEEE Vehicular Technology Conference Proceedings*, vol. 3, pp. 2004–2008 (2000a)
7. Molina-Garcia-Pardo, J.-M., et al.: Polarized indoor MIMO measurements at 2.45 GHz. In: Molina-Garcia-Pardo, J.-M., Castillo O.I., Egea-Garcia, F., Juan-Llacer, L. (eds.) *IEEE Antennas and Propagation Society International Symposium*, pp. 5335–5338 (2007)
8. Almers, P., Bonek, E., Burr, A., Czink, N., Debbah, M., Degli-Esposti, V., Hofstetter, H., Kosti, P., Laurenson, D., Matz, G., Molisch, A.F., Oestges, C., Özcelik, H.: *Survey of Channel and Radio Propagation Models For Wireless MIMO Systems (Research Article)*. *EURASIP J. Wirel. Commun. Networking* **2007** (2007)
9. Botonjia, A.: “MIMO channelmodels”, Diploma Thesis, Examensarbete utfört i Elektronikdesign vid Linköpings Tekniska Högskola, Campus Norrköping
10. Yu, K., Ottersten, B.: Models for MIMO propagation channels, a review. In: *Wiely Journal on Wireless Communications and Mobile Computing Special Issue on Adaptive Antennas and MIMO Systems*, 2002-07-08
11. Richard J.E., Oscar F., Torres, R.P.: Empirical analysis of 2×2 MIMO channel in outdoor–indoor scenarios for BFWA applications. *IEEE Antennas Propag. Mag.* **48**(6) (2006)
12. Rappaport, T.S.: *Wireless Communications (Principle and Practice)*, 2nd edn. Prentice Hall of India, India (2007)
13. Schott, J.R.: *Matrix Analysis for Statistics*, p. 24. Wiley, New York (1997)
14. Liu, L., Hong, W.: *Characterization of Line-of-Sight MIMO Channel for Fixed Wireless Communications*. 2007
15. Pérez F. F., Mariño E. P.: *Modeling The Wireless Propagation Channel*. Wiley, New york (2008)
16. Özcelik, H.: *Indoor MIMO channel models*. PhD Thesis, submitted to Institute für Nachrichtentechnik und Hochfrequenztechnik of the Technische Universität Wien. Dec 2004

Chapter 7

Design of Multi-band Antenna for Wireless MIMO Communication Systems

M.M. Saad, M.N. Husain, M.Z.A. Aziz, A.R. Othman, K.A.A. Rashid and M. Senon

Abstract This paper present the design of band antenna for wireless Multiple Input Multiple Output (MIMO) communication system. The advantage of MIMO system is able to enhanced the overall wireless channel capacity. Besides system performances, the size and mobility of the communication devices are also contributed the successfulness of proposed wireless communication system. The multi-band antenna design has been discussed by many researchers for various application. This paper present the design of multiband antenna by using slot techniques. The design was simulated by using microwave CST studio. The results show that, the proposed antenna can operate at 2.4, 3.5 and 5.2 GHz frequency band.

7.1 Introduction

High capacity is very crucial in wireless as it is required for future application. It is because, the data transferred by users are becoming more larger nowadays. For example, in the early years the data transferred consists of only text or voice only. As time goes by, the user is demanding higher channel capacity as they wish to send voice with picture, or text with video or text, picture, and video at the same time. These data are required more capacity to be transferred. One of the main challenges in wireless communication is to gain high capacity in order to fulfil future application necessities. It is important in increasing the data transferred simultaneously. Today's channel suffers from attenuation due to multipath in the channel [1]. The increasing demand for capacity in wireless systems has motivated considerable research aimed at achieving higher throughput on a given bandwidth. It is proven in [2–5] that wireless communication systems using multiple antennas

M.M. Saad (✉) · M.N. Husain · M.Z.A. Aziz · A.R. Othman · K.A.A. Rashid · M. Senon
Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia
e-mail: majdi@psp.edu.my

at both transmitter and receiver sides, denoted as Multiple Input Multiple Output (MIMO) antenna systems, enable great enhancements of channel capacity compared to the single input single output (SISO) system.

From [1–8], it is mentioned that multiple antennas at both transmitter and receiver can improve the wireless channel capacity within the same bandwidth and power received. Besides improving performance in terms of higher channel capacity which enables a wireless device to transfer and receive data with higher data rate, MIMO also proven to reduce the multipath fading. Also, it is mentioned that multiple antennas are proven to introduce robustness against channel fading and interference.

Current communication devices are portable and multi-applications supported such as Bluetooth, Wi-Fi, GPS, and so much more. These applications are used different operating frequencies in order to avoid interferences between them. The conventional way to support this demand is by placing two or more antennas with different resonant frequency in that device. Clearly, this method is not efficient because integrating two or more antennas proves difficult as they are likely going to couple to each other causing degradation of the received or transmitted signals [9].

Placing two antennas in a device also bring in the space issue since it is usually designed to be a portable device and by placing multiple antennas in the device required more spaces. Another way to support multi-frequency demand without facing any of those problems is by using multi-band antennas which consist only one simple structure antenna that can support dual frequency ranges and it is clearly being more efficient.

As mentioned before, in MIMO system, at least two antennas for each operating frequency should be placed in the wireless device in order to support the requirement of the multi-application device. This is quite challenging since generally, according to [10], placing two antennas at the distance of half wavelength is necessary to achieve good isolation. Then, if the antennas are in large size, more space required for the spacing and for the antennas itself which are not practical and inefficient. Then, an efficient antenna especially for MIMO application is an antenna that compact and can support multi-frequency operation.

7.2 Antenna Design

There are several design steps that has been done in designing the multiband antenna in this project. Dual-band antenna which is Design **A** is designed first and the method approached in designing a dual-band antenna in this project is by embedding a pair of slots on the antenna's patch. There are several slot shapes that designed and studied and they are divided to three designs which are Design **A1(i)**, **A1(ii)**, and **A1(iii)**. All of the designs are founded by the shape of slot in Design **A1(i)** which is C-shaped slots. Design **A1(ii)** is the Three-shaped slot which are double-up the C-shaped slots and Design **A1(iii)** is continuous from the

Three-shaped slot, but instead of having the equal diameter at both upper and lower slots, this design has a different diameter of the upper and lower slot and produce a pumpkin-like shape, hence named as The Pumpkin-shaped slot.

7.2.1 C-Shaped Slot Antenna [Design A1(i)]

The antenna structure of C-Shaped Slot Antenna and all other designs are similar to the basic rectangular patch antenna which consist patch, substrate (dielectric), and the ground plane and the material of each plane are similar. The structure of antenna's geometry is as shown in Fig. 7.1. There is a pair of slots that in C-shaped which gives the additional second frequency, then makes the antenna act as a dual-band antenna.

The effects of the antenna parameters such as the slot width and the slot radius, position of slots, and the separation between the slots are studied. But, in this section only the studies that give significant impact on the design presented, which are slot width, \mathbf{a} , distance between slots, \mathbf{x} , and the slot radius, \mathbf{r} .

From Table 7.1, it shows that the parametric study of slot width and distance between slot respectively. The second resonance frequency, $\mathbf{f2}$ tends to shift larger as the slot gets wider and as the slots get further from each other. The parameter of the slot width, \mathbf{a} is studied from 0.5 to 2 mm, and the distance between slots, \mathbf{x} is studied from 1 to 29.5 mm as it is the minimum and the maximum value, but the result presented is the one that gives the significant impact on the result. In terms of return loss, it gets smaller as the slots get wider but, conversely become worse when the slots gets further from each other.

From the results, it shows that as the slot radius, \mathbf{r} increases, the second resonant frequency, $\mathbf{f2}$ shifting to left, which is smaller, and the return loss becomes increases. The parameter studied included all values that possible, but radius of 3 to 7 mm gives the significance effects to the response. From all the studies, it occurs that all three parameters only affect the second resonance frequency, $\mathbf{f2}$ while first resonance frequency, $\mathbf{f1}$ remain constant or not much difference. Table 7.1. shows the summary of all three antenna parametric studies.

7.2.2 Three-Shaped Slot Antenna [Design A1(ii)]

Three-Shaped Slot Antenna is the design by combining two of the C-shaped slots in the previous design and creates a slot that looks like the shape of the number three. It also has the same structure and material used as the dielectric and conductor material which are FR-4 and copper respectively. Figure 7.2 shows the antenna's structure and in this design, all the antenna parameters that studied are the same as the design A1(ii). However, only two of them are considered giving the significance effects to the response, which are slot width, \mathbf{a} and the slot radius, \mathbf{r} .

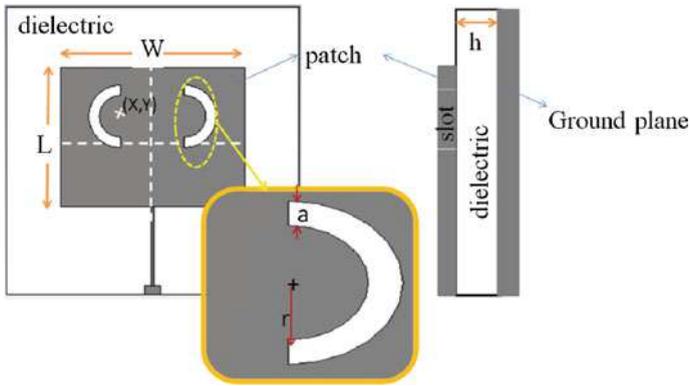


Fig. 7.1 C-Shaped slot antenna design

Table 7.1 Summary of the parametric studies of Design A1(i)

Parameters		f1 (GHz)	Return loss (dB)	f2 (GHz)	Return loss (dB)
Slot width, a (mm)	0.5	2.392	-26.33	5.158	-6.86
	1.0	2.386	-23.81	5.480	-12.13
	1.5	2.380	-23.38	5.590	-17.72
	2.0	2.374	-24.03	5.752	-20.77
Distance between slots, x (mm)	6	2.380	-23.25	5.368	-9.94
	8	2.380	-26.69	5.500	-14.52
	10	2.380	-23.38	5.590	-15.72
	14	2.380	-23.78	5.704	-19.11
Slot radius, r (mm)	3	2.404	-28.59	6.190	-21.84
	4	2.392	-28.77	6.022	-22.86
	5	2.380	-23.38	5.590	-15.72
	6	2.356	-20.56	3.196	-7.57
	7	2.338	-19.34	2.896	-4.29

Table 7.2 represents the analysis study of slot width, **a** that embedded on the patch and the studied are started with **a** equal to 0.5 until **a** equal to 2.0 mm. From figure, it shows that as slot width increases, the second resonance frequency, **f2** also increases, contrary to the return loss which decreases as the slots get wider.

Table 7.2 also shows the analysis study for the parameter of slot radius, **r**. The table shows that the second resonant frequency, **f2** will be increased as the radius increases. Noticed that the same effects occurred in the frequency of Design A1(i). As for the return loss parameter, it will decrease as the slot radius, **r** increases. All parameters are affecting only the upper frequency and not much impact to the lower one. Table 7.2 shows the summary of all parametric studies.

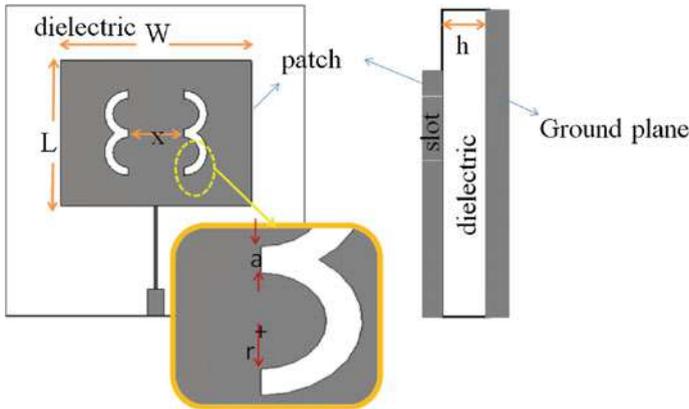


Fig. 7.2 Three-shaped slot antenna design

Table 7.2 Summary of the parametric studies of Design A1(ii)

Parameters		f1 (GHz)	Return loss (dB)	f2 (GHz)	Return loss (dB)
Slot width, a (mm)	0.5	2.434	-20.35	4.678	-12.42
	1.0	2.428	-20.89	4.634	-11.38
	1.5	24.160	-19.72	5.200	-40.26
	2.0	24.160	-20.04	5.530	-30.35
Slot Radius, r (mm)	3	2.434	-23.60	5.926	-20.21
	4	2.422	-20.48	5.278	-37.30
	5	2.404	-23.08	5.056	-29.17

7.2.3 Pumpkin-Shaped Slot Antenna [Design A1(iii)]

The Pumpkin-Shaped Slot Antenna is design based on Three-Shaped antenna. The only difference is the lower and upper part of the slot is not in the same radius which makes it looks like a pumpkin-shaped slot. This design has the same geometry of design structure and the same material used in the previous designs. Figure 7.3 shows the geometry of the design.

In this design, the parameters that give the significance impact to the response are the distance between the slots, x and the differences of slots radius, r . Table 7.3 shows that the same results are occurred as the slots get further from each other and the differences of slots radius become larger because both of them makes the second resonance frequency, f_2 shifting and become larger. However, it is not the same case for the return loss as it gets increased when the distance between slots, x increases and decreases when the differences of slots radius, r become larger, but still not affected by the distances between slots, x . Table 7.3 shows the overall analysis of all both parameters.

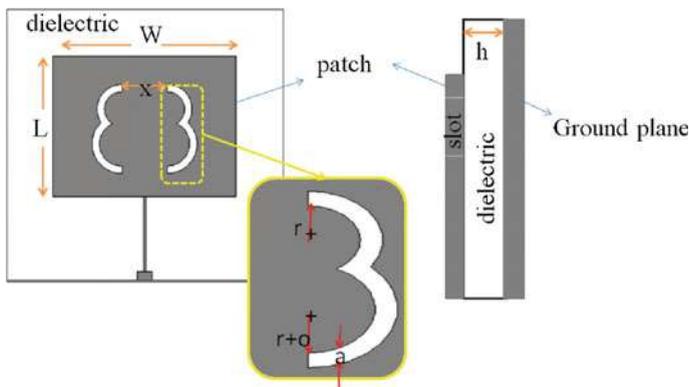


Fig. 7.3 Pumpkin-shaped slot antenna design

Table 7.3 Summary of the parametric studies of Design A1(iii)

Parameters		f1 (GHz)	Return loss (dB)	f2 (GHz)	Return loss (dB)
Slot Radius, r (mm)	0	2.506	-20.28	5.230	-9.47
	1	2.500	-12.74	5.326	-16.17
	2	2.500	-9.25	5.440	-22.99
Distance between Slots, x (mm)	5.75	2.512	-30.34	5.650	-37.51
	7.50	2.512	-29.40	5.554	-13.45
	9.25	2.506	-26.40	5.422	-5.58

7.3 Simulation and Measurement Results

In this section, the result of all dual-band antenna designs are presented and discussed. Table 7.4 shows the optimized value of design parameters for each antenna design. While Fig. 7.4 show the fabricated antenna that have been design. Figures 7.5, 7.6, 7.7 and 7.8 shows the simulation and measurement results of design antenna.

From the graph in Fig. 7.5, it shows that the most optimum design of Design A1(i) resulting resonance frequency of 2.386 GHz and 5.068 GHz whereas the design specification is designing dual-band antenna with resonance frequency 2.4 GHz and 5.2 GHz as **f1** and **f2** respectively. Because it is hard to achieve the specification using this design, it has been modified and improved as these next designs (Design A1(ii) and A1(iii)). From the graph, it also shows that there are shifting between the simulation and measurement resonance frequencies, **f1** and **f2** for Design A1(ii). That is may be caused by the dimension of fabricated design which is slightly inaccurate or not exactly same as the simulation. From the parametric studies of Design A1(ii) that have been performed earlier, it is noticed that the **f1** is dependence to the dimension of the patch, which is the width, **W** and

Table 7.4 Optimized dimension of Design A1

Parameter	A1(i)	A1(ii)	A1(iii)
Patch width, W	31.5 mm	33.8 mm	36.5 mm
Patch length, L	26.6 mm	26.12 mm	28.45 mm
Slot width, a	0.5 mm	1.5 mm	1.5 mm
Radii of slot, r	6.0 mm	4.2 mm	5 mm
Distance between slots, x	12 mm	5 mm	4.5 mm
Slot radius differences, o	NA		1 mm
Patch size changes	8.70 % (reduce)	3.80 % (reduce)	13.16 % (increases)
Coordinate of centre of circle, (X,Y)	(6,4)	NA	

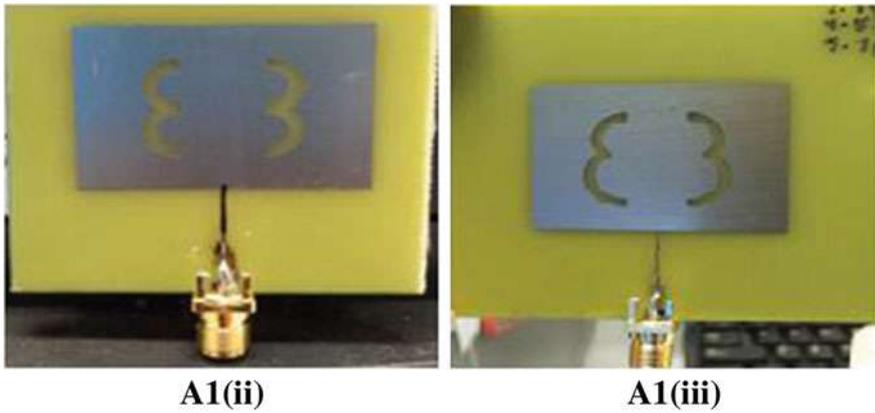


Fig. 7.4 Fabricated design of Design A1

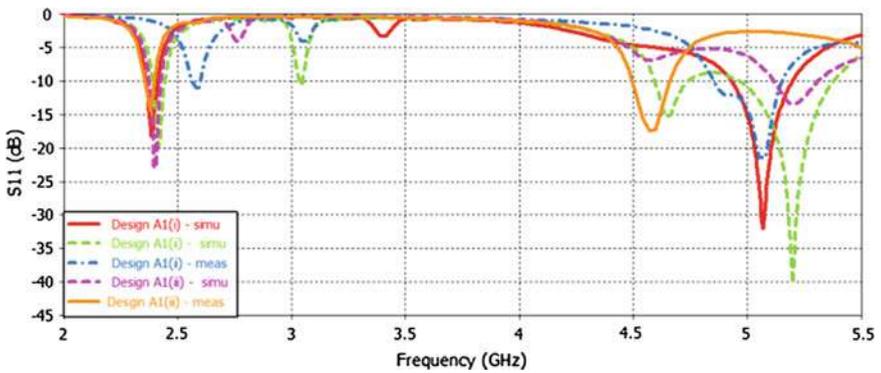


Fig. 7.5 Frequency response of Design A1

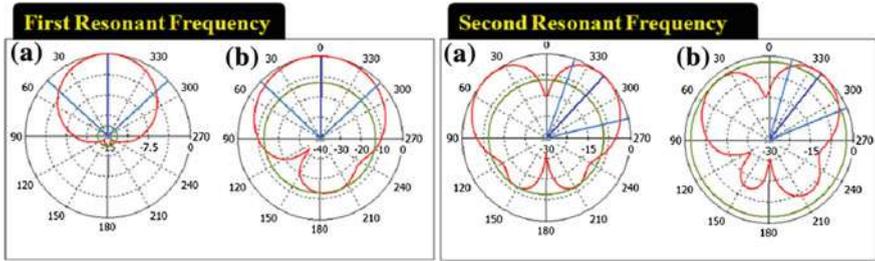


Fig. 7.6 Simulation radiation pattern of Design A1(i). a E plane. b H plane

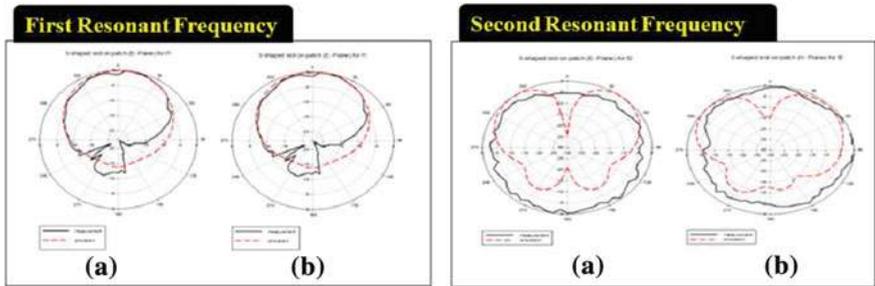


Fig. 7.7 Simulation and measurement radiation pattern of Design A1(ii). a E plane. b H plane

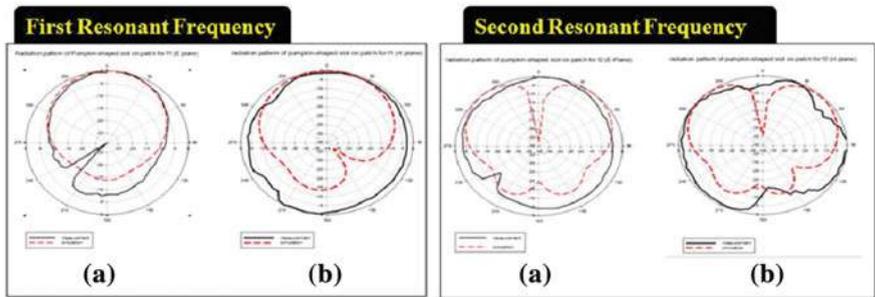


Fig. 7.8 Simulation and measurement radiation pattern of Design A1(iii). a E plane. b H plane

the length, L , while f_2 is dependence to the dimension and position of the slots. Due to lack of equipment, it is very difficult to maintain the accuracy of the dimension, especially at the slots part which only 1.5 mm width. Same case happened to Design A1(iii) but the frequency shifting are more greatly especially on second resonance frequencies, f_2 . That is also may be due to the inaccuracy of the slots dimension on the fabricated design. Consistent with the parametric studies that have been performed earlier which proved that f_2 is also dependence to the dimension and position of the slots.

As mentioned before, both Design **A1(ii)** and **A1(iii)** are the improvement and modification of Design **A1(i)** as it is hard to achieve the desired resonant frequency which are 2.4 GHz and 5.2 GHz. From Fig. 7.5, it seems like both Design **A1(ii)** and **A1(iii)** gives the similar result in their resonance frequency, but the Three-shaped Slot Antenna is slightly better than the Pumpkin-shaped Slot Antenna in terms of the return loss and bandwidth. The main factor that makes the Design **A1(ii)** is better is the size of the patch. Based on Table 7.4, Design **A1(ii)** gives size reduction by 3.89 percent but Design **A1(iii)** which is the Pumpkin-shaped Slot Antenna gives size increment by 13.16 % which are not fulfilling the design specification that required size reduction at least by 50 % of the basic rectangular microstrip patch designed earlier.

7.4 Conclusion

This paper present the design of multiband antenna by using C-shape slot techniques. Three design has been proposed based on different combination of C-shape slot. The design of double C-shape slot can produce the resonant frequency at 2.4 and 5.2 GHz. Besides that, the size of the double C-shape slot antenna can be reduced up to 50 % compared to the microstrip patch antenna. So, this design can be used to develop the compact mobile communication devices.

References

1. Daud, Z., Aziz, M., Suaidi, M., Rose, M., Kadir, M., Shah, M.: MIMO channel characterization and optimization. In: 6th National Conference on Telecommunication Technologies 2008 and 2008 2nd Malaysia Conference on Photonics, NCTT-MCP 2008. pp. 132–135 (2008)
2. Stuber, G., Barry, J., McLaughlin, S., Li, Y., Ingram, M., Pratt, T.: Broadband MIMO-OFDM wireless communications. *Proc. IEEE* **92**:271–294 (2004)
3. Keusgen, W.: On limits of wireless communications when using multiple dual-polarized antennas. In: 10th International Conference on Telecommunications, ICT 2003, pp. 204–210 (2003)
4. Jungnickel, V., Pohl, V., Nguyen, H., Kruger, U., Haustein, T., von Helmolt, C.: High capacity antennas for MIMO radio systems. In: The 5th International Symposium on Wireless Personal Multimedia Communications, vol. 86, pp. 407–411 (2002)
5. Nasr, A., Molina, J., Lienard, M., Degauque, P.: Optimisation of antenna arrays for communication in tunnels. In: 3rd International Symposium on Wireless Communication Systems, ISWCS '06, pp. 522–524 (2006)
6. Winters, J.: On the capacity of radio communication systems with diversity in a Rayleigh fading environment. *IEEE J. Sel. Areas Commun.* **5**:871–878 (1987)
7. Golden, G., Foschini, C., Valenzuela, R., Wolniansky, P.: Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture. *Electro. Lett.* **35**:14–16 (1999)

8. Ozdemir, M., Arslan, H., Arvas, E.: A mutual coupling model for MIMO systems. *IEEE Topical Conf. Wirel. Commun. Technol.* 306–307 (2003)
9. Matsunaga, M., Kakemizu, K., Candotti, M., Matsunaga, T.: An omni-directional multi-polarization and multi-frequency antenna. In: *IEEE International Symposium on Antennas and Propagation (APSURSI)*, pp. 2765–2768 (2011)
10. Luo, Q., Salgado, H., Pereira, J.: Printed C-shaped monopole antenna array with high isolation for MIMO applications.: In: *Antennas and Propagation Society International Symposium (APSURSI), 2010 IEEE*, pp. 1–4 (2010)

Chapter 8

Design of Linear Polarization Antenna for Wireless MIMO Application

K.A.A. Rashid, M.N. Husain, A.R. Othman, M.Z.A. Aziz,
M.M. Saad, M. Senon, M.T. Ahmad and J.S. Hamidon

Abstract This paper present the design of the linear polarized antenna for wireless MIMO communication system. It is impossible to fulfil the demand of the wireless communication system due to limitations in channel capacity on single input single output (SISO) systems. Multiple input multiple output (MIMO) system has become a famous research field for the next generation wireless communication system in order to overcome this problem. Since polarization diversity is effective to avoid the fading loss caused by multipath effects, therefore, polarization diversity becomes one of the most important techniques that can be used to enhance MIMO system performances. It can be utilized to improve the communications channel capacity and utilize the frequency spectrum with frequency reuse technique. Therefore, the development of linear polarized antenna is significant in order to improve the wireless MIMO system performance based on polarization diversity technique. Polarization diversity can be utilized to double the frequency spectrum to realize frequency reuse and improve the communications capacity. This project is to design an antenna that can provide linear polarization to reduce the signal losses.

8.1 Introduction

Antenna plays a crucial role in telecommunication field such as satellite communication, mobile phone and for military use [1]. The growth of mobile communications results in the increasing demand of smart phones, wireless internet and other broadband applications. The demand for high data rate and high capacity are increasing to satisfy the growth of mobile communications [2, 3].

K.A.A. Rashid (✉) · M.N. Husain · A.R. Othman · M.Z.A. Aziz · M.M. Saad · M. Senon · M.T. Ahmad · J.S. Hamidon
Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: P021110002@utem.edu.my; kariffin@psp.edu.my

The MIMO technology is become popular nowadays as the demand for high speed, high capacity and high quality transmission in wireless mobile telecommunication are increasing [3–5]. MIMO technology provide a very high spectral efficient and also increase the channel capacity without extra spectrum by using multiple transmitter and receiver [3, 5–7]. The loss of spectral efficiency occurs due to the spatial correlation between antennas [6, 7]. In order to decrease the spatial correlation, the spatial diversity is the most common used technology in MIMO system. The distances between MIMO antennas have to be at least half wavelength apart thus larger spacing is required for this technique [3, 7]. MIMO capacity increased proportionally with the number of antennas and thus higher spatial distance between antennas is required [3]. The mutual coupling between antennas occurs due to the narrow space for MIMO antenna design is the critical problem in MIMO system [1]. As a result, the polarization diversity becomes better solution for MIMO system.

In the modern wireless communication, the products such as smart phones and radio frequency identification (RFID) tags are designed in light weight and very compact size. The purpose of compact size and light weight enable the devices to be carried to everywhere. The antenna needs to be designed in a very small size so that the antenna is able to fit inside the compact size products. Moreover, the wireless devices are design to be portable; the antenna designed must be able to fit into the small device and are practical use. Thus, the size of an antenna is also one of the main criteria to take into account when designed the antenna [8–10].

8.2 Design Specification

The antenna is design for the frequency of 2.4 GHz, which is useful for the Bluetooth application. The 2.4 GHz frequency is an unlicensed band and is free for the public. The antenna is simulating using FR4 board and PEC or copper. FR4 board with thickness of 1.6 mm, tangent loss of 0.019 S/m and permittivity of 4.4 is used as the substrate. While the PEC with thickness of 0.035 mm is used as the conductor.

In this project, the antenna design is targeted to have gain from range of 2–4 dB for each of the polarization state. This is to avoid the problem of one of the polarization state provide higher gain and become the dominant mode. The range of the gain is referring to the works of previous researchers. Meanwhile the axial ratio for a circular polarization antenna must be below -2.5 dB and for linear polarization antenna must be above -3 dB. For this project, the antenna is design to fulfill the compact issue in terms of size by combined the radiator for both polarization states by using 2 ports. The radiation efficiency and the total efficiency for both polarization states must be below -3 dB to ensure that the antenna is able to transmit at least 50 % of the radiate power respectively. The matching is affect by the total efficiency, as the total efficiency getting closer to positive value the matching is getting better. While for return loss, the result must be below -10 dB to ensure that the transmission power of the antenna is more than 90 %. The specification of the design is listed at Table 8.1.

Table 8.1 Main specification of the design

Specification	Value
Frequency	2.4 GHz
Return loss	<-10 dB
Gain	2-4 dB [42, 51, 52]
Radiation efficiency	<-3 dB
Total efficiency	<-3 dB
Multipolarization	Linear polarization
Axial ratio	>3 dB for linear polarization

8.2.1 Antenna Design Process

The linear polarization antenna is first design to be a rectangular patch. The patch is then added double H-shaped slots. Lastly, the return loss of the design is then improved by adding stair notch to the patch.

8.2.2 Rectangular Patch with Slots

The rectangular patch is first drawn in CST microwave studio by using PEC (grey colour) and FR4 as the substrate (blue colour). The patch is connected to a CPW fed line. Then the slots are drawn on top of the patch and then substrate the slots from the patch as shown in Fig. 8.1. After the drawing of the patch is done, the setting for the simulation is set as same as the previous design. The effects of the slots dimension are obtained using parametric study method. From previous parametric study, the performance of the design is under performance. Thus, the gap of the ground and CPW fed are studied (Fig. 8.2).

From Table 8.2, the spacing between the CPW fed and the ground has significant effect on the return loss of the design. The return loss start to increased until $b = 0.6$ mm. In addition, spacing does not affect much on the axial ratio and gain of the design.

The parameter of c does not affect much on the axial ratio and gain of the antenna as shown in Table 8.3. However, the axial ratio of the antenna is decreasing as the c increasing. While for the return loss of the antenna is unstable when the c is increased (Table 8.4).

Slot4 length, d does not affect much on axial ratio and gain of the design. For return loss of the design, it did not have a stable trend as the d increased.

From Table 8.5, the return loss of the design do not showed a stable pattern. The return loss slightly increased after $e = 4$ mm. Furthermore, axial ratio and gain of the design do not affect much by the slot2 length. The gain only differs around 0.2 dB (Table 8.6).

The width of slot3, f does not affect much on the gain of the design. The return loss of the design decreased when the width getting wider. But the axial ratio

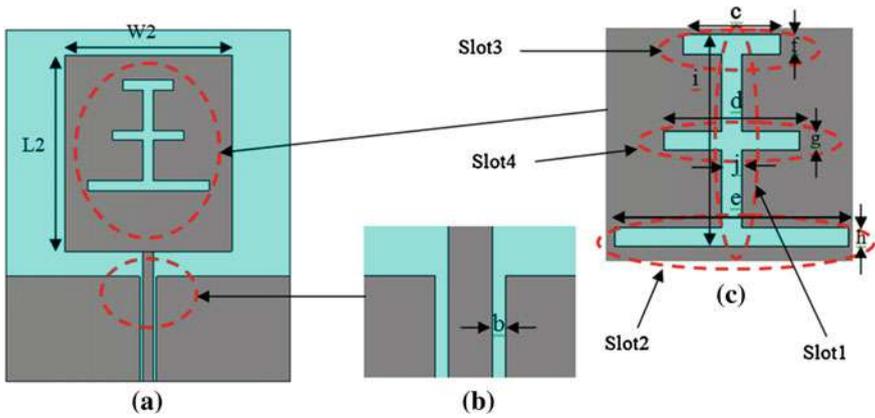


Fig. 8.1 Front view of the rectangular patch with slots

Fig. 8.2 Back view of the rectangular patch with slots

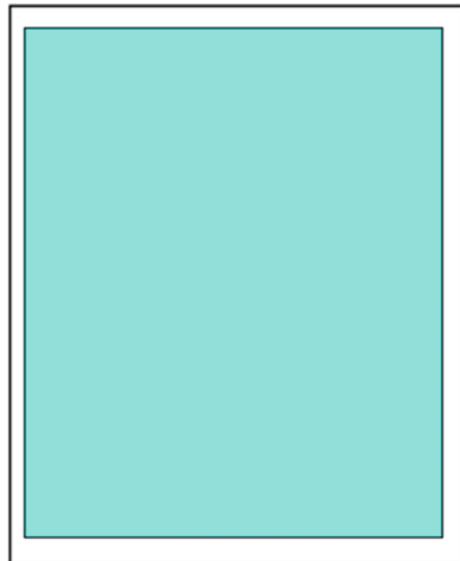


Table 8.2 Effects of spacing (b)

b (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-8.57	11.51	0.73
0.1	-9.36	11.57	0.77
0.2	-10.33	11.54	0.66
0.3	-13.98	11.53	0.97
0.4	-15.49	11.55	0.97
0.5	-19.95	11.57	0.96
0.6	-27.74	11.59	0.96
0.7	-17.62	11.63	0.94

Table 8.3 Effects of slot3 length (c)

c (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-15.61	11.93	1.25
2	-14.56	11.81	1.02
4	-18.64	11.68	0.87
6	-16.65	11.64	0.94
8	-15.14	11.57	1.01
10	-13.73	11.52	1.05

Table 8.4 Effects of slot4 length (d)

d (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-15.61	11.93	1.25
2	-14.85	11.85	1.08
4	-18.74	11.81	0.91
6	-16.83	11.85	1.02
8	-15.37	11.78	1.09
10	-13.98	11.71	1.11

Table 8.5 Effects of slot2 length (e)

e (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-15.61	11.93	1.25
2	-14.63	11.79	1.02
4	-18.69	11.87	0.86
6	-16.63	12.01	0.98
8	-15.21	11.91	1.01
10	-13.85	11.86	1.09

Table 8.6 Effects of slot3 width (f)

f (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-17.88	11.63	0.93
0.5	-18.36	11.64	0.99
1	-18.59	11.64	1.02
1.5	-18.76	11.64	1.03
2	-18.77	11.63	1.03

remains the same as the slot3 width increasing. The gain is still smaller even the slot3 width increasing.

As showed in the Table 8.7, width of slot4 did not affect much on the axial ratio, gain and return loss. But still the gain is getting better as the width increasing. The axial ratio of the design is still a linearly polarized antenna.

Table 8.7 Effects of slot4 width (g)

g (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-16.11	11.81	1.06
0.5	-16.54	11.82	1.12
1	-16.76	11.85	1.18
1.5	-16.74	11.86	1.18
2	-16.68	11.87	1.18

From Table 8.8, the results showed that the width of slot2 does not affect much on the return loss, axial ratio and gain of the design. As the width of slot2 increase, the gain is only increased from 1.02 to 1.18 dB. The axial ratio is also increased around 0.26 dB from 11.79 to 12.05 dB (Table 8.9).

As the slot1 length increasing, the return loss is slightly increased. However, the gain is only decreased a bit from 5.37 to 5.21 dB. In addition, the axial ratio is increased around 0.2 dB and the antenna is still a linearly polarized antenna (Table 8.10).

As the slot1 width increased, the return loss is increasing. While for the gain of the design is decreasing as the width of slot1 increasing. The axial ratio of the design is increasing from 11.93 to 12.04 dB.

Table 8.8 Effects of slot2 width (h)

h (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-16.47	11.79	1.02
0.5	-16.75	11.84	1.10
1	-17.01	12.02	1.13
1.5	-16.79	12.04	1.17
2	-16.58	12.05	1.18

Table 8.9 Effects of slot1 length (i)

i (mm)	RL (dB)	AR (dB)	Gain (dB)
2	-19.18	9.00	5.19
4	-18.77	8.99	5.23
6	-19.11	8.98	5.18
8	-18.94	9.03	5.21
10	-16.72	8.79	5.31

Table 8.10 Effects of slot1 width (j)

j (mm)	RL (dB)	AR (dB)	Gain (dB)
0	-15.61	11.93	1.25
1	-12.33	11.89	0.79
2	-12.21	11.87	0.95
2.5	-11.83	12.09	1.04

8.2.3 Result and Discussion

The final design of linear polarization antenna consists of a rectangular patch with double H-shaped slots and the design is added with stair notches. Figures 8.3 and 8.4 are the front view and back view of the linear polarization final design antenna. Table 8.11 is the optimum dimensions of the linear polarizes antenna.

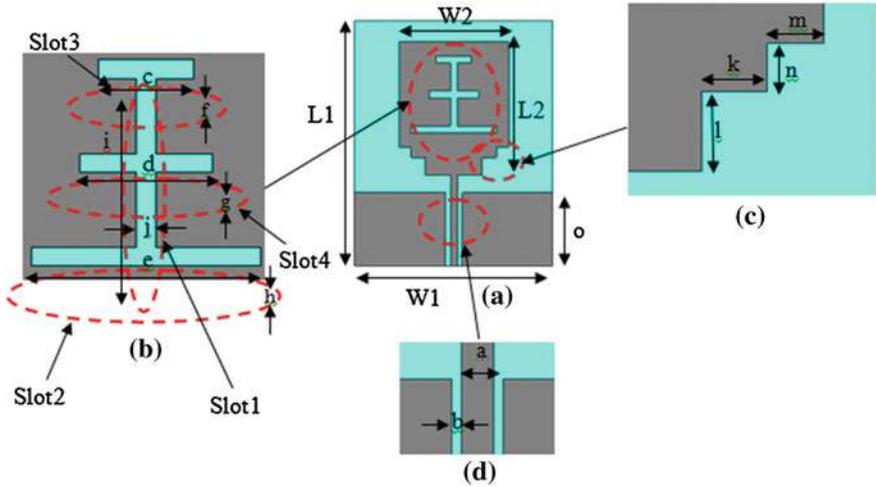


Fig. 8.3 Front view design of final linear polarization antenna

Fig. 8.4 Back view design of final linear polarization antenna

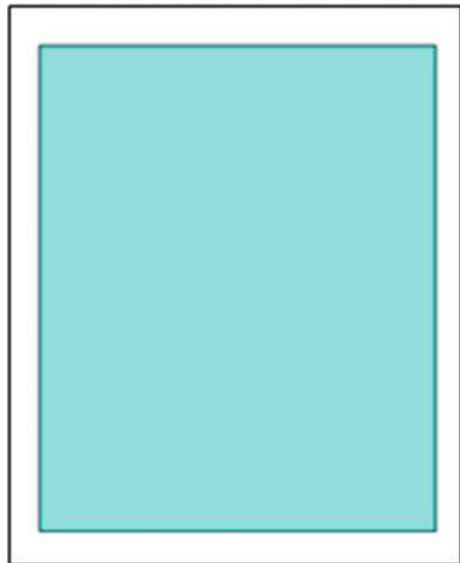


Table 8.11 Optimum dimension of linear polarization antenna

	Value(mm)
L1	35
W1	28
L2	19
W2	15.4
A	1
B	0.7
C	5
d	7
e	12
f	1
g	1
h	1
i	11
j	1
k	3.7
l	2.5
m	1.7
n	1.5
o	10.5

Figures 8.5, 8.6 and 8.7 are the comparison graphs of return loss, axial ratio and gain between the rectangular patch, rectangular patch with slots and rectangular patch with slots and notches. The return loss is improved with the used of notches from -8.9 dB at frequency of 2.398 GHz to -38.7 dB at frequency of 2.401 GHz.

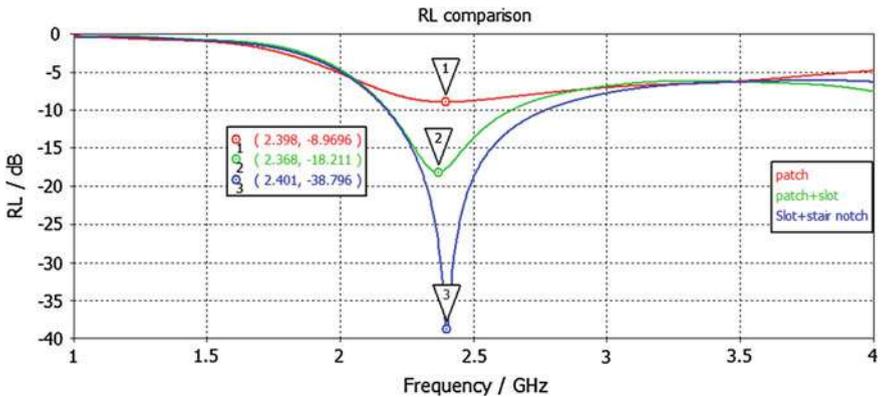


Fig. 8.5 Return loss comparison

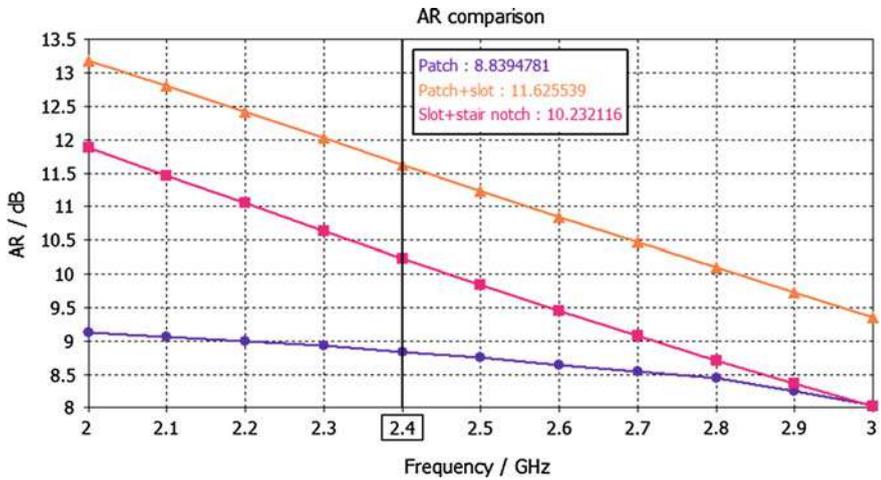


Fig. 8.6 Axial ratio comparison

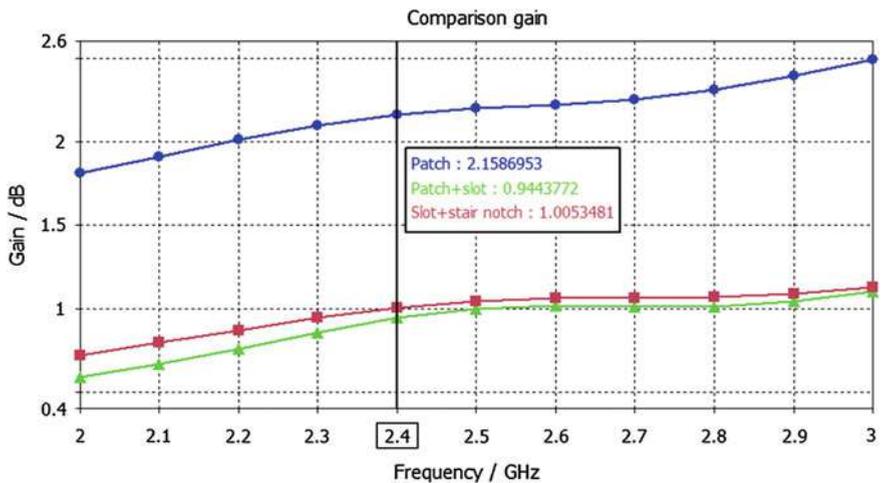


Fig. 8.7 Gain comparison

However, the gain of the patch with slots and notches is decreased from 2.16 to 1 dB. The axial ratio of these three antennas is all linear polarized antennas. The return loss is improved with slots and notches, but the gain is degraded.

Figure 8.8 is the comparison of return loss between the measurement and the simulation for linear polarized antenna. The measured return loss consists of ripple at the front and the end of the response. The bandwidth of the measured return loss is much wider than the simulation.

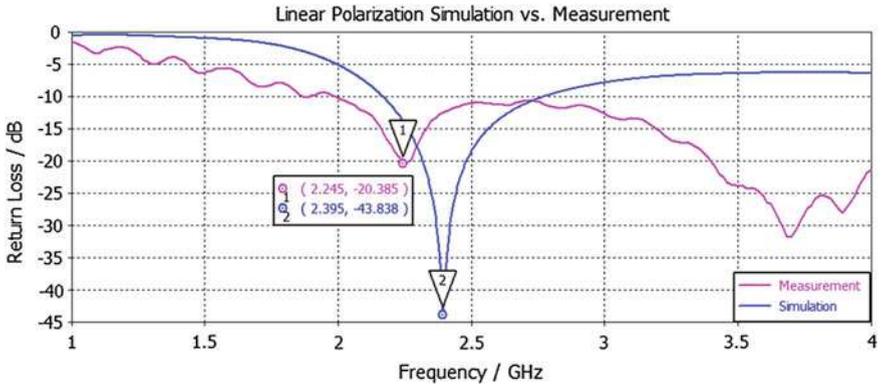


Fig. 8.8 Return loss comparison for simulation and measurement (linear polarization)

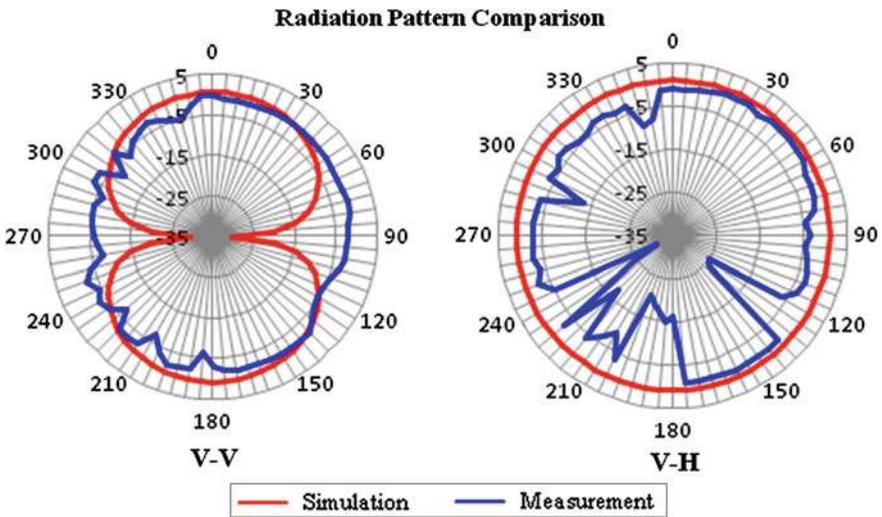


Fig. 8.9 Radiation pattern comparisons for linear polarization

Figure 8.9 is the radiation pattern for linear polarization in terms of vertical and horizontal orientation. The simulated radiation pattern for vertical is a ‘8’ shape, however the measured radiation pattern is more like a circle shape. While, the radiation pattern for horizontal orientation, the shape of the measured radiation pattern is almost the same with simulated radiation pattern (Fig. 8.10).

Type	Farfield
Approximation	enabled ($kR \gg 1$)
Monitor	farfield (f=2.4) [1]
Component	Abs
Output	Directivity
Frequency	2.4
Rad. effic.	-1.035 dB
Tot. effic.	-1.036 dB
Dir.	2.041 dBi
Gain	1.005 dB

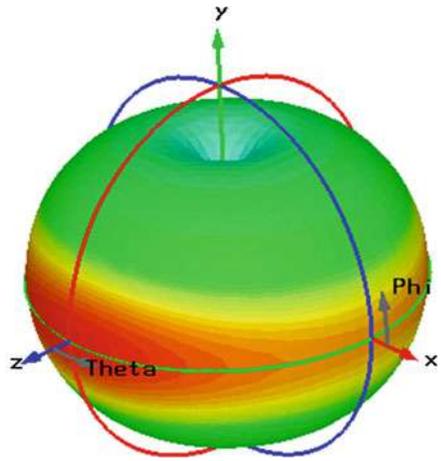


Fig. 8.10 Gain and directivity

8.3 Conclusion

This paper presents the design of Linear Polarization for wireless MIMO communication system. Wireless channel capacity seems very important nowadays in order to support the broadband applications. Then, the design of linear polarized antenna has been done. Rectangular patch with slots technique been discussed. The linear polarized antenna can be used in developing the polarization diversity in wireless MIMO communication system.

References

1. Chang, L.C., Tsai, C.H., Hsu, P., Liu, C.C.: A polarization diversity MIMO antenna design for WiMAX dongle application. In: Microwave Conference Proceedings (APMC), 2010 Asia-Pacific, pp. 762–765, Dec. 2010
2. Vergerio, S., Rossi, J.-P., Sabouroux, P.: A two PIFA antenna systems for mobile phone at 2 GHz with MIMO applications. *EuCAP 2006*, 1–5 (2006)
3. Elliot, P.G., Rosario, E.N., Davis, R.J., Rzhonov, A.E.: MIMO polarization diversity antenna with ultra-wide bandwidth and small size. In: IEEE International Symposium Phased Array Systems and Technology (ARRAY), 2010, pp. 559–566, Oct. 2010
4. Wang, X., Chen, W., Feng, Z., Zhang, H.: Compact dual-polarized antenna combining printed monopole and half-slot antenna for MIMO applications. In: IEEE APSURSI'09, pp. 1–4, June 2009
5. Hu, S., Pan, J., Qiu, J.: A compact polarization diversity MIMO microstrip patch antenna array with dual slant polarizations. In: IEEE APSURSI'09, pp. 1–4, June 2009
6. Chae, S.h., Choi, M.G., Park, S.O.: The realization and analysis of polarization diversity in WiBro MIMO antenna. In: Antennas and Propagation Society International Symposium, 2007 IEEE, pp. 2409–2412, June 2007

7. Chae, S.h., Choi, M.G., Park S. O.: The realization of pattern/polarization diversity by applying vertical excitation. In: TENCON 2007, pp.1–4, Nov. 2007
8. Lu, Y.C., Wu, Y.S., Chiu, C.T., Lin, Y.C.: A novel compact design of USB dongle antenna for bluetooth applications. In: IEEE RWS'09, pp. 31–34, Jan. 2009
9. Lee, C.J., Gummalla, A., Achour, M.: Compact dualband antenna subsystem for MIMO application. In: iWAT 2009, pp. 1–4, March 2009
10. Liu, X.F., Chen, Y.B., Jiao, Y.C., Zhang, F.S.: Design and optimization of a compact patch antenna. In: ICMMT'07, pp. 1–3, April 2007

Chapter 9

The Effect of IV Characteristics on Optical Control of SDR Si IMPATT Diode

T.S.M. Arshad, M.A. Othman, M.N. Hussain and Y.A. Rahim

Abstract Effect of light incident on the Si SDR IMPATT diode is investigated in this paper. The authors have used an IMPATT diode which consists of p+, n+ (contact region), n-well region and p-sub region. Since the n-well region is used to be the drift region of the structure, the light is shined on the top of the layer through tiny hole created on the SiO₂ layer. The results of the IV characteristics are compared to the dark current condition which no light will be supply on top of the structure. The result of the electric field and mobility in those two conditions are also observed in this paper.

9.1 Introduction

IMPATT diode is known as one of the most powerful millimeter wave sources. IMPATT stands for 'Impact Ionization Avalanche Transit Time' as it employ impact ionization and transit time properties of its semiconductor structures to produce negative resistance at microwave frequencies [1]. IMPATT diode has various structures such as mesa, planer, distributed and so on. Those structures consist of contact and drift region that are stacked perpendicular to the substrate

T.S.M. Arshad · M.A. Othman (✉) · M.N. Hussain · Y.A. Rahim
Microwave Research Group, Faculty of Electronics and Computer Engineering, Centre for Telecommunication Research and Innovation (CeTRI), Universiti Teknikal Malaysia Melaka (UTeM), Durian Tunggal, Melaka, Malaysia
e-mail: azlishah@utem.edu.my

T.S.M. Arshad
e-mail: teh.suriati@gmail.com

M.N. Hussain
e-mail: drmohdnor@utem.edu.my

Y.A. Rahim
e-mail: yahaya@utem.edu.my

surface [2]. Since this paper are focus on the optical control of IMPATT structure, the lateral structure has been choose since this structure are more suitable in coupling optical energy into the active region of the device. Compared to those structures, lateral IMPATT structure has a unique design which is the location of the contact and the drift region adjacent to the substrate [3].

The optical of microwave semiconductor devices have a large growing interest because of their very wide bandwidth, the inherent high dc and reverse signal isolation between the control and RF signals, and it also suitable to use with optical fiber links [4]. The basic operation of semiconductor devices when it relates to the direct optical control is the electron hole pairs (EHP) within the active region of the device when sufficient light photon energy is absorbed. When the dimension of depletion regions is modified, the main effect is to generate a photocurrent and change the built-in potential of the junction. Hence, the conductivity of the semiconductor material increases effect by the photoconductivity [4].

The authors proposed a lateral DDR IMPATT structure which suitable for millimeter-wave optical interaction [2]. The structure is designed to operate at 94 GHz frequency by using standard 0.18 μm CMOS technology. The structure consists of p+ and n+ layers (contact layers) and drift layers (p- and n-epitaxial layers) that adjacent to the substrate. By using this type of structure, light can be easily coupled to the drift layers through tiny holes created on the SiO₂ layer [5]. The paper shows two type of optical control which on p+ layer (to generate electron dominated photocurrent) and n+ layer (to generated hole dominated photocurrent). By comparing those two types of optical control, the hole dominated photocurrent has large value of efficiency than electron dominated photocurrent.

An indium phosphate IMPATT diode under optical illumination at 94 GHz has been proposed. The effect of the photo-generated predominated electron or hole components of the leakage current on the admittance and negative properties are also has been reported in the paper [6]. The result presented in the paper shows that the frequency shift as well as the admittance and negative resistance characteristics are more sensitive to hole-dominated photocurrent than to electron-dominated photocurrent [6]. However, result in the paper [7] shows that the DC and high-frequency parameters of the device are more sensitive to electron-dominated photocurrent (Top Mount structure) compared to the hole-dominated photocurrent (Flip Chip structure). It also observed that the avalanche response time decreases as more number of photon are incident of the active region of the device [7].

The authors proposed a lateral SDR IMPATT diode in an earlier paper [8] which will be use in this present paper. The simulation study will carry out to investigate the optical control performance. This present paper will focus on the effect of IV characteristics when dark current and has light incident. The authors also have made an attempt in this paper to study the effect of the electric field and mobility of the device in these two different conditions.

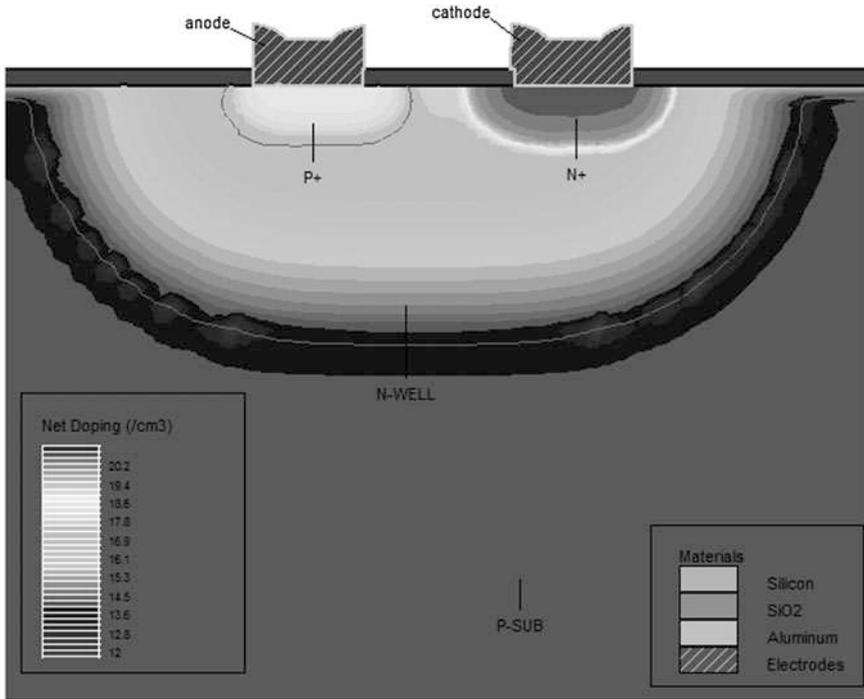


Fig. 9.1 Structure of SDR Si IMPATT diode

9.2 Design Structure

The structure proposed in this paper is design by using Silvaco TCAD ATHENA tools. Athena provides a suitable platform for simulating processes used in semiconductor industry which are ion implantation, diffusion, oxidation and so on [9]. So, it is a convenient tool to design the proposed IMPATT diode since this structure follows the standard fabrication process of complementary metal-oxide-semiconductor (CMOS) technology.

Figure 9.1 show the SDR Si IMPATT diode consists of p+ and n+ layer (contact layers), n-well region and p-sub region. The base material for this structure is silicon. This structure is designed for a particular frequency (f_d) from transit time formula [9] given by

$$W_{dep} = 0.37 v_s / f_d \tag{9.1}$$

where W_{dep} , v_s and f_d are total depletion layer width (n or p side), saturation velocity of electrons/holes and design frequency respectively. At first step, the doping concentration of n-well and p-sub region is initially chosen according to

Table 9.1 Design parameter for SDR Si IMPATT diode

W_n (μm)	N_D (m^{-3})	N_{n+}, N_{p+} (m^{-3})
0.16	5×10^{17}	1×10^{20}

previous paper [2–4]. Later the doping concentration and design parameter for both regions are adjusted to get the corresponding frequency design. Table 9.1 show the structure design of SDR Si IMPATT diode.

9.3 Simulation Result and Discussion

9.3.1 Dark Current Condition

This paper is focus on the effect of the current-voltage characteristics (IV characteristics) of the device in two different conditions which is in dark current (no light incident) and in light incident. Since IMPATT diodes are operated in reverse biased voltage, the simulation process supplies negative voltage to anode contact. By using the same structure that show in Fig. 9.1, the IV characteristic is obtained from the simulation process shown in Fig. 9.2. The result is for dark current condition which is no light incident on the surface of structure. From the graph, it shows that the breakdown voltage is -5 V and it produce value of current in range of $-6\text{e-}14$. The small value of current is because it is in reverse biased condition.

Figure 9.3 show the electric field of the structure when there's no light supplied on the structure. The value for the electric field is in the range from -83.5679 to -91272.6 V/cm. The values of electric field are shown in Table 9.2.

The simulation also has carry out the result for e- mobility and h+ mobility of the structure. The results are shown in Fig. 9.4. From the graph, it shows that the electron mobility are higher than hole mobility. From here [10], it state that the mobility of electrons and hole are normally in less or equal to $1,400$ and 450 cm^2/Vs respectively. Therefore, the values from the result are in same range since it used silicon for the structure. The values plotted in Fig. 9.4 are summarized in Table 9.2. Later, the value in Table 9.2 will be compared to the value in the next section which is simulated with light incident.

9.3.2 Light Incident Condition

Based on the structure in previous section, the light energy can be easily coupled into n-well region through tiny holes created on the SiO_2 which is located on the top of the structure. It can be shown in Fig. 9.5. From previous paper [11], there are two types of optical control which either on p+ layer or n+ layer. If light energy

Fig. 9.2 IV characteristics for SDR Si IMPATT diode (dark current)

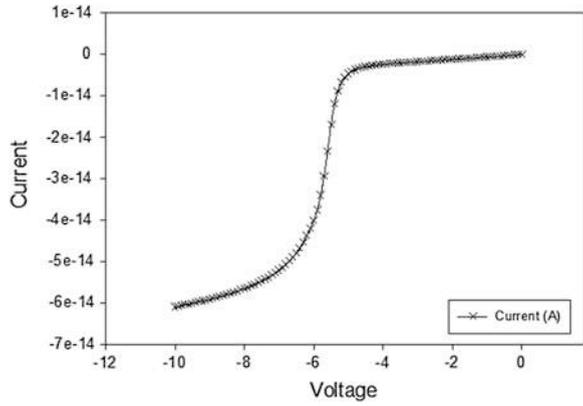


Fig. 9.3 Electric field for SDR Si IMPATT diode (dark current)

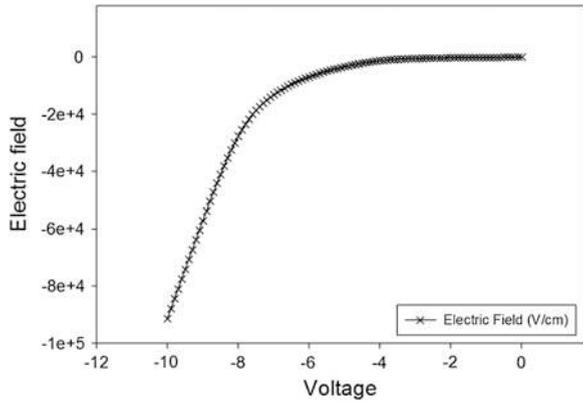


Table 9.2 Electric field, e-mobility and h+ mobility for Si IMPATT diode (no light)

Voltage (V)	Electric field (V/cm)	e- mobility (cm^2/Vs)	h+ mobility (cm^2/Vs)
0	-83.5679	449.435	252.094
-1	-120.611	449.432	251.867
-2	-254.553	449.411	251.049
-3	-535.642	449.317	249.348
-4	-1283.64	448.742	244.934
-5	-3393.47	444.636	233.286
-6	-6979.23	430.117	215.84
-7	-13159.1	390.217	191.198
-8	-27674.0	287.471	150.769
-9	-57083.8	168.137	105.548
-10	-91272.6	109.945	78.2606

Fig. 9.4 Mobility for SDR Si IMPATT diode (dark current)

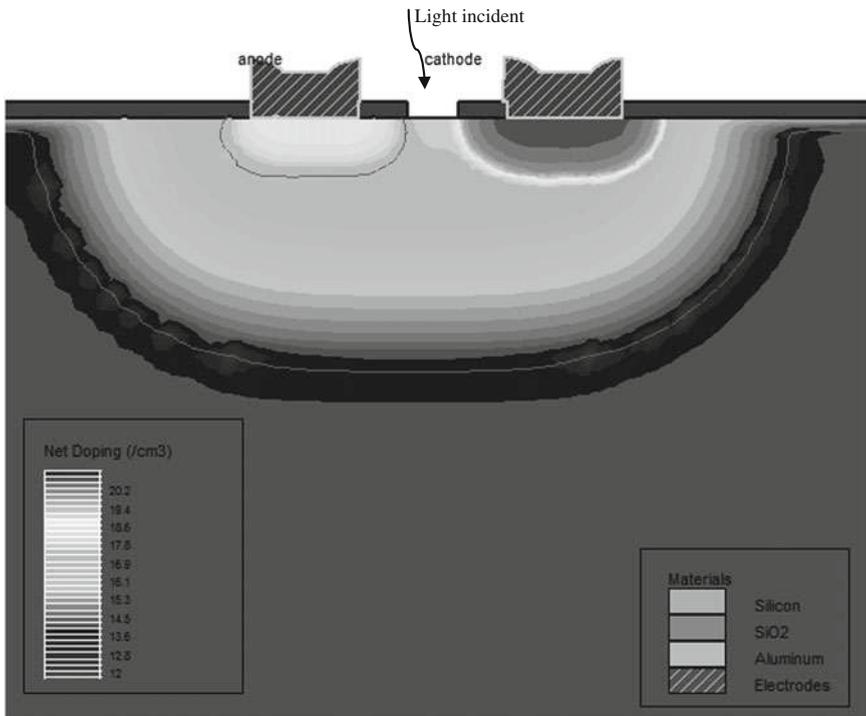
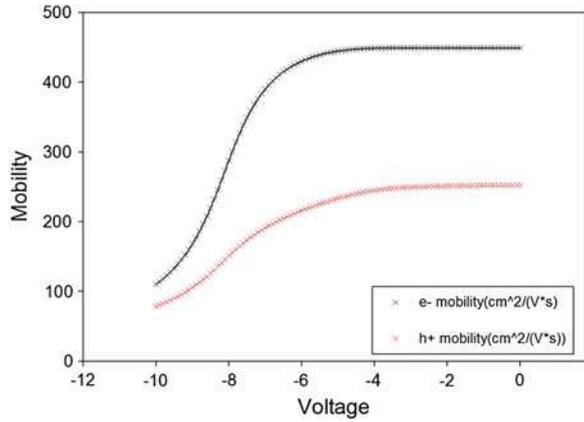
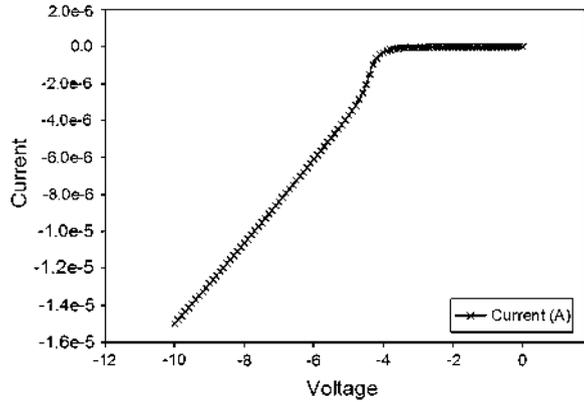


Fig. 9.5 Structure of SDR Si IMPATT diode (light incident)

is supply to the p+ layer, it will generate electron dominated photocurrent and if the light energy supply to the n+ layer, it will produce hole dominated photocurrent [5].

Fig. 9.6 IV characteristics for SDR Si IMPATT diode (light incident)



In this section will be presented on the effect of IV Characteristics in light incident condition. Since the light will be incident on the top of n-well layer (drift layer), the photo-generated current will be dominated by hole. From the equation given below,

$$\lambda_g = hc / E_g \tag{9.2}$$

where $h = 6.625 \times 10^{-34}$ Js, $c = 3 \times 10^8$ ms⁻¹, the corresponding cut off wavelength can be obtained which is $\lambda_g = 1.118 \mu\text{m}$ since the bandgap of Si at 300 K is $E_g = 1.111$ eV [12]. Therefore, the simulations use the wavelength of light less or equal to $1.118 \mu\text{m}$ for the absorption of light energy by the Si structure. This process will generate electron-hole pairs (EHP) in Si structure.

The simulations for light incident condition start with the observation of IV characteristics. The same value of voltage supply from previous section is used which is -10 V. From the graph from Fig. 9.6, it shows breakdown voltage of the structure is -3.7 V. From the observation, the presence of light reduced the voltage breakdown to produce conductivity.

The presence of light on semiconductor devices is exactly similar to the effect of temperature or heat on semiconductor devices. Light energy causes electrons to break their covalent bonds and then generated EHP [13]. The conductivity of the device is increase since the light energy decrease the resistance of the structure. Compared to the dark current condition in previous section, the resistance is high because there are few free electrons in the device. The resistance in this condition is called dark resistance [13]. So, it takes more times to generate EHP and produce conductivity (Table 9.3).

Figure 9.7 shows the graph of electric field of the device in presence of light. The value of electric field is almost similar compared to previous section. This is because the value of voltage supply uses in these two different conditions is same. The observation of electron and hole mobility can be shown in Fig. 9.8. There are also slightly similar values of these two mobility compare to previous section.

Table 9.3 Electric field, e-mobility and h+ mobility for Si IMPATT diode (with light)

Voltage (V)	Electric field (V/cm)	e- mobility (cm ² /Vs)	h+ mobility (cm ² /Vs)
0	-79.4343	449.436	252.120
-1	-96.1578	449.434	252.017
-2	-175.775	449.425	251.529
-3	-381.337	449.377	250.279
-4	-812.410	449.159	247.697
-5	-2047.64	447.672	240.584
-6	-4896.17	439.611	225.643
-7	-9674.43	414.355	204.354
-8	-19067.0	346.172	172.383
-9	-41523.6	217.968	125.457
-10	-75877.4	130.518	88.5718

Fig. 9.7 Electric field for SDR Si IMPATT diode (light incident)

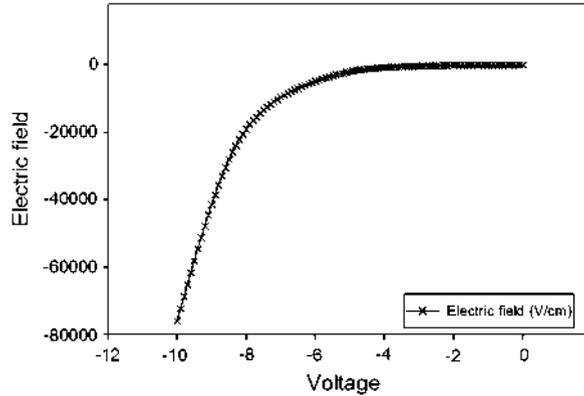
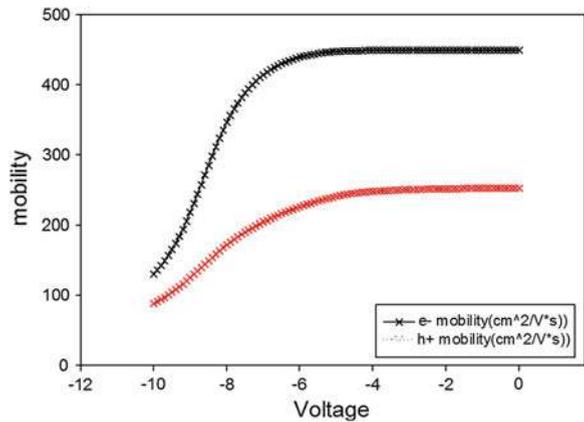


Fig. 9.8 Mobility for SDR Si IMPATT diode (light incident)



There are only slightly increased for mobility in light incident because the light energy causes the ions vibrating with sufficient movement to impede the movement of electrons. For this cases, the light that supplied on the structure give small effect to the lattice scattering to increase and the same time it is not affect to the reduction of mobility.

9.4 Conclusion

In this paper, the authors have made an attempt to investigate the effect of photo incident on IV characteristics of SDR Si IMPATT diode. This paper also has shown the result of electric field and mobility of the structure. The results are compared with dark current condition. From the IV curve, it shows that the breakdown voltage in light incident condition are faster than dark current condition. This is because the presence of light is similar to the effect of temperature. When light is supply on the structure, it causes electron to break their covalent bonds. With the energy of light, it will generate EHP and increase its conductivity. So, in light incident condition, it required few voltage supplies to start the conduction process. So, for the conclusion, the optical control is affect to the IV characteristics of the devices.

Acknowledgments The authors wish to thank Dato' Prof Dr. Mohd Nor Bin Husain for the professional guidance and useful critiques of this research work. I would also like to extend my thanks to Mr. Mohd Azlishah Bin Othman for the suggestion and patient guidance of this work. My thanks must also go to Mr. Yahaya bin Abdul Rahim (PJP/2012/FTMK(56D)S1063) and Mr. Zul Atfyi Fauzan bin Napiah (PJP/2012/FKEKK(13C)S01116) for their valuable support.

References

1. Srivastava, G.P., Gupta, V.L.: Microwave Devices and Circuit Design, p. 294. PHI Learning Pvt. Ltd, New Delhi (2006)
2. Acharyya, A., Banerjee, J.P.: A proposed lateral DDR IMPATT structure for better millimeter-wave optical interaction. In: International Conference on Devices, Circuits and Systems, March 2012, pp. 599–602
3. Stabile, P.J., Lalevic, B.: Lateral IMPATT diodes. *IEEE Electron Device Lett.* **10**(6), 249–251 (1989)
4. Seeds, A.J., Augusto, A.: Optical control of microwave semiconductor devices. *IEEE. Trans. Microw. Theory Tech.* **38**(5), 577–585 (1990)
5. Acharyya, A., Banerjee, S., Banerjee, J.P.: Optical control of millimeter-wave lateral double-drift region silicon IMPATT device. *Radioengineering* **21**(4), 1208–1217 (2012)
6. Banerjee, J.P., Mukherjee, R.: Effect of electron- and hole-dominant photocurrent on the millimetre wave properties of an indium phosphide IMPATT diode at a 94 GHz window under optical illumination. *Semicond. Sci. Technol.* **9**, 1690 (1994)

7. Acharyya, A., Banerjee, J.P.: Dependence of avalanche response time on photon flux incident on DDR silicon IMPATT devices. In: The 32nd PIERS in Moscow, Russia, pp. 867–872, August 19–23, 2012
8. Arshad, T.S.M., Othman, M.A., Yasin, N.Y.M., Taib, S.N., Napiyah, Z.A.F.M., Hussain, M.N., Rahim, Y.A., Pee, A.N.C., Ismail, M.M., Misran, M.H., Said, M.A.M., Sulaiman, H.A., Ramlee, R.A.: Variable junction temperature analysis in silicon IMPATT diode. In: 2013 3rd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME), pp. 76,79, 7–8 Nov. 2013
9. SILVACO TCAD ATHENA, http://www.silvaco.com/products/tcad/process_simulation/athena
10. Electronic Achieve: New Semiconductor Materials, Characteristics and Properties. <http://www.ioffe.ru/SVA/NSM/Semicond>
11. Vyas, H.P., Gutmann, R.J., Borrego, J.M.: The effect of hole versus electron photocurrent on microwave—Optical interactions in IMPATT oscillators. *IEEE Trans. Electron Device*. **26**(3), 232–234 (1979)
12. Varshni, Y.P.: Temperature dependence of the energy gap in semiconductor. *Physica* **34**, 149–154 (1967)
13. Bakshi, U.A., Godse, A.P.: *Semiconductor Devices & Circuits*, p. 25. Technical Publications, India (2008)

Chapter 10

Variable Intrinsic Region in CMOS PIN Photodiode for I–V Characteristic Analysis

M.A. Othman, N.Y.M. Yasin, T.S.M. Arshad, Z.A.F.M. Napiah, M.M. Ismail, H.A. Sulaiman, M.H. Misran, M.A. Meor Said and R.A. Ramlee

Abstract In this paper presented an investigation on I–V characteristic for CMOS PIN Photodiode. PIN diodes are widely used in optics and microwave circuits as it acts as a current controlled resistor at these frequencies. PIN diode performance is greatly influenced by the geometrical size of the device, especially in the intrinsic region. Two different I-layer thickness of PIN diode structure has been designed using Sentaurus Technology Computer Aided Design (TCAD) tools. The I-layer

M.A. Othman · N.Y.M. Yasin (✉) · T.S.M. Arshad · Z.A.F.M. Napiah · M.M. Ismail · H.A. Sulaiman · M.H. Misran · M.A. Meor Said · R.A. Ramlee
Faculty of Electronic and Computer Engineering, Centre for Telecommunication Research and Innovation (CeTRI), Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: yashidar@yahoo.com

M.A. Othman
e-mail: azlishah@utem.edu.my

T.S.M. Arshad
e-mail: yat_lampard08@yahoo.com

Z.A.F.M. Napiah
e-mail: zulatfyi@utem.edu.my

M.M. Ismail
e-mail: muzafar@utem.edu.my

H.A. Sulaiman
e-mail: asyrani@utem.edu.my

M.H. Misran
e-mail: harris@utem.edu.my

M.A. Meor Said
e-mail: maizatul@utem.edu.my

R.A. Ramlee
e-mail: ridza@utem.edu.my

thickness (or width) is varied from 4 to 8 μm in order to investigate its effects on the current-voltage (I - V) characteristics. These structures were design based on CMOS process.

10.1 Introduction

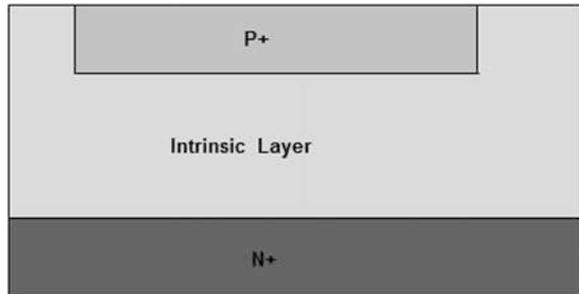
Nowadays a lot of microwave and optical devices have been design by academician, scientist and engineers. One of two terminal devices that get intention is PIN diode. PIN diodes name attribute to overall their structure where P is P type layer, I intrinsic layer and N is N type layer. The intrinsic layer is interesting layer since this layer makes the PIN diodes application comes to be as attenuators, RF switches and photodiode. Hence, PIN diodes are used extremely in RF and microwave applications due to ability to control the magnitude and phase of the signals [1, 2]. In addition, the ability to control RF and microwave signals while using a smaller level of dc excitation makes the PIN diode suitable for attenuating, limiting, phase shifting, modulating and microwave switching.

The I-layer or intrinsic layer is the one that gives a changes in-term of properties comparing with PN diode. The intrinsic layer comprises of undoped or virtually undoped semiconductor and in most PIN diodes and very thin between 1 μm up to 200 μm . PIN photodiode is a special case of the PN junction photodiode, in which a large intrinsic or lightly doped N semiconductor area is inserted in between the P and N region as shown in Fig. 10.1.

It is a well-known fact that in any PN diode, the depletion region extends more into the lightly-dope N region than into the heavily doped P region. This is because, in the heavily doped region, the number of free charge carriers available for conduction is quite large compared to that in the lightly doped region [3]. In a normal operation a reverse-bias voltage is applied across the device so that no free electrons or holes exist in the intrinsic region (Fig. 10.2).

Electrons in semiconductor materials are allowed to reside in only two specific energy bands. A forbidden region called energy gap separates the two allowed band.

Fig. 10.1 PIN photodiode structure



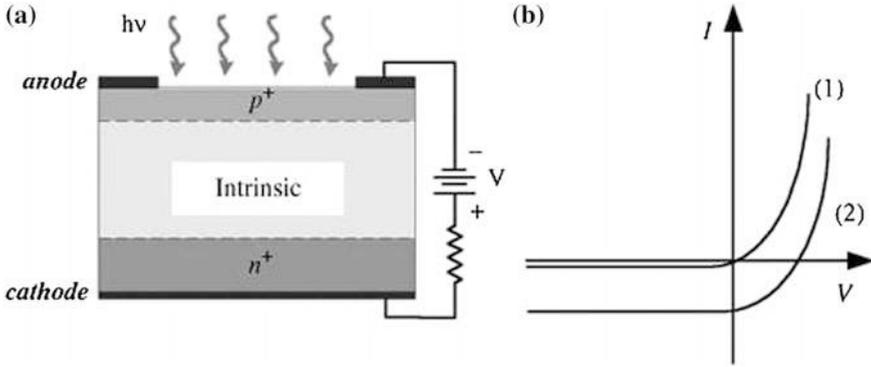


Fig. 10.2 (a) PIN photodiode operated in reverse bias, $h\nu$ is the energy of radiation, V is the bias, and RL is the local load resistance (b) I - V characteristics for a pin photodiode (I) with no light (2) with light

The energy difference between the top and bottom bands is referred to as the band gap energy [4]. In a generic photodiode, light enters the device through a thin layer whose absorption typically causes the light intensity to exponentially drop with penetration depth. For enhanced performance it is often necessary for the device to have a shallow junction followed by a wide depletion region where most of the photon absorption and electron-hole generation should take place.

The magnitude of the generated current is proportional to the intensity of the incident light. It can be used as a photon detector by operating it in the third quadrant of its electrical current-voltage (I - V) characteristics. PIN photodiode is usually functioning by applying a reverse-bias voltage. The magnitude of the reverse-bias voltage depends on the photodiode purpose, but naturally is less than a few volts. While no light is incident on the photodiode, a current is still produced [5]. The objectives of this paper are to study the theory behind PIN photodiode, to design and analyze PIN-Photodiode and to determine characteristics of PIN-Photodiode.

10.2 Design Procedure

The simulations were performed using Silvaco ATLAS. Silvaco ATLAS is a TCAD product that is physics based on modeling system. This allows TCAD to predict device performance based upon equations, which describe the physics within the structure of the device [6].

The structure under consideration is lightly doped with 5×10^{16} for N-epi doping, 1×10^{20} for N+ doping, 1×10^{20} for P+ doping, sandwiched with I+ doping with concentration 1×10^{21} . The incident light is absorbed in neutral P+ region and neutral N+ region as well as in the depleted N- region. All the calculations presented in

Fig. 10.3 Intrinsic with $2 \times 8 \mu\text{m}$

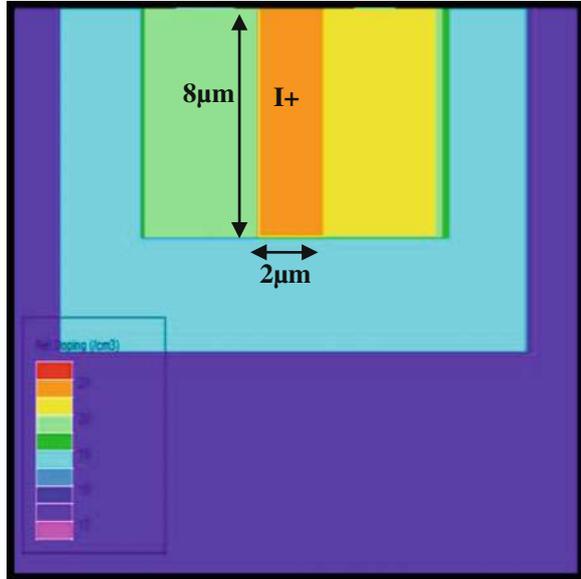
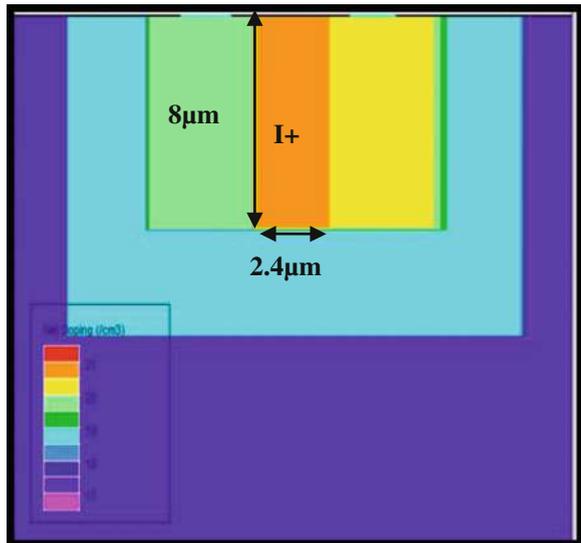


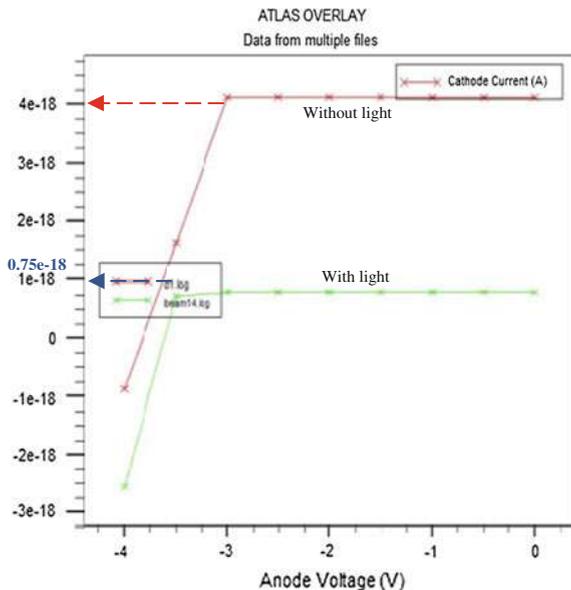
Fig. 10.4 Intrinsic with $2.4 \times 8 \mu\text{m}$



this work have been obtained with ATLAS software of SILVACO unless stated otherwise. The intrinsic region size had been varied with 2 different sizes which is 4 and $8 \mu\text{m}$ in width.

ATLAS enables device technology engineers to simulate the electrical, optical, and thermal behavior of semiconductor devices. Different models such as conmob, fldmob, srh, auger and bgn are used for effective implementation of changes occurring related to radiation damage [7] (Figs. 10.3, 10.4).

Fig. 10.5 Intrinsic with $2.4 \times 8 \mu\text{m}$



As shown in the figure, the intrinsic width for both structures had been increased to $0.4 \mu\text{m}$. By capturing most of the light in the depletion region (intrinsic layer), it will effect to improve the responsivity of the device. To fully benefit from the intrinsic layer or depletion region, the PIN Photodiode normally requires an applied reverse bias that ensures a depletion region extension all the way through this layer. Figures 10.5, 10.6 and 10.7 shows the simulated I–V characteristic of Silicon PIN Photodiode for the structure that had been created. The simulated results of p+, i+, n+ photo detector are obtained by developing a program in DECKBUILD window, interfaced with ATLAS, simulation software of SILVACO and TonyPlot.

In a generic photodiode, light enters the device through a thin layer. For Fig. 10.5, it shows the difference of IV characteristic when supply light and without supply light for $2.4 \times 8 \mu\text{m}$. The IV characteristic of a photodiode with no incident light is similar to a rectifying diode. When the photodiode is forward biased, there is an exponential increase in the current. When a reverse bias is applied, a small reverse saturation current appears. The light has the effect of shifting the IV Characteristics down into the fourth quadrant where power can be extracted from the photodiode.

Figure 10.6 show the difference of IV characteristic when supply light and without supply light for $2 \times 8 \mu\text{m}$. Basically, it use the same theory with the result show in Fig. 10.5, the difference is the size of depletion use based on the structure. For IV Characteristic in Fig. 10.5, it uses huge depletion size compared to IV Characteristics in Fig. 10.6. That is why the effect of shifting of IV Characteristics

Fig. 10.6 Intrinsic with $2.4 \times 8 \mu\text{m}$

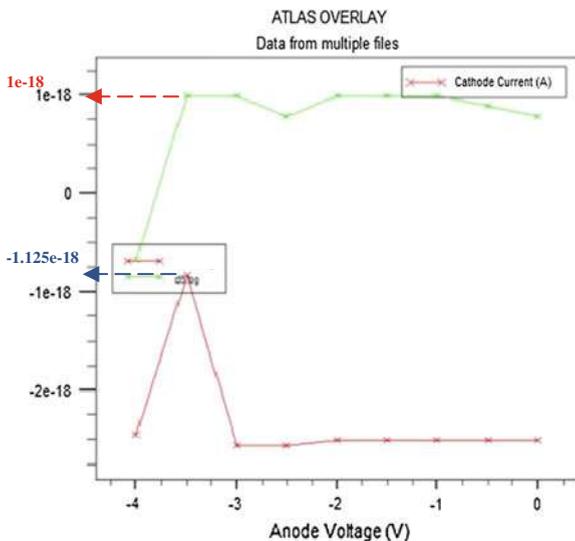
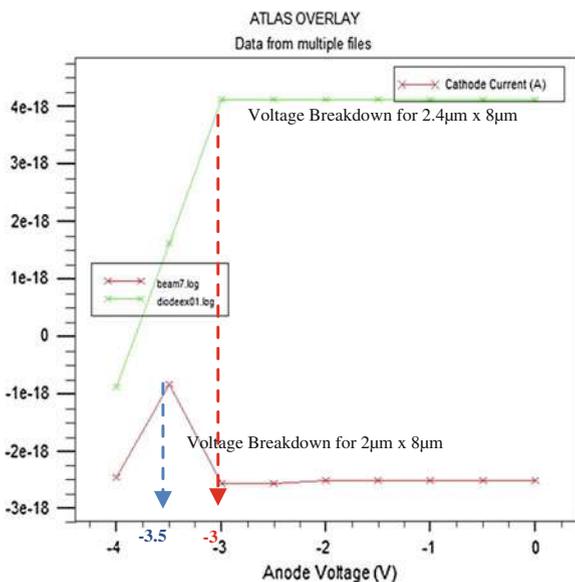


Fig. 10.7 Voltage Breakdown for $2 \times 8 \mu\text{m}$ and $2.4 \times 8 \mu\text{m}$



from the photodiode is smaller compared to IV Characteristic in Fig. 10.5 due to difference structure used.

Based on Fig. 10.7, it shows the voltage breakdown for $2 \times 8 \mu\text{m}$ and $2.4 \times 8 \mu\text{m}$. There is a difference for the voltage breakdown between each structure. Voltage breakdown for intrinsic for $2.4 \times 8 \mu\text{m}$ is faster compared to Voltage breakdown for intrinsic for $2 \times 8 \mu\text{m}$. This is due to the depletion region in the structure since the size of intrinsic is not the same.

Based on graph in Figs. 10.5 and 10.6, the shape of graph had a little bit difference in their variance. This is due to the size of the mesh that had been varied since the size of intrinsic layer had been changed. Thus, the mesh for Intrinsic with $2 \times 8 \mu\text{m}$ becomes closer and that is why variance appeared.

10.3 Conclusion

In this study, in order to observe the width of intrinsic or depletion region effect on the IV Characteristics performance, the two-dimensional silicon PIN photodiode with varying width has been simulated using the SILVACO TCAD tools. Through the simulations, it shows that there is a difference occurs during voltage breakdown for difference intrinsic size. PIN Photodiode give effect when light supplied to the photodiode by shifting the IV Characteristics down. In general, PIN photodiode need small voltage supply to have a better I–V performance (voltage breakdown).

Acknowledgments Authors would like to thank Centre for Telecommunication, Research and Innovation (CeTRi), Universiti Teknikal Malaysia Melaka for their equipment, assistance support and also financing the conference's paper.

References

1. Yashchyshyn, Y.: Reconfigurable antennas by RF switches technology. In: 5th International Conference on Perspective Technologies and Methods in MEMS Design, pp. 155–157, April 2009
2. Yang, J.G., Kim, M., Yang, K.: An InGaAs PIN-diode based broadband travelling-wave switch with high-isolation characteristics. In: IEEE International Conference on Indium Phosphide and Related Materials, pp. 207–209, May 2009
3. Somanathan Nair, B.: Electronic Devices and Applications, 3rd edn, pp. 330, Sept (2006)
4. Keiser, G.: Optical Communications Essentials. McGraw-Hill, NY pp. 108 (2003)
5. Piprek, J.: Optoelectronic Devices: Advanced Simulation and Analysis. Springer Media and Business, Berlin, pp. 382 (2005)
6. Maiti, C.K., Armstrong, G.A.: TCAD for Si, SiGe and GaAs Integrated Circuits. The Inst. Eng. Technol. 47 (2007)
7. Vasileška, D., Goodnick, S.M.: Computational Electronics. Morgan and Claypool Publishers, US pp. 98 (2006)

Chapter 11

Variable Depletion Region in CMOS PN Photodiode for I–V Characteristic Analysis

M.A. Othman, T.S.M. Arshad, Z.A.F.M. Napiah, M.M. Ismail, N.Y.M. Yasin, H.A. Sulaiman, M.H. Misran, M.A. Meor Said and R.A. Ramlee

Abstract In this paper, CMOS PN photodiode will be design and analyze for the application at 5 GHz optical communication. The paper will be divided in several section; the theory of CMOS PN photodiode and design with analysis of I–V characteristics of PN photodiode. A better understanding of the operation will be investigated through this. The PN photodiode will be design using Silvaco TCAD

M.A. Othman (✉) · T.S.M. Arshad · Z.A.F.M. Napiah · M.M. Ismail · N.Y.M. Yasin · H.A. Sulaiman · M.H. Misran · M.A. Meor Said · R.A. Ramlee
Faculty of Electronic and Computer Engineering, Centre for Telecommunication Research and Innovation (CeTRI), Universiti Teknikal Malaysia Melaka, 76100 Hang Tuah Jaya, Durian Tunggal, Melaka, Malaysia
e-mail: azlishah@utem.edu.my

T.S.M. Arshad
e-mail: yatmie02@gmail.com

Z.A.F.M. Napiah
e-mail: zulatfyi@utem.edu.my

M.M. Ismail
e-mail: yashidar@yahoo.com

N.Y.M. Yasin
e-mail: muzafar@utem.edu.my

H.A. Sulaiman
e-mail: asyrani@utem.edu.my

M.H. Misran
e-mail: harris@utem.edu.my

M.A. Meor Said
e-mail: maizatul@utem.edu.my

R.A. Ramlee
e-mail: ridza@utem.edu.my

and will be characterize and experimental in I–V Characteristic. The effects of I–V characteristic will be analyzed in term of changes the width and light. Further understanding of I–V characteristic will be presents in this paper.

11.1 Introduction

PN junction is one of the most simple of all semiconductor devices because it is much faster and very small in size and weight. It is the only most important device in the studies of modern semiconductors. It is the heart of the most photocells, rectifiers and transistors. It is also cheaper and has greater sensitivity [1].

PN junction forms a diode, and consequently a junction used as a photodetector is frequently called a photodiode [2]. It is formed by doping donor atoms on one side (N-side) and doping acceptor atoms on the other (P-side). Figure 11.1 shows the structure of PN photodiode [3]. In this paper, the authors concentrate on the geometry of PN photodiode using different size of P+ region and its effects toward the I–V characteristics PN photodiode. The structure of PN photodiode is simulated with different size of P+ region and the analysis of I–V characteristic, which is use Silvaco TCAD tools discussed in the next topic.

The PN junction operated in the reverse biased condition. The PN photodiode operation involving light and the light will be directed fall on the upper top of the portion of the photodiode. The depth where photons come into the depletion region depends on the wavelength of incident light. Electron-hole pairs will exist in the depletion region when the absorbed optical energy is sufficient. Electrons and holes are now attracted towards the opposite respective terminals of the battery. EHP that formed outside the depletion region will produce a photocurrent [4]. When a reverse bias voltage is applied to the PN junction, more electrons and holes are attracted to the contact. As a result, more donor and acceptor ions appear at the depletion region which in turn increases as well [5] (Figs. 11.2 and 11.3).

11.2 Design Procedure

In this project, SILVACO TCAD tool is used to design the PN photodiode and simulate the I–V characteristic. ATLAS is a physically based device simulator, providing general capabilities for 2D and 3D simulation of semiconductor devices [6]. In the Atlas project, it divided into 2 parts, which is the structure and the I–V curve simulation. In the structure part, mesh is needed to generate meshes that are suitable for the design of PN photodiode. There are some methods and model that used in the second part of the coding. It will use for the simulation of I–V curve.

Fig. 11.1 PN junction photodiode

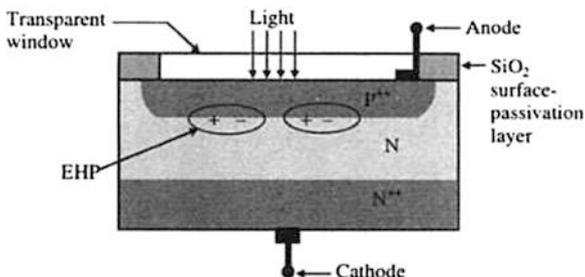


Fig. 11.2 I-V Characteristic of PN photodiode

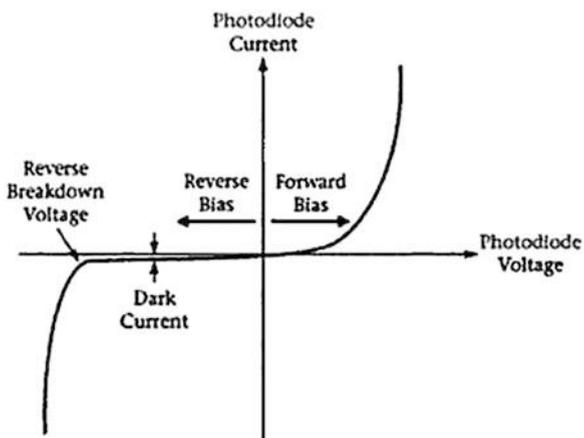
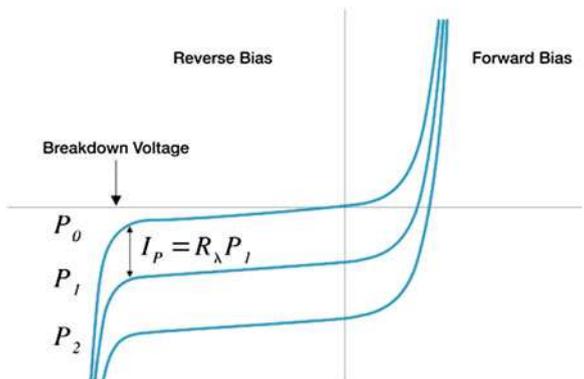


Fig. 11.3 I-V Characteristics with different light level



11.3 Result and Discussion

Figures 11.4 and 11.5 show the structure of the PN photodiode using Atlas tools. Two heavily doped P region and N region with doping concentration $1e + 20$ is located beside each other. Two contacts for PN photodiode called anode and

Fig. 11.4 PN Photodiode with width, $W = 1 \mu\text{m}$

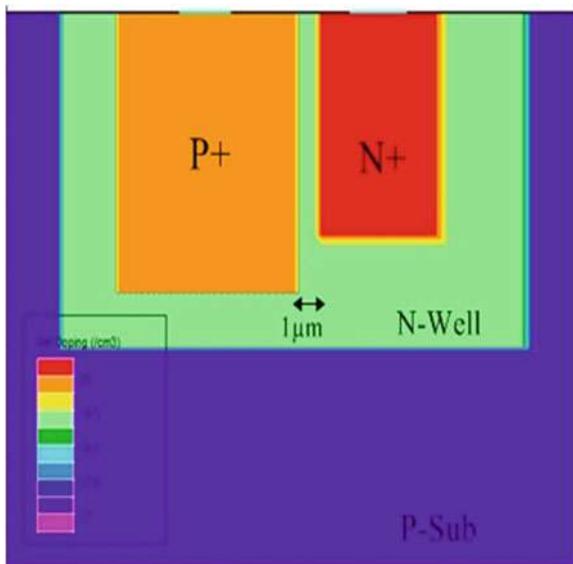
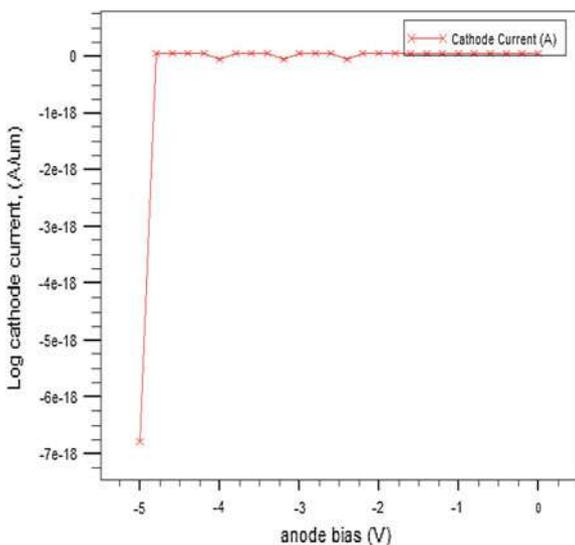


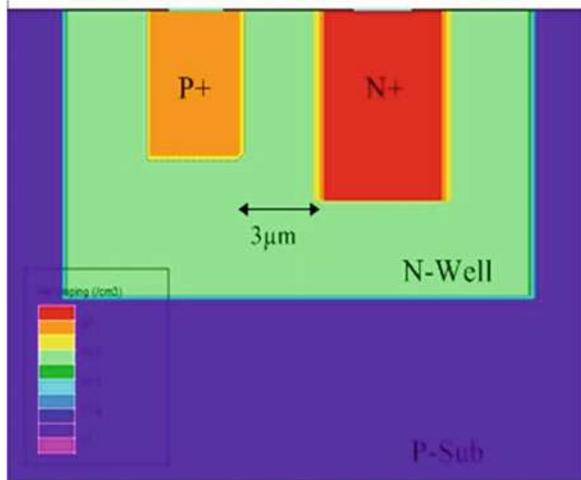
Fig. 11.5 I–V curve for PN Photodiode with width, $W = 1 \mu\text{m}$



cathode is located at the top of P-type and N-type region. It may be noted that N-type and P-type region is formed within the n-well region. There is also a p-sub region which is located below n-well.

The reduction of depletion region width with increased the electric field and reduced breakdown voltage. The breakdown voltage of the structure with width, $W = 3 \mu\text{m}$ is 4.9 V while the breakdown voltage for the structure with width,

Fig. 11.6 PN Photodiode with width, $W = 3 \mu\text{m}$



$W = 1 \mu\text{m}$ is 4.7 V. The applied electric field creates the depletion region on either side of the PN junction. Carriers—free electrons and holes—leave the junction area. The electric field and potential in the depletion region can affect the breakdown voltage. It can show from the equation below:

$$V_B = \frac{E_m W}{2} = \frac{\epsilon_S E_m^2}{2q} (N_B)^{-1} \tag{11.1}$$

From the Eq. (11.1), V_B is proportional to the W . As can be expected, reduce of depletion region can generate the small voltage breakdown. The structure that only uses less voltage supply will start the operation of reverse biased faster than the structure, which uses more voltage supply.

In reverse bias, the number of free carriers in the device depends primarily on the current in the device and the voltage is clamped at the knee voltage of the diode. Because of the I–V curve for the photodiode in reverse bias is exponential, a small change in V has a much larger impact to the device [7].

For the analysis of presence of light, the second structure is chose, shown in Fig. 11.6. From the analysis, some photons are incident on the surface of PN photodiode. Since the photon will pass through it and strike the junction. Light that is absorbed by the photodiode produces current flow through the entire external circuit (Fig. 11.7).

In the analysis, two different width of PN photodiode have been tested in the device structure which are 1 and 3 μm . By comparing both structures, the different of width is change due to the different size of P+. Since the size of P+ reduced, the width will also change. The I–V Curve is compiled together as it is easier to view the trend of voltage breakdown versus the width depletion region. Figure 11.8 illustrates the simulation for both structures. From the result, it is shown that both

Fig. 11.7 I–V curve for PN Photodiode with width, $W = 3 \mu\text{m}$

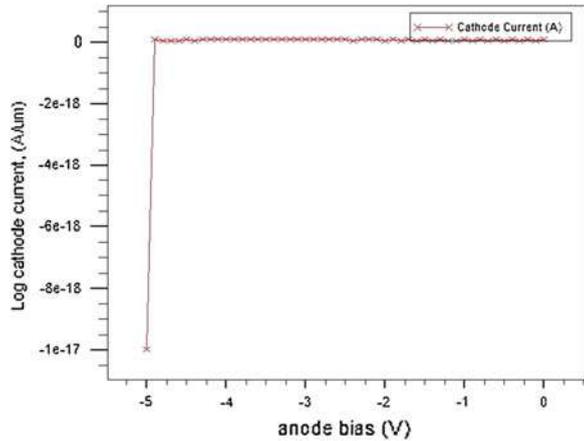
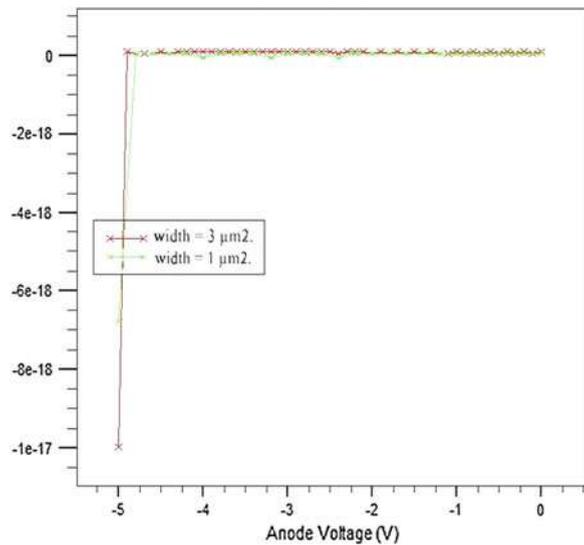


Fig. 11.8 I–V curve of PN Photodiode with different width



photodiode are start from $V = 0 \text{ V}$ and operated at the different reverse bias voltage.

The depth at which the photons reach into the depletion region depends on the incident of light. Figure 11.9 show the I–V curve with no light falling on the PN photodiode. When no light is falling on the photodiode, a very small current passes through the photodiode. This current is basically due to the reverse bias applied to the PN photodiode.

Figure 11.10 show the I–V curve with light falling on the PN photodiode. As the light is incident on the photodiode, photocurrent is developed. This photocurrent gets increased by increasing the light intensity.

Fig. 11.9 I-V curve for PN Photodiode without light

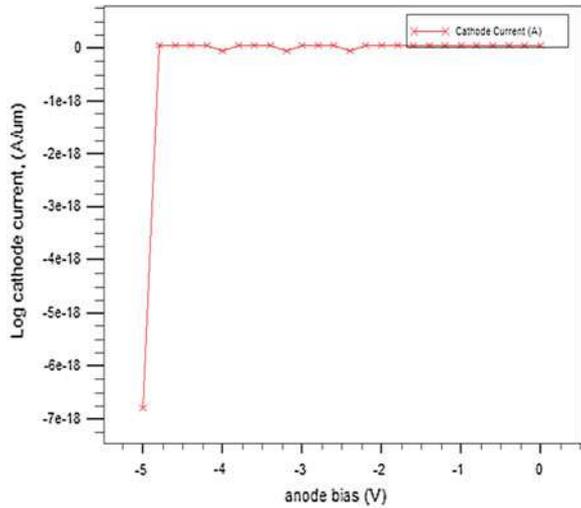
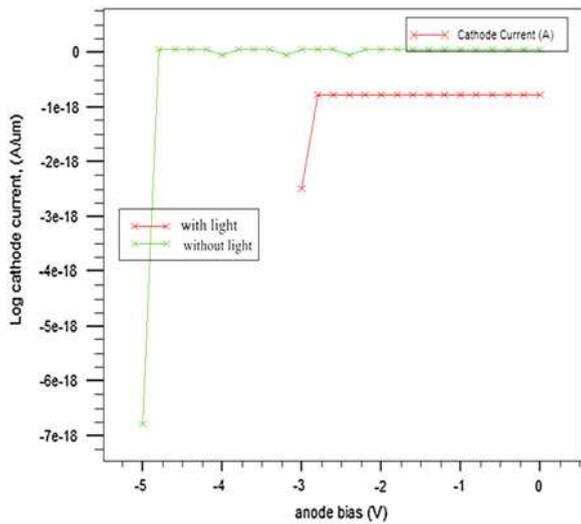


Fig. 11.10 I-V curve of PN Photodiode with presence of light



11.4 Conclusion

In this paper, the different width of depletion region effects of the PN photodiode on its I-V characteristics has been successfully simulated and verified. The presence of light also affects the I-V curve of the PN photodiode. In order to perform the I-V curve, the SILVACO TCAD is used. The result is compared

between the different of structure. It is proved that the smaller width of depletion region, the better the device. From this paper also can conclude that the photocurrent gets increased by increasing the photon energy or light intensity.

Acknowledgments The authors would like to thanks Universiti Teknikal Malaysia Melaka, Malaysia (UTeM) for their equipment, financial and assistance support. Last but not list, the author would like to thanks those involved direct and indirect in completing this project success fully.

References

1. Chiang, C.Y.-T., Yeow, Y.T.: Inverse modelling of two-dimensional MOSFET dopant profile via capacitance of the source/drain gated diode. *IEEE Trans. Electron Devices* **47**(7), 1385–1392 (2000)
2. Alexander, S.B.: *Optical Communication Receiver Design*. SPIE Press, Bellingham (1997)
3. Nair, B.S.: *Electronic Devices and Application*, p. 329. Prentice-Hall India, New Delhi (2006)
4. Chou, F., Wang, C., Chen, G., Sin, Y.: An 8.7 GHz Si photodiode in standard 0.18- μm CMOS technology. In: *OECC*, pp. 826–827 (2010)
5. Mohamad, M., Jubadi, W.M., Tugiman, R., Zinal, N., Zin, R.M.: Comparison on I–V performances of Silicon PIN diode towards width variations. In: *IEEE ICSE*, pp. 12–14 (2010)
6. Silvaco: *Atlas User's Manual: Device Simulation Software*. Silvaco International, Santa Clara (2000)
7. Sze, S.M.: *Physics of Semiconductor Devices*, 3rd edn. Wiley, New York (2007)

Chapter 12

U-Slot Rectangular Patch Antenna for Dual Band Application

Mohammad Shawkat Habib, I.M. Rafiqul, Khaizuran Abdullah
and M. Jamil Jakpar

Abstract Dual and multi-band rectangular microstrip antennas can be realized by cutting U-slots inside the patch. In this paper, the length and width of U-slots are optimized in order to achieve dual-band and multi-band operation. Computer Simulation Technology (CST) software was used to design, simulate and optimization of antenna. Two resonant frequencies at 1.8 and 2.4 GHz were found with reasonable gain. Additional resonant frequencies could also be achieved from 2.8 to 3.0 GHz using the similar approach.

Keywords Patch antenna · U-slot · Dual band

12.1 Introduction

Wireless local area network (WLAN) is one of the most important applications of the advancing wireless communication technology. Developed by the Institute of Electrical and Electronics Engineers (IEEE) and the 802.11 standard the wireless local-area network (WLAN) standard is a family of specifications for WLAN technology [1–3]. Most of the wireless devices are integrated with IEEE WLAN functionalities [4–6]. The emerging market of wireless devices like the laptops, tablet pc etc. has set off notable research activities on the design of cost-effective, multi band yet simple antennas. With the benefits of having low manufacturing

M.S. Habib · I.M. Rafiqul (✉) · K. Abdullah · M.J. Jakpar
Department of Electrical and Computer Engineering, International Islamic University
Malaysia, Jalan Gombak, 53100 Kuala Lumpur, Malaysia
e-mail: rafiq@iium.edu.my

M.S. Habib
e-mail: mshkanto@gmail.com

K. Abdullah
e-mail: khaizuran@iium.edu.my

cost and compatible in size, the planar antennas are good choice for the majority of the wireless LAN stations both on subscriber end and base station side. This paper presents a patch antenna with U shaped slot resonant at 2.4 GHz for WLAN application and 1.8 GHz for cognitive radio application.

Dual-band and multi-band rectangular microstrip antennas are realized by cutting U-slots, V-slots, or a pair of rectangular slots inside the patch. The technique for designing dual-band microstrip antenna is to cut slots of different shapes at an appropriate position inside the rectangular patch [7–9]. Since the slots are cut inside the microstrip antenna, they neither increase the patch size nor significantly affect the radiation pattern of the antenna. When the slots are cut very close to the radiating edge of the microstrip antenna, they alter the third-order-mode resonance frequency of the patch and, along with the fundamental mode; result in a dual-band response [10]. By integrating four slots inside the patch, a nine-band antenna, covering various cellular and TV bands, was reported in Ref. [11]. The analysis for studying the effects of a U-slot on the broadband or the dual-band response in a rectangular microstrip antenna was reported in Ref. [12].

In most of the design, depending upon where the slot is cut, the slot length is taken to be equal to either a quarter-wavelength or a half-wave length. However, these simpler approximations of slot length as a function of frequency do not give a close match for different slot lengths and their positions inside the patch. The surface currents and voltage distributions for a dual-band U-slot-cut on rectangular microstrip antennas are studied over a wide frequency range. It was observed that the slot does not introduce any mode, but reduces the higher-order orthogonal mode resonance frequency of the patch and, along with the fundamental mode, realizes the dual-band response. In this paper, formulation proposed by [13, 14] for U-slot were utilized and an antenna has been designed for 1.8 and 2.4 GHz dual-band applications with reasonable gain. The technique has been extended to design third resonance also at 2.95 GHz. Results are obtained by using Computer Simulation Technology (CST) software.

Section 12.2 describes the parameter values that were considered in the final design of the antenna. The simulated results of the final design were produced in Sect. 12.3. Section 12.4 overviews the fabrication of the antenna and its characteristics which were further contrasted with the simulated results in Sect. 12.5. The reason for the slight variation of the simulated and the fabricated result are also discussed in the later part of Sect. 12.5.

12.2 Antenna Design

The length, width, return loss, VSWR of the patch antenna can be calculated from the Eqs. (12.1)–(12.6) narrated in Ref. [15]. Where L and W are the length and width of the patch, c is the velocity of light, ϵ_r is the dielectric constant of substrate, h is the thickness of the substrate, f_o is the target center frequency, ϵ_e is the effective dielectric constant and ρ is the radiation coefficient.

$$W = \frac{c}{2f_o} \sqrt{\left(\frac{\epsilon_r + 1}{2}\right)} \quad (12.1)$$

$$L = \frac{c}{2f_o \sqrt{\epsilon_r}} - 2\Delta l \quad (12.2)$$

$$\epsilon_e = \frac{1}{2}(\epsilon_r + 1) + \frac{1}{2}(\epsilon_r - 1) \sqrt{\left(1 + \frac{10h}{W}\right)} \quad (12.3)$$

$$\Delta l = 0.412h \frac{(\epsilon_e + 0.3) \left[\frac{W}{h} + 0.8\right]}{(\epsilon_e - 0.258) \left[\frac{W}{h} + 0.8\right]} \quad (12.4)$$

$$VSWR = \frac{1 + \rho}{1 - \rho} \quad (12.5)$$

$$\text{Return Loss} = -10 \log\left(\frac{1}{\rho^2}\right) \quad (12.6)$$

While designing the antenna some of the parameters were varied and the effects were observed in order to tune the antenna resonant frequency to the desired configuration. With different values of L_v and L_h the return loss was observed. The variation of the vertical slot cut L_v was observed and the variation in return loss is displayed in Fig. 12.1. L_v was varied from 20 to 50 mm with a step of 10 mm keeping the horizontal slot cut L_h constant at 30 mm. It was observed that with the increase in value of L_v the return loss increases till $L_v = 40$ mm, and at

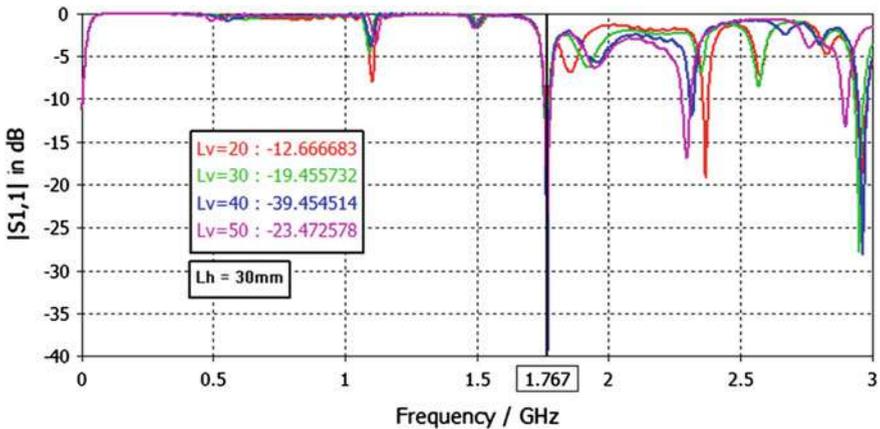
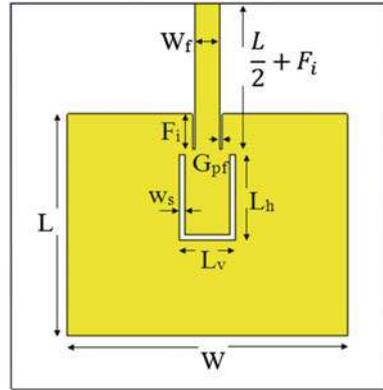


Fig. 12.1 Variation of return loss (S11) when L_h kept constant at 30 mm and L_v is varied from 20–50 mm

Fig. 12.2 Front view of U-slot rectangular microstrip antenna



$L_v = 50$ mm the value of return loss is less than -24 dB. Moreover the resonant frequency near to 2.4 GHz shifts left with the increase of L_v .

The final microstrip antenna has been designed by using a patch of length 80 mm and width of 100 mm whose top view is displayed at Fig. 12.2. The patch is having a U-slot craved within it, and the dimensions are 31 and 20 mm for horizontal and vertical slots respectively. The substrate used is of FR-4 with a dielectric constant of 4.3 and the length and width is twice the dimension of the copper patch and the thickness is of 1.6 mm.

The bottom of the substrate is ground layered with copper with a thickness of 0.1 mm and the dimension is same as the substrate's dimension and the assumed impedance of the line is 50 ohms. The patch antenna with the mentioned design is simulated with the CST Microwave Studio and required results are obtained. The detailed parameters are provided in the Table 12.1.

Table 12.1 Designed parameters of U-slot dual band microstrip antenna for 1.8 and 2.4 GHz

Parameter	Value (mm)
Length, L	80
Width, W	100
Feed slot cut, F_i	12.5
Horizontal slot cut, L_h	31
Vertical slot cut, L_v	20
Distance of the starting U-slot from the middle, X_f	25
Feed line width, W_f	8.7
Substrate thickness, h	1.6

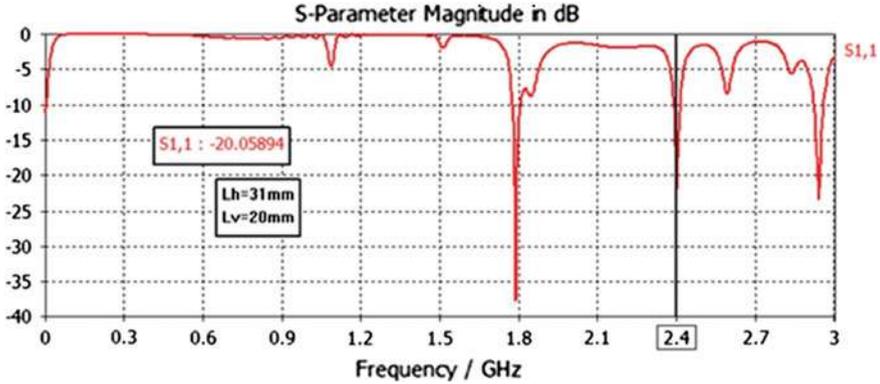


Fig. 12.3 Simulated return losses for the U-slot microstrip patch antenna

12.3 Simulation Result

Utilizing CST Microwave Studio the proposed design has been simulated and the obtained results are plotted in Fig. 12.3. The desired resonant frequencies were at 1.8 and 2.4 GHz and was obtained when the design parameters were $L_h = 31$ mm and $L_v = 20$ mm. At the first resonant frequency $f_1 = 1.8$ GHz, the return loss is equal to -35 dB and at the second resonant frequency $f_2 = 2.4$ GHz, the return loss is equal to -20 dB.

The radiation patterns of the antenna at 1.8 and 2.4 GHz are presented in Fig. 12.4a, b respectively. The antenna directivity for both frequencies of 1.8 and 2.4 GHz from the simulation output is presented in Table 12.2. Both of the frequency bands display effective results in terms of antenna gain which is above 5.0 dBi.

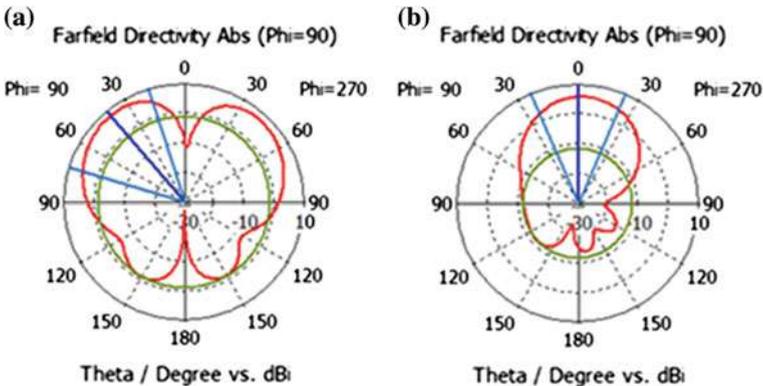
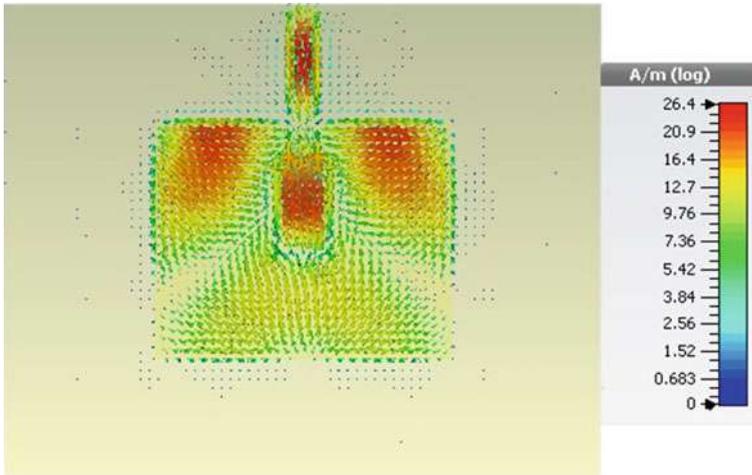


Fig. 12.4 Radiation pattern of a 1.8 GHz and b 2.4 GHz U-slot microstrip antenna

Table 12.2 Antenna characteristics

Frequency	1.8 GHz	2.4 GHz
Main lobe magnitude	8.0 dBi	5.6 dBi
Main lobe direction	40.0°	0.0°
Angular width (3 dB)	55.7°	45.9°
Side lobe level	-9.3 dB	-17.3 dB

**Fig. 12.5** Current distribution at 1.8 GHz

Current distribution of the proposed antenna is shown in Figs. 12.5, 12.6 and 12.7. Arrow sign is used to indicate current distribution. From the figures, it can be easily observed that current flow is the maximum to the microstrip line that is used as a feeding technique and to the insidious part of the U-slot along with the corners of the patch near to the feed line for the 1.8 GHz resonant frequency in Fig. 12.5. For the 2.4 GHz resonant frequency in Fig. 12.6, near the feed line connection of the patch, the current density is found to be the maximum. In 2.95 GHz resonant frequency, the current distribution is at its maximum inside the U-slot and around the two slots close to the connection of the feed line displayed in Fig. 12.7.

12.4 Fabrication and Test Result

The antenna being designed in the simulation section has been fabricated. The artwork design used in order to fabricate the antenna is shown in Fig. 12.8a. After following different steps for the fabrication, the final fabricated product of U-slot Microstrip Patch Antenna is shown in Fig. 12.8b.

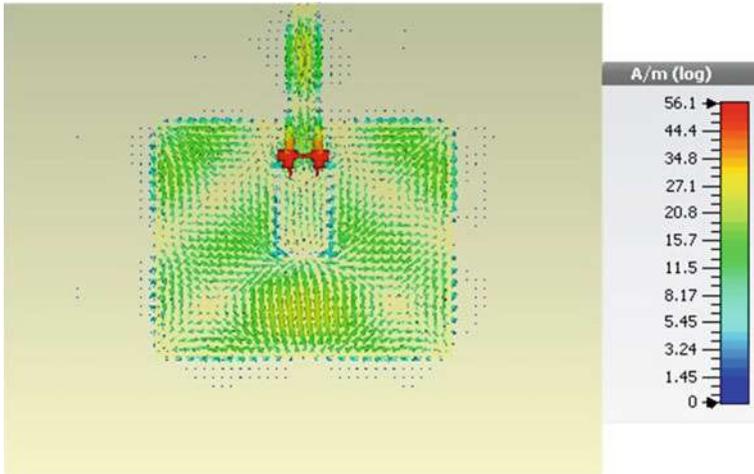


Fig. 12.6 Current distribution at 2.4 GHz

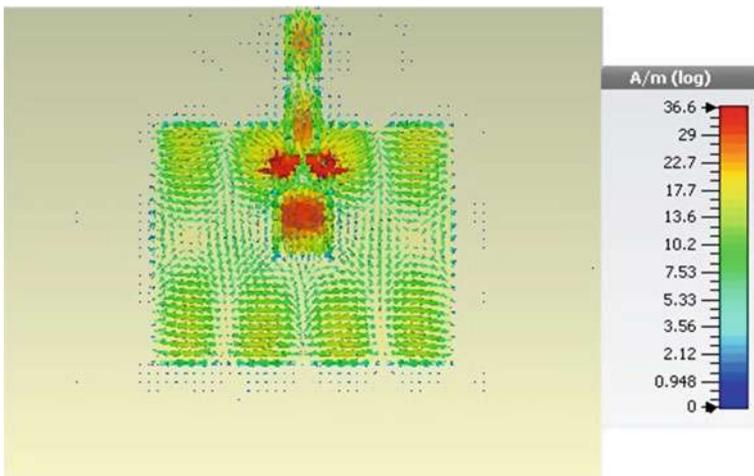


Fig. 12.7 Current distribution at 2.95 GHz

The fabricated antenna was connected with SMA-Female (Gold Type) connector in order to test the antenna using Vector Network Analyzer (VNA) at the lab. The VNA test results are presented in Fig. 12.9. Result shows that the operating frequency band of the antenna has been minutely shifted from 1.8 to 1.839 GHz with a return loss of -13.755 dB. The same shifting is observed in case of the second resonant frequency band that shifted from 2.4 to 2.48 GHz with

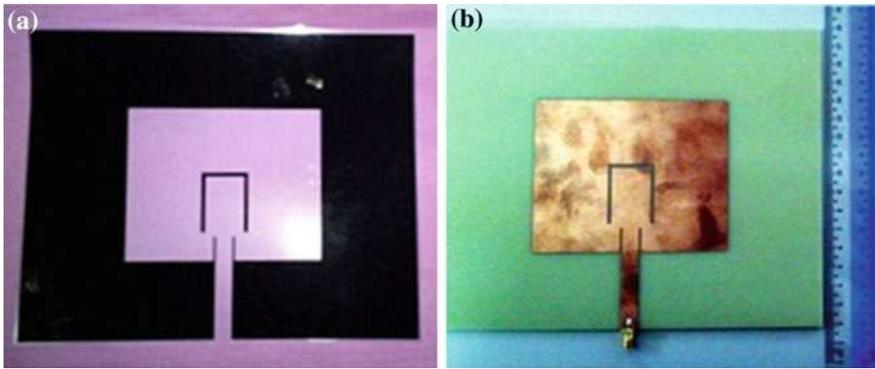


Fig. 12.8 a Prepared artwork design of the b Fabricated U-slot microstrip antenna

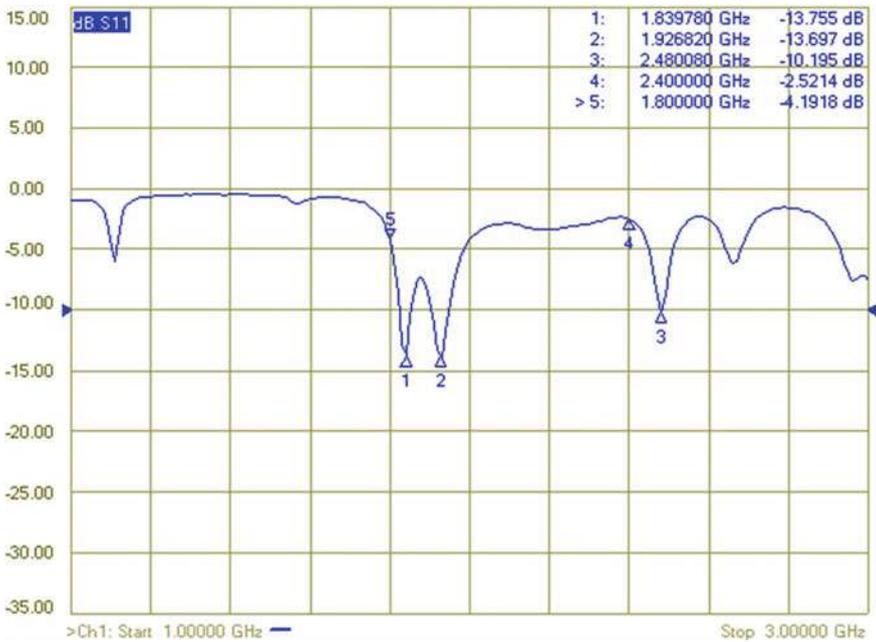


Fig. 12.9 Test result from vector network analyzer (VNA)

return loss of -10.195 dB. This deviation of the practical result with the simulated one may be due to the fabrication error and the different characteristics of substrate dielectric constant.

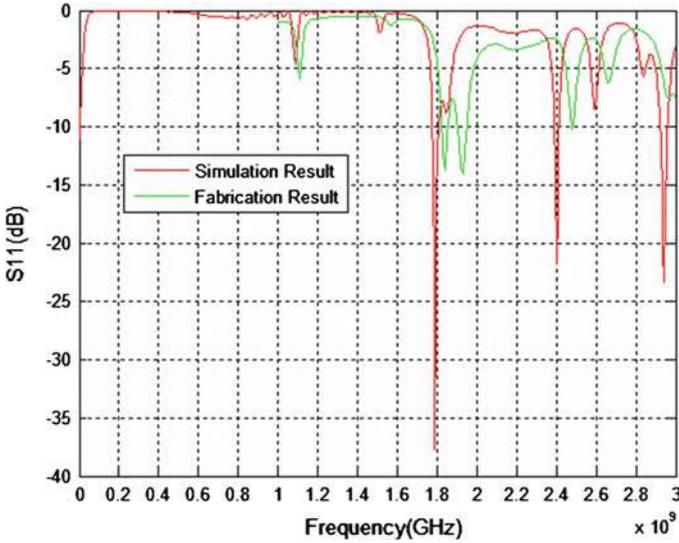


Fig. 12.10 Comparison of simulated and test result of U-slot microstrip antenna

12.5 Comparison Between Simulation and Test Result

The result of the design dimension is $L = 80$ mm, $W = 100$ mm, $L_h = 31$ mm and $L_v = 20$ mm. The Length L and width W of the antenna is quite big due to the formulation that was made in Ref. [14] which enabled the antenna to vary the resonant frequency from 900 MHz to 7 GHz. With the decrease of the values of the resonant frequencies, the wavelength of the radiation wave increases. Therefore in order to deal with such wavelength, the size of the antenna is increased. Through producing the U-slot in the microstrip patch antenna, the dual band operational at frequencies of 1.8 and 2.4 GHz was achieved.

The simulated and fabricated test results are plotted in Fig. 12.10 for comparison. From Fig. 12.10, it can be observed that the simulation result are almost identical to the one occurred in that of the fabrication test result. The only difference is that all the resonances are gradually shifted with different value of return loss. This shifting is witnessed due to several reasons.

The differences between the result of simulation and fabrication may be because of the value of substrate constant. In simulation process, the substrate constant value being used is $\epsilon_r = 4.3$ which varies slightly in available board at lab. Then, the dimension of the board and slot cut could not be extremely precise. The process of cutting the board is by using automated drilling machine in PCB lab which may also cause the differences of the results.

12.6 Conclusion

In this paper we successfully designed a microstrip patch antenna with U-slot that was resonant at 1.8 and 2.4 GHz with gains of 8.0 and 5.6 dBi respectively. It is also observed that the same structure can be used to design more resonances to achieve triple or quad band antenna for future works. Designed dual band antenna was fabricated and tested. The test result is similar to simulated result with slight shifting of resonant frequencies.

Acknowledgments Authors are grateful to Kulliyyah of Engineering and Research Management Center, International Islamic University Malaysia to support this research through grants.

References

1. Jordan, R., Abdallah, C.T.: Wireless communications and networking: an overview. *IEEE Trans. Antennas Propagat. Mag.* **44**(1), 185–193 (2002)
2. Joseph, M., Paul, B., Raj, R.K., Mohanam, P.: Compact wideband antenna for 2.4 GHz WLAN applications. *Electron. Lett.* **40**, 1460–1461 (2004)
3. Suo, Wei: Internal PIFAs for UMTS/WLAN/WiMAX multi network operation for a USB dongle. *Microw. Opt. Technol. Lett.* **48**(11), 2249–2253 (2006)
4. Karaboikis, M., Soras, C., Tsachtsiris, G., Makios, V.: Compact dualprinted inverted-F antenna diversity systems for portable wireless devices. *IEEE Antennas Wirel. Propag. Lett.* **3**, 9–14 (2004)
5. Eldek, A.A., Elsherbeni, A.Z., Smith, C.E.: Wideband bow-tie slot antennas for radar applications, 2003 IEEE Topical. In: Conference Wireless Communication Technology, Honolulu, Hawaii, (2003)
6. Jan, J.Y., Tseng, L.C.: Small planar monopole antenna with a shorted parasitic inverted-L wire for wireless communications in the 2.4-, 5.2-, and 5.8-GHz bands. *IEEE Trans. Antennas Propag.* **52**(7), 1903–1905 (2004)
7. Deshmukh, A.A., Ray, K.P.: Half U-slot loaded multi-band rectangular microstrip antennas. *Int. J. Microw. Opt. Technol.* **2**(2), 216–221 (2007)
8. Lee, K.F., Steven Yang, S.L., Kishk, A.A.: Dual and multi band U-slot patch antennas. *IEEE Antennas Wirel. Propag. Lett.* **7**, 645–647 (2008)
9. Deshmukh, A.A., Kumar, G.: Compact broadband U-slot loaded rectangular microstrip antennas. *Microw. Opt. Technol. Lett.* **46**(6), 556–559 (2005)
10. Maci, S.: Dual Band Slot Loaded Antenna. *IEEE Proc. Microw. Antennas Propag.* **142**, 225–232 (1995)
11. Boyle, K.R., Massey, P.J.: Nine band antenna system for mobile phones. *Electron. Lett.* **42**(5), 265–266 (2006)
12. Weigand, S., Huff, G.H., Pan, K.H., Bernhard, J.T.: Analysis and design of broadband single layer rectangular U-slot microstrip patch antenna. *IEEE Trans. Antennas Propag.* **AP-51**(3), 457–468 (2003)
13. Lee, K.F., Yang, S.L.S., Kishk, A.A., Luk, K.M.: The versatile U-slot patch. *IEEE Antennas Propag. Mag.* **52**(1), 71–88 (2010)
14. Deshmukh, Amit A., Ray, K.P.: Formulation of resonance frequencies for Dual-band slotted rectangular microstrip antennas. *IEEE Antennas Propag. Mag.* **54**(4), 78–97 (2012)
15. Islam, M.M., Islam, M.T., Faruque, M.R.I.: Bandwidth enhancement of a microstrip antenna for X-band applications. *ARPN J. Eng. Appl. Sci.* **8**(8), 591–594 (2013)

Chapter 13

Analysis of Synthetic Storm Technique Based on Ku-Band Satellite Beacon Measurements in Malaysia

Ali K. Lwas, I.M. Rafiqul, Mohamed Hadi Habaebi,
Ahmad F. Ismail, Mandeep Singh, Jalel Chebil,
Al-Hareth Zyoud and Hassan Dao

Abstract Most of the existing rain attenuation prediction models were proposed based on measurements taken in temperate climates. These models are found not accurate in tropical regions and were thus modified in order for such models to be applied in tropical regions. Synthetic Storm Technique (SST) is one of the most reliable methods to estimate rain attenuation time series in Europe. However, due to the lack of measured data in the tropical regions of the world, the above-mentioned method is yet to be validated for those regions. This paper aims to investigate SST validity in Malaysia by focusing on both rain events and the overall statistical behavior. Its performance is assessed based on concurrent measurement of Ku-band satellite beacon and rain rate over University of Science

A.K. Lwas · I.M. Rafiqul (✉) · M.H. Habaebi · A.F. Ismail · A.-H. Zyoud · H. Dao
Kulliyah of Engineering International Islamic University Malaysia, Jalan Gombak,
53100 Kuala Lumpur, Malaysia
e-mail: rafiq@iium.edu.my

A.K. Lwas
e-mail: alilawas@yahoo.com

M.H. Habaebi
e-mail: habaebi@iium.edu.my

A.F. Ismail
e-mail: af_ismail@iium.edu.my

A.-H. Zyoud
e-mail: alhmtz@yahoo.com

H. Dao
e-mail: sun.spu@gmail.com

M. Singh
Faculty of Engineering, University Kebangsaan, Bangi, Malaysia
e-mail: mandeep@eng.ukm.my

J. Chebil
Higher Institute of Transport and Logistics, University of Sousse, Sousse, Tunisia
e-mail: chebil8@hotmail.com

Malaysia (USM) campus at Tronoh. Preliminary analysis shows that SST is capable of providing details of time-series of many rain events to reflect the dynamics of rain fade. However, it is unable to predict the entire range of rain intensity.

13.1 Introduction

Most severe conditions of propagation impairments such as rain, clouds and atmospheric gases are occurring frequently in tropical regions. Furthermore, rain-induced attenuation is the major issue at frequencies above 10 GHz which faces the signal of satellite communications, more especially in this area which is subjected to heavy rainfall. To design a reliable earth-to-satellite link, the effects of these are required to measure and understand clearly. Scarcity of measured data at higher frequencies is also an issue in tropics. This makes it difficult to test available propagation prediction models or to develop new ones. Thus, this fact gives a real reason of why the researchers are still in their researches improve and update prediction models to give abstract view about the behavior of the rain rate and attenuation phenomena in tropical regions [1].

The synthetic storm technique is a method which can transform a rainfall rate time series directly into a rain attenuation time series. This technique developed by [2] based on rain rate time series recorded in Italia. Compared to nine prediction models, this model overcomes all of them in the three Italian sites. In addition, study [3] focused on the applicability of the SST on single rain events affecting V-band satellite links. It finds that the SST gives specifics of time series of single rain events as well as matches long-term statistics. Also, in paper [4] performance of SST is evaluated for Ka bands over a tropical location in India. The analysis indicated that SST is an appropriate to estimate the fade characteristics from the rain rate time series measurements over this region. A new prediction model is proposed in [5] that bases on advantages of SC-EXCELL (stratiform-convective rain discrimination) and of SST, named SC-SST (Stratiform-Convective SST). From analysis, it is found to be in good agreement with beacon measurements collected at Kuala Lumpur from MEASAT-1 satellite.

This paper aims to investigate applicability of SST in Malaysia by comparing measured rain attenuation with that predicted by SST from rain rate both measured concurrently. Rain attenuation and rain rate measurements were conducted at USM campus (4.390 N, 100.980 E) at Tronoh, Perak (200 km from Penang), which faces very heavy rainfall frequently. The receiver antenna is pointing towards Superbird-C at 40.1 elevation angle and diameter size of 2.4 m. The beacon signal frequency is 12.255 GHz. The data logging system has a sampling rate of one sample per second and the rain gauge is of 1-min integration time. Section two of the paper provides brief on the SST. Results are introduced and analyzed in Section three. Section four concludes the contribution of the paper.

13.2 Synthetic Storm Technique

Synthetic Storm Technique is a physical-mathematical radio propagation method that can be used to generate reliable rain attenuation time series by converting a rain rate time-series at a specified site into a rain attenuation time-series. Local parameters such as rain rate, length of signal path through rain cell and the rain cell velocity at the site under investigation are required in the SST [6]. Moreover, by applying the SST method, rain attenuation time series at any frequency and polarization can be generated, and for any slant path above approximately 10°, as long as the hypothesis of isotropy of the rainfall spatial field holds in the long term. The SST method is very useful for designing communication satellite systems and improving their performance [7] for calculating rain attenuation using the SST over a satellite path. The following are the basic assumptions, which are explained in Fig. 13.1 [2, 3, 7, 8]. During rain, the vertical structure of the troposphere separates into two layers. A is the rain layer and B is the melting layer as shown in Fig. 13.1. The signal attenuation in case of satellite path is obtained from specific attenuation at a point using the following expression:

$$A(x) = K_A \int_0^{L_A} R^{\alpha_A}(x_0 + \Delta x_0, \xi) d\xi + K_B r^{\alpha_B} \int_{L_A}^{L_B} R^{\alpha_B}(x_0, \xi) d\xi \tag{13.1}$$

where $A(x)$ is the attenuation at a specific point and ξ is the distance measured along the satellite path. K and α depend on the electromagnetic wave, frequency and polarization, and raindrop size distribution. They are given by [9] for water temperature of 20 °C, and Parson’s law drop size distribution and for 0 °C, are given by [10]. More details about the SST are explained in [6].

According to [2], the following equation was derived by applying the Fourier transform theory and some assumptions on Eq. (13.1) of the attenuation time series

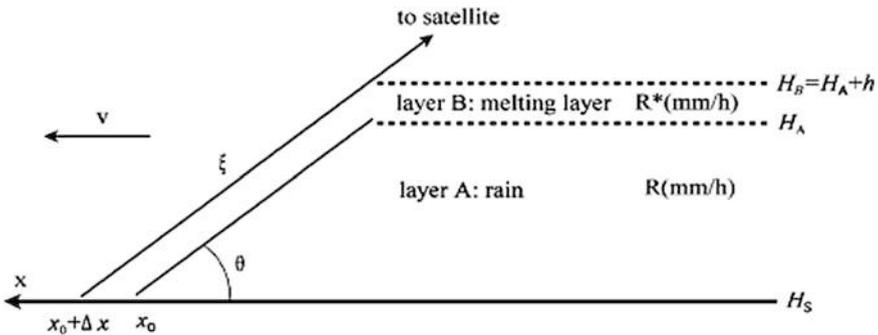


Fig. 13.1 Schematic diagram of rain structure for Synthetic Storm Technique [2]

$$A(t) = K_A R^{\alpha_A}(t) L_A + r^{\alpha_B} K_B R^{\alpha_B}(t) (L_B - L_A) \quad (13.2)$$

where $R(t)$ is rain rate time series. L_A and L_B are the radio path lengths and are given by

$$L_A = (H_A - H_S) / \sin(\theta) \quad (13.3)$$

$$L_B = (H_B - H_S) / \sin(\theta) \quad (13.4)$$

where H_S is the height above sea level of the Earth station, H_A is the height above sea level of the upper limit of layer A and H_B is the height above sea level of the upper limit of layer B. According to ITU-R Rec. 839-0 (1992) [11], H_B is given by

$$H_B = \begin{cases} 5, & \phi < 23 \\ 5 - 0.075(\phi - 23) \text{ km}, & \phi \geq 23 \end{cases} \quad (13.5)$$

The above equation indicates that H_B depends on the latitude (ϕ) of the Earth station. The author in [2] supposes the H_A is given by

$$H_A(\phi) = H_B(\phi) - h(\text{km}) \quad (13.6)$$

where $h = 0.4$ km is the thickness of the melting layer.

13.3 Results and Discussion

Conversion from rainfall rate to rain attenuation was performed by the Synthetic Storm Technique method using Eq. (13.2). The radio path lengths L_A and L_B are calculated according to Eqs. (13.3) and (13.4) respectively. The specific attenuation of melting region (B) is constructed according to [12] and the values for coefficients k and α are taken to be 0.2521 and 1.1635 respectively. The k and α values are related to rainy layer (A) and are selected from [10] to be 0.02514 and 1.24 respectively. As originally proposed in [2] the rain advection velocity is $v = 10$ m/s.

In Figs. 13.2 and 13.3, measured beacon rain attenuation and that predicted by SST for 12.255 GHz are plotted with respect to time using measured rain rate time series. It is observed that there is a good correlation between the measured rain rate and the measured rain attenuation. Predicted rain attenuation time series using the SST method followed the measured rain attenuation time series in both events. Many peaks are observed in Fig. 13.2. In the first peak, the value of rain attenuation is close to 20 dB by the SST method but the measurement is about 15 dB. Also, the value of rain attenuation peak observed in the last peak is about 20 dB by the SST method but the measurement is close to 14 dB. Figure 13.3 shows that the

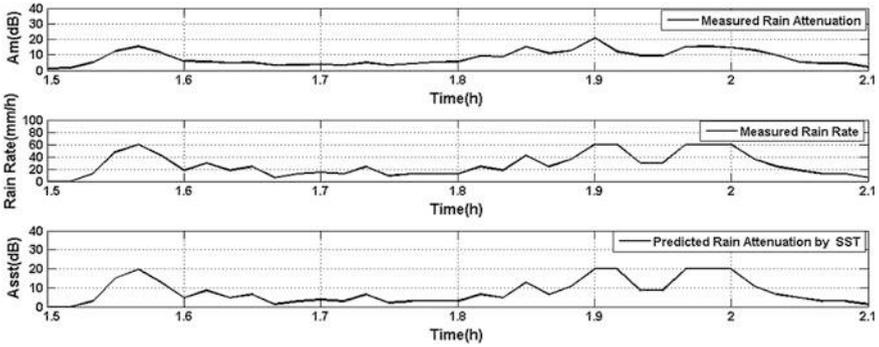


Fig. 13.2 Comparison between measured rain attenuation and that converted by SST for a rainy event on 10/8/2009

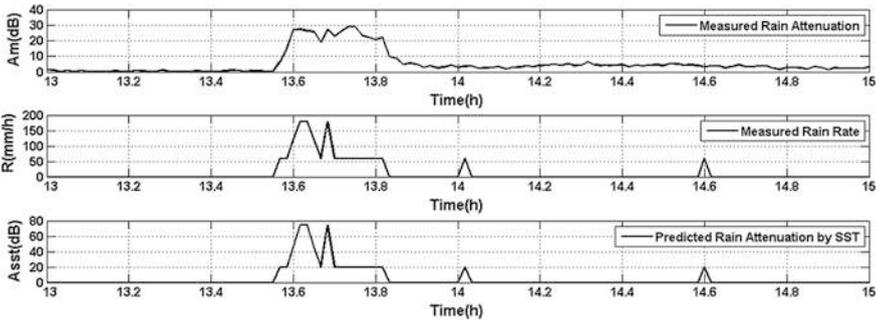


Fig. 13.3 Comparison between measured rain attenuation and that converted by SST for a rainy event on 26/8/2009

highest value of measured attenuation reaches to 30 dB while the SST prediction is close to 74 dB. One month statistics of measured rain attenuation for August 2009 and predicted rain attenuation are presented in Tables 13.1 and 13.2. The difference between measured and predicted rain attenuation for rain rate up to 60 mm/hr is shown in Fig. 13.4. 1 min rain rate at 60, 120 and 180 mm/h are considered for comparing the peak value of measured rain attenuation respectively. Three values of 60, 120 and 180 mm/h were measured for rain rate. Therefore, the resolution for this data was very low.

Figure 13.4 explains the differences between measured peak attenuation and attenuation predicted by SST for all events where rain rate are 60 mm/h. The percentage of error is calculated using Eq. (13.7). More than 100 peaks with 60 mm/hr rain rate are recorded during August 2009. Average of percentage error is 19 %, in other words, in most cases SST prediction is higher than measured attenuation.

Since the resolution of measured rain rate is only 1.0 mm/min or 60 mm/h, the peak value of 120 mm/h can be assumed average peak of 90 mm/h, while the peak value of 180 mm/h can be assumed average peak of 150 mm/h. According to

Table 13.1 Percentage of error between measured attenuation and that predicted by the SST method for Medium Rain Rate ($R \leq 120$ mm/h)

Date	A_{measur}	$A_{\text{SST}} (\text{max})$	Error (%)	$A_{\text{SST}} (R = 90)$	Error (%)
1/8/2009 (1)	23.5	45.56	94	32.25	37
1/8/2009 (2)	18.12	45.56	151	32.25	78
5/8/2009	20.11	45.56	126	32.25	60
9/8/2009	27.75	45.56	64	32.25	16
	25.32	45.56	80	32.25	27
	25.51	45.56	79	32.25	26
	26.44	45.56	72	32.25	22
	27.46	45.56	66	32.25	17
13/8/2009	28.63	45.56	59	32.25	12.6
	25.76	45.56	77	32.25	25.2
	27.9	45.56	63	32.25	15.6
17/8/2009	25.5	45.56	79	32.25	26.5
	28.19	45.56	62	32.25	14.4
	27.56	45.56	65	32.25	17.1
26/8/2009	26.95	45.56	79	32.25	19.7
27/8/2009	25.69	45.56	77	32.25	25.5

Table 13.2 Percentage of error between measured attenuation and that predicted by the SST method for High Rain Rate ($R \leq 180$ mm/h)

Date	A_{measured}	$A_{\text{SST}} (\text{max})$	Error (%)	$A_{\text{SST}} (R = 150)$	Error (%)
4/8/2009	28.23	74.2	163	59.59	111
26/8/2009	27.39	74.2	171	59.59	118
26/8/2009	27.145	74.2	173	59.59	120

Table 13.1, where the maximum rain rate is 120 mm/h and the assumed average peak is 90 mm/h, the differences between measured rain attenuation and the SST predictions are very high for rain rate when it equals 120 mm/h. It varies between 17 and 27 dB, which is 60–150 %. For the assumed average rain rate of 90 mm/h, the differences are 4–14 dB, which varies from 13 to 78 %. Furthermore, Table 13.2 shows that the differences between the measured attenuation and the SST predictions are very divergent for 180 mm/h (46.6 dB), while it is 32 dB for an average of 150 mm/h. In both cases, the errors are more than 100 %. The percentage of errors (E) between the measured attenuation (A_m) and the predicted attenuation (A_p) are calculated based on the following equations [13]:

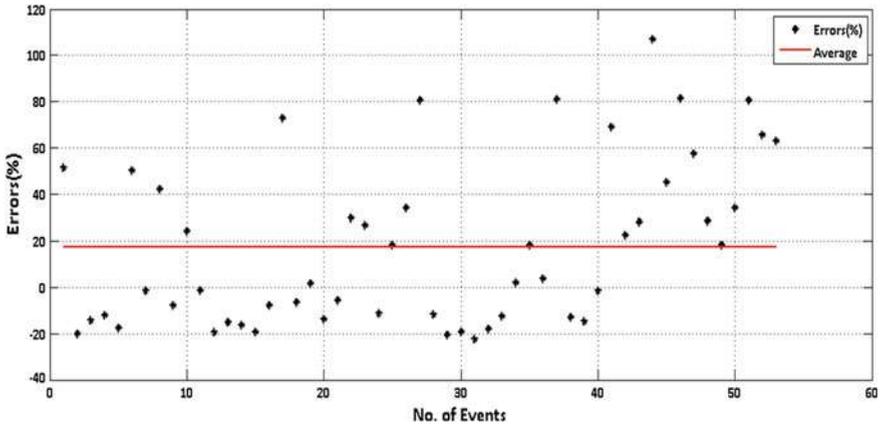


Fig. 13.4 Percentage of error between measured attenuation and that predicted by the SST method with respect to number of events

$$E = \frac{A_p - A_m}{A_m} \times \% \quad (13.7)$$

13.4 Conclusion

This paper focuses on the applicability of the SST in Malaysia, which has been established as an accurate method in Europe. The SST method is used to generate rain attenuation time series by using rain rate data measured for one month in Malaysia. The attenuation predicted by using the SST method is compared with measured rain attenuation at 12.255 GHz satellite beacon recorded at Satellite Lab, USM. It is found that the attenuation derived by SST method is close to the measured attenuation in low rain intensity but it is overestimated in medium rainfall and extremely overestimated in high rainfall rate. Despite these discrepancies, the SST method can be used for generating time-series of rain attenuation on satellite to earth links in tropical regions with proper modifications. However, this method is required to be modified based on long term measurements to be applicable for tropical regions.

Acknowledgments This work is supported by the Research Management Centre (RMC), project no. EDW B13-038-0923, International Islamic University Malaysia (IIUM).

References

1. Emiliani, D., Agudelo, J., Gutierrez, E., Restrepo, J., Mendez, F.: Development of rain attenuation and rain rate maps for satellite system design in the Ku and Ka bands in Colombia. *IEEE Antennas Propag. Mag.* **46**, 54–68 (2004)
2. Matricciani, E.: Physical-mathematical model of the dynamics of rain attenuation based on rain rate time series and a two-layer vertical structure of precipitation. *Radio Sci.* **31**, 281–295 (1996)
3. Sánchez-Lago, I., Fontán, F.P., Mariño, P., Fiebig, U.C.: Validation of the synthetic storm technique as part of a time series generator for satellite links. *IEEE Antennas Wirel. Propag. Lett.* **6**, 372–375 (2007)
4. Shukla, K.A., Das, S., Roy, B.: Rain attenuation measurements using synthetic storm technique over ahmedabad. In: *International Conference on Computers and Devices for Communication India* (2007)
5. Lam, Y.H., Luini, L., Din, J., Capsoni, C., Panagopoulos, A.D.: Investigation of rain attenuation in equatorial kuala lumpur. *IEEE Antennas Wirel. Propag. Lett.* **11**, 1002–1005 (2012)
6. Lwas, A., Islam, Md., Chebil, J., Habaebi, M., Ismail, A., Zyoud, A., Dao, H.: Rain attenuation analysis using synthetic storm technique in malaysia. In: *International Conference on Mechatronics (ICOM 2013) Malaysia* (2013)
7. Matricciani, E., Riva, C., Castanet, L.: Performance of the synthetic storm technique in a low elevation 5 slant path at 44.5 GHz in the French, Pyrénées, *European Conference Antennas Propagation (EuCAP 2006) Nice: France* (2006)
8. Panagopoulos, A.D., Arapoglou, P.D., MChatzarakis, G.E., Kanellopoulos, J.D., Cottis, P.G.: A new formula for the prediction of the site diversity improvement factor. *Int. J. Infrared Millimeter Waves* **25**, 1781–1789 (2004)
9. ITU-R P.838-3: Specific attenuation model for rain for use in prediction methods, *ITU-R Recommendations Geneva* (2005)
10. Maggiori, D.: Computed transmission through rain in the 1–400 GHz frequency range for spherical and elliptical drops and any polarization. *Alta Frequenza* **50**, 262–273 (1981)
11. ITU-R P.839-0: Rain height model for prediction methods. *Recommendation P. Series Fascicle Radio wave propagation International Telecommunication Union Geneva* (1992)
12. ITU-R P.839-3: Rain height model for prediction methods: *Recommendation P. Series Fascicle Radio wave propagation International Telecommunication Union Geneva* (2001)
13. Mandeep, J.S.: Prediction of signal attenuation due to rain models at TRONOH, Malaysia. *Mapan J. Metrol. Soc. India* **28**, 105–111 (2013)

Chapter 14

The Evolution of Double Weight Codes Family in Spectral Amplitude Coding OCDMA

N. Din Keraf, S.A. Aljunid, A.R. Arief and P. Ehkan

Abstract This paper presents the review of Double Weight (DW) codes family from perspective of codes evolution since beginning until now. First generation focuses on one-dimensional (1D) code which spreading has been carried out in time. Later, the second generation was introduced known as two-dimensional (2D) code where the encoding of data bit has spread in both wavelength and time domain. Codes construction and variety of detection techniques scheme used in DW codes family are also discussed. In this paper we focus on structure of among DW codes family as all the design of code sequence aim to eliminate Multiple Access Interference (MAI). Previous papers of DW codes family show a good quality of transmission and satisfactory the standard quality of Bit Error Rate (BER) above 10^{-9} .

Keywords Double weight · Optical code division multiple access · Bit error rate

N. Din Keraf · S.A. Aljunid · A.R. Arief · P. Ehkan (✉)
School of Computer and Communication Engineering, Universiti Malaysia Perlis,
Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: phaklen@unimap.edu.my; plen07@yahoo.co.uk

N. Din Keraf
e-mail: aiu_dnkrf@yahoo.com

S.A. Aljunid
e-mail: syedalwee@unimap.edu.my

A.R. Arief
e-mail: amirrazif@unimap.edu.my

N. Din Keraf
Department of Electrical Engineering, Politeknik Sultan Abdul Halim Mu'adzam Shah,
06000 Jitra, Kedah, Malaysia

14.1 Introduction

About last three decades, optical code division multiple access (OCDMA) has been rapidly growth [1] as a result of exhaustive research and to satisfy the hunger user with new services and applications especially for multimedia services. In OCDMA, an optical code represents a user address and signs or each user has its own codeword. Various types of OCDMA codes family such as optical orthogonal code (OOC), modified frequency hopping (MFH) and prime codes have been introduced for spectral amplitude coding [2]. Besides, DW codes family are also including in this type of coding. Several codes for DW have been proposed comprising of DW, Modified Double Weight (MDW) [3, 4] and Enhanced Double Weight (EDW) [5].

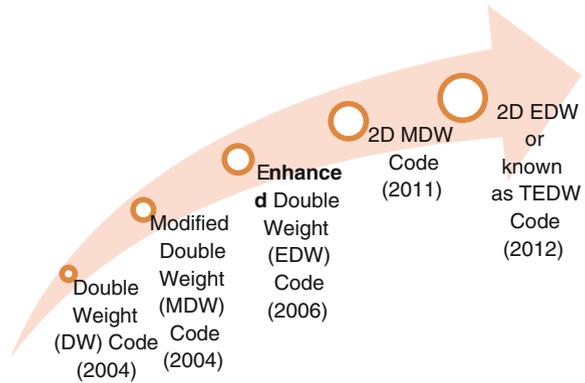
The principle source of noise and the major limit factor to system performance in OCDMA is called multiple-access interference (MAI) [6, 7]. As the number of simultaneous users increase, the effect of MAI is also increase in incoherent OCDMA. Spectral Amplitude Coding (SAC) is one of the methods used in incoherent OCDMA systems to eliminate MAI [2] by using code sequences. Most of the codes including DW codes family aim to overcome MAI as its effect a poor BER performance for the overall system. This paper is organized as the evolution of the codes is detailed in Sect. 14.2, structures of the codes are described in Sect. 14.3 and finally, conclusion of this paper in Sect. 14.4.

14.2 The Evolution of the Codes

DW codes family was first introduced by Aljunid [3] in year 2004. Initially, DW codes family are proposed for SAC OCDMA system which has a constant weight of two. Later, the development of this code family has become aggressively expanded as illustrated in Fig. 14.1. The first three codes were categorized as one-dimensional (1D) whereas the last two codes we will refer as two dimensional (2D) codes. For 1D code, the spreading has been carried out in time while for 2D encoding performs the frequency spreading in time and wavelength domain simultaneously.

The constraint of constant weight of two in DW code inspires the development of MDW code that has a variable weight greater than two. MDW codes are able to support simultaneous transmissions at different bit rates [4]. Another variation of DW code has been proposed by Hasoon et al. [5] known as EDW code which has a variable weight odd number greater than one. The extended of 1D DW codes family was proposed in 2011 [8] namely 2D MDW with the conscious that all those 1D codes suffer from a numerous limitations. Moreover, realizing the use of bandwidth must be optimized, two dimensional (2D) code has proposed to increase the number of simultaneous users compared to 1D code. Recently, 2D TEDW was proposed [9] and give more advantages such as lower bandwidth consumption compared to the conventional 1D EDW.

Fig. 14.1 The evolution of double weight codes



Construction techniques used to construct these codes family are matrix construction technique and mapping scheme [3, 10]. The details of code structure in DW code family will be described in Sect. 14.3. Other than that, selection of detection techniques has been introduced recent years such as complimentary subtraction, AND subtraction, NAND subtraction and Modified AND subtraction [11–13]. For conventional complementary subtraction technique, two different code sequences are modulated with data and sent to multiplexer. The received signal is divided into two branches of spectral chips; upper branch and lower branch [12]. These two branches of spectral signals are sent to a subtractor that computes the correlation difference [11]. On the other hand, for AND subtraction the received signal splits into two parts; one to the decoder that has an identical filter structure with the encoder and the other to the decoder that has the AND filter structures [11]. A subtractor is then used to subtract the overlapping data from the intended code. Furthermore, the modified-AND subtraction technique provides a better performance on higher data rate or a larger number of users [12]. The major advantage of modified-AND subtraction is to suppress MAI and the impacts of PIIN by dividing the spectrum of the used code sequence.

Moreover, DW codes family is also used in many applications [12] such as for instance using MDW code to support triple play services (voice, video, and data) by utilizing the different detection scheme in SAC-OCDMA. Furthermore, [14] proposed that MDW code is applicable to use in Wide Area Network (WAN). DW code was also implemented in local area network (LAN) environment which is applied in ring network as proposed by [15]. In addition, EDW codes are suitable to support multiple bit rate transmissions as suggested by [16].

There are three main properties that take into account for design a code [17] i.e. correlation properties, maximum supported user number and BER. In order to maintain as many as possible simultaneous users within an acceptable BER i.e. $BER = 10^{-9}$, correlation properties must be good enough because they are connected and restricted by each other. Thus for designing the DW is code, most authors that previously mentioned make a great effort to fulfill the requirement as much as possible. For instance the value of cross correlation are highly considered

in DW family code as the large value of cross correlation affects the system performance due to MAI. On the other hand, to maximize the number of user, 2D code was proposed. It is not only increases the flexibility of the code design but also improves the cardinality of codes dramatically [17].

14.3 Constructions of the Codes

In one dimensional DW code family, three important parameters are denoted as (N, W, λ) where N represents code length, W is code weight and λ for correlations. Three types of code consists in this category, they are DW, MDW and EDW. However in two dimensional code there are five parameters involved consists of number of wavelength M , temporal code length N , w is weight, λ_a and λ_c are auto and cross-correlation values are denoted by $(M \times N, W, \lambda_a, \lambda_c)$. In designing 2D code, cross correlation λ_c and auto correlation, λ_a are significantly concern as the cross correlation λ_c value implies the interference between users. On the contrary, the auto correlation value implies the signal power and the distinction in the presence of other users. As above mentioned, the major degrading factors which affect the overall performance of OCDMA networks are the multiple access interference (MAI) constraints. Accordingly, codes with minimum cross correlation value are preferred for OCDMA communications to reduce the effect of MAI [7].

14.3.1 DW Code

Reference [3] described the DW of (N, W, λ) family can be constructed by using the following steps:

Step 1:

The DW code can be represented by using a $K \times N$ matrix. In DW codes structures, the matrix K rows and N columns will represent the number of user and the minimum code length respectively. A basic DW code is given by a 2×3 matrix, as

$$H_1 = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{vmatrix}$$

Notice that H_1 has a chips combination sequence of 1, 2, 1 for the three columns (i.e. $0 + 1, 1 + 1, 1 + 0$).

Step 2:

A simple mapping technique is used to increase the number of codes as

$$H_2 = \begin{vmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{vmatrix} = \begin{vmatrix} 0 & H_1 \\ H_1 & 0 \end{vmatrix}$$

Note that as the number of user, K increases, the code length, N also increases. The relationship between the two parameters, K and N is given by

$$N = \frac{3K}{2} + \frac{1}{2} \left[\sin \frac{K\pi}{2} \right]^2 \quad (14.1)$$

A position of weight is very important to maintain in pairs as it reduces the number of filters use in the encoder and decoder. This way, a filter with the bandwidth twice of the chip width can be used, instead of two different filters.

14.3.2 MDW Code

MDW is the modified version of DW code. The MDW code weight can be any even number that is greater than two. As a family of DW code, MDW can also be represented by using the $K \times N$ matrix. Reference [3] described that the basic MDW can be developed by using the following steps:

Step 1:

The basic matrix for MDW codes also consists of a $K \times N$ matrix depending on the value of code weight. The general form of matrix for a MDW code is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix}$$

where;

- i. A consists of a $1 \times 3 \sum_{j=1}^{\frac{w}{2}-1} j$ matrix of zeros
- ii. B consists of a $1 \times 3n$ matrix containing the basic matrix of $[X_2]$ for every 3 columns (i.e. a $1 \times 3n$ matrix which is n repetition of $[X_2]$)
- iii. C is the basic code matrix for the next smaller weight, $W = 2(n - 1)$.
- iv. D is a matrix $n \times n$ consisting of basic matrix of $[X_3]$ arranged as

$$\begin{vmatrix} 000 & 000 & [X_3] \\ 000 & [X_3] & 000 \\ [X_3] & 000 & 000 \end{vmatrix}$$

And $n = W/2$, $W = 2, 4, 6, \dots R$ where X_1, X_2 and X_3 are the $[1 \times 3]$ matrix and consists of

$$\begin{aligned} X_2 &= [0 \quad 1 \quad 1] \\ X_3 &= [1 \quad 1 \quad 0] \\ X_1 &= [0 \quad 0 \quad 0] \end{aligned}$$

Step 2:

There are two basic components in basic matrix for MDW codes:

Code length,

$$N_B = 3 \sum_{j=1}^{\frac{W}{2}} j \quad (14.2)$$

Number of user,

$$K_B = \frac{W}{2} + 1 \quad (14.3)$$

The Eqs. (14.2) and (14.3) represent the basic matrix for MDW code, where N_B is the column (i.e. its represent basic code length) and K_B is the row (its represents basic number of user). The MDW matrix is consisting of $(K_B \times N_B)$. The basic MDW code with code length 9, weight 4 and an ideal in-phase cross correlation denoted by (9, 4, 1) is

$$\begin{array}{c} \mathbf{A} \quad \quad \quad \mathbf{B} \\ \left| \begin{array}{ccc|cccc} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \end{array} \right| \\ \mathbf{C} \quad \quad \quad \mathbf{D} \end{array}$$

Notice that similar structure of the basic DW code, H_1 is still maintained with a slight modification, whereby the DW pairs are maintained in a way to allow only two overlapping chips in every column. Thus, the 1, 2, 1 chips combination is maintained for every three columns as in the basic DW code. This is important to maintain $\lambda = 1$.

14.3.3 EDW Code

Reference [5] described EDW code as the enhancement version of DW code family. Its code weight can be any odd number that greater than one. The construction of EDW code with the weight of three can be described as

Step 1:

EDW codes also consists of a $K \times N$ matrix which K rows and N columns will represent the number of user and the minimum code length respectively. A basic EDW code is given by a 3×6 matrix, as

$$H_0 = \left| \begin{array}{cc|cc|cc} 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{array} \right|$$

Notice that similar structure of the basic DW code, H_1 is still maintained with a slight modification, whereby the double weight pairs are maintained in a way to allow only two overlapping chips in every column.

Step 2:

From the basic matrix, a larger number of K can be achieved by using a mapping technique as

$$H_1 = \left| \begin{array}{cccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right| = \left| \begin{array}{c|c} 0 & H_0 \\ H_0 & 0 \end{array} \right|$$

An EDW code with weight of 3 denoted by $(N, 3, 1)$ for any given code length N , can be related to the number of user K through

$$N = 2K + \frac{4}{3} \left[\sin \frac{K\pi}{3} \right]^2 + \frac{8}{3} \left[\sin \frac{(K+1)\pi}{3} \right]^2 + \frac{4}{3} \left[\sin \frac{(K+2)\pi}{3} \right]^2 \quad (14.4)$$

14.3.4 2D MDW Code

Development of 2-D MDW code is actually from the 1-D MDW code and denoted by $(M \times N, W, \lambda a, \lambda c)$ [8]. As mentioned earlier, 2D encoding performs the frequency spreading in time and wavelength domain simultaneously. Thus for 2D MDW, matrix of M row vectors $d_{k,N}^j$ is related to the temporal spreading; $d_{1,N}^j = [c_{k,1}^j, c_{k,2}^j, \dots, c_{k,N-1}^j, c_{k,N}^j]$ while k is the emitted wavelength $k \in \{1, \dots, M\}$ as described in Eq. (14.5).

$$C_{M,N}^j = \begin{bmatrix} d_{1,N}^j \\ d_{2,N}^j \\ \vdots \\ d_{M-1,N}^j \\ d_{M,N}^j \end{bmatrix} \quad (14.5)$$

The signals $r_{k,N}(t)$ represent the sum of the temporal spreading data of each user carried on the wavelength, λ_k . They are expressed as:

$$r_{K,N}(t) = \sum_{j=1}^{Fu} b_i^j(t) d_{k,n}^j \quad (14.6)$$

The M signals $r_{k,N}^j(t)$ are multiplexed and the total signal $R_{M,N}(t)$ is transmitted on the optical fiber. It can be represented by a matrix ($M \times N$):

$$R_{M,N}(t) = \begin{bmatrix} r_{1,N}(t) \\ r_{2,N}(t) \\ \vdots \\ r_{M-1,N}(t) \\ r_{M,N}(t) \end{bmatrix} \quad (14.7)$$

The signal $r_{k,N}^j(t)$ is expressed as $\sum_{j=1}^{Fu} b_i^j(t) d_{k,n}^j$ represent sum of the temporal spreading data user carried the wavelength, t . 2-D MDW network consists of M, N pairs of transmitters and receivers. $A_{g,h}$ is the code where $g \in (1, 2, 3, \dots, M-1)$ and $h \in (1, 2, 3, \dots, N-1)$. X_g is the spectral encoding and Y_k is the spatial encoding. Table 14.1 shows some examples of 2-D MDW code sequences.

The cross correlation of 2-D MDW code can be obtained by introducing the four characteristic matrices $A^{(d)}$, where $d \in (0, 1, \dots, 3)$ are defined as

$$A^{(0)} = Y^T X \quad (14.8)$$

$$A^{(1)} = Y^T \bar{X} \quad (14.9)$$

$$A^{(2)} = \bar{Y}^T X \quad (14.10)$$

$$A^{(3)} = \bar{Y}^T \bar{X} \quad (14.11)$$

Parameter \bar{X} and \bar{Y} are the complementary of X and Y respectively. The cross correlation of 2D MDW code $A^{(d)}$ and $A_{g,h}$ is expressed as

Table 14.1 2-D MDW code for $k_1 = 4$ and $k_2 = 2$ sequences

$X_{g,h}$	[000011011]	[011000110]	[110110000] Y_k
0	000000000	000000000	000000000
1	000011011	011000110	110110000
1	000011011	011000110	110110000
1	000011011	011000110	110110000
1	000011011	011000110	110110000
0	000000000	000000000	000000000

Table 14.2 Cross correlation of 2D MDW Code

$X_{g,h}$	$R^{(0)}(g, h)$	$R^{(1)}(g, h)$	$R^{(2)}(g, h)$	$R^{(3)}(g, h)$
$g = 0, h = 0$	$k_1 k_2$	0	0	0
$g = 0, h \neq 0$	k_1	k_1	0	0
$g \neq 0, h = 0$	k_2	0	$k_2 (k_1 - 1)$	0
$g \neq 0, h \neq 0$	1	1	$k_1 - 1$	$k_1 - 1$

$$R^{(d)}(g, h) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} a_{ij}^{(d)} a_{(i+g)(j+h)}^{(d)} \tag{14.12}$$

where $a_{ij}^{(d)}$ is the (i, j) th of $A^{(d)}$ and $a_{(i+g)(j+h)}$ is the (i, j) th of $A_{g,h}$; $g \in (0, 1, 2, \dots, M - 1)$ and $h \in (0, 1, 2, \dots, N - 1)$. Table 14.2 illustrates the cross correlation of 2-D MDW code generated from Eq. (14.12).

The derivation of new correlation functions can be expressed as

$$R^{(0)}(g, h) - R^{(1)}(g, h) - \frac{R^{(2)}(g, h)}{k_1 - 1} + \frac{R^{(3)}(g, h)}{k_1 - 1} = \begin{cases} k_1, k_2, & \text{for } g = 0 \text{ and } h = 0 \\ 0, & \text{otherwise} \end{cases} \tag{14.13}$$

14.3.5 TEDW Code

The generating algorithm steps and equations for TEDW code can be described as follows [9]:

(i) *Base Matrix Generating*

A 0's matrix is generated Bm and the matrix dimensions determined by

$$\text{Rows} = W \tag{14.14}$$

	A		B		D		
Users	C1	C2	C3	C4	C5	C6	
1	1	0	1	1	0	0	3
2	1	1	0	0	1	0	3
3	0	0	1	0	1	1	3
	2	1	2	1	2	1	

Fig. 14.2 EDW code matrix W = 3 [9]

$$\text{Columns} = N_a = \frac{\sum_{j=1}^w j}{w} \tag{14.15}$$

for all rows in the base matrix Bm but for the final row a round shift operation is performed by using the Eq. (14.16).

$$\sum_{j=1}^{N_a} (Bm[j,j] = Bm[j,j] + 1) \tag{14.16}$$

For Final row we use,

$$Bmn \text{ Final} = \sum_{j=1}^{N_a} (Bm[n,j] = Bm[j,j]) \tag{14.17}$$

By applying the Eqs. (14.14)–(14.17) we have Part A (Base Matrix) as shown in Fig. 14.2.

(ii) *Shift Level*

The number of shifts is determined as

$$\text{Number of Shifts} = \frac{W + 1}{2} \tag{14.18}$$

(iii) *Full Row shift and Full Base Matrix Shift*

The sequence order is determined by S = W and for the purpose of generalization, assuming X = Bm and X = [] (Row, Columns) = (W, S). To determine part B for the code from the previews matrix assume that y = X and use the simplified Eq. (14.19).

$$\left\{ \sum_{i=1}^s [X(1, i) = y(w, i)] \sum_{j=1}^{s-1} [X(j + 1, i) = y(j, i)] \right\} \tag{14.19}$$

Next shift X_{next} = the matrix from the previews part of the code and $y = X_{\text{next}}$ shifted. The full generation of the code is done by applying Eqs. (14.14) till (14.19) and following the sequence steps.

14.4 Conclusion

The DW codes family proved to be one of the successful codes in the spectral amplitude coding OCDMA systems to reduce MAI due to the ability of these codes family to perform the standard acceptable BER (i.e. $\text{BER} = 10^{-9}$). In addition, these codes family are also designed to obtain the codes with ideal cross correlation due to the fact that multiuser interference significantly affects the overall system performance. The 1D DW family codes has properties summarized as each code sequence has cross correlation, λ equal to 1, the weight chips are always in pairs and there must be relation between number of user and code length. On the other hand, to increase the cardinality in OCDMA system, various domain and parameters are combined to create multidimensional codes. However, multi-dimensional brings a lot of complexity to the system implementation and architecture and so the DW codes family only focuses on 2D codes. An interesting future research for DW family codes would be the 3D code which promises the better performance and support more active users simultaneously.

References

1. Prucnal, P.R.: Optical code division multiple access fundamental and applications. CRC Press, Florida (2006)
2. Yin, H., Richardson, D.J.: Optical code division multiple access communication networks. Theory and applications. Tsinghua University Press, Springer-Verlag GmbH, Beijing, Berlin (2007)
3. Aljunid, S.A., Ismail, M., Ramli, A.R., Ali, B.M., Abdullah, M.K.: A new family of optical code sequences for spectral-amplitude-coding optical CDMA systems. *IEEE Photonics Technol. Lett.* **16**(10), 2383–2385 (2004)
4. Aljunid, S.A., Samad, M.D.A., Othman, M., Hisham, M.H., Kasiman, A.H., Abdullah, M.K.: Development of modified double-weight code and its implementation. In: *IEEE International Conference*, pp. 288–192 (2005)
5. Hasoon, F.N., Aljunid, S.A., Abdullah, M.K., Shaari, S.: Spectral amplitude coding OCDMA systems using enhanced double weight code. *J. Eng. Sci. Technol.* **1**(2), 192–202 (2006)
6. Fouli, K., Maier, M.: OCDMA and optical coding: principles, applications, and challenges. *IEEE Commun. Mag.* **45**, 27–34 (2007)
7. Ghafouri-Shiraz, H., Massoud Karbassian, M.: *Optical CDMA networks: principles, analysis and applications*. Wiley-IEEE Press, New York (2012)
8. Arief, A.R., Aljunid, S.A., Anuar, M.S., Junita, M.N., Ahmad, R.B., Ghani, F.: Enhanced performance of new family modified double weight codes spectral amplitude coding optical CDMA system network. In: *IEEE International Conference on Control System, Computing and Engineering*, pp. 488–494 (2011)

9. Zahid, A.Z.G., Mandeep, J.S., Susthitha Menon, P., Bakarman, H., Hasoon, F.N., Bakar, A.A.A., Ali, M.A.M.: Performance analysis of multi-weight 2D OCDMA TEDW. In: 3rd International Conference on Photonics, pp. 204–208 (2012)
10. Mohammed, A., Saad, N.M., Aljunid, S.A., Safar, A.M., Abdullah, M.K.: Optical spectrum CDMA: a new code construction for double weight code family. In: International Symposium on Communications and Information Technology, pp. 812–815 (2006)
11. Norazimah, M.Z., Aljunid, S.A., Fadhil, H.A., Md Zain, A.S.: Analytical comparison of various SAC-OCDMA detection techniques. In: 2nd International Conference on Photonics, pp. 1–5 (2011)
12. Al-Khafaji, H.M.R., Aljunid, S.A., Amphawan, A., Fadhil, H.A.: Triple-play services using different detection techniques for SAC-OCDMA systems. In: 3rd International Conference on Photonics, October 2012, pp. 350–354 (2012)
13. Ahmed, N., Aljunid, S.A., Ahmad, R.B., Rashid, M.A.: Novel OCDMA detection technique based on modified double weight code for optical access network. *Elektron. Ir Elektrotech.* **18**, 117–121 (2012)
14. Radhi, I.F., Aljunid, S.A., Fadhil, H.A., Al-Khafaji, H.M.R.: Performance evaluation of spectral amplitude coding signature sequences for OCDMA systems. In: 2nd International Conference on Photonics, pp. 1–4 (2011)
15. Dayang, H.K., Aljunid, S.A.: Optical code division multiple access (OCDMA) using double weight (DW) codes for local area network. In: International Conference on Computer and Communication Engineering (ICCCE 2010) (2010)
16. Hasoon, F.N., Aljunid, S.A., Abdullah, M.K., Shaari, S.: Multi-rate transmissions on SAC-OCDMA system using new enhancement double-weight (EDW) codes. In: 2nd Information and Communication Technologies (ICTTA) (2006)
17. Zhang, M.: Design and performance analysis of novel signature code in two-dimensional optical CDMA systems (2012)

Chapter 15

Performance Evaluation of LTE Scheduling Techniques for Heterogeneous Traffic and Different Mobility Scenarios

Lukmanhakim Sukeran, Mohamed Hadi Habaebi,
Al-Hareth Zyoud, Musse Mohamud Ahmad, Shihab Hameed,
Amelia Wong and I.M. Rafiqul

Abstract In this paper five scheduling algorithms were investigated and their performance was evaluated in terms of Fairness Index, Peak Throughput, Average Throughput and Edge Cell User Throughput. A system level MATLAB simulator was used. The simulation takes into account different types of traffic for several mobility scenarios and propagation channel models. Results indicate that the scheduling algorithms showed some quality in certain parameter of evaluation but lack in other terms. While some scheduling algorithm take the moderate path but still be lacking especially in Edge Cell User Throughput necessitating the use of Relays or femtocells.

15.1 Introduction

Evolution of Universal Mobile Telecommunications System (UMTS) has not reached its end even though with the existence of High Speed Packet Access (HSPA). UMTS Long Term Evolution (LTE) has been introduced in 3rd

L. Sukeran · M.H. Habaebi (✉) · A.-H. Zyoud · M.M. Ahmad · S. Hameed · A. Wong · I.M. Rafiqul

Faculty of Engineering, Electrical and Computer Engineering Department,
International Islamic University Malaysia (IIUM), 53100 Gombak, Kuala Lumpur, Malaysia
e-mail: habaebi@iium.edu.my; habaebi@gmail.com

A.-H. Zyoud
e-mail: alhmtz@yahoo.com

M.M. Ahmad
e-mail: mussemoh@gmail.com

S. Hameed
e-mail: shihab@iium.edu.my

A. Wong
e-mail: amelia@iium.edu.my

I.M. Rafiqul
e-mail: rafiq@iium.edu.my

Generation Partnership Project (3GPP) Release 8 to guarantee the competitiveness of UMTS for the next coming years. The rapid grow of mobile data usage in the recent years such as gaming, mobile channel TV, and other streaming content have concerned in the (3GPP) leading to motivation on LTE. Therefore the work towards 3rd Generation Partnership Project (3GPP) Long Term Evaluation started in 2004 and the targets of LTE standard were set [1].

Orthogonal Frequency Division Multiplexing (OFDM) has been adopted as the downlink transmission scheme for the 3GPP LTE. OFDM is a multicarrier transmission scheme since it splits up the transmitted high bit-stream signal into different sub-streams and sends these over many different sub-channels [2]. OFDM simply divides the available bandwidth into multiple narrower sub-carriers and transmits the data on these carriers in parallel streams. Each sub-carrier is modulated using different modulation scheme, e.g. Quadrature Phase Shift Keying (QPSK), Quadrature Amplitude Modulation (QAM), 64QAM and an OFDM symbol is obtained by adding the modulated subcarrier signals [3].

The scheduling algorithm is the radio resource management technique that is used by the base station to manage and control the available radio resources and assign them efficiently to the available users to meet their service requirement. The minimum resources that could be assigned for a user are called Resource Block (RB). RB includes 12 adjacent OFDM subcarriers. The scheduler task is to assign these RBs to the users in the network. Many scheduling algorithms have been proposed in the literature. So far, Different studies have been conducted to investigate the performance of the proposed algorithms for different scenarios using several simulation platforms.

The work in [4] investigated the performance of five scheduling algorithms for video traffic using 3GPP LTE simulator. The results showed that Maximum-Largest Weighted Delay First (M-LWDF) algorithm performs better than other algorithms like Round Robin (RR), Exponential/Proportional Fair (EXP/PF), Maximum Rate (Max-Rate), and Proportional Fair (PF) in terms of throughput, number of users supported and fairness. In [5], a comparison of different scheduling algorithms for downlink channel was performed using NS-3 simulator. Similarly, NS-3 was used in [6] to evaluate the performance of scheduling algorithms for uplink scenarios. Moreover, LTE-Sim [7] was used in [8] to compare the performance of three different scheduling algorithms in video traffic scenarios. Habaebi et al. [9] evaluated three of the most known scheduling algorithms namely, RR, PF and Best Channel Quality Indicator (BCQI) using LTE system level simulator [10]. They found that the BCQI outperforms RR and PF in terms of throughput and Block Error Rate (BLER).

In this paper five types of scheduling are considered which are RR, Proportional Fair Sun (PFS), BCQI, Resource Fair Maximum Throughput (RF) and Max-Min Fairness (MaxMin). System level simulations were carried out to compare and evaluate the previous algorithms. The performance was evaluated in terms of fairness index, peak throughput, average throughput and cell edge throughput.

15.2 Scheduling Algorithms

The scheduling algorithms that are investigated in this paper are highlighted in the sub sections below:

15.2.1 Resource Fair Maximum Throughput

Resource Fair Maximum Throughput algorithm integrates the Max-rate and Proportional Fair scheduling. This algorithm efficiently employs available radio resource as user's packets are transmitted on a radio resource with a good channel condition. Users are treated according to the rank and the schedulers are either Max-rate or Proportional Fair [11].

15.2.2 Proportional Fair Scheduling

This scheduling algorithm is basically the improvement and less complexity form of the PF scheduling algorithm. Karush-Kuhn-Tucker (KKT) condition was used to reduce the complexity of the PF [12].

15.2.3 Max-Min Fairness

The algorithm key parameter is fairness. It distributes the resource block to achieve optimal fairness. It aims to provide the maximum resource to the minimum data rate for the receiver so that the data rate distribution is fair [11].

15.2.4 Best Channel Quality Indicator

BCQI scheduling policy is to allocate resource blocks to the user with the best channel condition. In order to perform scheduling, terminals send Channel Quality Indicator (CQI) to the base station [13].

15.2.5 Round Robin

RR is proposed to solve the problem of fairness that appears in the BCQI and RF algorithms. It allocates equal time for each user without priority option. Therefore, the channel condition has no impact on the user chance. Basically this algorithm rates the user in terms of first come first serve basis [13]. The fairness is improved, however the throughput is degraded significantly.

Table 15.1 Simulation parameters for LTE system level simulator

Parameter	Value
Frequency	2.14 GHz
Bandwidth	20 MHz
Simulation length	100 TTI
Inter eNodeB distance	500 m
eNodeB TX power	20 dBW
Number of UEs	1, 2, 5, 10, 20, 40 users
Antenna pattern	Omni-directional
eNodeB antenna gain	15 dBi
Uplink delay	3 TTIs
UEs speed	3 and 120 km/h

15.3 Simulation Environment

In this paper, 7 hexagonal base stations (eNodeBs) with various number of user equipments (UEs) are used. The users are randomly located in the eNodeB region of interest. The LTE system level simulator parameters are given in Table 15.1. The mobility is considered in all scenarios (pedestrian and vehicular). Five different traffic types have been considered: VoIP, Video, FTP, HTP, and Gaming.

15.4 Simulation Results and Discussions

Using the parameters presented in Sect. 15.3, the results were generated for different traffic types, for each mobility scenario and for various numbers of users. The figures in this section show the generated result for VoIP service only since there is not enough space to show all the generated results for all traffic types. However, all results were discussed at the end of the section. The results for VoIP traffic for pedestrian UE with speed 3 km/h are presented in Figs. 15.1, 15.2, 15.3, and 15.4.

In terms of fairness index as in Fig. 15.1, BCQI shows declination rapidly compared to other scheduling algorithms as the number of users increase. Other scheduling algorithms have uniform distribution of fairness index ranging in between the values of 0.64 and 0.87. However, in terms of peak throughput as in Fig. 15.2, BCQI scores the highest value of peak throughput 160 Mbps and this value is maintained up to 2 users before it dropped. Other scheduling algorithms showed decreasing trend as the number of user increase and only has the highest Peak Throughput at the smallest number of user which is 1.

For average throughput as in Fig. 15.3, all the scheduler showed a somewhat same behavior. Average throughput decreases as the number of user increases. BCQI comes to have the highest value in throughput, while other scheduling

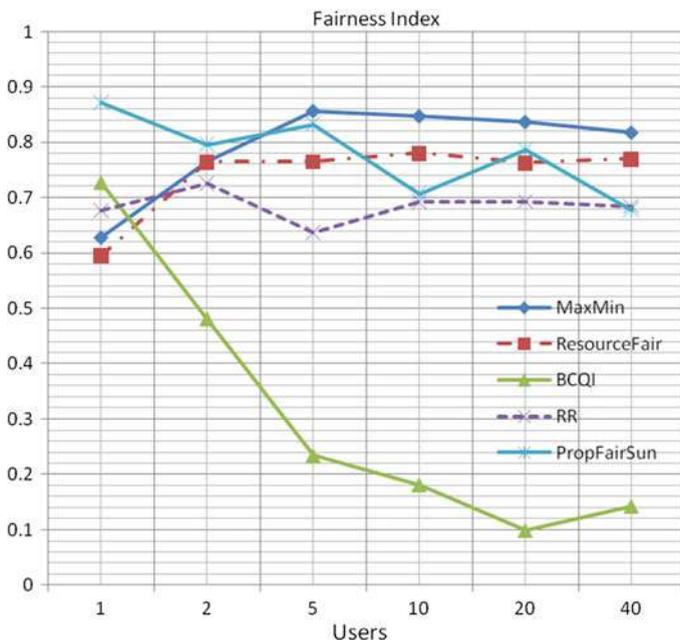


Fig. 15.1 Fairness index against number of users

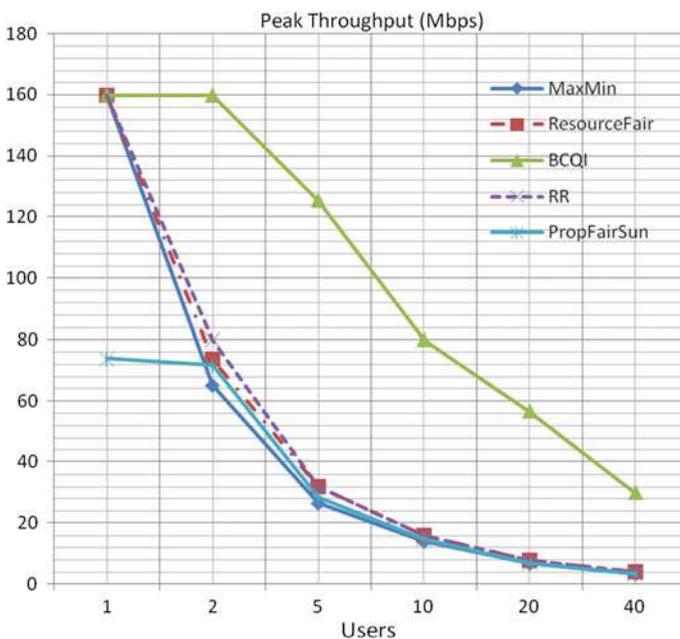


Fig. 15.2 Peak throughput against number of users

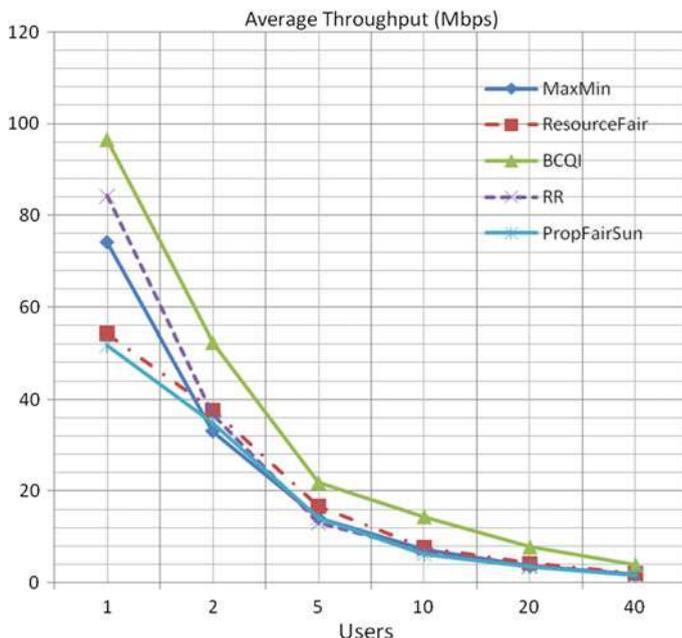


Fig. 15.3 Average throughput against number of users

algorithms are having about the similar values. For the edge users as in Fig. 15.4, PFS comes out as the best for 1 to 2 users before the MaxMin takes the lead for 5 and above number of users. On the other hand, BCQI only treats the edge user when there is only 1 user while as the number of user increases to 2, the edge users are no longer experiencing throughput.

In the second scenario, the speed of the UE was changed to be 120 km/h for VoIP service. Figure 15.5 shows the fairness as a function of number of users. MaxMin has the highest value among other schedulers in case of 1 and 2 UEs. Then Round Robin took the place as the number of users increase from 3 to 40. Round Robin has the smallest decrement of fairness index compare to other schedulers. Same as the first scenario, BCQI has the lowest value among the other schedulers. However, in terms of peak throughput as in Fig. 15.6, BCQI able to achieve the highest throughput value at 160 Mbps and maintain it until 2 users. PFS also able to achieve the value of 160 Mbps but dropped sharply as the number of user increase from 1 to 5 users and above. Other schedulers show almost the same behavior.

In addition, BCQI shows the best performance in terms of average throughput as in Fig. 15.7. The other schedulers have about the similar value to each other. All the schedulers showed the same trend as the value of Average Throughput decrease exponentially as the number of user increases from 1 to 40 in the cell.

Finally, for the edge users as in Fig. 15.8, RR and RF are able to support more than 5 users at the cell edge. BCQI shows the worst case where only can support 1 user.

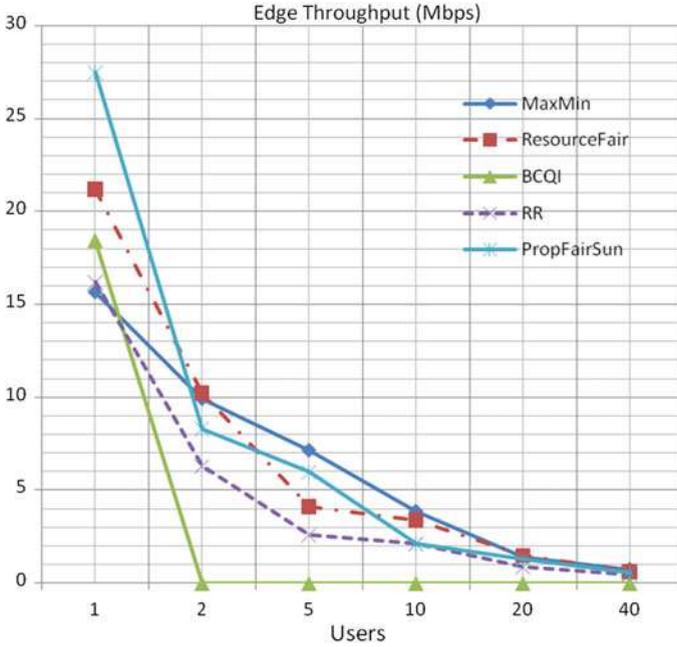


Fig. 15.4 Edge user throughput against number of users

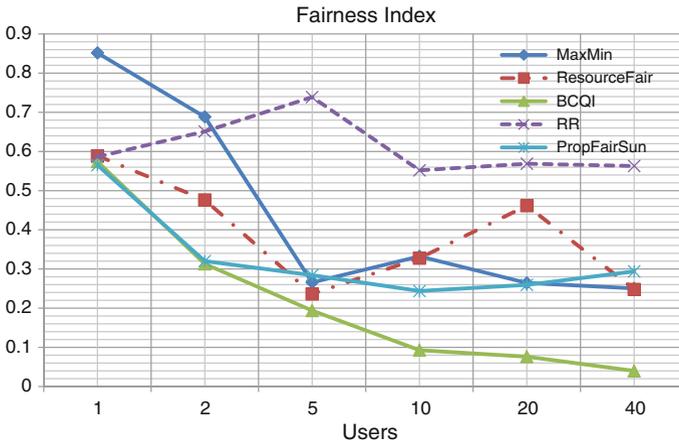


Fig. 15.5 Fairness index against number of users

In all the traffics types' scenarios, generally, all the results show that the throughput decreasing as the number of users increasing, same goes for the edge cell UE. Both of these parameters have inversely proportional relationship to each other. However in terms of the fairness index, RR algorithm able to maintain the

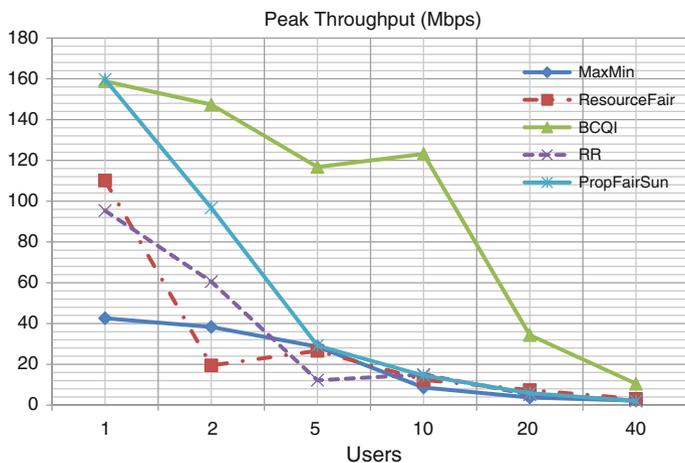


Fig. 15.6 Peak throughput against number of users

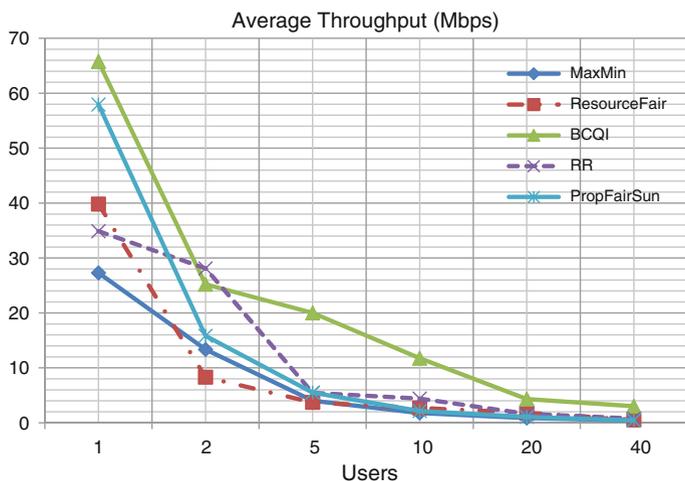


Fig. 15.7 Average throughput against number of users

fairness index value as the number of users increase in several scenarios. In pedestrian and vehicular scenarios, values of fairness index can be considered as stagnant and having fixed range except for BCQI algorithm. However, BCQI performs very well in terms of peak and average throughput.

To sum, BCQI is not suitable for real-time transmission since it seems to cannot provide the QoS for the VoIP when incomes to cell edge users compare to the other scheduler especially Round Robin, the reason behind that is the channel quality for the edge user is worse than the cell center users and using BCQI scheduler the edge user will not be able to get a good service.

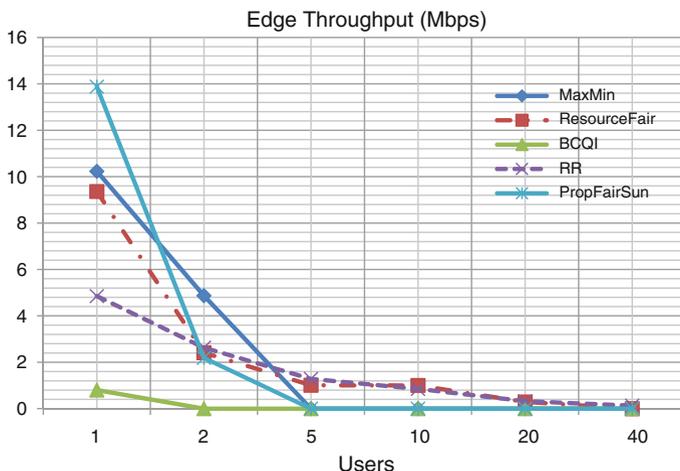


Fig. 15.8 Edge user throughput against number of users

In terms of efficiency, the BCQI can be seen as the most efficient scheduler compare to the other scheduler due to the peak throughput achieved has the highest value show that the spectral efficiency is the highest while, in term of effectiveness, Round Robin is the most effective scheduler since it still provide throughput to the edge users and for high number of users, since RR allocates equal time and data rate for each user in the network. The other scheduler performed moderately.

15.5 Conclusion

All the proposed scheduling algorithms; RR, PFS, RF, MaxMin, and BCQI has been investigated. Fairness index, peak throughput, average throughput, and edge users throughput also has been achieved and the entire scheduling algorithm has been tested for several traffic types. Each one of the scheduling algorithm has shown performance merit in certain criteria of evaluation. For instance, RR has shown that it is good for vehicular channel model, while BCQI has shown the best in achieving peak throughput. However, the best suited scheduling algorithm is still remains argumentative and in need for extensive and more comprehensive improvement not just the scheduling algorithm but the system infrastructure as a whole and also the simulation platform. However, the use of cell edge Relays and Femtocells is necessary.

Acknowledgments This work was funded by E-Science grant 01-01-08-SF0194 from Malaysian ministry of Science, Technology and Innovation (MOSTI).

References

1. 3GPP: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2. TS 36.300, 3rd Generation Partnership Project (3GPP) (2010)
2. Myung, H.G., Goodman, D.: Single Carrier FDMA New Air Interface for LTE. Wiley, New York (2008)
3. Goldsmith, A.: Wireless Communications. Cambridge University Press, Cambridge (2005)
4. Ramli, H., Basukala, R., Sandrasegaran, K., Patachaianan, R.: Performance of well known packet scheduling. In: Proceedings of the IEEE 9th Malaysia International Conference on Communications, Kuala Lumpur, pp. 815–820, (2009)
5. Zhou, D., Baldo, N., Miozzo, M.: Implementation and validation of LTE downlink schedulers for ns-3. In: Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques (SimuTools '13). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, pp. 211–218 (2013)
6. Safa H., Tohme, K.: LTE uplink scheduling algorithms: performance and challenges. In: 19th International Conference on Telecommunications (ICT), pp. 1–6, (2012)
7. Piro, G., Grieco, L.A., Boggia, G., Capozzi, F., Camarda, P.: Simulating LTE cellular systems: an open-source framework. *IEEE Trans. Veh. Technol.* **60**(2), 498–513 (2011)
8. Sahoo, B.: Performance comparison of packet scheduling algorithms for video traffic in LTE cellular network. *Int. J. Mob. Netw. Commun. Telematics (IJMNCT)* **3**(3), 9–18 (2013)
9. Habaebi, M.H., Chebil, J., Al-Sakkaf, A.G., Dahawi, T.H.: Comparison between scheduling techniques in long term evolution. *IJUM Eng. J.* **14**(1), 66–75 (2013)
10. Ikuno, J.C., Wrulich, M., Rupp, M.: System level simulation of LTE networks. In: Proceedings of the 71st IEEE Vehicular Technology Conference, Taipei, Taiwan, May (2010). http://publik.tuwien.ac.at/files/PubDat_184908.pdf
11. Manor R., Romsey: LTE MAC Scheduler and Radio Bearer QoS (2011)
12. Sun, Z., Yin, C., Yue, G.: Reduced-complexity proportional fair scheduling for OFDMA systems. In: Proceedings of the IEEE International Conference on Communications, Circuits and Systems, vol. 2 (2006)
13. Dahlman, E., Parkvall, S., Skold, J., Beming, P.: 3G Evolution HSPA and LTE for Mobile Broadband, 2nd edn. Elsevier, Amsterdam (2008)

Chapter 16

Design for Energy-Aware IP Over WDM Networks with Hibernation Mode and Group-Node Techniques

M.N.M. Warip, Ivan Andonovic, Ivan Glesk, R. Badlishah Ahmad, P. Ehkan, Mohamed Elshaikh Elobaid Said Ahmed, Shamsul Jamel Elias and Fazrul Faiz Zakaria

Abstract The focus of the paper is an investigation and evaluation of energy efficient solutions in IP over WDM core networks using as a foundation, a hierarchy of hibernation modes implementing different degrees of node groupings and fibre links establishment that support a sleep state. It seeks to embed this groups-nodes strategy into an intelligent control plane implementing routing schemes targeting energy consumption, adaptive signalling and traffic engineering. A Group-Nodes mechanism is proposed as a function of topology and node distribution based on a fixed (or geographical) and random (or ownership) principle. The impact of the proposed technique on energy saving and network performance

M.N.M. Warip · R. Badlishah Ahmad · P. Ehkan (✉) · M.E.E.S. Ahmed · S.J. Elias · F.F. Zakaria

School of Computer and Communication Engineering, Universiti Malaysia Perlis, Pauh Putra Main Campus, 02600 Arau, Perlis, Malaysia
e-mail: phaklen@unimap.edu.my

M.N.M. Warip
e-mail: nazriwarip@unimap.edu.my

R. Badlishah Ahmad
e-mail: badli@unimap.edu.my

M.E.E.S. Ahmed
e-mail: elshaikh@unimap.edu.my

F.F. Zakaria
e-mail: ffaiz@unimap.edu.my

I. Andonovic · I. Glesk
Department of Electronic and Electrical Engineering, University of Strathclyde,
204 George Street, Glasgow G1 1XW, UK
e-mail: i.andonovic@strath.ac.uk

I. Glesk
e-mail: i.glesk@strath.ac.uk

is assessed; results are presented and evaluated for various scenarios. Evaluation of this methodology indicates potential reduction in power consumption from 7 % up to 15 % at the expense of reduced network performance.

16.1 Introduction

With regards to zero carbon emission, the issues of green networks technology has become primary interest among researchers. Recent studies shows that rapid changes on the information and communication technology (ICT) devices and deployment of network infrastructures are having a critical effect on carbon footprint.

For this reason, the power consumption prediction as stated in [1], exhibit that the worldwide operation of network equipment accounts for 25GW (yearly average) of the total ICT consumption. Conversely, the joules/bit in telecommunication networks is decreasing with time, the joules/user keeps steadily increasing.

So far, the Smart Sleep Mode is the pinnacle of current green networks technology in which focusing on access networks, offering automated and low power design mechanisms. However, far too little attention has been paid to core networks.

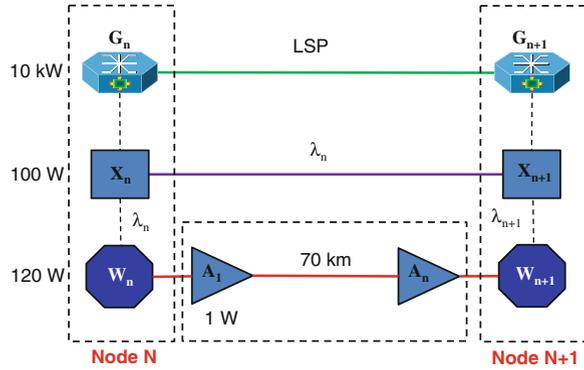
In this paper, the research centre on the development of energy saving schemes that support the evolution of greener core IP over Wavelength Division Multiplexer (WDM) networks. The cornerstone of the adopted strategy is various schemes underpinned by the hibernation state implemented through a modification of the control plane, in particular for transparent network architectures under different scenarios. The research evaluated the impact and constraints that arise under this strategy, to provide useful insights on the viability of the approach for practical energy efficient savings.

16.2 Potential Energy Conservation in Core IP Over WDM Networks

One of the most momentousness cost minimisation strategies in core IP over WDM networks (also sometimes known as IP over Optical or optical IP) [2, 3] in terms of provisioning, operation and maintenance is savings in energy consumption. The major power consumption contributors of equipment in the network are:

- WDM chassis; receiving/transmitting equipment such as transponders modules, short-reach optical interface transponders.
- Optical Switching (OXC) Chassis, opaque/transparent optical transponders, Optical-Electrical-Optical (OEO) conversion.
- Core Router: electronic processing, traffic grooming and aggregation in the IP layer.
- Optical Amplifiers: predominately the Erbium Doped Fibre Amplifier (EDFAs).

Fig. 16.1 An energy model design for IP over WDM networks



- 3R Regenerators: signal regeneration with re-timing, re-amplification and re-shaping operations.
- Control Plane: signalling and routing algorithm modification.

16.3 Network Energy Model Design

In our approach in order to evaluate overall network power consumption and consumed energy per a data bit we used so called equivalent network energy model (see Fig. 16.1) based on multilayer Internet Protocol/Generalized Multi-Protocol Label Switching (IP/GMPLS) over optical layers. In this model, a network carrier bandwidth of OC-192 and average energy consumption of 1019 nJ per bit was assumed following [4, 5, 6].

The parameter G_n denotes a Router's dissipated power of 10 kW within its energy consumption of 1000 nJ/bit. X_n represents Optical Cross-Connect (OXC) dissipating 100 W and consumes 10nJ/bit. W_n denotes Wavelength Division Multiplexing (WDM) part of the node with dissipating power of 120 W and energy consumption of 12 nJ/bit. A_n represents the consumption owing to Erbium-Doped Fibre Amplifiers (EDFA) within connection spans placed at 70 km intervals, the power consumption is estimated to be 1 W with energy of 0.1 nJ/bit.

16.3.1 Energy Per Bit

Further we define energy per bit consumed by the node as $E_b = P_T/C$ where P_T represents the node total power consumption and C is the bandwidth offered by the network link.

16.3.2 Energy Consumption and Power Consumption

The total power consumption of the link is given by:

$$P_{TOTAL} = \left[\underbrace{\Delta p_{CONTROL\ PLANE}}_{IPLAYER} \right] + \left[\underbrace{OXC + \Delta p_{WDM} + TRANSPOENDERS + \Delta p_{EDFA}}_{OPTICAL\ LAYER} \right] \quad (16.1)$$

where $\Delta p_{CONTROL\ PLANE}$, Δp_{OXC} , Δp_{WDM} , $\Delta p_{TRANSPOENDERS}$, and Δp_{EDFA} represent power consumed by the IP/GMPLS router, OXC, WDM, transponders/transmitters, and EDFA optical amplifier, respectively.

Similarly, the total energy consumption of the link is given by:

$$E_{TOTAL} = \left[\underbrace{\Delta e_{CONTROL\ PLANE}}_{IPLAYER} \right] + \left[\underbrace{OXC + \Delta e_{WDM} + TRANSPOENDERS + \Delta e_{EDFA}}_{OPTICAL\ LAYER} \right] \quad (16.2)$$

where $\Delta e_{CONTROL\ PLANE}$, Δp_{OXC} , Δp_{WDM} , $\Delta p_{TRANSPOENDERS}$, and Δp_{EDFA} represent power consumed by the IP/GMPLS router, OXC, WDM, transponders/transmitters, and EDFA optical amplifier, respectively.

The total energy consumption in IP over WDM networks calculated per data bit Ebit in order to support the network offered load can be defined as:

$$\sum_{k=0}^{\infty} E_{BIT} = \left[\underbrace{\frac{\sum_{k=0}^{n-1} CONTROL\ PLANE}{C}}_{IPLAYER} \right] + \left[\underbrace{\frac{\sum_{k=0}^{n-1} \Delta p_{WDM}}{C} + \frac{\sum_{k=0}^{n-1} \Delta p_{TRANSPOENDERS}}{C}}_{OPTICAL\ LAYER} \right] + \left[\underbrace{\sum_{k=0}^{n-1} EDFA + (\alpha + 1) \sum_{k=0}^{n-1} \Delta e_{OXC}}_{OPTICAL\ LAYER} \right] + \left[\underbrace{\beta}_{CONSTANT} \right] \quad (16.3)$$

where Δe_{EDFA} is the energy consumed by EDFA; Δe_{OXC} is the energy consumed by the node's OXC; β represent the noise factor associated with the Bit Error Rate (BER) and a heat transfer rate in network equipments; and finally, α is number of hops.

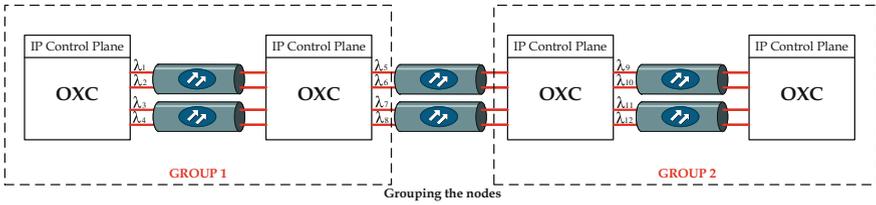


Fig. 16.2 Hibernation mode: grouped nodes structure

16.4 Hibernation Mode: Group Nodes Distribution Factors

The Group Node structure in relation to the hibernation mode is illustrated in Fig. 16.2. In this architecture, nodes consisting of IP Router and OXC are interconnected by point-to-point optical fibre links.

The exchange of messages updating energy consumption profiles can be described by a sequence diagram; Fig. 16.3 illustrates the messages sequence diagram for the transition from ON to OFF state. For end-to-end provisioning - from ingress to egress node—after path computation, the wavelength is reserved by signalling. Node 1 transmits a connection request along the link to reserve the wavelength and establish an end-to-end channel through the LSP *Path/Resv* Message. The LSP setup request is then forwarded to the next node until the message request is at the egress node. If there is an idle node at an intermediate node (in this case Node 3), after the nested hold-off timer expires, the hibernation notification message is propagated back along the path to release the reserved wavelength. The loopback LSP *Resv_Confirm* message will be transmitted back along the link until at the ingress node to request the suspension of the idle node. If Node 3 receives a LSP setup request message to place the node in Sleep state, it sends a *Resv_Err* message to acknowledge the ingress node that Node 3 is in powering off state. As a result, the network updates the routing table and TED topology. The ingress node releases a LSP by propagating *Path_Tear* message and powers down the connection to the idle node.

Figure 16.4 presents a message sequence diagram for the transition from OFF state to ON state. Node 3 detects traffic and changes its state to BUSY (active transition state) and full power operation is resumed, confirmed by sending a notify message to inform the adjacent node that it is in the process of waking-up. A *Resv_Tear* message is sent to the ingress node to notify that node is powering to ON state.

16.4.1 Fixed (Geographical) Nodes Effect

Fixed node or Geographical Node groups are defined as a grouping topology that contains selected neighbouring nodes and grouped as disjoint clusters.

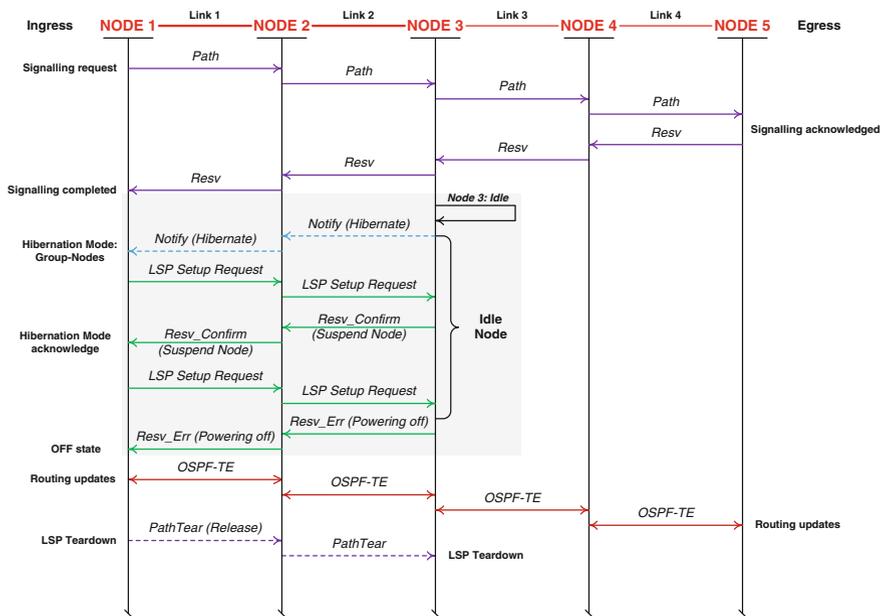


Fig. 16.3 Messages sequence diagram for transition from ON state to OFF state of hibernation: group-nodes

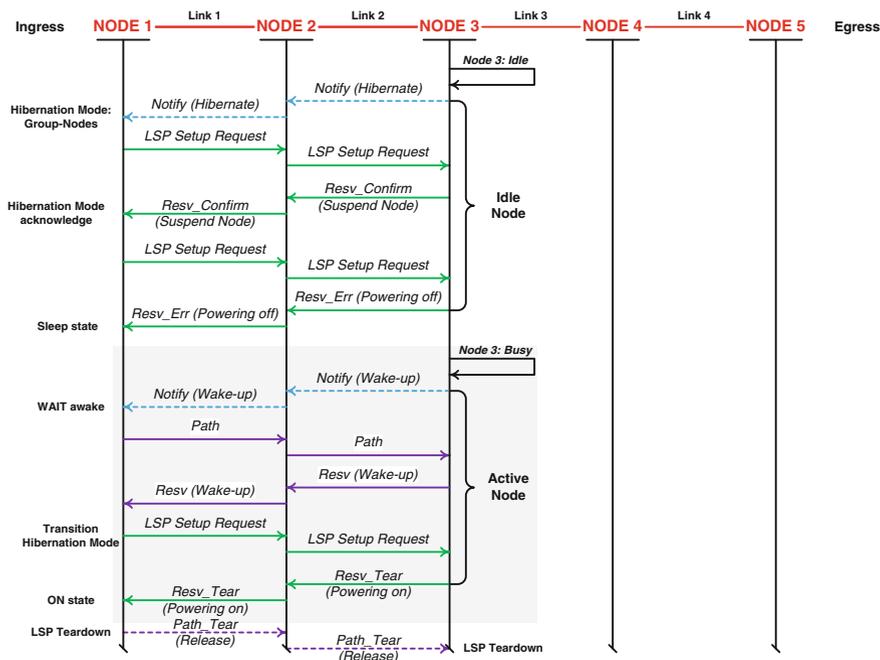


Fig. 16.4 Messages sequence diagram for transition from OFF state to ON state of hibernation: group-nodes

16.4.2 *Random (Ownership) Nodes Effect*

Random node or Ownership-based node groupings are defined as nodes belonging to the same owner (service provider) e.g. organization that having many entities under the same company name.

16.5 Simulation Results

The Group-Nodes hibernation mode that invokes cluster based architectures was evaluated and investigated. By dividing the nodes into several disjoint sets, as well as providing each node with geographical and ownership topology settings, produces clusters adopting sleep cycles to reduce power consumption. Figure 16.5 illustrates the network topology which is being utilised by the European Optical Network (EON network). The EON network has a Full Mesh Network Topology with 9 nodes and 20 bidirectional fibre links [7, 8]. The IP/GMPLS nodes are linked by bidirectional pairs of single mode fibres. The EON network topology was used in our simulations and was based on a discrete event modelling tool known as OMNet++ (Object Modular Network Tested in C++). It has been assumed that all links are equal in terms of number of wavelengths (eight), that the message length is fixed at 256 bytes, and a nodal processing delay is 20 ms. All EON network Nodes are capable to maintain information on their total power consumption as well as energy per bit consumed. Wherein, the standard GMPLS signalling and routing protocols are implemented following the Internet Engineering Task Force (IETF) standard [9, 10].

The performance metrics takes into account the average power consumption, blocking probability and average request blocking [9–13]. We also assumed that lightpath requests are uniformly distributed. Note that, the inter-arrival connection requests are independent Poisson processes with an arrival rate of α and the queue lengths exponentially distributed with the expected service rate time of $1/\mu$ measured in seconds. Therefore, the network offered load is α/μ .

Full Mesh Network Topology: European Optical Network (EON) (Fig. 16.5). The average power dissipation and energy consumption values assigned to each node are captured in the network energy model. In this architecture, the power consumption of nodes comprises the core router (10 kW), OXC (100 W), WDM (120 W) and EDFAs (1 W) placed at 70 km intervals along links. For example, the power consumption (reference value) between Node A and Node G (Fig. 16.5) linked by the single mode optical fibre across a distance of 2090 km at a data rate of 10 Gb/s is 29 W. Therefore, the total power consumption between Node A and Node G is 10.249 kW and the energy per bit is 1024.9 nJ (Eqs. 16.1–16.3).

The proposed Hibernation concept was verified on described EON by implementing group-nodes schemes. These nodes/links are then put into “hibernation” or an “SLEEP state”, in which nodes have suspended their unused functionalities

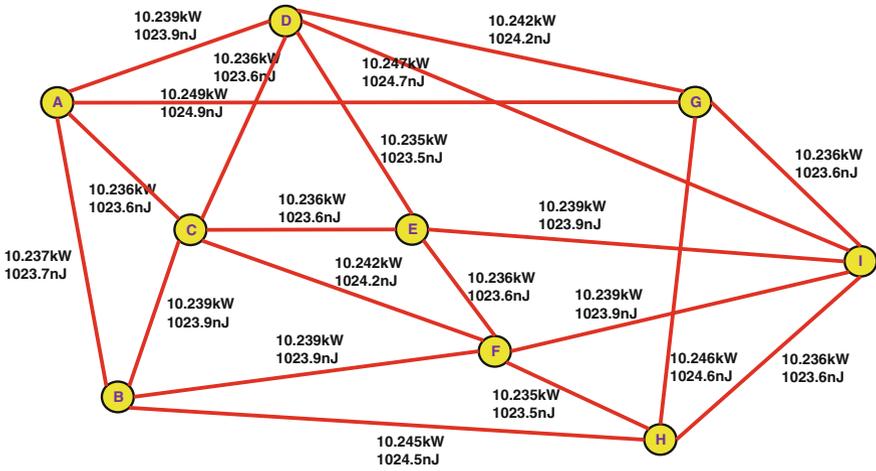


Fig. 16.5 Power consumption for full mesh European Optical Network (EON)

Table 16.1 EON node grouping categorization

Group	Geographical	Ownership-based
G1	Nodes {A, B}	Nodes {B, H}
G2	Nodes {C, D, E}	Nodes {D, E, G}
G3	Nodes {F, G, H, I}	Nodes {A, C, F, I}

(e.g. unused ports/interfaces, Mux/DeMux capabilities, signalling gates, unused wavelengths, etc.) and keep only the minimum network operation activities.

The aim is to evaluate the impact of a route to energy saving through hibernation in core optical IP network by performing the Fixed (Geographical) grouping and Random (Ownership) grouping for full mesh network topologies. Note that, these results are based on cross-layer optical/IP domains integration and previous research produced results based in the optical domain only. Therefore, results are difficult to compare in terms of power savings and network performance.

Using the full mesh EON (Fig. 16.5), hibernation settings are applied to network nodes based on the group membership type (see node grouping categorization in Table 16.1).

Figures 16.6 and 16.7 present the average power consumption and average request blocking for various “Geographical” (adjacent nodes) groupings as a function of offered network load for the EON network mesh topology. For “All Groups HM” (case when nodes’ unused functionalities are suspended) ~0.137 kW of power is saved per node’ (Fig. 16.6) but the probability of blocking is ~55 % (Fig. 16.7). For Groups “G2 and G3 = HM” or Groups “G1 and G2 = HM”, the power savings of 0.12 kW or 0.10 kW is obtained respectively (Fig. 16.6) with a corresponding blocking probability of 33 % or 10 %, respectively (Fig. 16.7).

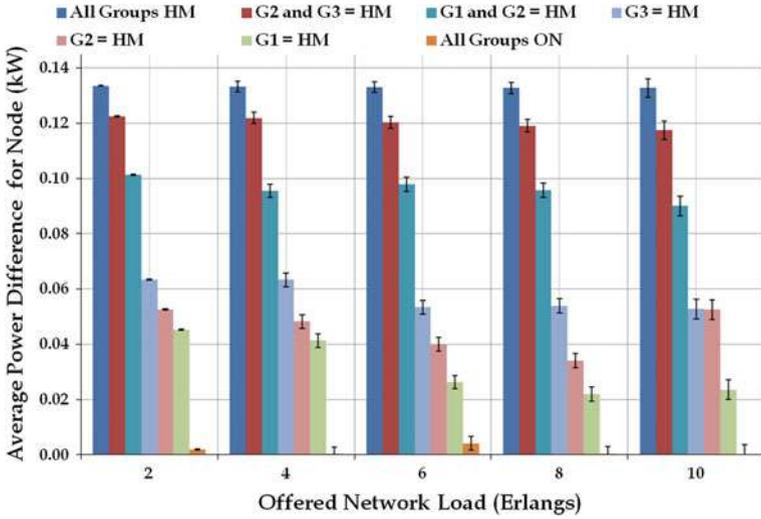


Fig. 16.6 Simulation results for different EON Fixed (geographical) node groupings. ‘HM’ stands for hibernation mode

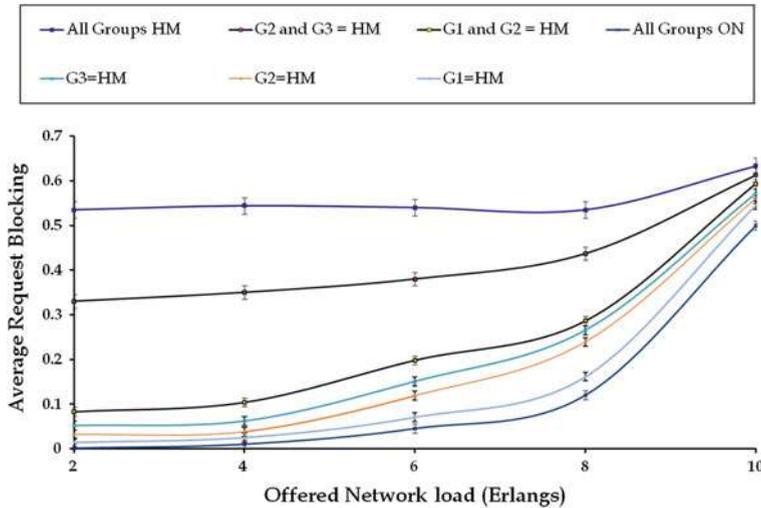


Fig. 16.7 Request blocking for EON Fixed (Geographical) grouping of nodes

“All Groups ON” yields the lowest blocking probability but the power savings per node are minimal. The EON network also becomes congested for network loads exceeding 8 Erlangs. The trade-off between a reduction in energy consumption and the probability of blocking is evident.

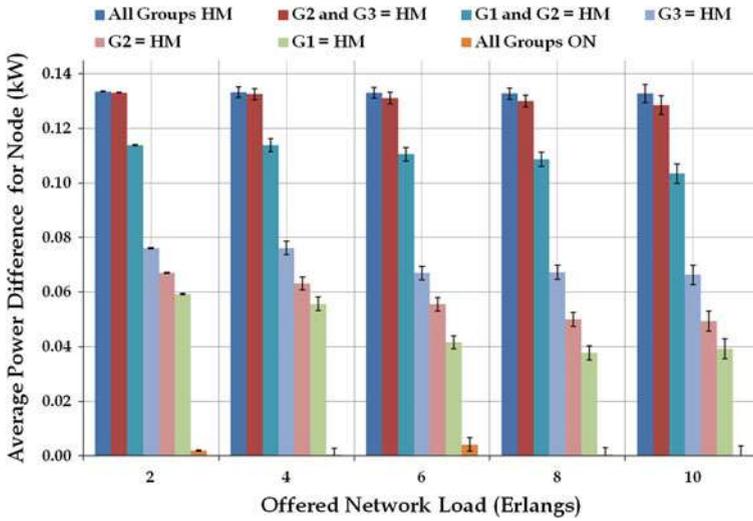


Fig. 16.8 Average power consumption difference per node as a function of network load for different EON random (Ownership) node grouping

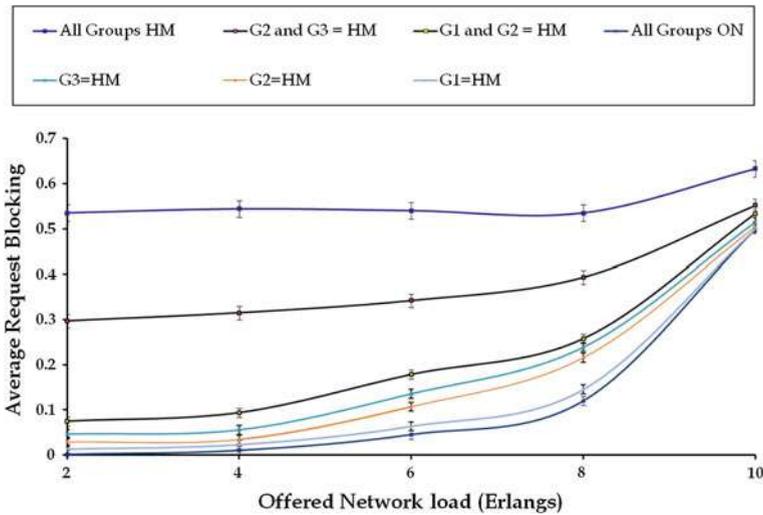


Fig. 16.9 Request blocking probability as a function of network load for EON random (Ownership) grouping of nodes

Figure 16.8 depicts the average power consumption for different groups with respect to offered network load and Fig. 16.9 presents the blocking probability for the EON ownership (random) grouping of nodes. As expected, savings in power with ownership grouping in the full mesh (EON) topology improves when

compared to the partial mesh (NSFnet) topology, particularly in the case of ownership grouping. In this case, the power savings for grouping “G2 and G3” or “G1 and G2” are most significant, the improvement being ~ 0.135 and ~ 0.11 kW respectively (Fig. 16.8); the corresponding blocking probability is 29 and 8 %, respectively (Fig. 16.9). The reason for this is that the optical bypass at intermediate nodes reduces the number of required core router electrical ports in the IP layer and thereby, the energy owing to this electrical equipment is saved. The EON ownership-based node groupings (Figs. 16.6 and 16.8), “All Groups = HM” and “All Groups = ON” deliver similar results as the geographical node groupings.

16.6 Conclusion

In this paper we have presented an evaluation of an approach to energy saving in IP over WDM networks-based GMPLS control plane based on hibernation. The impact of the grouping of nodes following the principles of hibernation on network performance as a function of energy saving was examined and quantified for two representative network topologies. The approach was implemented through two grouping network node strategies—Geographical and Ownership-based. Results show that hibernation has the potential to deliver energy savings at the expense of reduced network performance. Results show that the “Ownership-based node groupings for “All Groups = HM” (hibernate) and “All Groups = ON” delivers similar performance as the Geographical nodes groupings.

Acknowledgments The research leading to these results has received funding from the Ministry of Education Malaysia under grant Fundamental Research Grant Scheme (FRGS).

References

1. Pickavet, M., Vereecken, W., Dameyer, S., Audenaert, P., Vermeulen, B., Develder, Colle, D., Dhoedt, B., Demeester, P.: World energy needs for ICT: the rise of power-aware networking. In: Proceedings of the Advanced Networks and Telecommunication Systems, Bombay, India (2008)
2. Musumeci, F., Tornatore, M., Pattavina, A.: A power consumption analysis for IP-over-WDM core network architecture. *IEEE/OSA J. Opt. Comms. Netw.* **4**, 108–117 (2012)
3. Rajagopalan, B., Luciani, J.V., Awduche D.O.: IP over optical networks: a framework. RFC3717 IETF. (2004)
4. Balinga, J., Ayre, R., Hinton, K., Sorin, W.V., Tucker, R.S.: Energy consumption in optical IP networks. *IEEE J. Lightw. Technol.* **27**(13), 2391–2403 (2009)
5. Bathula, B.G., Alresheedi, M., Elmirghani, J.M.H.: Energy efficient architectures for optical networks. In: Proceedings of the London Communications Symposium, pp. 1–4. University College, London (2009)
6. Cisco Systems Data Sheets: <http://www.cisco.com> (2013)
7. Shen, G., Tucker, R.S.: Energy-minimized design for IP over WDM networks. *IEEE/OSA J. Opt. Comms. Netw.* **1**(1), 176–186 (2009)

8. Kim, Y., Lee, C., Kevin Rhee, J.K., Lee, S.: IP over WDM cross-layer design for green optical networking with energy proportionally consideration. *IEEE. J. Lightw. Technol.* **30**(13), 2088–2096 (2012)
9. Albarrak, S.: Failure recovery in distributed GMPLS-based IP-over-optical networks. Ph.D. thesis, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK (2008)
10. Chabarek, J., Sommers, J., Barford, P., Eitan, C., Tsiang, D., Wright, S.: Power awareness in network design and routing. In: *Proceedings of the 27th Conference on Computer Communications INFOCOM*, Phoenix, pp. 1130–1138 (2008)
11. Ben Yoo, S.: Energy efficiency in the future internet: the role of optical packet switching and optical-label switching. *IEEE J. Sel. Quantum Electr.* **17**(2), 406–418 (2011)
12. Berger, L.: Generalized Multi-Protocol Label Switching (GMPLS) signalling functional description. RFC3471. IETF (2003)
13. Lang, J.: Generalized Multi-Protocol Label Switching (GMPLS) signalling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) extension. RFC3473. IETF (2003)

Chapter 17

Content Based Image Retrieval Using Color Layout Descriptor and Generic Fourier Descriptor

Muhammad Imran, Rathiah Hashim and Noor Elaiza

Abstract Image databases have been a significant part of systems or applications in many fields including education, forensics and medical sciences. However, due to the increase in the number of digital resources especially images, the demand to manage these databases such as storing and retrieving the images effectively and efficiently from these databases is crucial. Content Based Image Retrieval (CBIR) is one of the best techniques used to retrieve similar images from the image database. Even though it is a good and well-researched technique, still there are some issues to be improved. One of the challenges in CBIR is to represent the image as a feature vector. In this paper, we proposed a new feature vector using Generic Fourier Descriptor (GFD) and Color Layout Descriptor (CLD) to increase the accuracy of the retrieval process. We tested our technique with the Coral Database and compared the results with other CBIR techniques. Proposed method achieved better results than previous techniques. The proposed technique can be used by the forensic department for identification of criminal suspect using forensic database.

Keywords CLD · CBIR · Color layout descriptor · GFD

M. Imran (✉) · R. Hashim
University Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
e-mail: malikimran110@gmail.com

R. Hashim
e-mail: radhiah@uthm.edu.my

N. Elaiza
University Teknologi MARA Malaysia, Shah Alam, Malaysia
e-mail: elaiza@tmsk.uitm.edu.my

17.1 Introduction

Due to powerful processors and cheap memories, a variety of applications are using large image databases. Different fields such as medicine, geography, advertising, architecture and design are using large image databases to help their users [1]. Therefore, an effective and efficient image retrieval technique is now a necessity. Two approaches to search the image are available. One is the Keyword Based Image Retrieval (KBIR) and the other is Content Based Image Retrieval (CBIR). The KBIR searches the images based on keywords tagged with an image. KBIR faces many problems—keywords are manually annotated, the person adding the keyword may have a different perception about the image than the one who is searching the image. To overcome these drawbacks, researchers introduced CBIR. In CBIR, searching is performed by the visual features of the images. Therefore, there is no further need for manual keywords annotation.

CBIR used color, texture and shape features of the image. In CBIR, features of all images in the image database are extracted and saved as feature database. When a user input query image to the system, system checks the similarity between the query image features and image database feature through any similarity measure such as Euclidean distance, Manhattan distance, earth movers distance etc. and display ranked result to the user. Due to the high demand for searching from an image database, CBIR becomes a popular image searching technique. Particle swarm optimization (PSO) with relevance feedback was used by the Broilo [2] to enhance the performance of CBIR. Broilo formulates the image retrieval processes as an optimization problem and applied PSO on CBIR. PSO was proposed by Kennedy and Eberhart [3] in 1995 and modified by different research works such as [4–6].

However, CBIR is facing problems and accuracy achieved up to a limit only. In this paper, we proposed a new signature development approach for CBIR. We used the color and shape features to search similar images. Color Layout Descriptor (CLD) is used for the color feature extraction and Generic Fourier Descriptor (GFD) is used for shape feature extraction. Both feature vectors are combined and then, similarity measure is performed using Manhattan distance. Section 17.2 discusses about the related work and proposed approach is illustrated in Sect. 17.3. Sections 17.4 and 17.5 are reserved for the results and analysis. Finally, the paper is concluded in Sect. 17.5.

17.2 Related Works

Bhuravarjula and kumar [7] used color moments to enhance the accuracy of CBIR. They divided the image into four segments and then extracted the color moments from all segments and clustered them into four classes. Finally, they calculated the means moments for each class which was further used for the similarity measure

between the query image and the database images. Coral Database was used to perform experimental work. Precision and recall were used as performance metrics and results were compared with the previous CBIR feature extraction techniques.

Imran et al. [8] used color histogram for searching similar images. Center moment is adopted to describe the histogram. Each image is divided into 4×4 sub images and each sub images is divided into HSV components to generate the histogram. Experiments are performed on Coral Database using precision as performance metric.

Zhang and Zou [9] used the color feature and edge direction features for CBIR. For edge direction feature, Edge Histogram Descriptor (EHD) of MPEG-7 was used. Experiments were performed using Coral Database to assess the performance of proposed technique. Recall and Precision are calculated as performance metrics. The combination of color and texture feature was applied by Soman et al. [10]. Discrete Cosine Transforms (DCT) was used for texture feature extraction while mean, deviation and skewness from color moments were used for the color feature extraction. The proposed technique was tested on the Coral Database having 1000 images and achieved better accuracy then previous CBIR techniques. Hema-chandran and Singh [11] combined color moments with Gabor Texture descriptor. The image is divided into horizontally three non-overlapping regions to extract first three moments of the color distribution. Performance of proposed technique is tested using Coral Database, which contains images from 10 different categories and each category has 100 images.

Abubacker et al. [12] used color, texture and shape features for the image. For the color feature, they used the spatial based color moments. First they divided the image into 25 blocks then calculated the Red Green Blue (RGB) values of each block. RGB values are converted to Hue, Saturation Intensity (HSI). According to the author the three color moments; mean, variance and skewness are effective and efficient for the color distribution of images. The formula for mean, variance and skewness are given below:

$$\text{Mean}(\mu_i) = \frac{1}{N} \sum_{j=1}^N f_{ij} \quad (17.1)$$

$$\text{Variance}(\sigma_i) = \left(\frac{1}{N} \sum_{j=1}^N (f_{ij} - \mu_i)^2 \right)^{1/2} \quad (17.2)$$

$$\text{Skewness}(S_i) = \left(\frac{1}{N} \sum_{j=1}^n (f_{ij} - \mu_i)^3 \right)^{1/3} \quad (17.3)$$

where, f_{ij} is the value of the i th color component of the image block j and N is the number of blocks in the image. For the texture feature, author used the Gabor filter. Author applied the 2D Gabor function to obtain the set of Gabor filter with different scale and orientation. By using Gabor filter, author performed convolution

on the image to obtain the Gabor transform. Invariant shape features were used to extract the shape features. Following are the steps taken by the author to extract the shape feature.

1. Based on the threshold value the image is converted to the binary image.
2. Using canny algorithm the edges of the binary image are detected.
3. The centroid of the object is obtained by arranging the pixels in clockwise order and forming a closed polygon.
4. The centroid distance and complex coordinate function of the edges is found.
5. The farthest points are found and Fourier transform is applied on them.

17.3 Proposed Approach

To represent the image as feature vector we used Shape and Color features. To extract the shape features, GFD is used. About the detail of GFD, we refer our readers to Zhang and Lu [13]. GFD containing 36 features reflecting 4 radial frequencies and 9 angular frequencies are selected to index the shape. The selected 36 features build a feature vector which is combined with the CLD feature vector to form the final feature vector. Color feature is one of the basic features used to retrieve the images. It is most intuitive and obvious feature of the image and color is easily extracted. To describe color information different methods are available. In this paper, we have used the CLD from MPEG-7 standard to extract the color information from the image. Following are the steps to extract the color feature vector from the image.

1. Divide the image into 8×8 blocks
2. Single representative color is selected from each block
3. The selection results in a tiny image icon of size 8×8
4. The color space is converted from RGB to YCbCr
5. The luminance (Y), blue and red chrominance (Cb and Cr) are transformed by 8×8 DCT
6. A zigzag scanning is performed on these three sets of DCT coefficients
7. As a result we obtain three matrixes for each block of Y, Cb and Cr color space
8. Take sum of each matrix to get three feature vectors
9. Finally horizontally concatenate three feature vectors to obtain a final feature Vector for an image

The feature vector calculated by GFD and CLD are combined in one feature vector which represent the each image in the database. When user input query image to the system to search similar images, the similarity between the query image and the database images is calculated using Manhattan Distance and results are sorted, ranked and presented to the user.

17.4 Results

To assess the proposed technique, experiments were performed using the Coral Database. The database contains images from 10 different categories and each category has 100 images, so a total of 1000 images are available in the Coral Database. To validate the proposed technique, the results were compared with Variance Segment Method [7] and Histogram based taken from [14]. The proposed approach is implemented in Matlab 2010b. For performance evolution precision is used as performance metric, which is reported by previous work such as Hiremath et al. [15], Banerjee et al. [14] and Wang et al. [16]. Following expression is used to calculate the precision.

$$Precision = \frac{Number\ of\ True\ Positive}{Number\ of\ True\ Positive + False\ Positive}$$

In our experiments, we select 10 random images from each category as the query image and perform retrieval process. The process is repeated for 10 times and category wise average precision is calculated. Average precision is calculated for 20, 30 and 40 top retrieval where 20 top retrieval means 20 most similar images are displayed and this can be represented as P@20. The results achieved by the proposed feature vector technique are shown in Table 17.1 (Fig. 17.1).

17.4.1 Comparison with Previous Methods

Table 17.2 illustrates the comparison of the proposed technique with Simple Hist and Variance Segment Method.

Table 17.1 Performance of proposed technique at different top retrieval

Class	P@40	P@30	P@20
Africa	0.27	0.32	0.27
Beach	0.44	0.34	0.48
Buildings	0.22	0.26	0.29
Buses	0.37	0.36	0.44
Dinosaurs	1.00	1.00	1.00
Elephant	0.49	0.50	0.64
Flower	0.82	0.83	0.86
Horses	0.62	0.68	0.75
Mountains	0.40	0.19	0.44
Food	0.23	0.29	0.28
Avg	0.48	0.50	0.54

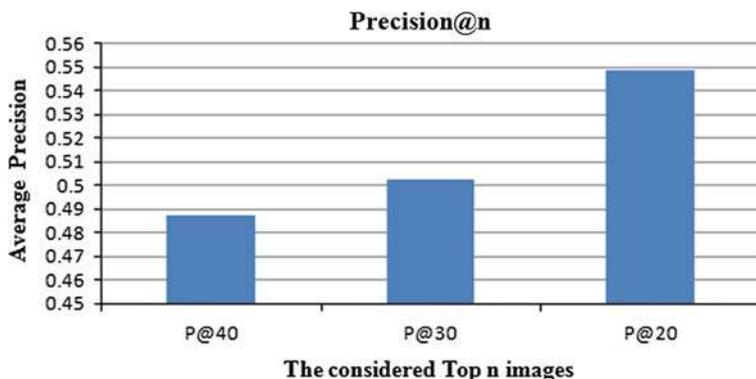


Fig. 17.1 Average precision of proposed method at different top retrieval

17.5 Analysis

From Table 17.2 and Fig. 17.2, it is obvious that the proposed method has better results than Histogram based and variance segment method. The overall average precision of Histogram based method, Variance Segment and proposed method are 0.36, 0.306 and 0.50 respectively. Histogram based method achieved 36 %, Variance Segment method achieved 30 % while proposed technique achieved 50 % accuracy rate. Figure 17.2 illustrates the comparison of proposed method with Simple Hist where average accuracy of proposed method is calculated for top 30 and 20 retrieval results. The graph depicted that the proposed method has better precision than both previous methods. Category wise performance comparison of the proposed method with Variance Segment method is presented in Fig. 17.3.

Table 17.2 Comparison of proposed method with previous methods

Class	Simple hist	Variance segment	Proposed method
Africa	0.3	0.13	0.32
Beach	0.3	0.26	0.34
Buildings	0.25	0.11	0.26
Buses	0.26	0.17	0.36
Dinosaurs	0.9	0.96	1.00
Elephant	0.36	0.34	0.5
Flower	0.4	0.49	0.83
Horses	0.38	0.2	0.68
Mountains	0.25	0.25	0.4
Food	0.2	0.15	0.29
Avg	0.36	0.306	0.5

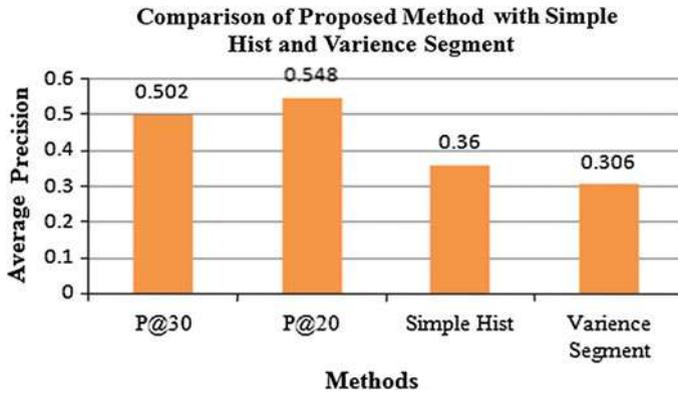


Fig. 17.2 Comparison of the proposed method with simple hist and variance segment

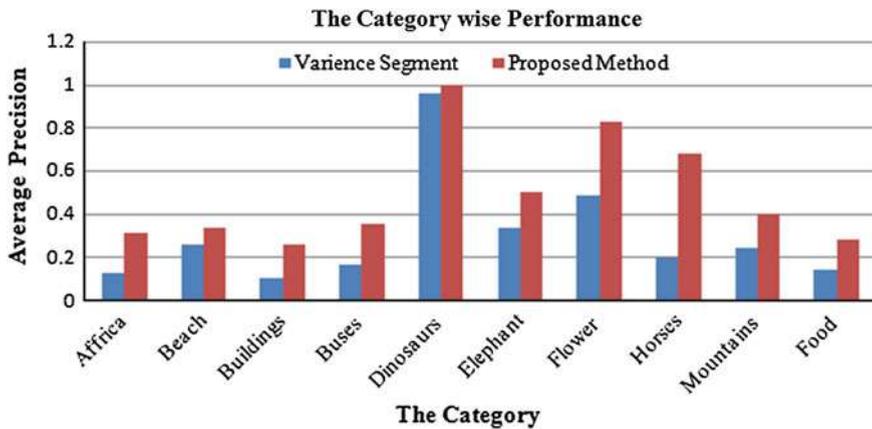


Fig. 17.3 Category-wise comparison of proposed method (P@30) with previous methods

For all categories, proposed method achieved higher average precision than Variance Segment method. For the class dinosaurs, the proposed method has 100 % retrieval rate as the precision reaches 1.0.

17.6 Conclusion

This paper proposed and implemented a new signature technique to improve the performance of CBIR. This new signature technique is the combination of the color and shape features. Color features are extracted using CLD while GFD is used to extract the shape features. The performance of proposed technique is tested on coral Database containing the images from 10 different categories. To validate

the proposed technique, results were compared with the previous CBIR techniques. The proposed technique increased the performance of CBIR in term of accuracy. Previous techniques achieved 30 % and 36 % while proposed technique achieved 50 % accuracy. For future work, the performance of the proposed system, texture feature can be combined with color and shape features.

Acknowledgments The researchers would like to thank University Tun Hussein Onn Malaysia (UTHM) for supporting this project under Project Vote No 1315

References

1. Lee, S.M., Bae, H.J., Jung, S.H.: Efficient content-based image retrieval methods using color and texture. *ETRI J.* **20**(3), 272–283 (1998)
2. Broilo, M., Natale, F.G.B.D.: A stochastic approach to image retrieval using relevance feedback and particle swarm optimization. *IEEE Trans. Multimedia* **12**, 11 (2010)
3. Kennedy, J., Eberhart, R.: Particle swarm optimization In: *IEEE International Conference on Neural Networks*, pp. 1942–1948 (1995)
4. M. Imran, et al., “Modified Particle Swarm Optimization with student T mutation (STPSO),” in *Computer Networks and Information Technology (ICCNIT)*, 2011 International Conference on, 2011, pp. 283–286
5. Imran, M., et al.: Particle swarm optimization (PSO) variants with triangular mutation. *J. Eng. Technol.* (2013)
6. Imran, M., et al.: Opposition based particle swarm optimization with student T mutation (OSTPSO). In: *Data Mining and Optimization (DMO)*, 2012 4th Conference on, pp. 80–85 (2012)
7. Bhuravarjula, H., Kumar, V.: A novel content based image retrieval using variance color moment. *Int. J. Comput. Electron. Res.* **1**(3), 93–99 (2012)
8. Imran, M., et al.: New Approach to Image Retrieval Based on Color Histogram. In: Tan, Y., et al. (eds.) *Advances in Swarm Intelligence*, vol. 7929, pp. 453–462. Springer, Berlin Heidelberg (2013)
9. Zhang, J., Zou, W.: Content based image retrieval using color and edge direction features. In: *IEEE 2nd International Conference on Advanced Computer Control (ICACC)*, vol. 5, pp. 459–462 (2010)
10. Soman, S., Ghorpade, M., Sonone, V., Chavan, S.: Content based image retrieval using advanced color and texture features. In: *International Conference in Computational Intelligence (ICCI)*, vol. 3, (2012)
11. Singh, S.M., Hemachandran, K.: Content-based image retrieval using color moment and gabor texture feature. *Int. J. Comput. Sci. Issues (IJCSI)* **9**(5) 299 (2012)
12. Abubacker, K., Indumathi, L.: Attribute associated image retrieval and similarity re ranking. In: *International Conference on Communication and Computational Intelligence (INCOCCI)*, pp. 235–240 December 2010
13. Zhang, D., Lu, G.: Generic fourier descriptor for shape-based image retrieval, in proceedings. In: *2002 IEEE International Conference on Multimedia and Expo*, pp. 425–428 (2002)
14. Banerjee, M., Kundu, M.K., Maji, P.: Content-based image retrieval using visually significant point features. *Fuzzy Sets Syst.* **160**(23), 3323–3341 (2009)
15. Hiremath, P., Pujari, J.: Content based image retrieval using color boosted salient points and shape features of an image. *Int. J. Image Process.* **2**(1), 10–17 (2008)
16. Wang, J., Li, J., Wiederhold, G.: Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(9), 947–963 (2001)

Chapter 18

Pilot Based Pre FFT Signal to Noise Ratio Estimation for OFDM Systems in Rayleigh-Fading Channel

A.M. Khan, Varun Jeoti and M. Azman Zakariya

Abstract In design of adaptive orthogonal frequency division multiplexing (OFDM) transmission, Signal-to-Noise Ratio (SNR) is well-known to be an informative performance measures. Channel impairments, specifically in wireless communication degrade the performance of accurate SNR measurement and said problem becomes more serious in low SNR regime. As a result it reduces the performance of adaptive transmission in OFDM systems. To alleviate this problem we develop a pilot based Pre FFT (time domain) Signal-to-Noise Ratio (SNR) estimator in the presence of frequency selective Rayleigh fading channel. A novel time domain SNR estimation technique is proposed that accurately measures the Signal-to-Noise Ratio in low SNR regime, where signal plus noise power is evaluated by using autocorrelation of received signal and signal power is indirectly estimated from pilot power using cross correlation. Simulation results show that proposed method has very less bias and very close to the actual SNR values.

18.1 Introduction

OFDM has been applied widely in wireless communication systems due to its high data rate transmission capability with high bandwidth efficiency and its robustness to multipath delay. OFDM introduces Cyclic Prefix (CP), which eliminates the Inter-Symbol-Interference (ISI) between OFDM symbols [1] and its high data rates are available without having to pay for extra bandwidth. With these advantages, OFDM is widely accepted in numerous wireless standards such as Digital Video

A.M. Khan (✉) · V. Jeoti · M. Azman Zakariya
Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS,
Tronoh, Malaysia
e-mail: abidmk77@yahoo.com

V. Jeoti
e-mail: varun_jeoti@petronas.com.my

Broadcasting (DVB), Digital Audio Broadcasting (DAB), Wireless Metropolitan Area Network (WMAN) and Wireless Local Area Network (WLAN).

In order to exploit all these advantages and optimize the performance of OFDM systems; channel state information (CSI) plays a vital role. A Signal-to-noise ratio (SNR) gives a comprehensive knowledge of CSI, which allows us to decide whether a transition to higher bit rates would be favorable or not. Similarly, SNR estimation is required for power control, adaptive modulation, coding and channel estimation. Therefore major input parameter for fourth generation adaptive system is a good SNR estimator.

Mainly there are two types of SNR estimation [2], which are as follows: (1) Data-aided (DA-SNR) estimator. (2) Non-data aided (NDA-SNR) estimator. *DA* is further divided into two types: (1) pilot-aided and (2) training sequence. In pilot-aided scheme, known information is transmitted together with data. While, in training sequence known information is transmitted over one or more OFDM symbols without data. In *NDA* estimator, no known data is transmitted and therefore SNR is estimated at the receiver blindly. In this work, an estimator of type DA-SNR (pilot-aided) is proposed because it works well in multipath fading channel.

One of the most significant current discussions in SNR estimation is Pre FFT (time domain) and Post FFT (frequency domain) estimators. Several studies have been produced SNR estimation in frequency domain, but little attention has been paid to time domain. Popular SNR estimators reported in literature are Boumard's, Ren's and Minimum Mean Square Error (MMSE) [3].

In [4], it was assumed that the channel conditions throughout observation are same. However, in highly frequency selective multipath channels, the assumption is not valid, and the performance is degraded greatly. Therefore this problem is overcome in [5], but its complexity is similar in terms of addition and multiplications. In addition, pilot aided MMSE [6] is investigated in frequency selective fading channel but it suffers when noise level increases and SNR fluctuates from its threshold value.

Recently, in [7] Pre FFT pilot aided SNR estimation is proposed where low MSE is achieved in flat fading channel. Therefore, to the best of our knowledge no work has been published for pilot aided Pre FFT SNR estimation over frequency selective Rayleigh fading channel that works well in low SNR environment.

A novel estimator presented here differs from previous methods in a way that it is pilot aided SNR estimation in time domain, by using autocorrelation of received signal. In addition, proposed method can accurately measure SNR which has very less bias and very close to the actual SNR values. It can also be seen that proposed SNR estimation performs better at low SNR.

This paper has been organized in the following way. In Sect. 18.2, overview of proposed Pre FFT SNR estimation technique is explained. Section 18.3 addressed the methodology used by proposed estimator. Section 18.4 described the channel model used in simulation. Results and analysis are discussed in Sect. 18.5. Conclusion of the work is presented in Sect. 18.6.

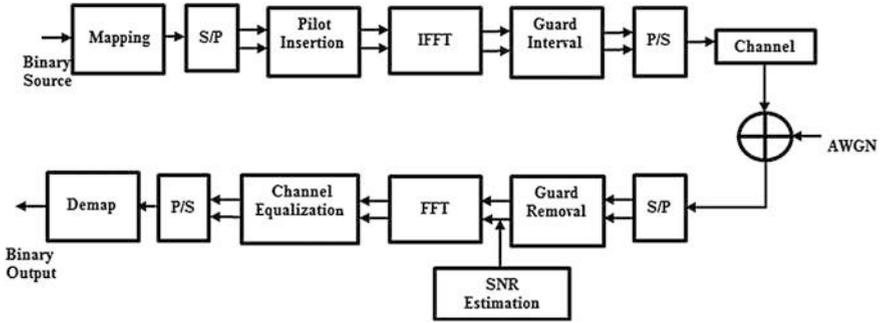


Fig. 18.1 Typical OFDM baseband transceiver [8]

18.2 Proposed Technique

Pre FFT OFDM system model is shown in Fig. 18.1. It begins with “signal mapping” block where binary information is mapped according to modulation. Then serial to parallel conversion takes place. After that, pilot sub-carriers $P(k)$ are inserted along with data sub-carriers $D(k)$, arrangement of pilots and data sub-carriers are shown in Fig. 18.2. Therefore $S(k)$ can be written as

$$S(k) = D(k) + P(k) \quad (18.1)$$

Then $IFFT$ is applied on $S(k)$ samples that transform frequency domain samples $S(k)$ into time domain $s(n)$, which can be shown as

$$s(n) = IFFT\{S(k)\} = \sum_{k=0}^{N-1} S(k) e^{\frac{j2\pi kn}{N_{FFT}}} \quad (18.2)$$

where N_{FFT} is no of FFT points and $s(n)$ can also be written as

$$s(n) = d(n) + p(n) \quad (18.3)$$

After passing through multi-path frequency selective fading channel, receiver input signal is first passed through serial to parallel converter and then guard band removal takes place. Therefore signal $r(n)$ is obtained, which can be shown as

$$r(n) = h(n) \otimes s(n) + w(n) \quad (18.4)$$

where $h(n)$ is channel impulse response and $w(n)$ is additive white Gaussian noise. The $h(n)$ can be expressed as

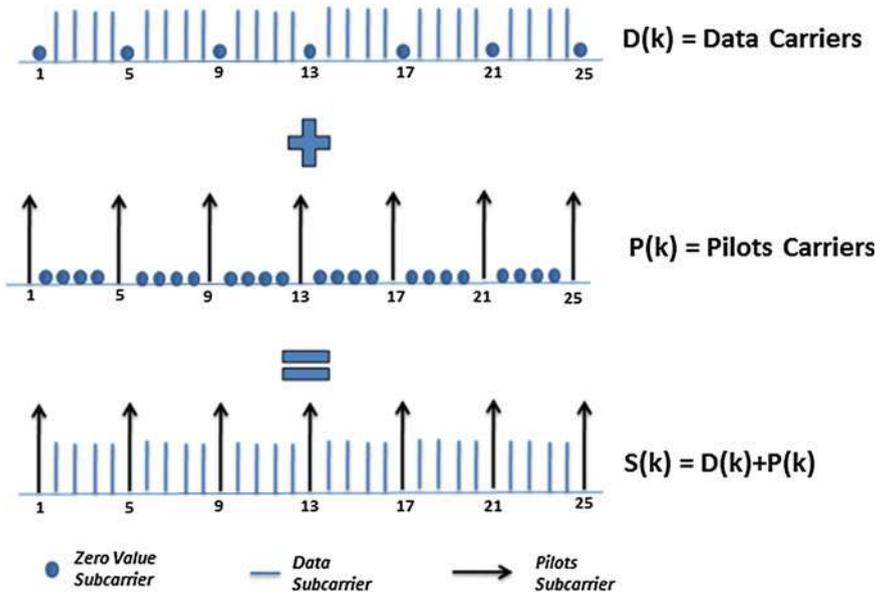


Fig. 18.2 Proposed pilots and data sub carriers arrangement

$$h(n) = [h_0 h_1 h_2 \dots h_{T-1}] \tag{18.5}$$

where $h_0, h_1 \dots h_T$ are the coefficient of channel impulse response and T is total no of channel taps. By substituting Eq. (18.3) in Eq. (18.4), $r(n)$ is given by

$$r(n) = h(n) \otimes [d(n) + p(n)] + w(n) \tag{18.6}$$

The SNR estimation technique is such that it attempts to estimate the signal power using a priori known pilots and determines the noise power from it using measurement of total received power. Towards that end, autocorrelation of the received signal is performed to measure the total received power which is the value at lag zero. And in order to estimate the signal power, pilot power is estimated with the help of cross correlation and indirectly the signal power is deduced from it. Hence, moving on, we calculate the autocorrelation of complex valued received signal, written as

$$R_{rr}(l) = \sum_{n=-\infty}^{n=+\infty} r(n)r^*(n+l) \tag{18.7}$$

where

$$r^*(n+l) = h^*(n+l) \otimes [d^*(n+l) + p^*(n+l)] + w^*(n+l) \quad (18.8)$$

By substituting the Eqs. (18.6) and (18.8) in Eq. (18.7), the autocorrelation $R_{rr}(l)$ is written as

$$R_{rr}(l) = \sum_{n=-\infty}^{n=\infty} \{h(n) \otimes [d(n) + p(n)] + w(n) \otimes [d^*(n+l) + p^*(n+l)] + w^*(n+l)\} \quad (18.9)$$

$$\begin{aligned} R_{rr}(l) = & \sum_{n=-\infty}^{n=\infty} \{R_{hh}(l) \otimes [R_{dd}(n) + R_{dp}(n)] + h(n) \otimes R_{dw}(n) \\ & + R_{hh}(l) \otimes [R_{pd}(n) + R_{pp}(n)] + h(n) \otimes R_{pw}(n) \\ & + h^*(n+l) \otimes [R_{wd}(n) + R_{wp}(n)] + R_{ww}(n)\} \end{aligned} \quad (18.10)$$

At zero lag when $l = 0$

$$R_{dp}(0) = R_{pd}(0) \cong 0 \quad (18.11)$$

$$R_{pw}(0) = R_{wp}(0) \cong 0 \quad (18.12)$$

$$R_{dw}(0) = R_{wd}(0) \cong 0 \quad (18.13)$$

By using Eqs. (18.11), (18.12) and (18.13), $R_{rr}(l)|_{l=0}$ is given as

$$R_{rr}(l) = R_{hh}(l) \otimes [R_{dd}(l) + R_{pp}(l)] + R_{ww}(l) \Big|_{l=0} \quad (18.14)$$

where Eq. (18.14) represents the received signal power as function of autocorrelation. It can also be in written form of K

$$R_{rr}(l) = R_{hh}(l) \otimes [K + 1]R_{pp}(l) + R_{ww}(l) \Big|_{l=0} \quad (18.15)$$

where

$$K = \frac{R_{dd}(l)}{R_{pp}(l)} \Big|_{l=0} \quad (18.16)$$

Therefore total Signal plus Noise Power P'_{S+N} can be written as

$$P'_{S+N} = R_{hh}(l) \otimes [K + 1]R_{pp}(l) + R_{ww}(l) \Big|_{l=0} \quad (18.17)$$

Similarly, the Signal Power P'_S can be expressed as

$$P'_S = R_{hh}(l) \otimes [K + 1]R_{pp}(l)|_{l=0} \quad (18.18)$$

To evaluate channel impulse response $h(n)$, the cross correlation of Eq. (18.6) is determined, with locally generated pilot signal. Therefore R_{rp} can be written as

$$R_{rp}(l) = h(l) \otimes [R_{dp}(l) + R_{pp}(l)] + R_{wp}(l)|_{n=l} \quad (18.19)$$

where R_{rp} is the cross correlation between received signal and pilot carriers. Similarly R_{dp} and R_{wp} are cross correlations of data and noise with pilots respectively. By substitution of Eqs. (18.12) and (18.13) in Eq. (18.19), $R_{rp}(n)$ is given by

$$R_{rp}(n) \cong h(n) \otimes R_{pp}(n)|_{l=n} \quad (18.20)$$

Equation (18.20) is the convolution of $h(n)$ and $R_{pp}(n)$, which can be shown in the form of impulses

$$h(n) = [h_0\delta(n) + h_1\delta(n - 1) \cdots h_T\delta(n - T)], \quad (18.21)$$

$$R_{pp}(n) = [\alpha_0\delta(n) + \alpha_1\delta(n \pm L) + \cdots \alpha_N\delta(n \pm NL)] \quad (18.22)$$

where $\alpha_0 \dots \alpha_N$ are coefficients of $R_{pp}(n)$. Therefore $R_{rp}(n)$ can be written as

$$\begin{aligned} R_{rp}(n) = & [\alpha_0 h_0 \delta(n) + \alpha_0 h_1 \delta(n - 1) + \cdots \alpha_0 h_T \delta(n - T)] + \cdots \\ & \alpha_1 h_0 \delta(n \pm L) + \alpha_1 h_1 \delta(n \pm L - 1) + \cdots \alpha_1 h_T \delta(n \pm L - T) + \cdots \\ & \alpha_N h_0 \delta(n \pm NL) + \alpha_N h_1 \delta(n \pm NL - 1) + \cdots \alpha_N h_T \delta(n \pm NL - T)|_{n=l} \end{aligned} \quad (18.23)$$

Equation (18.23) shows that the channel coefficients $h_0, h_1 \dots h_T$ are available at different location on both sides of zero lag ($\pm L$), but only zero lag position is used at ($n = 0$) due to its high energy. Therefore Eq. (18.23) can be written as

$$R_{rp}(0) = \frac{[\alpha_0 h_0 \delta(0) + \alpha_0 h_1 \delta(-1) + \cdots \alpha_0 h_T \delta(-T)]}{\alpha_0} \Big|_{n=0} \quad (18.24)$$

and it gives

$$h(n) = [h_0 \ h_1 \ h_2 \ \dots \ h_{T-1}] \quad (18.25)$$

Finally the estimated Signal to Noise Ratio (SNR') can be written as

$$SNR' = \frac{P'_S}{P'_N} \quad (18.26)$$

$$\overline{SNR}' = \frac{P'_S}{P'_{S+N} - P'_S} \quad (18.27)$$

$$SNR' = \frac{R_{hh}(l) \otimes [K + 1]R_{pp}(l)}{R_{rr}(l) - R_{hh}(l) \otimes [K + 1]R_{pp}(l)} \Big|_{l=0} \quad (18.28)$$

The Mean Square Error (MSE) is calculated by using following expression

$$MSE = \sum_{i=1}^I (SNR'_i(l) - SNR)^2 \Big|_{l=0} \quad (18.29)$$

where ($I = 1,000$) is the number of iteration used for simulation. $SNR'_i(l)$ and SNR are the estimated and actual SNR respectively.

18.3 Methodology

In proposed algorithm, signal-to-noise ratio (SNR) is estimated in time domain. An initial step of receiver is to remove the guard band from the receiver input signal $r(n)$. After the removal of guard band, signal $r(n)$ is provided to autocorrelation processor, where received signal is autocorrelated $R_{rr}(n)$. To estimate the Signal plus Noise Power P'_{S+N} and Signal Power P'_S , we required the autocorrelation of channel impulse response $R_{hh}(n)$. Therefore channel impulse response $h(n)$ is evaluated by taking cross correlation of $r(n)$ signal with time domain pilot signal $p(n)$. In final step, time domain SNR' is estimated by using Signal plus Noise Power P'_{S+N} and Signal Power P'_S . OFDM simulation parameters are mentioned in Table 18.1. It is also assumed that there is perfect synchronization between transmitter and receiver, so no Inter-Carrier Interference (ICI) and ISI are present in OFDM symbols. This is justified if the estimation error is sufficiently low at low SNRs.

18.4 Channel Model

In this work, six multi-paths fading channel model (COST 207 Typical Urban Reception TU6) for DVB-T applications are utilized in the simulations.

The taps of channel follow the Rayleigh statistics, whose parameters are shown in Table 18.2. The static case Impulse response of channel can be written as:

Table 18.1 OFDM simulation parameters

No of FFT points (N_{FFT})	2048 (2 K-Mode)
Length of cyclic prefix (N_G)	512
Total no of subcarriers (N_{OFDM})	2,560
Total no of symbol (N_{SYS})	1,000
Pilots spacing (N_F)	4
Numbers of pilots (N_P)	512
Pilot arrangement	Comb type
Pilot and data constellation	BPSK
Data per OFDM symbol	1,536
Bandwidth	8 MHz
Data constellation	BPSK

Table 18.2 COST 207 typical urban reception (TU6) for DVB-T application [9]

Tap number	Average power (dB)	Delay(us)
1	-3.0	0
2	0	0.2
3	-2.0	0.5
4	-6.0	1.6
5	-8.0	2.3
6	-10.0	5

$$h(n) = \begin{bmatrix} 0.5\delta(n) + \delta(n-3) + 0.63\delta(n-8) \\ + 0.25\delta(n-25) + 0.15\delta(n-36) + 0.1\delta(n-80) \end{bmatrix} \quad (18.30)$$

18.5 Results and Analysis

Figure 18.3 shows the cross correlation R_{dp} between pilots and data carriers. Pilots carriers are inserted into data in such a manner that it produces a value of zero when (lag = 0). It is also depicted that the value of $1.574e^{-18}$ is produced at zero lag, which is very small as compared to other time instant values and therefore it is neglected.

Figure 18.4 shows a cross correlation $R_{wp}(n)$ of noise signal with pilot carriers. In SNR estimation $R_{wp}(n)$ plays a very pivotal role because channel MSE is

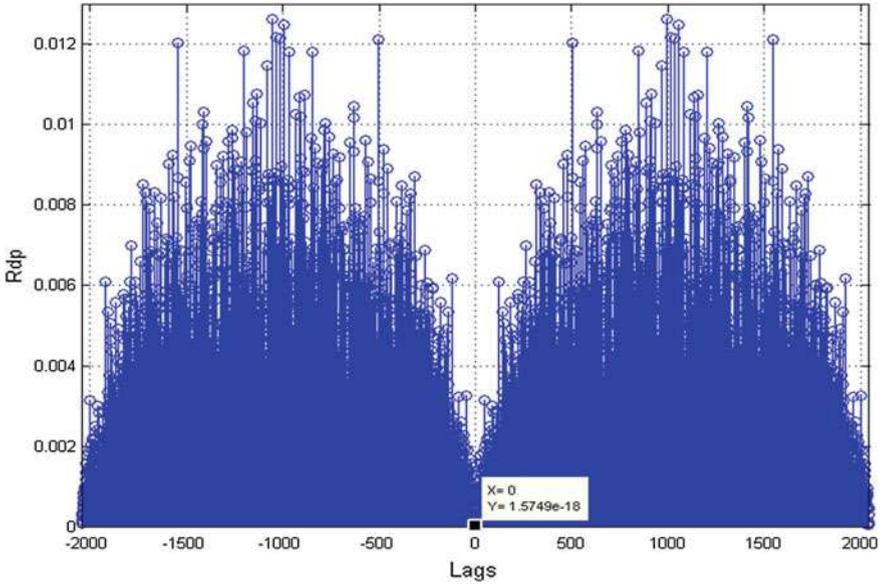


Fig. 18.3 (R_{dp}) Cross correlation between data and pilot sub-carriers at 10 dB SNR

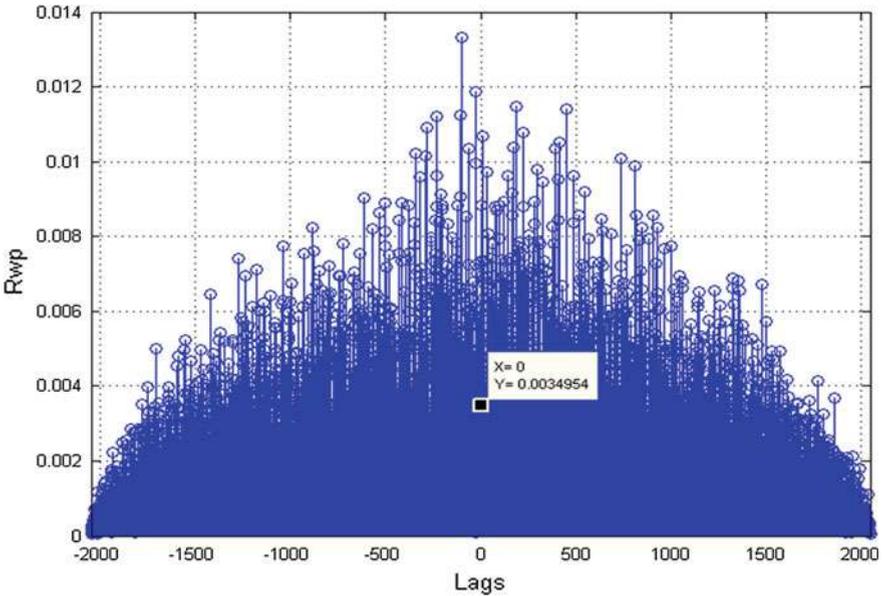


Fig. 18.4 (R_{wp}) Cross Correlation between noise and pilot sub-carriers at 10 dB SNR

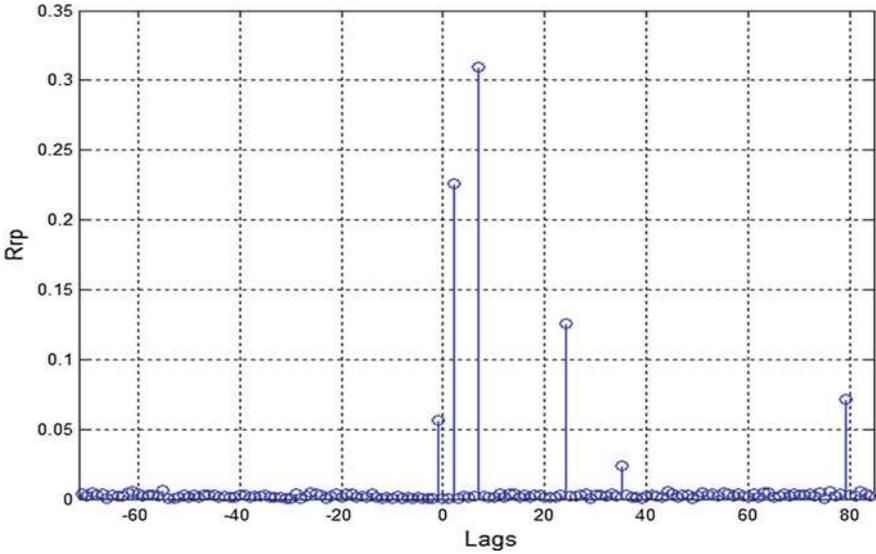


Fig. 18.5 (R_{rp}) Cross correlation between received signal and pilot sub-carriers

proportional to $R_{wp}(n)$. It is assumed that $R_{wp}(n)$ is expected to be relatively small due the arrangement of pilot carriers, which is also found at zero lag position.

An important feature of the proposed Pre FFT SNR estimation is investigated in Fig. 18.5 [i.e. $R_{rp}(n)$]. It has been shown that $R_{rp}(n)$ is the convolution of channel impulse response $h(n)$ and pilots autocorrelation $R_{pp}(n)$. Equation (18.20) also reflects this property of $R_{rp}(n)$. It is found that the channel co-efficient $h_0, h_1 \dots h_T$ are present at every L number of lags. For illustration purpose only the first L instants co-efficient values ($L = N_{fft}/4$) are shown in Fig. 18.5. Channel co-efficient is estimated at first L instants because at this location high energy and low MSE values are present.

In Fig. 18.6 the estimated SNR is plotted verses actual SNR for Rayleigh fading channel. To improve the estimation in low SNR regime, -8 to 20 dB range is used for simulation. It is depicted that the proposed SNR estimator has very small deviation with respect to the actual SNR values. It is also shown that Pre FFT SNR estimation performs better at low SNR.

In Fig. 18.7, MSE performance is analyzed over Rayleigh fading channel. The range of SNR used for simulation is $[-20$ to 20 dB]. Contrary to common sense it is observed that MSE of estimator is large at high SNR. This can be explained from the following observation. The SNR estimation is obtained primarily from signal power estimation. Noise power P'_N is estimated from signal power P'_S as shown in Eq. (18.27). So, at large SNR the signal power and total power are nearly same. Therefore, at large SNR, the noise power, estimated from it, is of the same order as the signal power estimation error, hence the SNR estimation error

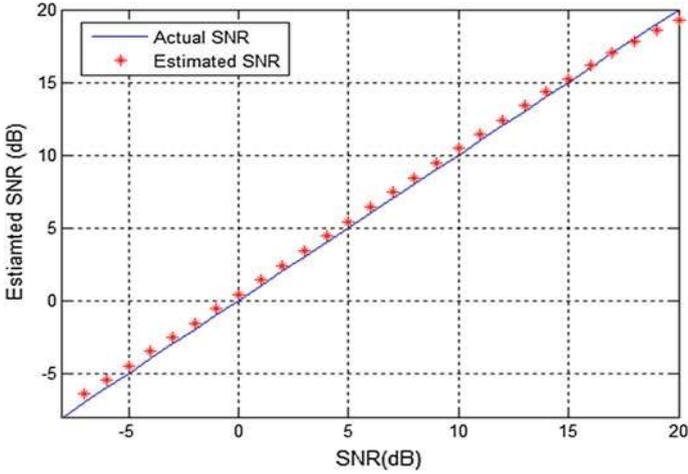


Fig. 18.6 Actual SNR versus estimated SNR for Rayleigh fading channel

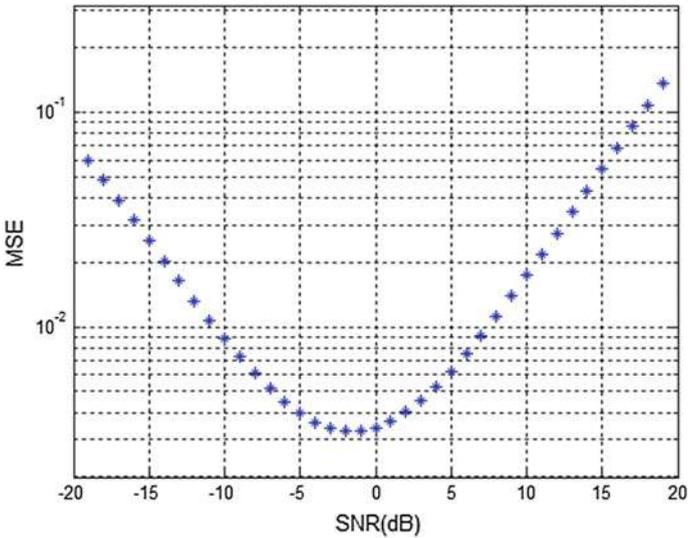


Fig. 18.7 MSE performance of proposed SNR estimated technique for Rayleigh fading channel

increases at large SNR. In order to be statistically accurate; mean is obtained over 1,000 samples. It is shown that proposed estimator performs better with multipath channel especially at low SNR environment.

18.6 Conclusion

This paper investigates a novel pilot-based (DA) Pre FFT SNR estimation, where autocorrelation of received signal is used to estimate the signal plus noise power. Proposed estimator performs well over frequency-selective Rayleigh fading channel at low SNR regime. The standardization DVB-T parameters are setup for OFDM simulation. From the simulation results it is found that a low MSE is achieved by exploiting proposed estimator. The amount deviation corresponds to each SNR values is measured in term of MSE and it is observed that estimated SNR has very less bias and very close to the actual SNR values.

References

1. Ozdemir, M.K., Arslan, H.: Channel estimation for wireless ofdm systems. *IEEE Commun. Surv. Tutor.* **9**, 18–48 (2007)
2. Wiesel, A., et al.: SNR estimation in time-varying fading channels. *IEEE Trans. Commun.* **54**, 841–848 (2006)
3. Ramchandrarao, S.R.M., Anil Kumar, P.T., Srinivas Rao, K.: Comparative study of SNR estimation techniques for Rayleigh and Rician channel. *Int. J. Sci. Eng. Technol. Res. (IJSETR)* **2**, 2091–2094 (2013)
4. Boumard, S.: Novel noise variance and SNR estimation algorithm for wireless MIMO OFDM systems. In: *Global Telecommunications Conference, GLOBECOM '03*. IEEE, vol. 3, pp. 1330–1334, 2003
5. Guangliang, R., et al.: SNR estimation algorithm based on the preamble for OFDM systems in frequency selective channels. *IEEE Trans. Commun.* **57**, 2230–2234 (2009)
6. Adachi, K.T.a.F.: SNR estimation for pilot-assisted frequency domain MMSE channel estimation. In: *Paper Presented at the IEEE VTS APWCS, Hokkaido, Japan, 2005*
7. Gafer, A.Y. et al.: Front-end signal to noise ratio estimation for DVBT fixed reception in flat-fading channel. In: *2012 4th International Conference on Intelligent and Advanced Systems (ICIAS)*, pp. 296–300, 2012
8. Coleri, S., et al.: Channel estimation techniques based on pilot arrangement in OFDM systems. *IEEE Trans. Broadcast.* **48**, 223–229 (2002)
9. Khan, A.M. et al.: Improved pilot-based LS and MMSE channel estimation using DFT for DVB-T OFDM systems. In: *2013 IEEE Symposium on Wireless Technology and Applications (ISWTA)*, pp. 120–124, 2013

Chapter 19

The Use of Convolutional Code for Narrowband Interference Suppression in OFDM-DVBT System

Aizura Abdullah, Muhammad Sobrun Jamil Jamal,
Khaizuran Abdullah, Ahmad Fadzil Ismail and Ani Liza Asnawi

Abstract The problem of mitigating narrowband interference (NBI) due to coexistence between Digital Video Broadcasting-Terrestrial (DVB-T) and International Mobile Telecommunication-Advanced (IMT-A) system is considered. It is assumed that a spectrum of IMT-A system between 790 and 862 MHz interfere the spectrum of the OFDM signal in DVB-T band. Two types of convolutional code (CC) which are non-systematic convolutional code (NSCC) and recursive systematic convolutional code (RSCC) are proposed to mitigate NBI. The performance of the two techniques is compared under additive white Gaussian noise (AWGN) channel. It is observed that NSCC has a better bit error rate (BER) performance than RSCC. The result showed good performance for low SNR (≤ 5 dB).

Keywords OFDM · Convolutional code · Narrowband interference · DVB-T

A. Abdullah (✉) · M.S.J. Jamal · K. Abdullah · A.F. Ismail · A.L. Asnawi
Department of Electrical and Computer Engineering, International Islamic University
(IIUM), Kuala Lumpur, Malaysia
e-mail: aizura.abdullah@gmail.com

M.S.J. Jamal
e-mail: ssobrun_88@yahoo.com

K. Abdullah
e-mail: khaizuran@iium.edu.my

A.F. Ismail
e-mail: af_ismail@iium.edu.my

A.L. Asnawi
e-mail: aniliza@iium.edu.my

19.1 Introduction

Orthogonal Frequency Division Multiplexing (OFDM) is a popular multiplexing scheme used for transmission of high data rates in various communication standards such as DVB-T, WLANs, WMANs and WiMAX [1, 2]. For certain standards such as WLANs and WMANs, an OFDM system has the ability to operate in unlicensed frequency bands. As a result, there is a possibility that they have to share the same frequency band with other communication systems such as cordless telephones, garage door openers and baby monitors. This leads to narrowband interference (NBI) in the systems [3]. Another example of systems sharing the same frequency band is WiMAX and UWB systems. According to [4], UWB system is required to modify its spectrum to avoid interference with WiMAX. In WRC-07 conference, ITU-R has allocated the 790–862 MHz frequency for IMT-A system. This also suggests that the DVB-T system which operates between 470 and 862 MHz have to share its upper frequency band with the IMT-A system [5].

There are several techniques proposed to mitigate NBI such as using orthogonal codes, frequency domain cancellation, receiver windowing and excision filtering [3, 4]. Although orthogonal codes is found to give better performance compared to error control code (ECC), this method does not comply with the current OFDM standard such as DVB-T and IEEE [3]. On the other hand, frequency domain cancellation technique is not suitable to be implemented in broadcasting because the channel and interference information from the receiver needed to be fed back to the transmitter periodically for update. A limitation of receiver windowing method is that it is suitable to be used together with frequency domain cancellation to reduce the effect of sinc shape side lobes from spreading to adjacent channel while excision filtering method provides less benefit with quadrature amplitude modulation (QAM) [4].

ECC is a suitable candidate to mitigate NBI as it is able to protect the data using a specific code. The data which is corrupted during transmission in the noisy channel will be recovered by the specific decoding method. Theoretically, ECC has the capability to lower the bit error rate of an uncoded system by a certain coding rate [6]. In brief, there are three types of ECC known as block, convolutional and modern codes. In this work, convolutional code (CC) is proposed as it is suitable to be used in broadcasting, deep space communication, digital speech and also for Gaussian channel condition [7, 8]. Two types of convolutional code proposed to mitigate NBI in DVB-T system are non-systematic convolutional code (NSCC) and recursive systematic convolutional code (RSCC).

Section 19.2 describes the OFDM system and NBI model used. Section 19.3 explains about the proposed ECC techniques. Section 19.4 provides the simulation results and discussion while Sect. 19.5 concludes this paper.

19.2 System Model

The OFDM simulation model of a DVB-T system under narrowband effect is as shown in Fig. 19.1. It is referred from a MATLAB simulation by [9]. The simulation model is modified by adding ECC as a narrowband mitigation technique and using different carrier frequency. ECC acts as encoder in the transmitter and decode the signal back for recovery in the receiver.

At baseband, ECC is applied at the stream of binary data $k = \{k_1 k_2 k_3 \dots k_n\}$. Then, the coded binary data $c = \{c_1 c_2 c_3 c_4 \dots c_n\}$, is converted into symbols to be modulated with M number of Quadrature Amplitude Modulation (QAM). Each serial modulated symbols $S = \{S_1 S_2 S_3 S_4 \dots S_n\}$, are mapped into N number of parallel subcarriers. The modulated symbols $X(k)$, appeared as a complex signal in frequency domain:

$$X(k) = R(k) + jI(k) \quad (19.1)$$

The modulated symbols are passed to inverse fast Fourier transform (IFFT) processing block to create a time domain OFDM signal for transmission. $2N$ -IFFT processing is used to center the subcarriers and processed the discrete signal $x(n)$,

$$x(n) = 1/N \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N} \quad (19.2)$$

where $n = 0, 1, \dots, N - 1$, $k = 0, 1, 2, 3, \dots, N - 1$; N being the number of subcarriers.

An OFDM symbol of N subcarriers is to be transmitted in an OFDM symbol period duration. The next processing block is to sample the OFDM discrete signal $x(n)$, within the OFDM symbol period duration. It undergoes filtration process in digital-to-analog (DAC) converter to obtain continuous time domain signal $x(t)$. Finally, the signal $x(t)$, is modulated with its RF transmit signal carrier f_c , and ready for transmission. The receiver system is the reverse process of the transmission system. After demodulation, a decoder recovered the data based on ECC scheme applied. After the data is recovered, it is compared with the original data for bit error rate (BER) calculation. From Fig. 19.1, the received signal and the effect of channel can be written as follows:

$$r(t) = x(t) + n(t) + i(t) \quad (19.3)$$

where $r(t)$ is the received signal consist of transmitted signal $x(t)$, Gaussian noise (AWGN) $n(t)$ and narrowband interference $i(t)$.

Figure 19.2 shows the theoretical model of OFDM-DVB-T band adopted from [2] which is used to represent general scenario in this work. For all the channels that are used for transmission, there are 49 channels in the DVB-T frequency band. From Eq. (19.3), $i(t)$ has a frequency range f_i , between 790 and 862 MHz

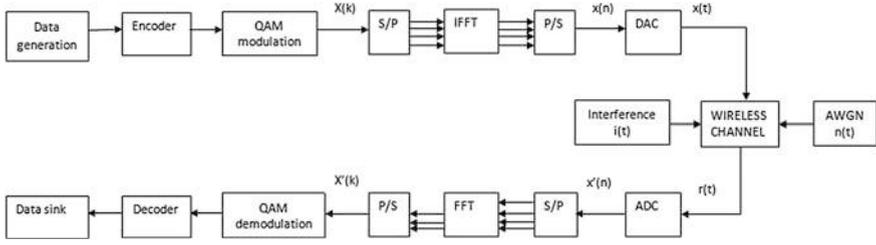


Fig. 19.1 OFDM simulation model with DVB-T parameters

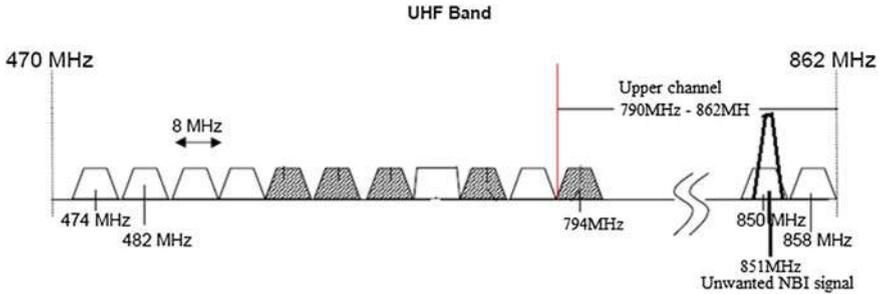


Fig. 19.2 Theoretical model of OFDM-DVB-T band adopted from [2]

interfered with the upper channel in DVB-T band. In this work, the 48th channel in the DVB-T band which has carrier frequency of 850 MHz is chosen as simulation parameter with the unwanted NBI signal of frequency 851 MHz.

The narrowband interference (NBI) signal is modeled as sinusoidal signal $i(t)$,

$$i(t) = I \cos(2\pi f_i t + \theta), \quad 0 < \theta < 2\pi \tag{19.4}$$

where I is the amplitude of the NBI signal and θ is the phase angle. The value of θ considered in this work is $\theta = \pi$. Substituting (19.4) into (19.1), the received signal $r(t)$, is derived as:

$$r(t) = A \cos(2\pi f_c t) + n(t) + I \cos(2\pi f_i t + \theta) \tag{19.5}$$

where A is the amplitude and f_c is the carrier frequency of the OFDM signal. The performance of an OFDM system is degraded when a strong NBI signal f_i , with carrier frequency close to the OFDM signal's carrier frequency f_c , overlapped,

$$f_i = f_c + \Delta f \tag{19.6}$$

and that the amplitude of the NBI signal is greater than the amplitude of the OFDM signal ($I > A$). Further details can be found in simulation part 4.

19.3 The Proposed Technique

19.3.1 Non-systematic Convolutional Code

Figure 19.3 shows the block diagram of 1/2 rate NSCC encoder which consist of m number of memory registers. It is used to store previous binary input data. If a binary data k , enters an encoder, it produces n coded bits at the output with code rate $R = k/n$. The code representation is written as (n, k, m) . The design of NSCC can be found in literatures such as [7, 8, 10] with different generator polynomials. In this work, the generator polynomials used are different compared to the ones used in example [7, 8, 10] because based on simulation result, it gave BER performance curve closest to the OFDM system without NBI effect. The generator polynomials used for 1/2 rate NSCC are $g_1 = [111]$ and $g_2 = [011]$.

An input bit 1 which entered the encoder will be modulo-2 added with stored values in memory register to generate the coded bits $u_1 u_2$. The generator polynomials determine which stored values in memory register needed to be modulo-2 added with the input bit. Assuming the initial state of memory register is 000, the output is shown in Table 19.1 below. The input bit is then moved into shift register m_1 with all the bits in the memory register shifted. The oldest stored bit in m_3 is disappeared. The next input bit entered is modulo-2 added with stored values in memory register which is 100 and the process is repeated.

The system is then extended to 1/3 rate by addition of another generator polynomial, $g_3 = [101]$ as shown in Fig. 19.4. In the case of 1/3 rate encoder, one bit entered the encoder produced three output bits.

NSCC gave better performance when Viterbi decoder is used [8]. The possible path that the encoder has undergone is represented in Trellis diagram as shown in Fig. 19.5. All the memory register's possible state is written at the first column. The input bit (in bracket) is written next to the matched output bits referred to encoder's truth table. Usually, the process will start at state 000. The branch metric is calculated by comparing the number of bit agreement with the coded bits. The process is repeated for all the coded bits. The path which has the highest number of branch metric is chosen as 'survivor path' and the decoded bits are determined.

19.3.2 Recursive Systematic Convolutional Code

RSCC, also known as turbo code is developed from NSCC. Compared to NSCC, RSCC is formed by concatenating in parallel two RSCCs separated by an interleaver. It is a systematic coding because one of the message bit itself is called systematic bit and the other two are the parity bits generated by the two RSCC

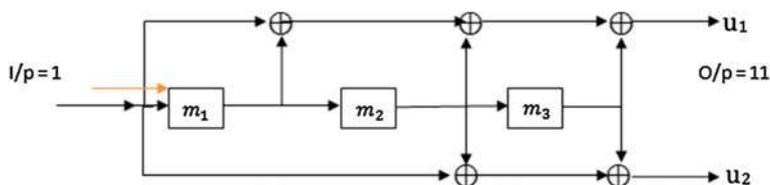


Fig. 19.3 1/2 rate convolutional encoder (2, 1, 3)

Table 19.1 Example of truth table for 1/2 rate convolutional encoder (2, 1, 3)

Start state	Input	End state	U1	U2	Output
000	0	000	$(0 \oplus 0 \oplus 0 \oplus 0) = 0$	$(0 \oplus 0 \oplus 0) = 0$	00
000	1	100	$(1 \oplus 0 \oplus 0 \oplus 0) = 1$	$(1 \oplus 0 \oplus 0) = 1$	11
100	0	010	$(0 \oplus 1 \oplus 0 \oplus 0) = 1$	$(0 \oplus 0 \oplus 0) = 0$	10
100	1	110	$(1 \oplus 1 \oplus 0 \oplus 0) = 0$	$(1 \oplus 0 \oplus 0) = 1$	01

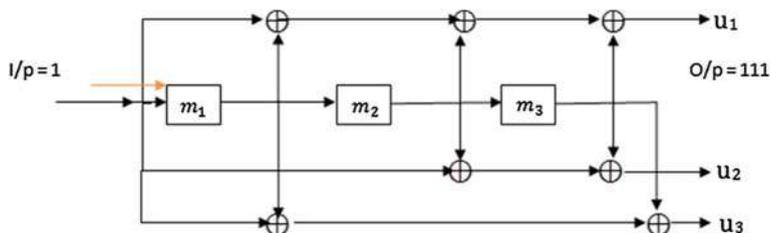
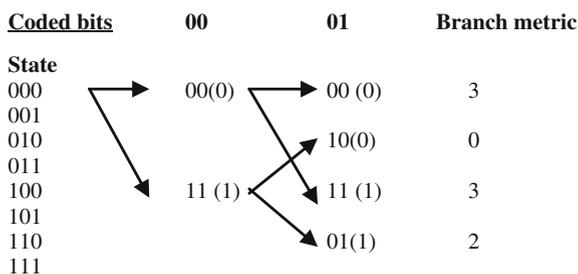


Fig. 19.4 1/3 rate convolutional encoder (3, 1, 3)

Fig. 19.5 Example of Trellis diagram for 1/2 rate convolutional code



encoders. Since the aim of this research is to determine which type of convolutional code performs better in mitigating NBI, the design of RSCC used in this paper is adopted from [11]. Its coding rate R , is 1/3 with generator polynomial $g_1 = [101]$, $g_2 = [111]$, and $g_3 = [101]$. Iterative decoding is used to decode the message.

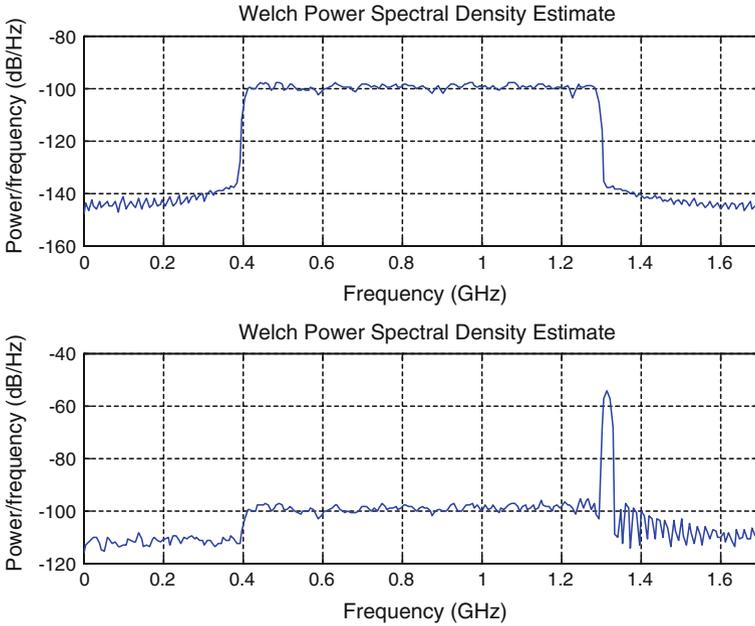


Fig. 19.6 An OFDM signal spectrum appears between 0.4 and 1.25 MHz along the x-axis for 850 MHz carrier frequency, showing OFDM transmitted signal spectrum (*top*) and OFDM received signal spectrum with Gaussian noise and the presence of 851 MHz NBI signal (*bottom*)

19.4 Simulation Results

The performance of convolutional codes in OFDM system under NBI effect is simulated using MATLAB based on the DVB-T parameters for 2 k mode [9]. Figure 19.6 shows the illustration of the OFDM signal in the presence of Gaussian effect and unwanted NBI signal.

The BER performance curve of the OFDM system with and without the presence of NBI is presented in Fig. 19.7. The initial OFDM curve (without interference) has average signal power value of -10 dB. The value obtained is calculated based on simulation according to [1]. The NBI signal is added to the initial OFDM system and simulated in two conditions i.e. with sinusoidal amplitude $I = 10$ V and $I = 20$ V. Referring to [6], the NBI signal power P_i , for the case of Eq. (19.4) is,

$$P_i = I^2/2 \tag{19.7}$$

where I is the amplitude of the narrowband sinusoidal signal.

The NBI signal power for sinusoidal amplitude $I = 10$ V and $I = 20$ V are approximately 17 and 23 dB respectively. For NBI signal power less than 17 dB, it

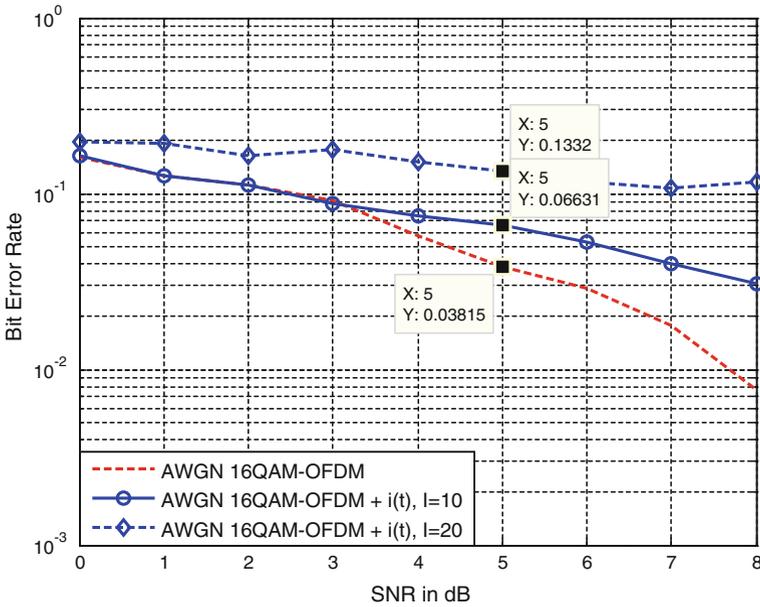


Fig. 19.7 BER performance of OFDM system with NBI effects (blue) and initial system without interference (red). The simulation also shows comparison of the system performance when the unwanted sinusoidal amplitude is varied ($I = 10$ V and $I = 20$ V), having different signal power

has a small effect on the OFDM system. In Fig. 19.7, at SNR = 5 dB, the bit error rate is 0.06631 for system under NBI signal power of 17 dB and 0.1332 for NBI signal power of 23 dB compared to 0.03815 for initial system. The difference of about 0.02816 between system with 17 dB NBI power and initial system is due to the fact that the system performance is affected by the NBI signal. When the NBI’s signal amplitude is increased to $I = 20$, the difference with initial system is 0.09505 which implies further degradation of the system performance due to the increased in NBI signal power.

Figure 19.8 shows the performance of convolutional coded (CC) OFDM with the presence of 17 dB NBI signal power. For the case of SNR = 5 dB, the bit error rate is about 0.04577 for 1/2 rate NSCC which is close to the initial system. As the SNR increased, the performance of 1/2 rate NSCC did not follow the curve of initial OFDM system. RSCC obtained BER of 0.09507 at SNR = 5 dB which showed worst performance compared to the OFDM system with NBI. On the other hand, 1/3 rate NSCC gave faulty result because it achieved BER of 0.01819 at SNR = 5 dB which is lower than the BER of initial system.

The convolutional coded OFDM under 23 dB NBI signal power effect is shown in Fig. 19.9. Based on the figure, 1/3 rate NSCC outperformed 1/2 rate NSCC and RSCC. At SNR = 5 dB, an error of 0.04401 is obtained by 1/3 rate NSCC which

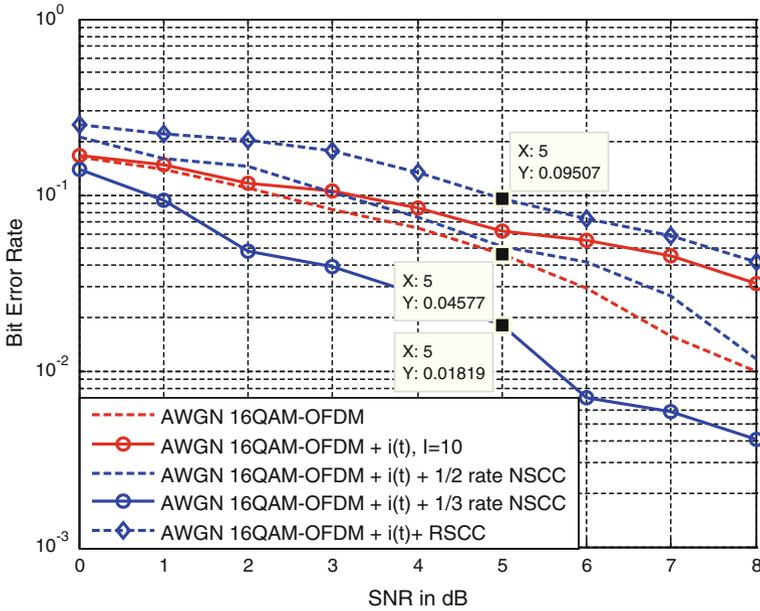


Fig. 19.8 Performance comparison of convolutional coded OFDM (blue) with the presence of 17 dB NBI signal power with uncoded system (red)

is similar to the initial system. As the SNR increased, the performance of 1/3 rate NSCC also did not follow the curve of the initial system. 1/2 rate NSCC showed less performance as it followed the curve of OFDM system under NBI effect whereas RSCC showed the worst performance with error rate 0.2019 at SNR = 5 dB.

Based on observations on Figs. 19.8 and 19.9, at low SNRs (≤ 5 dB), the performance curve of NSCC followed the curves of initial system (without interference) compared to high SNR (>5 dB). This implies that the performance of NSCC with Viterbi decoder is different at low and high SNR. According to [7], for a convolutional code, the error correction and detection capability t , is

$$t = (d_{\text{free}} - 1)/2 \tag{19.8}$$

where d_{free} is free distance which is the smallest Hamming distance between all possible code sequences of the code. At high SNR, the performance is limited by the capability of Viterbi decoder to correct more than t number of errors in n bits. Power, bandwidth constraint and nature of noise in the channel can also affect the performance of the coding scheme. Based on the simulation results, the proposed

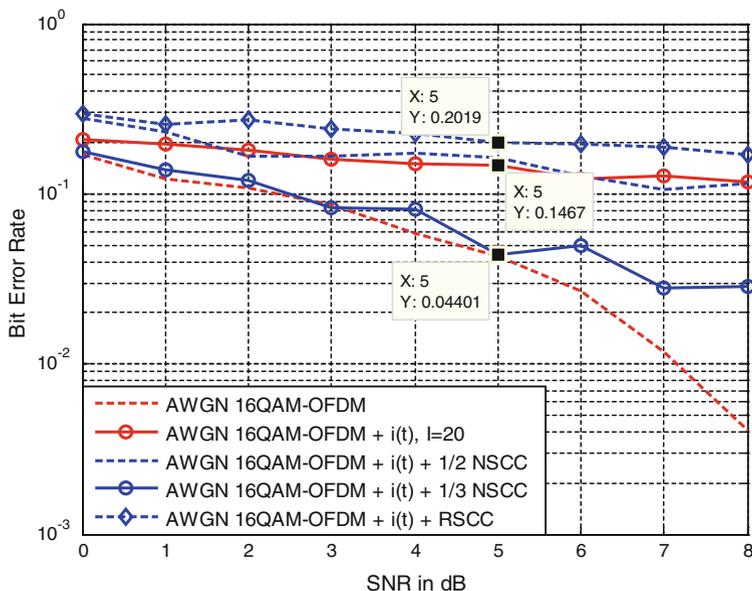


Fig. 19.9 Performance comparison of convolutional coded OFDM (blue) with the presence of 23 dB NBI signal power with uncoded system (red)

NSCC has good performance for low SNR (≤ 5 dB) to mitigate NBI compared to RSCC. The performance of this code also showed considerable result at low SNR when compared with time windowing method for NBI mitigation under DVB-T system [9].

19.5 Conclusion

In this paper, a narrowband mitigation technique is proposed using convolutional code. NSCC and RSCC are presented as two different ECC techniques suitable for NBI mitigation for DVB-T transmission under Gaussian channel. The interference is assumed coming from IMT-A services affected the upper channel of the DVB-T band. The simulation showed that 1/2 rate NSCC can mitigate the 17 dB NBI signal power while 1/3 rate NSCC suited for 23 dB NBI signal power at low SNR (≤ 5 dB). The performance result for RSCC showed that it is not effective in mitigating the NBI for this work.

Acknowledgments The authors would like to thank the Malaysian government, Ministry of Higher Education (MOHE) for sponsoring this work under FRGS grant no. FRGS13-026-0267.

References

1. Cho, Y.S., Kim, J., Yang, W.Y., Kang, C.-G.: MIMO-OFDM wireless communications with MATLAB. Wiley, Singapore (2010)
2. E. B. Union: Etsi en 301 701 (2000)
3. Coulson, A.J.: Bit error rate performance of OFDM in narrowband interference with excision filtering. *IEEE Trans. Wirel. Commun.* **5**(9), 2484–2492 (2006)
4. Batra, A., Zeidler, J.R.: Narrowband interference mitigation in OFDM systems. In: *Military Communications Conference MILCOM 2008*, pp. 1–7 (2008)
5. Shamsan, Z.A., Rahman, T.A., Al-hetar, A.M.: Point-point fixed wireless and broadcasting services coexistence with IMT-advanced system. *Prog. Electromagn. Res.* **122**, 537–555 (2012)
6. Lathi, B.P.: *Modern digital and analog communication systems*. Oxford University Press, New York (1998)
7. Jiang, Y.: *A practical guide to error control coding using MATLAB*. Artech House, Boston (2010)
8. Sweeney, P.: *Error control coding from theory to practice*. Wiley, England (2004)
9. Abdullah, K.: *Interference mitigation techniques for wireless OFDM*. Ph.D. dissertation, RMIT, Melbourne (2009)
10. Arshad, N., Basit, A.: Implementation and analysis of convolutional codes. *Int. J. Multidiscip. Sci. Eng.* **3**(8), 9–12 (2012)
11. Arshad, N., Jamal, M.A.: Implementation and analysis of turbo codes using MATLAB. *J. Expert Syst. (JES)* **2**(1), 115–118 (2013)

Chapter 20

Two-Elements Crescent Shaped Printed Antenna for Wireless Applications

Wan Noor Najwa Wan Marzudi, Zuhairiah Zainal Abidin,
Ma Yue and Raed A. Abd-Alhameed

Abstract This study presents an investigation of the mutual coupling between two printed elements antenna for a multiple-input-multiple-output (MIMO) antenna performance. It consists of two crescent shaped radiators placed symmetrically, and a neutralization line is applied to improve the mutual coupling. Theoretical and experimental characteristics are presented and compared. The antenna yields an achieved impedance bandwidth of 18.67 % (over 2.04–2.46 GHz) with a reflection coefficient < -10 dB and mutual coupling minimization of < -20 dB in addition to a reasonable and stable radiation pattern and envelope correlation.

Keywords Multiple-input-multiple-output (MIMO) · Neutralization line · Mutual coupling

W.N.N.W. Marzudi (✉) · Z.Z. Abidin
Research Center of Applied Electromagnetic, Universiti Tun Hussein Onn Malaysia, Johor,
Malaysia
e-mail: wannooranjwa@gmail.com

Z.Z. Abidin
e-mail: zuhairia@uthm.edu.my

M. Yue
National Astronomical Observatories (NAOC), Chinese Academy of Sciences,
Beijing, China

R.A. Abd-Alhameed
Mobile and Satellite Communication Research Centre, University of Bradford,
Bradford, UK

20.1 Introduction

The potential for MIMO antenna systems to improve reliability and enhance channel capacity in wireless mobile communications has generated great interest [1]. A major consideration in MIMO antenna design is to reduce correlation between the multiple elements, and in particular the mutual-coupling electromagnetic interactions that exist between multiple elements are significant, because at the receiver end this effect could largely determine the performance of the system. Lower mutual coupling can result in higher antenna efficiencies and lower correlation coefficients [2]. The effect of mutual coupling on antenna diversity performance of the MIMO antenna array has been reported in [3–10]. In [3], a shorting strip and isolation stub was used to reduce mutual coupling for portable wireless devices with isolation values lower than -25 dB at 2.45 GHz. Authors in [4] introduced I-shaped conductor in a modified ground plane that reached -14 dB of mutual coupling across 1.6–2.8 GHz. Inserting stubs is one of the method used to enhanced the isolation and minimize mutual coupling between elements [5]. In addition, by inserting neutralization line technique between antenna elements is was also promising method [6, 7]. Low mutual coupling can also be achieved through defected ground structure [8]. Other methods to reduce mutual coupling and enhanced isolation of the MIMO antenna, such as protrude branch and T-slot etched on ground plane [9] and inserting slits into ground plane [10].

In this paper, two-element crescent shaped MIMO antenna presented for the purpose of wireless applications. The MIMO antenna consists of two crescent shaped radiators placed symmetrically with respect to ground plane with neutralization line (NL) connected in between of the two antennas. The total dimensions of this antenna are $100 \times 45 \times 1.6$ mm³. Both simulated and measured result of the fabricated prototype details reported and discussed.

20.2 Antenna Design Concept

The proposed antenna geometry is illustrated in Fig. 20.1. The antenna system comprises with two crescent shaped radiators that is similar to that in [11] deployed on an economically FR-4 substrate with relative permittivity of 4.4 and a thickness of 1.6 mm operating at 2.4 GHz. The radiators are separated by 0.147λ (18.36 mm) for the minimization of mutual coupling. While, 83×45 mm² ground plane placed on the other side of the substrate as shown in Fig. 20.1b. The overall dimensions of the proposed antenna are $100 \times 45 \times 1.6$ mm³ which is suitable for wireless application such as a network card or mobile device.

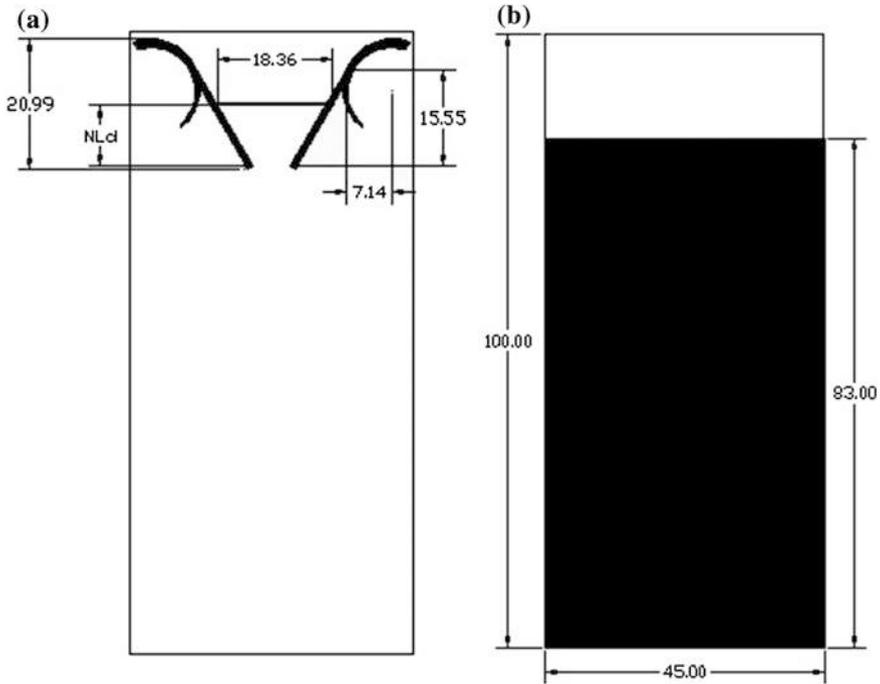


Fig. 20.1 Geometry of the proposed antenna (in mm) **a** *Top view*, and **b** *Bottom view*

To minimize the mutual coupling and enhanced the isolation of the antenna, neutralization line was inserted in between of the radiating element. In general, the neutralization line transfer the signal from the first antenna element and the second antenna element receive the signal in order to cancel out existing coupling between two elements.

20.3 Parametric Study

To clarify the effectiveness of the neutralization line (NL) of the proposed antenna, the parametric study of the location of the neutralization line, NL_d was carried out with the width of the NL is kept at 0.5 mm. From Fig. 20.2, it can be observed that the optimal distance of NL_d is at 9.79 mm which gives the lowest mutual coupling at 2.4 GHz.

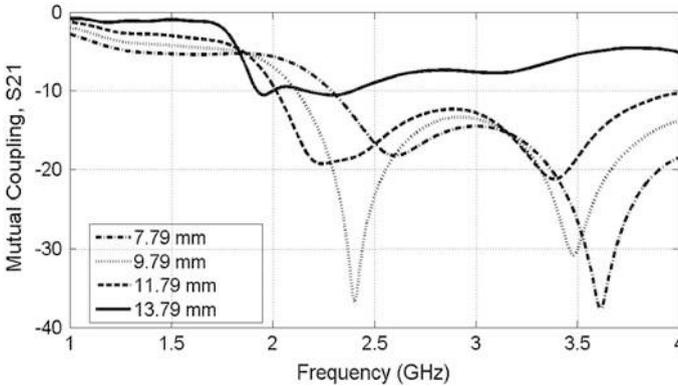


Fig. 20.2 Simulated transmission coefficient, S_{21} of the various distance location of neutralization line

20.4 Simulated and Measured Performance

Figures 20.3 and 20.4 show the simulated and measured S-parameters output for the proposed antenna with and without neutralization line, respectively. As can be observed, the measured return loss, $|S_{11}|$ and mutual coupling, $|S_{21}|$ for both figures (Figs. 20.3 and 20.4) are reasonably good agreement with the simulated results. The resonance frequency is slightly shifted between the simulated and measured results and this is probably due to the discrepancy of SMA connector and fabrication tolerance. It is apparently seen that by implementing the neutralization lines, the mutual coupling, $|S_{21}|$ of the proposed antenna can be improved.

To validate the simulated results, the physical prototypes of the proposed antenna with and without the neutralization line were fabricated and tested, as shown in Fig. 20.5. The S-parameters of the antenna were measured by Vector Network Analyser 8722ET (VNA). The measured return loss $|S_{11}|$ and mutual coupling $|S_{21}|$ are plotted in Fig. 20.6. As can be seen, when the neutralization line was inserted, the mutual coupling has been reduced around 7.14 dB (from -14.63 to -21.77 dB) with an impedance bandwidth of 18.67 % (over 2.04–2.46 GHz). The bandwidth achieved fully covered the wireless application such as network card at 2.4 GHz.

To evaluate the capabilities of MIMO/diversity antenna, the envelope correlation coefficient (ECC) is an important criterion to be presented. Basically, envelope correlation can be computed by using S-parameters or radiation pattern of the antenna. The envelope correlation of the MIMO antenna system can be expressed by using the following expression [12]:

$$\rho_e = \frac{|S_{11}^* S_{12} + S_{21}^* S_{22}|^2}{(1 - |S_{11}|^2 - |S_{21}|^2)(1 - |S_{22}|^2 - |S_{12}|^2)} \quad (20.1)$$

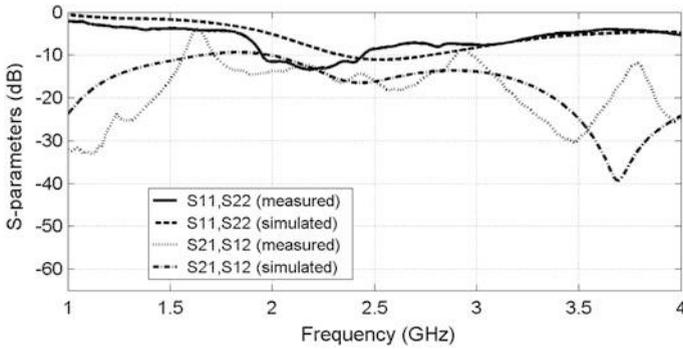


Fig. 20.3 Comparative plot of S-parameters output for simulated and measured results for the proposed antenna without neutralization line

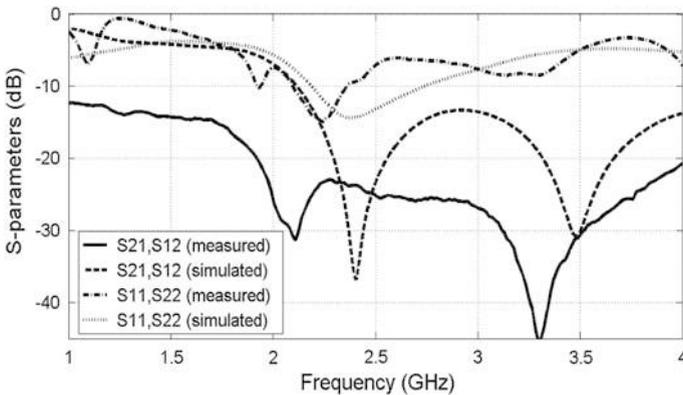


Fig. 20.4 Comparative plot of S-parameters output for simulated and measured results for the proposed antenna with neutralization line

The simulated envelope correlation coefficient of the proposed antenna with and without neutralization showed in Fig. 20.7. An improvement of the ECC can be seen after the neutralization line was inserted and it fulfils the characteristic of diversity $p_e < 0.5$ [13]. Therefore, the proposed antenna is suitable candidate for MIMO application.

The simulated and measured radiation patterns of the proposed antenna in the X–Z plane (E-plane) and Y–Z plane (H-plane) with port 1 excited while port 2 terminated with 50Ω load plotted in Figs. 20.8 and 20.9, respectively. The antenna shows a stable omnidirectional pattern in the E-plane and H-plane over the operating frequency of 2.4 GHz.

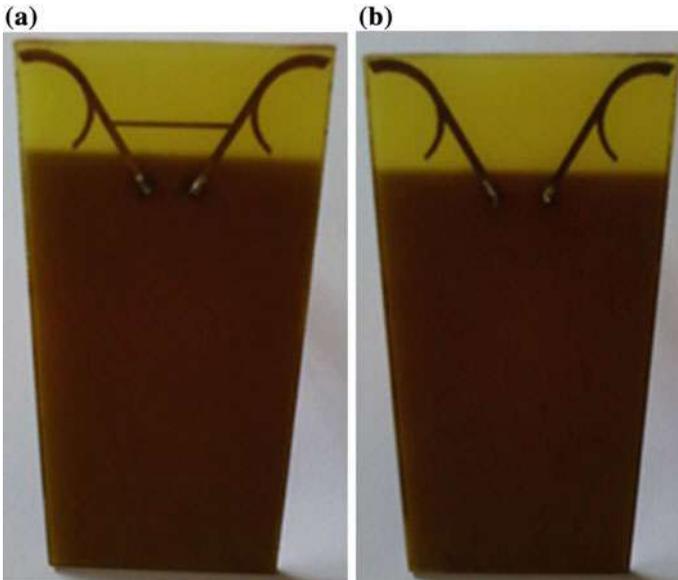


Fig. 20.5 Practical prototype of the proposed antenna **a** with neutralization line **b** without the neutralization line

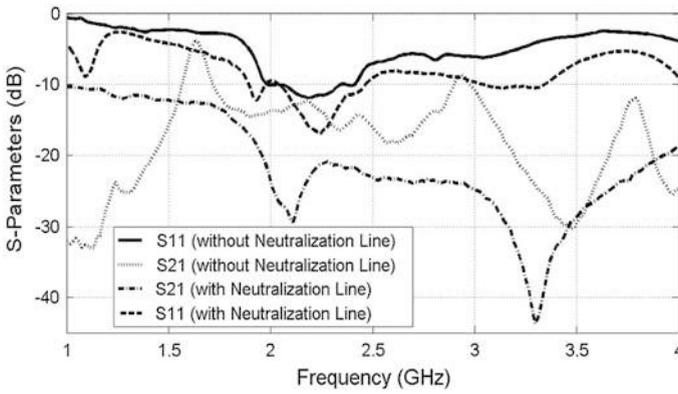


Fig. 20.6 Measured S-parameters of the proposed antenna with and without the neutralization line

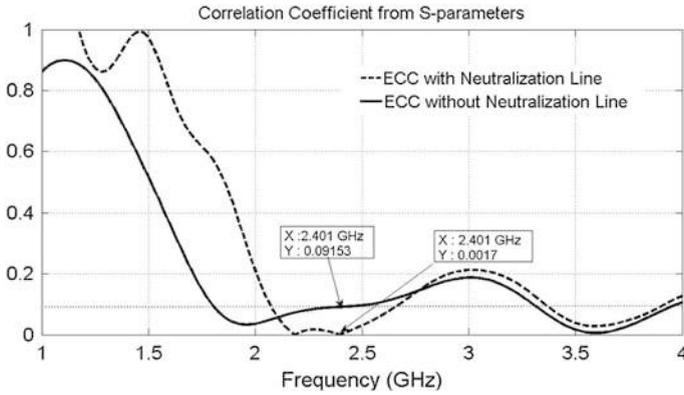


Fig. 20.7 Simulated envelope correlation coefficient for the proposed antenna with and without neutralization line

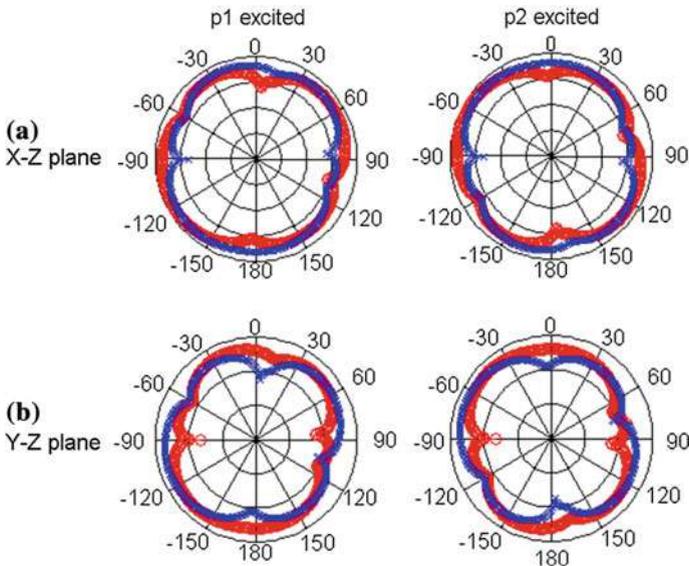


Fig. 20.8 Simulated radiation patterns for the proposed antenna for two planes at 2.4 GHz; **a** X-Z plane. “xxxx” (blue) simulated co-polarization, “oooo” (red) simulated cross-polarization. **b** Y-Z plane. “xxxx” (blue) simulated cross-polarization, “oooo” (red) simulated co-polarization port 1 (left) excited and port 2 (right) terminated in 50 Ω (colour figure online)

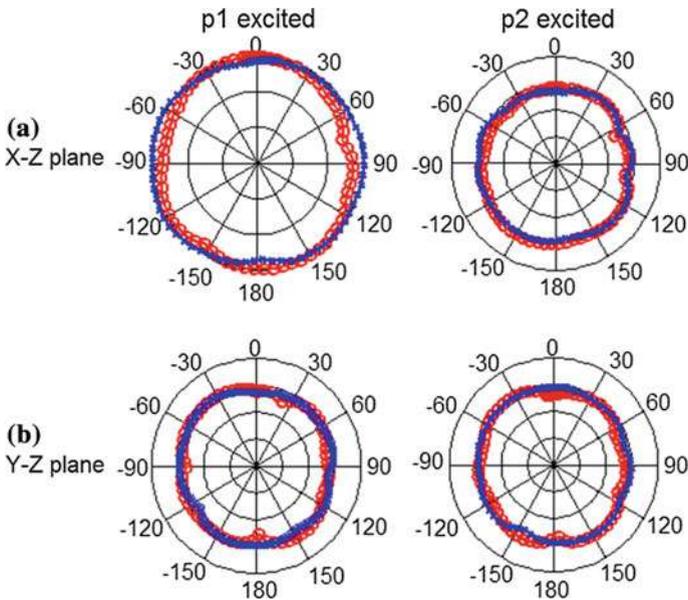


Fig. 20.9 Measured radiation patterns for the proposed antenna for two planes at 2.4 GHz; **a** X-Z plane. “xxxx” (blue) simulated co-polarization, “oooo” (red) simulated cross-polarization. **b** Y-Z plane. “xxxx” (blue) simulated cross-polarization, “oooo” (red) simulated co-polarization port 1 (left) excited and port 2 (right) terminated in 50Ω (colour figure online)

20.5 Conclusion

A two elements crescent shaped printed MIMO antenna for covering 2.4 GHz wireless applications is presented. Neutralization line is applied to meet the requirement of MIMO in term of low mutual coupling parameter. Simulated and measured results show that the antenna achieves an impedance bandwidth of 18.67 % (over 2.04–2.46 GHz) with a reflection coefficient < -10 dB and mutual coupling minimization of < -20 dB which is suitable for wireless applications.

Acknowledgement The authors of this paper wish to acknowledge the funding of this project by Universiti Tun Hussein Onn Malaysia under short term grant Vot 0992.

References

1. Foschini, G.J., Gans, M.J.: On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Pers. Commun.* **6**, 311–335 (1998)
2. Thaysen, J., Jakobsen, K.B.: Design considerations for low antenna correlation and mutual coupling reduction in multi antenna terminals. *Eur. Trans. Telecommun.* **18**, 319–326 (2007)
3. Ling, X., Li, R.: A novel dual-band MIMO antenna array with low mutual coupling for portable wireless devices. *IEEE Antennas Wirel. Propag. Lett.* **10**, 1039–1042 (2011)

4. See, C., Abd-Alhameed, R., McEwan, N., Jones, S., Asif, R., Excell, P.: Design of a printed MIMO/diversity monopole antenna for future generation handheld devices. *Int. J. RF Microw. Comput. Aided Eng.* **24**, 348–359 (2013)
5. Liu, L., Cheung, S., Yuk, T., Wu, D.: A compact ultrawideband MIMO antenna. In: 2013 7th European Conference on Antennas and Propagation (EuCAP), pp. 2108–2111 (2013)
6. See, C.H., Abd-Alhameed, R.A., Abidin, Z.Z., McEwan, N.J., Excell, P.S.: Wideband printed MIMO/diversity monopole antenna for WiFi/WiMAX applications. *IEEE Trans. Antennas Propag.* **60**, 2028–2035 (2012)
7. Wang, Y., Du, Z.: A wideband printed dual-antenna system with a novel neutralization line for mobile terminals. *IEEE Antennas Wirel. Propag. Lett.* **12**, 1428–1431 (2013)
8. Prasanna, K.M., Behera, S.: A hexagonal MIMO antenna system with defected ground structure to enhance bandwidth and isolation (2013)
9. Xia, X.-X., Chu, Q.-X., Li, J.-F.: Design of a compact wideband MIMO antenna for mobile terminals. *Prog. Electromagn. Res. C* **41**, 163–174 (2013)
10. Wu, Y.-T., Chu, Q.-X., Yao, S.-J.: A dual-band printed slot diversity antenna for wireless communication terminals. In: 2013 IEEE International Wireless Symposium (IWS), pp. 1–3 (2013)
11. See, C.H., Abd-Alhameed, R.A., Zhou, D., Lee, T.H., Excell, P.S.: A crescent-shaped multiband planar monopole antenna for mobile wireless applications. *IEEE Antennas Wirel. Propag. Lett.* **9**, 152–155 (2010)
12. Xiong, L., Gao, P.: Compact dual-band printed diversity antenna for Wimax/WLAN applications. *Prog. Electromagn. Res. C* **32**, 151–165 (2012)
13. Li, J.-F., Chu, Q.-X., Huang, T.-G.: A compact wideband MIMO antenna with two novel bent slits. *IEEE Trans. Antennas Propag.* **60**, 482–489 (2012)

Chapter 21

Wideband Linearly Polarized Printed Monopole Antenna for C-Band

Touhidul Alam, Mohammad Rashed Iqbal Faruque
and Mohammad Tariqul Islam

Abstract This paper presents a printed wideband elliptical patch antenna on FR-4 substrate for C-band applications. High-frequency structural simulator (HFSS) based on the finite element method (FEM) and Computer simulation technology (CST) based on the finite difference time domain (FDTD) softwares have been used in this research and a wide bandwidth of 4.34 GHz (3.38 to 7.72 GHz) was achieved. The parametric study and equivalent impedance matching circuit of the proposed antenna has been investigated.

21.1 Introduction

The elliptical patch antenna has been playing an important role in communication system due to its some special characteristics like ultra wideband, easy fabrication and compact in size. However, the conventional antenna exhibits narrow bandwidth [1–6], but the present communication system demands compact antenna with wider bandwidth. To meet these demands, designers are looking for methods to increase the bandwidth of the antenna. For example, in [7], Behdad et al. proposed a new method to increase wideband by using slot. And Deshmukh et al. used proximity-fed in rectangular microstrip antenna for achieving wider bandwidth in [8]. Moreover, to improve bandwidth Rafi et al. proposed a V-shaped slot antenna in [9]. Parasitic elements have been used to increase the bandwidth in [10, 11].

T. Alam (✉) · M.R.I. Faruque
Space Science Center (ANGKASA), Research Centre Building, Universiti Kebangsaan
Malaysia, 43600UKM Bangi, Selangor, Malaysia
e-mail: touhid13@yahoo.com

M.T. Islam
Department of Electrical Electronic and System Engineering, Universiti Kebangsaan
Malaysia, 43600UKM Bangi, Selangor, Malaysia

Luis et al. proposed elliptical antenna array for wider bandwidth in [12], but the antenna size was 98×28 mm for operating frequency band of 5.3 to 6.4 GHz.

In this paper, we presented a wideband elliptical patch antenna for C-band applications. We achieved a wide bandwidth of 3.43 GHz, from 3.38 to 7.72 GHz, with the presented antenna size of $30 \times 34 \times 1.6$ mm³. Parametric analysis has been performed and analysed LC matching circuit of the presented antenna.

This paper is presented as follows: In Sect. 21.2, detail in geometry of the presented antenna, in Sect. 21.3, parametric study of the presented antenna will be discussed. In Sect. 21.4, LC matching circuit of the antenna and in Sect. 21.5, result and discussions will be presented. Finally in Sect. 21.6, conclusion will be presented.

21.2 Antenna Design

The design layout of the proposed antenna is given in Fig. 21.1. The proposed antenna has been printed on 1.6 mm thick FR-4 substrate materials with material the properties of relative permittivity of 4.6, relative permeability of 1 and dielectric loss tangent of 0.02. The dimensions of the antenna are shown in Table 21.1.

21.3 Parametric Studies

The effects of the proposed antenna parameters have been investigated. The parametric study on surface current distribution and reflection coefficient has been performed.

21.3.1 For Different Substrate Materials

The effect of substrate material properties of reflection coefficient has been observed, which is shown in Fig. 21.2. The material properties of different types of substrate material are shown in the Table 21.2.

21.3.2 Substrate Thickness

The optimum value of substrate thickness for the desired frequency band has been obtained at 1.6 mm, which is shown in Fig. 21.3.

Fig. 21.1 Design layout of the presented antenna

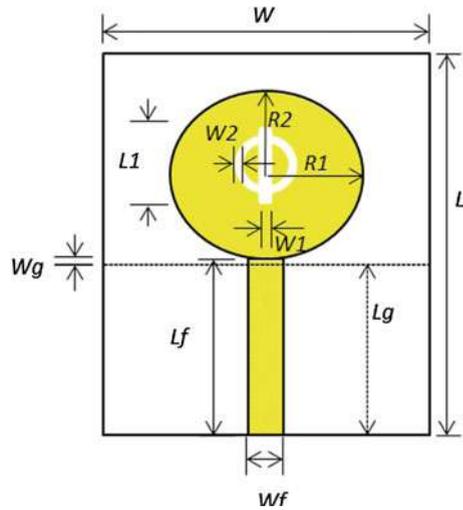


Table 21.1 Antenna design specifications

Parameters	Values (mm)
Substrate length (L)	34
Substrate height, H	1.6
Substrate width (W)	30
Patch radius R1	8
Patch radius R2	11
Feed gap (Wg)	0383
Feed line length (Lf)	16.383
Feed line width (Wf)	1.885
Ground plane length (Lg)	16
Rectangular slot length (L1)	7
Rectangular slot width (W1)	1
Circular slot width (W2)	0.6

21.3.3 The Effect of Slotting in the Patch

The effect of slotting in the elliptical patch has been investigated. The surface current of the antenna for three different frequencies without slotting and with circular slotting are shown in Figs. 21.4 and 21.5, respectively.

Fig. 21.2 Reflection coefficient for different types of substrate materials

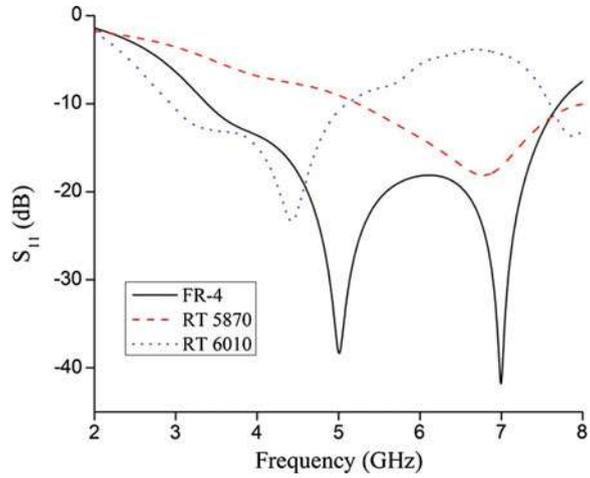
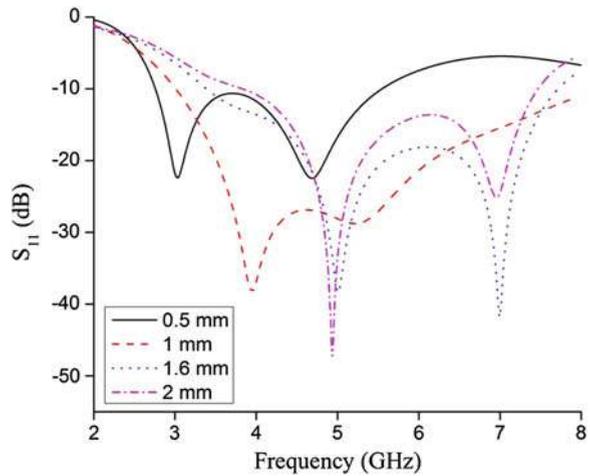


Table 21.2 Material properties of different substrates

Substrate name	Permittivity	Loss tangent	Substrate thickness
FR4	4.6	0.02	1.6
RT 5870	2.33	0.0012	1.6
RT 6010	10.2	0.002	1.6

Fig. 21.3 Reflection coefficient for different thickness of FR-4 substrate material



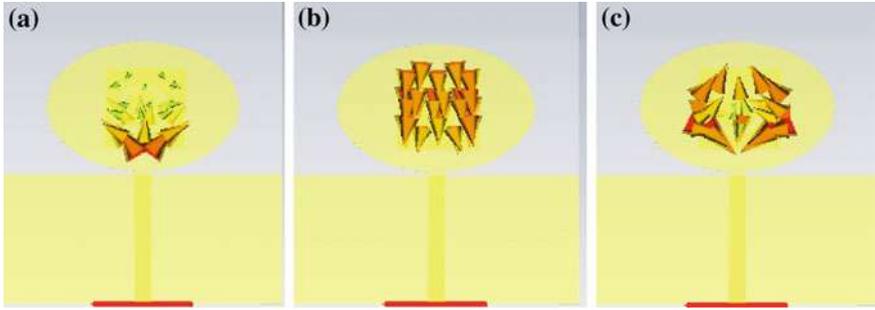


Fig. 21.4 Surface current distribution without slotting—**a** at 4 GHz, **b** at 5 GHz, and **c** at 7 GHz

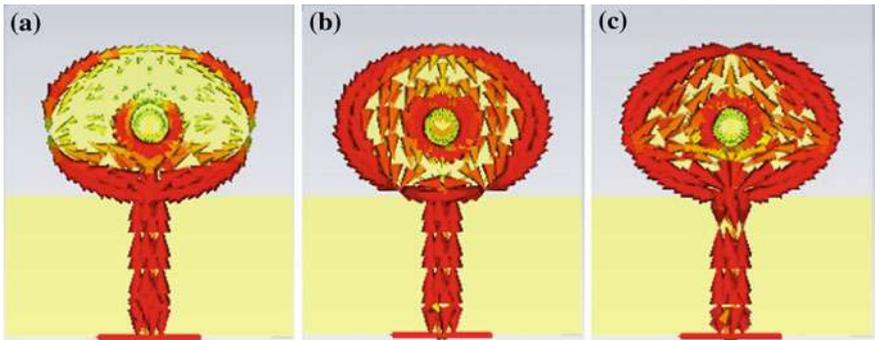


Fig. 21.5 Surface current distribution with circular slotting—**a** at 4 GHz, **b** at 5 GHz, **c** at 7 GHz

Fig. 21.6 LC matching circuit of the proposed antenna

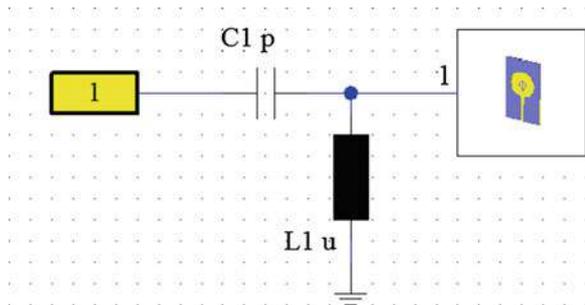
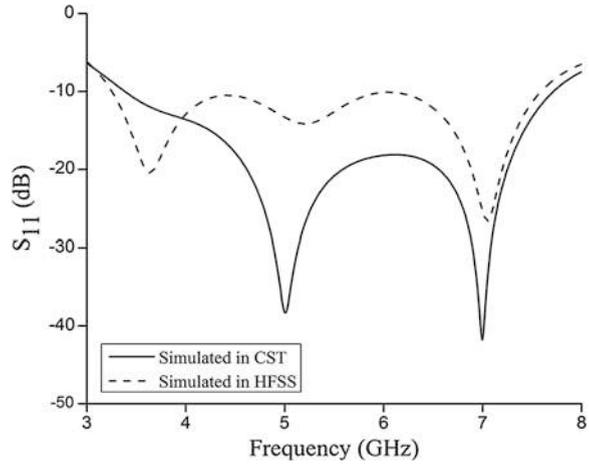


Fig. 21.7 Reflection coefficient of the presented antenna



21.4 LC Matching Circuit

Input impedance matching is an important factor in antenna designing. If the antenna is not properly matched, then the input impedance will fluctuate with the transmission line length. That's why sufficient power will not be delivered to the antenna. LC matching of the presented antenna is shown in Fig. 21.6. The value of the capacitor and inductor are 0.229 pF and 0.00184 μ H, respectively.

21.5 Results and Discussions

The reflection coefficient of the presented antenna has been investigated by using CST microwave studio and Ansoft HFSS simulation softwares and compared two results, which is presented in Fig. 21.7. Moreover, the farfield directivity of the antenna has been analyzed at 4 GHz, 5 GHz, 6 GHz and 7 GHz, which is presented in Fig. 21.8.

21.6 Conclusions

A wideband circularly polarized elliptical patch antenna for C-band application has been realized. A bandwidth of the presented antenna is 4.34 GHz, which is 78.19 % of fractional bandwidth with respect to centre frequency of 5.55 GHz.

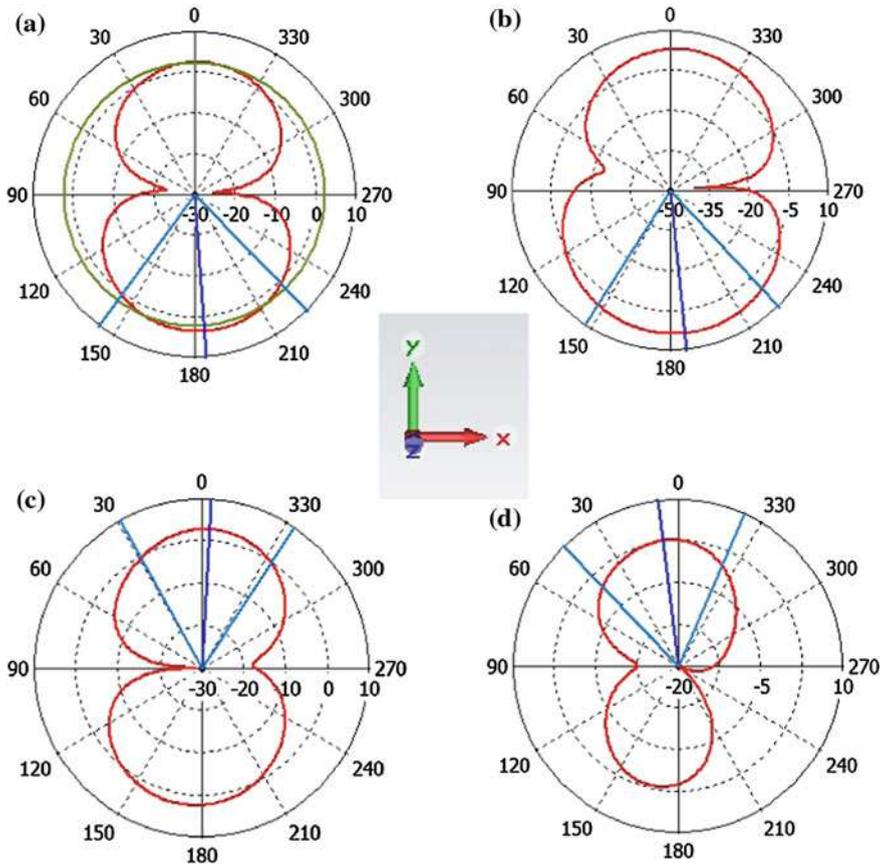


Fig. 21.8 Farfield directivity of the presented antenna—a at 4 GHz, b at 5 GHz, c at 6 GHz and d at 7 GHz

The input impedance matching of the antenna has been analysed and found approximately 50Ω of input impedance. So, the proposed antenna can be a suitable one to implement in modern communication system for C-band applications.

References

1. James, J.R., Hall, P.S.: In: Handbook of Microstrip Antennas, pp. 825–849. Peregrinus, London (1989)
2. Islam, M.M., Islam, M.T., Faruque, M.R.I.: Dual-band operation of a microstrip patch antenna on a duroid 5870 substrate for ku- and k-bands. *Sci. World J.* **2013**, 1–10 (2013) (Article ID 378420)
3. Faruque, M.R.I., Islam, M.T.: Novel design of triangular metamaterial for electromagnetic absorption in human head. *Prog. Electromagnet. Res. PIER* **141**, 463–478 (2013)

4. Faruque, M.R.I., Islam, M.T., Misran, N.: Design analysis of new metamaterial for EM absorption reduction. *Prog. Electromagnet. Res. PIER* **124**, 119–135 (2012)
5. Islam, M.T., Faruque, M.R.I., Misran, N.: Specific absorption rate analysis using metal attachment. *Informacije MIDEM* **40**(3), 238–240 (2010)
6. Faruque, M.R.I., Islam, M.T., Misran, N.: Evaluation of specific absorption rate (SAR) reduction for PIFA antenna using metamaterials. *Frequenz J.* **64**(7/8), 144–149 (2010)
7. Behdad, N., Sarahandi, K.: A multi-resonant single element wideband slot antenna. *IEEE Antennas Wirel. Propag. Lett.* **3**(1), 5–8 (2004)
8. Deshmukh, A.A., Ray, K.P.: Broadband proximity-fed modified rectangular microstrip antennas. *IEEE Antennas and Propag. Mag.* **53**(5), 41–56 (2011)
9. Rafi, Gh.Z., Shafai, L.: Wideband V-slotted diamond-shaped microstrip patch antenna. *Electron. Lett.* **40**(19) 1166–1167 (2004)
10. Wong, K.L.: *Compact and Broadband Microstrip Antennas*. Wiley, New York (2002)
11. Priyashman, V., Jamlos, M.F., Lago, H., Jusoh, M., Ahmad, Z.A., Romli, M.A., Salimi, M.N.: Effects of parasitic ring on the performance of an elliptical shaped antenna. In: *IEEE Symposium on Wireless Technology and Applications (ISWTA)*, 2012
12. Brás, L., Carvalho, N.B., Pinho, P.: Circular polarized planar elliptical antenna array. In: *Antennas and Propagation (EuCAP)*, 7th European Conference, pp 891–893 2013

Chapter 22

A Novel Anti-collision Protocol for Optimization of Remote Sensing in Dense Reader Network

Faiza Nawaz and Varun Jeoti

Abstract Passive Radio Frequency Identification (RFID) network with several reader placed densely and close to each other are susceptible to reader collision problem. In this paper, a novel and efficient RFID reader's anti-collision protocol is proposed based on Neighbor Friendly Reader Anti-collision (NFRA) mechanism by revising its contention procedure to provide higher throughput in dense reader network. The behavior of the algorithm is evaluated through a set of simulation experiments which demonstrates that the algorithm is 15 % efficient than NFRA. It also has higher fairness as compared to NFRA and other state-of-the-art proposals.

Keywords Radio frequency identification · Anti-collision protocol · NFRA · Dense reader environment

22.1 Introduction

The Radio Frequency Identification (RFID) [1] part of the Automatic Identification and Data Capture (AIDC) group has revolutionized many applications including the industrial environment monitoring. In such time critical environments, reliable and energy efficient monitoring is often required. The RFID tagging and sensing technology is becoming popular in a variety of fields, such as health care [2] warehouse inventories [3], object tracking [4] sports [5], food traceability [6] chain management [7] etc. An RFID network consists of RFID tags/sensor, readers, air interface, and backend servers. The reader communicates with the tags by means

F. Nawaz (✉) · V. Jeoti

Department of Electrical and Electronic Engineering, Universiti Teknologi Petronas,
31750 Tronoh, Perak, Malaysia
e-mail: faiza.mudassar@gmail.com

V. Jeoti

e-mail: varun_jeoti@petronas.com.my

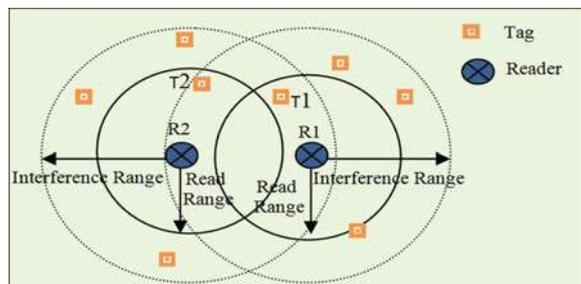
of radio signals, in a finite space, known as the interrogation zone. A passive tags have no battery, and they acquire the necessary power from the electromagnetic field of the reader. The basic idea of functioning is backscatter. These tags are cost effective and durable than active battery operated tags, but have a shorter interrogation range and limited processing capabilities. A Dense Reader Environments (DRE) consists of several interrogators closely placed in an area to be monitored using passive tags/sensors. In DRE scenarios, readers collect tags information from time to time by using a reader-to-tag identification protocol like ETSI EN 302 208, ISO/IEC 18000 and EPCglobal Class-1 Gen-2 [8]. EPCglobal is the most widely accepted standard for RFID network in which the information collected by the reader is sent to a central server (CS) by mean of a wired or wireless link using Low Level Reader Protocol (LLRP), whereas for reader to tag communication slotted ALOHA is used. However EPCglobal is not designed to deal explicitly with reader to reader collision problem that is a major issue in DRE which negates the network performance.

The collisions in DRE affect the throughput and efficiency of the system. Three types of collisions can happen in a DRE. Tag to tag collision (TTC) occurs when multiple tags respond simultaneously to same reader. Reader to Tag collision (RTC) occurs when two or more readers try to read the same tag because of an overlap in their read ranges. Reader-to-Reader Collision (RRC) happens when the signal generated by one reader interferes with the reception system of other readers [9]. Figure 22.1 shows a RRC scenario; reader 1 (R1) attempts to read data from tag 1 (T1) and reader 2 (R2) is trying to read data from tags tag 2 (T2) using the same channel. The weak response signal of T1 is degraded because of the strong interference signal of R2.

Numerous approaches have been proposed in the literature to minimize RFID network collisions [9]; however the solutions to this problem are limited. This research work proposes a new anti-collision algorithm for DRE and its operational performance is illustrated using a set of simulation iterations.

The rest of the paper is organized as follows: in Sect. 22.2, related work in the field of DRE anti-collision is discussed. The approaches are classified as distributed and centralized. This section also gives a detailed overview of Neighbor

Fig. 22.1 Reader to reader interference



Friendly Reader Anti-collision (NFRA) approach. In Sect. 22.3, the proposed algorithm is described. Section 22.4 evaluates the results of proposed algorithm under different scenarios. Section 22.5 presents a conclusion of this work.

22.2 Related Work for DRE Anti-collision Management

Implementation of reader to reader anti-collision mechanisms is very important in DRE. Two broad classifications of RRC algorithm are centralized algorithms and distributed algorithms. In distributed schemes every reader communicates directly with its neighbors and do not rely on a centralized device for network resources allocation. The most common distributed technique for RRC, based on carrier sense multiple access (CSMA) or listen before talk (LBT) is Pulse [10]. In Pulse random back off time is introduced, to avoid simultaneous transmission by more than one reader. Pulse protocol has less overhead but when compared with Time division Multiple Access (TDMA) based approaches it lacks in efficiency and throughput. An extension of Pulse protocol is [11] in which Slot Occupied Probability (SOP) is used for reducing the number of collisions. The algorithm effectively mitigates reader collisions in dense reader mode. Distributed tag access with Collision avoidance (DiCa) is another distributed algorithm based on listen before talk and focusing both on RTC and RRC reduction [12]. It is an energy saving system, however readers' energy consumption has a minor impact on operational cost, and high performance and low complexity are more desirable features in RFID reader networks. Reader Anti-Collision Algorithm for Multi-channel Mobile RFID Networks (RAC-Multi) [13] has separate data and control channels for communication. Adjacent channel interference is avoided by introducing even and odd numbered data channels. Distributed Color Scheme (DCS) protocol uses a single frequency for all the readers, with same frame sizes. Time is divided in predetermined identification cycles, which are subdivided into time-slots named as colors [14]. Probabilistic DCS (PDCS) [15] is the extension of this work intended for increasing the low performance of DCS. In PDCS a probability is considered for choosing new colors for collided readers. It reduces the number of collisions and also the number of readers that change color with time. Colorwave [14] is proposed with the aim of improving the low performance of DCS. In this approach each identification cycle has a variable number of colors and the number of colors per cycle increases when RRC are extremely high. Colorwave performs better than ALOHA algorithms in less dense networks and slightly worse in highly loaded networks. Expowave [16] is also a novel approach that outperforms Colorwave and DCS.

In centralized RRC mechanism, server is the acting agent for all communications. All readers are connected to a server that stores both the communications and the readers information. Among the most notable, and high performance centralized RRC avoidance algorithms is Neighbor Friendly Reader Anti-collision (NFRA) [17]. An extension of NFRA with emphasis on high fairness and less RRC

is proposed in [18], but the protocol is not compatible with EPCglobal Class-1 Gen-2, and its implementation in real world requires the use of extra wireless network at 433 MHz. Hierarchical Qlearning algorithm (HiQ) is intended for finding dynamic solutions to the reader collision problem by mapping collision patterns among readers [19]. The main shortcoming of this approach is that readers have to manage a huge amount of data, also the final result depends on the quality of the neural network training. Resource allocation based on genetic algorithm (RA-GA) [20] is also a FDMA-TDMA technique based on a HiQ algorithm. It uses the SNR constraint of each reader to appropriately assign spectral and temporal resources. However, there is no description given about how it satisfies the requirements of standards and regulations. Distributed adaptive power control (DAPC) [21] is a novel solution which used a back off algorithm to improve coverage. Distributed color no cooperative selection (DCNS) is a high throughput solutions for static RFID networks [22]. It uses the killer configuration and dynamic priority management for improving the performance of RFID readers as compare to other state of the art reader to reader anti-collision protocols. Distributed color no cooperative selection (DCNS) protocol is based on Colorwave [23], with additive killer configuration. It reduces the unused time slots to increase throughput of the network. It does not require any prior deployment knowledge, and it is appropriate for low cost RFID readers. It has reduced channel control overhead compare to Colorware by employing a new color update mechanism. DCNS provides 16 percent higher throughput than NFRA. Geometric distribution reader anti-collision (GDRA) [24] is a new centralized scheduler which exploits the Sift geometric probability distribution function to minimize reader collision problems.

The results presented in literature shows that centralized algorithms are more efficient in term of throughput as compared to distributed techniques. The work proposed in this paper is based on NFRA [17] in which a polling server is designated to divides the time into identification rounds. Every round begins by an arrangement command AC (random numbers from 1 to maximum number, MN) broadcasted to all the readers. The readers that receive the AC, generate its own random number. The server then issues an ordering command (OC); the readers then compare their random numbers with the value in the OC. If both values are same, the readers broadcasts their beacons signals to determine whether a collision occurs or not. If a readers does not detect any collisions, it send overriding frame (OF) to the neighboring readers. The OF prevents the neighboring readers from receiving the next OC from the server. The neighboring readers which do not identify the next OC due to the OF or which detect a collision of beacons do not actively operate. Communications between a Reader and Tags (CRT) is performed by successful readers only. The throughput of NFRA is affected by the random selection of MN. In many situations a collision detected at the beginning of the round could no longer be valid at the end. Next section describes the details of proposed algorithm based on modified NFRA contention phase.

22.3 Proposed NFRA-C Algorithm

This Section illustrates the proposed RRC avoidance algorithm. This anti-collision algorithm extends NFRA, using counters, in dense reader networks. Each reader maintains the history of his successful communication with tags in the form of counters. The counter is incremented one time for each successful communication. Counters are exchanged along with beacons to detect the presence of other readers within the range. Algorithm 1 describes the pseudo code of the proposed mechanism. Whenever a collision of beacons is detected, counter of both colliding readers are compared and the reader with a lower value of counter is permitted to participate in that particular round.

Consider the scenario shown in Fig. 22.2. After a collision is detected between reader 1 and 2, counters are compared and the reader with lower value of counter participates in that round, unlike NFRA which does not permit any reader to communicate with tags until a new AC is received. Each identification round begins by the server broadcasting AC to all the readers. Upon receiving the AC readers generate a random number between 1-MN, and wait for OC from server. As soon as the OC is received each reader compares the OC with its random number and if the generated random number is equal to OC number reader broadcasts beacons to its neighbors confirming its intension to participate in that round. If a single collision of beacon is detected at a reader then counters are compared. If the value of counter broadcasted by the reader is less than the received value; the reader broadcasts an OF to its neighbors and conduct

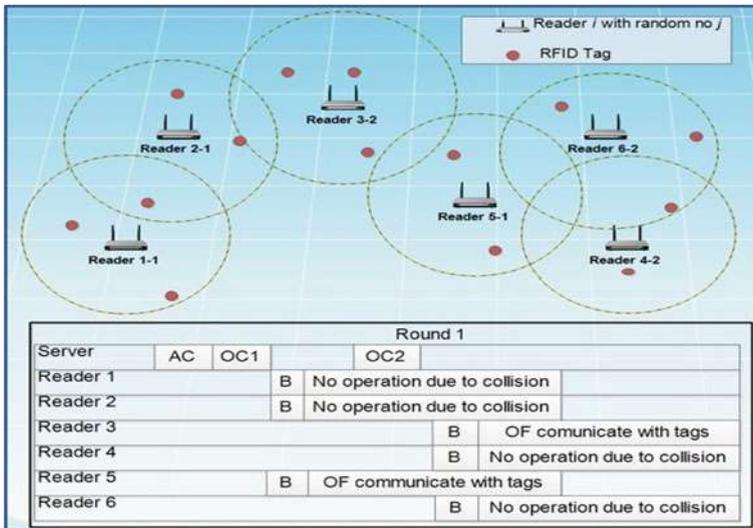


Fig. 22.2 NFRA network scenario

identification of tags during CRT time. Else if multiple collisions at a reader are detected than that reader has to wait until next AC from the server is received.

Algorithm 1 NFRA-C anti-collision algorithm

```

1: Loop
2: While a signal is not received from the server do
3:   No operation
4: End while
5: If a reader receives signal = AC then
6:   Generate a random number  $r$  among  $[1, MN]$ 
7:   Wait for OC from the server
8:   Decode every received OC to extract number  $j$ 
9: If (  $j = r$  ) then
10:  Broadcast a beacon to neighbor readers
11: If a single beacon collision is detected then
12:  Compare counter
13:  If send counter < received counter
14:  Broadcast OF to neighbor readers
15:  Conduct identification during CRT
16: Else if send counter > received counter || send counter = received counter
17:  Wait until next AC signal from server
18:  End if
19: Else if multiple collisions of beacons is detected then
20:  Wait until next AC from server
21:  End if
22: Else
23:  Broadcast OF to neighbor readers
24:  Conduct identification during CRT
25:  End
26: Else If OF is received
27:  Wait until the next AC from server
28: Else
29:  Wait until the next OC from server is received
30:  End
31: End
32: End Loop

```

The proposed algorithm is evaluated using a number of simulation scenarios. The first measurable parameter investigated in the algorithm is system efficiency which depends on the successful queries performed by the readers. It is determined as (22.1)

$$\text{Efficiency} = \frac{\text{Total successful queries}}{\text{Total no. of queries}} \times 100 \quad (22.1)$$

Until, total number of queries = total number of readers

Throughput measurement is defined as the number of successful reader transmissions per second. The throughput is evaluated by using Eq. (22.2)

$$\text{System Throughput} = \frac{\text{Total successful queries}}{\text{Total Time}} \quad (22.2)$$

The Jain's fairness index [25] is used to rate the fairness of n readers. It evaluates the fairness of throughput distribution among the readers. The fairness index ranges from 0 (min fair) to 1 (max fair). It is given in Eq. (22.3)

$$j(x_1, x_2, \dots, x_n) = \frac{(\sum_{i=1}^n x_i)^2}{n \sum_{i=1}^n x_i^2} \quad (22.3)$$

where x_i is the throughput of the i th reader, and n is the quantity of readers.

22.4 Results and Performance Evaluation

The proposed algorithm is analyzed and compared with the state-of-the-art anti-collision protocols using various performance evaluation scenarios. OMNET++ is used to evaluate the performance of 100 readers, placed in a grid of 500×500 m. Efficiency of proposed algorithm is demonstrated in Fig. 22.3, with different values of MN, i.e. 16, 32 and 64. The performance with $MN = 32$ is highest. With $MN = 64$, delay becomes huge hence a lower efficiency is achieved. The performance of $MN = 64$ got better when number of readers were increased from 100.

NFRA-C shows higher percentage efficiency than NFRA and other state of the art anti-collision protocols as shown in Fig. 22.4. For 100 readers it has a 15 % increased efficiency as compare to NFRA, 21 % increased efficiency as compare to PULSE and 43 % increased efficiency as compare to Colorware. The collision rate for different RRC mechanism are compared in Fig. 22.5.

NFRA-C achieved a maximum fairness value for number of readers >75 , also providing a significant increase in throughput among readers. In Pulse the performance of readers is dramatically reduced by two neighbors that take turns with each other at querying tags. In Colorwave, each reader varies the number of colors used in each round; hence there is a significant difference of the frequency used by the readers to query tags, resulting in a severe impact on fairness. The fairness of NFRA is affected by the random number and no history maintenance of successful rounds as shown in Fig. 22.6.

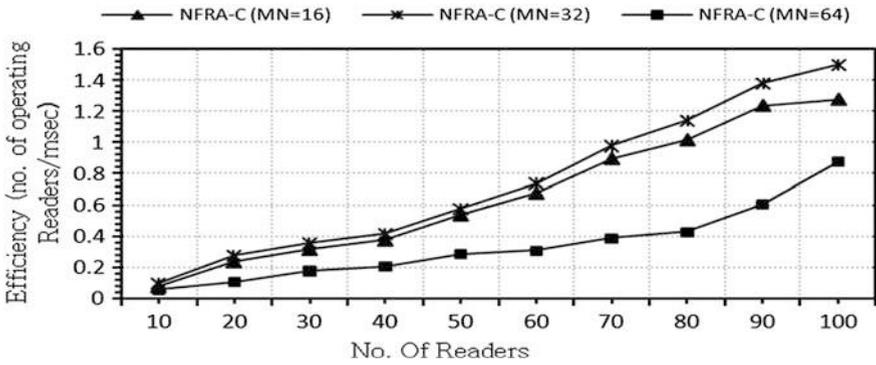


Fig. 22.3 Efficiency of NFRA-C with MN = 16, 32, 64

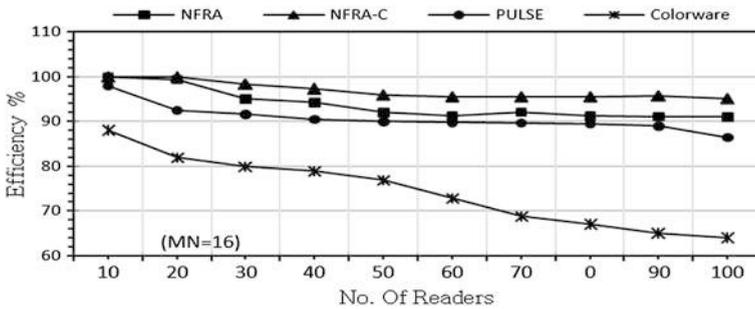


Fig. 22.4 Efficiency comparison of NFRA-C with MN = 16

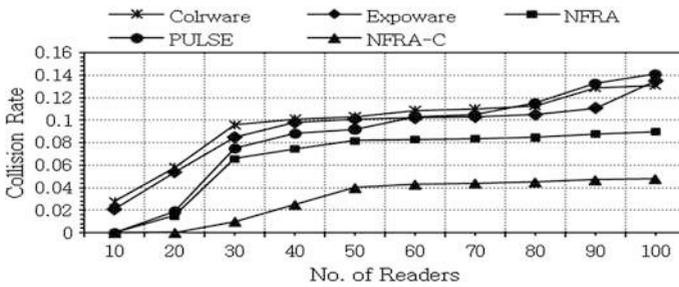


Fig. 22.5 Collision Rate comparison of NFRA-C

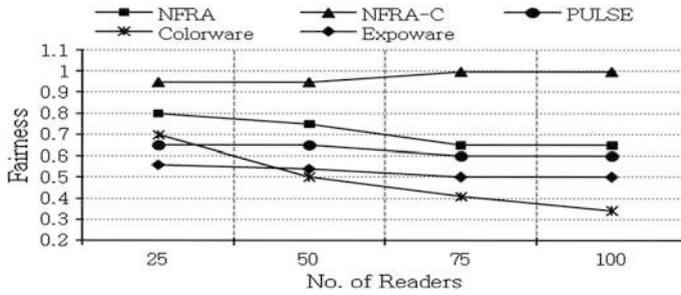


Fig. 22.6 Fairness comparison of NFRA-C

22.5 Conclusion

NFRA-C is a novel reader to reader anti-collision approach that provides high throughput and fairness in dense RFID reader environment. Among the existing protocols, NFRA achieves higher network throughput however, its performance is dependent on number of neighbors and probability of collision. The proposed NFRA-C ensures high throughput and better fairness by introducing counters in the contention process. Readers that have not participated in tags identification process for a long time gets a higher priority. NFRA-C has been compared with NFRA, Colorware, PULSE and Expoware by analyzing collisions, efficiency, and fairness. The analysis shows that counters maintained by each reader helps in increasing the performance and fairness of the network, and providing a throughput better than state of the art DRE anti-collision approaches. The fairness reaches to a value of 1(maximum) for number of readers >75 showing that all readers receive almost same allocation and the efficiency is 15% increased as compare to NFRA.

References

1. López, T.S.: RFID and sensor integration standards: State and future prospects. *Comput. Stand. Interfaces* **33**(3), 207–213 (2011)
2. Chena, Y.Y., Tsaib, M.L.: An RFID solution for enhancing inpatient medication safety with real-time verifiable grouping-proof. *Int. J. Med. Inform.* **83**(1), 70–81 (2014)
3. Vitaz, J., Buerkle, A., Sallin, M., Sarabandi, K.: Enhanced detection of on-metal retro-reflective tags in cluttered environments using a polarimetric technique. *IEEE Trans. Antennas Propag.* **60**(8), 3727–3735 (2012)
4. Geng, L., Bugallo, M., Athalye, A., Djuric, P.: Indoor tracking with RFID systems. *IEEE J. Sel. Top. Signal Process.* **8**(1), 96–105 (2014)
5. Bialkowski, A., Lucey, P., Carr, P., Denman, S., Matthews, I., Sridharan, S.: Recognising team activities from noisy data. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2013)

6. Exposito, I., Gay-Fernandez, J.A., Cuinas, I.: A complete traceability system for a wine supply chain using radio-frequency identification and wireless sensor networks [wireless corner]. *IEEE Antennas Propag. Mag.* **55**(2), 255–267 (2013)
7. Chena, J.C., Chengb, C.-H., Huangb, P.B.: Supply chain management with lean production and RFID application: a case study. *Expert Syst. Appl.* **40**(9), 3389–3397 (2013)
8. EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Protocol for Communications at 860 MHz–960 MHz, Jan 2005
9. Nawaz, F., Jeoti, V., Awang, A., Drieberg, M.: Reader to reader anticollision protocols in dense and passive RFID environment. In: 11 IEEE Malaysian International Conference on Communication, Kuala Lumpur, Malaysia, Nov 2013
10. Birari, S.M., Iyer, S.: Mitigating the reader collision problem in RFID networks with mobile readers. In: Proceedings of the 13th IEEE International Conference on Networks, Nov 2005
11. Song, I., Hong, S., Chang, K.: An improved reader anti-collision algorithm based on pulse protocol with slot occupied probability in dense reader mode. In: IEEE 69th Vehicular Technology Conference, VTC Spring 2009, April 2009
12. Kwang-il, H., Kyung-tae, K., Doo-seop, E.: Distributed tag access with collision avoidance among mobile RFID readers. In: International Conference on Computational Science and Engineering, Vancouver, Canada (2009)
13. Shin, K., Song, Q.: RAC-multi: reader anti-collision algorithm for multichannel mobile RFID networks. *Sensors* **10**, 84–96 (2009)
14. Waldrop, J., Engels, D.W., Sarma, S.E.: Colorwave: an anticollision algorithm for the reader collision problem. In: IEEE International Conference on Communications (2002)
15. Gandino, F., Ferrero, R., Montrucchio, B., Rebaudengo, M.: Probabilistic DCS: an RFID reader to reader anti-collision protocol. *J. Netw. Comput. Appl.* **34**(3), 821–832 (2011)
16. Konstantinou, N.: Expowave: an RFID anti-collision algorithm for dense and lively environments. *IEEE Trans. Commun.* **60**(2), 352–356 (2012)
17. Eom J.; Yim S.; Lee T.: An efficient reader anti-collision algorithm in dense RFID networks with mobile RFID readers. *IEEE Trans. Ind. Electron.* **56**(7), 2326–2336 (2009)
18. Ferrero, R., Gandino, F., Motrucchio, B., Rebaudengo, M.: Fair anti-collision protocol in dense RFID networks. In: Third International EURASIP Conference on RFID Technology, Spain (2010)
19. Ho, J., Engels, D.W., Sarma, S.E.: HiQ: a hierarchical Qlearning algorithm to solve the reader collision problem. In: International Symposium on Application and the Internet Workshops (2006)
20. Seo, H., Lee, C.: A new GA-based resource allocation scheme for a reader-to-reader interference problem in RFID systems. In: IEEE ICC, Cape Town, South Africa (2010)
21. Cha, K., Jagannathan, S.: Adaptive power control protocol with hardware implementation for wireless sensor and RFID reader networks. *IEEE Syst. J.* **1**(2), 145,159 (2007)
22. GandinGandino, F., Ferrero, R., Montrucchio, B., Rebaudengo, M.: DCNS: an adaptable high throughput RFID reader-to-reader anticollision protocol. *IEEE Trans. Parallel Distrib. Syst.* **24**(5), 893–905 (2013)
23. Waldrop, J., Engels, D., Sarma, S.: Colorwave: an anticollision algorithm for the reader collision problem. In: IEEE International Conference on Communications (2002)
24. Bueno-Delgado, M.V., Ferrero, R., Gandino, F., Pavon-Marino, P.: A geometric distribution reader anti-collision protocol for RFID dense reader environments. *IEEE Trans. Autom. Sci. Eng.* **10**(2), 296–306 (2013)
25. Jain, R., Chiu, D., Hawe, W.: A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report DEC-TR-301, Digital Equipment Corporation, Maynard, Mass, USA (1984)

Chapter 23

Double Square Loop Frequency Selective Surface (FSS) for GSM Shielding

Nur Khalida Binti Abdul Khalid and Fauziahanim Binti Che Seman

Abstract This paper proposes the deployment of Frequency Selective Surface (FSS) in any close building that aims to block the signals from the mobile phones without disrupting other types of communication. The proposed FSS is designed as a band-stop filter to attenuate the Global System for Mobile Communications (GSM) frequency bands operating at 900 and 1,800 MHz. The structure consists of a periodic array of double square loop elements etched on FR-4 substrate. The FSS shows a stable frequency response for the angles of incidence ranging from 0–60° with an attenuation of at least 16.7 dB. The measured results are shown to be in a very good agreement with the simulated results.

23.1 Introduction

The increasing use of wireless communication systems can lead to an electromagnetic interference. This includes an interference of the widely used GSM signals with the highly sensitive equipments employed in the buildings such as hospitals, airports and military camps. This interference not only can degrade the system's performance but also can harm the security of the system [1]. In order to reduce the interference level, an environment without the radiation from the GSM sources needs to be provided. This can be done by using few approaches available such as the mobile phone jammer and the shielding paint techniques [2, 3]. Practically, the mobile phone jammer can be used in any location since it can be easily carried due to the size that is fairly small. However, this electronic device

N.K.B.A. Khalid (✉) · F.B.C. Seman
Faculty of Electrical and Electronic Engineering, University Tun Hussein Onn Malaysia,
86400 Batu Pahat, Johor, Malaysia
e-mail: nur.khalida.khalid@gmail.com

F.B.C. Seman
e-mail: fauziahs@uthm.edu.my

requires power supply to operate [2]. Moving towards the environmental friendly approach, the needs for power supply can be eliminated by using the shielding paint technique. However, this technique is against the operation of the jammer as it blocks all other microwave signals that might be useful for other types of communication [3].

Therefore, in this study, a double square loop FSS that acts as a spatial filter with the ability to attenuate the GSM signals is proposed to overcome the limitations of the previous shielding techniques [4]. In Malaysia, the commercially available GSM frequency bands are GSM900, GSM1800 and IMT2000 [5]. In this paper, a prototype of the FSS is designed to reflect and block the incident signals at GSM900 and GSM1800 only. GSM is the second generation (2G) of mobile communication standard used for cellular network and introduced in 1990s. Since then, few newer standards are implemented such as the Universal Mobile Telecommunication System (3G) and the Long Term Evolution (LTE) Advanced (4G). However, the operating frequency of GSM900 and GSM1800 are still applicable. The bandwidth requirements provided by Malaysian Communications and Multimedia Commission (MCMC) are taken into account when designing the FSS.

The FSS is widely used as antenna radomes [6], spatial filters [1], dichroic reflectors [7] and many other applications. In order to design the FSS, there are various types of element that can be used such as square loop [1], circular loop [8], cross dipole [9] or convoluted [10]. The application of the proposed FSS is similar to [4], where the FSS is designed as a dual-band to attenuate GSM frequency bands and transmit other microwave signals. However, in [4], the authors introduced a new dual-band element of the square loop type and did not emphasize the stability of the FSS under various angles of incidence.

This paper presents the performance of double square loop FSS for normal and oblique angles of incidence up to 60° . The design details, including the dimensions of square loop FSS are shown in Sect. 23.2. The simulation by using commercially available Computer Simulation Technology (CST) software and discussion are described in Sect. 23.3. The fabrication and the measurement setup conducted in an anechoic chamber are discussed in Sect. 23.4. Lastly, in Sect. 23.5 the conclusion of the study is presented.

23.2 FSS Design Considerations

The square loop element is chosen due to its superior performance in terms of the stability under various angles of incidence, cross-polarization, bandwidth and level of band separation [11]. The unit cell dimensions of square loop FSS are shown in Fig. 23.1a. The square loop element is printed on dielectric substrate with the dielectric permittivity, $\epsilon_r = 4.3$ and thickness, $t = 1.6$ mm. The effective dielectric permittivity is calculated as $\epsilon_{eff} = 0.5(\epsilon_r + 1)$ [12]. The resonance frequency

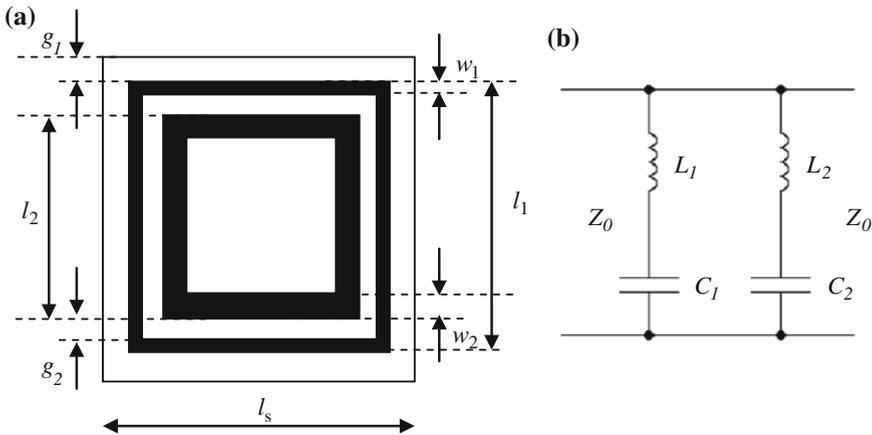


Fig. 23.1 **a** Unit cell of square loop FSS, $l_s = 46.6$ mm, $l_1 = 44.6$ mm, $l_2 = 38.0$ mm, $w_1 = 1.0$ mm, $w_2 = 2.5$ mm, $g_1 = 1$ mm, $g_2 = 3.3$ mm **b** An equivalent-circuit model for double square loop FSS

will decrease as the dielectric permittivity increases due to the loading effect of the dielectric [4]. Therefore, it is very crucial to take into account the dielectric properties of the substrate when designing the FSS.

Another important design parameters to consider is the dimensions of both square loops FSS. Generally, a double square loop FSS can be represented by a parallel equivalent circuit with a capacitive component in series with an inductive component [1] as illustrated in Fig. 23.1b. The resonance frequency of the FSS is inversely proportional to \sqrt{LC} in which contributed by the periodicity, length and width of the conductor loop. In CST, the optimization for dimensions of the outer and inner square loops FSS are performed separately. Note that the outer and inner square loops provide the resonance frequency at 900 and 1,800 MHz respectively. Further optimization is required after combining the inner and outer square loops in a single unit cell due to the effect of mutual coupling between both elements. A higher packing density FSS per unit area is preferable due to the insensitiveness to oblique angles of incidence.

The length and width of the outer square loop element is optimised to 44.6 and 1.0 mm respectively to tune the resonance frequency at 900 MHz. On the other hand, the inner square loop that contributes to a higher resonance frequency at 1,800 MHz is optimised to the length of 38 mm and width of 2.5 mm. Both of the square loops are designed to provide sufficient attenuation at the GSM operating bands from 880 to 960 MHz and 1,710 to 1,880 MHz [5] in which the bandwidth requirements are 80 and 170 MHz respectively.

23.3 Simulated Results and Discussion

The proposed double square loop FSS as shown in Fig. 23.1 is simulated by using commercially available CST Microwave Studio software. The frequency domain solver is chosen to emulate an infinite size and highly resonant structures. Since the FSS can be seen as an infinite periodic structure, therefore, it is only necessary to analyse a single unit cell of the FSS.

Figure 23.2 shows the simulated transmission frequency response that is obtained for TE polarization under various angles of incidence. At a normal incidence, the FSS provides a maximum attenuation of 34.1 and 37.7 dB at the resonance frequencies of 920 and 1,830 MHz respectively. The -10 dB bandwidth for GSM900 frequency band is 290 MHz which is from 745 to 1,035 MHz. On the other hand, the -10 dB bandwidth for the second frequency band, GSM1800 is 660 MHz which is from 1,590 to 2,250 MHz. As stated in the previous section, the bandwidth requirements for GSM900 and GSM1800 are 80 and 170 MHz respectively. These indicate that the proposed FSS is capable to provide adequate attenuation for the GSM signals far beyond the requirements. By referring to the MCMC operating band requirement, the attenuation varies between

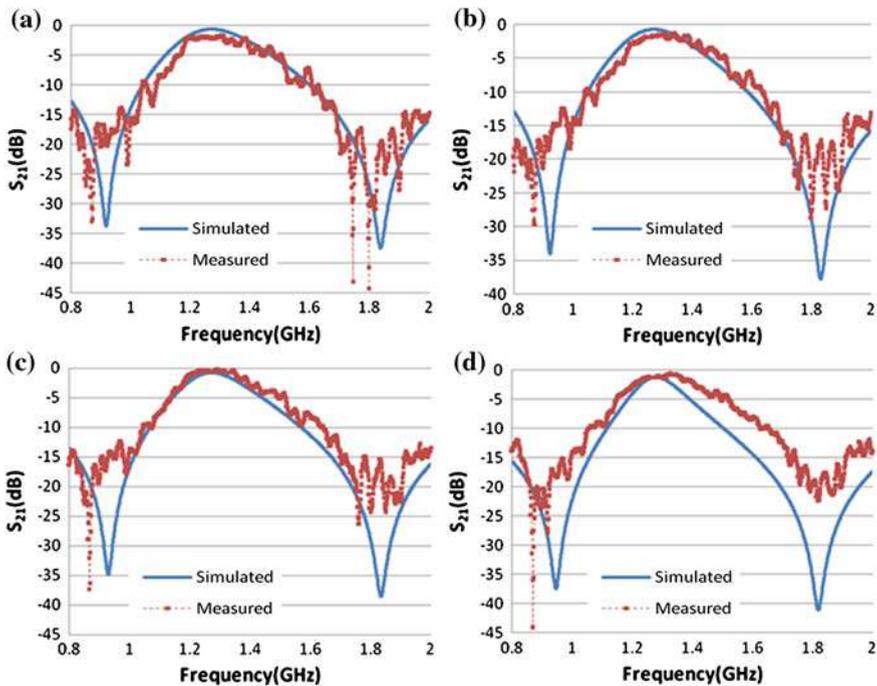


Fig. 23.2 Simulated and measured transmission frequency response **a** at normal incidence **b** $\theta = 20^\circ$ **c** $\theta = 40^\circ$ and **d** $\theta = 60^\circ$

Table 23.1 The attenuation of the FSS for TE polarization under various angles of incidence

Frequency band provided by MCMC	Attenuation			
	0°	20°	40°	60°
GSM900 (880–960 MHz)	19.6–34.1 dB	20.3–34.1 dB	21.5–35.1 dB	21.8–37.5 dB
GSM1800 (1,710–1,880 MHz)	16.7–37.7 dB	17.2–37.7 dB	17.8–38.7 dB	21.1–40.8 dB

19.6 and 34.1 dB for 880–960 MHz while 16.7–37.4 dB for 1,710–1,880 MHz as tabulated in Table 23.1.

As the angle of incidence increases to 20°, both resonance frequencies as well as the attenuation remain the same. However, this trend slightly changes as the angle of incidence varies to 40° and 60°. At 40°, the resonance frequency for the lower band increases to 930 MHz with an attenuation of 35.1 dB. The upper resonance frequency remains the same as 1,830 MHz. However, the maximum attenuation at that resonance frequency improves to 38.7 dB. As the angle of incidence increases up to 60°, the lower resonance frequency shifts right by 2.8 %–946 MHz. In contrast, the upper resonance frequency shifts left by 0.5 %–1,820 MHz. The attenuation for both resonance frequencies, 946 and 1,820 MHz at the oblique angle, 60° greatly improves to 37.5 and 40.8 dB respectively. For the GSM900 frequency band, the –10 dB bandwidth performance increases to 453 MHz which is from 6,641 to 1,114 MHz. On the other hand, for the GSM1800 frequency band, the –10 dB bandwidth slightly decreases to 636 MHz which is from 1,518 to 2,154 MHz as the angle of incidence varies to 60°. By referring to the MCMC operating band requirements, at 60°, the attenuation varies between 21.8 and 37.5 dB for 880–960 MHz while 21.1–40.8 dB for 1,710–1,880 MHz as shown in Table 23.1. These results show that the proposed FSS manages to cover the required bands provided by the MCMC under various angles of incidence.

23.4 Fabrication and Measurement

The square loop FSS was etched on FR-4 dielectric substrate to emulate the required dielectric permittivity of 4.3. The FR-4 substrate was used due to the availability of the material in the laboratory although it is expected that the dielectric substrate with lower dielectric permittivity offers a better attenuation. Since the FSS does not require a ground plane, a single layer PCB, where the copper is only mounted on one side of the PCB was used. The fabricated square loop FSS is shown in Fig. 23.3. Since the size of the fabricated FSS was small due to the limitation of the UV exposure machine, several pieces of the FSS were fabricated. All the fabricated FSSs were properly aligned on the plywood to immitate an infinite size sheet as defined in the computer model.

The bi-static measurement technique was employed in order to measure the fabricated FSS. The measurement setup consists of two horn antennas with the

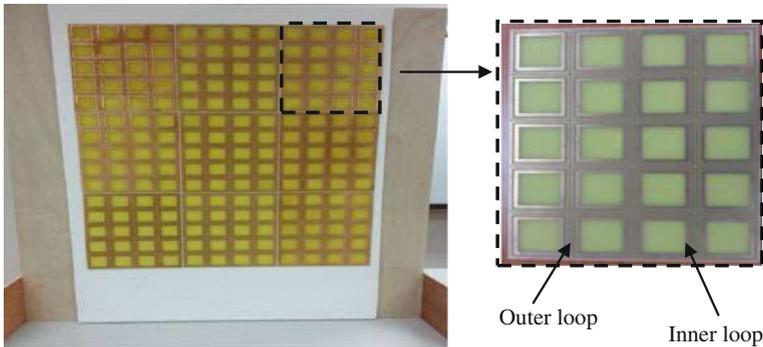
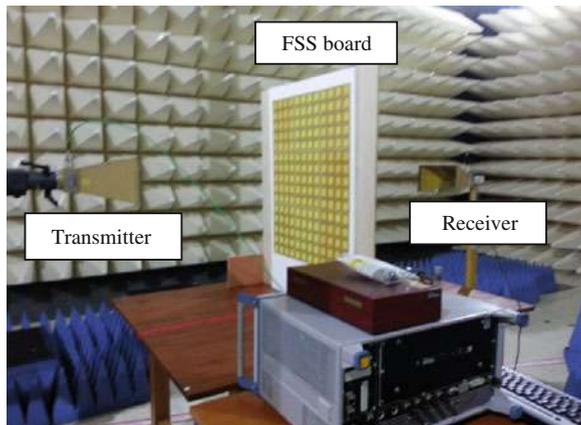


Fig. 23.3 The fabricated square loop FSS

Fig. 23.4 The measurement set up with the FSS placed in between the transmitter and the receiver



gain of 15 dB for transmitting and receiving, with both were connected to the network analyzer by using the coaxial cable. These two horn antennas were separated about 1 m away to ensure that the rule to operate in far-field region was obeyed. This is based on the equation, $d_{farfield} \geq 2D^2/\lambda$ where D is the horn antennas' maximum dimension and λ is the wavelength [13]. The frequency range of the network analyzer was set up to operate from 800 to 2,000 MHz. The fabricated FSS was positioned in the middle between the transmitter and the receiver as shown in Fig. 23.4.

During the measurement, the loss due to the propagation path was taken out in order to ensure that the attenuation of the microwave signal was contributed merely by the FSS. This was done by subtracting the receiving signal without the FSS with the receiving signal with the FSS. The comparison between the simulated and the measured results is shown in Fig. 23.3. The fluctuation appears in the measured results is expected due to the scattering of the microwave signals from the surrounding. In general, the measured results showed a good agreement with the simulated results.

23.5 Conclusions

The performance of the FSS with the double square loop elements is demonstrated. The proposed FSS shows a good shielding performance in terms of the bandwidth and the sensitivity towards different angles of incidence ranging from 0° to 60° for TE polarization. A maximum reflection at 900 and 1,800 MHz with an attenuation of at least 16.7 dB is achieved. The measured results are shown to be in a very good agreement with the simulated results. Our future work will focus on the performance of the FSS for TM polarization.

Acknowledgments The authors would like to thank the Ministry of Education Malaysia for supporting this study under the Exploratory Research Grant Scheme (ERGS/1/2012/TK06/UTHM/02/1/E005). Much appreciation also goes to the Research Center for Applied Electromagnetics, Universiti Tun Hussein Onn Malaysia for providing the measurement facilities.

References

1. Sung, G.H., et al.: A frequency-selective wall for interference reduction in wireless indoor environments. *IEEE Antennas Propag. Mag.* **48**(5), 29–37 (2006)
2. Mishra, N.K.: Development of GSM—900 mobile jammer: an approach to overcome existing limitation of jammer. In: Proceedings of the Fifth IEEE Conference Wireless Communication and Sensor Networks (WCSN), pp. 1–4. (2009)
3. Y-shield EMR-Protection, *Shielding Paints* [Online]. Available: <http://www.yshield.com/shielding-paints.html>
4. Kiermeier, W., Biebl, E.: New dual-band frequency selective surfaces for GSM frequency shielding. In: Proceedings. 37th Eur. Microwave Conference, pp. 222–225. (2007)
5. Malaysian Communications and Multimedia Commission, *Spectrum Allocation* [Online]. Available: <http://www.skmm.gov.my/Spectrum/Spectrum-Allocation-List/Spectrum-Allocation.aspx>
6. Zhao, J., Xu, X.: Study of the effect of a finite FSS radome on a horn antenna. In: IEEE International Conference on Microwave Technology and Computational Electromagnetics (ICMTCE), pp. 74–76. (2011)
7. Pasian, M., et al.: Accurate modeling of dichroic mirrors in beam-waveguide antennas. *IEEE Trans. Antennas Propag.* **61**(4), 1931–1938 (2013)
8. Taylor, P.S., et al.: A passively switched dual-band circular FSS slot array. In: Proceedings of the IEEE-APS Topical Conference Antennas and Propagation in Wireless Communications (APWC), pp. 648–651. (2011)
9. Kiani, G.I., et al.: Cross-dipole bandpass frequency selective surface for energy-saving glass used in buildings. *IEEE Trans. Antennas Propag.* **59**(2), 520–525 (2011)
10. Parker, E.A., et al.: Frequency selectively screened office incorporating convoluted FSS window. *Electron. Lett.* **46**(5), 317–318 (2010)
11. Wu, T.K.: *Frequency Selective Surface and Grid Array*. Wiley, New York (1995)
12. Munk, B.A.: *Frequency Selective Surfaces: Theory and Design*. Wiley, New York (2000)
13. Raspopoulos, M., Stavrou, S.: Frequency selective buildings through frequency selective surfaces. *IEEE Trans. Antennas Propag.* **59**(8), 2998–3005 (2011)

Chapter 24

Analysis of the Active Region of Archimedean Spiral Antenna

Abdirahman Mohamoud Shire and Fauziahanim Che Seman

Abstract The paper elaborates the current distribution of the Archimedean Spiral Antenna (ASA), demonstrating the concept of frequency dependent active region and this determines the effective radiation area on the spiral arm. The band theory is used to explain the theoretical principles of the operation of the spiral antenna. In this paper, a two arm Archimedean spiral antenna is designed using Computer Simulation Technology, (CST). The properties of the active region of the spiral antenna are analyzed in different types of dielectric substrates. The calculated and simulations results show the position of the active region is very dependent to the operating frequency and dielectric permittivity of the substrates. The maximum surface current induced to the spiral arm increases as the operating frequency and permittivity of the dielectric substrates reduces.

24.1 Introduction

Frequency independent antennas currently receive huge interest [1, 2] due to their special properties in wideband applications. The term frequency-independent (FI) is reserved for antennas with electrical characteristics that vary insignificantly over an extremely wide operating frequency range [3]. This properties is valid on its own and an improper insertion of the ground plane (GP) behind the antenna might disrupt the FI belongingness [4]. In this paper, the key important parameters of the FI of an Archimedean Spiral Antenna backed by a metallic plate has been analysed accordingly. In ASA the FI property is contributed by the position of the active

A.M. Shire (✉) · F.C. Seman
Wireless and Radio Science Centre (WARAS), Faculty of Electrical and Electronic
Engineering, University Tun Hussein Onn Malaysia, 86400 Parit Raja, Johor, Malaysia
e-mail: shire_288@hotmail.com

F.C. Seman
e-mail: fauziahs@uthm.edu.my

region on the spiral arm where the effective radiation pattern of the ASA take places. Then the GP is carefully inserted in order to provide a unidirectional radiation pattern and a higher gain [5] and later, the possible change of the active region indicated by the current distribution on the spiral arm is analysed. Furthermore, the variation of the magnitude of the current induced to the spiral arm with operating frequency range and dielectric substrates is investigated.

The organization of the paper is described as follows; in the theoretical operation section of the ASA, the current distribution along the spiral arm is analyzed numerically. Later, the maximum current induced on the spiral arm associated by the position of the active region is analyzed based on the numerical and the simulation results.

24.2 Theoretical Operation for Spiral Antenna

Considering a two-arm spiral antenna operating in Mode 1, which is fed from its center at ports X and X' as illustrated in Fig. 24.1. The *band theory* (radiating ring theory) is applied to explain the theoretical principles behind the operation of the antenna. The current distribution on the antenna is divided into three areas, which are *feeding region*, *transition region* and *decay region* [6].

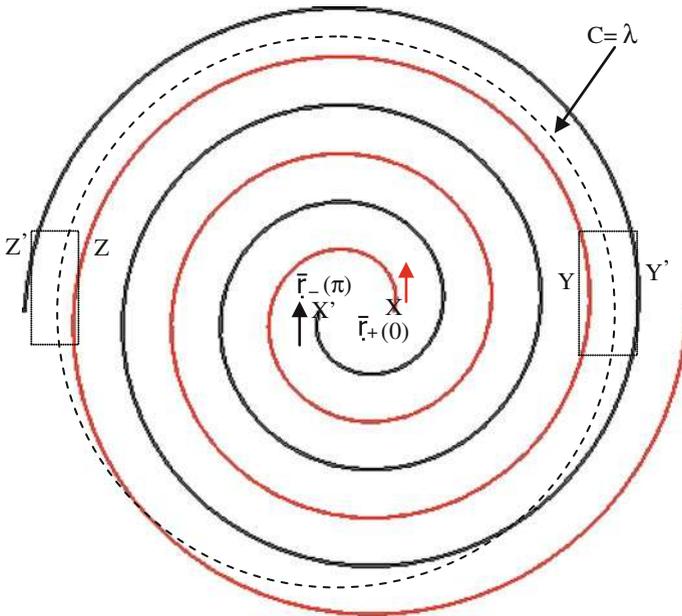


Fig. 24.1 Active region of spiral antenna based on Radiation band theory

The spiral antenna is fed at $\bar{r}_+(0)$ and $\bar{r}_-(0)$, which presents the time harmonic excitation of the spiral. Then, the current at the point $\bar{r}_\pm(\theta)$ and the phase of the current on the spiral antenna is expressed by the travel time from the feeding is defined in Eqs. (24.1) and (24.2) respectively [3, 7, 8]:

$$I_\pm(\theta, t) = i_\pm(\theta)e^{j\omega t}. \quad (24.1)$$

$$t_d = s_\pm(\theta)/v_p. \quad (24.2)$$

Where the real current of the spiral is the real part of $I_\pm(\theta, t)$, $i_\pm(\theta)$ is a complex of θ , v_p is the phase velocity and $s_\pm(\theta)$ is the distance along the spiral arms and it is found as :

$$s_\pm(\theta) - s_\pm(0) = \int_0^\theta \left| \frac{d\bar{r}_\pm(\theta)}{d\theta} \right| d\theta = \int_0^\theta \sqrt{\left(\frac{dr(\theta)}{d\theta} \right)^2 + \left(r \frac{d\theta}{d\theta} \right)^2} d\theta = (e^{a\theta} - 1) \sqrt{1 + \frac{1}{a^2}}. \quad (24.3)$$

The phase of the current is

$$\begin{aligned} \angle I_\pm(\theta, t) &= \angle I_\pm(\theta, t - t_d) = \angle I_\pm(\theta)e^{j\omega t} + \angle e^{-j\omega t} \\ &= \angle I_\pm(\theta, t) - \frac{2\pi f s_\pm(\theta)}{v_p} = \angle I_\pm(\theta, t) - \frac{2\pi f s_\pm(\theta)}{\lambda}. \end{aligned} \quad (24.4)$$

In the *Feeding region*, the spiral antenna is fed in a balanced feeding at points X and X' as illustrated in Fig. 24.1, and the feeding diameter is very small compared to the wavelength, due to that, the current these points are anti-phase and it is calculated as:

$$\angle I_+(\pi, t) - \angle I_-(0, t) \approx \angle I_+(0, t) - \angle I_+(\pi, t) = \pi. \quad (24.5)$$

The simplified mathematical formulation of active region of spiral antenna is analyzed and presented:

$$r(\theta) = e^{a\theta+b}. \quad (24.6)$$

The spiral arm rotates accordingly which can be represented as a curve [7, 8] and in polar coordinate, the constants a determines the rate of wrapping and scales the curve, constant b scales the curve, while θ is angle of the phase and calculated as:

$$\theta = \frac{1}{a} \log \left(\frac{\lambda}{2(e^{a\pi} - 1)\sqrt{1 + a^{-2}}} \right). \quad (24.7)$$

Spiral with tight wrapping can facilitate to describe the active region in simple way, so letting b tend to zero; it is obtained an approximate radiation condition from Eqs. (24.6) and (24.7):

$$r(\theta) = e^{a\theta} = \frac{a\lambda}{2(-1 + e^{a\pi})\sqrt{1 + a^2}} \approx \frac{a\lambda}{2(-1 + 1 + a\pi + \dots)1} \approx \lambda/(2\pi),$$

$$2\pi r(\theta) = \lambda, \tag{24.8}$$

Based on “radiation band” theory, Eq. (24.8) clearly elaborates the Mode 1 radiation predominantly occurs from the area whose circumference is approximately one of wavelength ($C = 1\lambda$) [8–10] as indicated in Fig. 24.1 at points Y and Y’, which belong to neighbouring arms, because the current flowing in contiguous arms is in phase leading to coherent or constructive radiation in the far-field. On the other side, the same conditions occurred diametrically opposite points Z and Z’. This area defines the position of the active region, where most of the radiation takes place due to the decaying of the in-phase current.

Outside these regions the current is not in-phase and, therefore, the radiated field interferes destructively [8–10]. The nonradiated traveling wave currents will flow past this region, and if the size of the spiral permits, radiate in the next properly phased section in which the in-phase current conditions will show up at odd wavelength circumferences of the spiral and higher order modes will radiate (this will occur at a circumference equal to three wavelengths; mode 3 for a two-arm spiral). If the spiral is not large enough, the currents will reach the end of the spiral arms where they are either absorbed or reflected back toward the spiral’s center. The position of the active region vary accordingly with the changes of the dielectric permittivity as defined by Eq. (24.9).

$$r_{eff} = \frac{\lambda_{eff}}{2\pi} = \frac{\lambda_0}{2\pi\sqrt{\epsilon_{eff}}}. \tag{24.9}$$

24.3 Archimedean Spiral Antenna Design

The ASA is designed using CST MWS where the operating frequency range is chosen to be in the 2–12 GHz range and this frequency range is available for UWB applications. The arm width (w), arm spacing (s), inner radius (r_1) and the outer radius (r_2) of the ASA are calculated using equations in [5]. The r_2 is optimized to be 35 mm while the other parameters are optimized to be 1.3 mm. The number of turns of each arm is selected as 8 turns and in computer model, the ASA is fed at the center using discrete port. The spiral arm is cascaded to an electrically $\lambda/4$ thick dielectric substrate backing by a metallic plate.

24.4 Results and Analysis

Figure 24.1 shows the current distribution on the spiral arm in which indicating the current behavior in the *feeding region*, *transition region* and *decay region*. In the *feeding region*, the current at neighbor arms is anti-phase as proven in Eq. (24.5) thus there is no radiation takes place and this is illustrated in Fig. 24.1a. The calculated phase at Eq. (24.5) is 180° , so since the phase difference at points X and X' is 180° out phase, the radiation is negligible because the two current waves of the two neighbor arms are cancelling each other due to the anti-phase condition. According to the current distribution theory of spiral antenna, the current reaches its maximum at the end of the feeding region which is clearly observed in the simulated current on the spiral antenna, in which the maximum current magnitude is 82.57 A/m. The current starts to decay to 69.24 A/m at the *transition region* as shown in Fig. 24.2b, but still no effective radiation takes place in this region since the currents at the neighbor arms are anti-phase.

Identification of the position of the active region in *decay region* is done based on the (i) calculation (See Table 24.1, third column) by referring to Eq. (24.8) and (ii) current distribution of the CST simulation results (Refer to Table 24.1, fourth column). Note that the spiral arm is attached on the free space so the dielectric permittivity, ϵ_r is 1. Figure 24.1 shows that at the lower operating frequency, the active region of the spiral is located around the perimeter (see Fig. 24.2a) of the antenna and as the frequency increases the position of the active region moves inwards the center feeding (see Fig. 24.2c). The simulated active regions of the active region are shown in Fig. 24.2. The calculated and simulated locations of the active region are tabulated in Table 24.1 and the comparisons of the two results are shown in Fig. 24.3.

The ASA spiral antenna is simulated on four different substrates permittivity, ϵ_r in order to examine the effect of the dielectric substrate on the characteristics of the active region of the ASA. It is clearly seen in Table 24.2 that as ϵ_r increases from 1 to 2.33, the maximum current induced at the active region reduces from 82.57 to 70.41 A/m. The maximum current reduces 12.7 % as ϵ_r increases from 4.3 to 10.2. This is due as the currents propagate along the spiral arm, a higher dielectric permittivity assimilates more input power fed to the antenna and therefore leads to reduction of the current distribution on the spiral arms [10]. The current magnitude also degrades as the operating frequency of the ASA increases because spiral antenna behaves like inductive circuit at higher frequencies, so the impedance of the inductive circuit increases as the frequency increases which leads to the reduction of the current magnitude [11].

The location of the active region of the ASA changes as the dielectric permittivity increases from 1 to 10.2. Numerically, based on Eq. (24.9), as the substrate permittivity increases, the location of the active region shrinks towards the

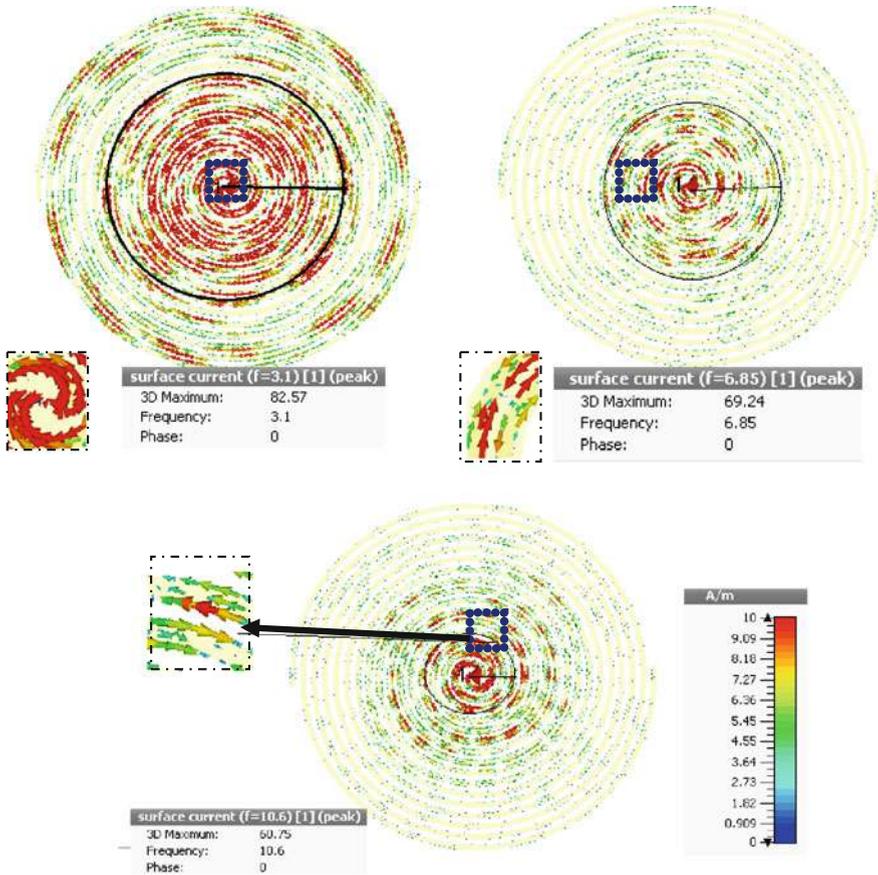


Fig. 24.2 Current distribution of two arms Archimedean spiral antenna is shown as *black band* for different frequencies **a** 3.1 GHz, **b** 6.85 GHz and **c** 10.6 GHz

center feeding of the ASA. This is proven at the operating frequency of 2 GHz when $\epsilon_r = 2.33$ the active region is positioned 18.5 mm from the center fed and this significantly reduces 10.1 mm when $\epsilon_r = 10.2$ as demonstrated in Fig. 24.4. As the operating frequency increases to 12 GHz, the position of the active region occurs 1.68 and 3.08 mm away from the feeding point for $\epsilon_r = 10.2$ and $\epsilon_r = 2.33$ respectively. Again, the higher dielectric substrate slows the traveling wave in which leads to shrinkage of the active region area and increases the coupling between the neighboring arms [5]. Both calculation and simulation results demonstrates excellent agreement as shown in Fig. 24.4.

Table 24.1 Details position of the active region

Frequency (GHz)	λ (mm)	$r(\theta)$ (mm) Calculation	$r(\theta)$ (mm) Simulation
2	150	23.87	25
3	100	15.92	16
4	75	11.94	12.5
5	60	9.4	9.5
6	50	7.96	8.2
7	42.86	6.82	8
8	37.5	5.96	7.5
9	33.33	5.31	6
10	30	4.78	5.5
11	27.27	4.34	4.8
12	25	3.98	4

Fig. 24.3 Calculation and simulation results of position of the active region of ASA

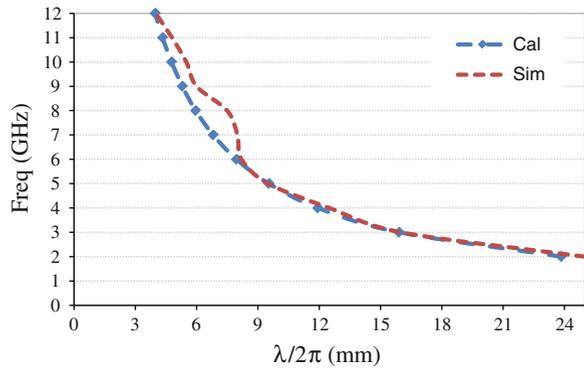


Table 24.2 The maximum current distribution (A/m) at the active region

Substrate Permittivity (ϵ_r)	3.1 GHz (A/m)	6.85 GHz (A/m)	10.6 GHz (A/m)
1	82.57	69.24	60.75
2.33	70.41	62.38	56.3
4.3	65.75	57.83	52.93
10.2	57.39	53.21	49.51

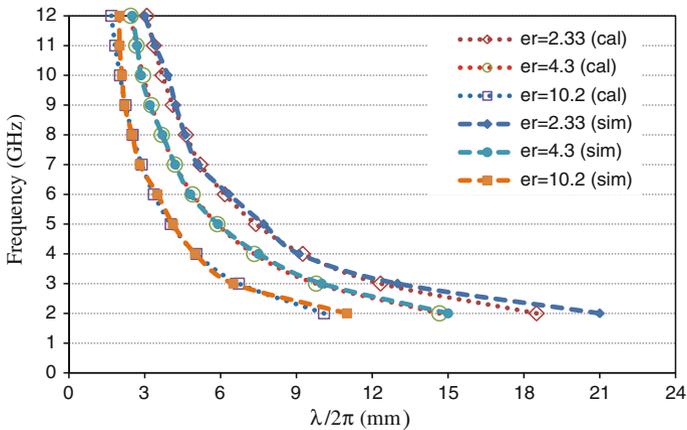


Fig. 24.4 Calculated and simulated results of the position of the active region of ASA on different dielectric substrate

24.5 Conclusion

The characteristics of the active region based calculation and simulation results are presented in this paper. The ASA is designed in CST to operate between 2 and 12 GHz for UWB applications. It demonstrated that the position of the active region depends on both operating frequency and dielectric substrate. As the operating frequency and the substrate permittivity increases, the locality of the active region shrinks inward the antenna.

References

1. Dyson, J.D.: Frequency-independent antennas (survey of development), *Electronics*, **1-01 35**, 39–44 (1962)
2. Jordan, E.C., Deschamps, G.A., Dyson, J.D., Mayes, P.E.: Developments in broadband antennas. *IEEE Spectr.* **1**, 58–71 (1964)
3. McFadden, M.: Analysis of equiangular spiral antenna. Ph.D. Dissertation, Georgia Institute of Technology (2009)
4. Nakano, H., Nogami, K.: A spiral antenna backed by a conducting plane reflector. *IEEE Trans. Antennas Propag.* **AP-34**, 1417–1423 (2007)
5. Shire, A.M., Seman, F.C.: Effects of dielectric substrate on the performance of UWB Archimedean spiral antenna. In: *Proceeding of the 2013 IEEE International Conference on Space Science and Communication (IconSpace)*, pp. 412–415. (2012)
6. Yeh, Y.S., Mei, K.K.: Theory of conical equiangular spiral antennas, part ii- current distributions and input impedances. *IEEE Trans. Antennas Propag.* **16**, 14–21 (1964)
7. Rumsey, V.H.: Frequency independent antennas. *IRE National Convention Record* **5**, 114–118 (1957)

8. Kaiser, J.A.: The Archimedean two-wire spiral antenna. IRET Trans. Antennas Propag. **AP-8**, 312–323 (1960)
9. Bawer, R., Wolfe, J.J.: The spiral antenna. IRE Int. Convention Rec. **8**, 84–95 (1960)
10. Burdine, B.H.: The spiral antenna Massachusetts Institute of Technology, Research Laboratory Technical Report, (1955)
11. Rosu, I.: Microstrip, Strip Line and CPW Design RF Technical Artical. 2014. <http://www.qsl.net/va3iul>

Chapter 25

Optimization of BER Performance in the MIMO-OFDMA System for Mobile WiMAX System Using Different Equalization Algorithm

Azlina Idris, Norhayati Abdullah, Nor Azlizan Hussein and D.M. Ali

Abstract Combination of Multiple Input Multiple Output (MIMO) and Orthogonal Frequency Division Multiple Access (OFDMA) is implemented to offer a simple and high performance system as to increase channel capacity and serve high data rate. Even though the OFDMA concept is simple in its basic principle, but it suffers one of the most challenging issues, which is synchronization error that introduces the inter-symbol interference (ISI), thus degrades the signal performance. The goal of this paper is to provide a method to mitigate this ISI by employing the equalizers at the receiver end and using Space Time Frequency Block Codes (STFBC) to improve the Bit error rate (BER) performance and to achieve a maximum diversity order in MIMO-OFDMA by using simulation based on the platforms of MATLAB software. As a result, the BER performance is improved when implementing equalizers at the receiver with STFBC outperforms the conventional system without equalizer with a maximum diversity order and an efficient bandwidth in the Mobile WiMAX system.

Keywords Inter-symbol interference (ISI) · Multiple input multiple output (MIMO) · Orthogonal frequency division multiple access (OFDMA) · Bit error rate (BER) · Space time block codes (STFBC) · Worldwide interoperability for microwave access (WiMAX)

A. Idris · N. Abdullah (✉) · N.A. Hussein · D.M. Ali
Faculty of Electrical Engineering, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
e-mail: hayatie.abdullah@yahoo.com

25.1 Introduction

Successful deployment of wireless voice communication systems promises a bright future for wireless high data rate services such as internet access or multimedia applications [1]. Orthogonal Frequency Division Multiplexing (OFDM) provides such high data rate services and considered as a good choice of this matter due to its ability to overcome multipath fading. Currently, there is a strong interest in extending the OFDM concept to multiuser communication scenarios. A prominent example of this trend is orthogonal frequency division multiple access (OFDMA) technology, which results from the combination of OFDM with a frequency division multiple access (FDMA) protocol [2]. Many broadband wireless networks have now included the MIMO option in their protocols. In principle, OFDMA and MIMO can be combined to offer the benefits of simplicity, high performance system [3] and exploitation of the multipath diversity which increases the achievable rate and enhances link reliability [4]. The proposed method is to test the diversity performance of this system by using the Alamouti code technique which is the simplest compared to the others. Basically, in this OFDMA system, there are two basic diversity order systems which are the Space-Time Block Codes (STBC) and Space-Frequency Block Codes (SFBC), while the Space-Time-Frequency Block Codes (STFBC) is the combination of both. STFBC can offer spatial, temporal and frequency diversity MIMO channels. The coding distributes symbols along transmit antennas, time slots and at different frequencies. This STFBC may contain several OFDM symbols which can increase diversity order [5].

Even OFDMA has a lot of advantages, yet there are still some disadvantages exist in this system. For instance, different users share available subcarriers in OFDMA thus, synchronization becomes a difficult task. The receiver must estimate a number of parameters and need to compensate inter-symbol (ISI) interference [6]. The cyclic prefix (CP) can be added to overcome this matter but ISI may still exist if channel delay spread is larger than the CP and this will severely affect the system performance. Thus, by adapting the ideal equalizer at the receiver, performance degradation and ISI can be reduced. There were several work done previously for instance, in [5], the researcher investigated BER of system performance using STFBC with intercarrier interference self-cancellation scheme (ICI-SC) without equalizer to reduce ICI only but not ISI.

Besides, in [7], the author introduced a diversity technique, but it is applied for MIMO-OFDM system using a new ICI-SC technique subcarrier mapping scheme without equalizers. However, it is difficult to obtain frequency diversity gain and suffers ICI and ISI. In [8], the researcher studied the performance evaluation of BER using STFBC MIMO-OFDM using equalization algorithm. However, the system could not achieve the maximum diversity order and an efficient bandwidth. In [9], the researcher evaluates system performance using pair-wise error probability with specific subcarrier mapping and a linear equalizer for MIMO-OFDM system. So far, there is no literature on performance evaluation of BER using equalization and diversity order technique (STFBC) in MIMO-OFDMA system.

Therefore, this research paper is proposed. The objectives of this paper are to simulate the BER performance of MIMO-OFDMA using MATLAB software, to mitigate the inter-symbol interference (ISI) in the OFDMA system, to implement the equalizers at the receiver and to evaluate the different type of diversity order in this system.

25.2 System Model

An OFDMA system is defined as one in which each terminal occupies a subset of subcarriers (termed an OFDMA traffic channel), and each traffic channel is assigned exclusively to one user at any time [10]. In OFDMA, users are not overlapped in frequency domain at any given time [11]. However, the frequency bands assigned to a particular user may change over the time.

In order to mitigate the presence of ISI, the implementation of equalization at the receiver can be made in frequency or time domain to diminish such interference. This paper will focus on implementing three different types of equalizers which are Zero-Forcing Equalizer (ZF), Minimum Mean Square Error (MMSE) Equalizer and Maximum Likelihood Sequence Estimation (MLSE) Equalizer. In this paper, the BER performance will be compared through with and without equalization at the receiver OFDMA system. Figure 25.1 above shows an example of a baseband model of OFDMA system. The block diagram basically comprising of three major parts namely transmitter, channel and receiver. The data input at the transmitter side are random data which are being produced within the MATLAB command language. Then, the random data will be generated in serial format to perform serial to parallel conversion. The serial data stream represents the data information to be transmitted.

The parameters such as number of subcarriers used and the FFT size are using Mobile WiMAX wireless communication standards in OFDMA technology system. Quadrature Amplitude Modulation (QAM) is being used to perform modulation on parallel stream. Each symbol is presented by complex number in phase and quadrature phase vector [12]. The selection of modulation scheme applied to each sub-channel depends solely on the compromise between the data rate requirement and transmission robustness [12]. The samples of the transmitted OFDM signal can be obtained by performing an IFFT operation on the group of data symbols to be sent on orthogonal sub-carriers [13]. The IFFT is used to convert the frequency domain data into time domain signal while maintaining the orthogonality of subcarriers [14]. Cyclic prefix consists of a block of redundant samples at the beginning of each transmitted frame and it is also a cyclic extension of the symbol to eliminate ISI effects on original symbols.

Additive white Gaussian noise (AWGN) channel is the most common channel model but it does not work well due to multipath propagation. It is static in real environment and applied in simulation in the MATLAB software. Practical channel that is being used is Rayleigh channel which may introduce a different

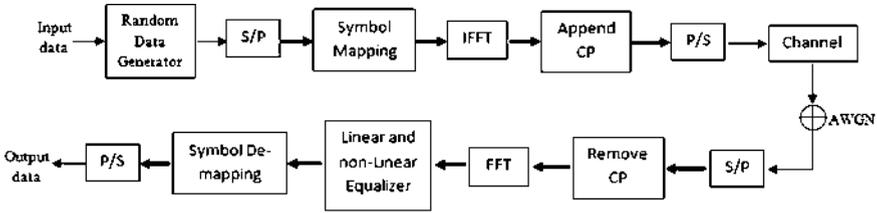


Fig. 25.1 OFDMA system block diagram with equalizer

phase, amplitude attenuation, delay and Doppler shift to the signal. At the receiver part, serial input data is converted to parallel form and the symbol transformation is performed by FFT [12]. The output from the FFT will contain interferences or distortions including ISI. Equalization is implemented to mitigate this ISI. The equalized output is demapped, deinterleaved and then convolutionally decoded to get back original data words [12]. To get the original message information, the data words will then be multiplexed.

25.2.1 Zero Forcing Equalizer

Zero Forcing Equalizer is a form of linear equalization algorithm which applies the inverse of the channel frequency response to the received signal, to restore the signal after the channel in communication system [10]. This form of equalizer was first proposed by Robert Lucky. It has many useful applications. For example, it is applied for IEEE 802.16e (Mobile WiMAX) in MIMO, where knowing the channel allows recovery of the two or more streams which will be received on top of each other on each antenna. The name Zero-Forcing corresponds to bringing down the intersymbol interference (ISI) to zero and will be useful when ISI is significant compared to noise [11].

Figure 25.2 above shows an example to show that there are different parameters used in Zero-Forcing equalization. Let $C_{ZF}(k)$ be the equalizing circuit filter. The LTI filter with transfer function, $C_{ZF}(k)$ is considered to be the ZF equalizer, that can be realized by multiplying the OFDMA received signal as an Eq. (25.1) with the vector $1/H(k)$ which produces the Eq. (25.2). In this case, the equalizer filter compensates for the channel induced ISI as well as the ISI, but this is not eliminating all ISI because the filter is of finite length.

$$y(k) = x(k)H(k) + w(k) + I(k) \tag{25.1}$$

$$C_{zf}(k) = \frac{1}{H(k)} \tag{25.2}$$

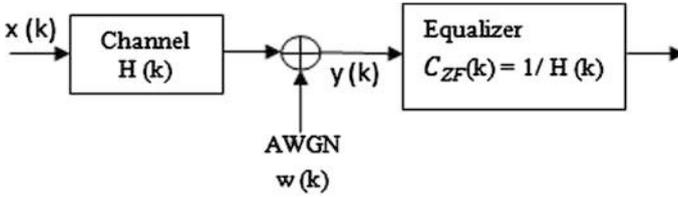


Fig. 25.2 Block diagram of zero-forcing equalizer

where $y(k)$ is complex coefficient envelope of FFT, $x(k)$, $H(k)$ and $w(k)$ are frequency domain equivalents of input data, channel impulse response and AWGN noise respectively.

25.2.2 Minimum Means Square Error Equalizer

A more balanced linear equalizer in this case is the minimum mean-square error equalizer, which does not usually eliminate ISI completely but instead minimizes the total power of the noise and ISI components in the output.

Since this scheme can minimize the mean square error (MSE) between the desired equalizer output and the actual equalizer output, it is found that adaptive MMSE equalizer is an effective and feasible method to mitigate the serious effects of multipath dispersion.

Figure 25.3 above shows a block diagram whereby a few parameters involved which are, $H(k)$ is the channel impulse response, $X_p(k)$ is the actual output while the $Y_p(k)$ is the desired output. The e_k is the error between these two output of the system. In this type of equalizer, the tap weights are chosen as they minimize the mean-square-error (MSE) of all the ISI terms and the noise power at the output of the equalizer. MMSE is the expected value of the squared difference between the desired data symbol and the estimated data symbol. Error between desired and actual output is given by [12].

$$e_k = x_p(k) - Y_p(k)^T W_k \tag{25.3}$$

where W_k is the weight vector of filter. Mean square error is the square of Eq. (25.3) which produces Equation as followed,

$$MSE = E [X_p(k)]^2 + W_k R W_k - 2P^T W_k \tag{25.4}$$

R and P are the correlation and auto-correlation matrices. To minimize ISI we have to find filter weights which minimized when $R = P W_k$. The vector W_k corresponds to the number of taps in the equalizing filter. The equalizer correction term

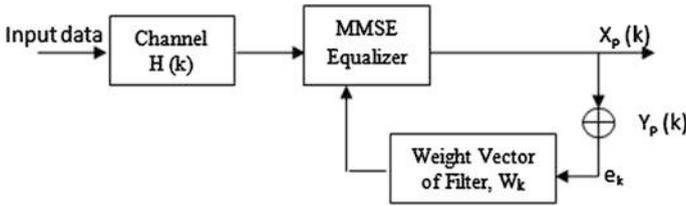


Fig. 25.3 Block diagram of MMSE equalizer

which is the inverse of channel response multiplying with the Eq. (25.1) produces the equalizer output as below,

$$C_{MMSE} = 1/[H(k)+N_o] \frac{1}{H(k) + N_o} \tag{25.5}$$

where the $H(k)$ is the channel impulse response, while the N_o is the noise in the system.

25.2.3 Maximum Likelihood Sequence Estimator

Among of all those equalizers, MLSE exhibits the strongest capability in compensating ISI. However, its main problem is the complexity that increases exponentially with the memory length of the channel. The MLSE Equalizer block uses the Viterbi algorithm to equalize a linearly modulated signal through a dispersive channel. The block processes input frames and outputs the maximum likelihood sequence estimate (MLSE) of the signal, using an estimate of the channel modeled as a finite input response (FIR) filter.

Figure 25.4 above shows the architecture of the MLSE equalizer. The receiver compares the time response with the actual received signal and determines the most likely signals. The problem to be solved is to use the observations $\{r(k)\}$ to create a good estimation of $\{x(k)\}$. In this system, the $r(k)$ is received signal, $h(k)$ denotes the overall channel response of the system, whereas $z(k)$ is the output of the received signal passed to the match filter, and the Viterbi algorithm is obtained by computing the recursive relation iteratively and produce the estimated sequence, $\hat{x}(k)$ which is defined to be sequence of values which maximize the functional [12].

$$C_{MLSE} = p(r|x) \tag{25.6}$$

where $p(r|x)$ denotes the conditional joint probability density function of the observed series $r(k)$ given that the underlying series has the values $x(k)$.

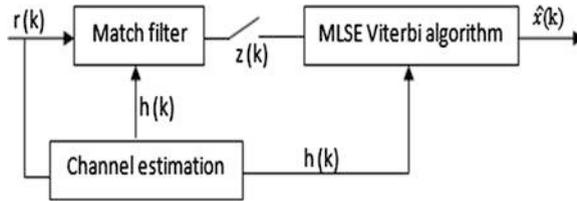


Fig. 25.4 Block diagram of MMSE equalizer

25.3 Simulation Results

One of the important parameter in wireless communication for quality measurement of recovered data is the performance of BER. It is observed that different equalization techniques can give low BER performance under multipath fading environment, but with a slight changes in BER the quality changes many folds [13]. Simulation results are plotted for bit error rate (BER) performance of MIMO-OFDMA system and being compared with and without the implementation of three different equalizers. Besides, BER performance also can be observed when the three different diversity methods used in the system. Table 25.1 shows the system parameters for MIMO-OFDMA for Mobile WiMAX system that have been used in this project [14].

25.3.1 Zero-Forcing, MMSE, MLSE Equalizer

Figure 25.5 above shows that BER performance can be compared between different equalizers. From the simulation, it can be observed that the MLSE equalizer gives the best performance which produces the less interference (ISI) in the system compared to MMSE and Zero-Forcing equalizer. So, as the interference is decreased, the BER is decreased as well but SNR is increased. This is because of MLSE evaluates a sequence of received data samples to determine the most likely correct transmitted sequence. This is proven in the Eq. (25.6), as the most likely transmitted signal, $x(k)$ is increased, the equalizer output, C_{MLSE} is also increased. So, it can successfully minimize the interference in the signal. That is why changes in BER can be seen clearly at high SNR = 17 dB. On the other hand, the MMSE equalizer gives better performance compared to ZF equalizer because it is not only equalizing the channel but also suppressing the noise as in Eq. (25.5) which proves that when the channel impulse response, $H(k)$ and noise, N_o is decreased, the output equalizer, C_{MMSE} will increased. Besides, by applying the ZF equalizer in the system, the performance is improved too as in Eq. (25.2) whereby the channel impulse response, $H(k)$ is decreased, the output equalizer C_{ZF} will increased which tends to produce an increment of SNR, as the interference and BER is reduced.

Table 25.1 MIMO-OFDMA parameter [14]

System bandwidth (MHz)	1.25	2.5	5	10	20
Sampling frequency (MHz)	1.4	2.8	5.6	11.2	22.4
FFT size	128	256	512	1024	2048
Subcarrier spacing (kHz)	10.94				
OFDM symbol duration (μ s)	102.86				
Useful symbol time (μ s)	91.43				
Cyclic prefix (μ s)	11.43				

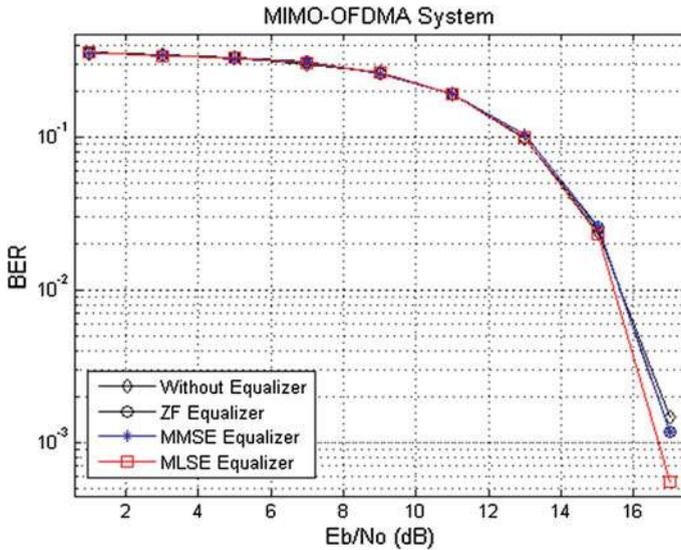


Fig. 25.5 BER comparison using ZF equalizer, MMSE equalizer, MLSE equalizer and without equalizer

But it does not perform as well as the MLSE and MMSE equalizer since this equalizer forces the ISI to zero when only the ISI is significant but does not consider the noise.

25.3.2 STF, ST, SF Diversity

Figure 25.6 above shows the BER performance comparison when using different diversity order system when implementing the MLSE equalizer since this is the best equalizer among the others. It can be seen that by using space, time, frequency diversity (STF), it offers the maximum diversity order compared to space, time

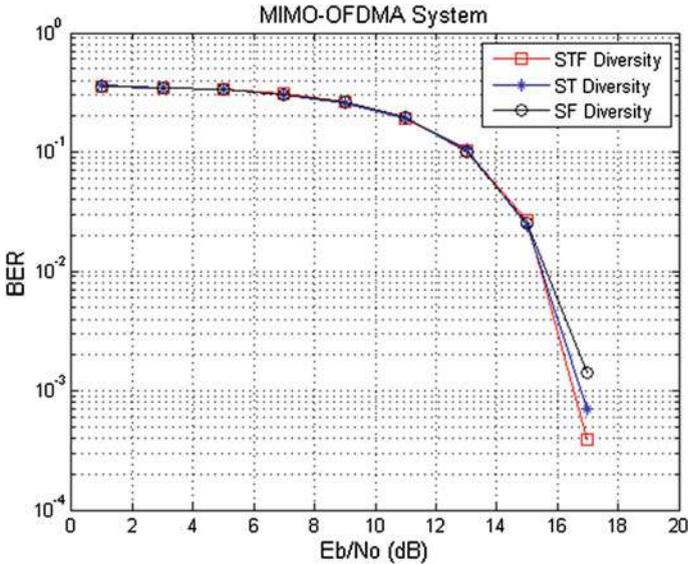


Fig. 25.6 BER performance with three different diversity methods

(ST) diversity and space, frequency (SF) diversity because it can transmit data in different time and frequency slot which improved the performance of the system as the SNR is increased while BER and ISI is decreased. It means that, there is least amount of ISI when STF diversity order is applied in the system. At higher SNR = 17 dB, BER value is at 7×10^{-3} , which gives the least amount of ISI. This is because in the Eq. (25.6), as the most likely transmitted signal, $x(k)$ is increased, the equalizer output, C_{MLSE} is also increased. So, it can successfully minimize the interference in the signal.

25.4 Conclusion

The multipath propagation causes fading of received signal power which leads to ISI. The equalizers are used to improve the distorted received signal caused by ISI. This paper compares the performance of equalized system with the unequalized system to observe which equalizer is the best among three different equalizers that can generate the least ISI in the system and at the same time, it will improve the BER performance within the OFDMA system. From the simulation results, it can be proved that with the equalization system, the ISI can be mitigated, the maximum diversity order can be achieved and the BER performance also can be improved.

Future Work

This project basically applies the basic concept in Mobile WiMAX system. In future works, Quadrature Phase Shift Keying (QPSK) as the modulation scheme can be used instead of using QAM. Moreover, the BER system performance also can be investigated for other types of equalizer such as Decision Feedback Equalizer, Blind Equalizer, and Linear Equalizer.

Acknowledgments This work was supported in part by the Research Management Institute University Teknologi MARA under Excellent fund (Research Intensive Faculty) grant number 600-RMI/DANA 5/3/RIF(86/2012).

References

1. Rhee, W., Cioffi, J.M.: Increase in capacity of multiuser OFDM system using dynamic subchannel allocation. In: Proceedings of the IEEE 51st Vehicular Technology Conference, 2000 (VTC 2000-Spring), Tokyo, pp. 1085–1089 (2000)
2. Morelli, M., Kuo, C.-C., Pun, M.-O.: Synchronization techniques for orthogonal frequency division multiple access (OFDMA): a tutorial review. *Proc. IEEE* **95**, 1394–1427 (2007)
3. Xu, J., Kim, J., Paik, W., Seo, J.-S.: Adaptive resource allocation algorithm with fairness for MIMO-OFDMA system. In: Proceedings of the IEEE 63rd Vehicular Technology Conference, 2006 (VTC 2006-Spring), pp. 1585–1589 (2006)
4. Zheng, L., Tse, D.N.C.: Diversity and multiplexing: a fundamental tradeoff in multiple-antenna channels. *IEEE Trans. Inf. Theory* **49**, 1073–1096 (2003)
5. Idris, A., Syed Yusof, S.K.: Performance evaluation of intercarrier interference self-cancellation schemes for space times frequency block codes MIMO-OFDM system. *IEEE J.* (2008)
6. Love, D.J., Heath Jr, R.W.: Diversity performance of precoded orthogonal space-time block codes using limited feedback. *IEEE Commun. Lett.* **8**, 305–307 (2004)
7. Idris, A., Dimiyati, K., Syed Yusof, S.K., Ali, D.: Diversity technique for multiple input multiple output orthogonal frequency division multiplexing (MIMO-OFDM) system using a new subcarrier mapping scheme. *Int. J. Phys. Sci.* **6**(16), 3879–3884 (2012)
8. Idris, A., Jais, S.M., Salih, N.M., Yusof, A.L., Ali, D.M.: Performance evaluation of STFBC MIMO-OFDM system using different equalization algorithm. *J. Eng. Technol.* **2**(2), 62–67 (2012). ISSN 2231-8798
9. Idris, A., Abdullah, N., Ali, D., Yaacob, N., Mohamad, H.: Reduction of interference with linear equalizer using quarter subcarrier mapping scheme. Paper presented at the IEEE Symposium on Wireless Technology and Applications (ISWTA), (2013)
10. Morelli, M.: Timing and frequency synchronization for the uplink of an OFDMA system. *IEEE Trans. Commun.* **52**, 296–306 (2004)
11. Mark, J., Zhuang, A.W.: *Wireless Communications and Networking*. Prentice Hall, New Jersey (2003)
12. Sharma, P.: Performance analysis of zero-forcing equalizer for ISI reduction in wireless channels. *Int. J. Eng. Res. Technol. (IJERT)* **1**, 3 (2012)
13. Gupta, M., Nema, R., Mishra, R.S., Gour, P.: Bit error rate performance in OFDM system using MMSE & MLSE equalizer over Rayleigh fading channel through the BPSK, QPSK, 4 QAM & 16 QAM modulation technique
14. Rajesh Goel M.S., Bansal, P.K.: On importance of bit error rate in wireless communication
15. Wang, F., Ghosh, A., Sankaran, C., Fleming, P., Hsieh, F., Benes, S.: Mobile WiMAX systems: performance and evolution. *IEEE Commun. Mag.* **46**, 41–49 (2008)

Chapter 26

Performance Analysis of Polling Delay in Transparent and Non-transparent Multi-hop Relay WiMAX Network

Mohd Daud A. Hassan, Habibah Hashim and D.M. Ali

Abstract The relay task group has extended the IEEE 802.16e-2005 to a new standard known as IEEE 802.16j Mobile Multi-hop Relay (MMR). The deployment of relay station (RS) in WiMAX network will overcome the increase in cost and improve the economic viability of the system. Furthermore, the MMR standard also addresses the problem of limited spectrum, low SINR at the cell edge and coverage hole of the previous standard. However, the introduction of relay may cause longer delays and degrade the performance of resource efficiency. In this paper, we investigate the performance analysis of polling delay for transparent and non-transparent relay modes in MMR WiMAX. We derived and analyzed the average polling delay of both modes using M/M/1 and tandem queue respectively. Numerical result has shown that the average polling delay decreases when the number of poll increases for both relay modes. Moreover, the delay for non-transparent relay mode is higher due to the signals overhead.

26.1 Introduction

The IEEE 802.16j standard was introduced to support multi-hop operation and relay station (RS) relay packets between a base station (BS) and subscriber station/mobile station (SS/MS). In mobile multi-hop relay (MMR) WiMAX system, the

M.D.A. Hassan (✉)

Faculty of Electrical Engineering, Universiti Teknologi MARA, 13500 Penang, Malaysia
e-mail: mohddaud106@ppinang.uitm.edu.my

H. Hashim · D.M. Ali

Faculty of Electrical Engineering, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia
e-mail: habib350@salam.uitm.edu.my

D.M. Ali

e-mail: darma504@salam.uitm.edu.my

framing information of MS is allowed to route through intermediate RSs to reach the BS, which differs from the single hop. There are benefits of the introduction of relay in WiMAX network as compared to conventional repeater in a cellular system such as easy network deployment, and reduction in infrastructure cost. A relay does not need any connection to the access network wired backhaul hence can be deployed to a location where the wired connection does not exist or when it is complicated to install. Furthermore, throughput and capacity enhancement can be achieved. The relay that is used in the IEEE 802.16j standard is also inadvertently compatible with the IEEE 802.16e for SS/MS.

The history of IEEE 802.16j was created in March 2006 [1] and the first technical contribution was made 8 months later. In order to support mobile multi-hop relay specification, mesh mode was removed from the IEEE 802.16 -2009 standard. The specification is an amendment of the IEEE 802.16e standard for achieving throughput enhancement and coverage extension.

A study on the performance of the IEEE 802.16j network has been conducted from various aspects. An overview of the relay based technology which focuses on modes of operation, framing structure and network entry procedure is described in [2]. In [3], the paper investigates the impact of the location of the RS and the possibility of RS to increase the network capacity. In addition, the author analyzed the signaling that arises within such system and compared it with a single hop implementation. In [4], an adaptive polling approach is proposed for the real-time Polling System (rtPS) in order to improve on the average delay. However, the author has performed the simulation in a standard conventional WiMAX environment without considering the multi-hop relay. The performance analysis of both relay modes in MMR is demonstrated in [5], and yet the analysis is limited to sustain the rate of uplink transmission. The adaptation to the traffic pattern was conducted by Nie et al. [6] which satisfied the delay constraints of most real-time applications. However this approach only performed the multi-hop bandwidth request mechanism in terms of overall spectrum efficiency. In [7], the authors investigated the usage of RS in improving the per-user throughput. The investigation of the Non-Transparent relay mode has also been discussed in [8] and the simulation results have shown an improvement in the transmission delay. To the best of the author's knowledge, there is no current research that specifically examines the queuing analysis of polling based on multi-hop WiMAX for both relay operation modes.

We present the analysis of the polling based on MMR WiMAX using Markov chain which focuses on the delay centralized scheduling for transparent and non-transparent modes. We investigated the polling delay affected by the numbers of poll (NOP) and the total numbers of MS and only considered unicast polling mode in the analysis.

26.2 Relay Technologies

The standard of IEEE 802.16j, defines 2 operation modes of RS which differ in terms of the usage. These modes are known as the transparent and non-transparent relay modes. The difference between both modes is basically on how the framing information is transmitted. Generally, in transparent mode operation, the relay does not transmit/forward frame header information. The frame header contains essential scheduling information (such as preamble, frame control header (FCH), and media access protocol (MAP) and the nodes used to determine when they can transmit and receive information [2]. However in the non-transparent mode the RS generates their own framing information and forward those provided by the BS (depending on scheduling approach).

There are also two different modes of resource allocation which are centralized and distributed. For centralized, BS determines the scheduler and the scheduling packets are transmitted in a collision freeway within the scheduling control sub-frame and the transparent relay mode operates using this type of scheduling. While in distributed mode, RSs have some autonomy and can make scheduling decision for the nodes which they communicate with. Thus, the non-transparent relay has a choice between distributed or centralized modes to operate based on the scheduling approach.

The frame structure of multi-hop relay consists of downlink (DL) and uplink (UL) sub-frame, where it includes at least one access zone (AZ) and may include one (transparent zone, TZ) or more relay zone (RZ). This structure enables RS to operate in either transmit or receive mode. For the polling purpose, BS polls individual MSs by using UL-MAP in the DL sub-frame.

Figure 26.1 shows the frame structure of a two-hop transparent relay which contains the DL sub-frame (AZ and TZ) on the left side, and UL sub-frame (AZ and RZ) on right side of the frame. The RS serving can decode the control information of the BS and is not required to transmit control information themselves. This type of serving can be utilized to decode the central information in order to achieve higher throughput. Figure 26.2 shows the non-transparent relay frame structure which for the non-transparent mode contains DL sub-frame (AZ and RZ), and UL sub-frame (AZ and RZ). Meanwhile, due to the fact that the relay serving MSs cannot decode the BS control information, they must transmit control information themselves.

In transparent relay mode, AZ for DL sub-frame, BS sent out data burst to RS and MS, respectively. Similarly in TZ, RS forward data burst received from BS to MSs. The DL sub-frame starts with preamble which is used by the PHY layer for synchronization and equalization followed by the frame control and data section. The UL transmission sub-layer frame operation is identical to that of DL. The description of the frame structure in transparent and non-transparent relay can be referred in [9]. We focus on the UL sub-frame which is divided into 3 parts: initial ranging, bandwidth request and data transmission period [10].

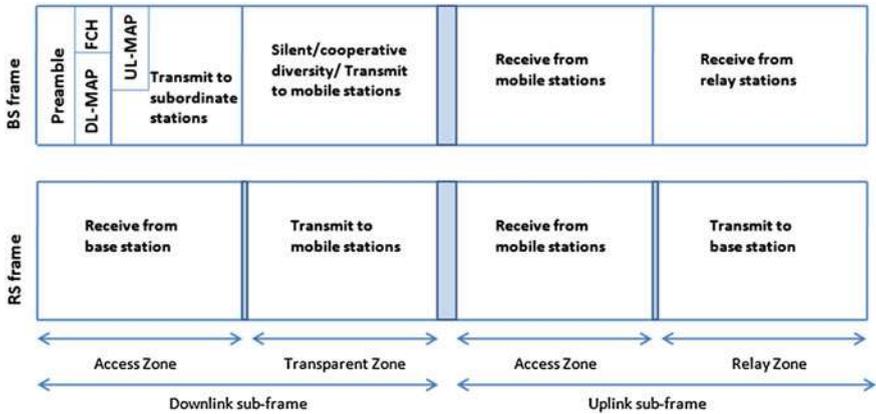


Fig. 26.1 A transparent frames structure for both the BS and RS

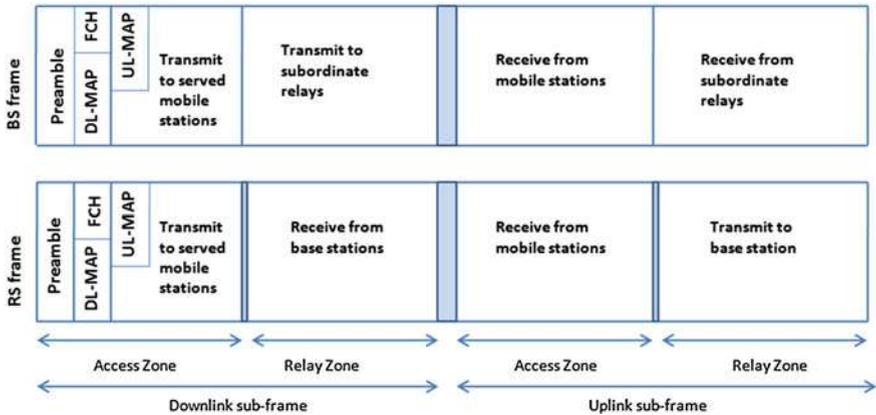


Fig. 26.2 A non-transparent frames structure for the BS and RS

Assumed that for zone of UL (AZ and RZ) are combined as depicted in Fig. 26.3. In standard WiMAX network, each MS that has data to be sent out must request for bandwidth which is provided by the BS. The MS will request the bandwidth based on the corresponding service class (e.g., data, voice, video or real time services) by replying to the polling message from the BS before the data uplink transmission. The BS will accept the bandwidth request if resources are sufficient enough. The information about polling control is stored in the UL-MAP within the downlink sub-frame which includes uplink channel identifier, uplink channel descriptor, number of information elements to map, allocation start time and map information elements. Therefore in multi-hop relay WiMAX network polling plays an important role to accomplish an efficient performance particularly those that involve the transmission of data and video.

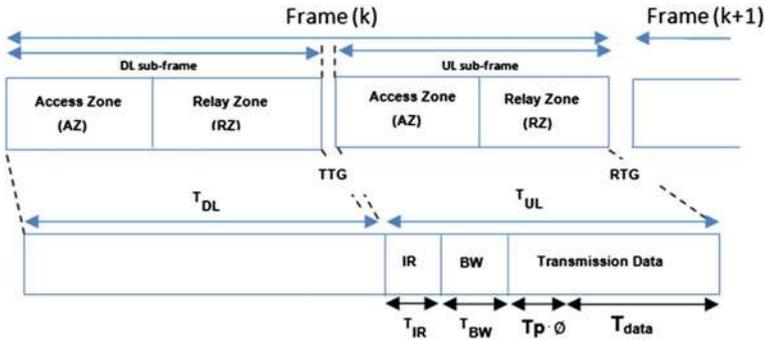


Fig. 26.3 Generic frame structure

The analysis uses the following notations;

- T_f frame duration time
- $T_{f_{BS}}$ base station frame duration time
- $T_{f_{RS}}$ relay station frame duration time
- T_{DL} downlink sub-frame duration time
- T_{UL} uplink sub-frame duration time
- T_{IR} initial ranging period for uplink sub-frame
- T_{BW} bandwidth request period for uplink sub-frame
- T_P polling period for MSs to send bandwidth request
- N_1 and N_2 numbers of MS in the two system in the steady state
- T_{ave} average polling delay
- R data rate
- P_t packet transmission probability
- ϕ no. of polls within a frame period

26.3 Network Model Analysis

This part presents the analysis of a transparent and non-transparent polling for centralized scheduling. The number of MSs served in the network is the parameter that affects the network performance. As afore mentioned, the signal from BS are not forwarded to the RS in the transparent mode. Thus, the UL-MAP will be transmitted directly to MS. We model the polling based of the multi-hop relay WIMAX network using the Markov chain model as in [10], where the number of served MSs is the state of model. The analysis of the model is a birth-death process as shown in Fig. 26.4, with a birth rate of λ_n and death rate of μ_n .

By using the global balance equations, derived for steady-state solution of an irreducible, homogeneous continuous Markov chain;

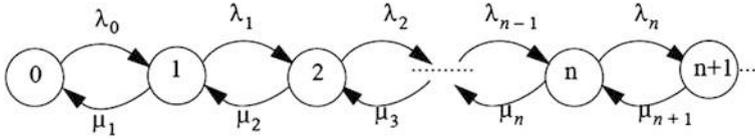


Fig. 26.4 State transition diagram of the number of MSs in the MMR network

$$\sum_{k \neq j} p_k q_{kj} - q_j p_j = 0 \tag{26.1}$$

and combine with the equation

$$\sum_j p_j = \mathbf{1} \tag{26.2}$$

where p_j is the steady-state limiting probability of the system being in state j and q_{kj} and q_j are the transition rates. Then applying this to birth-death

$$-(\lambda_j + \mu_j)p_j + p_{j+1}\mu_{j+1} + p_{j-1}\lambda_{j-1} = 0 \tag{26.3}$$

$$-\lambda_0 p_0 + \mu_1 p_1 = 0 \tag{26.4}$$

Equation (26.4) is the balance equation derived for the birth-death as illustrated in Fig. 26.4. Assumed that the arrival rate is the same and denoted as λ and the service rate is also the same and denoted as μ . The number of MSs in a system of an M/M/1 queue is a homogeneous, irreducible birth-death in which $\lambda_i = \lambda$ for \forall_i and $\mu_i = \mu$ for \forall_i . Furthermore, traffic intensity of $\rho = \lambda/\mu$ and $\rho < 1$ is defined for a stable system. The mean number of MSs in the system in the steady-state

$$\begin{aligned} E[N] &= \sum_{n=0}^{\infty} n p_n \\ &= \sum_{n=0}^{\infty} n \rho^n (1 - \rho) \\ &= \rho(1 - \rho) \sum_{n=0}^{\infty} n \rho^{n-1} \\ &= \rho(1 - \rho) \frac{1}{(1 - \rho)^2} \\ &= \frac{\rho}{(1 - \rho)} \end{aligned} \tag{26.5}$$

The average polling delay of multi-hop WiMAX depends on the number of polls (NOP), ϕ . Every MS is polled by the BS in order for the MS to send its packet

which is then, mapped in the UL-MAP. Subsequently after receiving the poll, MS will request for bandwidth through the slot in the uplink sub-frame. Thus, the average polling delay of MMR WiMAX for transparent mode is

$$\begin{aligned}
 T_{ave} &= T_f \cdot \frac{1}{\emptyset} \cdot E[N] \\
 &= (T_{DL} + T_{UL}) \cdot \frac{1}{\emptyset} \cdot E[N] \\
 &= (T_{DL} + T_{IR} + T_{BW} + \emptyset T_p + T_{data}) \cdot \frac{1}{\emptyset} \cdot \left[\frac{\rho}{1 - \rho} \right]
 \end{aligned} \tag{26.6}$$

For the non-transparent relay mode, the signal received from BS is forwarded by the RS to MS. Therefore we model the polling based of a multi-hop relay WiMAX as a two-stage tandem queue in queuing system. This tandem refers to an arrangement of objects in which they are lined up one behind the other, all facing the same direction and multiple job classes, one after another and an arriving packet undergoes each job class before leaving the system [11]. Figure 26.5 depicts the two stage queue in a tandem system.

The arrival of the first stage is define as λ while in the second stages, it's arrival is the of the departure of the MS from the first stage. The state of the first stage ($N_1 = n_1$) is independent from the previous sequence of departure and it is these departure that defines the arrivals to the second system. Therefore if N_1 and N_2 represent the numbers of served MSs in the two systems in then steady state

$$p(n_1, n_2) = \Pr(N_1 = n_1, N_2 = n_2) = \Pr(N_1 = n_1) \Pr(N_2 = n_2) \tag{26.7}$$

thus from result for the M/M/1 queue,

$$\begin{aligned}
 p_n &= (1 - \rho)\rho^n \\
 p(n_1, n_2) &= (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2}; n_1, n_2 \geq 0
 \end{aligned} \tag{26.8}$$

where $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$ and $0 < \rho_1, \rho_2 < 1$
by summing over n_1 , obtained that

$$\Pr(N_2 = n_2) = (1 - \rho_2)\rho_2^{n_2} \text{ and } E[N_2] = \frac{\rho_2}{1 - \rho_2} \tag{26.9}$$

Similarly,

$$\Pr(N_1 = n_1) = (1 - \rho_1)\rho_1^{n_1} \text{ and } E[N_1] = \frac{\rho_1}{1 - \rho_1} \tag{26.10}$$

and then $E[N_1 + N_2]$

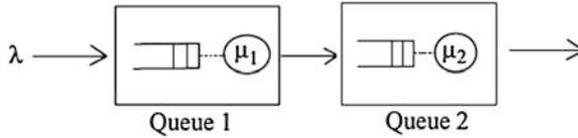


Fig. 26.5 Two stages Markovian queue in tandem

$$\begin{aligned}
 &= \sum_{n_1} \sum_{n_2} (n_1 + n_2)(1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2} \\
 &= \sum_{n_1} n_1(1 - \rho_1)\rho_1^{n_1} \cdot \sum_{n_2} (1 - \rho_2)\rho_2^{n_2} + \sum_{n_2} n_2(1 - \rho_2)\rho_2^{n_2} \cdot \sum_{n_1} (1 - \rho_1)\rho_1^{n_1} \\
 &= \frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2} \\
 &= E[N_1] + E[N_2]
 \end{aligned} \tag{26.11}$$

As described in Sect. 26.2, the RS received the signal from the BS and forward it to MS. Therefore the equation is derived based on the operation of the transparent mode by doubling the time frame to represent the two hops. Thus, the equation for an average polling delay for a non-transparent network is

$$\begin{aligned}
 T_{ave} &= T_f \cdot \frac{1}{\emptyset} \cdot E[N_1 + N_2] \\
 &= (T_{f_{BS}} + T_{f_{RS}}) \cdot \frac{1}{\emptyset} \cdot \left[\frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{1 - \rho_2} \right]
 \end{aligned} \tag{26.12}$$

Both average polling delays show the inverse to the number of polls. This will result in, reduced in the average polling delay when the number of polls in the system is increased. However, it will increase the bandwidth waste of polling if the MS has no packet to send.

26.4 Numerical Result

The number of served MS in MMR WiMAX is a part of the important parameter that affects the system. The BS polls the MS in every frame period and this will cause an increase in polling delay.

In this analysis, the MSs are assumed to arrive at the MMR WiMAX network using Poisson distribution with arrival rate of $\lambda = 0.3$. The ratio for the DL:UL is 2:1. Table 26.1 summarizes the parameters for the analysis. Figures 26.6, 26.7, and 26.8 illustrates the result for MMR WiMAX based on the M/M/1 and tandem queue. Figure 26.6 shows the average polling delay for transparent and non-transparent mode when the numbers of polls (NOP) (ranging from 1 to 11) are varied.

Table 26.1 Parameter for analysis

Parameter	Values
Number of polls (NOP), \emptyset	1 ~ 20
Numbers of mobile station (MS)	1 ~ 100
Frame periods, T_f	5 ms
Data rates, R	50 Mbps
Total number of slots in frame, L_f	630
Number of slots in DL sub-frame, L_{DLf}	420
Number of slots in UL sub-frame, L_{ULf}	210
Polling bandwidth slots	3
Required bandwidth slots	32
MS mean service time, μ	0.5

Fig. 26.6 Average polling delay versus no. of polls

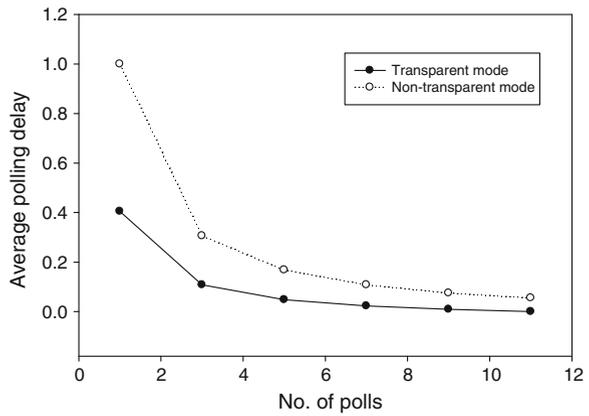


Fig. 26.7 Average polling delay versus no. of MS (for transparent)

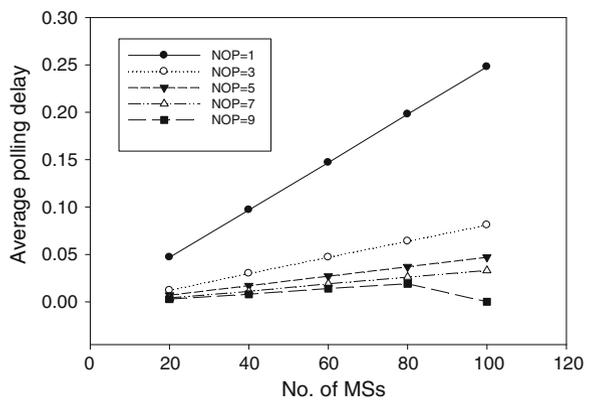
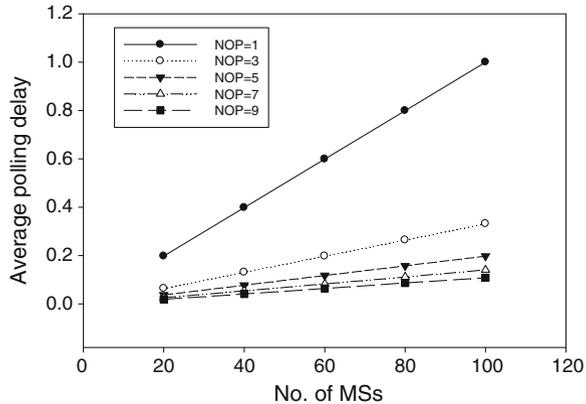


Fig. 26.8 Average polling delay versus no. MS (for non-transparent)



The average polling delay shows a decrease trend when the NOP increases for both modes. The transparent mode has less delay than the non-transparent one. This is because the decision is only done by the BS whereas the non-transparent mode has to undergo two decisions which are BS and RS. The average polling delay of transparent mode and non-transparent mode with increasing number of MS are shown in Figs. 26.7 and 26.8 respectively. The delay will vary in line with the increased of the NOP and the no. of MS. When NOP is increased, the average polling delay for the transparent mode will decrease and the same goes for non-transparent mode.

However, the delay for the non-transparent mode is higher than the transparent mode due to the signaling messages received by the RS from the BS which is then forwarded to the MS. This is proven by Eq. (26.12). Although the average polling delay increases with the number of MS, it decreased significantly for both modes when the NOP is increased from 1 to 3. For instance, when MS = 100, the average polling delay decreased approximately to 70 %. However there is a slight changes to the average polling delay when the NOP is increased.

This paper presents a queuing model to evaluate the delay in the polling based operation of MMR WiMAX IEEE802.16j network. The average polling delay for both modes show the inversed to the number of polls. As a result, the polling delay is reduced if more polls in the system are created. However, this will result in waste of bandwidth if the MS has no packet to send. For future direction we will validate the mathematical analysis derive with the results of the simulation. We will also focus on optimizing the number of polls by the BS in the network while reducing the average polling delay and increasing the throughput of the system.

Acknowledgments This research was supported by the Universiti Teknologi MARA (UiTM) under Exploratory Research Grant Scheme (ERGS).

References

1. Peters, S.W., Heath, R.W.: The future of WiMAX: multihop relaying with IEEE 802.16j. *IEEE Commun. Mag.* **47**(1), 104–111 (2009)
2. Genc, V., Murphy, S., Yu, Y., Murphy, J.: IEEE 802.16j relay-based wireless access networks: an overview. *IEEE wirel. Commun.* **15**(5), 56–63 (2008)
3. Genc, V., Murphy, S., Murphy, J.: Performance analysis of transparent relays in 802.16j MMR networks. In: 6th International Symposium on Modeling and Optimization in Mobile, Adhoc and Wireless Network and Workshops, pp. 273–281. (2008)
4. Chang, B.J., Chou, C.M.: Adaptive polling algorithm for reducing polling delay and increasing utilization for high density subscribers in WiMAX wireless networks. In: 10th IEEE Singapore International Conference on Communication Systems, pp. 1–5. (2006)
5. Yusoff, R., Baba, M.D., Rahman, R.A., Ibrahim, M., Mat Isa, N.: Performance analysis of transparent and non-transparent relays in MMR WiMAX networks, In: IEEE Symposium on Industrial Electronics and Application (ISIEA), pp. 237–240. (2011)
6. Nie, C., Korakis, T., Panwar, S.: A multi-hop polling service with bandwidth request aggregation in IEEE 802.16j networks, In: IEEE Conference on Vehicular Technology, pp. 2172–2176. (2008)
7. Izza, W.N., Baba, M.D., Ali, D.M.: Performance study on relay station usage in IEEE 802.16j mobile multi-hop relay network. In: IEEE Symposium on Computer Application and Industrial Electronics (ISCAIE), pp. 218–223. (2012)
8. Sayenko, A., Alanen, O., Martikainen, H.: Analysis of the non-transparent in-band relays in the IEEE 802.16 multi-hop system. In: IEEE Wireless Communication and Networking Conference, pp. 1–6. (2010)
9. Mohd Daud, A.H., Habibah, H., Darmawaty M.A., NurHidayat A.: A queueing analysis of polling based for mobile multihop relay WiMAX networks. In: IEEE International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), pp. 322–327. (2013)
10. Chang, B.J., Chou, C.M., Liang, Y.H.: Markov chain analysis of uplink subframe in polling-based WiMAX networks. *Comput. Commun. J.* **31**(10), 2381–2390 (2008)
11. Ng, C.H., Soong B.H.: *Queueing Modeling Fundamentals with Application in Communication Network*, 2nd edn. Wiley, New York (2008)

Chapter 27

Blind Source Computer Device Identification from Recorded Calls

Mehdi Jahanirad, Ainuddin Wahid Abdul Wahab
and Nor Badrul Anuar

Abstract This study investigates the use of blind source computer device identification for forensic investigation of the recorded VoIP call. It was found that a combination of mel-frequency cepstrum coefficients (MFCCs) and entropy as an intrinsic audio feature captures the specific frequency response due to the tolerance in the nominal values of the electronic components associated to individual computer device. By applying the supervised learning techniques such as naïve Bayesian, linear logistic regression, neural networks (NN), support vector machines (SVM) and sequential minimal optimization (SMO) classifier to the Entropy-MFCC features, state-of-the-art identification accuracy of above 99.8 % has been achieved on a set of 5 iMacs and 5 desktop PCs from the same model.

27.1 Introduction

Audio forensics mostly focus on situations that impose trust in authenticity and integrity of audio signals. An example for these scenarios is forensic acquisition, analysis and evaluation of admissible audio recording as crime evidence in court. Authenticity of audio evidence is important as part of a civil and criminal law enforcement investigation or as part of an official inquiry into an accident or other civil incidents. In these processes, authenticity analysis determines whether the recorded information is original, contains alterations, or has discontinuities

M. Jahanirad (✉) · A.W.A. Wahab · N.B. Anuar
Faculty of Computer Science and Information Technology, University of Malaya,
50603 Kuala Lumpur, Malaysia
e-mail: mehdijahanirad@siswa.um.edu.my

A.W.A. Wahab
e-mail: ainuddin@um.edu.my

N.B. Anuar
e-mail: badrul@um.edu.my

attributed to recorder stops and starts. Authenticity evaluations by using the device-based techniques are attracting widespread interest in fields such as: (a) identification of computer-generated audio from the original audio recording file [1]; (b) identification of the source brand or model of the recording devices, such as telephone handsets, microphones [2, 3], and cell phones [4]; (c) identification of the speech codecs [5]. The most related works to this approach are the recording source forensics.

Kraetzer et al. [6] published the first practical evaluation to identify a device. They adopted the statistical pattern recognition technique to determine the source microphone and its recording environments. Buchholz et al. [7] focused on microphone classification through the histogram of Fourier coefficients and extracted the coefficients from near-silent frames to capture microphone properties. With the use of a suitable context model for microphone forensics, Kraetzer et al. [8] extended the works in [6] by reevaluating their sample data separately. Kraetzer et al. [9] extended the proposed context model in [8] toward better generalization and constructed a new application scenario model for microphone forensic investigations with the aim of detecting playback recordings. In addition to microphone identification, Garcia-Romero and Espy-Wilson [2] extended device identification by identifying landline telephone handsets. A similar landline telephone handset identification method proposed by Panagakos and Kotropoulos [3] improved the accuracy of the identification by using sparse representation classifier (SRC). Haniłçi et al. [4] proposed a cell phone identification method that uses SVM by identifying the audio source based on the cell phone brand and model. However, these approaches lack sufficient study to eliminate convolution by speech context. Moreover, they only focused on identifying source recording devices.

This paper focuses on the novel idea of recognizing the communicating acquisition devices based on the recorded call. We present a case study that records calls by using stationary Notebook that makes VoIP calls to computer devices of the same model. The motivation for this case study is based on the fact that the combination of anonymity, ease of access and free offerings of VoIP services provide fertile ground for criminal activity. Thus, identifying the computer devices used in VoIP call can help the forensic investigator to reveal useful information in addition with authenticity of the recorded call. This study identifies source brand and model of the computer devices from recorded calls based on entropy-mel-frequency cepstrum coefficient (MFCC) features. Extracting the entropy of MFCCs adds an advantage to the application of MFCCs in [4] by eliminating the effects of speech contents. The method evaluates the feasibility of entropy-MFCC features and its robustness against speech signals by using classifier benchmarking.

The remainder of this paper is organized as follows: Sect. 27.2 discusses an overview of the methodology. Section 27.3 outlines the recording setup. Section 27.4 describes the experiments and evaluates the performance of the proposed method. Finally, Sect. 27.5 discusses further implications of the practical study, its limitations, and future applications.

27.2 Source Computer Device Identification Scheme

In our implementation we followed [10] by using audio mining techniques: (a) creates blocks through preprocessing the recorded samples, (b) determines intrinsic computer device fingerprint through feature extraction and (c) uses supervised learning techniques known as classification.

27.2.1 Preprocessing

Preprocessing method uses two different approaches as illustrated by Fig. 27.1 to create blocks from: (a) original speech signal, and (b) near-silent segments. This is to justify the robustness of the proposed method against speech signals.

Original Speech Signal. The pre-processing stage includes sampling, framing, windowing and cleaning the signal. The signals become more distinct when noise is eliminated. Thus, we adopted cleaning to remove the noise generated by environment. The algorithm splits clean signals into overlapping audio frames of length 40 samples. In other words the output is a matrix with 40 columns, when each column represents one block for feature extraction.

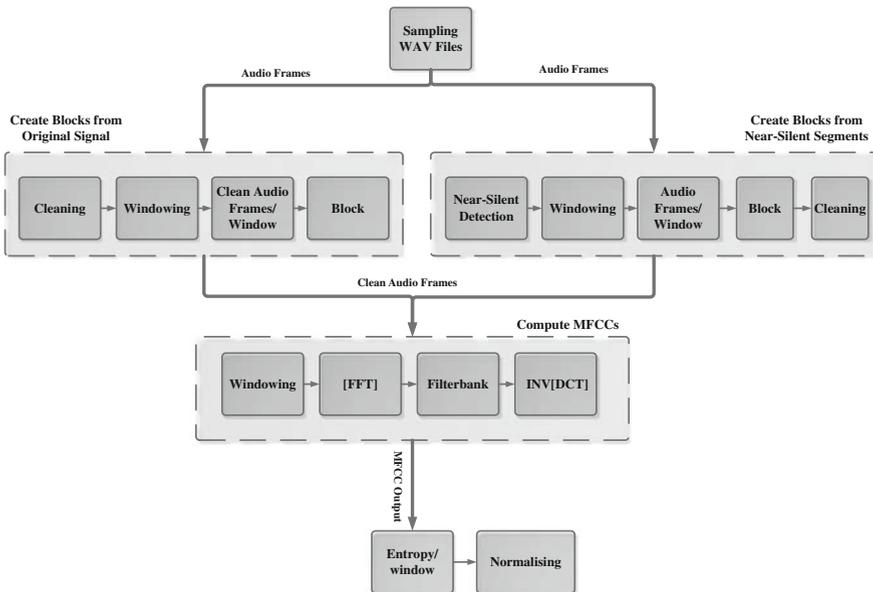


Fig. 27.1 Flow chart of the proposed feature extraction

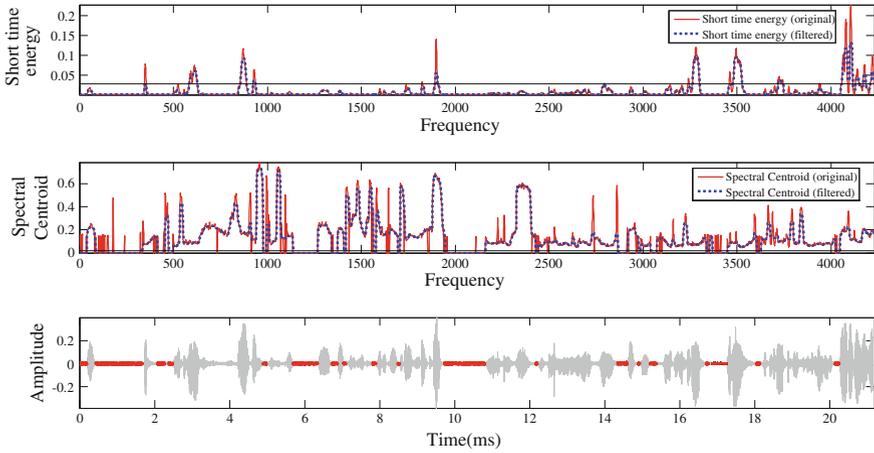


Fig. 27.2 Visualization of Near-Silent Detection algorithm. *Top* and *middle* plots represent the histogram of short time energy and spectral centroids of the signal, respectively. The *horizontal line* shows the estimated threshold. The *bottom* plot shows the spectrum of the original signal, when *red* color determines the near-silent segments

Near-Silent Segments. This approach uses simple segmentation of the recorded signal in order to extract the near-silent segments as demonstrated in Fig. 27.2. We implemented the near-silent detection algorithm based on the silent removal approach in [11], however our objective is to select the silent segments instead of removing them. This method uses two audio features known as signal energy and spectral centroid [12]. This algorithm includes following steps: (a) extracts two feature sets from the original recorded signal, (b) computes the histogram of the feature series values, (c) applies the smoothing filter on histograms, (d) estimates two thresholds for each histogram, (e) applies a simple thresholding criterion to the feature sets, (f) and detects silent segments based on the assigned criterion. Finally, the algorithm assigns equal number of samples from near silent segments into blocks, and then performs cleaning on each block prior to feature extraction.

27.2.2 Intrinsic Computer Device Fingerprints

Assuming that the computer device is a linear time-invariant system, its influence on the recorded call is modeled by the convolution of its impulse response and the original speech. The convolution means the spectrum of any recorded speech segment is the product of the spectrum of the original speech signal and the device frequency response. Well known MFCC features are selected to capture the device frequency response because of the fact that the convolution in time domain is

represented by the summation in cepstrum domain. Thus, MFCCs produce inherent invariance toward linear spectral distortions. Furthermore, the entropy of MFCCs reduces the dimensionality of the feature space. According to information theory, this value increases for silent segments that contain uncertainty and reduces for the speech segments that contain information [13]. As a result this paper computes Entropy-MFCC features as the intrinsic computer device fingerprints through three stages: (a) computes the MFCCs, (b) computes the entropy of MFCCs, and (c) normalizes the output.

MFCCs. The complex cepstrum of the signal is defined as the Fourier transform of the log of the signal spectrum. The signal was transferred to a linear frequency scale by using fast Fourier transform and was converted to the mel frequency scale by using a filter bank. Finally, the MFCCs were determined based on the analysis of the short-time mel frequency log spectrum. This analysis includes computing the inverse discrete cosine transform (DCT) of the log spectrum of the signal. The mel-cepstrum output consists of N frames and 12 coefficients.

Entropy. The entropy of the MFCC vectors was computed in two stages. First, the spectrum was normalized into the probability mass function (PMF), where c_{ld} is the mel-cepstrum coefficient d in frame l and $p(c_{ld}) = p_d$ is the PMF of the signal. In the second stage, the entropy \mathcal{H} was computed by using

$$\mathcal{H}_d = - \sum_{l=1}^N p_d \log_2 p_d. \quad (27.1)$$

Overall, 12 entropy-MFCC features were extracted by using MATLAB functions.

Normalization. The last and one of the most important steps in preparing the features is normalization. This step reduces large differences between the maximum and minimum data values.

27.2.3 Supervised Learning Techniques

Supervised learning techniques known as classification problems can be implemented through different learning algorithms, such as statistical modeling, linear, non-linear, and ensemble learning models. However, to determine which approach is the most efficient for a particular problem, evaluation metrics are required. We selected five simple classifiers based on the fact that the simplicity-first methodology is the best choice for analyzing practical datasets [14]. Furthermore, these algorithms demonstrated high performances for similar works in the literature. Table 27.1 details the employed classifiers in the experiment and their specifications.

Table 27.1 Classification algorithms used in the experiment

Classification algorithms	Specifications
Naïve Bayesian	Works based on Bayes’ rule of conditional probability that assumes independence
Linear logistic regression	Measures goodness of fit by using the log-likelihood of the model
Neural network	Primarily learns the network structure and the connection weights by fixing the network structure to determine the weights
Support vector machine	Uses the LibSVM wrapper that implement a multi-class SVM classifier with a radial basis function (RBF) kernel [15]
Sequential minimal-optimization	Implements the sequential minimal-optimization algorithm for training a SVM classifier using a Polynomial kernel

27.3 Recording Setup

The implemented setup enabled recording of VoIP communication between computer devices and the single stationary inside faculty building. The stationary user made Skype call to computer user, then recorded signals in mono with WAV format by using Pamela for Skype-Version 4.8 [16]. The setup collected recordings with respect to computer devices including: (a) five iMacs of identical model located in Multimedia Research Lab, (b) five desktop PCs of identical model located in Micro Lab. At the same time the stationary was located inside the open corridor. All conversations were conducted between the same male and female over the experiment. Table 27.2 indicates the specifications for computer devices that employed in the setup. All iMacs were 21.5-inch, Late 2012 model built with 3.1GHz Intel Core i7 processor, stereo speakers, and dual microphones. All desktop PCs were Lenovo ThinkCentre M81 7518 model with 3.3 GHz Intel Core i3 processor that attached to the same external microphone. The desktop PCs use 32 bit, Windows 7 professional Service Pack 1 as operating system.

Table 27.2 iMac and desktop PC devices and class names

iMac			Desktop PC	
Serial no.	Software no.	Class	Serial no.	Class
C02JWXXJDXXX	OS X 10.8.4 (12E55)	i1	XXXXXXXXXXXXX3D6B	PC1
C02JWXXDDXXX	OS X 10.8.5 (12F45)	i2	XXXXXXXXXXXXX3DMK	PC2
C02JWXXGDXXX	OS X 10.8.2 (12C3103)	i3	XXXXXXXXXXXXX3AME	PC3
C02JWXXEDXXX	OS X 10.8.3 (12D78)	i4	XXXXXXXXXXXXX3CGM	PC4
C02JWXXKDXXX	OS X 10.8.2 (12C3103)	i5	XXXXXXXXXXXXX2ZRB	PC5

27.4 Experiments and Results

The experiments evaluated the feasibility of the source computer device identification method through features that extracted from original speech signal against features that extracted from near silent segments. Moreover the performance of the proposed scheme was evaluated for inter and intra-model device identification using five classification algorithms that implemented in data mining tool Weka Version 3.6 [17].

27.4.1 Experiment on Original Speech Sample

The total of 1,320 and 1,400 blocks were extracted from original speech samples collected from each iMac and desktop PC devices, respectively. The first part of the experiment used the histogram of all 12 features to visualize the discriminatory effect of entropy-MFCC features among five identical iMacs and desktop PCs, as demonstrated in Fig. 27.3. For further investigation we selected three different pair of iMacs and examined the squared Euclidean distance between their entropy-MFCC feature vectors in Fig. 27.4. The result of this measurement indicates the considerable distances between feature vectors corresponding to pair of iMacs, therefore justifying the effectiveness of entropy-MFCC features in differentiating individual iMac devices. The second part of the experiment employed classification benchmarking to evaluate the classification performance among both desktop PCs and iMac devices of identical models.

Table 27.3 indicates the results of three sets of experiments on original speech signal with all five classifiers, 10-fold cross-validation and default parameters. For

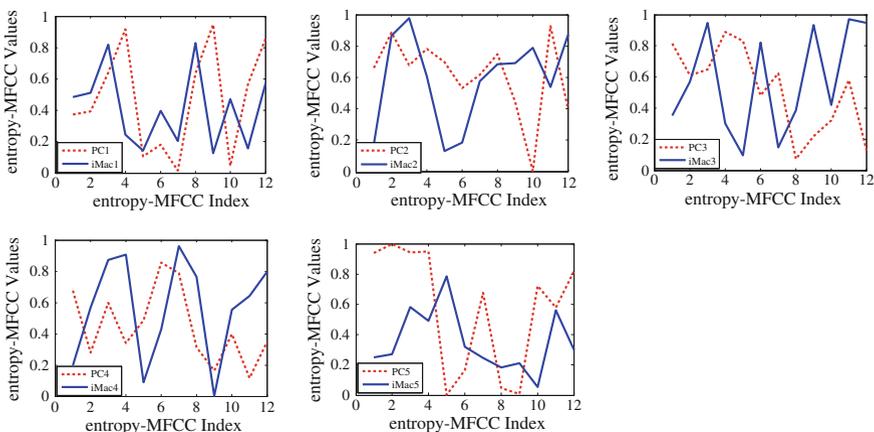


Fig. 27.3 Histogram of entropy-MFCC features for each iMac and desktop PC devices

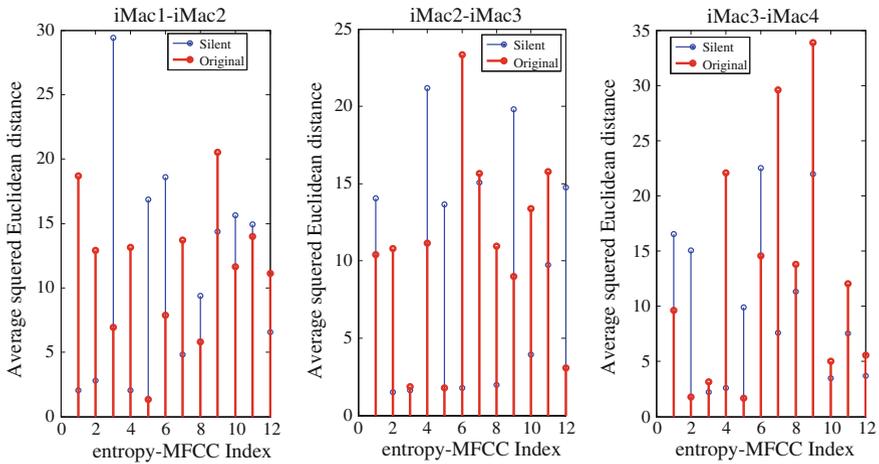


Fig. 27.4 Average squared Euclidean distances of each entropy-MFCC features on three different iMac pairs

Table 27.3 Performance on intra-model identification using original speech signal

Intra-model identification based on iMacs					
Classification algorithms	TCB	FCB	RMSE	ACC (%)	Elapsed time (s)
Naïve Bayesian	6,600	0	0	100	2.6
Linear logistic regression	6,588	12	0.0218	99.82	35.1
NN	6,591	9	0.0202	99.86	72.7
LibSVM	6,593	7	0.0206	99.89	5.5
SMO	6,596	4	0.3157	99.94	41.5
Intra-model identification based on desktop PCs					
Naïve Bayesian	7,000	0	0	100	0.51
Linear logistic regression	6,983	17	0.029	99.76	36.53
NN	6,987	13	0.0224	99.81	140.81
LibSVM	6,993	7	0.02	99.90	6.0
SMO	6,989	11	0.3163	99.84	1.18
Source brand/model Identification Based on all computer devices					
Naïve Bayesian	13,600	0	0	100	1.11
Linear Logistic Regression	13,567	33	0.0194	99.76	146.63
NN	13,570	30	0.0183	99.78	450.22
LibSVM	13,579	21	0.0176	99.85	15.81
SMO	13,575	25	0.2719	99.82	3.03

Table 27.4 Confusion matrix of linear logistic regression classifier based on iMacs

ACC = 99.82 %		Predicted				
		i1	i2	i3	i4	i5
Actual	i1	1,318	0	0	1	1
	i2	1	1,316	1	2	0
	i3	1	1	1,318	0	0
	i4	1	1	1	1,317	0
	i5	0	1	0	0	1,319

Table 27.5 Confusion matrix of linear logistic regression classifier based on PCs

ACC = 99.76 %		Predicted				
		PC1	PC2	PC3	PC4	PC5
Actual	PC1	1,397	2	0	1	1
	PC2	1	1,396	0	3	0
	PC3	0	0	1,400	0	0
	PC4	1	3	0	1,395	1
	PC5	2	0	0	3	1,395

intra-model identification based on iMacs, the overall results proved the feasibility of the entropy-MFCC features with high classification accuracies (ACC) (99.82–100 %) and minimal root mean square error (RMSE). Naïve Bayesian classifier performed higher classification accuracy and computational efficiency. SMO classifier outperformed LibSVM classifier with respect to classification accuracy, but produced larger RMSE. Table 27.4 shows the confusion matrix for the linear logistic regression classifier, with a total of 12 falsely classified blocks for intra-model device identification among five iMacs, where the truly and falsely classified blocks (TCB and FCB) are in diagonal and non-diagonal cells, respectively. In Table 27.3 the performance for intra-model identification based on desktop PCs strongly confirmed the previous result, even though the sound was transferred through the same headset microphone for all PCs. Table 27.5 shows the confusion matrix for the linear logistic regression classifier, with a total of 17 falsely classified blocks for intra-model device identification. Finally in Table 27.3 the results for source brand/model identification based on all computer devices showed that increasing the number of blocks amplifies the computation time, however the classification accuracy maintained in the same range. Moreover, we achieved lower RMSE values because increasing the number of blocks reduces the average classification error. Table 27.6 shows the confusion matrix for the linear logistic regression classifier, with a total of 33 falsely classified blocks for source brand/model device identification based on all devices.

Table 27.6 Confusion matrix of linear logistic regression classifier based on all devices

Actual		Predicted										
		i1	i2	i3	i4	i5	PC1	PC2	PC3	PC4	PC5	
i1		0	0	0	0	0	0	1	0	0	1	1
i2		0	1,317	0	1	0	1	0	0	0	0	1
i3		1	0	1,318	0	0	0	1	0	0	0	0
i4		1	0	0	1,316	0	1	1	0	0	0	1
i5		0	1	0	0	1,317	0	0	1	1	1	0
PC1		0	0	0	1	0	1,396	1	0	0	1	1
PC2		1	0	2	0	0	1	1,396	0	0	0	0
PC3		1	0	0	0	0	0	1	1,398	0	0	0
PC4		3	0	0	0	0	1	0	0	1,395	0	1
PC5		0	0	0	0	1	0	0	0	2	0	1,397

27.4.2 Experiment on Near-Silent Segments

This experiment aims to evaluate the performance of the proposed scheme without the interference of the speech signal. The experiment extracted the near-silent segments according to the algorithm that discussed in Sect. 1.1, and created a total of 1,319 and 1,400 blocks with respect to each iMac and desktop PC devices. At first the average squared Euclidean distances were calculated as in previous experiment for the same pair of iMacs to allow comparison, as in Fig. 27.4. The second part of the experiment repeated the evaluation experiments in previous Sub-Section using blocks created from near-silent segments. In overall the results in Table 27.7 shows slight improvements in classification accuracy, computation time and RMSR with comparison to Table 27.3. However, for intra-model identification based on iMacs through NN classifier the computation time was noticeably longer with compare to the results for the original speech samples. It is plausible that creating the blocks from near-silent segments is a practical option for eliminating the effects of signal variations by different speakers in real-time scenarios. However, with adaptation of the entropy-MFCC features the performance that obtained for source computer device identification based on original speech signal was in good agreement with near-silent segments. This justified our hypothesis in Sect. 27.2.2 for selecting entropy-MFCC features.

Table 27.7 Performance on proposed entropy-MFCC Features using near-silent segments

Intra-model identification based on iMacs					
Classification algorithms	TCB	FCB	RMSE	ACC (%)	Elapsed time (s)
Naïve Bayesian	6,595	0	0	100	0.54
Linear logistic regression	6,583	12	0.0218	99.82	36.28
NN	6,587	8	0.0202	99.88	116.8
LibSVM	6,591	4	0.0206	99.94	5.6
SMO	6,593	2	0.3157	99.97	0.87
Intra-model identification based on desktop PCs					
Naïve Bayesian	7,000	0	0	100	0.51
Linear logistic regression	6,994	6	0.0164	99.91	33.15
NN	6,995	5	0.0158	99.93	139.33
LibSVM	6,996	4	0.0151	99.94	5.30
SMO	6,995	5	0.3163	99.93	1.09
Source brand/model identification based on all computer devices					
Naïve Bayesian	13,595	0	0	100	1.17
Linear logistic regression	13,567	19	0.0152	99.86	144.80
NN	13,574	21	0.0183	99.85	450.60
LibSVM	13,578	17	0.0158	99.88	14.10
SMO	13,578	17	0.272	99.88	2.10

27.5 Conclusions

A blind source computer device identification scheme is developed based on the entropy-MFCC feature set and recorded VoIP calls. MFCC and entropy features identify the distinguishing pattern amongst individual computer of the same model as well as computer devices of different brands. This feature exhibited high performance to capture characteristics of the transfer function of the computer devices even with the existence of speech signal's transfer function. The naïve Bayesian classifier always achieved the highest classification accuracy of 100 % for blind source computer device identification. However, the scheme was set up in control condition and further studies are required to implement this approach on more real case forensic scenarios including large number of devices.

Acknowledgments This work is fully funded by the Malaysian Ministry of Higher Education under the University of Malaya High Impact Research Grant UM.C/625/1/HIR/MOHE/FCSIT/17.

References

1. Keonig, B.E., Lacey, D.S.: Forensic authenticity analysis of the header data in re-encoded WMA files from small Olympus audio recorders. *J. Audio Eng. Soc.* **60**, 255–265 (2012)
2. Garcia-Romero, D., Epsy-Wilson, C.Y.: Automatic acquisition device identification from speech recordings. In: *Proceedings of ICASSP*, pp. 1806–1809. Dallas, Texas (2010)
3. Panagakis, Y., Kotropoulos, C.: Telephone handset identification by feature selection and sparse representations. In: *Proceedings of WIFS*, pp. 73–78. Tenerife (2012)
4. Haniççi, C., Ertaş, F., Ertaş, T., Eskidere, Ö.: Recognition of brand and models of cell-phones from recorded speech signals. *IEEE Trans. Forensics Secur.* **7**, 635–634 (2012)
5. Jenner, F.: Non-intrusive identification of speech codecs in digital audio signals. ProQuest (2011)
6. Kraetzer, C., Oermann, A., Dittmann, J., Lang, A.: Acm: digital audio forensics: a first practical evaluation on microphone and environment classification. In: *Proceedings of the Multimedia & Security Workshop 2007 Mm&Sec'07*, pp. 63–73 (2007)
7. Buchholz, R., Kraetzer, C., Dittman, J.: Microphone Classification Using Fourier Coefficients. *Information Hiding, LNCS*, vol. 5806, pp. 235–246. Springer, Berlin (2009)
8. Kraetzer, C., Qian, K., Schott, M., Dittmann, J.: A context model for microphone forensics and its application in evaluations. In: *Proceedings of SPIE-IS&T*, San Francisco, CA (2011)
9. Kraetzer, C., Qian, K., Dittmann, J.: Extending a context model for microphone forensics. In: *Proceedings of SPIE*, vol. 8303, p. 83030S, 83012, Burlingame, CA (2012)
10. Bhatt, C.A., Kankanhalli, M.S.: Multimedia data mining: state of the art and challenges. *Multimed. Tools Appl.* **51**, 35–76 (2011)
11. Giannakopoulos, T.: A method for silence removal and segmentation of speech signals, implemented in Matlab, pp. 1–3. University of Athens, Athens (2010)
12. Giannakopoulos, T.: Study and application of acoustic information for the detection of harmful content, and fusion with visual information. Department of Informatics and Telecommunications, vol. PhD. University of Athens, Greece (2009)
13. Beigi, H.: *Fundamentals of Speaker Recognition*. Springer, New York (2011)
14. Witten, I.H., Frank, E., Hall, M.A.: *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edn. Morgan Kaufmann, Boston (2011)

15. Chang, C.-C., Lin, C.-J.: LIBSVM : a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011)
16. Pamela for Skype—Professional Edition 4.8, <http://www.pamela.biz/en/>
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *SIGKDD Explor.* **11**, 10–18 (2009)

Chapter 28

Visibility for Network Security Enhancement in Internet Protocol Over Ethernet Networks

W.K. Alzubaidi, Longzheng Cai, Shaymaa A. Alyawer
and Erika Siebert-Cole

Abstract This study addresses the naming architecture security problems arising in Internet protocol (IP) over Ethernet networks. These problems arise because of the compatibility issue between two different compositional protocols: the IP and Ethernet protocol. The findings of this study have given rise to proposals for modifications. A reduction in the current naming architecture design is advocated, which led to utilize Ethernet frame to provide a visibility for IP over Ethernet networks. The use of the IP address, as one flat address for the naming architecture, is proposed instead of using both the IP and Media Access Control (MAC) addresses. The proposed architecture has shown a promising in network security enhancement.

28.1 Introduction

IP over Ethernet network has become the primary network used by the Internet. In this network, the data link layer (Layer 2) security problems, has not yet been adequately addressed. The motivation behind addressing the compatibility problem is to improve the security of networks by studying the link compatibility between the Ethernet and IP protocols. The Ethernet was not established to work with a specific network layer (Layer 3) protocol. Likewise, the IP protocol was not

W.K. Alzubaidi (✉) · E. Siebert-Cole
Information Technology Department, University of Tun Abdul Razak, Selangor, Malaysia
e-mail: waleed@ieee.org

L. Cai
International University, Selangor, Malaysia
e-mail: charles_cai@unitar.my

S.A. Alyawer
Computer Science Department, Baghdad College, Baghdad 645, Iraq
e-mail: shaymaa@ieee.org

designed to work with a specific Layer 2 protocol [1]. This scenario clearly indicates that the relationship between the IP and Ethernet protocols is not fully compatible because such networks are not dedicated to each other, which can give rise to numerous security problems. Resolving the IP to Media Access Control (MAC) address and the encapsulation of the IP packet into the Ethernet frame are some requirements to link the IP and Ethernet protocols. Therefore, flat IP address was proposed to represent the naming architecture [2]. The use of the IP address, as one flat address for the naming architecture, is proposed instead of using both the IP and MAC addresses.

In this study, a new concept is introduced to reveal the origin of the private LAN addresses or of an undeclared Layer 2 address. Internet visibility is not possible with the current naming architecture in the IP over Ethernet networks [3]; thus, a new solution to current security problems is introduced in this study.

The visibility in the network provides new methods to handle the various network security threats. For instance, a client can connect to and request services from a Web server in the WAN without revealing private IP and MAC addresses [4] (even with a public IP with a proxy case) in the current architecture. As a result, several types of network attacks can be accomplished without disclosing the source of the attack. The IP address from Layer 3 is not detected because of the existence of a NAT in the subsequent router that replaces the original private IP address with the router's public IP address. The Layer 2 MAC address cannot be revealed in any side point because the destination and the source MAC addresses are replaced with new addresses in each hop. Determining the source of the IP address, in case it is a public IP address, is possible in Layer 3; however, even this IP type, which has network proxies, cannot be revealed. This study introduces the main procedures for the visibility concept proposed for the networks and focuses on the factors that enable this concept. The dependency and the requirements of the visibility mechanism are also revealed, and the Visibility Merging Address (VMA) is also explained in the next sections.

This paper is organized as follows. Next, in Sect. 28.2, we describe utilizing source hardware address in Ethernet frame header. Section 28.3 presents the visibility merging address VMA. Section 28.4 describes the VMA construction mechanism. Section 28.5 describe the visibility architecture and Sect. 28.6 the conclusions.

28.2 Utilizing Source Hardware Address in Ethernet Frame Header

In a LAN, the ARP is used to map the IP address to the MAC address. The destination of the MAC address should be obtained by a source machine using a destination IP address to construct and transmit the Ethernet frame in IP over Ethernet networks [5]. This task is performed by the ARP through a broadcast

request for mapping the IP to the MAC address and through a stored reply in a memory space called an ARP cache table. ARP works as follows: An application attempts to send data to an IP address of a machine. The IP packet is created by the network stack and then encapsulated into the Ethernet frame. The destination of the MAC address is required to transmit this frame [6]. Therefore, the network stack verifies the IP in the ARP cache table to locate the destination of the MAC address. If such information is not there, the broadcast ARP request is sent in the network. Each machine in the network examines the ARP request and checks if the requested IP is owned. The machine that owns this IP will create an ARP reply containing the MAC address. A unicast reply would then be sent to the originator of this request. The originator uses this address in the destination MAC address field to complete and transmit the frame. This simple protocol does not have any type of security to bind the IP to the MAC and may result in serious breaches in security. For instance, the ARP poisoning attack uses unsolicited ARP reply messages. Network devices cannot verify the ARP sender and whether the message comes from the correct device. The ARP does not provide any security measures, but is based on broadcast messages on the LAN.

28.3 Network Visibility Concept

Providing an address that represents the source of the IP and the MAC addresses on the Internet clearly informs the end point regarding the origin of the delivered information. Thus, the two ends of the connection to obtain precise information about each other. If the VMA concept is utilized, the server clearly identifies the location of a client in the client server architecture and vice versa. The role of the source hardware address field in the Ethernet frame in this approach is to provide this representative source address. Therefore, the transfer of the source MAC address and the source IP address is proposed to provide and enable visibility between the two ends of the connection. The two addresses are represented in the merged address form concept under the condition that 48 bits of the source hardware address field must be fitted [7].

The VMA address is created by combining the least significant bits of each source IP address and MAC address. An original MAC address and an IP address are transferred to the outer side of the LAN, which provide precise source addresses for the data that are being transmitted. The transparency and visibility between the two points of the connection help to prevent unauthorized modifications of the source MAC and IP address pairs and prevents anonymous attacks. For instance, a DoS attack such as example defines the source of the starting attack points by providing the original source MAC address and the private IP address [8]. The link between the original MAC source and the private IP address source is critical in providing security and aids in the clear identification of the other party's connection. The DHCP server may assign a different Layer 3 IP address to a single network node each time it is attached to the network [9], even though the node and

its MAC address remain the same. Visibility plays a main role in security, as even the dynamic IP address may change and may not be guided to the right attack node. Visibility can provide a Layer 2 source MAC address directed at the precise network machine that began the event. Thus, visibility can identify the source that is using the addresses of the two layers.

The proposed VMA address uses the existing size of the MAC address 48-bit to fit the source hardware address field and to avoid break the standard. The VMA address also provides advantages such as avoiding the creation of a non-standard Ethernet frame format. Avoiding the difficulties of adopting a new scheme entails fresh efforts and additional costs. Each router in the network path must follow the new procedure when re-encapsulating the arrived frame to resend it instead of following the current procedure included in a current router's MAC address. This procedure maintains the travel of the VMA through the networks until it arrives at the destination node.

28.4 The VMA Construction Mechanism

The VMA is entered into the 48-bit source hardware address field. The six-byte length of the MAC address is divided into two three-byte lengths. The first three least significant bytes from the source IP address replace the first three least significant bytes in the source hardware address field, whereas the first three least significant bytes from the source MAC address replace the last three most significant bytes in the source hardware address field. In other words, the 48-bit VMA is constructed when the first three least significant bytes from the source IP address replace the first three least significant bytes. The last three most significant bytes from the 48-bit VMA replace the first three least significant bytes in the source MAC address (as shown in Fig. 28.1).

However, revealing the source's private IP address and MAC address may not be desirable in most cases. Thus, the sending process can proceed without enabling the VMA and providing the source hardware address containing a VMA in the frame header; the process may maintain the proposed scheme by providing the source IP address in the source hardware address field. In the destination MAC address, the most significant bits are occasionally used to indicate broadcast station. A broadcast station that checks or uses the source MAC address in the source hardware address field is available; in other words, a broadcast address is not used in the source hardware field of the Ethernet frame header. The VMA address in the proposed scheme uses the source hardware address field in the Ethernet frame as a carrier. To recognize whether the VMA address has been utilized or not, the last most significant byte included in the broadcast bits in the VMA address is examined. If this byte is used, the VMA address was utilized.

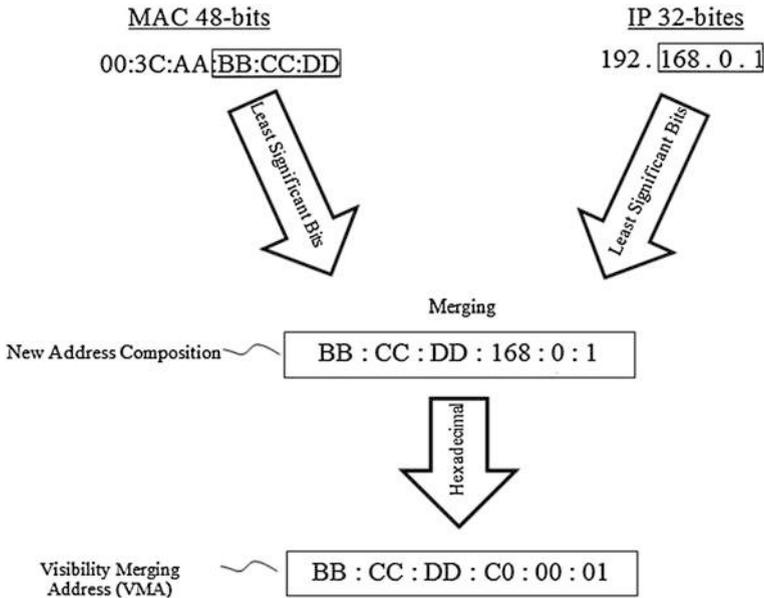


Fig. 28.1 Generation mechanism for the Visibility Merging Address (VMA)

After converting the source IP address into the hexadecimal form, take the first three octets (the least significant bits) to be the first three octets from the VMA address. Taking the last three octets from the original MAC address (the most significant bits) to take the last three octets position from the VMA address that is want to be construct. The VMA address will carry in the destination hardware address field in Ethernet frame header (Fig. 28.2).

28.5 Visibility Architecture

The transmission of a Layer 2 MAC address to the outer side of the LAN to reach the second end of the connection on the WAN side is proposed. This proposal considers the necessary level of Internet visibility. The transfer of the MAC address outside the LAN has not been proposed previously because the MAC address is typically used to guide Ethernet frame travel within the LAN; thus, the MAC address does not exit its LAN. The transfer of the original MAC address from one network to another using the source hardware address field in the Ethernet frame is proposed in this study. The source hardware address field no longer plays a main role given the previous description of the new Layer 2 transmission mechanism. Therefore, a new concept is introduced to utilize the

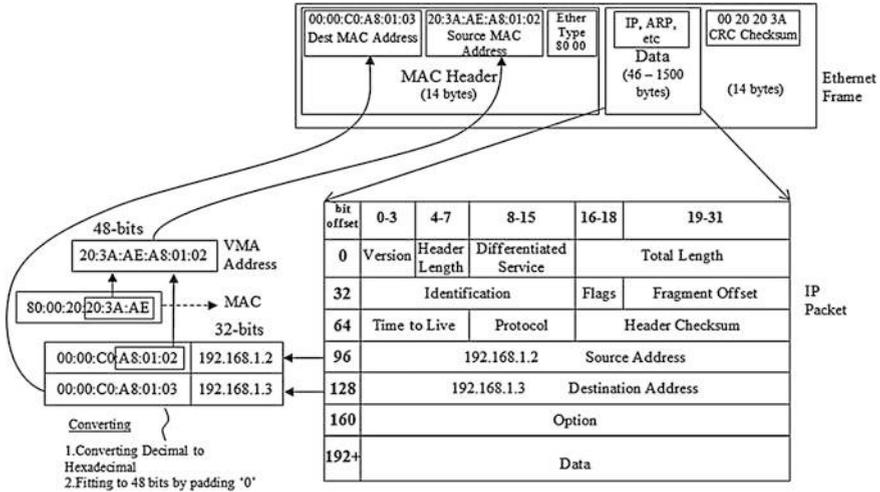


Fig. 28.2 Ethernet frame with VMA for the proposed visibility architecture

source hardware address field to carry the information regarding the original MAC and IP address source. In the current scheme, the online destination IP address travels through the networks in a packet until it arrives at the target network node. The source IP address in the packet does not change if it is a public IP address type. If the source IP is a private IP address, it changes at the next router in the NAT process.

In this approach, the source hardware address in the Ethernet frame remains fixed until it reaches the end of the connection along with the destination IP address fields in the packet. As the destination MAC address in the Ethernet frame header changes, each hop obtains the value of the next destination. The destination IP address is used as a destination address for Layer 2 in the next hop instead of the destination MAC address in the current scheme, as previously described, on the condition that the content of the source hardware address is maintained without changing the visibility of the merging VMA with each intermediate node. This process has a significant effect on the private network, especially on the NAT function.

In terms of security, the visibility helps to identify attacks. Therefore, the visibility concept has a significant effect on the enhancement of network security, especially with elements, such as the MAC address and the original source private IP address, that are not revealed to the outside LAN. Visibility does not introduce any additional complexity to the naming system. The only amendments made are to the contents of the Ethernet frame header fields. This mechanism includes encapsulation with the new VMA concept. The size of the frame header field is also unaffected because the proposed VMA fits the 48-bit source hardware address.

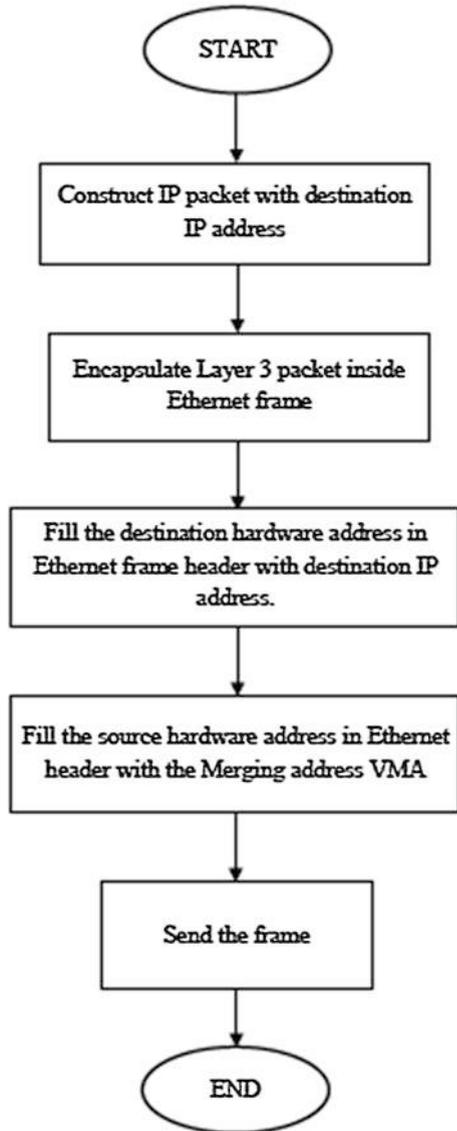
In the proposed scheme, the source hardware address field in the Ethernet frame header is proposed as a carrier of information to the WAN side to provide the original source MAC and IP address. This information is viewed as a VMA that consists of a combination of part of the source MAC address and part of the IP address that conveys the 48-bit standard source hardware address field. All of the information for the original source MAC address is removed at the first router gateway and a new Layer 2 header is added because the Ethernet frame is used within the LAN. The IP and MAC addresses in the outgoing packet/frame from the router do not provide any information regarding the original source MAC address used in the frame. The Layer 2 encapsulation typically carries information only within the LAN. In the network node procedure, the Ethernet frame in each delivered node is de-capsulated and re-encapsulated with the new control information. In the current procedure, the routers do not allow the source MAC address to cross to another communication end point. The routers also do not allow the end points to know the original source of the private IP and MAC addresses. Each intermediate node removes the Layer 2 information and re-capsulate the information with its MAC address as a source address. As shown in Fig. 28.6, how a sending and receiving in the visibility was enabled with a Private IP Address and NATing in the proposed naming architecture. In private IP networks, the router has a NAT that removes the original source private IP address in the packet. Therefore, the maintenance of the source hardware address field in the Ethernet frame header, which contains an address generated by the merging of parts of the MAC and IP addresses in the source node, is proposed. In this approach, the intermediate routers do not change the parameters of the source hardware address field. In the intermediate router, the Ethernet frame is de-capsulated and then re-encapsulated on the condition that the VMA address remains unchanged and is resent to the next node.

As a result of this approach, visibility on the Internet is provided to the original Layer 2 device address and the Layer 3 private IP address. Therefore, the original MAC and IP address link is provided with protection to avoid unauthorized changes.

The destination hardware address field is examined to determine whether the delivered frame is for the right node or not because the IP address is used as a source and destination address instead of the MAC address in Layer 2, as described in the proposal. The flowcharts describing sending and receiving procedures are shown in Figs. 28.3 and 28.4.

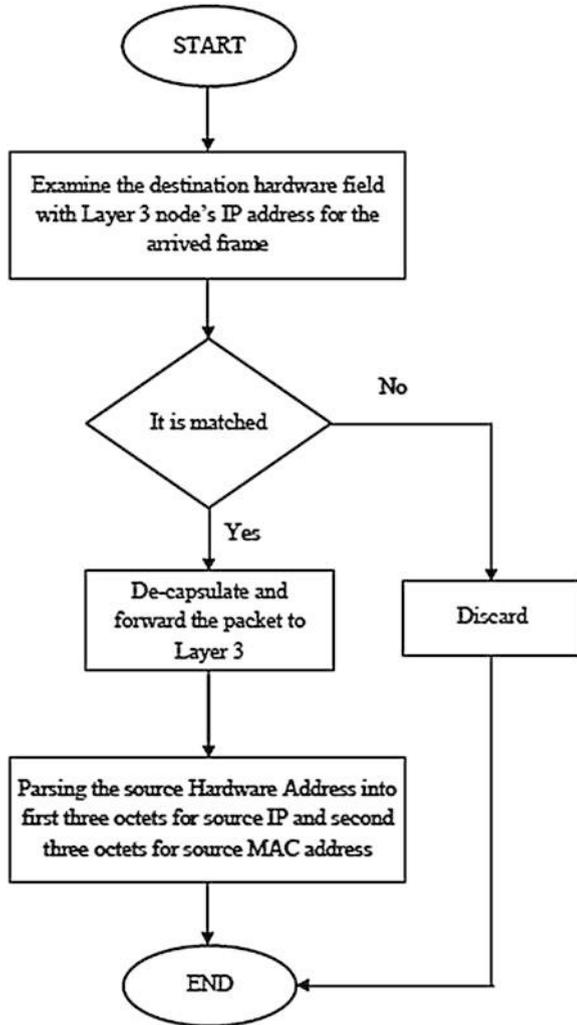
In this approach, the reply of the network node to the source of the delivered frames is dependent on the source IP address field in the packet to construct the frame using the MAC address that was generated from the source IP address. This mechanism traditionally enters the source MAC address into the source hardware address field in the frame, and the reply uses the source hardware address as the destination address in the new frame. The algorithm used to generate the Layer 2 MAC address based on the source IP address in the header of the delivered packet shown as flowcharts as following.

Fig. 28.3 Flowchart sending data with visibility enabled



The generated MAC address is used as a Layer 2 destination address and includes the constructed frame to be sent. As previously mentioned the procedure does not use the ARP cache table but depends only on the Layer 3 IP address as a flat address to guide the frame in Layer 2 transmission [12]. The source hardware address field in the Ethernet frame header does not play a role at present because

Fig. 28.4 Flowchart receiving data with visibility enabled



the source IP address is used as a reference in case the reply message requires knowledge regarding the source of the delivered frame.

Visibility also provides security advantages. In the event that a server is susceptible to any type of attack (such as Distributed Denial of Service or DDoS), the exact source of each delivered packet is easily determined. Thus, the identity of any anonymous attacker can be revealed as seen the description for the visibility architecture in Fig. 28.5 (Fig. 28.6).

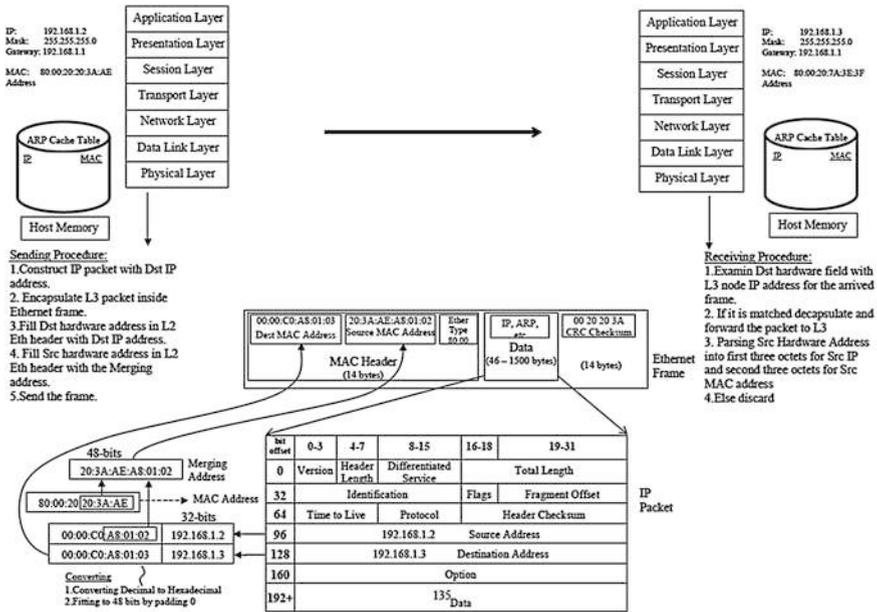


Fig. 28.5 Ethernet frame with VMA for the proposed visibility architecture

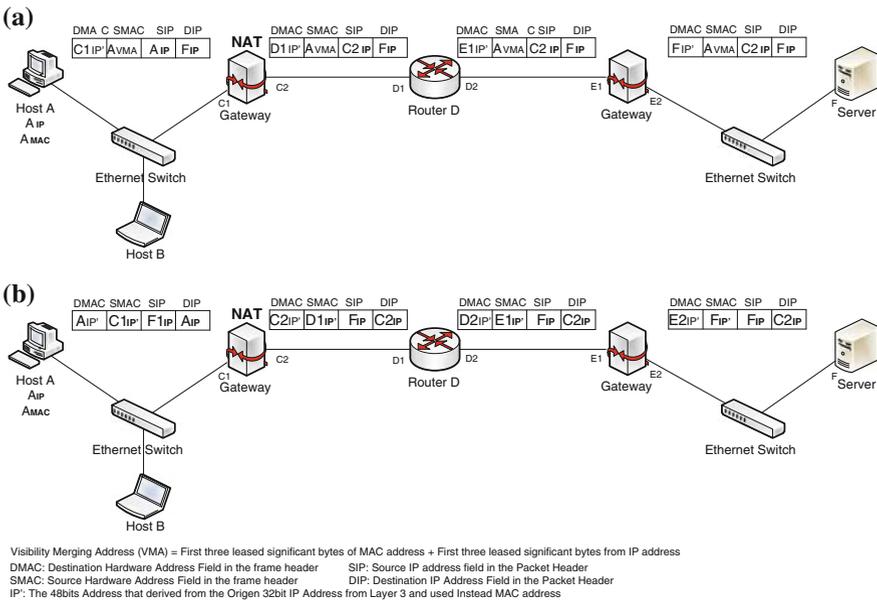


Fig. 28.6 Visibility enabled with a private IP address and NATing in the proposed naming architecture. a Sending, b reply message

28.6 Conclusions

A new Internet visibility architecture is presented to enhance the security level by raising the Ethernet working level to match the WAN side. A visibility mechanism is proposed to provide the private IP and MAC addresses to the WAN side of the network. The private IP address is generally hidden behind the NAT in the router. In the proposed architecture, the private IP address is revealed to the second end node in the communication. This visibility causes the Internet to become more visible and to be easily tracked. The visibility also eliminates Internet attacks such as anonymous and DDoS attacks. The limitation is in the scope of this study, the current visibility concept is presented under the Ethernet protocol. This concept can also be applied to Layer 2 WAN protocols and provides visibility to for all of the architectures in the LAN and WAN networks. The application of the visibility concept to ATM and frame relay as WAN protocol examples may require further study. The influence of the proposed architecture and its effectiveness in terms of Internet visibility are theoretical evaluated.

References

1. Postel, J.: Internet protocol (1981)
2. Alzubaidi, W.K., Cai, L., & Alyawer, S.A.: Enhance the security and performance of IP over ethernet networks by reduction the naming System Design. *Int. J. Comput. Netw. (IJCN)* **4**(5) (2012)
3. Wulf, V., Hartmann, A.: The ambivalence of network visibility in an organizational context. In: Paper presented at the NetWORKing (1993)
4. Forouzan, B.A.: TCP/IP Protocol Suite. McGraw-Hill, New York (2002)
5. Plummer, D.C.: RFC 826: An ethernet address resolution protocol. InterNet Network Working Group (1982)
6. Alzubaidi, W.K., Cai, L., Alyawer, S.A.: A new verification method to prevent security threads of unsolicited message in IP over ethernet networks. *Int. J.* **4**(6) (2012)
7. Spurgeon, C.E.: Ethernet: The Definitive Guide. O'Reilly, Sebastopol (2009)
8. Mirkovic, Jelena, Dietrich, Sven, Dittrich, David, Reiher, Peter: Internet Denial of Service: Attack and Defense Mechanisms (Radia Perlman Computer Networking and Security). Prentice Hall PTR, New Jersey (2004)
9. Strebe, Matthew: Windows 2000 Server 24seven. Wiley, New York (2006)
10. Hayawi, Kadhim, Al Braiki, Arwa, Mathew, Sujith Samuel: Network Attacks and Defenses: a Hands-on Approach. CRC Press, New York (2012)
11. Kiravuo, T., Sarela, M., Manner, J.: A Survey of Ethernet LAN Security
12. Alzubaidi, W.K., Cai, L., Alyawer, S.A.: A framework for optimizing IP over ethernet naming system. *Int. J. Comput. Sci. Issues (IJCSI)* **9**(6) (2012)

Chapter 29

Comparative Analysis of Different Single Cell Metamaterial

Pankaj Rameshchandra Katiyar
and **Wan Nor Liza Binti Wan Mahadi**

Abstract Artificially constructed metamaterial have become of considerable interest due to their negative properties. Under normal circumstance the electromagnetic energy inside metamaterial flows in the reverse direction. In this research, we compare seven different metamaterial structures against their negative frequency, simulation time, memory requirement and mesh cell. We also present a comparative analysis on tunability of single cell metamaterial structure to a different frequency band of interest. To come up with proper analysis we have restricted cell size of $2.5 \text{ mm} \times 2.5 \text{ mm}$ and simulated structure from 0 Hz to 20 GHz. SRR structure has a negative frequency response in a lower band while Jerusalem cross takes less simulation time. S structure has lowest memory and mesh cell requirement, but it is hard to tune. However, CLS loaded split ring resonator has advantages over all the parameters of comparison and is found best.

29.1 Introduction

V.G. Veslago in the negative direction and hence also called as Left handed Material. J.B. Pendry in 1968 introduced the concept of metamaterial and described their distinct negative properties [15]. Under normal condition the electromagnetic energy in metamaterial flows fabricated first metamaterial by arranging metal rod into an array [12]. This array when exposed to electromagnetic energy was capable of focusing energy into a narrow beam. J.B. Pendry also

P.R. Katiyar (✉) · W.N.L.B.W. Mahadi
EMRD, Department of Electrical Engineering, Faculty of Engineering,
University of Malaya, 50603 Kuala Lumpur, Malaysia
e-mail: pankaj.katiyar@siswa.um.edu.my

W.N.L.B.W. Mahadi
e-mail: wnliza@um.edu.my

showed that negative permittivity and permeability can be obtained using Split Ring Resonator [13].

Due to their unique negative property Metamaterial find many applications in optimizing electromagnetic waves. One of the first application as proposed by Pendry was to create an electromagnetic lens of sub wavelength resolution. Ran [14, 16] carried out several experiments on metamaterial such as Power transmission experiment, Prism refraction, Beam Shifting experiment, and focusing experiment. These experiments form the basis of all the application areas of metamaterial [1, 4, 5, 7, 17]. Specific experiments have been conducted to enhance gain of antenna by using metamaterial in different forms and places surrounding the antenna.

We present comparison of 7 types of metamaterial structures. The comparison is done against their negative frequency range by keeping substrate size and its properties constant. Simulation of each model is performed using CST and corresponding 2 port S-Parameter is extracted. NRW method is used to extract constitutive parameter from simulated S-Parameter. Nicholson-Ross-Weir (NRW) [10] method provides a direct calculation of both permittivity and permeability from the s-parameter. However, the method diverges for low loss material at frequencies corresponding to integer multiples of one-half wavelength in the sample which is due to the phase ambiguity, Hence it is restricted to an optimum sample thickness of $\lambda g/4$ and used preferably for short samples. The comparison is also extended to analyze tunability, size as compared to resonance frequency, memory consumption while simulating and meshing, simulation time and number of mesh cell required. This comparison is helpful in determining the type of metamaterial cell for desired application.

29.2 Metamaterial Design and Simulation

This paper shows a comparison between following four metamaterial structures.

- A. Split ring resonator structure
- B. Omega structure
- C. S structure
- D. Symmetrical ring structure
- E. Jerusalem cross
- F. Offset fed split ring [DSRR]
- G. CLS loaded SRR

To come up with fair comparison, several constrain on simulation and modeling single cell structure is enforced. The modeling constraints such as substrate size should not exceed $2.5 \text{ mm} \times 2.5 \text{ mm}$, substrate is 0.25 mm thick regular FR4. The copper pattern used on top and bottom layer of PCB has 0.017 mm thickness. Regular FR4 has permittivity (ϵ) of 4.35 and loss tangent ($\tan \delta$) of 0.025 at 20 MHz. All the single cell metamaterial structures are made to fit in the above set

Fig. 29.1 Single cell split ring resonator structure

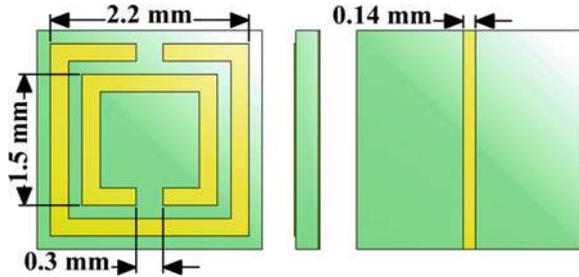
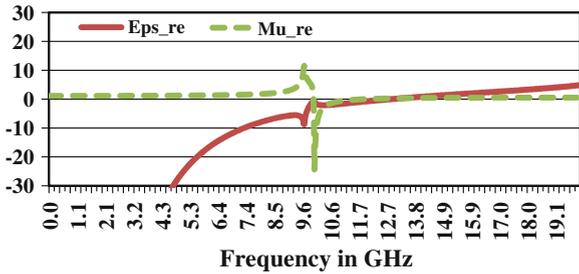


Fig. 29.2 Extracted negative permittivity and permeability



physical parameter. Simulation constrains such as the space surrounding meta-material is constant and is vacuum, the boundary condition is kept same.

29.2.1 Split Ring Resonator Structure

Split Ring resonator has two rings on top layer and one strip at the back acting as a rod. The outer ring has opening on top side while the opening of inner ring is rotated by 180°. The outer ring has length is 2.2 mm and the inner ring length is 1.5 mm. The trace width of both rings is 0.2 mm. Inner ring and outer ring are separated by 0.15 mm. The gap between inner ring and outer ring is 0.3 mm. The trace on bottom layer of PCB is of 2.5 mm length and the width of trace is 0.2 mm.

The inductance is provided by the ring pattern formed on top layer while the capacitance is provided by the separation between inner ring and outer ring. Additionally the strip on the bottom side of the PCB provides some more inductance and capacitance (Figs. 29.1 and 29.2).

29.2.2 Omega Structure

Omega structure [6] consists of an omega shape structure on top and bottom surface of the PCB. The ring traces width is 0.2 mm with inner ring radius of 0.9 mm.

Fig. 29.3 Single cell *omega* structure

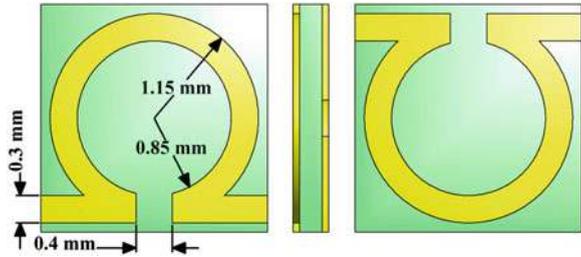


Fig. 29.4 Simulated results of Omega structure

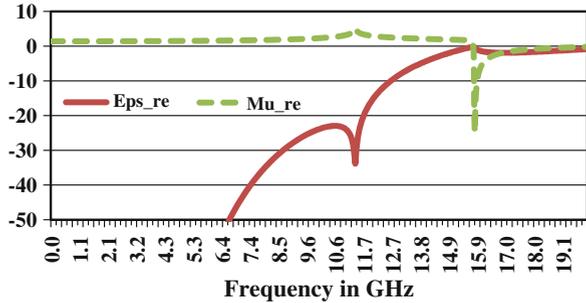
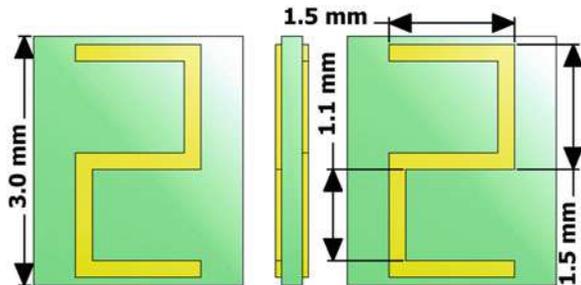


Fig. 29.5 Single cell *S* structure



The trace at the lower section of omega is also 0.2 mm wide. The gap at the lower center of the structure is 0.4 mm wide. The ring of Omega structure provides inductance. Since the ring is printed on both the side of PCB they couple to provide capacitance. Additional capacitance is provided by the 0.4 mm split in the ring (Figs. 29.3, 29.4).

29.2.3 S Structure

Figure 29.5 shows the simulated single cell S-Structure [2]. The trace width used to create S-Structure is 0.2 mm. S-Structure is placed in substrate of 3 mm × 2.5 mm. The width of the substrate used here is 0.25 mm. It is very difficult to fit S-Structure in 2.5 mm × 2.5 mm substrate and get a negative response below 20 GHz.

Fig. 29.6 Simulated results of S structure

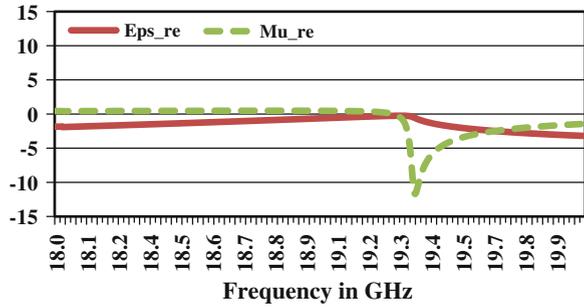
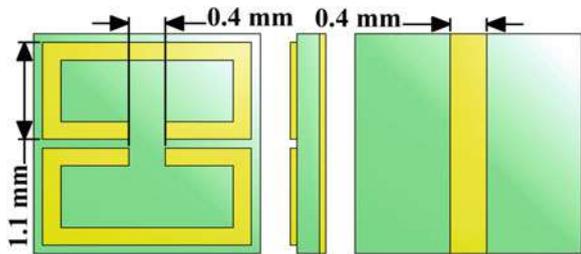


Fig. 29.7 Symmetrical ring structure



The inductance is provided by the trace forming S pattern while capacitance is provided by coupling between S pattern on top and bottom layer. In this case, the mutual coupling is limited due to limited overlap of pattern between top and bottom layer, hence the negative permeability response is weak. The inductance providing negative permittivity response is weaker as compared to Omega pattern (Fig. 29.6).

29.2.4 Symmetrical Ring Structure

Another type of metamaterial structure which exhibits negative permittivity and permeability is symmetrical ring structure as shown in Fig. 29.7. This structure is made to fit in the physical constrain set earlier [11].

Trace width 0.2 mm is used to construct a symmetrical ring. Two symmetrical ring of 1.1 mm in length are placed facing each other at a gap of 0.1 mm. The rod on the bottom layer of the PCB is 0.4 and 2.5 mm in length.

The major part of the inductance of symmetrical ring structure is due to two ring structure on top while the capacitance is due to the coupling between one side of both the rings on top layer. The split on both the ring provides additional capacitance. Additionally the strip on the back has the same effect as to strip in SRR structure (Fig. 29.8).

Fig. 29.8 Simulated results of symmetrical ring structure

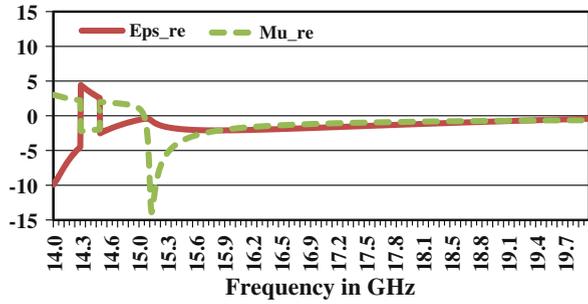
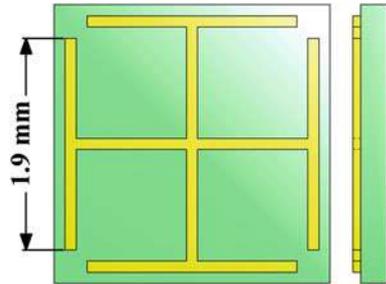


Fig. 29.9 Jerusalem cross structure



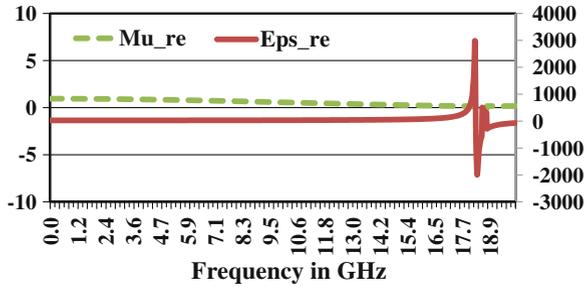
29.2.5 Jerusalem Cross

Jerusalem cross metamaterial is very similar to Jerusalem cross and hence the name [3]. The design used for our simulation is a single layer design with 0.1 mm wide trace constructing cross. Each arm of the cross is 1.9 mm long as shown in Fig. 29.9. The simulated permittivity and permeability is as shown in Fig. 29.10. Jerusalem cross is a complex structure in analyzing capacitance and inductance. The 0.1 mm traces forming part of the Jerusalem cross structure provides inductance while the capacitance is provided by the coupling formed on four corners of the Jerusalem cross. If the width and height of the trace are increasing the capacitance will also increase. Since Jerusalem cross has pattern on top layer it has limited capacitance and inductance. If the Jerusalem cross is used in array the capacitance will increase exponentially due to coupling with outer arm of the cross.

29.2.6 Offset Fed Diamond Shape Split Ring Resonator [DSRR]

DSRR is similar to split ring resonator but rotated by 45° angle anti clockwise [8]. Due to limitation applied on substrate size the dimensions of the rings are changed. The trace is 0.16 mm wide. The outer ring is 1.72 mm long while the inner ring is

Fig. 29.10 Simulated results of Jerusalem cross



1.17 mm long. The gap between inner and outer ring is 0.2 mm as shown in Fig. 29.11. Simulated permittivity and permeability is shown in Fig. 29.12.

29.2.7 Capacitive Loaded Strip with Split Ring Resonator

As the name suggests capacitive loaded strip SRR has two additional strips on sides of SRR [9]. The capacitive strip is I in shape with 1.32 mm in length and 1.15 mm wide. The gap between two capacitive strips is 0.05 mm as shown in Fig. 29.13. The simulated permittivity and permeability is shown in Fig. 29.14.

Fig. 29.11 DSRR metamaterial

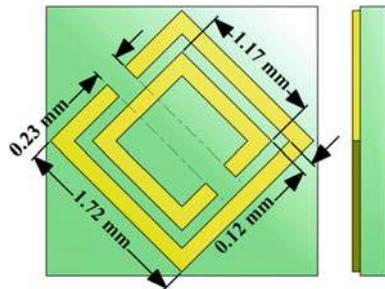
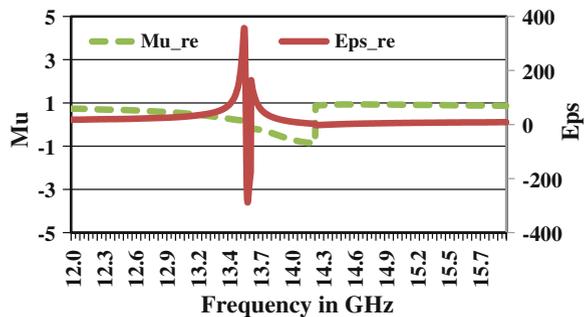


Fig. 29.12 Simulated results of DSRR metamaterial



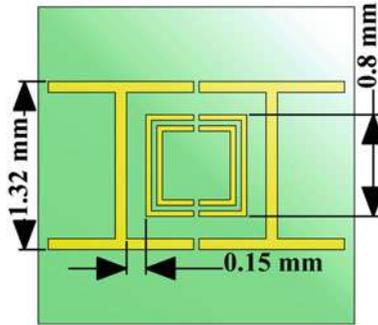


Fig. 29.13 Capacitive loaded strip with SRR

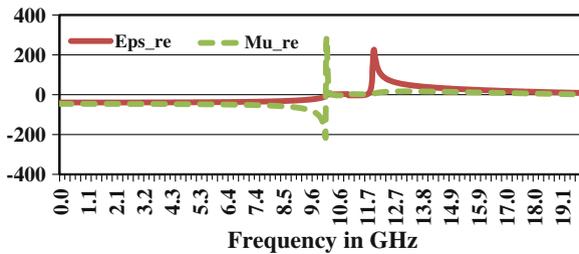


Fig. 29.14 Simulated results of capacitive loaded strip with SRR

CLS loaded SRR has the advantage of SRR and I pattern. Additional I pattern increased the amount of capacitance exponentially. The coupling of I pattern to I pattern and both I pattern to outer ring brings the resonance frequency down. The SRR structure is 3.5 times smaller but the negative response and resonance are close to SRR.

29.3 Comparison and Observation

To have fair comparison all the structure is constrained to be designed for 2.5 mm × 2.5 mm × 0.25 mm FR4 substrate. The only exception here is S-Structure which was about 2.5 mm × 3.0 mm × 0.25 mm. We are comparing the simulation result against their resonance frequency, negative permittivity and permeability frequency and ease of tuning structure to operate at the desired frequency band. We are also listing the amount of memory consumed while simulating along with number of mesh cell generated and simulation time. While simulating care is taken to maintain structure orientation and the space surrounding structure constant.

Table 29.1 Comparison of simulated parameters

	Resonance frequency (GHz)	Negative permittivity (GHz)	Negative permeability (GHz)	Tunability
Split ring resonator	10.4	≤ 10	≥ 10	Easy
Omega structure	16.8	Entire band	≥ 15.7	Hard
S structure	19	≥ 15.7	15.7	Hard
Symmetrical ring structure	15.5 and 19.1	≥ 14.3	≥ 14.3	Hard
CLS loaded SRR	Low loss, entire bandwidth	≤ 10.2	≤ 10.2	Easy
DSRR	14.4	13.6 ~ 13.7	13.6 ~ 14.2	Hard
Jerusalem cross	No clear resonance	≥ 18.3 GHz	18.3 ~ 18.4	Moderate

Table 29.2 Comparison of simulation time and memory requirement

	Simulation time (Sec)	RAM memory (MB)	Mesh cell
Split ring resonator	121	78	195,250
Omega structure	367	71	213,120
S structure	111	50	107,712
Symmetrical ring structure	92	57	150,672
CLS loaded SRR	88	54	149,124
DSRR	680	88	248,400
Jerusalem cross	72	59	110,400

The two most important parameters for antenna design are simulation time and size. In terms of simulation time CLS loaded SRR and Jerusalem cross takes the least amount of time. The amount of memory consumed by these structures is also low. However, Jerusalem cross has negative permittivity and permeability at higher frequencies than CLS. In terms of negative parameters SRR and CLS loaded SRR provides negative permittivity and permeability at the lowest frequency, so has a huge size advantage. Omega structure, S Structure and Symmetrical structure are complex to tune and shows negative parameter at higher frequency (Table 29.1).

Simulation time of DSRR is highest. Symmetrical ring has 2 resonance frequency once at 15.5 GHz and second at 19.1 GHz but it is hard to tune and simulate. CLS loaded SRR and Jerusalem cross has very wide resonance frequency and hence no clear peak is observed. CLS-SRR has low loss over the entire frequency band (Table 29.2).

29.4 Conclusion

CLS loaded SRR offers the best solution for small size and lowest simulation time. The next alternative is the Split Ring Resonator which is easily tunable, but takes 33 s more to simulate. DSRR on the other side takes 680 s to simulate and has higher memory requirement. The negative response of DSRR is also on the higher side.

References

1. Buell, K., Mosallaei, H., Sarabandi, K.: A substrate for small patch antennas providing tunable miniaturization factors. *IEEE Trans. Microw. Theory Tech.* **54**(1), 135–146 (2006)
2. Chen, H., Ran, L., Huangfu, J., Zhang, X., Chen, K., Grzegorzcyk, T.M., Kong, J.A.: Left-handed materials composed of only S-shaped resonators. *Phys. Rev. E* **70**(5), 057605 (2004)
3. Dongying, L., Szabo, Z., Xianming, Q., Li, E.P., Zhi Ning, C.: A high gain antenna with an optimized metamaterial inspired superstrate. *IEEE Trans. Antennas Propag.* **60**(12), 6018–6023 (2012). doi:[10.1109/TAP.2012.2213231](https://doi.org/10.1109/TAP.2012.2213231)
4. Faruque, M., Islam, M., Misran, N.: Analysis of electromagnetic absorption in mobile phones using metamaterials. *Electromagnetics* **31**(3), 215–232 (2011)
5. Faruque, M.R.I., Islam, M.T., Misran, N.: Design analysis of new metamaterial for EM absorption reduction. *Prog. Electromagnet. Res.* **124**, 119–135 (2012)
6. Huangfu, J., Ran, L., Chen, H., Zhang, X-m, Chen, K., Grzegorzcyk, T.M., Kong, J.A.: Experimental confirmation of negative refractive index of a metamaterial composed of Ω -like metallic patterns. *Appl. Phys. Lett.* **84**(9), 1537–1539 (2004)
7. Islam, M., Faruque, M., Misran, N.: SAR reduction in a muscle cube with metamaterial attachment. *Appl. Phys. A* **103**(2), 367–372 (2011)
8. Joshi, J., Pattnaik, S., Devi, S., Lohokare, M., Vidyasagar, C.: Offset fed diamond shaped split ring (DSSR) planar metamaterial antenna. In: *IEEE Applied Electromagnetics Conference (AEMC)*, pp. 1–4 (2009)
9. Majid, H.A., Abd Rahim, M.K., Masri, T.: Microstrip antenna's gain enhancement using left-handed metamaterial structure. *Prog. Electromagnet. Res. M* **8**, 235–247 (2009)
10. Nicolson, A., Ross, G.: Measurement of the intrinsic properties of materials by time-domain techniques. *IEEE Trans. Instrum. Meas.* **19**(4), 377–382 (1970)
11. O'Brien, S., Pendry, J.: Magnetic activity at infrared frequencies in structured metallic photonic crystals. *J. Phys.: Condens. Matter* **14**(25), 6383 (2002)
12. Pendry, J.B.: Negative refraction makes a perfect lens. *Phys. Rev. Lett.* **85**(18), 3966 (2000)
13. Pendry, J.B., Holden, A.J., Robbins, D., Stewart, W.: Magnetism from conductors and enhanced nonlinear phenomena. *IEEE Trans. Microw. Theory Tech.* **47**(11), 2075–2084 (1999)
14. Ran, L.-X., Huang-Fu, J.T., Chen, H., Zhang, X.-M., Chen, K.S., Grzegorzcyk, T.M., Kong, J.A.: Experimental study on several left-handed matamaterials. *Prog. Electromagnet. Res.* **51**, 249–279 (2005)
15. Vesslago, V.: The electrodynamics of substances with simultaneously negative values of ϵ and μ . *Sov Phys Usp* **10**, 509–514 (1968)
16. Wu, B.-I., Wang, W., Pacheco, J., Chen, X., Grzegorzcyk, T.M., Kong, J.A.: A study of using metamaterials as antenna substrate to enhance gain. *Prog. Electromagnet. Res.* **51**, 295–328 (2005)
17. Ziolkowski, R.W.: Metamaterial-based antennas: Research and developments. *IEICE Trans. Electron.* **89**(9), 1267–1275 (2006)

Chapter 30

Distributed Video Coding with Frame Estimation at Decoder

Kin Honn Chiam and Mohd Fadzli Mohd Salleh

Abstract In distributed video coding, input video stream is split into group of pictures of various lengths. Longer group of pictures is preferred as there is more temporal redundancy. However, the system time might increase as more frames need to be transmitted and be stored in the buffer. In this paper, we try to reduce the system time by proposing a model where only the odd-numbered Wyner-Ziv frames shall be transmitted from the encoder. At the decoder, the missing even-numbered Wyner-Ziv frames shall be estimated for full reconstruction of the video sequence. The simulation results show that the proposed model is more efficient as the output video sequence could be obtained in a shorter time. The estimated missing frames at the decoder are also of acceptable quality compared to the original frames at the encoder.

30.1 Introduction

Distributed video coding (DVC) is a new video coding paradigm for video transmission, based on the Slepian-Wolf and Wyner-Ziv information theoretic results. In the conventional video compression techniques, decoder is of low-complexity. The primary objective of DVC is to achieve low-complexity encoding, by shifting the bulk computation to the decoder [8]. There are couple of early architectures and implementations of DVC as stated in the publications [1, 3, 7].

Based on the cumulative motion crossing a pre-defined threshold, the incoming video sequence is sampled into consecutive frames in groups. The number of

K.H. Chiam (✉) · M.F.M. Salleh
School of Electrical and Electronic Engineering, Universiti Sains Malaysia, Seri Ampangan,
14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: kinhonn1985@yahoo.com

M.F.M. Salleh
e-mail: fadzlisalleh@eng.usm.my

frames in each group is called group of pictures (GOP). The first frame in a GOP is known as the key frame while the remaining frames are called the Wyner-Ziv (WZ) frames. Depending on the input video, the GOP can be of various sizes. Higher GOP size increases the number of WZ frames between key frames. However, this reduces the data rate [4].

In this study, we proposed a model where only certain frames in the GOP will be transmitted from the encoder. An interpolator is used at the decoder to estimate the missing nontransmitted frames. The quality of the estimated frames is measured by comparing the frames with the corresponding input frames.

Hence, together with the estimated frames, there is still a complete GOP at the decoder. All frames will then be used to reconstruct the output video sequence. The efficiency of the model is determined by measuring the time needed by the system to complete processing and transmitting a video sequence.

Prior to transmission, we utilized the Reed-Solomon (RS) codes as the forward error correction (FEC) technique to protect the message. These codes are word oriented rather than bit oriented [9]. This stems the codes the capacity to correct burst errors, where a series of bits in the codeword are received in error. The burst error is relatively common in wireless communication due to fading. A bit oriented code would treat this situation as many independent single-bit errors. However, for RS codes, a single error means any or all incorrect bits within a single word [6].

The Sect. 30.2 highlights the basic concepts of DVC, RS codes, and linear interpolation method. In Sect. 30.3, the details of the proposed DVC codec model, methodology, and implementation are presented. The results with the C++ codes implementation are in the Sect. 30.4. Finally, the Sect. 30.5 states the conclusions and further work.

30.2 Foundation of DVC, RS Codes, and Linear Interpolation

30.2.1 Concepts of Distributed Video Coding

During the encoding process of DVC, incoming video sequence is first split into the consecutive frames, to form the GOP. First frame in the group, the key frame, is intra-coded with the conventional way, providing side information (SI) at the decoder. Other frames, the WZ frames, are interframe encoded, with only the parity bits sent across the network.

In contrast, the DVC decoding process is relatively more complicated. The received key frame will first be decoded for reconstruction of the SI.

The SI is an estimation of the WZ frame, which is only available at the encoder. Together with the parity bits, SI is used to complete the decoding process. If the decoding fails, more parity bits will be requested via the feedback channel. This is iterated until successful decoding is obtained. The detailed operation of the DVC could be found in other publications [2, 4].

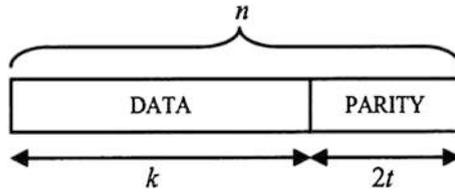


Fig. 30.1 Typical Structure of RS Codeword [6]

30.2.2 Reed-Solomon Codes

The RS codes are a subset of BCH codes and class of linear block codes. Normally, RS codes are specified with the notation, $RS(n, k)$ with m -bits symbols. A RS encoder takes k -blocks of digital data and adds extra redundant symbols, t to create n -blocks of codeword [6]. The RS decoder processes each codeword block and attempts to correct the errors occurred during transmission, recovering the original data. The number and type of errors that can be corrected depends on the characteristics of that particular RS codes and the number and distribution of errors.

- (1) If $2s + r < (n - k)$ (s -errors, r -erasures), then the original transmitted codeword will always be recovered.
- (2) The decoder is not able to recover the original codeword and indicate this fact.
- (3) The decoder will mis-decode and recover an incorrect codeword without any indication.

Figure 30.1 shows a typical RS codeword. There are more details about the RS coding scheme in the publication [9].

30.2.3 Linear Interpolation

Linear interpolation can be used as a method to estimate the missing values of a function by curve fitting or assuming that it is a straight line between two values. Generally, the linear interpolation function is given by (30.1)

$$f(x) = y_1 + \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1). \quad (30.1)$$

where x is an unknown point on a line with (x_1, y_1) and (x_2, y_2) are two other points on the same line.

On a regular two dimensional grid, bilinear interpolation is used to interpolate a bi-variate function. Linear interpolation is first performed in one direction, and then again in the other direction. The estimated value of the desired point, $f(x, y)$ on a xy -plane is given by (30.2)

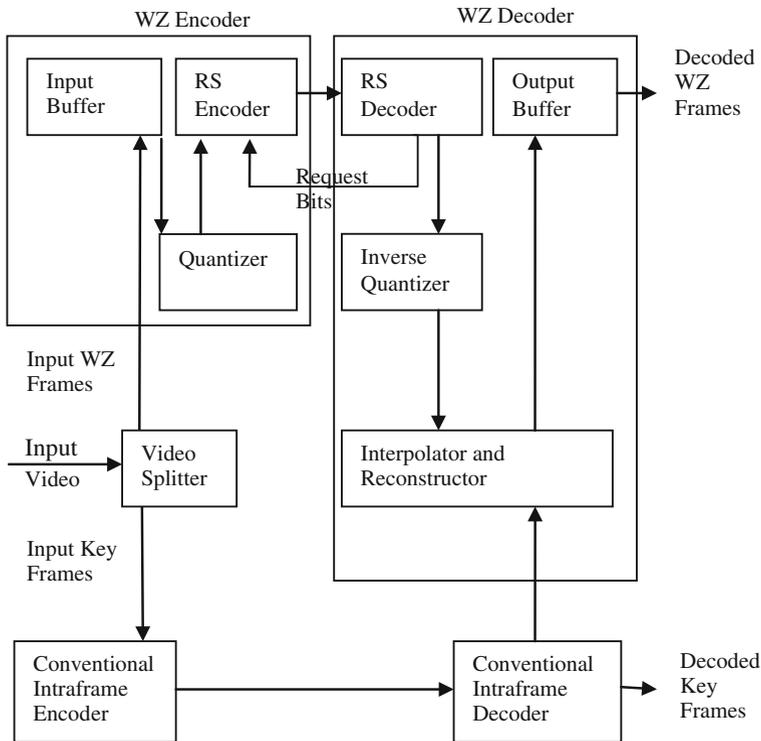


Fig. 30.2 Proposed distributed video coding model

$$f(x, y) = \frac{1}{(x_2 - x_1)(y_2 - y_1)} [f(x_1, y_1)(x_2 - x)(y_2 - y)f(x_2, y_1)(x - x_1)(y_2 - y) + f(x_1, y_2)(x_2 - x)(y - y_1) + f(x_2, y_2)(x - x_1)(y - y_1)]. \tag{30.2}$$

where (x_1, y_1) , (x_1, y_2) , (x_2, y_1) and (x_2, y_2) are four other points on the same plane.

30.3 Methodology

In this section, the proposed model as shown in Fig. 30.2 and the detailed methodology are presented. Figure 30.3 describes the operational flow of the model. Firstly, the input video sequence is split into a GOP of eight frames with the video splitter. All WZ frames will be stored in the input buffer. Only the odd-numbered WZ frames will be sent for quantization and encoded with the RS codes. At the decoder, the received frames will be first decoded and stored in the output buffer. The missing even-numbered WZ frames will be estimated by using the

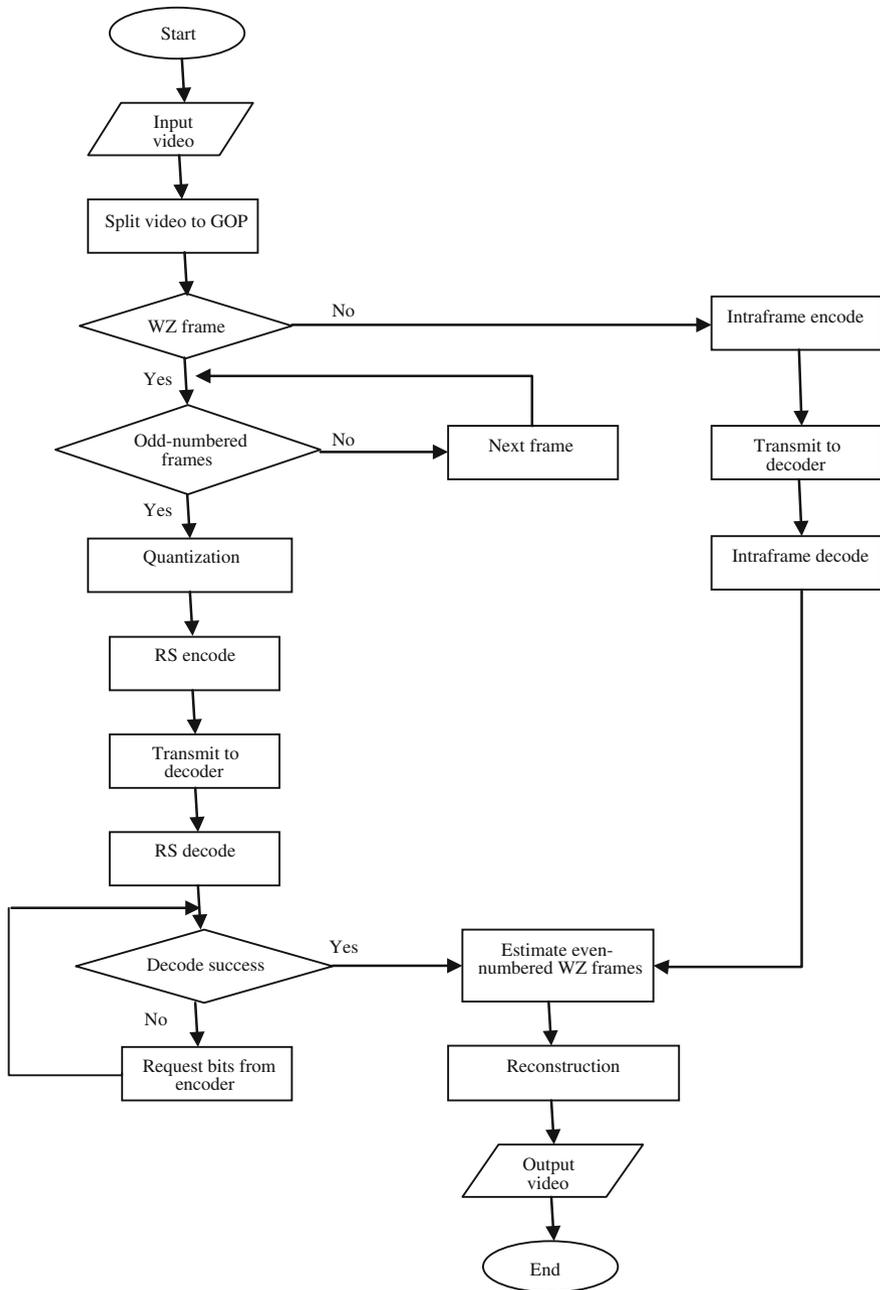


Fig. 30.3 Flowchart for the proposed distributed video coding model

Table 30.1 Rs Parameter

Parameter	Meaning	Value or range
m	Number of bits per symbol	Integer between 3 and 16
n	Number of symbols per codeword	Integer between 3 and $2^m - 1$
k	Number of symbols per message	Positive integer $< n$, where $(n-k)$ is even
t	Error correction capability of the codes	$\frac{(n-k)}{2}$

interpolation function. Each decoded or interpolated output frame will be stored in the output buffer and compared with the corresponding input frame. Based on this comparison, the peak signal to noise ratio (PSNR) is calculated. The quality of the output frame is considered good if the value of the PSNR is above 30 dB.

On the other hand, the key frame is encoded using the conventional intraframe encoding method. The chosen technique in this proposed model is the H.264 encoding scheme. The model is tested with a video sequence of low motion activity.

The performance of the model is measured by calculating the time required by the model to finish processing the GOP of eight frames. As a reference, the system will first transmit all WZ frames. Then, the time taken for the system to complete processing the GOP till the complete reconstruction of the output video sequence is recorded. After that, the same procedure is repeated with the system is now transmitting only the odd-numbered WZ frames. To ensure that the results are compared fairly, the same video sequence shall be used for both scenarios.

Prior to transmission from the encoder, all frames in the GOP shall be indexed with a number. The first frame (key frame), F_0 shall always be sent. On the other hand, the p -th WZ frame, F_p will only be transmitted if and only if p is an odd number. Therefore, p will first be verified using the (30.3).

$$q = p \% 2. \quad (30.3)$$

The F_p is only transmitted if and only if $q \neq 0$.

The RS function, $RS(n, k, m)$ has the following parameters: message length, k , codeword length, n , and number of bits per symbol, m . Table 30.1 summarizes the meaning of each parameter and the allowable positive integers of the RS codes.

In this work, eight bits are chosen for each symbol, $m = 8$. The data rate, r is 15/31, with $n = 31$ and $k = 15$. For Galois field of the form of $GF(2^m)$, the primitive element, β is 2. The method used to encode the message using the $RS(n, k, m)$ is by varying the primitive polynomial of the Galois field that contains that symbol, using an input argument in Galois field as given by the Matlab function in (30.4).

$$\text{encode} = \text{rsenc}(\text{msg}, n, k). \quad (30.4)$$

This function encodes the message, msg with the $RS(n, k)$ codes using a narrow sense generator polynomial [5]. The message is Galois array of symbol having

m -bits each. Each k -element row of msg represents a message word, with the leftmost symbol as the most significant symbol. At most, the codeword length, $n = 2^m - 1$. However, if n is not as exact as $2^m - 1$, shortened RS codes will be used. Parity symbols are added at the end of each word in the output Galois codes.

Another Matlab function (30.5),

$$\text{decode} = \text{rsdec}(msg, n, k). \quad (30.5)$$

will decode the received message, msg with the narrow sense generator polynomial [5]. The codes used at the decoder are a Galois array of symbols with m -bits each. Each n -element row of codes represents a corrupted systematic codeword, with the leftmost symbol as the most significant symbol.

If there are more than $\frac{(n-k)}{2}$ errors detected in a row of the received message, a decoding failure occurs. In this case, the function in (30.5) forms a corresponding row of decoded codeword by merely removing $(n - k)$ symbols from the end of the row of the received message.

To determine whether decoding is successful, the Hamming distance, between the received message and the SI generated using the previously decoded codeword, is calculated. If the Hamming distance is non-zero, then the decoder proceeds to the next iteration to request more parity bits from the encoder via the feedback channel. On the other hand, if the Hamming distance is zero, the decoding operation will then be verified with the 8-bits cyclic redundancy check (CRC) sum. Should the CRC sum computed on the decoded plane match the value received from the encoder, decoding is considered successful and the decoded codeword is sent for reconstruction.

The missing even-numbered WZ frames are estimated with the interpolation method. An even-numbered WZ frame must exist between two consecutive odd-numbered WZ frames which had been decoded successfully previously. Each pixel value in the even-numbered frames is interpolated from the same points of the two dimensional grid of the consecutive odd-numbered WZ frames.

The pixel value, $f(x_p, y_p)$ at the coordinate (x_p, y_p) , of a p -th WZ frame can be estimated using the linear interpolation equation in (30.6).

$$f(x_p, y_p) = f(x_{p-1}, y_{p-1}) + \frac{1}{2} [f(x_{p+1}, y_{p+1}) - f(x_{p-1}, y_{p-1})]. \quad (30.6)$$

where $2 \leq p \leq 6$ and p is even number. This function is iterated for a number of times equal to a frame size to reconstruct the whole frame.

30.4 Simulation Result

The DVC model proposed in section III is completely implemented with the C++ programming language. Both encoding and decoding functions of RS codes are imported from the Matlab library and converted to the C++ programming

Table 30.2 PSNR Values FOR Each Frame

Frame number	PSNR (dB)	
	Transmits only odd WZ frames	Transmits all frames
0	36.1473	36.1477
1	36.3032	36.3022
2	29.4138	36.3240
3	36.2582	36.2575
4	30.9052	36.2931
5	36.2789	36.2795
6	28.9724	36.2923
7	36.2521	36.2524

language. All calculations to determine the odd or even frame number and to estimate the missing even-numbered frames follow (30.3) and (30.6) respectively, and are written in the C++ programming language.

The model is evaluated with the Hall monitor QCIF video sequence. This video sequence has very low amount of motion activity.

Each reconstructed output frame in the GOP is compared with the original input frame. Based on the comparison, the PSNR is calculated using (30.7).

$$PSNR = 10 \log_{10} \frac{255^2}{d}. \quad (30.7)$$

where d is the distortion between the two frames. Table 30.2 and Fig. 30.4 present the simulation results.

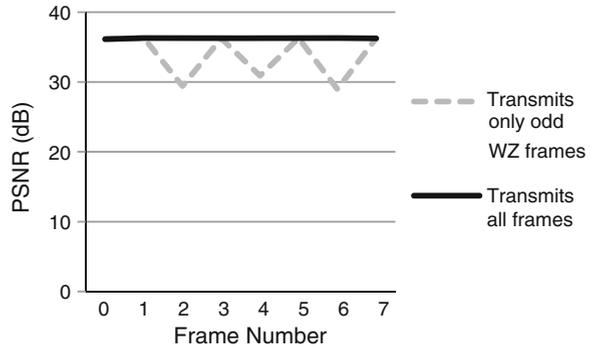
From the data, it can be seen that if all frames are transmitted, the PSNR value for each frame is above 36 dB. This also proves that there is very good decoding of the output frames at the decoder although there is no prior information available at the decoder. This finding matches the results in [1].

For the simulation with only odd-numbered WZ frames transmission, the odd-numbered WZ frames those are transmitted are decoded perfectly with the similar PSNR values like the simulation with all frames transmission. However, those nontransmitted even-numbered WZ frames which are estimated at the decoder show poorer PSNR values. This shows that the linear interpolation at the decoder is still unable to estimate the frames perfectly like the conventional inter-frame decoding.

Nevertheless, the PSNR values of the estimated frames are still close to the acceptable threshold values of 30 dB. As this input video sequence is of low motion activity, the temporal redundancy between the frames is more. Hence, the linear interpolation method can be a good approximation of the missing frames.

The simulation, where all frames are transmitted, needs an average time of 3447 s to finish processing the video sequence. On the other hand, the simulation,

Fig. 30.4 PSNR values for each frame



where only odd-numbered WZ frames are transmitted, only needs 3286 s to finish processing the same video sequence. There is a reduction of about 4.90 % processing time.

Longer processing time might due to more frames need to be transmitted and be put waiting in the buffer. Moreover, the inter-frame decoding requires the parity bits request from the encoder and also the SI from decoded key frames. Hence, the more WZ frames are to be decoded, the longer the waiting time for the parity bits request.

Contrary, interpolation method only needs two consecutive odd-numbered frames to estimate the missing even-numbered frames. All these information is readily available at the decoder and the decoder does not need to wait any information from the encoder.

30.5 Conclusion and Further Work

In this paper, the model is designed such that only the odd-numbered WZ frames in the GOP are transmitted across the network. The missing even-numbered WZ frames are recovered at the decoder with estimation using the interpolation method. Although this algorithm is less efficient in estimating the frames compared to the conventional inter-frame decoding method, the proposed model involves less complicated calculation and is able to produce the result in a shorter time. Nevertheless, the estimated frames at the decoder are also of acceptable quality compared to the corresponding input frames.

To improve the quality of the estimation of the missing even-numbered WZ frames, interpolation with motion estimation can be applied at the decoder. In [4], the authors did research about the 8×8 bidirectional motion estimation. This method can achieve better rate distortion (RD). The algorithm shall also be included in the proposed model to investigate whether there is improvement to the RD.

The proposed model shall also be tested with video sequences of different motion activity for RD comparison. This is because the RD performance of the

DVC is inconsistent with different video streams, as stated in [10]. Lower or medium activity video sequences give better RD performance, contrast to video sequences of significant motion activity. For this reason, the low motion activity video sequence of Hall sequence is selected for this work.

References

1. Aaron, A., Thang, R., Girod, B.: Wyner-Ziv coding of motion video. In: Proceedings of Asilomar Conference on Signals and Systems, Pacific Grove, CA, pp. 240–244 (2002)
2. Aaron, A., Setton, E., Girod, B.: Towards practical Wyner-Ziv coding of video. In: Proceedings of IEEE International Conference on Image Processing ICIP, Barcelona, Spain, pp. 869–872 (2003)
3. Artigas, X., Ascenso, J., Dalai, M., Klomp, S., Kubasov, D., Oualet, M.: The discover codec: architecture, techniques, and evaluation. In: Proceedings of Picture Coding Symposium (PCS). doi: 10.1.1.131.1402
4. Girod, B., Aaron, A., Rane, S., Rebollo-Monedero, D.: Distributed video coding. In: Proceedings of IEEE, Special Issue on Video Coding and Delivery, pp. 71–83 (2005) *Invited Paper*
5. Matlab Help, Matlab v7.4.0.287, The Math Works, Inc (R2007a)
6. Maggs, B.: Decoding Reed-Solomon codes [Online]. Available: http://www.cs.duke.edu/courses/spring10/cps296.3/decoding_rs_scribe.pdf Accessed 24 Oct 2000
7. Puri, R., Ramchandran, K.: PRISM: a new robust video coding architecture based on distributed compression principles. In: Proceedings of Allerton Conference on Communication, Control, and Computing, Monticello, IL (2002)
8. Riccardo, B., Roberto, R., Pamela, Z.: Distributed video coding: principals and evaluation of wavelet-based schemes. In: Sudhakar R (ed.) Effective Video Coding for Multimedia Applications. Intech, Rijeka, Croatia, pp. 111–115 (2011)
9. Shu, L., Daniel, J., Error Control Coding: Fundamentals and Applications. Prentice Hall, Eaglewood Cliffs (1983)
10. Varodayan, D., Aaron, A., Girod, B.: Rate-adaptive codes for distributed source coding. J. Sig. Proc. **86**, 3123–3130 (2006)

Chapter 31

Performance Analysis of an OCDMA System Based on SPD Detection Utilizing Different Type of Optical Filters for Access Networks

Sarah G. Abdulqader, Hilal A. Fadhil, S.A. Aljunid
and Anuar Mat Safar

Abstract This paper aims to study the performance of spectral-amplitude coding optical code-division multiple-access (SAC-OCDMA) systems based on single photo-diode (SPD) detection under various types of optical filters decoder schemes (Optical Gaussian filter, Thin film and Fiber Bragg Grating (FBG)). In our work, we utilized Modified Double weight (MDW) code as one of the SAC codes; the proposed system is performed by using single photodiode for each user. The results characterizing the bit-error-rate (BER) of 10^{-12} with respect to the data rate show that the SDP offers a significant improved performance for long haul applications (wide area network (WAN)) over other types of SAC-OCDMA detection techniques. Furthermore, the FBG filters have higher dispersion than Gaussian and thin film filters. The comparison states that the optical Gaussian filters have better filter amplitude response leading to lower dispersion bandwidth reduction, while FBG filters have higher dispersion which could reduce the goal of 622 Mbps channels in SPD detection scheme.

31.1 Introduction

There has been a huge interest in applying Optical Code Division Multiple Access (OCDMA) techniques to fiber optic communication access networks. This technique is one of the multiple access schemes that are becoming popular because their advantages such as the flexibility in the allocation of channels, ability to operate asynchronously, enhanced privacy and increased capacity in bursty networks [1–3]. Moreover, single photo-diode (SPD) detection based on OCDMA

S.G. Abdulqader (✉) · H.A. Fadhil · S.A. Aljunid · A.M. Safar
School of Computer and Communications Engineering, University Malaysia Perlis,
Kangar, Perlis, Malayasia
e-mail: eng.saragasaan@gmail.com

systems have been investigated recently to apply for high speed Wide Area Network (WAN) as multiple users allow to access network simultaneously [1–3].

Recently, many detection techniques have been proposed for SAC-OCDMA systems such as AND-detection technique, modified AND, and spectral direct detection techniques [3–6]. However, among all SAC-OCDMA detection techniques and based on the previously published papers at [3, 5]. The SPD detection technique has been proven as a good solution for ultra-high speed transmission, Multi-Access Interference (MAI) suppression and cost-effective [3]. In an optical access networks, the OCDMA signals pass through a number of optical components such as multiplexer, de-multiplexer, add-drop multiplexer, and passive optical router. However, each of these components include optical filter to separate and select the wavelengths. Choosing a proper optical filter for an OCDMA system can increase the efficiency of the system by reducing the effective bandwidth making the system less susceptible to noise, filter passband misalignment and dispersion effects. Therefore, in this study it is desirable to investigate the different type of optical filters (Fiber Bragg Grating (FBG), thin film and optical Gaussian filters) in order to be applicable with the various access networks such as MAN, and WAN. This paper is organized as follows: Sect. 31.2 provides a detailed explanation of our proposed system. Next, in Sect. 31.3, we focus on performance analysis followed by the discussion of the results. Finally, Sect. 31.4 summaries our conclusions.

31.2 System Design

31.2.1 Modified Double Weight Code

The code structure is based on Modified Double –Weight (MDW) code families for SAC-OCDMA systems. The MDW codes have a large number of weight can be developed based on double weight (DW) code of weight two, the MDW code is the modified version of DW code [6]. The MDW code possesses ideal cross-correlation properties and exists for every natural number [5, 6]. However, the MDW code weight can be any even number that is greater than 2. Moreover, the MDW codes can also be represented by using a $(K \times N)$ matrix as shown in Fig. 31.1. The details of code structure and code parameters have been presented in [6].

Figure 31.1 shows that we can increase the number of user from 1 to 3 while the weight is still fixed at 4. An MDW code with weight of 4 denoted by $(N, 4, 1)$ for a given code length N , can be related to the number of user K through:

$$N = 3K + \frac{3}{8} \left[\sin \left(\frac{k\pi}{3} \right) \right]^2 \quad (31.1)$$

$$\begin{vmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \end{vmatrix}$$

Fig. 31.1 The basic MDW code with code length 9, weight 4, and an ideal in-phase cross correlation

31.2.2 SPD Detection Based Gaussian Optical Filters

The proposed SAC-OCDMA receiver diagram of this technique is shown in Fig. 31.2. The received optical signal is decoded by the decoder, which has an identical spectral response to the intended encoder for the data to be received. The remainder of the signal from the decoder is then transmitted to the subtractive decoder (s-Decoder) to cancel out signals with mismatched signatures, i.e., interferers. The output from the s-Decoder is either zero power unit for active user or cross-correlation power unit for interferers. The proposed technique can be performed using inexpensive optical Gaussian filter to decode the received signal. Moreover, other types of filters such us Fiber Bragg Grating (FBG) and thin film filters are also used as the main part in the corresponding SPD implementations.

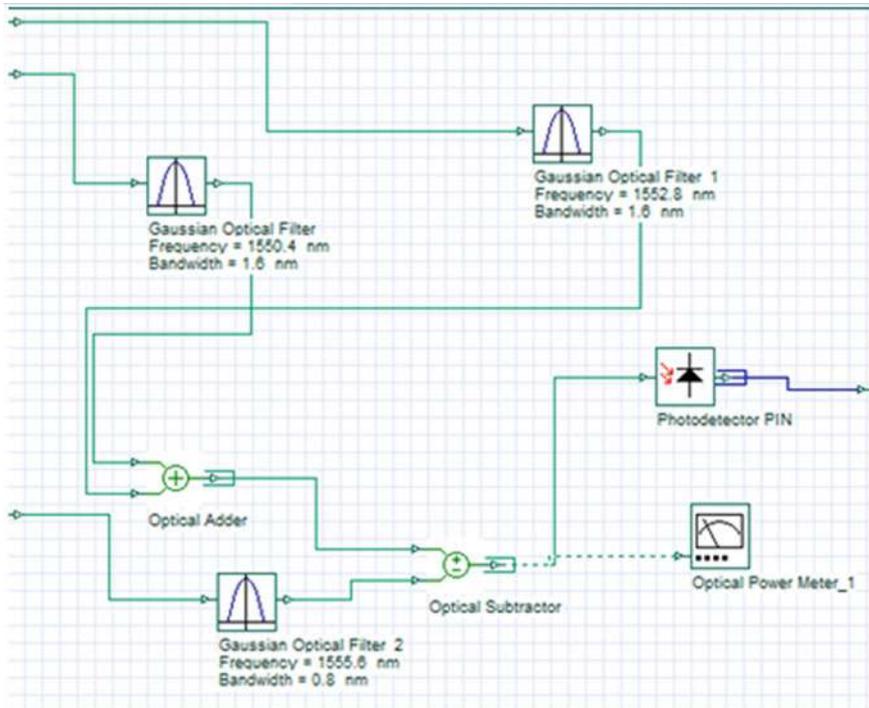


Fig. 31.2 SPD detection based on MWD code

After optical subtraction, the output is either code weight power unit for active user or zero power units for interferers. This implies that the interference signals are suppressed in the optical domain before the conversion of the signals to the electrical domain, as a result, the proposed SPD scheme alleviates both PIIN and MAI in the optical domain [3]. Moreover, the two interference signals at the optical subtractor are assumed to be equal and cancel each other out. However, practically, the interference signals differ slightly at the optical subtractor and results in a small amount of optical power to reach the photodiode. The main advantage of using the SPD is that the cancellation of the interference signals in the optical domain allows the use of only a single photodiode rather than two photodiodes as in typical subtraction detection schemes [7]. This reduces the amount of optical-to-electrical conversion and shot noise generated at the receiver part. The proposed detection technique can also be implemented with any fixed in-phase cross-correlation SAC codes with differ spectral chips distribution of the s-Decoder, depending on the structure of the SAC codes itself. Finally, after the desired signal is detected by a photodiode, the data-carrying electrical signal is low pass-filtered by a Bessel-Thompson filter [8].

The signal-to-noise (SNR) for an electrical signal is defined as the average signal to noise power $SNR = [I^2/i_t^2]$, where i_t^2 is the variance of noise source (note: the effect of the receiver's dark current and nonlinear noises are neglected in the analysis of the proposed system), given by

$$\sigma^2 = \langle i_{shot}^2 \rangle + \langle i_{PIIN}^2 \rangle + \langle i_{thermal}^2 \rangle, \quad (31.2)$$

Equation (31.2) can be expressed as

$$\sigma^2 = 2eBI + I^2B\tau_C + \frac{4K_bT_nB}{R_L}, \quad (31.3)$$

where the symbols used in Eq. (31.3) bear the following meaning.

- e Electron charge;
- I Average photocurrent;
- I^2 The power spectral density for I ;
- B Electrical bandwidth;
- K_b Boltzmann Constant;
- T_n Absolute receiver noise temperature;
- R_L Receiver load resistor.

The formula used to calculate the bit-error-rate (BER) with Gaussian approximation can be expressed as [7, 8]

$$BER = P_e = \frac{1}{2} \operatorname{erfc} \left(\sqrt{\frac{SNR}{8}} \right), \quad (31.4)$$

where erfc is the complementary error function

31.2.3 Different Decoders Architecture Schemes

Different decoder architecture scheme are based on different type of filters used in the SPD implementation. In this sub-section only three type of filter will be investigated and analysed which are the Optical Gaussian filter, thin film filter, and FBG filter based Gaussian apodization profile. It is well known that the apodized and linearly chirped Bragg gratings reduced the side lobe level in the reflectivity response and also the group delay response ripple [9]. The FBG based Gaussian apodization profile can be presented as [9]:

$$T(z) = \exp\left[-G\left(\frac{z}{l}\right)^2\right] \quad (31.5)$$

The parameters z and l are used to control the apodization sharpness parameter. Also, the optical Gaussian filter is based on Gaussian distribution expressed as:

$$T(z) = \frac{1}{2\pi} \exp\left\{\frac{-1}{2} z^2\right\} \quad (31.6)$$

The thin film filters are known to have a Butterworth response and are commonly modeled as 3rd order Butterworth filters [10]. However, the different decoder architecture schemes have the different impact; it would be very useful to investigate and to analyse the different filter functions to find out most significant one for dispersion compensation (long haul applications) and access network categories as well. Table 31.1 shows the system parameters used in the simulation model (*Optisystem*TM), these parameters based on the previously published papers at Hamza et al., Norazimah et al., and Ghafouri-Shiraz [3, 5, 8].

31.3 Simulation Results

Figure 31.3 shows the average BER for three users against the fiber length at data rates of 622 Mbps and 155 Mbps. It can be seen that the average BER value increases with the increasing of the transmission length. Further, for a fiber length of 55 km the BER are 1×10^{-12} and 1×10^{-18} for the data rate of 622 and 155 Mbps, respectively. Moreover, the FBG based Gaussian apodization of 622 Mbps allows short transmission in the fiber length as compared with data rate of 155 Mbps. However, this type of filter possesses easily controllable dual-wavelength narrow transmission peaks. It provides a simple and low cost approach of achieving the dual-wavelength fiber laser operation [9].

Figure 31.4 shows average BER for three users against the data rate at transmission distance of 30 and 50 km. It can be seen that the average BER value increases with the increasing of data rate for both distances. Moreover, for a given

Table 31.1 System parameters used in the simulation

Wavelength	1,550 nm
Thermal noise	1.8×10^{-23} W/Hz
Dark current	5 nA
Responsively	1 A/W
Optical bandwidth	$B_0 = 3.75$ THz
Attenuation	0.25 dB/km
No. of user	3
Code weight	4
SPD filter	FBG, thin film, and Gaussian

Fig. 31.3 BER versus fiber length at data rate 622,155 Mbps for FBG based Gaussian adopization

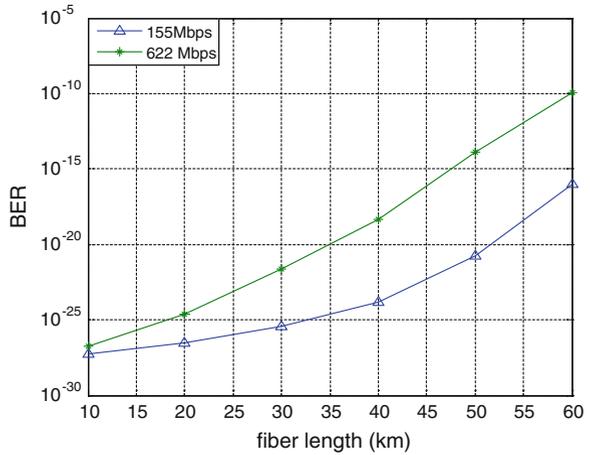


Fig. 31.4 BER versus data rate at different fiber length for FBG based Gaussian adopization

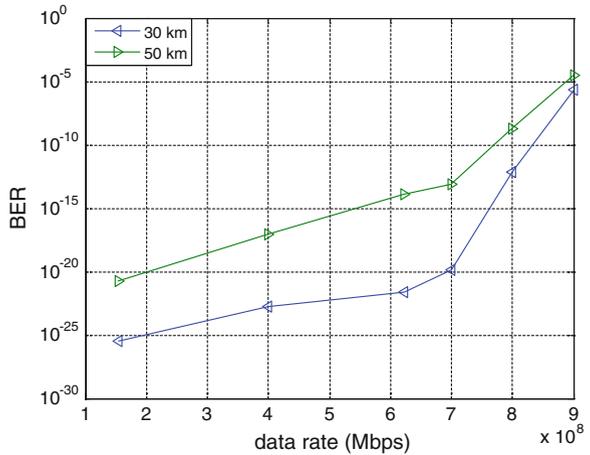
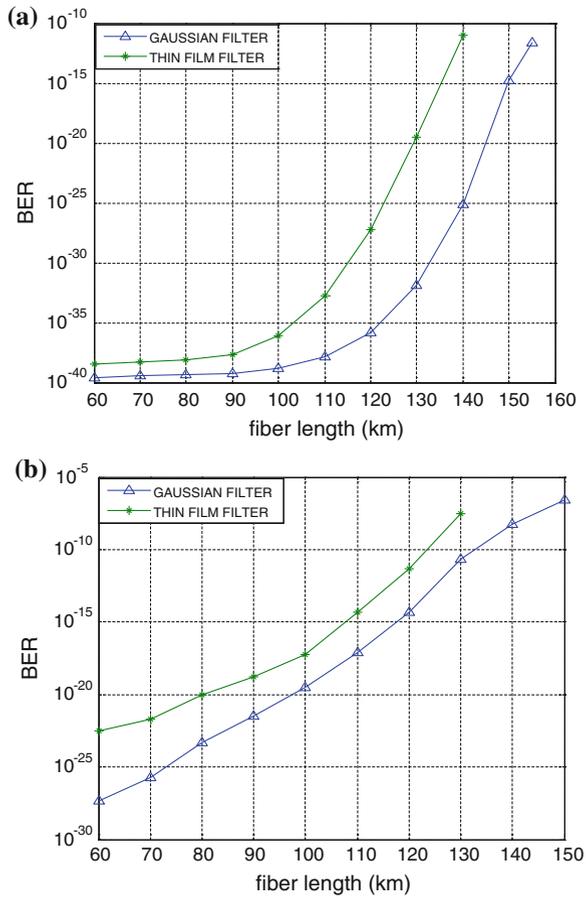


Fig. 31.5 BER versus fiber length for thin film filter and Gaussian filter at data rate of **a** 155 Mbps; **b** 622 Mbps



data rate at transmission distance of 30 km the proposed system offers better performance in terms of BER than at 50 km. This result indicates that there is a significant improvement in performance or, for a fixed BER, could accommodate a higher data rate for greater capacity.

Figure 31.5 shows average BER against the fiber length at different data rate of 155 and 622 Mbps. It can be seen that the average BER value increases with the increasing of the transmission length. Further, for a BER of 1×10^{-12} with data rate of 155 Mbps, optical Gaussian filter allows extra transmission of 20 km in the fiber length and better performance as compared with thin film filter, whereas for a BER of 1×10^{-12} with data rate of 622 Mbps, Gaussian filter allows longer transmission of 12 km in the fiber length as compared with thin film filter. This improvement can be explained as the optical Gaussian filters offers passband sharp which exhibits the lowest loss at the peak compared with thin film filter.

Fig. 31.6 Comparison between thin film filter and Gaussian filter at fiber length **a** 70 km; **b** 120 km

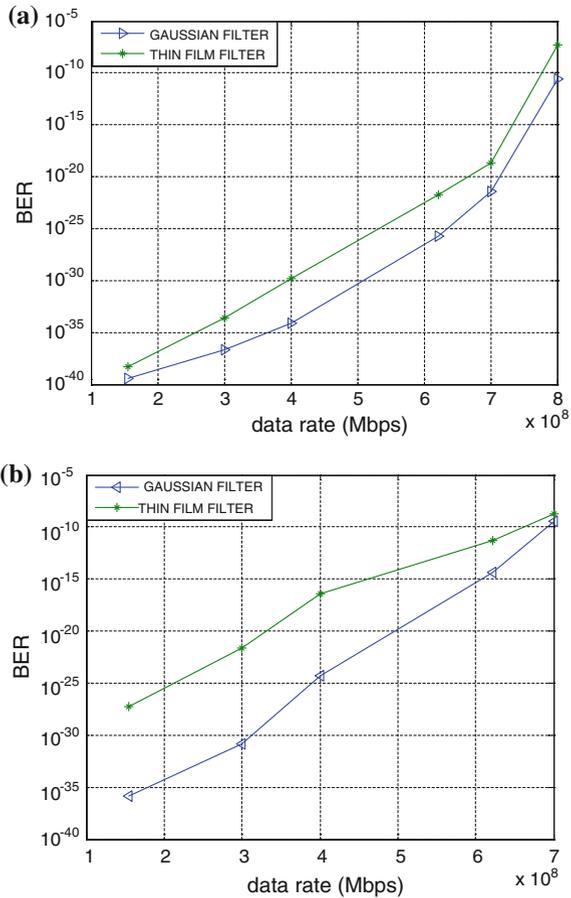


Figure 31.6 shows average BER against the data rate at different fiber length of 70 and 120 km. It can be seen that the average BER value increases with the increasing of transmission data rate. Further, for a BER of 1.5×10^{-12} with a fiber length of 70 km allows higher transmission rate than a fiber length of 120 km. However, the optical Gaussian filter is considered an ideal solution to achieve long haul transmission (120 km) for SAC-OCDMA system with minimized dispersion affects compared with thin film filter. Moreover, an SAC-OCDMA system is considered an ideal solution for different transmission rates on the access links [11].

The Eye pattern diagrams for the various proposed filters are shown in Fig. 31.7. Figure 31.7 clearly shows the dispersion effects by FBG based Gaussian adpotionzation compared with other filters.

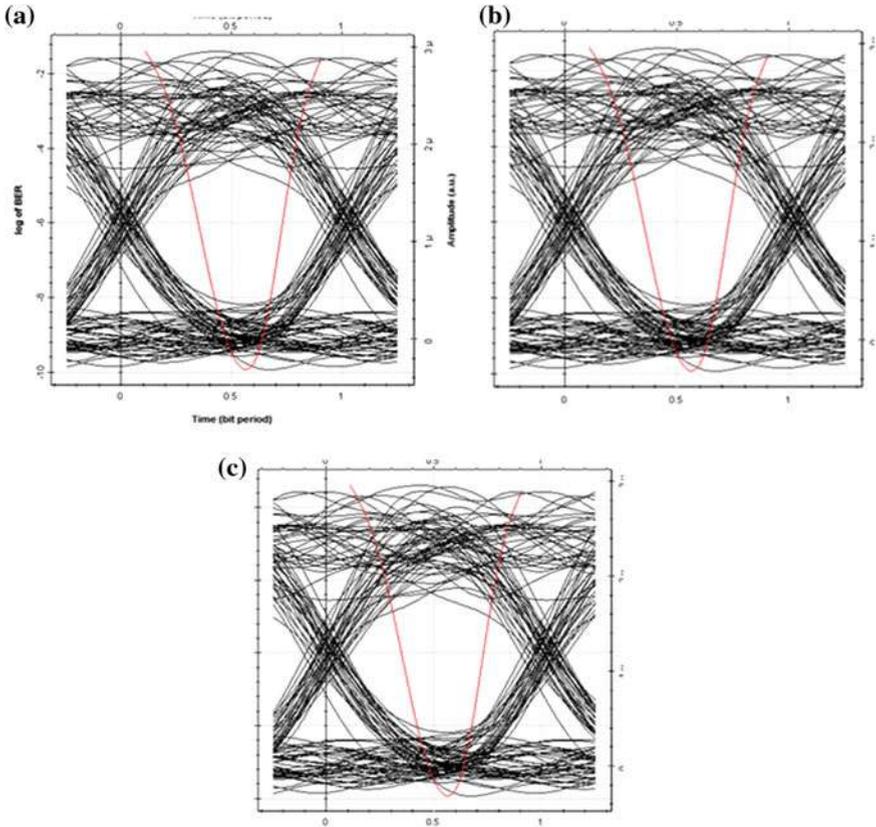


Fig. 31.7 Eye diagram for the data rate of 622 Mbps at fiber length of 60 km of one of the: **a** FBG based Gaussian adpoziation BER = 1.1×10^{-10} ; **b** Optical Gaussian filter, BER = 4.5×10^{-15} ; **c** thin film filter BER = 4.7×10^{-9}

Table 31.2 Applications of proposed system based different decoder architecture

Application	Decoder type	Fiber length (km)	BER
WAN	Gaussian filter	140	$\approx 10^{-12}$
WAN	Thin film filter	130	$\approx 10^{-12}$
MAN	FBG Gaussian	55	$\approx 10^{-12}$

Table 31.2 depicts that our proposed system based on different decoder architecture schemes can be applied to interconnection ranging from localized links within an equipments rack to links that span continents or Fiber-to-Building (FTTB) applications. As Table 31.2 illustrates, networks are divided into two broad categories based on filter types and transmission distance (WAN and MAN).

31.4 Conclusions

In this study, we have tested optical Gaussian, FBG and thin film filters to assess the suitability of these filter types as decoding devices in OCDMA networks using SPD detection. The FBG filters have higher dispersion than Gaussian or thin film filters (Thin film filters and Gaussian filter both have low dispersion). Thin film and Gaussian filters have better filter amplitude response leading to lower bandwidth reduction, while FBG filters have higher dispersion which could reduce the goal of 622 Mbps channels in SPD detection scheme. The properties of this system are described and discussed with the related equations. However, based on the simulation results, these filter can be implemented in the SPD decoder architecture at different access categories (WAN, MAN) as a specific BER of 10^{-12} is required in order to achieve different quality-of-service (QoS) requirements. Furthermore, the proposed study is employed for access network applications could be an excellent candidate for applying in the next generation OCDMA networks.

References

1. Tarhuni, N.G., Korhonen, T.O., Mutafungwa, E., Elmusrati, M.S.: Multiclass optical orthogonal codes for multiservice optical CDMA networks. *J. Lightwave Technol.* **24**(2), 694–704 (2006)
2. Kwong, W.C., Guu-Chang, Y.: Design of multilength optical orthogonal codes for optical CDMA multimedia networks. *IEEE Trans. Commun.* **50**(8), 1258–1265 (2002)
3. Hamza, M.R., Al-Khafaji, : Reducing BER of spectral-amplitude coding optical code division multiple-access systems by single photodiode detection technique. *Europ. Opt. Soc. Rap. Public.* **8**, 13022 (2013)
4. Fadhil, H., Aljunid, S., Badlisha, R.: Triple-play services using random diagonal code for spectral amplitude coding OCDMA systems. *J. Opt. Commun.* **30**(3), 155–159 (2009)
5. Norazimah, M.Z., et al.: Performance comparison of different detection techniques in Long-Haul Fiber SAC-OCDMA systems. In: 3rd International Conference on Photonics, Malaysia (2012)
6. Aljunid, S.A., et al.: A new code For optical code division multiple access systems. *Malays. J. Comput. Sci.* **17**(2), 30–39 (2004)
7. Prucnal, P.: *Optical Code Division Multiple Access: Fundamentals and Applications*. Taylor and Francis, Boca Raton (2006)
8. Ghafouri-Shiraz, H.: *Optical CDMA Networks, Principle, Analysis and Applications*. Wiley, Chichester (2012)
9. Sher Shermin, A., et al.: Determination of the best apodization function and grating length of linearly fiber bragg grating for dispersion compensation. *J. Commun.* **7**(11), 840–846 (2012)
10. Fadhil, H., Aljunid, S., Ahmed, B.: Performance of OCDMA Systems using random diagonal code for different decoders architecture schemes. *Int. Arab J. Inf. Technol.* **7**(1), 7 (2010)

11. Fadhil, H.A., et al.: Multi-rate transmissions on spectral amplitude coding optical code division multiple access system using random diagonal codes. *Optica Applicata* XXXIX(2) (2009)
12. Khaleghi, S., Pakravan, M.R.: Quality of Service Provisioning in Optical CDMA Packet Networks. *J. Opt. Commun. Netw.* **2**(5), 283–292 (2010)

Chapter 32

Deployment of Optimized Algorithm for MPEG-4 Data Over Wireless Multimedia Sensor Network

Norlezah Hashim, Sharifah Hafizah Syed Ariffin, Farizah Yunus, Fakrulradzi Idris and Norsheila Fisal

Abstract Typical Wireless Sensor Network (WSN) always deals with scalar data such as temperature. These types of data are suitable for low rate networking technology such as IEEE802.15.4. Transmitting a video for IEEE 802.15.4 raised other challenges for bandwidth limited sensor networks like WSN. The optimization method able to determine the optimal limit for quantization scale, group of picture and frame per second in order to suit the WSN environment. Changing these parameters affects the bandwidth requirement and video quality in term of Peak Signal to Noise Ratio (PSNR). This project aims to create an embedded code in TelG mote according to this concept. Results from experiment show improvements in packet delivery ratio of 28 % for Akiyo *qcif* file, 27 % for Foreman *qcif* file and 9 % for Mobile *qcif* file. This work proved that the method has successfully increased the network performance.

32.1 Introduction

Analysis in Wireless sensor network (WSN) often includes the study of network properties of wireless communication with small battery powered sensors. Nowadays some of WSN applications include disaster prevention, environmental monitoring, and logistics tracking. Most WSN uses IEEE 802.15.4 standard, is one

N. Hashim (✉)

Faculty of Engineering Technology, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Melaka, Malaysia
e-mail: norlezah@utem.edu.my

S.H. Syed Ariffin · F. Yunus · N. Fisal

UTM-MIMOS Center for Telecommunication Technology, UTM, 81310 Skudai, Malaysia

F. Idris

Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Melaka, Malaysia

of the interest due to its capability in producing low cost with low complexity but still maintaining the good quality of video. However, transferring a video is more challenging since it requires bigger bandwidth to carry bigger data and WSN only allow the maximum bandwidth of 250 kbps. Therefore a compression method such as MPEG-4 is needed before a video being transferred through WSN to reduce the requirement of its bandwidth.

The researchers in [1] have developed the benchmark for the settings of video encoding which can provide better video quality suitable for WSN application. They mentioned three parameters, namely quantization scale, group of picture, and frame per second that are important in determining the quality of a video received. The researchers used Network Simulator 2 (NS2) as an approach to the study. Even though simulation is a reliable approach to study the network properties, the real scenario of WSN is much more complicated. Therefore, a real world testbed implementation is needed to justify the results obtained using simulation. In this paper, the design, deployment and evaluation of WSN testbed using optimized algorithm to transmit MPEG-4 data will be discussed. The deployment will be done using TelG mote and WiseOS operating system. Results on packet delivery ratio and total packet loss at different distances are presented. Moreover, comparisons of results between MPEG-4 data with optimized algorithm and MPEG-4 data with non-optimized algorithm are further elaborated. Lastly, comparisons of results between experiment and simulation are presented to the readers.

The rest of this paper is organized as follows; Sect. 32.2 briefly describes the design concept of optimized parameter for MPEG-4 data, followed by testbed implementation in Sect. 32.3. In Sect. 32.4, the results obtained are elaborated and finally the conclusions are given in Sect. 32.5.

32.2 Design Concept

The standard for video compression which are MPEG-1 [2] and MPEG-2 [3], are not versatile enough to efficiently address the requirement of multimedia application. Therefore, Moving Picture Expert Group (MPEG) developed MPEG-4 standard that provides a platform for a wide range of multimedia applications [4]. In [5, 6], MPEG-4 video is already proven suitable to be transmitted over IEEE 802.15.4 standard. Many researchers focused research on MPEG-4 transmitted for IEEE802.11 standard that can support high bandwidth, such as in [7, 8], and only a few researches done for MPEG-4 over IEEE802.15.4.

In optimized method, the video need to be encoded without quantization scale with a default value of 30 frames per second and 30 groups of frame. Next, the bandwidth of video is checked and if it does not meet the requirement condition of WSN which is 250 kbps; the video will be quantized with the increasing value of the quantization until 31. If the bandwidth still does not meet the requirement, the value for frame per second will be reduced until it meets the bandwidth requirement. Next the PSNR is checked to ensure the video is still in good quality after

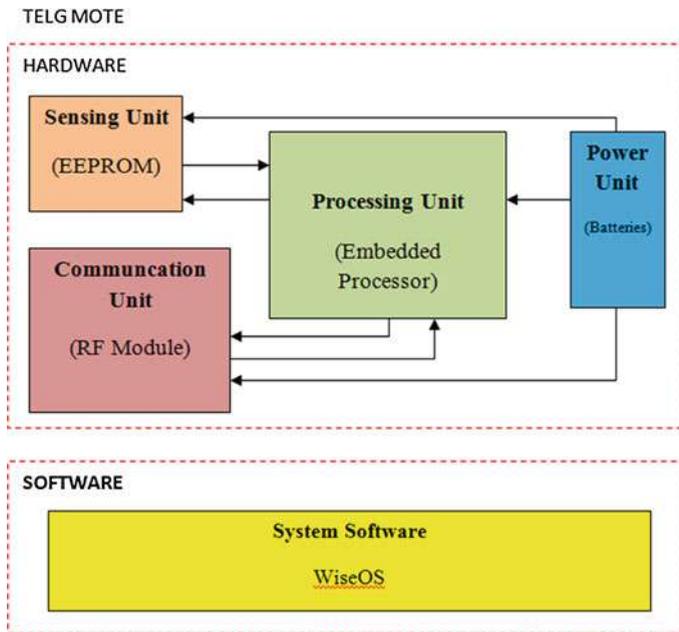


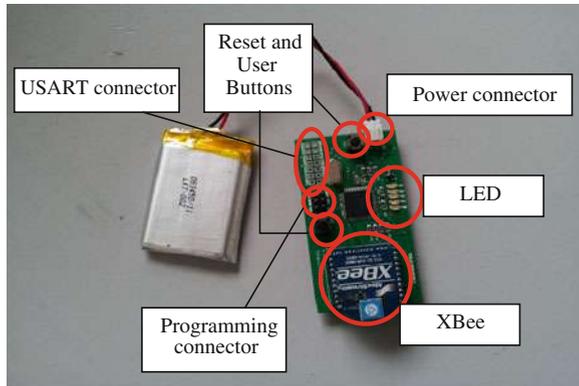
Fig. 32.1 TelG system architecture

several adjustments have been made to the video. In order to get better video quality, the PSNR values need to be increased by decreasing the group of frame value while maintaining the bandwidth requirement. This step will be repeated until the optimal value for group of frame, frame per second, and quantization scale is obtained but still within the bandwidth requirement of WSN and good PSNR value.

32.2.1 TelG System Architecture

Figure 32.1 shows TelG system architecture used in testbed deployment. TelG hardware system architecture consists of sensing unit, communication unit, processing unit, and power unit [9]. TelG is particularly designed where programming, computation and communication are integrated onto a single device. Each TelG can be reprogrammed on the board using ISP.

In sensing unit, the electrically erasable programmable read only memory (EEPROM) is used to store MPEG-4 data before transmitting. ATmega644PV microcontroller from Atmel Corporation is used in TelG as the processing unit. ATmega644PV has its own advantage in term of low power consumption used [10]. In the communication unit, the wireless device is using the XBee module

Fig. 32.2 TelG device

from MaxStream Inc. XBee is an IEEE 802.15.4 compliant radio based on Carrier Sense Multiple Access CSMA. TelG is battery powered, with a cutoff voltage of 1.8 V [10].

WiseOS was developed by a group of researchers from Telekom Research Group (TRG) of UTM. The programming language used for WiseOS is called nesC, which is C with some addition language features for components and concurrency. WiseOS is an event-driven operating system. Figure 32.2 shows TelG device used in experiment.

32.2.2 Experiment Setup

The experiment location is chosen at UTM block P19 as shown in Fig. 32.3. Figure 32.4 shows the configuration setup which consists of three TelG points spaced equally among them. TelG node A acts as the source node, TelG node B is the intermediate node and base station behaves as the sink node. Besides, base station also collects all data and sent them to a Personal Computer for analysis purposes.

32.3 Methodology

The video frame format used in the testbed is known as Quarter Common Intermediate Format (*qcif*) where this format has a resolution of 176×144 pixels that represents different color spaces [1]. Three types of video samples used in the testbed are Akiyo, Foreman and Mobile. These video samples appear as low, medium and high motion and scene complexity respectively. These video samples are commonly used among researchers. Table 32.1 shows the summary of the data used in this research where six different sets of MPEG-4 data were conducted in

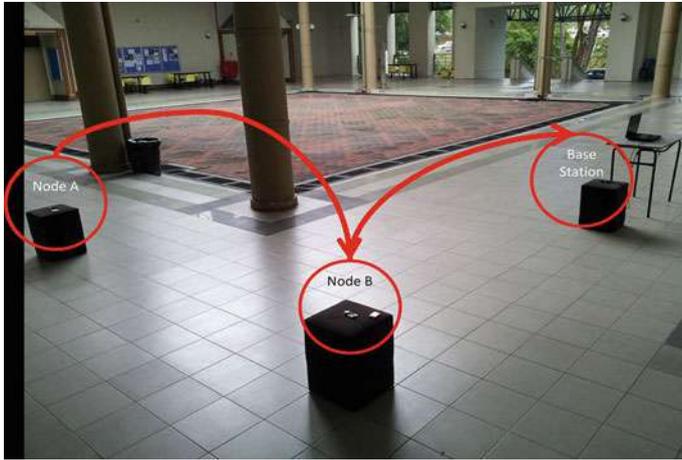


Fig. 32.3 Experiment setup at P19, UTM

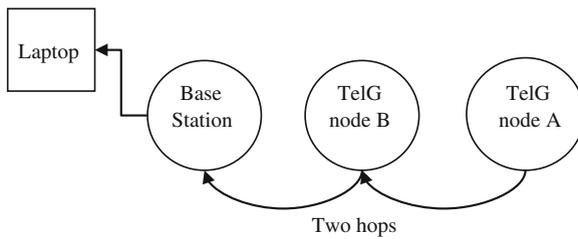


Fig. 32.4 Configuration setup

Table 32.1 Summary of the data used in experiment

Type of motion		Quantization scale	Frame per second	Group of frame	Bandwidth
Optimized	Akiyo	2	30	20	248
	Foreman	4	15	10	229
	Mobile	4	5	10	249
Non-optimized	Akiyo	2	30	10	293.5
	Foreman	3	15	10	310.7
	Mobile	3	5	10	324.8

the testbed. It can be seen when any value from the optimal value is varied, will result in the change of bandwidth requirement.

Before making any measurement, each TelG needs to be programmed according to the algorithm to test the proposed method. Figure 32.5 shows the overall flowchart of the proposed method. The flow chart is divided two parts.

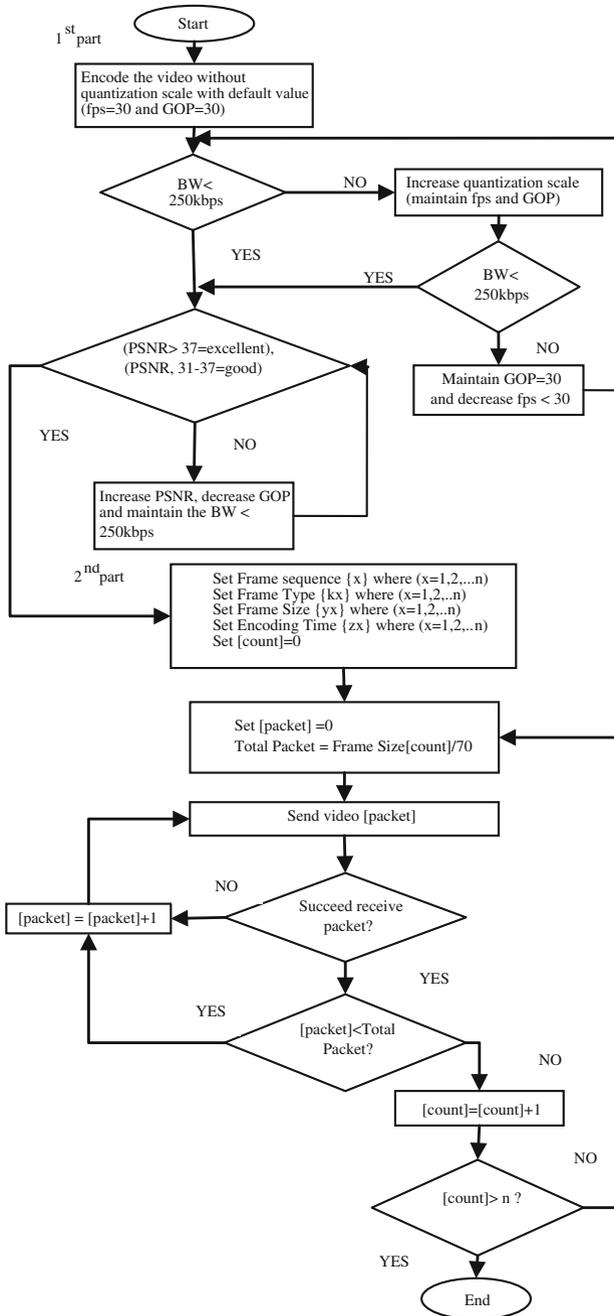


Fig. 32.5 Project flowchart

The first part is to select the optimal values for frame per second, quantization scale and group of frame while the second part is to transfer the video packets through WSN channel.

32.4 Result and Discussions

Figure 32.6 shows the packet delivery ratio for three samples namely Akiyo, Foreman, and Mobile *qcif* data. In general, the packet delivery ratio was showing a decreased pattern when the distances between nodes are increased. As the distances are increased, it is reasonable to experience more losses because the probability a data to loss is increased as the signal strength is decreased. From the graph, packet delivery ratio for Akiyo optimized data improves an average of about 28 % compared to non-optimized data, 27 % differences for Foreman and 9 % differences for Mobile.

It is observed from the graph, the improvement for Foreman is smaller than Akiyo as Foreman contains more packets which is twice the Akiyo's total packets. This will increase the probability of packets dropping since the network is getting busy with packets transferring. From Table 32.1, it can be seen that in order to maintain the network bandwidth requirement not more than 250 kbps and to maintain the video quality in term of PSNR, Foreman's file contains a smaller group of frame compared to Akiyo's generated from the optimized method. Smaller group of frame increased the quantity of I frames, which constitutes to bigger file size and lower packet delivery ratio.

For Mobile, the optimized data improves the network of about 9 % in averages compared to non-optimized data. Compared to Foreman and Akiyo, Mobile has the smallest improvement in term of packet delivery ratio. Mobile file represents a high motion and scene complexity of a horizontally moving toy train with a calendar moving vertically in the background. Therefore, Mobile file's size is bigger than Akiyo and Foreman which results in lowest packet delivery ratio.

In order to maintain the bandwidth requirement of Mobile below the maximum allowable rates of WSN, the value of quantization scale is increased to 4 (refer to Table 32.1). As we know quantization scale represents how much compression is done to the video. Increasing quantization scale will reduce the requirement of bandwidth but will introduce a lower video quality. However, by using the optimized method, the quality of the video is still maintained at a good PSNR value above 31 according to Mean Opinion Score (MOS) [1].

The packet loss versus distance graphs for Akiyo, Foreman and Mobile were shown in Fig. 32.7. In overall, packet loss was showing an increased pattern when the distance between nodes are increased. From this figure, it is clearly observed that optimized MPEG-4 data improved the overall network performance for Akiyo, Foreman and Mobile by reducing the total packet loss.

Fig. 32.6 Packet delivery ratio for Akiyo, Foreman, and mobile *qcif* data

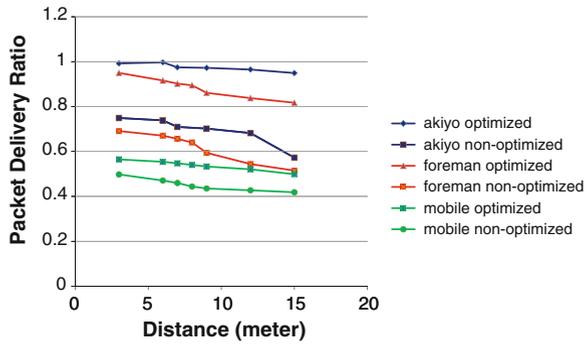
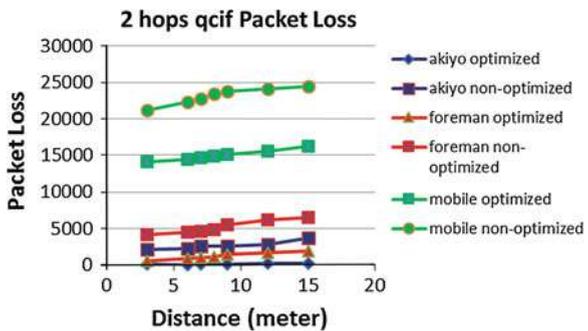


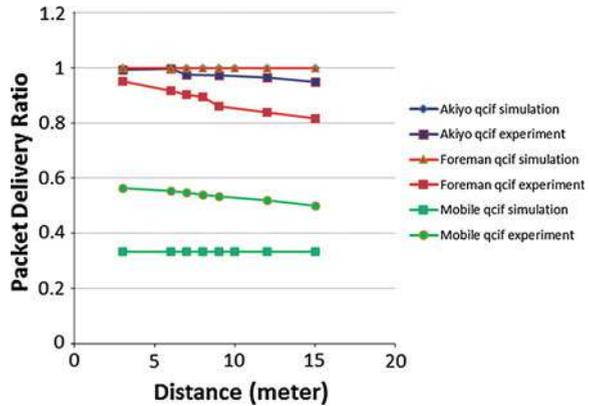
Fig. 32.7 Packet loss for Akiyo, Foreman, and mobile *qcif* data



Akiyo *qcif* shows the highest improvement with the ability to reduce the packet loss to an average of 2 % compared to non-optimized data where the average packet loss is 31 %. The average packet loss for Foreman *qcif* is 12 % compared to the average packet loss for non-optimized data which is 38 %. For Mobile *qcif*, the average packet loss is 46 % compared to non-optimized data which is 55 %. There are higher possibilities to decode optimized Akiyo and optimized Foreman compared to optimized Mobile since the average packet loss for both files is smaller. When packet loss is small, the probability of receiving most of I frame which serves the most important information are higher. However, for non-optimized data, it is almost impossible to decode the data as the packet losses are very high. Packet losses are more critical as the distances are increased.

Figure 32.8 shows the comparison of results between simulation using NS2 and experiment. From the graph, it can be seen the results observed from experiments are more or less the same with simulation. The average differences between experiment and simulation results are 2.3 % for Akiyo, 11.5 % for Foreman, and 10.3 % for Mobile *qcif*.

Fig. 32.8 Simulation versus experiment packet delivery ratio for Akiyo, Foreman and mobile *qcif* data



32.5 Conclusions and Future Works

In this work, we have deployed two hops wireless sensor network testbed and investigated its performance. The results show that the optimized algorithm for MPEG-4 data improves packet delivery ratio and reduces packet losses in WSN. However, among the three MPEG-4 *qcif* data, the highest improvement is observed for Akiyo, followed by Foreman and Mobile.

In future, the optimized method is suggested to be tested on TelG using 802.11 g WiFi module. The maximum allowable bandwidth in WiFi is bigger in 802.11 g compared to 802.15.4 which will further improve the network performance.

Acknowledgments The authors would like to thank Ministry of Higher Education (MOHE), UTM-MIMOS Center for Telecommunication Technology, Universiti Teknologi Malaysia (UTM) and Research Management Center (UTM-RMC) for the financial support of this project under GUP research grant no. Q.J130000.2523.04H39. The first author would also like to thank Universiti Teknikal Malaysia Melaka (UTeM) for providing the scholarship.

References

1. Farizah, Y., Sharifah, H.S.A., Sharifah, K.S.Y., Nor Syahidatul N.I., Abdul Hadi, F.A.H., Norsheila, F.: Optimum parameters for MPEG-4 data over wireless sensor network. *Int. J. Eng. Technol.* (0975-4024), **5**(5), 13p, (2013) 4501-4513
2. Information technology—Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s—Part 3: Audio, International Standard ISO/IEC 11172-3:1993 (1993)
3. Information Technology—Generic Coding of Moving Pictures and Associated Audio: Part 2—Video, International Standard ISO/IEC 13818-2:1995 (1995)
4. Information technology—Coding of audio-visual objects—Part 4: Conformance Testing, International Standard ISO/IEC 14496-4:2004 (2004)

5. Antonio Javier, G.S., Felipe Garcia, S., John Garcia, H.: Feasibility study of MPEG-4 transmission on IEEE 802.15.4 networks. In: International Conference on Wireless and Mobile Computing, Networking and Communication (2008)
6. Antonio Javier, G.S., Felipe Garcia, S., John Garcia, H.: A Cross-layer solution for enabling real-time video transmission over IEEE 802.15.4 Networks. In: Multimedia Tools and Application (2010)
7. Pilgyu, S., Kwangsue, C.: A Cross-layer based rate control scheme for MPEG-4 Video transmission by using efficient bandwidth estimation in IEEE 802.11e. In: International Conference Information Networking ICOIN (2008)
8. Zemin, M., Jinhe, Z., Tonghai, W.: A cross-layer QoS scheme for MPEG-4 streams. In: Wireless Communications, Networking and Information Security (WCNIS), 2010 IEEE International Conference, pp. 392–396 (2010)
9. Rozeha, A.R., Norsheila, F., Abdul Hadi, F.A.H.: Wireless multimedia sensor network platform for low rate image/video streaming. *Jurnal Teknologi. Special Edition (Instrumentation and Sensor Technology)* **54**, 1–12 (2011)
10. Abdul Hadi Fikri, F.A.H., Rozeha, A.R., Norsheila, F., Sharifah, K.S.Y., Sharifah, H.S.A.: Development of IEEE802.15.4 based wireless sensor network platform for image transmission. *Int. J. Eng. Technol.* **9**(10), 112–118 (2009)

Chapter 33

Partially Compensated Power Control Technique for LTE-A Macro-Femto Networks

Sawsan Ali Saad, Mahamod Ismail and Rosdiadee Nordin

Abstract Femtocells are expected to remarkably enhance network capacity, indoor coverage and introduce brand-new services. However, in co-channel deployment scenarios, and with the femtocell being applied as a closed subscriber group (CSG) access, femtocells can cause severe interference to the neighbouring cells and derive users to the outage. In this paper a downlink power control scheme without additional signalling exchange is proposed, which adjusts the transmit power subject to user's pathloss measurements. However the minimum level of power transmission is constrained to the acquired target SINR of femtocell user that could be set according to the QoS requirements. The performance is confirmed via system level simulations. The results show that, the proposed scheme reduces the outage probability of the macrocell users, and the spectral efficiency of femtocell users is improved. Furthermore, reducing the transmit power level helps in saving more power towards green femtocell networks.

33.1 Introduction

The concept of femtocell has attracted much attention in the wireless communication industry; as a solution for indoor coverage problems. Femtocell also known as Home enhanced NodeB (HeNB) in the third generation partnership project (3GPP) long term evolution-advanced (LTE-A) standard; has many attractive

S.A. Saad (✉) · M. Ismail · R. Nordin
Department of Electrical, Electronic and Systems Engineering,
National University of Malaysia, 43600 Bangi, Malaysia
e-mail: Sawsan_3@eng.ukm.my

M. Ismail
e-mail: mahamod@eng.ukm.my

R. Nordin
e-mail: adee@eng.ukm.my

features of plug-and-play, low deployment cost, traffic offload from macrocell and providing enhanced indoor coverage with high data rates [1].

The transmit power level of a femtocell base station affects its coverage range and the amount of interference it generates in the network. Properly selecting the femtocell base station transmit power level can help manage the interference from the femtocells to the macro-users, while maintaining femtocells performance. Although femtocells are low power base stations, the massive deployment of them will result in increasing the total power consumption of the network. Therefore, efficient methods, such as cell zooming, sleep mode and power control are required to reduce the power consumption while maintaining the performance of femtocells [2].

A method that discriminates between indoor and outdoor is presented in [3]. The location information gained from the discrimination procedure together with the required SINR for the HUE is used to perform power control for the HeNB. The proposed scheme outperforms the random power scheme by providing higher SINR for the home user equipment (HUEs). However the performance gap between the proposed scheme and the random method has decreased in environments with higher shadowing effect. This is due to the decreased accuracy of positional state discrimination.

The power control problem is formulated in [4] considering the worst case scenario, assuming no dominating interferer. A heuristic distributed algorithm is executed to determine the optimum power level, if no feasible solution for a sub-channel the algorithm determines the admissible subset of users on the given sub-channel by eliminating the interferers causing excessive interference level. The scheme delivers suboptimal results which are shown to be close to the optimal. However the effect of the user mobility is neglected as stationary network is assumed.

The co-channel deployment of femtocells with macrocells is investigated in [5]. A macro user (MUE) assisted HeNB power control scheme is proposed; that adjusts the transmit power of the HeNB when receiving an interference message from an MUE. Two timers are used to control the decrease and increase of the transmit power. The scheme can reduce the outage probability for the victim MUEs and also avoid unnecessary throughput loss of the HeNBs. However the MUE needs to send interference messages to HeNB, yet there is no direct connection between HeNB and MUE, which implies possible delay and low reliability in transmitting the control information.

This paper proposes a technique to mitigate the downlink interference in co-channel multi-tier macro-femto networks, while maintaining the performance of the HUEs. The scheme allows the implementation of the concept: coverage follows HUE; without requiring HUE positioning technique and hence saving the transmission of such information. Instead the position of the HUE is estimated using the pathloss between the HUE and the HeNB; the pathloss is partially compensated, which helps more in mitigating the interference to the neighbouring cells. However, the minimum level of power transmission is set adaptively to guarantee the required target SINR of the HUE which in consequence conserves the QoS requirements. The performance of the scheme is evaluated via system level simulation for the metrics of SINR, outage probability and the spectral efficiency.

The rest of this paper is organized as follows: Sect. 33.2 presents the system model. Section 33.3 provides details about the proposed power control scheme and describes the simulation scenario. Section 33.4 is dedicated to the performance evaluation results and discussion. Concluding remarks are given in Sect. 33.5.

33.2 System Model

The system model is used to study the interference impact on the performance MUEs and indoor HUEs, in terms of the SINR, the outage probability of MUEs and the spectral efficiency, due to the introduction of the power control approach for femtocell base stations.

33.2.1 Interference Scenario

In the downlink, the received signal at the user equipment (UE) contains the OFDMA transmitted symbols of the serving base station plus the interference induced by nearby femto and macro base stations. Figure 33.1 shows the interference scenario used in this study. Two types of interference are considered: (i) Cross-tier interference between macrocell and femtocell, in which the cell-edge MUE may suffer from severe interference. (ii) Co-tier interference between the femtocells.

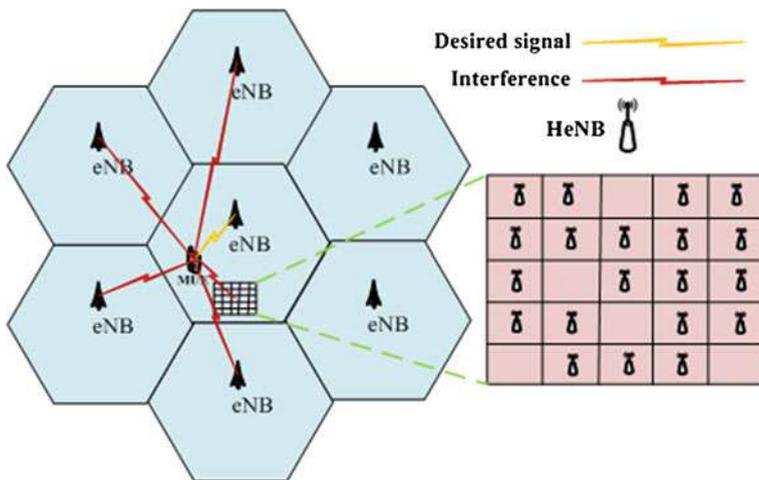


Fig. 33.1 Interference scenario

33.2.2 Propagation Models and SINR Computing

In this study we adopt the 3GPP LTE-Advanced path loss models for urban deployments [6], in which the path loss between the eNB and the UE is calculated as follows:

$$PL_{macro}(dB) = \begin{cases} 15.3 + 37.6\log_{10}R(\text{outdoor UE}) \\ 15.3 + 37.6\log_{10}R + L_{ow}(\text{indoor UE}) \end{cases} \quad (33.1)$$

where R is the distance between the UE and eNB in meters and L_{ow} is the penetration loss of an outdoor wall, which is 20 dB.

While the path loss between HeNB and UE within or outside an apartment for 5×5 grid scenario, is formulated as:

$$PL_{femto}(dB) = 127 + 30\log_{10}(R/1000) \quad (33.2)$$

The received SINR of the MUE in the downlink can be expressed as [7]:

$$SINR_{MUE} = \frac{PT_m g_{mj}}{\sum_{m \neq k} PT_k g_{kj} + \sum_{f \in F} PT_f g_{fj} + N_{th}} \quad (33.3)$$

$k \in M$

where PT_m , PT_k and PT_f are the transmit power of serving and interfering MeNB and HeNB respectively, M and F is the set of eNBs and HeNBs respectively, g_{mj} , g_{kj} and g_{fj} are the link gain between MUE j and the eNB and HeNB. N_{th} is the thermal noise.

The downlink received SINR of the HUE is given by [7]:

$$SINR_{HUE} = \frac{PT_f g_{fh}}{\sum_{k \in F} PT_k g_{kh} + \sum PT_m g_{mh} + N_{th}} \quad (33.4)$$

$k \neq f$

where PT_f and PT_k are the transmit power of the serving and interferer HeNBs respectively.

33.2.3 Spectral Efficiency Calculations

The spectral efficiency of a channel can be approximated by the attenuated and truncated form of Shannon bound [8]. Given a particular SINR, the spectral efficiency can be determined by the following:

$$\text{Throughput, } Thr[\text{bps/Hz}] = \begin{cases} Thr = 0 & SINR < SINR_{\min} \\ Thr = \alpha \cdot S(SINR) & SINR_{\min} < SINR < SINR_{\max} \\ Thr = Thr_{\max} & SINR > SINR_{\max} \end{cases} \quad (33.5)$$

where $S(SINR) = \log_2(1 + SINR)$ is Shannon bound, α is the attenuation factor, $SINR_{\min}$ and $SINR_{\max}$ are the minimum and maximum SINRs supported by the available AMC scheme. Thr_{\max} is the maximum spectral efficiency. The parameters α , $SINR_{\min}$, $SINR_{\max}$ and Thr_{\max} are set to 0.6, -10dB , 23 dB and 4.4 bps/Hz in this study [8].

33.3 Proposed Power Control Scheme

The basic requirement for a HeNB is to provide a strong enough signal for its HUEs. On the other hand the transmit power should not be too large, as to create strong interference to the neighbouring femtocells or MUEs. In the fixed power setting scheme (i.e. no power control) all the HeNBs transmit at their highest available transmit power without considering any surrounding information.

33.3.1 Analytical Model

In this paper, we propose a distributed HUE-assisted based power control scheme. In which the transmit power (P_{tx}) is adjusted according to the distance between the HUE and the HeNB as in (33.6):

$$P_{tx} = \min \left(\max \left\{ P_{t\max} \times \left(\frac{d}{R} \right)^K, P_{t\min} \right\}, P_{t\max} \right) \quad (33.6)$$

where d is the distance between the HUE and HeNB. R is the radius of the femtocell. $P_{t\max}$, $P_{t\min}$ are the maximum and minimum transmitting power of HeNB respectively. $P_{t\min}$ depends on the SINR target related to the QoS requirements of the HUE. K is an exponent that controls the dynamic range of power control.

However, it is more convenient and practical to estimate the parameter d from the reported path loss between HUE and the corresponding HeNB, without acquiring HUE position information. The path loss can be given as follows:

$$PL = x(f) + N \log_{10}(d) + x_{\sigma} \quad (33.7)$$

where $x(f)$ represents the dependence of the path loss on the frequency. N is a coefficient related to the type of environment. x_{σ} denotes the shadow fading, is a Gaussian random variable with zero mean σ^2 variance.

The distance (d) can be expressed as an exponential function:

$$d = 10^{\frac{1}{N} \times (PL - x(f) - x_{\sigma})} \quad (33.8)$$

This can be written as:

$$d = 10^{\frac{1}{N} \times P_{UE}} \quad (33.9)$$

Considering (33.8), the formula (33.5) can be modified in the following manner:

$$P_{tx} = \min \left(\max \left\{ P_{tmax} \times \left[10^{\frac{1}{N} \times (P_{UE} - P_{max})} \right]^K, P_{tmin} \right\}, P_{tmax} \right) \quad (33.10)$$

The P_{tmin} is calculated so as to guarantee the target SINR of the HUE. Furthermore the pathloss is partially compensated [9] as in (33.11):

$$P_{tmin} = \max(\min(\text{SINR}_{tar} + I_{serv} + \alpha PL_{serv}, P_{tmax_HeNB}), P_{tmin_HeNB}) \quad (33.11)$$

where I_{serv} is the average link interference, PL_{serv} is the the pathloss between HUE and the serving HeNB, α is a fractional compensation parameter between 0 and 1 note that if $\alpha = 1$, then we have full pathloss compensation. P_{tmin_HeNB} is the minimum allowed power that could be transmitted by HeNB which is set to 0 dBm as in 3GPP standard.

33.3.2 Algorithm Description

The proposed scheme makes use of local status information of the HeNB; obtained measurement reports from connected HUEs. The HeNB utilizes this information and based on the defined algorithm routines, makes a local decision.

Each HUE has a target SINR (SINR_{tar}) that satisfied the quality of service (QoS) for him. The minimum transmit power of the HeNB is set so as to guarantee SINR_{tar} for the HUE. To eliminate the burden of frequent P_{tx} update according to

the mobility of all the attached HUEs, Ptx update is associated with the arrival of new HUE. However, there is also a periodical update if there is no new user for a dedicated period of time. Two routines are used to calculate the transmit power: (i) Cal_PWR_SINRtar, in which Ptx is calculated according to (33.10). (ii) Dec_power, in which the Ptx is decreased by a step ΔPtx that is equal to the difference between the current SINR of the HUE and his SINRtar. The algorithm for the proposed scheme is shown in the following pseudo code.

Algorithm of the proposed scheme

```

1: for i=1:no_HeNB do
2:   if (new HUE attached to HeNB) then
3:     sort descending the distance of all the connected UE
4:     store the sorted distances in an array  $K_i$ 
5:     pick the HUE with the farthest distance  $k_i$ 
6:     if ( $k_i$  is the distance for the HUEnew) then
7:       if (SINR of HUEnew < SINRtar) then
8:         go to routine: Cal_PWR_SINRtar
9:         increase the number of power updates by 1
10:      elseif (SINR of HUEnew > SINRtar) then
11:        go to routine: Dec_power
12:        increase the number of power updates by 1
13:      end if
14:    else
15:      break
16:    end if
17:  else
18:    break
19:  end if
20:  each 10 Sec check the following
21:  if (no. of PWR updates=0) then
22:    repeat steps 3,4 and 5
23:    if (SINR of HUE with distance  $k_i$  > SINRthr) then
24:      repeat step 8
25:    else
26:      repeat step 11
27:    end if
28:  else
29:    reset the no. of PWR updates to zero
30:  end if
31: end for

```

Table 33.1 Simulation parameters

Parameter	Macrocell	Femtocell
Cell radius	500 m	10 m
HeNB transmitter power	46 dBm	20 dBm
HeNB antenna gain	14 dBi	5 dBi
Log normal shadowing standard deviation	8 dB	10 dB
Carrier frequency	2.0 GHz	
Bandwidth	10 MHz	
Antenna pattern	Omnidirectional	
Thermal noise	−174 dBm/Hz	

33.3.3 Simulation Scenario

The downlink macrocell/femtocell scenario as in Fig. 33.1 is considered with seven eNB at the centre of the macrocell. A typical 5×5 grid scenario of a dense urban area for femtocells with 25 houses each house has the size of $10 \text{ m} \times 10 \text{ m}$ is located at the edge of the centred macrocell. Twenty HeNBs are located randomly and uniformly in the houses. The closed access method is considered for HeNBs. This means only closed subscriber group (CSG) is allowed to connect to the femtocell. HUEs are dropped randomly and uniformly in the femtocells. The SINR target range for HUEs is $-4 \sim 20 \text{ dB}$. Table 33.1 shows the main simulation parameters confirmed with 3GPP TR 36.814 [10].

33.4 Results and Discussion

The outage probability of MUEs is defined as the ratio of MUEs whose SINR bellow -6 dB to the total number of MUEs [12]. Figure 33.2 shows the average outage probability of MUEs when conducting the simulation with different scenarios of no power control is applied and with different values of α for power control. For $\alpha = 1$ (i.e. full compensation for pathloss) the reduction in outage is not significant, however, with small (α) a clear improvement in MUE outage is obtained. As we consider the closed access mode for femtocells; the less the HeNB transmit power, the more the interference reduction could be obtained and in consequence the MUE outage is minimized.

The proposed scheme not only reduced the cross-tier interference, but furthermore the co-tier interference between femtocells is mitigated and that would explain the enhancement in the spectral efficiency (SE) of the HUEs despite the reduction of the transmit power as in Fig. 33.3. For small (α) the HUE SE is still better than no power control scenario, but it is less than when $\alpha = 1$.

Fig. 33.2 Average outage probability of MUEs

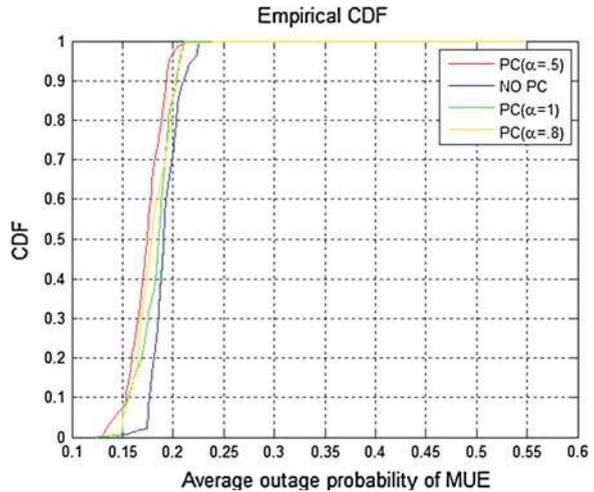
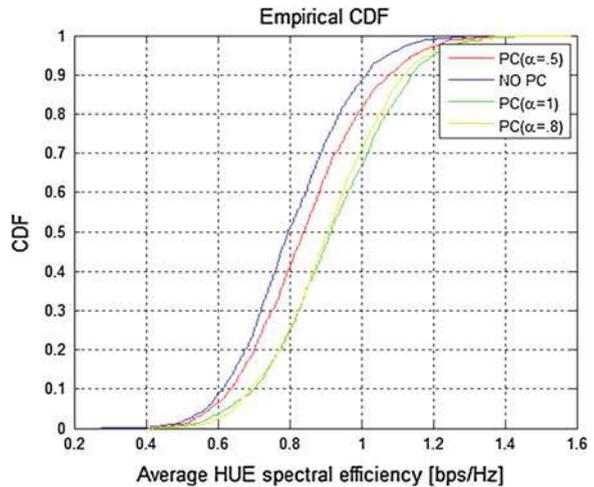


Fig. 33.3 Average spectral efficiency of HUE



As a result of mitigating the cross-tier interference, the signal quality is improved for MUEs as shown in Fig. 33.4. Higher MUE SINR is obtained with $\alpha = 0.5$. Improving SINR for MUEs indicates the ability to enhance the MUE throughput.

As the transmit power of the base station represents the major part that contributes to the power consumption. Applying the proposed scheme provides power consumption reduction of 17 % with $\alpha = 1$ compared to no power control scenario where all the HeNBs transmit at their maximum power. While a power reduction of 86.5 % is obtained with $\alpha = 0.5$. These savings are necessary for deploying the future femtocells in a cost effective and green way (Fig. 33.5).

Fig. 33.4 Average SINR of MUEs

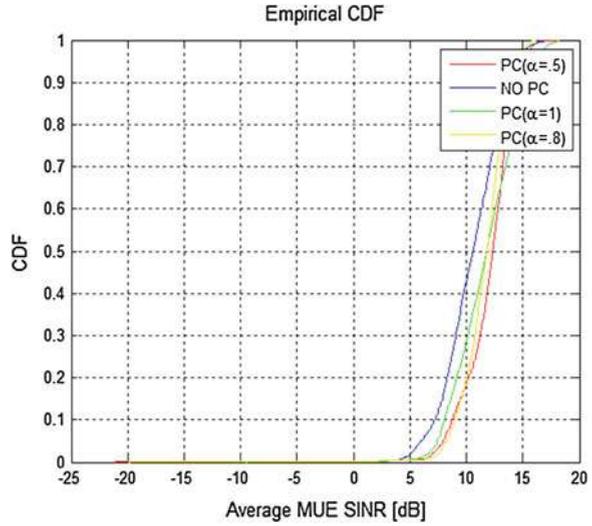
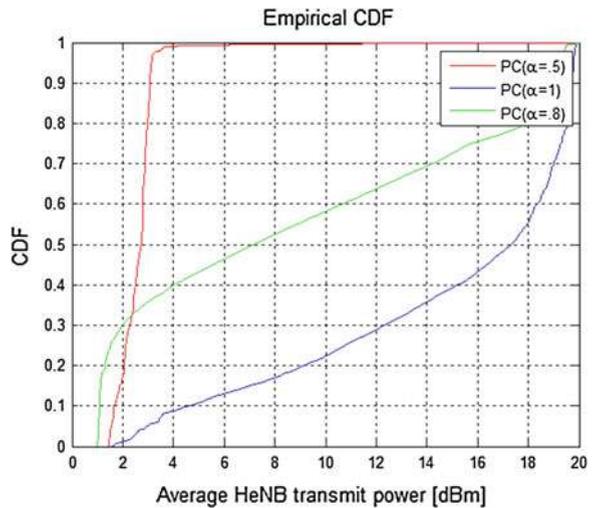


Fig. 33.5 Average HeNB transmit power



33.5 Conclusion

In this paper a power control scheme based on HUE measurements is proposed. In this scheme, the transmit power of the HeNB is adjusted based on the location of the HUE that is estimated from the path loss measurements for simplicity. Furthermore, the minimum level of power transmission is constrained to the target SINR that could be set according to the QoS requirements. In addition the minimum transmit power level is calculated based on partial compensation for the

pathloss which helps more in mitigating the interference to the neighbouring cells. The system level simulation results show that the proposed scheme has reduced the outage probability for the MUEs. Moreover, minimizing the transmit power, according to the position of HUE helps to reduce the power consumption for greener deployment of femtocells while maintaining good performance for HUEs in terms of spectral efficiency. Our next work will focus on finding the optimal value of α that compromise between the outage of the MUEs and the SE of the HUEs.

Acknowledgments This research was supported by The National University of Malaysia under grant DPP-2013-006 and Ministry of Science, Technology and Innovation Malaysia under grant 01-01-02-SF0788.

References

1. Chandrasekhar, V., Andrews, J.G., Gatherer, A.: Femtocell networks: a survey. *IEEE Commun. Mag.* **46**(9), 59–67 (2008)
2. Saquib, N., Hossian, E., Lee, L.B., Kim, D.I.: Interference management in OFDMA femtocell networks: issues and approaches. *IEEE Wireless Comm. Mag.* **19**(3), 86–95 (2012)
3. Cho, K.-T., Kim, J., Jeon, G., Ryu, B.H., Park, N.: Femtocell power control by discrimination of indoor and outdoor users. *Proceedings of the IEEE Wireless Telecommunications Symposium (WTS)*, pp. 1–6 (2011)
4. Akbudak, T., Czyliw, A.: Distributed power control and scheduling for decentralized OFDMA networks. In: *Proceedings of International ITG Workshop on Smart Antenna*, pp. 59–65 (2010)
5. Wang, Z., Xiong, W., Dong, C., Wang, J., Li, S.: A Novel downlink power control scheme in LTE heterogeneous network. In: *Proceedings of IEEE International Conference on Computational Problem-Solving (ICCP)*, pp. 241–245 (2011)
6. GPP tech. rep. 36.814. v.9.0.0. Mar. 2010
7. Wang, M., Zhu, X., Zeng, Z., Wan, S., Li., W.: System performance analysis of OFDMA-based femtocell networks. In: *Proceedings of IET International Conference on Communication Technology and Application (ICCTA)*, pp. 405–410 (2011)
8. GPP tech. rep. 36.942. v.11.0.0. Sept. 2012
9. Rao, A.M.: Reverse link power control for managing inter-cell interference in orthogonal multiple access systems. In: *Proceedings of IEEE 66th Vehicular Technology Conference (VTC)*, pp. 1837–1841 (2007)

Chapter 34

Design and Development of the Visible Light Communication System

Anuar Musa, Mazlaini Yahya, Nazaruddin Omar,
Mohd Kamarulzamin Salleh and Noor Aisyah Mohd Akib

Abstract This paper presents the development of a Visible Light Communications (VLC) system for a local area network (LAN) based on the IEEE 802.15.7 standard. The proposed VLC system enables new indoor applications by providing simultaneous illumination and networking capability between the optical wireless devices. The VLC system uses an AVAGO Technologies HSDL-4230 LED as the transmitter and an OSRAM SFH203FA photodiode as the receiver. The distance in which the optical link was measured was up to 1 meter and achieved about 9.2 Mb/s throughput. Therefore, it is suitable for most of the applications for the home or office environment such as Internet communication, multimedia streaming, etc.

34.1 Introduction

In recent years, the interest in optical wireless as a favorable complementary technology for the radio frequency (RF) communication has gained significant momentum with substantial deployments using solid state lighting technology and the released of the IEEE 802.15.7 draft standard [1]. The IEEE 802.15.7 compliant networks consist of two types of nodes, namely, a coordinator which initializes and

A. Musa (✉) · M. Yahya · N. Omar · M.K. Salleh · N.A.M. Akib
TM Innovation Centre, TM Reserach and Development Sdn. Bhd., Lingkaran Teknokrat
Timur, 63000 Cyberjaya, Selangor, Malaysia
e-mail: anuarmusa@tmrnd.com.my

M. Yahya
e-mail: mazlaini@tmrnd.com.my

N. Omar
e-mail: nazar@tmrnd.com.my

M.K. Salleh
e-mail: kamarulzamin@tmrnd.com.my

N.A.M. Akib
e-mail: aisayah@tmrnd.com.my

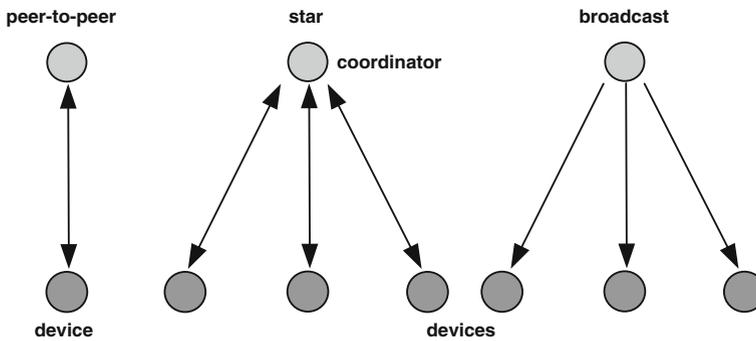


Fig. 34.1 IEEE 802.15.7 standard defines three types of topologies namely peer-to-peer, star, and broadcast

manages the network and a remote device which communicates with each other via the coordinator.

The standard defines three types of topologies consisting peer-to-peer, star, and broadcast as shown in Fig. 34.1. The standard also defines the first two layers of ISO/OSI stack protocol that includes Physical and Medium Access Control (MAC) layers. IEEE 802.15.7 MAC is a Carrier Sense Medium Access/Collision Avoidance (CSMA/CA) based protocol.

The basic concept of the VLC is simple. It uses a visible light medium to carry data through the air or space. Likewise, the VLC link architectures are very similar to an optical fiber communication links, without the optical fibers, lights are deployed as a transmission medium. The VLC is also similar to the RF wireless transmission with the radio waves replaced with a light-wave medium and the antennas are replaced with the free-space optical transceivers.

Despite this superficial similarity between the VLC and the RF links, the VLC exhibits several appealing attributes when compared to the RF system. The VLC links are inherently broadband. The optical frequencies are in the light-wave infrared and visible spectrums that are neither regulated nor required license. The optical components are also cheaper and consume less electrical power than the high-speed RF components. In fact, the VLC links do not suffer from severe multipath fading and interferences like the RF transmission since the transmission beam is focused and requires direct line-of-sight (LOS) transmission. These advantages do not, however, imply that the VLC is a universal replacement for the RF communications system.

The applications of the VLC systems are quite limited when considering area coverage and user mobility where the RF technologies are proven invaluable. In addition, the VLC systems operate under strict eye safety regulations. The VLC receiver is less sensitive than their RF counterparts because of their photo-electric conversion mechanisms and the impact of ambient light noise sources.

Substantial researches have been conducted into the VLC technology and have created a number of test beds presented in the journals [2–4]. Many of these test

beds do not follow IEEE 802.15.7 standard. In this paper, we propose to transmit a standard LAN 10/100BASE-T (10/100 Mb/s) Ethernet signal over a VLC link as a proof-of-concept for future homes, offices, hospitals as well as for providing an optical wireless LAN (VLC-LAN) for the end users.

This paper focuses on the design and the development of the IEEE 802.15.7 system by combining both lighting and communication attributes. This paper also will outline the requirements to achieve an effective VLC-LAN system and documents the properties of the proposed system. The rest of the paper is organized as follows. In Sect. 34.2, we present the system design and implementation. System performance is discussed in Sect. 34.3. Finally, we present our conclusions in Sect. 34.4.

34.2 System Design and Implementation

The proposed proof-of-concept (POC) system architecture is shown in Fig. 34.2. The VLC system uses an AVAGO Technologies HSDL-4230 [5] LED as the transmitter and an OSRAM SFH203FA [6] photodiode as the receiver.

The proposed transceiver block diagram is outlined in Fig. 34.3. At the transmitter end, the output of the laptop passes the packets through a modem, as an interface between the laptop and the optical transceiver and to the amplifier. To prevent flickering and brightness changes due to the transmission of other messages than the idle pattern, such as data messages, we use Manchester Coding [7] to distribute the occurrences of ‘1’ and ‘0’ more evenly; the human observer will perceive an always-on optical signal unnoticeable of very high speed flickering.

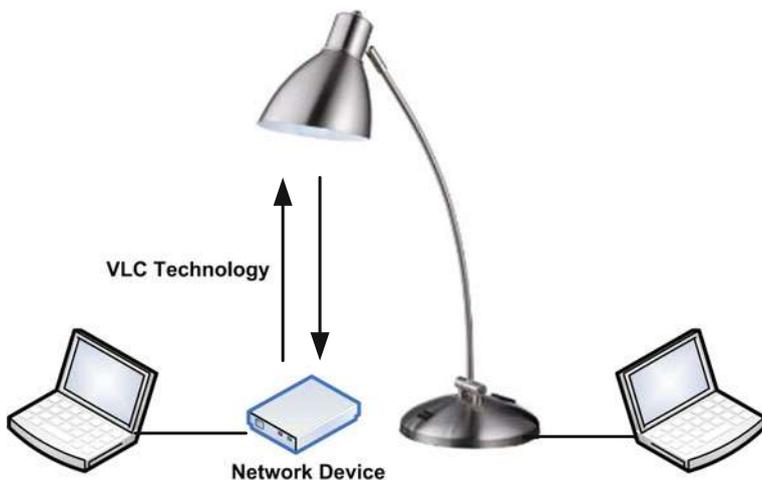


Fig. 34.2 High level system architecture for the full-duplex visible light communication (VLC) system test bed

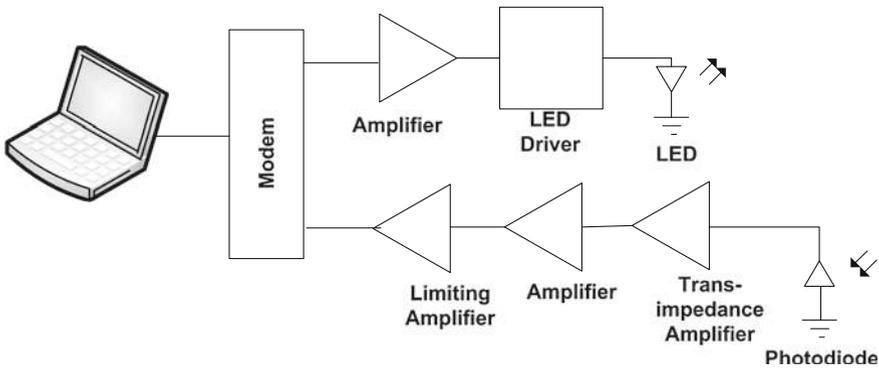


Fig. 34.3 Block diagram of the receiver and the transmitter for the uplink. Similar block diagram for the downlink with minor different in LED and Photodiode due to different wavelength used to avoid interference. Tables 34.1 and 34.2 provide the key implementation parameters

The amplifier amplifies the signal before sending the signal to the LED driver. The LED converts the electrical signal into optical signal for transmission. At the receiver, a photodiode (PD) followed by a transimpedance amplifier (TIA) and an amplifier are used to recover the input signal. The PD shown in the receiver circuit (Fig. 34.3) is OSRAM SFH203FA. These devices offer a fast response and a large 7 mm active area useful for home environment where the practical distance is more than a few centimeters. At the end of the signal recovery, a limiting amplifier limits the receiver output signal from clipping effect before entering into the modem.

Tables 34.1 and 34.2 provide the specifications of the hardware used for the implementations of the transmitter and the receiver for the uplink and downlink channels, respectively.

Table 34.1 Keys Parameters of the Uplink VLC System

Device	Property	Value
LED	Infrared LED	HSDL 4230 [5]
	Power	75 m/sr
	Beam width	17°
	Voltage	2.5 V
	Current	100 mA
	Size	3 mm
Photo diode	Photo sensitivity	0.59 A/W
	Spectral response range	800-1100 nm
	Peak sensitivity wavelength	900 nm
	Active Area	1 mm

Table 34.2 Keys Parameters of the downlink VLC System

Device	Property	Value
LED	Pure white	STAR LED [8]
	Power	75 m/sr
	Beam width	127°
	Voltage	3.25 V
	Current	700 mA
	Size	8 mm
Photo diode	Photo sensitivity	0.46 A/W
	Spectral response range	340–1040 nm
	Peak sensitivity wavelength	760 nm
	Active area	7 mm

34.3 System Evaluation

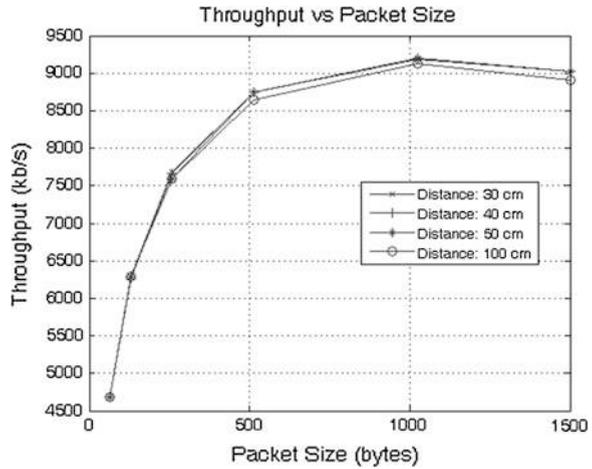
This paper provides the design and implementation of the VLC system for the desk lamp environment. When the system is deployed to the end-user, the information source can be as far away as 100 cm (desk lamp to receiving plane) in range. The actual system setup is shown in Fig. 34.4.

The UDP throughputs as function of packet size for the VLC system are shown in Fig. 34.5. The UDP throughput increases with packet size and reaches the maximum rate at 9.2 Mb/s. This is expected as the payload increase the UDP/IP overhead decreases; thus, the average throughput will also increase.

Fig. 34.4 Visible Light Communications (VLC) test bed for the desk lamp environment that is capable of transmitting and receiving a HD quality video



Fig. 34.5 Measured UDP throughput per packet size for VLC system in a standard office environment with different distances



34.4 Conclusion

This paper has presented a POC for a VLC LAN link. The distance achieved is one meter, which makes possible for a LAN internet link from an LED based desk lamp to the PC connection as a complementary technology to Wi-Fi. The link is successfully achieved by using a VLC driving circuit that consists of a series of amplifiers and feeds directly to the HSDL-4230 LED device. The incident light passes through the SFH203FA photodiode device, allowing 9.2 Mb/s communications. Future work will extend the POC with the design and development of the MAC layer to support multi-users for the communication system for practical usage in the homes and the offices environment.

References

1. Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications, IEEE 802.15.7-2011 draft8 standard
2. Elgala, H., Mesleh, R., Haas, H.: Indoor broadcasting via white LEDs and OFDM. *IEEE Trans. Consum. Electron.* **55**(3), 1127–1134 (2009)
3. Omega: The home gigabit access project. <http://www.ict-omega.eu/>, viewed 12-01-2012
4. Schmid, S., Gorlatova, M., Giustiniano, D., Vukadinovic, V., Mangold, S.: Networking Smart Toys with Wireless ToyBridge and ToyTalk. In: Poster Session, Infocom (2011)
5. HSDL-4230 datasheet. <http://www.farnell.com/datasheets/95278.pdf>
6. SFH203FA datasheet. <http://www.farnell.com/datasheets/1672046.pdf>
7. Proakis, J.G.: *Digital Communications*. McGraw-Hill Series in Electrical and Computer Engineering, New York (2001)
8. STAR-LED datasheet. <http://www.farnell.com/datasheets/1444587.pdf>

Chapter 35

The Embroidered Antenna on Bending Performances for UWB Application

M.S. Shakhirul, A. Sahadah, M. Jusoh, A.H. Ismail
and Hasliza A. Rahim

Abstract The performances of embroidered UWB wearable textile antenna due to the effect in bend conditions is discussed in this paper. The radiating element of the antenna is made of silver plated nylon thread by using embroidery technique, while the antenna substrate is cotton fabric and Nora dell as a ground plane. This UWB antenna is designed for operating frequency from 3.1 up to 10.6 GHz with a center frequency of 6.85 GHz. This research demonstrates the antenna bending effect at seven different angles in terms of the simulated and measured reflection coefficient.

35.1 Introduction

Nowadays, a lot of investigations on technology textile antenna to carry out the wearable monitoring system. The monitoring is a necessary for risky activities or environment such as mining, diving, mountain climbing as well as another sort of military applications. Moreover, the wearable antennas will be attached on the body or into clothing or may be worn as a button antenna. Hence, all these broad

M.S. Shakhirul (✉) · A. Sahadah · M. Jusoh · A.H. Ismail · H.A. Rahim
School of Computer and Communication Engineering, Universiti Malaysia Perlis
(UniMAP), Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia
e-mail: shakhirul@yahoo.com

A. Sahadah
e-mail: shahadah@unimap.edu.my

M. Jusoh
e-mail: muzammiljusoh11@gmail.com

A.H. Ismail
e-mail: abd.hafizh@gmail.com

H.A. Rahim
e-mail: haslizarahim@unimap.edu.my

applications of monitoring with data transmission function can be achieved by using wearable antennas that do not force the wearer to abandon the comfort zone with such compact and durable materials [1].

The commercial use of frequency bands from 3.1 to 10.6 GHz for ultra-wide-band (UWB) system has been approved by the FCC in 2002 [2]. With the advantage of life long battery, the UWB reduces the wearable device size since it uses low power signal to transmit data [3]. Therefore, this research applies the UWB textile technology with an embroidery technique for WBAN application.

The worn antenna needs to be low profile and hidden for the convenience of the user. Furthermore it is impossible to keep the wearable antennas in flat condition all the time because the antennas will bend quite often due to human body structure and movement. The bending could affect the performance characteristics of the antennas as its resonant length altered [4]. Therefore, an investigation of the effects antenna performance characteristics due to bending condition carried out in this paper.

35.2 Materials and Methods

The cotton fabric is used as substrate materials in this project instead of using rigid circuit boards. While the patch antenna is made of silver plated nylon thread and the conductive fabric “Nora dell” is used as a ground plane. The designing process of the antenna required the specification of the substrate materials such as dielectric constant, ϵ_r and the thickness, while the conductive required electrical permittivity. Table 35.1 shows the details of specification materials used in this paper [4].

35.2.1 Circular UWB Patch Antenna

A circular shape antenna is mostly suited to the specification of UWB application which provides several advantages such as omni-directional, low transmit power and large channel capacity [5, 6, 7]. Hence, this research proposed a circular antenna designed by Eq. (35.1), where a is radius of the circular patch antenna in millimeter; f_r is the resonance frequency and ϵ_r is the relative permittivity of the

Table 35.1 Specification of materials used

Materials	Thickness (mm)	Values
Cotton fabric	0.5	$\epsilon_r = 1.6$
Conductive fabric “Nora dell”	0.13	$1.538e + 6$
Silver plated nylon thread	0.105	–

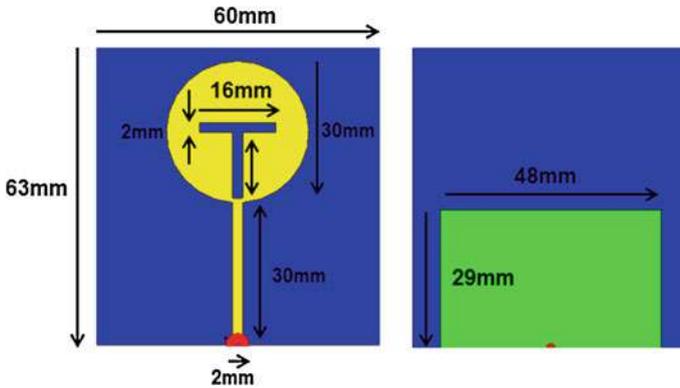


Fig. 35.1 Optimum dimension of antenna designed (The left sight is a front view while the right side is a back view of the antenna)

textile substrate material [1]. The resonance frequency and dielectric constant of this paper is 6.85 GHz and 1.6 respectively.

$$a = \frac{87.94}{f_r(\sqrt{\epsilon_r})} \tag{35.1}$$

All designs and simulations have been carried out using the CST simulation tool, however some modification has been made to suit the requirements. Figure 35.1 shows the geometry and the dimension for the proposed patch antenna.

35.2.2 Antenna Design in Bend Conditions

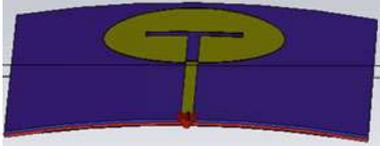
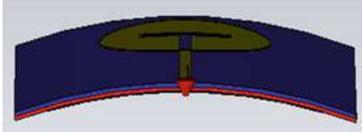
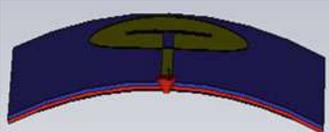
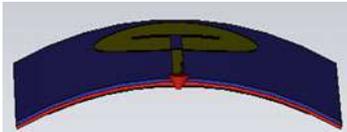
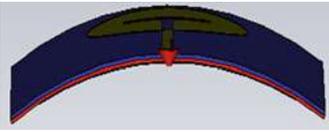
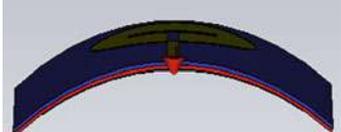
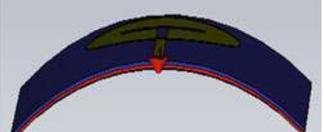
In this project, the bending effect of UWB textile antenna is investigated for seven different angles; 20°, 30°, 40°, 50°, 60°, 70° and 80°. The cylinder is used as reference structure of the antenna in order to be bent. Hence the Eq. (35.2) is used to find the outer radius of the cylinder [7].

$$S = r\theta \tag{35.2}$$

- Where: S = Arc length,
- r = Radius of the circle,
- θ = Measure of the central angle in radians.

The width of the substrate antenna was used as the arc length S is 60 mm. Based on the Eq. (35.2), the cylinder radius value of the intended angle was calculated in Table 35.2.

Table 35.2 The UWB antenna calculation for bending purpose

20° bending	
	
$\theta = \frac{20 \times \pi}{180} = 0.349$ (degree in radian)	
$r = \frac{s}{\theta} = \frac{60}{0.349} = 171.92$ mm (radius of cylinder)	
30° bending	40° bending
	
$r = 114.59$ mm	$r = 85.94$ mm
50° bending	60° bending
	
$r = 68.75$ mm	$r = 57.29$ mm
70° bending	80° bending
	
$r = 49.11$ mm	$r = 42.97$ mm

35.2.3 Antenna Fabrication

The embroidered UWB textile antenna is fabricated using a computerized sewing machine. The particular machine has sewn the silver plated nylon thread on the cotton fabric. In embroidery technique, thread is sewn on the substrate will be penetrated at the back of the substrate. Hence, another layer of cotton will be inserted as to separate between the conductive thread and Nora dell as ground plane which means there are two layers of cotton fabric.

The process is continued by sewn the second layer of cotton with the ground plane denotes as “Nora dell” fabric. The first layer cotton is the embroidery circular patch antenna while the second layer cotton is sewn with the ground plane of “Nora dell”. The special glue called silver conductive epoxy is used to solder

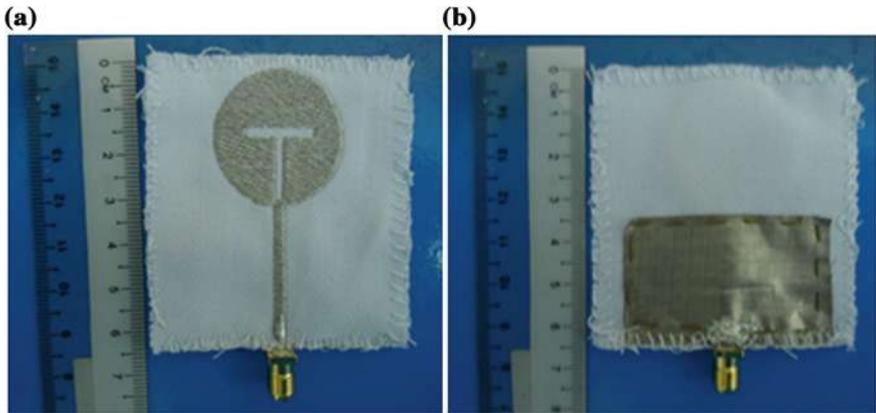


Fig. 35.2 Prototype of an embroidered UWB antenna. **a** Front view **b** Back view

the SMA connector with the embroidery antenna. Figure 35.2 shows the prototype of UWB embroidered antenna.

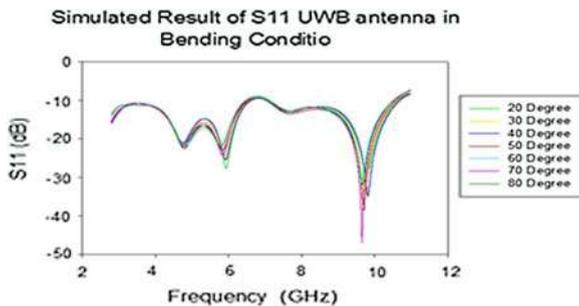
35.3 Results and Discussions

All measurements are done in free space using the Programmable Network Analyzer functioned between 100 and 20 GHz. The antenna reflection coefficient performance is evaluated based on the bending condition.

35.3.1 Simulated Result of UWB Antenna in Bending Conditions

The simulated reflection coefficient is done between 2.8 and 11 GHz under seven different bending angles; 20°, 30°, 40°, 50°, 60°, 70° and 80°. Figure 35.3 shows

Fig. 35.3 The simulated comparison of S-Parameter result for antenna in bends condition



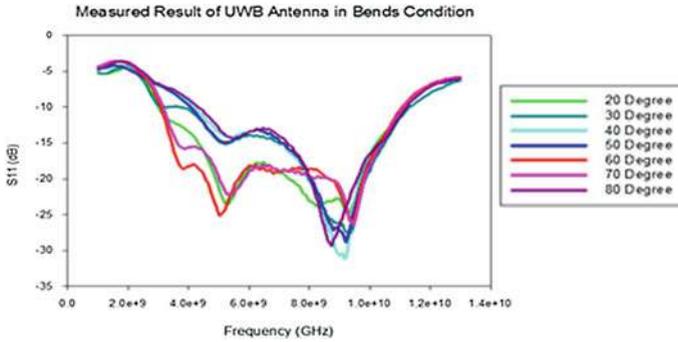


Fig. 35.4 The measured comparison of S-Parameter result for antenna in bends condition

the simulated comparison of S11 (dB) result for antenna in bends condition. The antenna shows the similar pattern of return loss at several degrees of antenna bends. At 70° degree bending, the antenna achieved a significant reflection coefficient of -46.75 dB at frequency 9.65 GHz. The antenna at 70° also has a lower resonance where the reflection coefficient is -12.53 dB at frequency 7.7 GHz.

35.3.2 Measured Result of UWB Antenna in Bending Conditions

Figure 35.4 shows the measured comparison of S-parameter result for antenna in bends conditions. By referring to the measured S-parameter result of antenna in bend condition, the antenna bends at 60 and 70° have a similar pattern of reflection coefficient follow by antenna bend at 20°. The antenna in these bends conditions which is 20° (green line), 60° (red line) and 70° (pink line), shown the better reflection coefficient matching. While the others bend angles shared a similar reflection coefficient pattern. However, the proposed antenna still can performed under tolerable reflection coefficient of less than -10 dB.

35.4 Conclusion

This paper presents the development, fabrication and analysis of the embroidered UWB antenna performance between frequency ranges of 3.1 GHz to 10.6 GHz. The investigation focused on the bending effect towards the antenna reflection coefficient result. There are seven bending angles considered; 20°, 30°, 40°, 50°, 60°, 70° and 80°. The result shows that the antenna has different performances

according to the bending angles. Regardless of the bending angles, both simulated and measured results have slightly difference in reflection coefficient. The higher the bending angles, the better the impedance matching of the proposed antenna.

Acknowledgments Acknowledgements are expressed to Research and Development (RND) Centre, School of Computer and Communication Engineering and University of Malaysia Perlis for providing the lab facilities and short term grant which enabled the publication of this article.

References

1. Osman, M.A.R., Rahim, M.K.A., Samsuri, N.A., Elbasheer, M.K., Ali, M.E.: Textile UWB Antenna bending and wet performances. *Int. J. Antennas Propag.* **2012**(251682) 12 (2012)
2. Sanz-Izquierdo, B., Huang, F., Batchelor, J.C.: Convert dual-band wearable button antenna. *Electron. Lett.* **42**(12668–12670) 2006
3. Osman, M.A.R., Rahim, M.K.A., Azfar, M., Kamardin, K., Zubir, F., Samsuri, N.A.: Design and analysis UWB wearable textile antenan. In: *Proceedings of the 5th European Conference on Antenna and Propagation (EUCAP)* (2011)
4. Shieldex website: <http://itp.nyu.edu/~kh928/sensorreport/ShieldexNoraDell.pdf>
5. Klemm, M., Troster, G.: Textile UWB antenna for on-body communications. In: *Proceedings of 'EUCAP 2006, Nice, France. 6–10 Nov 2006*
6. Soh, P.J., Vandenbosch, G.A.E., Javier H.-O.: Design and evaluation of flexible CPW-fed Ultra Wide Band (UWB) Textile antennas. In: *IEEE International RF and Microwave Conference (RFM 2011), 12–14 Dec 2011*
7. Dey, S., Saha, N., Biswas S.: Design and Performance analysis of UWB Circular disc monopole textile antenna and bending consequences. In: *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP), 11–15 Apr 2011*

Chapter 36

The Embroidered Wearable Antenna for UWB Application

M.S. Shakhirul, A. Sahadah, M. Jusoh, A.H. Ismail, C.M. Nor and F.S. Munirah

Abstract This paper discusses the design of a wearable textile antenna for Ultra Wide Band (UWB) application. This embroidered antenna addresses the issues of miniature size, wide bandwidth and low power consumption. A textile cotton has been chosen as a substrate and silver nylon plated yarn as a conductive element embroidered on cotton substrate. Simulated and measured results show that the proposed antenna design meets the requirements of wide working bandwidth with compact size and flexible material. The performances in terms of reflection coefficient, impedance bandwidth, current distribution as well as gain and antenna efficiency are compared between simulations and measurements and good agreement was observed.

M.S. Shakhirul (✉) · A. Sahadah · M. Jusoh · A.H. Ismail · C.M. Nor
School of Computer and Communication Engineering, Universiti Malaysia Perlis
(UniMAP), Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia
e-mail: shakhirul@yahoo.com

A. Sahadah
e-mail: shahadah@unimap.edu.my

M. Jusoh
e-mail: muzammiljusoh11@gmail.com

A.H. Ismail
e-mail: abd.hafizh@gmail.com

C.M. Nor
e-mail: cmnor@unimap.edu.my

F.S. Munirah
Fakulti Kejuruteraan Elektrik, Universiti Teknologi Mara (UiTM), 13500 Permatang Pauh,
Penang, Malaysia
e-mail: sitimunirah.f@gmail.com

36.1 Introduction

The development of wearable textile antenna becomes popular due to wide application of personal communication, wireless sensor and medical application. Other advantages that textiles antenna offer are, washable, flexibility, light weight, low cost and reliability. Generally, there are two manufacturing techniques to produce wearable textile antenna [1]; (i) Using conductive fabric fixed on the non conductive textile fabric and (ii) Using conductive textile yarns to weave, knit or embroider the conductive pattern of the antenna on the textile fabric.

The designing wearable textile antenna is quite challenging due to the effect of lossy environment such as humidity and temperature. Dimension wise, the selection of a proper textile in terms of dielectric permittivity, conductivity and loss tangent have to be determined [2, 3].

The performance of the embroidered UWB textile antenna is studied at a frequency range between 3.1 and 10.6 GHz. The embroidery technique being preferred due to its robustness and esthetic value. Our study focuses on the performance effectiveness cause by the T-slot. The antenna embroidered circular antenna with and without T-slots performance is verified by comparing measurement and simulation result.

36.2 Materials and Method

Instead of using rigid circuit boards, cotton fabric is used as substrate, conductive fabric “Nora dell” as a ground plane and silver plated nylon thread as patch antenna. The properties of selected conductive fabrics may optimize the characteristics of the designed textile antenna in a specific application. Some of the electro-textiles properties they are flexible for deformation when worn, low electrical resistance to minimize losses, lightweight and comfortable. Therefore, the “Nora dell” has been selected. There are three Nora dell elements which are nickel, copper and nylon silver. Nora dell proposed a highly protective from galvanic corrosion and extremely flexible to the harsh environment of 90 °C temperature [3].

Regarding fabrication techniques for textile antenna, a conductive thread need to be used in the embroidery techniques. Silver plated nylon thread is chosen as conducting materials that provide high quality conducting thread. According to the manufacturer specifications, this conducting material provides superior strength, ability to resist the normal conditions of use such as multiple deformations for wearable applications. In fact, the conducting thread can be washed with the ability to resist temperature up to 150 °C [4]. In embroidery technique, thread is sewn on the substrate will be penetrated at the back of the substrate. Hence, another layer of cotton will be inserted as to separate between the conductive thread and Nora dell as ground plane which means there are two layers of cotton fabric. Table 36.1 depicted the material used and their dimensions.

Table 36.1 Detailed of material used

Antenna components	Material	Dimension (mm)
Circular patch antenna	Silver plated nylon thread embroidered throughout the pattern on a cotton material	Radius = 15
Substrate	Cotton	Thickness, $h = 0.5$
Ground plane	Conductive fabric—"Nora dell"	Thickness, $h = 0.13$
Microstrip feed line	Silver plated nylon thread	30×2

Table 36.2 Dimension of Modified Patch Antenna

Parameter	Value
Patch width, W	60 mm
Effective dielectric constant, ϵ_{eff}	1.5025
Patch length extension, ΔL	0.3 mm
Patch length, L	63 mm
Effective patch length, Le	17.91 mm

It is important to determine the relative permittivity and thickness of the textile materials. The device used to measure dielectric constant is E8362B Portable Network Analyzer (PNA) operates from 10 MHz to 20 GHz while a vernier caliper is used to measure the thickness of the textile materials.

A circular shape is mostly suited to UWB antenna which provides several advantages such as omni-directional, low transmit power and large channel capacity that suited the specification of UWB characteristic [5–7]. In this paper, two prototypes are simulated and their performances are compared.

A CST simulation tool has been used in designing the UWB antennas, however some modification has been made to suit the requirements. Initially a circular shape patch antenna was designed using cotton as a substrate with dielectric constant ϵ_r of 1.6. The resonance frequency, f_r is 6.85 GHz. The radius is calculated by using Eq. (36.1) while the patch antenna dimensions are calculated using basic equations for microstrip patch antenna [8]. Table 36.2 shows the summarized dimension for patch antenna and Fig. 36.1 shows the geometry of a modified design antenna for first and second designs respectively.

$$a = \frac{87.94}{fr(\sqrt{\epsilon_r})} \tag{36.1}$$

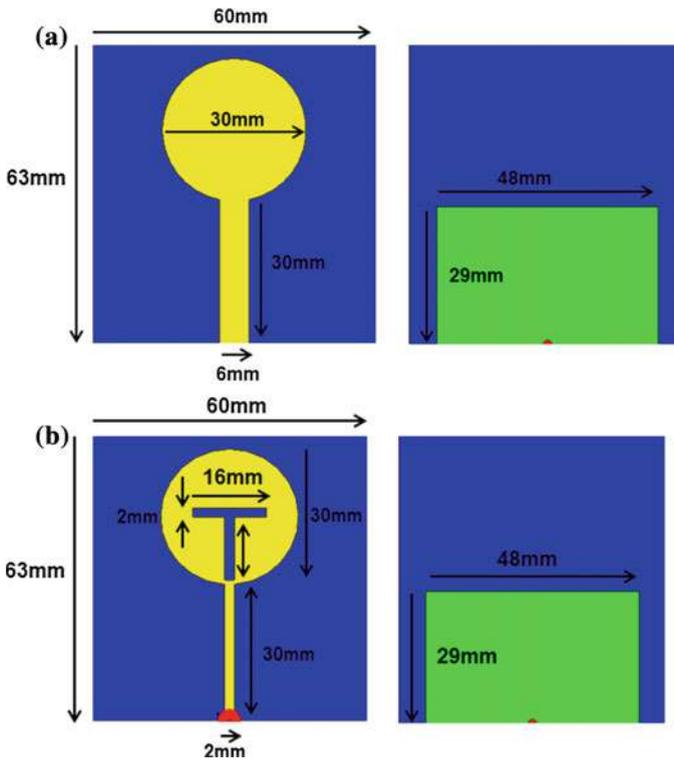


Fig. 36.1 Optimum antenna dimension. **a** First design the left sight is a front view while at the right sight is a back view of the antenna, **b** second design the left sight is a front view while at the right sight is a back view of the antenna

36.3 Results and Discussions

All designs and simulations of both antennas are simulated by using CST. The performance of the antenna is evaluated based on the reflection coefficient, gain, radiation pattern and the impedance bandwidth.

Figure 36.2 shows a reflection coefficient comparison between UWB embroidered antenna without T-slot and with T-slot. The simulated result is plotted from 2 to 11 GHz. It shows fluctuates pattern with a narrowband under the acceptable reflection coefficient of less than -10 dB. Moreover, the impedance bandwidth is 2.47 GHz.

Gain is one of the significant parameters that can determine the performance of the antenna. Table 36.3 shows the variation of frequencies against directivity, gain and efficiency for both antennas. First design indicates the lowest gain of 3.342 dBi at 6 GHz while the highest gain is 5.384 dBi at 8 GHz. The highest efficiency is 93 % at 4 GHz while at 10 GHz, the efficiency recorded the lowest

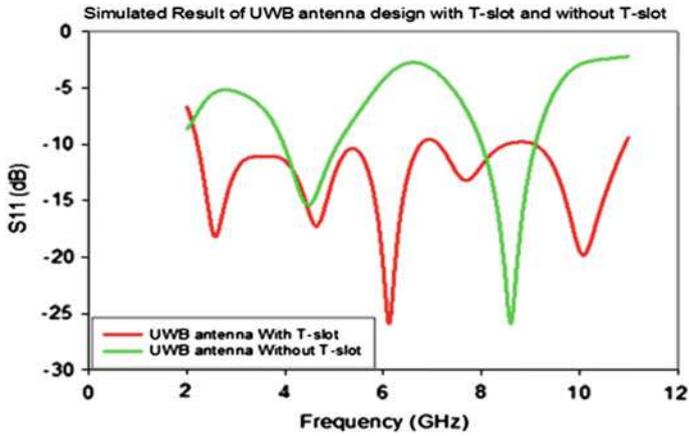
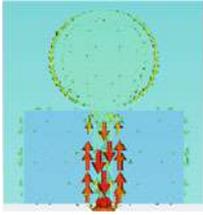
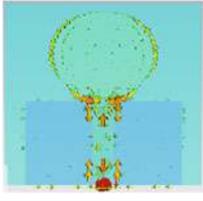
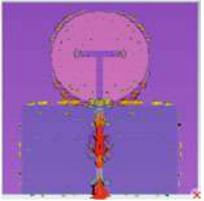
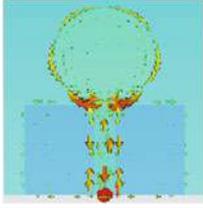


Fig. 36.2 Comparison of simulated reflection coefficient (S_{11}), between with T-slot and without T-slot

Table 36.3 Comparison of gain, directivity and % efficiency

Frequency (GHz)	Gain (dB)	Directivity (dBi)	% Efficiency
<i>First design of UWB embroidered antenna without T-slots</i>			
3	4.079	4.075	73
4	4.042	4.051	93
5	3.624	3.661	90
6	3.342	3.436	63
7	4.778	4.899	57
8	5.384	5.479	92
9	5.219	5.336	89
10	5.347	5.515	52
<i>Second design of UWB embroidered antenna with T-slot</i>			
3	3.753	3.793	93
4	3.719	3.748	92
5	3.507	3.581	93
6	3.053	3.118	97
7	4.778	4.597	88
8	4.234	4.310	92
9	4.006	4.068	89
10	4.535	4.608	96

Table 36.4 Current distribution for first and second design

Frequency (GHz)	First design: UWB embroidered antenna without T-slots	Second design: UWB embroidered antenna with T-slot
3		
6		
9		

value of 52 %. Generally, some of the gains are more than 5 dBi that depicted some frequencies using high power consumption. Therefore, some modification of the first design should be introduced to achieve low gain with large bandwidth and low power consumption that can cover up the entire frequency of UWB applications. The enhancement considered on the port width, partial ground and T-slot.

Antenna with T-slot achieved frequency range between 3 and 10 GHz. Maximum gain of 4.778 dBi is obtained at 7 GHz while the minimum gain is 3.053 dBi at 6 GHz. The second design has 97 % of highest antenna efficiency at 6 GHz while the lowest efficiency is 88 % at 7 GHz. Overall, the average gain is less than 5 dBi throughout the frequencies ranges from 3 to 10 GHz. Thus, the antenna has low power consumption and meets the specifications of the design.

Table 36.4 presented the current flows for both antennas designed at different frequencies; 3 GHz as low frequency, 6 GHz as medium frequency and 9 GHz as high frequency. High strength of current is radiating along the transmission line and the boundary of the patch after applying T-slot at the circular patch antenna. Moreover, the boundary of a partial ground plane also is a significant radiation area. Hence, the partial ground plane provides greater impedance bandwidth and used as low power consumption that suited the UWB applications. The UWB is achieved by considering suitable gap between partial ground planes and the antenna.

36.4 Conclusion

This paper presents the development, fabrication and evaluation of the antenna's performance between frequency ranges of 3.1–10.6 GHz. Two embroidered UWB antennas using cotton as substrate were considered. The investigation focussed on the performances of both antennas with and without T-slots. Based on the comparison of both designs, it shows that the antenna with T-slot achieved a better impedance matching with a wider bandwidth compared to the antenna without T-slot. In terms of gain and efficiency, the second design provides low gain with better performance. The gain is less than 5 dBi and basically used low power consumption. Hence, the modified second design provides low gain, omni-directional pattern and used low power which suited the performance of the UWB specifications. Moreover, the antenna with T-slot has a high efficiency of data rate. Further work has been done in another conference paper that focusing more on the bending issue.

Acknowledgments Acknowledgements are expressed to Research and Development (RND) Centre, School of Computer and Communication Engineering and University of Malaysia Perlis for providing the lab facilities and short term grant which enabled the publication of this article.

References

1. Maleszka, T., Kabacik, P.: Bandwidth properties of embroidered loop antenna for wearable applications. In: Proceedings, 3rd European Wireless Technology Conference, 2010
2. Van Langenhove, L.: Smart Textiles for Medicine and Health Care. CRC Press, Cambridge (2007)
3. Rahmat-Samii, Y.: Wearable and implantable antennas in body centric communications. Los Angeles, CA 90095, USA, 2010
4. Shieldex website: <http://itp.nyu.edu/~kh928/sensorreport/ShieldexNoraDell.pdf>
5. Osman, M.A.R., Rahim, M.K.A., Samsuri, N.A., Salim, H.A.M., Ali, M.F.: Embroidered fully textile wearable antenna for medical monitoring applications. *Prog. Electromagn. Res.* **117**, 321–337 (2011)
6. Sanz-Izquierdo, B., Batchelor, J.C., Sobhy, M.I.: UWB Wearable Button Antenna. Department of Kent, The University of Kent, Canterbury, Kent, 2006
7. Hsu, S.-H., Chang, K.: Ultra-thin CPW-fed rectangular slot antenna for UWB applications. In: Proceeding, IEEE International Symposium Antennas and Propagation Society, 2006
8. Shobanasree, R., Radha, S.: Ultra wideband wearable textile antenna. In: Proceedings, International Conference on Computing and Control Engineering, 2012

Chapter 37

Bowtie Shaped Substrate Integrated Waveguide Bandpass Filter

Z. Baharudin, M.Z.U. Rehman, M.A. Zakariya, M.H.M. Khir,
M.T. Khan and J.J. Adz

Abstract A planar bandpass filter based on a technique that utilizes substrate integrated waveguide (SIW) is presented. The SIW based bandpass filter is implemented using a bowtie shaped resonator structure. The bowtie shaped filter exhibits similar performance as found in rectangular and circular shaped SIW based bandpass filters. This concept reduces the circuit foot print of SIW, along with miniaturization high quality factor is maintained by the structure. The design for single-pole resonator structure is presented; and by coupling the resonators a two-pole bowtie shaped SIW bandpass filter is achieved for the frequency of 7 GHz. The simulation result reveals that the filter's insertion loss is better than 1 dB and return loss is less than 30 dB.

Z. Baharudin (✉) · M.Z.U. Rehman · M.A. Zakariya · M.H.M. Khir · M.T. Khan · J.J. Adz
Communication Cluster, Electrical and Electronics Engineering Department,
Universiti Teknologi PETRONAS, Tronoh, Perak, Malaysia
e-mail: zuhairb@petronas.com.my

M.Z.U. Rehman
e-mail: zaka_g01951@utp.edu.my

M.A. Zakariya
e-mail: mazman_zakariya@petronas.com.my

M.H.M. Khir
e-mail: harisk@petronas.com.my

M.T. Khan
e-mail: talhakhhan2002@hotmail.com

J.J. Adz
e-mail: adzjamros@petronas.com.my

37.1 Introduction

Filters have received a particular attention with the advent of various wireless systems, this interest has dramatically increased with the introduction and development of new millimeter waves applications over the past decade. Various applications have been recently proposed including wireless local area networks [1], radars [2], intelligent transportation systems [3] and imaging sensors [4]. Efficient filters demand has also increased with the development of chip-sets operating at 60 GHz or even higher frequencies by a number of semiconductor industries [5].

Filters based on Substrate Integrated Waveguide (SIW) structures are achieved through incorporating the rectangular waveguide structure into the microstrip substrate [6]. SIWs are dielectric filled and are formed from the substrate material utilizing two rows of conducting vias connecting bottom and top metal plates, these vias are embedded in dielectric filled substrate; hence providing easy combination with other planar circuits and a reduction in size. The size reduction along with involving dielectric filled substrate instead of air-filled reduces the quality factor (Q), but the entire circuitry including waveguide and microstrip transitions can be realized by using printed circuit board (PCB) technology or other techniques, like LCP [8] and LTCC [9].

The design of an SIW bandpass filter can either utilize a design methodology based on coupling matrix method, or it can also follow a methodology used for designing air filled waveguide filters. The design of an SIW filter based on the methodology adopted in a rectangular waveguide, a shunt inductive coupling realization is adopted. Vias of irregular diameters placed in the center of the cavity may possibly occur in an inductive post filter; which is based on a requirement of control couplings. Large couplings might occur in the use of a small diameter. The utilization of shunt inductive vias at the couplings of the filter realizes a shunt inductive coupling filter as depicted in r (a) or an iris (aperture) coupling post as shown in Fig. 37.1b. A detailed literature on the development of SIW filters has been reported in [11].

A three pole structure of a SIW bandpass filter based on shunt inductive vias is shown in Fig. 37.1a. It utilizes four coupling vias placed in the center of the cavity

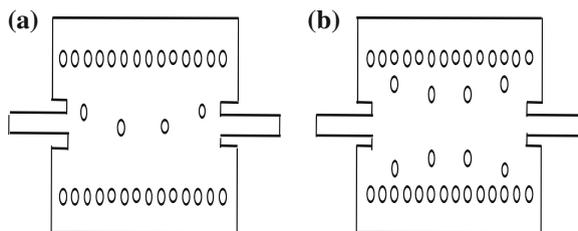


Fig. 37.1 a Shunt inductive coupling post filter b Iris coupling based filter

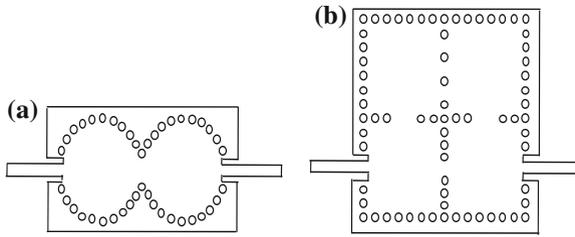


Fig. 37.2 Cavity filters with **a** circular cavities **b** rectangular cavities

of the filter, the small two vias provide facilitation in input and output coupling while the large vias provide coupling between the resonators. The transitions from SIW to microstrip are placed at both the input and the output of the SIW bandpass filters.

An SIW bandpass filter based on iris coupling posts is shown Fig. 37.1b; the apertures form three resonators. The filter's structure is such that the three cavities of half wavelength are formed in the center while SIW to microstrip transition are on the two edges of the filter; such a filter operational at 60 GHz has been presented in [12].

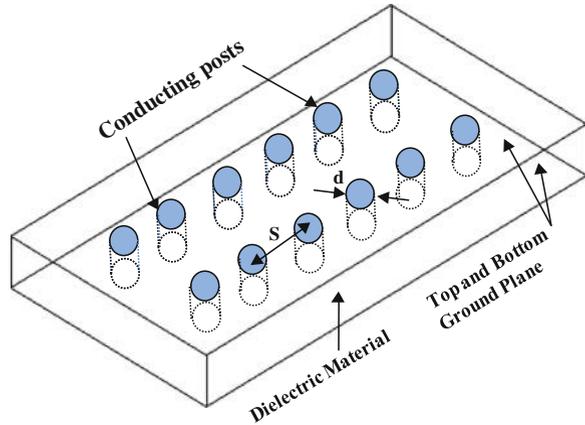
Cavity filters with circular [13] as presented in Fig. 37.2a and rectangular cavities [14] as depicted in Fig. 37.2b has been observed in literature. These variants of SIW allow more design variations and transmission zeros are also introduced due to cross coupling, better selectivity is also presented by these designs.

Various SIW filters structures have been proposed in the literature; however there still exists a need to further miniaturize the structure. Furthermore the cavities are only either in circular or rectangular shape. In this paper a triangular resonators based bandpass filter is presented. This resonator structure is a miniaturized form of SIW cavity, and exhibits similar low-loss and high quality characteristics as found in its other counterparts. The proposed bandpass filter configuration is suitable for integration with planar devices and its small footprint area allows other devices to be easily integrated on a single board.

37.2 Substrate Integrated Waveguide

Design and Implementation of SIW filters are being performed through defined Practical methods so far. The most common technique is to form the SIW cavity through metallic sidewalls [10] as shown in Fig. 37.3. A dielectric substrate having width of h forms the resonator and the resonator is of length L . The bottom and top of the resonator are constructed through placing metallic plates and conducting posts/vias going through the substrate connects the top and bottom plates; hence forming the sidewalls of the cavity. The vias are of diameter d and the separation

Fig. 37.3 Substrate integrated waveguide structure geometry



between two neighboring vias is given as s . The choice of diameter and separation between the two vias forms the basis of the SIW filters, therefore these should be selected in a manner that minimum radiation loss is exhibited.

The Dealandes and Wu [7] study reveals two primary design rules for SIW structures as given in Eq. (37.1); these rules are followed in order to ensure same design and modeling methodology adopted for rectangular waveguides. These rules pertain to the diameter d of the via posts and the via post spacing s :

$$d < \frac{\lambda_g}{5} \quad (37.1)$$

$$s \leq 4d \quad (37.2)$$

In our design d and s are chosen to be 0.8 and 2 mm respectively, these values ensure less radiation losses and the SIW cavity acts closely to a rectangular waveguide. For the TE_{101} mode, the dimensions of the SIW resonator structure are calculated by using the relation in Eq. (37.3) [7].

$$f_{TE_{101}} = \frac{c}{2\sqrt{\mu_r \epsilon_r}} \sqrt{\left(\frac{1}{W_{eff}}\right)^2 + \left(\frac{1}{L_{eff}}\right)^2} \quad (37.3)$$

W_{eff} and L_{eff} denote the effective width and length of the SIW resonator, respectively, and are given as:

$$W_{eff} = W - \frac{d^2}{0.95s}, \quad L_{eff} = L - \frac{d^2}{0.95s} \quad (37.4)$$

where W and L are the real width and length of the SIW resonator, c is the velocity of light in free space. In this design the width and length of the triangular resonator

Fig. 37.4 Initial triangular SIW structure

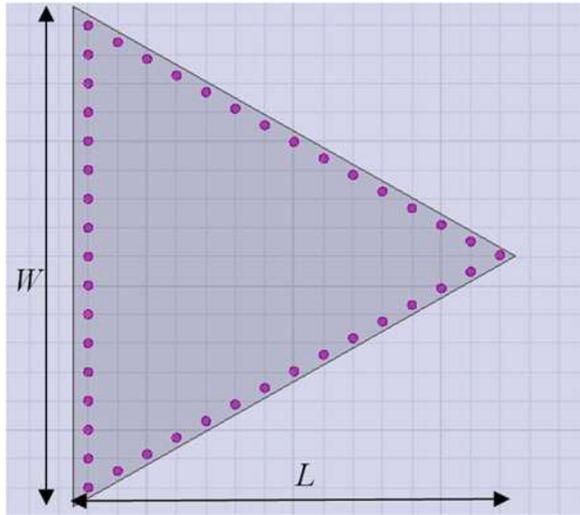


Table 37.1 Design specifications of the bandpass filter

Notation	Values
Passband centre frequency	7 GHz
Passband return loss, S_{11}	< -25 dB
Passband insertions loss, S_{21}	> -2 db
Passband bandwidth at -3 dB	> 300 MHz
Stopband rejection	> 30 dB

structure is computed using Eqs. (37.3) and (37.4) as shown in Fig. 37.4. Utilizing this method the cavity is designed for the specifications laid out in Table 37.1.

The triangular shaped resonator and subsequent bowtie shaped two pole filter is designed using Roger RT/Duriod 5880 material substrate having dielectric constant of 2.2 and substrate height of 787 μm . Theoretically, the resonance frequency does not depend on the thickness of the substrate. However, it has been observed in literature that it does play a role on the loss (mainly on radiation loss). The thicker the substrate the lower is the loss or higher Q. It has been shown that slight increase or decrease in the substrate thickness changes the unloaded quality factor.

To accomplish two pole bandpass SIW filter design, once the triangular resonator is created for a specific resonant mode, the design methodology closely resembles conventional simulation-based microstrip filter design [7]. Two single cavity resonators are coupled together through capacitive coupling along the RF input and output microstrip-to-SIW transition areas and a capacitive coupling located in-between two SIW cavities as can be seen in Fig. 37.5 (Table 37.2).

Fig. 37.5 Bowtie shaped two pole bandpass filter dimensions

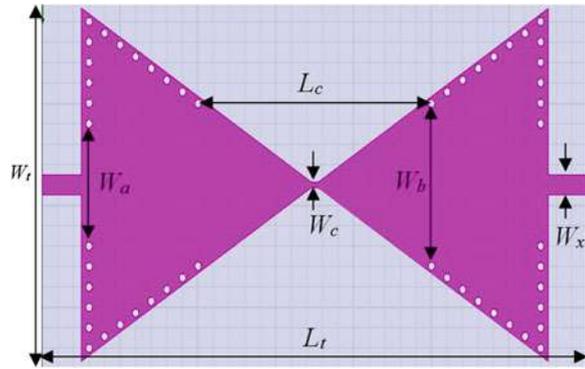


Table 37.2 Bandpass filter parameters dimensions and descriptions

Notation	Value (mm)	Description
Wt	36.0	Bowtie filter full width
Lt	70.0	Bowtie filter full length
L	30.0	Resonator's length
W	35.0	Resonator's width
Wa	12.0	I/O coupling
Wb	15.9	Coupling opening width
Wc	0.5	Inter-resonator coupling
Wx	2.0	Transition width
Lc	30.0	Coupling opening length

External quality factor and coupling coefficients are calculated from derived expressions based on lowpass prototype parameters [10].

$$M_{1,2} = \frac{FBW}{\sqrt{g_1 g_2}}, \quad Q_{e1} = \frac{g_0 g_1}{FBW}, \quad Q_{e2} = \frac{g_2 g_3}{FBW} \quad (37.5)$$

These values are then compared to the simulated extracted external quality factors and coupling coefficients for a particular cavity geometry.

$$Q_{ext} = \frac{f_0}{\Delta f_{-3\text{dB}}} \quad (37.6)$$

Iterations and adjustments to the dimensions of the coupling areas of the filter are performed until the calculated values match the extracted values from full-wave simulation, providing the desired filter.

37.3 Results and Discussions

The desired fixed filter described in the specifications in Table 37.1 and its corresponding designs structure shown in Fig. 37.5 are realized with the responses shown in Figs. 37.6 and 37.7. The simulation to obtain the filter responses from the designed structure shown in Fig. 37.5 is conducted using ANSYS High Frequency Structure Simulator (HFSS). In addition Agilent Vector Network Analyzer (VNA) is utilized for the measurements of the fabricated filters.

Figure 37.6 shows the response of the simulated two pole filter designed for 7 GHz. Response of the filter is obtained through realizing the resonator structure shown in Fig. 37.5, the resonator designed based on the requirements using the equations presented in Sect. 37.2. The simulated S_{21} and S_{11} response of the bowtie shaped bandpass filter reveals that the S_{11} value at the center frequency of 6.9 GHz is less than -30 dB, whereas the S_{21} response is greater than -0.5 dB and the passband bandwidth at -3 dB is greater than 500 MHz.

Fig. 37.6 Simulated response of the bowtie shaped two pole bandpass filter

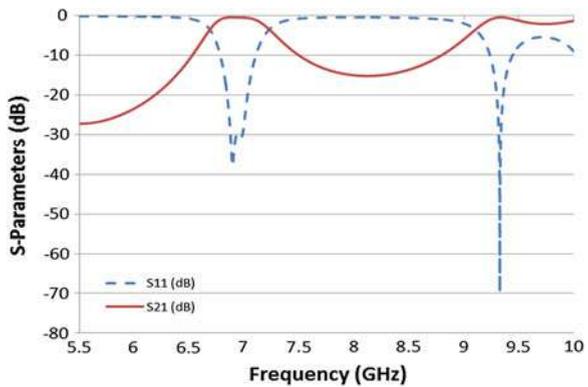


Fig. 37.7 Measured response of the bowtie shaped two pole bandpass filter

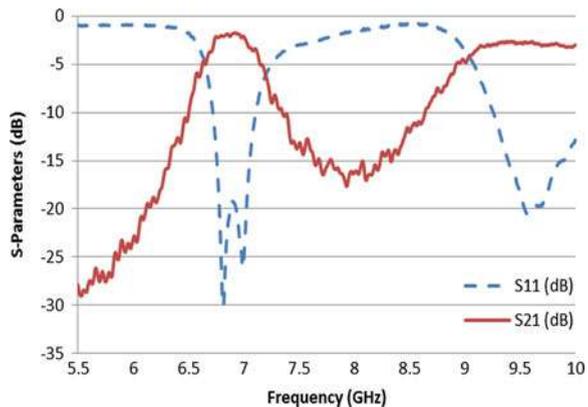
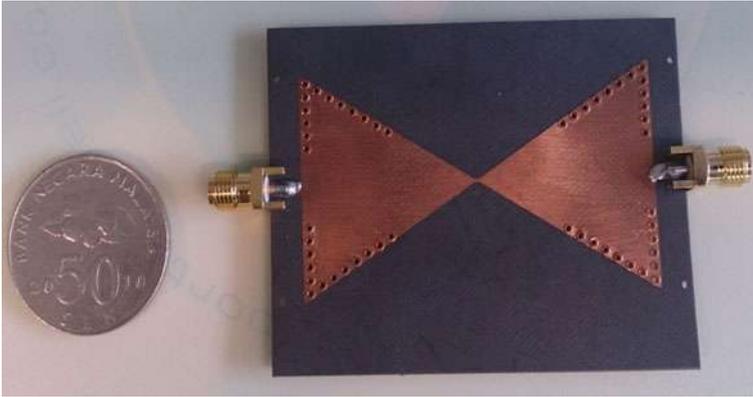


Table 37.3 Summary of the bandpass filter performance

Key parameters	Simulated	Measured
Passband centre frequency (GHz)	6.9	6.93
Passband return loss, S_{11} (dB)	<30	<25
Passband insertions loss, S_{21} (dB)	>-0.50	>-1.7
Stopband rejection (dB)	>30	>30
Passband bandwidth at -3 dB level	0.5 GHz	0.35 GHz

**Fig. 37.8** Fabricated design of the bowtie shaped two pole filter

Measured S_{21} and S_{11} response of the two pole bandpass filter is presented in Fig. 37.7. The S_{21} response at the centre frequency of 5.9 GHz is better than -1.7 dB and its corresponding S_{11} response at the centre frequency is less than -25 dB. The passband bandwidth at -3 dB is greater than 350 MHz, whereas the lower and upper stopband rejections are better than -25 dB. These responses are summarized in Table 37.3.

The performance of measured response in terms of the insertion loss is due to the quality of the SMA connectors and wires used for recording readings. The upper transmission zero produced due to cancellation of the TE_{101} and TE_{201} modes is distanced enough from the filter passband, hence a pure chebyshev response is observed. However high order resonant modes presents spurious at 9.5 GHz. Consequently suppression of spurious is achievable by employing lowpass filters at the input and output of the resonators. Therefore the cut-off for the lowpass filters has to be matched with the upper transmission zero of the bandpass filter (Fig. 37.8).

37.4 Conclusion

A 7 GHz bowtie shaped bandpass filter based on triangular structure SIW is proposed in this paper, the filter has presented good performance and a miniaturized version of the SIW structure is exploited in the design process. The filter has also presented broad bandwidth at the -3 dB level. This filter presents good performance in its small size, and it can be readily integrated with planar circuits and devices.

References

1. James, J., Shen, P., Nkansah, A., Liang, X., Gomes, N.J.: Millimeter-wave wireless local area network over multimode fiber system demonstration. *IEEE Photonics Technol. Lett.* **22**, 601–603 (2010)
2. Wang, L., Glisic, S., Borngraeber, J., Winkler, W., Scheytt, J.C.: A single-ended fully integrated SiGe 77/79 GHz receiver for automotive radar. *IEEE J. Solid-State Circuits* **43**, 1897–1908 (2008)
3. Yan, X., Zhang, H., Wu, C.: Research and development of intelligent transportation systems. In: 11th International Symposium on Distributed Computing and Applications to Business Engineering and Science (DCABES), vol. 19–22, pp. 321–327, October 2012
4. Wilson, J.P., Schuetz, C.A., Martin, R., Dillon, T.E., Yao, P., Prather, D.W.: Polarization sensitive millimeter-wave imaging sensor based on optical up-conversion scaled to a distributed aperture. In: 37th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), vol. 1, pp. 23–28, September 2012
5. Niknejad, A.M., Hashemi, H.: *Millimetre-Wave Silicon Technology: 60 GHz and Beyond*. Springer, Berlin (2008)
6. Hirokawa, J., Ando, M.: Single-layer feed waveguide consisting of posts for plane TEM wave excitation in parallel plates. *IEEE Trans. Antennas Propag.* **46**(5), 625–630 (1998)
7. Deslandes, D., Wu, K.: Accurate modeling, wave mechanisms, and design considerations of a substrate integrated waveguide. *IEEE Trans. Microw. Theory Tech.* **54**(6), 2516–2526 (2006)
8. Yang, K.S., Pinel, S., Kwon, K., Laskar, J.: Low-loss integrated-waveguide passive circuits using liquid-crystal polymer system-on-package (SOP) technology for millimeter-wave applications. *IEEE Trans. Microw. Theory Tech.* **54**(12), 4572–4579 (2006)
9. Xu, J., Chen, Z.N., Qing, X., Hong, W.: 140-GHz planar broadband LTCC SIW slot antenna array. *IEEE Trans. Antennas Propag.* **60**(6), 3025–3028 (2012)
10. Tang, H.J., Hong, W., Hao, Z.C., Chen, J.X., Wu, K.: Optimal design of compact millimetre-wave SIW circular cavity filters. *Electron. Lett.* **41**(19), 1068–1069 (2006)
11. Ur Rehman, M.Z., Baharudin, Z., Zakariya, M., Khir, M., Khan, M., Weng, P.W.: Recent advances in miniaturization of substrate integrated waveguide bandpass filters and its applications in tunable filters. In: *Business Engineering and Industrial Applications Colloquium (BEIAC)*, 2013 IEEE, pp. 109–114, 2013
12. Hirokawa, J., Ando, M.: Efficiency of 76-GHz post-wall waveguide-fed parallel-plate slot arrays. *IEEE Trans. Antennas Propag.* **48**(11), 1742–1745 (2000)
13. Tang, H.J., Hong, W., Hao, Z.C., Chen, J.X., Wu, K.: Optimal design of compact millimetre-wave SIW circular cavity filters. *Electron. Lett.* **41**(19), 1068–1069 (2005)
14. Chen, X.-P., Wu, K.: Substrate integrated waveguide cross-coupled filter with negative coupling structure. *IEEE Trans. Microw. Theory Tech.* **56**(1), 142–149 (2008)

Chapter 38

Logical Topology Design with Low Power Consumption and Reconfiguration Overhead in IP-over-WDM Networks

Bingbing Li and Young-Chon Kim

Abstract Reducing the power consumption of networks has drawn a lot of attention in recent years. To improve the energy efficiency, selectively turning off network components during light traffic periods could be an effective method. However, turning off/on network devices adapting to traffic load may result in the reconfiguration of logical topology in optical wavelength routed networks. In this paper, we study the logical topology design problem under multi-period traffic load in IP-over-WDM networks. To solve the problem, a mixed integer linear programming (MILP) model is proposed with the objective to minimize the overall network power consumption and the reconfiguration overhead between consecutive time periods. The proposed model is evaluated and compared with conventional schemes in terms of power consumption and reconfiguration overhead through a case study.

Keywords Energy efficiency · IP-over-WDM · Logical topology reconfiguration · MILP

38.1 Introduction

With the exponential growth of end users and the emergence of large-capacity required services, the Internet traffic has increased by 50–100 times during the last decade. It is expected that this increment speed will be kept in the near future.

B. Li (✉) · Y.-C. Kim

Department of Computer Engineering, Chonbuk National University, Jeonju, Korea
e-mail: batmangshock@hotmail.com

Y.-C. Kim

e-mail: yckim@jbnu.ac.kr

Y.-C. Kim

Smart Grid Research Center, Jeonju 561-756, Korea

To meet the traffic demand, network should be deployed with more transmission and switching equipment with higher capacity, consequently causing more power consumption. The network infrastructures are estimated to account for 12 % of total Internet power consumption at present and this portion will increase to 20 % by 2020 [1]. Hence, improving the energy efficiency of the Internet becomes a challenging issue nowadays.

Due to the high speed, huge capacity, low signal attenuation and other advantages of optical fibers, they are now widely used as the transmitting infrastructure in communication networks. It has been proven that optical component and equipment are much more energy efficient, comparing to their electronic counterparts. In particular, with the development of wavelength division multiplexing (WDM) technology, transmitting IP packets directly over WDM channel (IP-over-WDM) is considered as a promising paradigm. IP-over-WDM network can be implemented in different ways: namely IP with no Bypass, Transparent IP with Bypass and Grooming, Opaque IP with Bypass and Grooming [2], etc. Among these schemes, Transparent IP with Bypass and Grooming is the most energy efficient solution since the wavelengths can bypass at some intermediate nodes and low demand traffic flows can be groomed onto high-speed wavelength channels and transmitted integrally. As a result, the electronic processing at some nodes is avoided and the utilization of wavelength channels is improved. In such wavelength routed network, traffic demands are serviced by connection-oriented optical circuits, which are called lightpaths. Lightpaths can be established based on given network physical topology and corresponding traffic matrix, constructing a logical topology. Thus, designing an energy efficient IP-over-WDM network can be translated into logical topology optimization problem [3, 4]. There has been several researches focusing on IP-over-WDM network design with minimum power consumption [2, 5, 6]. However, they consider power consumption as the only parameter to be optimized. No or little attention is given to the changes of logical topology.

Following the user behavior over different periods in 1 day, the Internet traffic pattern can be approximated to sinusoidal function [7]. Figure 38.1 shows the traffic load of 2 days (from 8:00 23rd to 8:00 25th June, 2013), monitored from Amsterdam Internet Exchange. The peak time occurs at around 21:00. Usually, the logical topology is designed with capability to provision the heaviest network load. To reduce the power consumption, we can switch off some network elements when traffic load is light, like in deep night or early morning [8]. Consequently, the logical topology needs to be reconfigured adapting to immediate network status. Reconfiguration may include tearing down existing lightpaths and setting up new ones. The change from one logical topology to another will result in disruption of network, introducing data loss or delay [9]. Even though network resource (in our case, the power consumption) is optimized to fit the changing traffic load, the quality of service (QoS) in network is deteriorated. Hence, the traditional reconfiguration optimization solutions try to minimize the average number of hops encountered by a packet, total number of lightpaths, or total number of physical links used [3, 4, 9]. Reference [10] proposed a model with the objective to

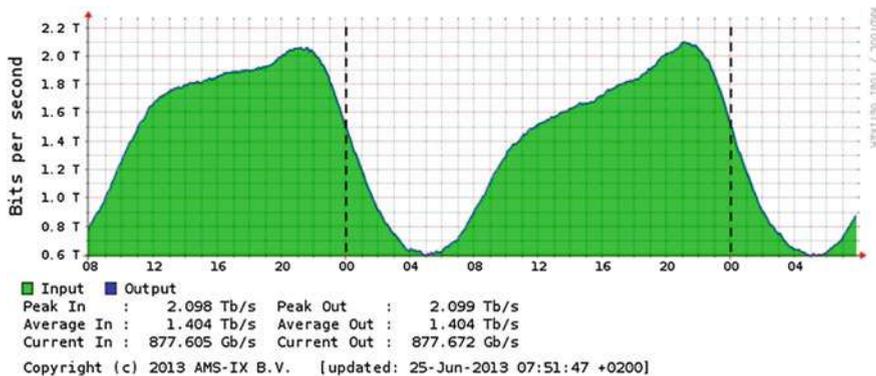


Fig. 38.1 Daily traffic load (from AMS-IX)

minimize power consumption while the reconfiguration is restricted via constraints. Nevertheless, none of these optimization models considers power consumption as well as reconfiguration at the same time.

In this paper, we consider a wavelength-routed optical backbone network based on the transparent IP-over-WDM network architecture. When network load is low, lightpaths need tearing down and corresponding elements can be turned off to save power. On the contrary, when network load increases, network elements are activated and new lightpaths should be established to provision increased traffic. In any case of network status change, influenced flows need rerouting and traffic interruption may occur. To solve the problem, we propose a MILP formulation with the objective to minimize the sum of total power consumption and the reconfiguration overhead multiplied with a weighting factor. This formulation considers minimizing both the power consumption in current status and the difference between two logical topologies of current and previous network status.

The rest of this paper is organized as follows: In Sect. 38.2, the mathematical formulation is presented and explained. In Sect. 38.3, the MILP models are evaluated and compared by illustrative examples; and the numerical results will be analyzed. Finally, we conclude the paper in Sect. 38.4.

38.2 Proposed Mathematical Model

38.2.1 Network Assumption

IP-over-WDM network consists of two layers: electrical layer and optical layer. IP routers are deployed at network nodes and constitute the IP layer (electrical layer). The functions of IP router is to generate (as a source node), process (as a grooming node) and drop (as a destination node) IP services. They are connected with an

optical cross-connect (OXC) via transponders which are used to emit and terminate lightpaths. Two adjacent OXCs are connected by an optical fiber and responsible for switching lightpaths. Each optical fiber can support many wavelength channels. All the OXCs and optical fibers construct the WDM layer (optical layer). IP packets are groomed at IP layer and then transmitted directly on optical WDM channels. Based on the transparent node architecture with bypass and grooming function, the power contributors are: (1) IP routers, for electronically processing traffic when grooming is needed; (2) Transponders, for establishing lightpaths; (3) OXCs, for optical switching wavelengths. The traffic processed at source node is not considered since it is fixed for a given traffic matrix. Note that the electronic processing is dependent with traffic amount while the power consumed by optically switching a lightpath is fixed, independent with the amount of traffic traveling on that lightpath. The power consumption of a transponder is also constant if it is activated, no matter the established lightpath is busy or idle.

On the other hand, we assume that the less reconfiguration happens to the logical topology, the less traffic needs rerouting, which leads to less data loss or delay, i.e. the less QoS degradation. The reconfiguration overhead is defined as the number of changes in physical links involving in forming a new logical topology comparing to the previous network status. Note that the change of wavelength assignment is also considered, which means that even if the route of a lightpath traverses the same physical links, different wavelength assignment decision is also viewed as a change and can affect those traversed links.

38.2.2 MILP Model

According to the network assumption mentioned above, we propose a mixed integer linear programming (MILP) formulation. The proposed MILP formulation deals with a static network design problem. The physical topology and corresponding estimated traffic matrix should be given in advance. To reflect traffic variation, 1 day can be divided into several periods. In each period, MILP model is run to find optimal solution with minimal power consumption for current period and least logical topology change from previous period. For variable indexing rule, s and d index the source and destination nodes of a traffic demand, m and n represent the nodes in physical topology, and i and j indicate the nodes in logical topology. The notations are summarized in Table 38.1.

According to the notations defined, the objective function can be written as:

$$\text{Minimize } PC + \alpha \times RO \quad (38.1)$$

Table 38.1 Summary of notations

Parameter	Meaning
$G(V, E)$	Network physical topology, consisting of node set V and edge set E
λ^{sd}	Amount of traffic demand from source s to destination node d , $s, d \in V$
$T = [\lambda^{sd}]$	Traffic matrix, set of traffic demands, $s, d \in V$
C	Capacity of each wavelength
W	Set of wavelengths on a fiber
C_{ep}	Routing capacity of an IP router
T_i	Maximum number of transmitters at node i
R_i	Maximum number of receivers at node i
P_{tr}	Power consumption of a transponder
P_{os}	Power consumption for optical switching one wavelength
P_{ep}	Power consumption for electronic processing per traffic unit (in Gbps)
PC	Power consumption of whole network
$N'_{ij,w}$	The number of lightpaths between node i and j , using wavelength w (in previous period)
$P'_{mn}{}^{ij,w}$	The number of lightpaths between node i and j , being routed through physical link (m, n) , using wavelength w (in previous period)
Variable	Meaning
f_{ij}^{sd}	Traffic amount of λ^{sd} that travels on lightpath l_{ij}
$N_{ij,w}$	The number of lightpaths between node i and j , using wavelength w
N_{ij}	The number of lightpaths between node i and j
$P_{mn}{}^{ij,w}$	The number of lightpaths between node i and j , being routed through physical link (m, n) , using wavelength w

where

$$\begin{aligned}
 PC = & P_{os} \times \sum_w \sum_i \sum_{j,j \neq i} \left(\sum_{(m,n)} P_{mn}{}^{ij,w} + N_{ij,w} \right) + 2 \times P_{tr} \times \sum_i \sum_{j,j \neq i} N_{ij} + P_{ep} \\
 & \times \sum_i \sum_j \sum_{s,i \neq s} \sum_d f_{ij}^{sd}
 \end{aligned} \tag{38.2}$$

$$RO = \sum_i \sum_j \sum_m \sum_n \sum_w |P_{mn}{}^{ij,w} - P'_{mn}{}^{ij,w}| \tag{38.3}$$

subject to the following constraints:

$$\sum_j f_{ij}^{sd} - \sum_j f_{ji}^{sd} = \begin{cases} \lambda^{sd}, & \text{if } i = s \\ -\lambda^{sd}, & \text{if } i = d \\ 0, & \text{otherwise} \end{cases} \quad \forall i, s, d \in V \tag{38.4}$$

$$\sum_n P_{mn}^{ij,w} - \sum_n P_{nm}^{ij,w} = \begin{cases} N_{ij,w}, & \text{if } m = i \\ -N_{ij,w}, & \text{if } m = j \\ 0, & \text{otherwise} \end{cases} \quad \forall i, j, m \in V, \forall w \in W \quad (38.5)$$

$$\sum_s \sum_d f_{ij}^{sd} \leq C \times N_{ij}, \quad \forall i, j \in V \quad (38.6)$$

$$\sum_j \sum_{s, i \neq s} \sum_d f_{ij}^{sd} \leq C_{ep}, \quad \forall i \in V \quad (38.7)$$

$$\sum_w \sum_j N_{ij,w} \leq T_i, \quad \forall i \in V \quad (38.8)$$

$$\sum_w \sum_j N_{ji,w} \leq R_i, \quad \forall i \in V \quad (38.9)$$

$$\sum_w N_{ij,w} = N_{ij} \quad \forall i, j \in V \quad (38.10)$$

$$\sum_i \sum_j P_{mn}^{ij,w} \leq 1 \quad \forall (m, n) \in E, \forall w \in W \quad (38.11)$$

In the MILP formulation, Eq. (38.1) gives the objective function, which minimizes the total power consumption (PC) as well as the reconfiguration overhead (RO). The power is consumed by OXCs and transponders in optical domain and routers in electrical domain of the IP-over-WDM networks. The RO is defined as the total number of physical links involved in lightpath establishment different from that in previous logical topology. A weight, α , is assigned to RO so as to make two factors mutually comparable. Equations (38.4) and (38.5) guarantee the flow balancing in network in the view of traffic flow and physical link, respectively. Constraint (38.6) limits that the total traffic amount transmitting on all lightpaths cannot beyond total capacity they offer. Constraint (38.7) guarantees that all traffic electronically processed at a node is restricted by the maximum capacity of IP router. The number of usable transmitters and receivers are limited by constraints (38.8) and (38.9), respectively. Equation (38.10) calculates the number of lightpaths between node i and node j . Constraint (38.11) ensures that each wavelength on a physical link can be used to establish at most one lightpath.

To compare the proposed model with conventional schemes, other two MILP models are presented. First one tries to minimize the total PC of network (shortly, *Min PC*), while the other one considers to minimize the RO from previous logical topology (*Min RO*). Both comparing models can share the same constraints with our model. Because our model considers minimizing both PC and RO, it is represented shortly as “*Hybrid*” in the following part.

38.3 Numerical Results

The numerical results will be shown and analyzed in this section. To evaluate the performance of *Hybrid* and compare it with *Min PC* and *Min RO*, we apply three schemes to a case study. Our results are obtained via optimization software IBM ILOG CPLEX Optimization Studio Version 12.5 on the computer with Intel Core 2 (TM) i5-2500 CPU (3.30 GHz) and 8 GB RAM.

38.3.1 Network Topology and Traffic

The case study is implemented in Pan-European COST239 network, shown as Fig. 38.2. Physical topology consists of 11 nodes and 26 links. Nodes are connected by bi-directional links, one fiber on each direction. One fiber can maximally support 16 wavelength channels, each with capacity 40 Gbps. Hence, the network can supply total capacity 33.28 Tbps. Based on the network topology, we assume the traffic matrix at peak time as shown in Table 38.2. The unit of traffic demand for each source-destination pair is Gbps and the total traffic amount is 1 Tbps.

To reflect realistic network load, we refer to the data from AMS-IX (24-h from 0:00 24th to 0:00 25th June, 2013) as shown between two dotted lines in Fig. 38.1. Considering the variation of traffic amount in 1 day, we divide 24 h into 12 time periods. During each 2-h period, the traffic amount is constant, while the amount is different from one period to another. We generate non-peak time traffic as a fraction of traffic at peak time. By multiplying each entry in peak time traffic matrix with a factor [0.72, 0.44, 0.31, 0.42, 0.61, 0.72, 0.77, 0.82, 0.87, 0.95, 1.0, 0.98], we can represent an approximated daily traffic pattern which is shown as Fig. 38.3.

Fig. 38.2 Pan-European COST239 network topology

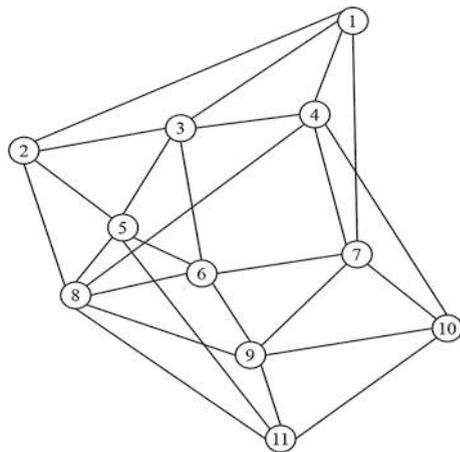


Table 38.2 Traffic matrix of peak load

Node	1	2	3	4	5	6	7	8	9	10	11
1	0	1	1	3	1	1	1	35	1	1	1
2	1	0	5	14	40	1	1	10	3	2	3
3	1	5	0	16	24	1	1	5	3	1	2
4	3	14	16	0	6	2	2	21	81	9	9
5	1	40	24	6	0	1	11	6	11	1	2
6	1	1	1	2	1	0	1	1	1	1	1
7	1	1	1	2	11	1	0	1	1	1	1
8	35	10	5	21	6	1	1	0	6	2	5
9	1	3	3	81	11	1	1	6	0	51	6
10	1	2	1	9	1	1	1	2	51	0	81
11	1	3	2	9	2	1	1	5	6	81	0

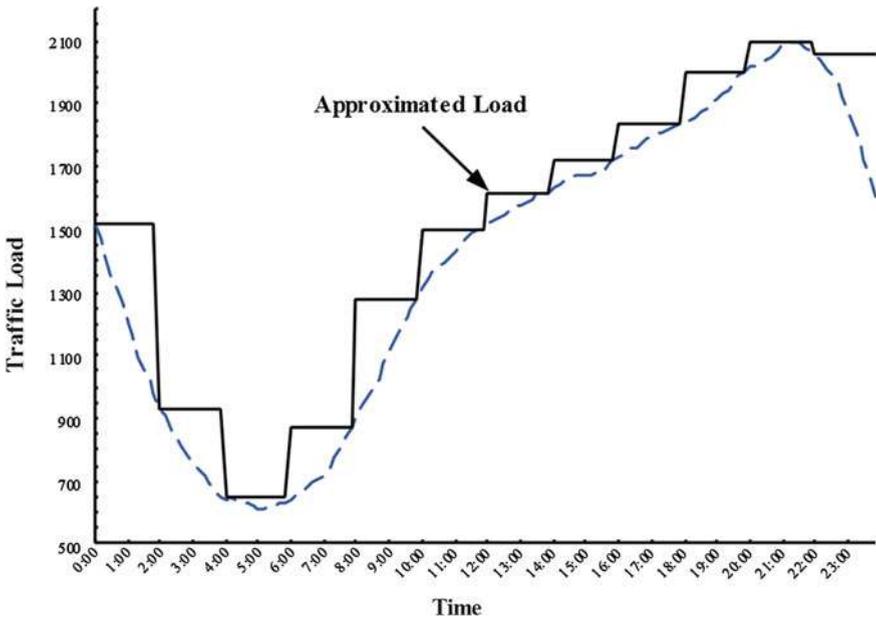


Fig. 38.3 Approximation of daily traffic amount

38.3.2 Power Consumption and Other Parameters

The PC considered is given in Table 38.3, referring to some literatures and data sheets of commercial products [6, 10, 11]. For IP router, Cisco CRS 16-Slot Carrier Routing System is considered. The total routing capacity per chassis is

Table 38.3 Power consumption

Device	Power Consumption
40 Gbps transponder	175 W
Optical switching	4 W per one 40 Gbps wavelength
Electronic processing	14.5 W (per 1 Gbps)

$C_{ep} = 4480$ Gbps. For OXC, the MEMS-based optical switch is considered. At each node, maximum number of transmitters/receivers is 16 ($T_i = 16, R_i = 16$). To investigate the influence resulting from the weight of RO, we implement Hybrid with different values of α ($\alpha = 1, 10, 100, 350, 1000$), shortly represented as “Hybrid_1”, “Hybrid_10”, “Hybrid_100”, “Hybrid_350” and “Hybrid_1000”.

38.3.3 Results

Figure 38.4 shows the reconfiguration overhead according to different time periods. Among all cases, *Min RO* needs smallest change of the logical topology when traffic varies from a period to another. In particular, the logical topology has no change during six periods. *Hybrid_1000* shows similar feature with *Min RO*. With reducing the importance of RO factor, more changes are needed. In the view of the cumulative changes in whole day, *Min RO* makes 43 changes of physical links involved in lightpath establishment; *Hybrid_1000* and *Hybrid_350* perform similarly to *Min RO* and need 45 and 50 changes, respectively; *Hybrid_100*,

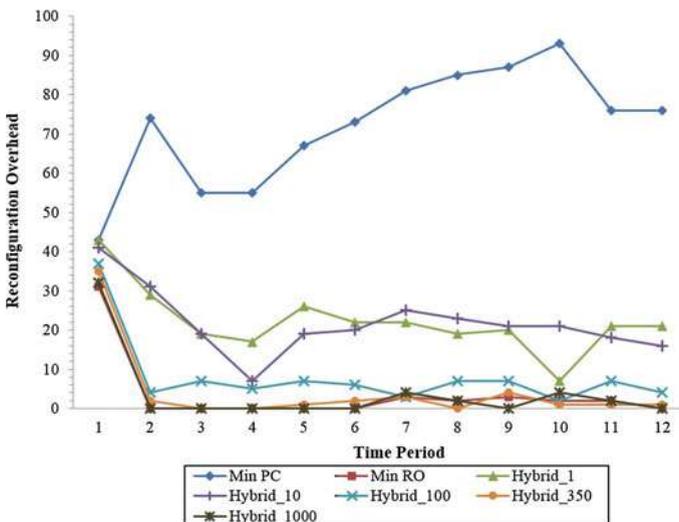


Fig. 38.4 Reconfiguration overhead versus time period

Hybrid_10 and *Hybrid_1* result in 76, 261 and 266 changes, respectively; while *Min PC* results in 865 changes. To limit the RO, the rate of change (ROC) can be defined as the ratio between the number of wavelengths involving in reconfiguration to the number of total wavelengths occupied for constructing the initial logical topology in the first period. The value of ROC can indicate the strictness of QoS. In our case study, it is ruled that the ROC cannot be larger than 0.2. Then *Hyrid_100*, *Hyrid_350*, *Hyrid_1000* and *Min RO* can satisfy this requirement among all tested cases.

As another important metric, the total PC of network in different time periods is given in Fig. 38.5. *Min PC* achieves the lowest PC, because the model essentially tries to minimize the PC factor. In the view of total PC in whole day, *Hybrid_1*, *Hybrid_10*, *Hybrid_100*, *Hybrid_350* and *Hybrid_1000* consume 0.8, 1.4, 4.4, 10.4 and 19.5 % more power than *Min PC*, respectively. *Min RO* achieves the highest PC, 1.43 times of the power consumed by *Min PC*. This is resulted from the extreme effort of *Min RO* on minimize RO. To guarantee logical topology unchanged, many lightpaths cannot be torn down and the corresponding transponders should be kept in “on” state when traffic load is light, for instance, from second to fifth period.

Obviously, even though *Min PC* or *Min RO* can reach the best performance in the view of single objective, they both perform the worst if the other metric is evaluated. *Min RO* model leads to large power waste: at first, many single-hop lightpaths are established; after network initialization, logical topology keeps nearly no change and transponders have to be turned on in light load periods. On the other hand, *Min PC* configures the logical topology for each period, independent with previous network status. The wavelengths used to establish lightpath

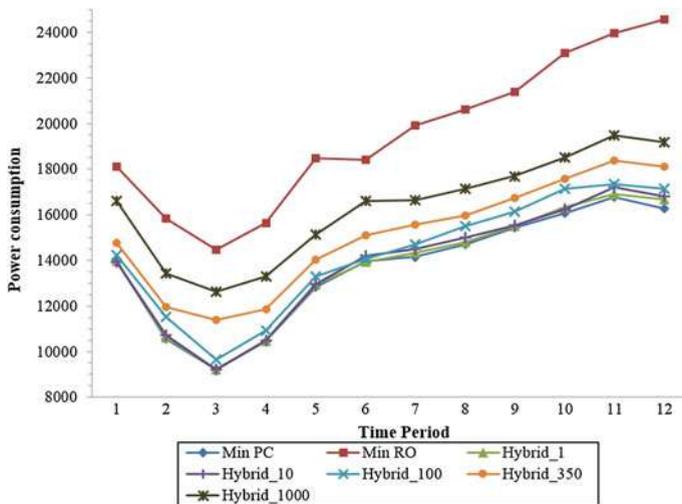


Fig. 38.5 Power consumption versus time period

need to be assigned again, making the overhead to configure a brand new logical topology considerably large. Different from the schemes mentioned above, the proposed model can obtain substantial reduction of PC and limit RO. However, the threshold of ROC could strongly influence the choice of α . By limiting $ROC \leq 0.2$, *Hybrid_100*, *Hybrid_350*, *Hybrid_1000* and *Min RO* can satisfy the requirement, among which *Hybrid_100* achieves the lowest PC. Hence, $\alpha = 100$ is reasonable suggestion for the case study.

38.4 Conclusion

Considering the traffic variation, turning off some network elements in low-load period can effectively reduce power consumption. Based on this load-adaptive scheme, logical topology needs to be reconfigured. However, reconfiguration may cause data delay and loss resulted from traffic interruption and rerouting. To solve such problem, we proposed a MILP model (*Hybrid*) for designing logical topology with low PC and RO in IP-over-WDM networks. Then the proposed model was evaluated and compared to conventional schemes with univocal objective (*Min PC* and *Min RO*) via illustrative case study. In addition, different values of α were investigated. With the definition of ROC, $\alpha = 100$ was suggested based on the tested network because it achieved the least power consumption without breaking the rule that ROC should be less than 0.2. The numerical results showed that our proposed model could achieve low reconfiguration overhead while keeping power consumption low.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIP) (2010-0028509).

References

1. The Climate Group, http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf
2. Musumeci, F., Vismara, F., Grkovic, V., Tornatore, M., Pattavina, A.: On the energy efficiency of optical transport with time driven switching. In: IEEE International Conference on Communications, pp. 1–5. Kyoto (2011)
3. Banerjee, D., Mukherjee, B.: Wavelength-routed optical networks: linear formulation, resource budgeting tradeoffs, and a reconfiguration study. *IEEE/ACM Trans. Netw.* **8**(5), 598–607 (2000)
4. Almeida, R.T.R., Calmon, L.C., Oliveira, E., Segatto, M.E.V.: Design of virtual topologies for large optical networks through an efficient MILP formulation. *Opt. Switch. Netw.* **3**(1), 2–10 (2006)
5. Shen, G., Tucker, R.S.: Energy-minimized design for IP over WDM networks. *J. Opt. Commun. Netw. IEEE/OSA* **1**(1), 176–186 (2009)

6. Idzikowski, F., Luca Chiaraviglio, L., Portoso, F.: Optimal design of green multi-layer core networks. In: Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet. ACM (2012)
7. Chiaraviglio, L., Mellia, M., Neri, F.: Energy-aware backbone networks: a case study. In: First International Workshop on Green Communications. Dresden, Germany, June 2009
8. Yayimli, A., Cavdar, C.: Energy-aware virtual topology reconfiguration under dynamic traffic. In: 14th International Conference on Transparent Optical Networks, pp. 1–4. Coventry, UK (2012)
9. Ramamurthy B., Ramakrishnan, A.: Virtual topology reconfiguration of wavelength-routed optical WDM networks. In: IEEE Global Telecommunications Conference, vol. 2, pp. 1269–1275. San Francisco (2000)
10. Zhang, Y., Tornatore, M., Chowdhury, P., Mukherjee, B.: Energy Optimization in IP-over-WDM Networks. *Opt. Switc. Netw. (OSN)* **8**(3), 171–180 (2011)
11. Idzikowski, F.: Power consumption of network elements in IP over WDM networks. TKN Technical Report, TKN-09-006 (2009)

Part II

Computer

Chapter 39

Systematic Analysis on Mobile Botnet Detection Techniques Using Genetic Algorithm

M.Z.A. Rahman and Madihah Mohd Saudi

Abstract Nowadays smart phone has been used all over the world and has become as one of the most targeted platforms of mobile botnet to steal confidential information especially related with online banking. It is seen as one of the most dangerous cyber threat. Therefore in this research paper, a systematic analysis on mobile botnet detection techniques is further investigated and evaluated. A case study was carried out to reverse engineering the mobile botnet codes. Based on the findings, this mobile botnet has successfully posed itself as a fake anti-virus and has the capability to steal important data such as username and password from the Android-based devices. Furthermore, this paper also discusses the challenges and the potential research for future work with relate of the genetic algorithm. This research paper can be used as a reference and guidance for further study on mobile botnet detection techniques.

Keywords Mobile botnet · Genetic algorithm · Reverse engineering · Android

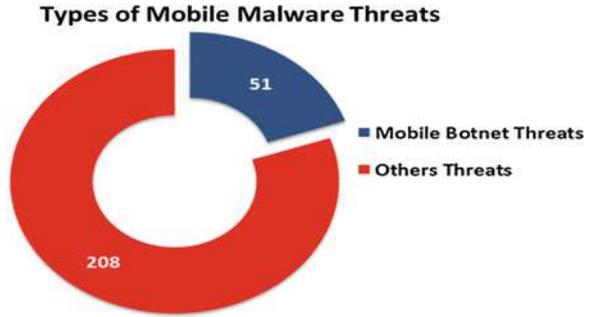
39.1 Introduction

Statistics taken from F-Secure [1] as displayed in Fig. 39.1 show that in every five malware threats, there is one mobile botnet threat. Botnet is seen as a new emerged cyber threat that attacks mobile devices, where all confidential information such as bank account number, username and password for online banking, credit card number, are kept in the smart phone. This kind of information makes mobile

M.Z.A. Rahman (✉) · M.M. Saudi
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia
e-mail: zuhairabdrahman@gmail.com

M.M. Saudi
e-mail: madihah@usim.edu.my

Fig. 39.1 Mobile bot threat 2013. [Adapted from F-Secure mobile threat report (Q3 2013)]



devices now targeted by attacker out there. There are several major mobile operating systems such as iOS, Android and Windows phone. Among them, Android is the most popular in term of attack as it is an open source operating system [2]. Earlier Google Play acts as an official market for Android platform and the main source for the users but nowadays there are several third-party market arise as an alternative such as Yandex market [3].

Usually, botnet infects victim's device by installing untrusted third-party application that acts as a trojan horse, entering the device and dropping botnet in it. By executing botnet such as Zeus, FakeDefender or Waledac, every data such as banking credential, email and SMS in victim's mobile device can be accessed within a second. Surprisingly, Google Play does not have firm control over its content where in June 2013, malwares reported have been hosted in the market. F-Secure reported there are applications that imitate the original application but repacked with malicious code and changed permission and has been downloaded between 100 and 22,000 times [4].

Therefore based on the botnet implication, this research is formed. The objectives of this research paper are to investigate and to evaluate mobile botnet detection techniques and to improve the gaps identified in the previous work. To achieve these objectives, a reverse engineering was conducted to the mobile botnet codes to see how the mobile botnet works. The findings from the evaluation and the reverse engineering are used as the basis to develop the mobile botnet architecture and detection algorithm. Later, these are used as the basis to produce an effective model to detect the mobile botnet attacks, but will not be discussed in this paper.

This paper is organised as follows. Section 39.2 presents the related works with mobile botnet. Section 39.3 explains the methodology used in this research paper which consists of static and dynamic analyses. Then follows by the architecture of the controlled laboratory architecture for the mobile botnet testing. Section 39.4 presents the research finding which consists of proof of concept of mobile botnet implication. Section 39.5 concludes and summarises the challenges and future work of this research paper.

39.2 Previous Works

A botnet is a set of computers that are infected by a specific malware or software which gives an attacker known as attacker the ability to remotely control the infected computers [5]. The infected computers are used by the attacker to launch the cyber-attack, such as sending spam messages, interruption, distributed denial of services (DDoS) and collecting sensitive information.

The term mobile botnet refers to a group of infected mobile devices and remotely controlled by the attacker via Command and Control (C&C) channel to do malicious tasks [6]. In 2004, the earliest case was reported as Cabir was detected and identified infecting Symbian OS [7] and this proves that mobile devices can be infected by malware. However it can only perform limited malicious tasks and spreads to other mobile devices using Bluetooth medium. As in 2009, SymbOS.Exy.C [8] and Ikee.B [9] show how mobile malware has evolved as it has the capability to connect back to C&C server and steal valuable information from the infected device. In 2011, DroidKungFu was discovered attacking Android smartphones and has the capability of rooting vulnerable Android phone and manage to evade the mobile antivirus detection [10]. It spreads via repackaging the mobile application, which has been downloaded from third party android market. It collects automatically OS version, IMEI number and phone model. Then it will contact the attacker by sending HTTP Post.

In the past few years, many studies have been done on mobile botnets such as by [11–13]. Traynor et al. [13] studied the possibility of using bluetooth as the command and control (C&C) channel of a botnet. As for Mulliner, he proposed SMS-HTTP command and control system in which the attacker created command and then the command was sent to bots via SMS [12]. The command is then being uploaded to a designated website in an encrypted file. Then, each bot will download, decrypt the file, and send out the commands to other bots via SMS. While Zeng and his colleagues designed a SMS-P2P hybrid botnet which uses SMS as the C&C channel, and the peer-to-peer (P2P) network as the underlying structure [11]. Botnet communicates by obtaining commands in a P2P fashion by sending and receiving SMS messages.

Yusoff and Jantan [14] implemented Genetic Algorithm (GA) to improve classification and the accuracy rate of PE file that failed to be classified by decision tree classifier. While work done by [15], implemented (GA) to the layered system to detect and filter http botnet attack and has succeeded provide less false positive rate. Monther and Rami also used (GA) to expand the training dataset through mutations. It managed to reduce training time and increased the detection from 77 to 80 % [16]. Most of the existing works discussed above were not focusing on mobile botnet and targeted for Windows platform only.

In this research paper, the mobile botnet sample was being reverse engineered using static and dynamic analyses, prior the formation of new mobile malware detection model. GA is chosen for classification as it simulates the natural processes. It uses selection, recombination and mutation acting on a genotype that is

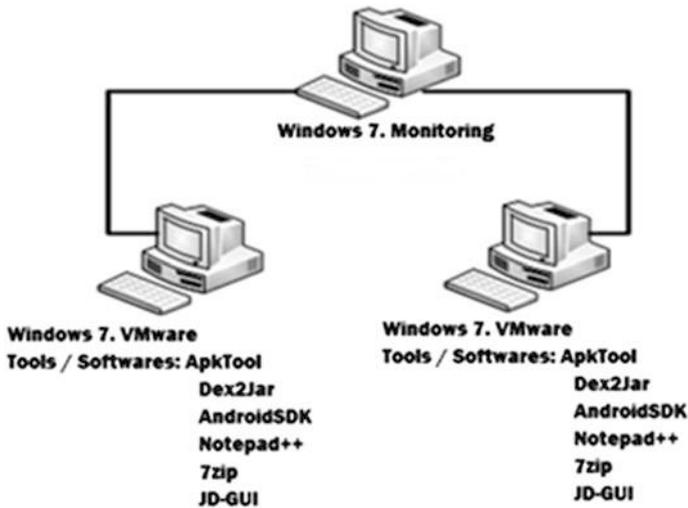


Fig. 39.2 Mobile malware controlled lab architecture

decoded and evaluated for fitness [17]. Furthermore, it is suitable for optimization of complex systems, robust and gives good result in real-time dynamic environment compared to other artificial intelligence such as Artificial Neural Network [18]. For this new model, it is expected to produce a better accuracy rate and lower false negative rates, but will not be discussed in this research paper.

39.3 Methodology

For reverse engineering, an isolated lab was set up, with no outgoing connection to internet as displayed in Fig. 39.2. It is a controlled lab environment and almost all software used in this experiment are an open source or available on free basis as displayed in Table 39.1. The training dataset in this research consists of different types of mobile malware and was downloaded from Android Malware Genome

Table 39.1 Software used in the testing

Software	Function
VMWare/VirtualBox	To build up virtual operating systems in a computer
ApkTool	To conduct the static analysis
Dex2Jar	To conduct the static analysis
JD-GUI	To conduct the static analysis
Notepad+++	To conduct the static analysis
AndroidSDK	To conduct the dynamic analysis
Wireshark	To monitor the network traffic generated from the infected computer
7zip	To unzip compressed file

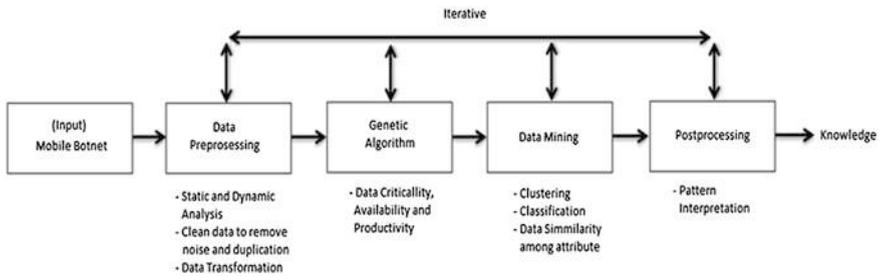


Fig. 39.3 KDD processes integrated with genetic algorithm

Project [19, 20]. As for the testing and evaluation, this experiment used dataset from Contagio [21]. The static and dynamic analyses were used in this isolated environment. As for static analysis, the files and its content associated with the mobile botnet were monitored without running the code or the program. While dynamic analysis includes executing the mobile botnet in a sandbox and observes its actions.

Knowledge Data Discovery (KDD) is an overall process of discovering useful knowledge from data and it has been applied in this research as well [22]. Figure 39.3 displays the KDD that has been improved by integrating an improved genetic algorithm to classify and to detect mobile botnet more efficiently with higher positive rate and lower false positive rate.

The basic Genetic Algorithm equation is defined based on the f function and the binary strings of length l and it is known as string fitness (refer to Eq. 39.1). A new string is created from the current population. The probability that a parent string H_j will be selected from N strings H_1, H_2, \dots, H_N as displayed in Eq. (39.1).

$$p(H_j) = \frac{f(H_j)}{\sum_{n=1}^N f(H_n)} \tag{39.1}$$

The string with greater fitness will be selected. The average fitness of all string in the population f_μ as displayed in Eq. (39.2). As for mobile botnet, the mobile botnet is transformed into chromosome containing the representation of the mobile botnet class, which is based on its target and operation behavior. In the mobile botnet analysis, a comparison based on different mobile malware features, which are the infection, behavior, payload, activation, operating algorithms and propagation were conducted. Mobile botnet then being categorized based on these features. Further, the data matching process was becoming easier when payload is used as the unique key identification. Once the data matching process completed, the eradication steps are identified for each botnet.

$$f_\mu = \frac{1}{N} \sum_{i=1}^N f(H_i) \tag{39.2}$$

Based on Eqs. (39.1) and (39.2), it has been integrated and mapped into mobile botnet classification as displayed in Eq. (39.3).

$$f(x) = \sum_{n=1}^{\infty} \left(\frac{\text{class}_{\text{score}}}{\text{max}_{\text{score}}} \right) \times W_1 \quad (39.3)$$

where:

$\text{class}_{\text{score}}$ Botnet attribute based on botnet classification (Payload, infection, activation, propagation, operation algorithm)

$\text{max}_{\text{score}}$ Total $\text{class}_{\text{score}}$ (In this case, payload + infection + activation + propagation + operation algorithm = 5)

W_1 Weight of group of selected individuals/class, to get W_1 , equation used: avg (botnet_class *4).

Then for selection method, Roulette wheel is used to create new population with better fitness value. The formula in Eq. (39.4) is used to find the probability of the selected chromosome. Genetic Algorithm (GA) is a heuristic search that simulates the process of natural evolution.

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j} \quad (39.4)$$

where:

f_i The fitness
 $\sum_{j=1}^N f_j$ The Sum of all fitness.

39.4 Findings

In this section a proof of concept on how the mobile botnet works is discussed. The Android.Fakedefender code was downloaded from Contagio [21] and it is a type of fake antivirus software [23]. Once installed, the application displays an icon as *Android Defender* and it will keep asking the victim to grant *Device Admin Privilege*. Once the *Device Admin Privilege* is granted, the mobile botnet will monitor and control the victim's smartphone and the reverse engineered codes as displayed in Fig. 39.4.

It started by displaying fake security status of a smartphone and tried to convince the victim to purchase a full version of the software, which is used to clean non-existing infection or fake infection. The message keeps on pop-up until the payment is made or until the infection is removed. Until payment is made, it will

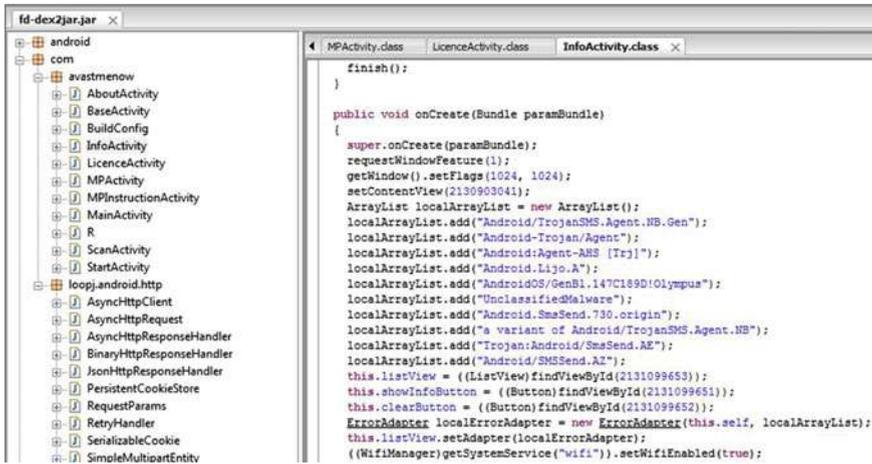


Fig. 39.4 Reverse engineered code

prohibit victim from doing anything on the phone. Next, it collected information such as SMS messages in the device's inbox, phone number, OS version, device manufacturer, location, and sends it to Command and Control (C&C) server. Based on the experiment conducted, it can be concluded that in order to produce an effective mobile botnet detection algorithm, all the mobile botnet behaviors should be further investigated and evaluated to identify the pattern used in existing mobile botnet for future prediction of the new mobile botnet.

39.5 Conclusions and Future Work

This paper presented a systematic analysis for mobile botnet detection techniques where gaps in the existing works have been identified for further improvement. Based on the findings for further improvement of the mobile botnet systematic analysis and experiment conducted, it can be concluded that mobile botnet has its own pattern and mechanism to spread and infect the victim smartphone. Therefore, further investigation needs to be carried out and this research paper has proposed an improved genetic algorithm to be integrated for a better solution of the mobile botnet classification and detection mechanism. This paper is part of larger project to design mobile botnet self-destruction model.

Acknowledgments The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Universiti Sains Islam Malaysia (USIM) grant [PPP/FST/SKTS/30/12712] and [PPP/GP/FST/SKTS/30/11912].

References

1. F-Secure: F-Secure Mobile Threat Report Q3. http://www.f-secure.com/static/doc/labs_global/Research/Mobile_Threat_Report_Q3_2013.pdf (2014). Accessed 24th Feb 2014
2. Teufl, P., Ferik, M., Fitzek, A., Hein, D., Kraxberger, S., Orthacker, C.: Malware detection by applying knowledge discovery processes to application metadata on the Android market (Google Play). *Secur. Commun. Netw.* doi:10.1002/sec.675 <http://dx.doi.org/10.1002/sec.675> (2013). Accessed 24th Feb 2014
3. Yandex: Yandex | Android apps market: download free and paid Android application. <http://market.yandex.ru/> (2014). Accessed 24th Feb 2014
4. Sullivan, S.: Bad bad piggies on google play. <http://www.f-secure.com/weblog/archives/00002566.html> (2013). Accessed 24th Feb 2014
5. Grizzard, J.B., Sharma, V., Nunnery, C., ByungHoon Kang, B., Dagon, D.: Peer-to-peer botnets: overview and case study. In: *Proceedings of First Workshop on Hot Topics in Understanding Botnets (2007)*
6. Xiang, C., Binxing, F., Lihua, Y.: Andbot: towards advanced mobile botnets. In: *4th Usenix Workshop on Large-scale Exploits and Emergent Threats (2011)*
7. Lee, W.K., Wang, C., Dagon, D.: *Botnet Detection: Countering the Largest Security Threat*. Springer, New York (2007)
8. Irfan, A.: Could sexy space be the birth of the sms botnet? <http://www.symantec.com/connect/blogs/could-sexyspace-be-birth-sms-botnet> (2013). Accessed 24th Feb 2014
9. Porras, P.A., Saidi, H., Yegneswaran, V.: An analysis of the iKee.B iPhone botnet. In: *Proceedings of the 2nd International ICST Conference on Security and Privacy on Mobile Information and Communications Systems (Mobisec) (2010)*
10. Security Alert: New sophisticated Android malware DroidKungFu found in alternative Chinese app market. <http://www.csc.ncsu.edu/faculty/jjiang/DroidKungFu.html>. Accessed 24th Feb 2014
11. Zeng, Y., Hu, X., Shin, K.G.: Design of SMS commanded-and-controlled and P2P-structured mobile botnets. University of Michigan Technical Report (2010)
12. Mulliner, C., Seifert, J.P.: Rise of the iBots: owning a telco network. In: *Proceedings of MALWARE 2010, France*, pp. 71–80 (2010)
13. Traynor, P., Lin, M., Ongtang, M.: On cellular botnets: measuring the impact of malicious devices on a cellular network core. In: *Proceedings of CCS 2009, Chicago, USA (2009)*
14. Yusoff, M.N., Jantan, A.: Optimizing decision tree in malware classification system by using genetic algorithm. *Int. J. New Comput. Arch. Their Appl. (IJNCAA)* **1**(3), 694–713 (2011)
15. Mathew, S.E., Ali, A., Stephen, J.: Genetic algorithm based layered detection and defense of HTTP botnet. *ACEEE Int. J. Netw. Secur.* **5**(1), 50–61 (2014)
16. Monther, A., Rami, A.: MALURLS: a lightweight malicious website classification based on URL features. *J. Emerg. Technol. Web Intell.* **4**(2) (2012)
17. Noreen, S., Murtaza, S., Zubair Shafiq, M., Farooq, M.: Evolvable malware. In: *GECCO '09 Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (2009)*
18. Lei, L., Wang, H., Wu, Q.: Improved genetic algorithms based path planning of mobile robot under dynamic unknown environment, mechatronics and automation. In: *Proceedings of the 2006 IEEE International Conference (2006)*
19. Zhou, Y., Jiang, X.: Dissecting Android malware: characterization and evolution. In: *Proceedings of the 33rd IEEE Symposium on Security and Privacy (Oakland 2012)*, San Francisco, CA, May 2012
20. Zhou, Y., Jiang, X.: Android malware genome project. <http://www.malgenomeproject.org/> (2012). Accessed 24th Feb 2014

21. Contagio Mobile: Mobile malware mini dump. <http://contagiominidump.blogspot.com/> (2013). Accessed 24th Feb 2014
22. Dunham, M.H.: Data mining: Introductory and Advanced Topics. Prentice Hall, New Jersey (2002)
23. Symantec Corporation: Android.Fakedefender. http://www.symantec.com/security_response/writeup.jsp?docid=2013-060301-4418-99 (2013). Accessed 24th Feb 2014

Chapter 40

An Empirical Study of the Evolution of PHP MVC Framework

Rashidah F. Olanrewaju, Thouhedul Islam and N. Ali

Abstract Commercial, social and educational importance of web technology has tremendously increased research activities in web programming/scripting. Several methods for writing PHP codes such as Object Oriented Programming (OOP), Procedural PHP coding and Model View Controller (MVC) pattern have been proposed. Model View Controller (MVC) which is one of the most powerful method for developing PHP application has many variant such Laravel, Symfony, CodeIgniter, CakePHP etc. However, selection of best MVC framework among the variants is of concern to the programmers as well as project managers, especially when managing big applications. Hence, performance evaluation criterions are required. This paper discusses the MVC based most famous PHP frameworks, evaluate their performance and it was found that Laravel outperforms other MVC framework, hence Laravel is proposed as the most suitable PHP framework for future web technology.

Keywords MVC · Laravel · PHP framework · CakePHP · CodeIgniter · Symfony

R.F. Olanrewaju (✉) · T. Islam
Department of Electrical and Computer, Kulliyah of Engineering, International Islamic University Malaysia, P.O. Box 10, 50728 Kuala Lumpur, Malaysia
e-mail: frashidah@iiu.edu.my

T. Islam
e-mail: tisuchi@gmail.com

N. Ali
College of Information Technology, Universiti Tenaga Nasional Malaysia, Selangor, Malaysia
e-mail: shikin@uniten.edu.my

40.1 Introduction

The rapid development of internet for web based application indicates a higher demand of reliability, scalability, security and maintainability of coding methodology. PHP, a scripting tool for web that enable dynamic interactive web development such intuitive, compiled fast, cross platform, open source, flexibility as well as required minimal setup [1]. This became one of the important web development language thus, PHP is one of the most powerful programming language in the web world. Several developers choose to deploy application based on PHP putting all the issues such as data access, business logic, and data representation layer together [2]. This in turn create development problems especially for big projects. To solve this problem, MVC design pattern brings an effective ways to separate code in layers from each other based on each layer activities.

MVC design pattern is a proven effective way to develop application such as CakePHP, CodeIgniter, Laravel, Symfony. The main methods of MVC to spilt an application into separate layer that can work separately and produce same result. The advantage of using MVC pattern are:

- Standard, consistency and predictability
- Software components or building-blocks so that developers can share and reuse code [3]
- A model or standard architecture that allows easy visualization of how the entire system works [4]
- Reusable and thoroughly tested code in the libraries, classes and functions [5].
- Well-structured code using architectural pattern [6].
- Security, interoperability and Maintenance.

Although MVC based framework (CakePHP, Laravel, CodeIgniter) has number of advantages [7], however, selecting of best PHP framework is still a concern. This is because all of the framework does not cover all aspect of web applications. This study evaluate most famous PHP frameworks based on MVC design model and it performance as well as proposed the best efficient PHP MVC framework for future web development.

40.2 Materials and Methods

The operation of Model View Controller (MVC) method is to spilt or separate the different parts of code into layers such as view, data access, controlling user's requests and forward request to relevant layers [8].

The MVC pattern's title is a collation of three core parts: Model, View, and Controller. A visual representation of a complete and correct MVC pattern looks like the following in Fig. 40.1.

Fig. 40.1 MVC mechanism

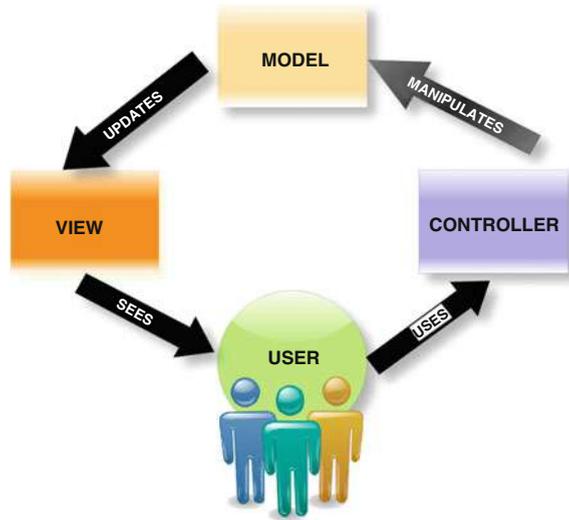


Figure 40.1 shows the pattern and interaction with the user and the application itself. It is a single flow layout of data, how it’s passed between each component, and finally how the relationship between each component works.

40.2.1 Model

The Model is the name given to the permanent storage of the data used in the overall design. It must allow access for the data to be viewed, or collected and written to, and is the bridge between the View component and the Controller component in the overall pattern [9].

One important aspect of the Model is that it’s technically “blind”—by this, the model has no connection or knowledge of what happens to the data when it is passed to the View or Controller components [10]. It neither calls nor seeks a response from the other parts of the component; its main purpose is to process data into its permanent storage, seek and prepare data to be passed along to the other parts.

The Model cannot simply be assumed as a database toolkit only, or a gateway to another system which handles the data process [11]. The Model represents a gatekeeper to the data itself, asking no questions but accept all requests which comes its way. Often this most complex part of the MVC system, the Model component is also the pinnacle of the whole system since without it there will be no connection between the Controller and the View.

40.2.2 View

The View is a module where data, requested from the Model is viewed and its final output is determined. Traditionally in web application use MVC for development, the View is the part of the system where the HTML is generated and displayed. The View also ignites reactions from the user, who then goes on to interact with the Controller. The basic example of this is a button generated by the View, which a user clicks and triggers an action in the Controller.

40.2.2.1 Misconceptions About View

There are some misconceptions held about View components, particularly by web developers using the MVC pattern to build their application. For example, many mistaken the View as having no connection whatsoever to the Model and that all of the data displayed by the View is passed from the Controller. In reality, this flow disregards the theory behind the MVC pattern completely. Fabio Cevasco's article, *The CakePHP Framework: Your First Bite* [1] demonstrates this confused approach to MVC in the CakePHP framework.

In order to correctly apply the MVC architecture, there must be no interaction between models and views: all the logic is handled by controllers [13]. Furthermore, the description of Views as a template file is inaccurate. The View is really much more than just a template, the modern MVC inspired frameworks have bastardised the view almost to the point that no one really cares whether or not a framework actually adheres to the correct MVC pattern or not. It's also important to mention that the View part is never given data by the Controller. There is no direct relationship between the View and the Controller without the Model in between them.

40.2.3 Controller

The third component of the triad is the Controller. Its job is to handle data the user submits as well as update the Model accordingly. The Controller can be summed up as a collector of information, which then passes it on to the Model to be organized for storage, and does not contain any logic other than collecting user input [9]. The Controller is also only connected to a single View and to a single Model, making it a one way data flow system, with handshakes and signoffs at each point of data exchange. Controller is only given tasks to perform when the user interacts with the View first, and that each Controller function is a trigger, set off by the user's interaction with the View [14]. The most common mistake made by developers is confusing the Controller as a gateway, and ultimately assigning it functions and responsibilities that the View should do (this is normally a result of the same developer confusing the View component as a template).

Additionally, it's a common mistake to assign the Controller functions that gives it the sole responsibility of crunching, passing, and processing data from the Model to the View. Nonetheless, the MVC pattern relationship should be kept between the Model and the View.

40.3 MVC Design Frameworks

40.3.1 CodeIgniter

CodeIgniter is an open source rapid development web application framework, for use in building dynamic websites with PHP. Its goal is to enable to develop projects much faster than writing code from scratch, by providing a rich set of libraries for commonly needed tasks, as well as a simple interface and logical structure to access these libraries. The first public version of CodeIgniter was released on February 28, 2006, and the latest stable version 2.1.4 was released July 8, 2013 [15].

CodeIgniter is loosely based on the popular Model-View-Controller development pattern. While view and controller classes are a necessary part of development under CodeIgniter, models are optional. CodeIgniter is most often noted for its speed when compared to other PHP frameworks.

40.3.2 CakePHP

CakePHP is an open source web application framework. It follows the Model-View-Controller (MVC) approach and is written in PHP, modeled after the concepts of Ruby on Rails, and distributed under the MIT License [16].

CakePHP uses well-known software engineering concepts and software design patterns, as Convention over configuration, Model-View-Controller, Active Record, Association Data Mapping, and Front Controller.

40.3.3 Symfony

Symfony is a PHP web application framework for MVC applications. Symfony is free software and released under the MIT license. The symfony-project.com website was launched on October 18, 2005 [17].

40.3.4 *Laravel*

Laravel is a free, open source PHP web application framework, designed for the development of MVC web applications. Laravel is released under the MIT license, with its source code hosted on GitHub.

The key design points of Laravel are:

- Bundles provide Laravel with a modular packaging system, and numerous bundled features are already available for easy addition to applications.
- Eloquent ORM is an advanced PHP implementation of the active record pattern, providing internal methods for enforcing constraints to the relationships between database objects [18].
- Application logic is part of developed applications, either by using controllers, or as part of route declarations. Syntax used for definitions is similar to the one used by Sinatra framework.
- Reverse routing defines a relationship between links and routes, making it possible for later changes to routes to be automatically propagated into relevant links. When links are created by using names of existing routes, appropriate URIs are automatically created by Laravel [19].
- Restful controllers provide an optional way for separating the logic behind serving HTTP GET and POST requests.
- Class auto loading provides automated loading of PHP classes, without the need for manual maintenance of inclusion paths. On-demand loading prevents loading of unnecessary components; loaded are only those components which are actually used [20].
- View composers are logical code units that can be executed when a view is loaded.
- Migrations provide a version control system for database schemas, making it possible to associate changes in the application's code base and required changes in the database layout, easing deployment and updating of applications [21].
- Unit testing plays an important role in Laravel, which itself has a large number of tests for detecting and preventing regressions. Unit tests can be run through the artisan command-line utility [22].
- Automatic pagination simplifies the task of implementing pagination, replacing the usual manual implementation approaches with automated methods integrated into Laravel [23].

40.4 Benchmarking

To evaluate the performance of four mentioned PHP frameworks; CodeIgniter (CI), Symfony, CakePHP and Laravel. The best way to do the benchmarking is by applying several criteria such as request per second, system load average,

memory usage, number of function calls and number of files required on each of the MVC. To evaluate these four frameworks, web design which contains “hello word” was carried out on apache (ab -c 200 -n 50,000), below are results of each evaluation criteria.

40.4.1 Request per Second

This benchmarking is based on apache (ab -c 200 -n 50,000). Figure 40.2 shows the performance comparison among four MVC: CI, CakePHP symphony and Laravel. It can be seen that Laravel outperforms other MVC in terms of request person. It was able to handle 3,000 request per second compare to others. In this case, bigger output indicates perfect result hence, it denotes best performance.

40.4.2 System Average Load

The system average load is measure in relation to time. that is, in 1 min when Apache Benchmark is complete, the smaller is better in the condition of (ab -c 200 -n 50,000). Figure40.3 shows the comparison of the four MVC in terms of system load within 1 min. In this graph, Laravel contain lowest times (0.98) where CakePHP contain maximum load time 5.1 per minute to load system. Based on average time, lowest average time is better to run MVC application.

40.4.3 Memory Usage

This benchmarking checks how much memory is used in a one word ‘hello world’ page display. The smaller number of KB the better the memory usage. Figure 40.4 indicated that Laravel is about 518 KB compare to CI which is which 725 Kb

Fig. 40.2 Request per second among PHP framework

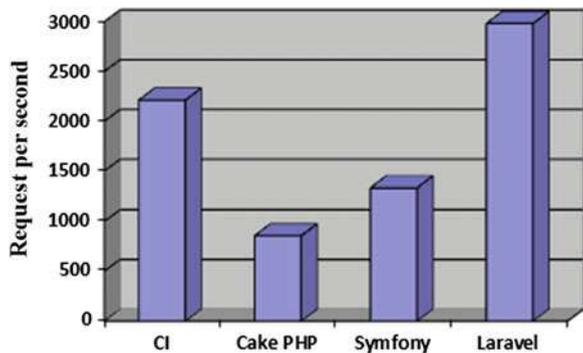


Fig. 40.3 System average load time

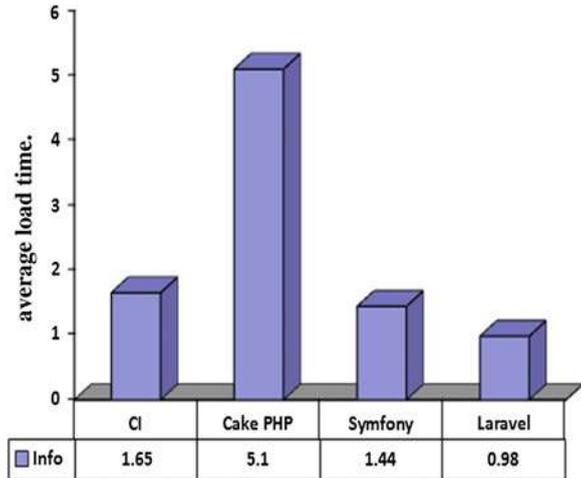
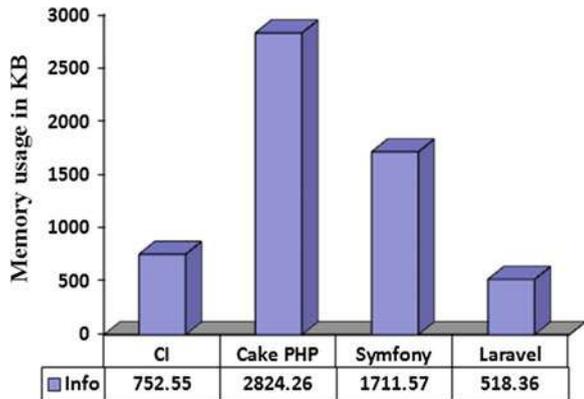


Fig. 40.4 Memory usage



follow by symphony with memory usage of about 1,711 then CakePHP with 2,824 KB. Hence, it can be concluded that Laravel memory usage is efficient.

40.4.4 Response Time

The time of page request to response from framework is one of the most important criteria to evaluate MVC performance. It is calculated in millisecond. The lower number of millisecond calculated the better performance. Figure 40.5 depicts the result of all the four MVC used. Among all, Laravel came out to be with the least response time, 4.46 ms compare to CI with 7.2 followed by symphony with 12 then CakePHP with about 14 ms.

Fig. 40.5 Response time for various MVC

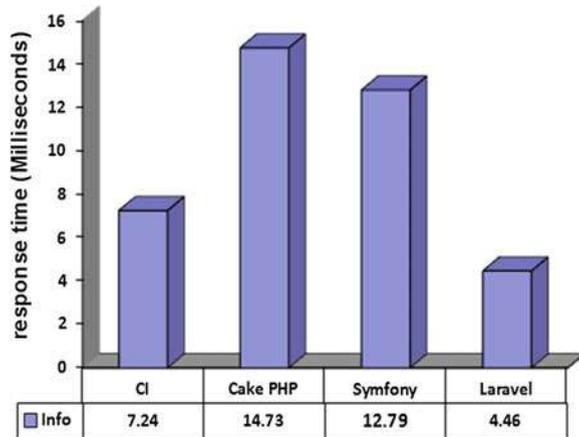
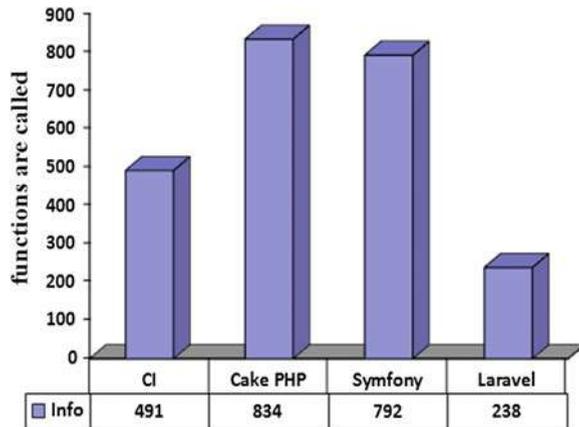


Fig. 40.6 Numbers of functions called



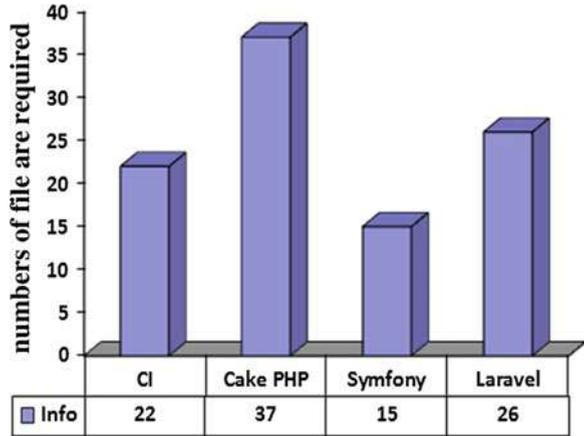
40.4.5 Number of Function Calls

This test checks how many functions are calling for one “hello world” pages in terms of Facebook XHProf [24]. Here smaller number of functions is most effective for PHP framework evaluation. It can be seen from Fig. 40.7, Laravel outperforms other MVC framework with the minimal number of function calls with 238 calls compare to CakePHP 834 and others (Fig. 40.6).

40.4.6 Number of Files

The number of files included or required in one ‘hello world’ page. Less amount of required files represent that such framework will be loaded first in-terms of file

Fig. 40.7 Numbers of file are required by various MVC



running. Smaller numbers of required files are highly appreciable. As shown in Fig. 40.7, CI comes up with the least file of 22, then symphony with 15 files followed by Laravel with 26, while CakaPHP loaded 37 files. Here CI outperforms other frameworks

40.5 Results and Discussion

Based on different core criteria of PHP framework such as benchmark, pattern, database access, field database, session, cache and library, this research compare among four top PHP framework (Laravel, CakePHP, CodeIgniter and Symfony) and their performance. Table 40.1 shows comparisons the frameworks in terms of facilities.

From the Table 40.1, it is assumable that, Laravel has advantage over other MVC PHP framework.

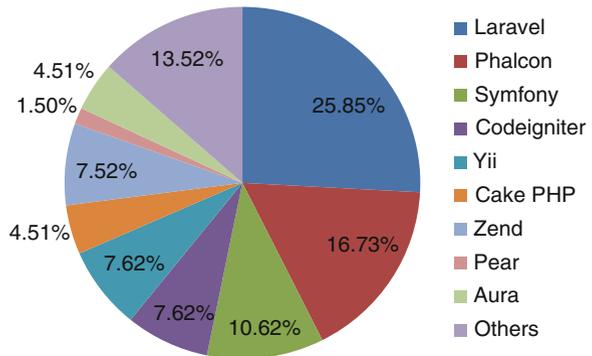
The four frameworks was also measured based on current trends for the future PHP framework that will lead the next generation of web. Figure 40.2 shows the details.

From Fig. 40.8, it is shown clearly that, Laravel took almost 26 % place in world web development in 2013 by MVC pattern framework. Based on this analysis, it is understandable that, Laravel is going to be one of the most famous MVC pattern framework in PHP world with huge flexibility of deployment as well as manage.

Table 40.1 Comparison among four MVC framework

Specification	CakePHP	CodeIgniter	Laravel	Symfony
MVC	✓	✓	✓	✓
DB	✓	✓	✓	✓
Fielddb	✗	✗	✓	UK
Auth	✓	✓	✓	✓
Validate	✓	✓	✓	✓
Session	✓	✓	✓	✓
Cache	✓	✗	✓	✓
Ajax	✓	✓	✓	UK
Db	✗	✗	Eloquent	✓
MVC type	PMVC	UMVC	HMVC	UMVC
MVC db	AR	AR	ORM	ORM
Upload	✓	✓	✓	✓
Form	Objects	Procedural	Objects	Objects
Xml	✓	✓	✓	✓
library	✗	✗	✗	UK

Fig. 40.8 Compare among all MVC-PHP framework based on current trends



40.6 Conclusion

An empirical study on major MVC pattern for PHP framework has been evaluated in this paper.

The results obtained from evaluating the four frameworks: CodeIgniter (CI), Symfony, CakePHP and Laravel using criteria such as request per second, system load average, memory usage, number of function calls and number of files required as well as available facilities in each framework. It shows that the Laravel outperforms other MVC framework. A request per second of as high as 3,000 was recorded for Laravel compare to others like CakePHP with as low as 750 request per second. The results obtained for all other parameters such as storage, function calls, number of files, response time etc. indicated that Laravel has huge flexibility

of development of web application, it has more facilities for programmers that makes it acceptable to all web programmer in terms of different criteria such as intuitive, compiled fast, cross platform, open source and flexibility. It enable easy migration, enriched library, template system, eloquent ORM and wide range of community support that helps to develop application smoothly. All the criteria and facilities of Laravel prove that, it is of opinion that Laravel would be the best choice to deploy next generation PHP based web application.

References

1. Bergmann, S., Kniesel, G.: GAP: generic aspects for PHP. In: Proceedings of EWAS'06 (2006)
2. Bakken, S.S., Aublich, A., Schmid, E., et al.: PHP manual (The PHP Documentation Group). <https://php.net/manual/en/index.php>, Accessed 10 March 2014
3. Nakajima, S., Hokamura, K., Ubayashi, N.: Aspect-oriented development of PHP-based web applications, 34th annual IEEE computer software and applications conference workshops (2010)
4. Veglis, A., Leclercq, M., Quema, V.: PHP and SQL made simple distributed systems online, Volume 6 Issue 8, August 2005, Page 4
5. <http://webcoderpro.com/blog/top-6-most-popular-php-frameworks-of-2013/>. Accessed 21 Jan 2014
6. <http://www.catswhocode.com/blog/top-10-php-frameworks-for-2014>. Accessed 23 Jan 2014
7. <http://www.sitepoint.com/best-php-frameworks-2014/>. Accessed 20 Jan 2014
8. Ricca, F., Tonella, P.: Analysis and testing of web applications. In: Proceedings of 23rd ICSE, pp. 25–34 (2001)
9. <http://www.tonymarston.net/php-mysql/model-view-controller.html>. Accessed 13 Dec 2013
10. Cui, W., Huang, L., Liang, L.J., Li, J.: The research of PHP development framework based on MVC pattern, 4th international conference on computer sciences and convergence information technology (2009)
11. <http://www.sitepoint.com/the-mvc-pattern-and-php-1/>. Accessed 15 Jan 2014
12. <http://www.sitepoint.com/application-development-cakephp/>. Accessed 17 Jan 2014
13. Enderlin, I., Giorgetti, A., Bouquet, F.: A constraint solver for PHP array, 6th international conference on software testing, verification and validation workshops (2013)
14. <http://matrix.include-once.org/framework/simplese>. Accessed 19 Jan 2014
15. <http://en.wikipedia.org/wiki/CodeIgniter>. Accessed 10 Mar 2014
16. http://en.wikipedia.org/wiki/MIT_License. Accessed 28 Feb 2014
17. <http://en.wikipedia.org/wiki/Symfony>. Accessed 11 Mar 2014
18. <http://www.developed.be/2013/07/16/php-frameworks-which-to-choose/>. Accessed 16 Jan 2014
19. <http://brianretterer.com/why-laravel-is-the-best-php-framework/>. Accessed 15 Jan 2014
20. <http://www.ruilog.com/blog/view/b6f0e42cf705.html>. Accessed 8 Mar 2014
21. <http://www.webdesignermag.co.uk/features/laravel-a-modern-php-framework/>. Accessed 17 Jan 2014
22. Merlo, E., Letarte, D., Antoniol, G.: Automated protection of PHP applications against SQL-injection attacks, 11th European conference on software maintenance and reengineering (2007)
23. [http://en.wikipedia.org/wiki/Laravel_\(framework\)](http://en.wikipedia.org/wiki/Laravel_(framework)). Accessed 9 Mar 2014
24. <http://www.php.net/manual/en/intro.xhprof.php>. Accessed 10 Mar 2014

Chapter 41

Evolutionary Approach of General System Theory Applied on Web Applications Analysis

Aneta Bartuskova, Ondrej Krejcar and Kamil Kuca

Abstract This paper reviews evolution stages of websites and presents framework for websites analysis, based on the evolutionary approach of general systems theory. Development of websites and web-based applications is discussed, according to their historical emergence, usage, increasing complexity and integration of new aspects and principles. Resulting individual stages of this development are suggested in accordance with the evolutionary approach. Framework for websites analysis and evaluation is then presented in a form of criteria list for each defined stage. Website's maturity (in the meaning of internet's evolution) can be then determined by confronting these criteria.

Keywords Web evolution · Websites evaluation · General systems theory

41.1 Introduction

The internet belongs to major providers of information and has a significant impact on our lifestyle. It functions also as a platform for communication, used by a wide range of users [1–4]. Considering an important role and rapid evolution of the internet, it is useful to organize our knowledge of this evolution. By identifying individual stages of complexity in the field of website development, we can evaluate any website or web-based application according to their evolution or maturity.

A. Bartuskova (✉) · O. Krejcar · K. Kuca

Faculty of Informatics and Management, Center for Basic and Applied Research, University of Hradec Kralove, Rokitanskeho 62, 50003 Hradec Kralove, Czech Republic
e-mail: Aneta.Bartuskova@uhk.cz

O. Krejcar
e-mail: ondrej.krejcar@remoteworld.net

K. Kuca
e-mail: kamil.kuca@uhk.cz

Framework for such evaluation, presented in this paper, is based on the evolutionary approach of general systems theory. Use of these principles for websites is justified by Yourdon's application of general systems theory to the information technology and systems [5].

General systems theory was invented by biologist Ludwig von Bertalanffy, who identified similar principles across many fields of knowledge, such as biology, social sciences, engineering or management [6]. Purpose of his theory was identification of laws pertaining these many branches and creation of suitable conditions for their collective development [7]. One of the key aspects of this approach is investigating systems as organizational units. Boulding defined two possible approaches to general systems theory, which are more complementary than competitive [8]. First approach relies on picking out general phenomena across various disciplines and create theoretical models. Key idea of the second approach is an arrangement of relevant constructs or empirical fields in a hierarchy, which defines organization of individual units within the system. The second approach will be used in this study and used for defining a hierarchy of evolution stages of websites and web-based applications.

41.2 The Evolutionary Approach of General Systems Theory

The second approach towards general systems theory was defined by Boulding as a systematic approach leading to system of systems. Rapoport specified it as the "evolutionary approach", since levels of abstraction are increasingly complex, marking the evolution of knowledge [9]. Each of these levels can be also defined by input, output, throughput or process, feedback, control, environment and goal or purpose, known as common elements of a system, which originated from Bertalanffy's types of finality [10, 11, 12]. The input means an energy or a material which is transformed by the system through some process, resulting in an output as a product of system's processing. The feedback is also a product of the process, which returns to the system as an input. An evaluation of the input, process and output is encapsulated in a control element, an environment denotes the area around a system and a goal is a purpose of the system. Individual levels are described in Table 41.1.

41.3 The Evolution of Websites: Review of Stages

We can distinguish individual stages of development in the field of websites and web-based applications. The stages were identified by authors in accordance with the evolutionary approach and appropriately to their complexity, usage and

Table 41.1 Arrangement of levels by Vossen [8], further specified by Bertalanffy [9]

Level	Name	Included entities
1	The structure level	Static structure, framework, arrangement
2	The simple dynamic level	Predetermined motions, simple machines
3	The control mechanism	Transmission and interpretation of information
4	The open system	Self-maintenance, self-reproduction
5	The genetic social level/ Level of the cell	Division of labour, differentiated parts
6	The animal level	Increased mobility, teleological behaviour, self-awareness, specialized information-receptors
7	The human level	Self consciousness, self-reflexivity, speech
8	Human society	Social organizations, units as a role in society
9	The transcendental level	Ultimates and absolutes and the unknowables

historical appearance. Numbering of stages follows complexity of web development and comply with an arrangement of levels in general systems theory by Vossen and Bertalanffy [8, 9]. Individual stages will be reviewed in this section as a basis for next research part of the paper. Descriptive figures are appended to illustrate processing of the website on particular levels of complexity.

41.3.1 Static Website

The internet had started its existence with basic functionality—requesting and displaying HTML files. Its proper functioning was ensured as a collaboration of three fundamental technologies, specified by Penhaker [13]. These are: HTML, URI and HTTP. A markup language, such as HTML or XML, forms a structure of a web document. Such website is considered static from a view of a user, as it does not change once loaded, until a user clicks on a next hypertext link [14, 15].

The first of evolutionary levels is defined by Boulding as a static structure, also a framework or an arrangement. Static website fulfils this concept and therefore it was identified as the first evolutionary level of websites.

41.3.2 Interactive Website

We consider an interactive website as a static website with capability of client-side interaction. It is usually powered by a combination of HTML, CSS and JavaScript. CSS is primarily used for styling HTML documents, but it can also convey interaction like a hover effect. The latest specification of CSS—CSS3—also brought a wide range of interactive features. It is essential that this interaction does not change the website as it does not involve communication with server.

Temporary change of appearance as a result of user interaction is only visible on this particular user's browser [16, 17].

The evolutionary approach describes the second stage as a simple dynamic level with predetermined motions. This level is represented by simple machines. In relation to web applications, the second level corresponds with an interactive website. Interaction here proceeds as a reaction of a website to a predetermined event on the client side, followed by predetermined effect [18].

41.3.3 Dynamic Website

Dynamic website is defined by server-side scripting, which is often accompanied by a database system. Server-side scripting requires server-side language, such as PHP, ASP, Java or Perl. As a database system is usually used relational database MySQL or MSSQL. Additional value of dynamic websites is possibility to use the same page structure and design for dynamically loaded content, and also to use different parts of structure and functionality depending on the request. This made possible content management, search engines and variety of applications with preserving and maintaining user data like webmail or online stores [19].

The third evolutionary level is defined as the control mechanism, with a purpose of transmission and interpretation of information. The dynamic website corresponds with this description, as it enables inserting, updating and a retrieval of information. It dynamically delivers information and it also provides mechanisms for data validation. Content management system is a great example of a control mechanism, which is a principal description of the third evolutionary level in the Boulding's hierarchy (Fig. 41.1).

41.3.4 Social Web Application

In this stage, internet has evolved from provider of information to a socialization platform, where every user can be a contributor to its content. Socialization of the internet is covered by term Web 2.0. It is technologically associated with AJAX, enabling rich functionality. Other feature are mash-ups, joining multiple data sources or services to create a new service. Core aspects of Web 2.0 are then data (mash-ups), functionality (AJAX) and socialization (community) [20, 21]. This approach leads to a functionality dependence and information redundancy, as the same data occur in many variations across the web, with little or none unified content organization.

The open system is a fourth stage according to Boulding, defined by an ability of self-maintenance and self-reproduction. A Web 2.0 application can be perceived as such an open system. Every user can be a contributor to this kind of application, so its content is growing and is maintained without central interventions. Social networks and wiki sites are great examples on self-maintenance. Activity of

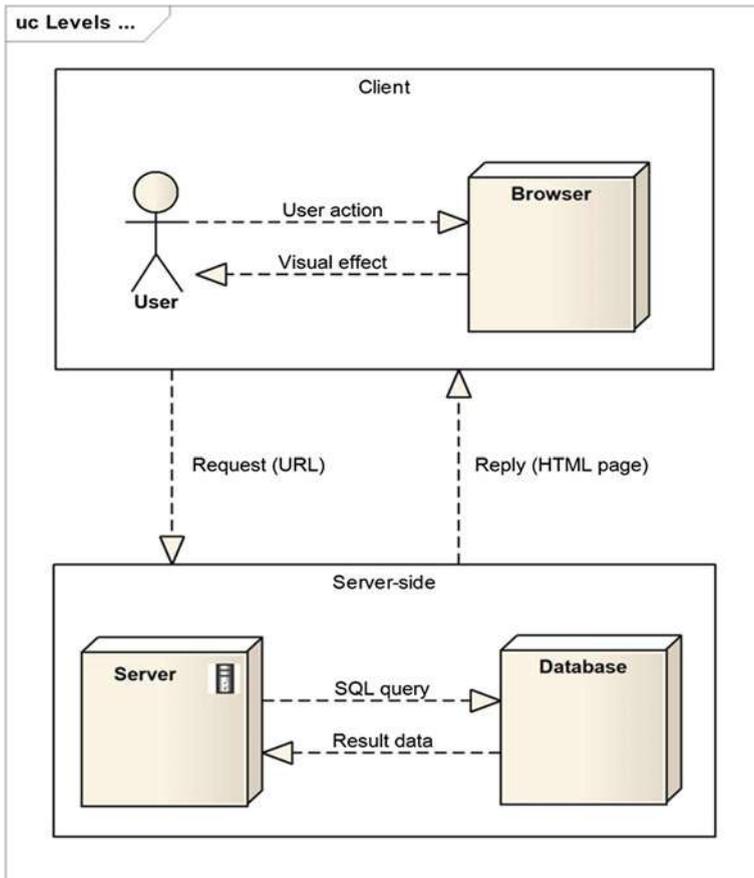


Fig. 41.1 Schema of first three levels of website development

sharing, a social aspect with none or loose terms, is typical for this stage. Self-reproduction can be viewed e.g. in the form of mash-ups. Boulding also defined this stage as a level at which life begins to differentiate itself from not-life [8]. In accordance with this, social networks create living systems, which are changing our social behaviour [20].

41.3.5 Semantic Web Application

Fifth level of websites can be represented by an aspiration for semantic web, which aims to implement a logical structure with help of taxonomies and ontologies. The term Web 3.0 is emerging as a possibility of combining today's web with semantic architectures [22]. Embedding web content in a logical structure enables not only

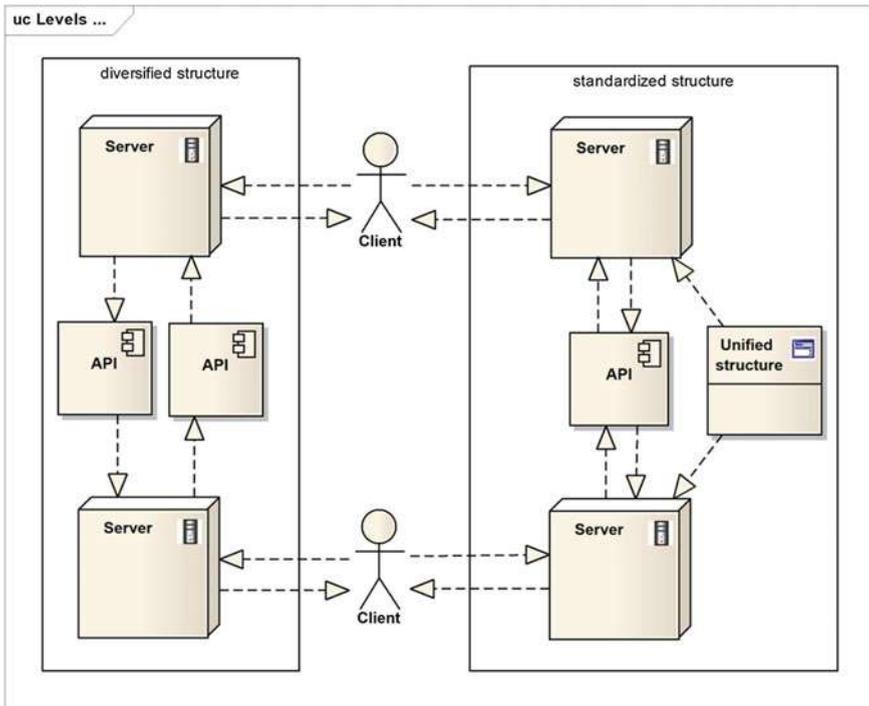


Fig. 41.2 Schema of the fourth level (on the *left*) and the fifth level (on the *right*)

machine-readable data, but also machine-understandable data [23]. Semantic web aspires for providing information models and languages that embed semantic contexts and metadata to enable automated processing of data [20].

The genetic social level is the fifth level of the evolutionary approach, defined by a division of labour and differentiated and mutually dependent parts of a system [8]. The semantic web application can represent such a model. By integrating and encapsulating data, they can be handled differently and these differentiated parts of a system enable a division of labour. Another aspect of the genetic social level and also semantic web is a collaboration, as of a community, which needs to be active in order to implement linguistic and structural concepts on the web. A collaboration here represents a more elaborate social aspect with established terms and structure (Fig. 41.2).

41.3.6 Adaptive Web Application

Adaptive web applications are suggested as the sixth level. Considering expansion of mobile devices with internet access, an adaptation of web is necessary. This can

be relevant to visual appearance, since desktops, notebooks, tablets, mobile phones etc. have different range of screen dimensions and control possibilities [24]. Adaptation is also providing different functionalities according to capabilities of chosen device or personalized content. New techniques are emerging to deal with these requirements, such as HTML5 API. Evolution of websites is connected to a development in ambient intelligence, ubiquitous computing and intelligent user interfaces [25]. Web applications are becoming context-aware systems with three basic functionalities—sensing, thinking and acting [26]. A research on sensors is also closely connected with this stage, e.g. sensing movement, light, location, proximity or biological signals [27].

The sixth level as the animal level is characterized by an increased mobility, teleological behaviour, self-awareness and specialized information-receptors. Adaptive web applications correspond with this stage as context-aware systems [26]. Mobility can be understood as an ability to function appropriately in different environments. Teleological behaviour ensures different functionalities according to capabilities of chosen device. Adaptive web application is aware of its capabilities, and of the relevant environment, which determines use of these capabilities (Fig. 41.3).

41.4 Summary of the Web Evolutional Stages

Evolutional stages of websites and web-based applications were reviewed in Sect. 41.3. A summary of conclusions, completed with remaining stages of Boulding's hierarchy (but yet expected in web evolution) is presented in Table 41.2.

To the authors best knowledge, the current state of the internet can be placed past levels 1–3, in level 4 and in the beginning of both levels 5 and 6. The Boulding's hierarchy has three more levels, 7 the human level, 8 human society and 9 the transcendental level. According to their features and a position in the hierarchy, we can roughly predict associated future stages of web applications. The seventh level as the autonomous web systems, which encapsulate wide range of functionality and are capable of complex decisions by their expert systems. The eighth level as the cooperative web systems, capable of communication among autonomous applications and delivering desirable performance without human intervention.

41.5 Framework Proposal for Websites Analysis

On the basis of reviewed evolutional stages, list of criteria can be defined, which are typical for particular stage. Some of the criteria are of course applicable also in earlier stages, but they have been deliberately allocated to the stages, where their

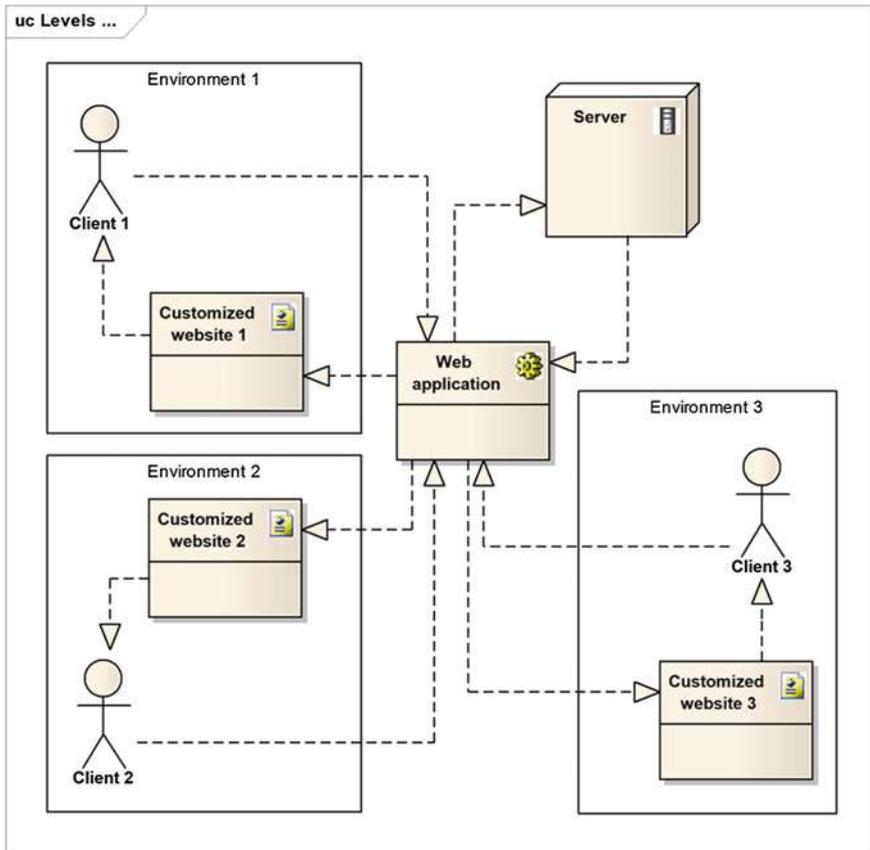


Fig. 41.3 Schema of the sixth level of website development

effect is most pronounced or it can be completely performed by means available in this particular stage. By confronting these criteria with factual state of evaluated website, we can determine its maturity in the sense of the internet evolution.

These criteria should be in theory hierarchically based, with one level as necessary requirement for the next level. This is not usually true with real web applications, as there is no enforcement on the internet regarding quality of development and poorly designed sites get as much space as standards-complying websites. Also to a certain extent, it is possible to ignore possibilities of particular stage and target more recent and complex issues. Finally, requirements may differ for specific web projects (Table 41.3).

At the first level of static website, three main components are responsible for result—structure, visual design and content. Structure is regarded here from pure technical aspect via HTML. Visual design considers layout, colors, fonts etc. ensured by CSS. Content has a quality perspective, especially its relevancy,

Table 41.2 Stages of web applications in relation to the Boulding’s hierarchy

Level	Type of website/ output	Essential feature/input
1	Static website	Markup language, creating a static structure
2	Interactive website	Scripting client-side language, styling language (hover), creating simple dynamics
3	Dynamic website	Scripting server-side language and a database, creating a control mechanism
4	Social web application	Creating and sharing content in a community, forming a living system
5	Semantic web application	Established logical structure, ensuring differentiation of the content types
6	Adaptive web application	Adaptation and context-awareness—ability to function correctly in any environment
7	Autonomous web systems	Encapsulating functionality and decision-making with internal expert systems
8	Cooperative web systems	Communication among autonomous web systems without human intervention

Table 41.3 List of criteria, which apply to individual evolutionary stages of websites

Level	Type of website (technologies/concepts)	Principal criteria
1	Static website (HTML, CSS)	Structure (technical aspect) Visual design Content
2	Interactive website (CSS, JavaScript)	User-website interaction Performance issues Strategy of fallback
3	Dynamic website (PHP/ASP/..., MySQL/MSSQL)	Content management Navigation and usability Search engine optimization
4	Social web application (social aspect—sharing)	Customization Communication Rights and security issues
5	Semantic web application (social aspect—collaboration)	Structure (logical semantic aspect) Machine-understandable
6	Adaptive web application (HTML5 API, JavaScript)	Responsiveness to device Context-awareness (sensors)

correctness and amount. These three criteria remain essential throughout the whole hierarchy at every level.

The second level adds CSS effects and JavaScript to create interaction between a user and website. Additional criteria to consider in this stage are connected to this interaction.

The third stage, dynamic website, uses server-side scripting language and database primarily to ensure content (or functionality) management. With this

approach comes increased need for viable navigation, usability and search engine optimization.

At the fourth level, the development is more of social nature than technological. Website complying with this evolutionary stage should facilitate customization i.e. support for personalization and user content, and some form of communication and feedback. Important back-end criteria are rights management and security issues, which were already important in the third stage, but reach increasing importance here, considering that users have much greater access to web application than before [28].

The fifth stage, semantic web application, is defined by more complex social aspect in the form of collaboration. This social aspect takes place on back-end part of the internet in order to create underlying structure and rules [29].

Adaptive web applications on the sixth stage are, in current state, enabled primarily by HTML5 API and JavaScript, along with infrastructure equipment as sensors. General criteria for this evolutionary stage are responsiveness to device parameters and capabilities and context-awareness (which is on client-side delivered by sensors).

41.6 Conclusions

This paper reviewed evolution stages of websites and web-based applications according to their historical emergence, usage, complexity and integration of new aspects. The authors believe that this classification clarified the evolvement of internet from simple static websites to complex web applications. Individual stages of this development were discussed with connection to the evolutionary approach of general systems theory. Framework for websites analysis and evaluation was then proposed in a form of criteria list for each defined stage. By confronting these criteria with factual state of the evaluated website, we can define its maturity in the meaning of internet's evolution and available possibilities and necessities, which are connected with individual evolution stages. Elaboration of this framework along with concrete computation of websites evolution index is planned to be subject of further studies.

Acknowledgment This work and the contribution were supported by project “SP/2014/05—Smart Solutions for Ubiquitous Computing Environments” from University of Hradec Kralove.

References

1. Shneiderman, B.: Universal design. *Commun. ACM* **43**, 84–91 (2010)
2. Yourdon, E.: *Modern Structured Analysis*. Yourdon Press, Prentice-Hall International, Englewood Cliffs, New Jersey. ISBN 978-0135986240 (1989)

3. Weinreich, H., Obendorf, H., Herder, E., Mayer, M.: Not quite the average: an empirical study of web use. *ACM Trans. Web*, 2(1) (Article 1) (2008)
4. Boulding, K.: General Systems theory—the skeleton of science. *Manag. Sci.* vol. 2 No. 3, 197–208 Apr 1956 Reprinted in *E:CO* 6(1-2), 127–139 (1956/2004)
5. World Wide Web Foundation.: History of the Web. <http://www.webfoundation.org/vision/history-of-the-web/> (2008–2013)
6. Liou, C.Y., Cheng, W.C.: Manifold construction by local neighborhood preservation. In *Springer Lect. Notes Comput. Sci.* **4985**, 683–692 (2007)
7. Chen, A.Q., Harper, S.: Web Evolution: Method and Materials. Technical Report, University of Manchester. <http://wel-eprints.cs.man.ac.uk/74> (2008)
8. Hofkirchner, W.: Ludwig von bertalanffy. forerunner of evolutionary systems Theory. In: Gu, J., Chroust, G. (eds.) *The New Role of Systems Sciences For a Knowledge-based Society, Proceedings of the First World Congress of the International Federation for Systems Research*, Kobe, Japan, 6 (2005)
9. Loke, S.: *Context-Aware Pervasive Systems: Architectures for a New Breed of Applications*. Taylor and Francis Group (2007). ISBN 978-0849372551
10. Gillies, D. A.: *Understanding General Systems Theory. Nursing Management A Systems Approach*, pp. 56–74. W. B. Saunders Company, Philadelphia (1982)
11. Doyle, M.: *Beginning PHP 5.3*. Wiley, Indianapolis, Indiana. ISBN 978-8126527977 (2010)
12. Bertalanffy.: An outline of general system theory. *Br J Philos Sci* 1(2), 134–165 (1950)
13. Hofkirchner, W.: *General System Theory. The origins of General System Theory (GST)*. <http://www.hofkirchner.uti.at/wp-content/uploads/2010/10/GSTcombined.pdf> (2010)
14. Hercik, R., Slaby, R., Machacek, Z., Koziorek, J. Correlation methods of OCR algorithm for traffic sign detection implementable in microcontrollers. *Adv. Int. Syst. Comput.* **189**, 381–389 (2013). ISSN: 2194-5357
15. Krawiec, J., Penhaker, M., Krejcar, O., Novak, V., Bridzik, R., Web system for electrophysiological data management. In: *Proceedings of 2010 Second International Conference on Computer Engineering and Applications ICCEA 2010*, 19–21 Mar 2010, p. 404–407. Bali Island, Indonesia, vol. 1 (2010)
16. Machaj, J., Brida, P.: Performance comparison of similarity measurements for database correlation localization method. *Lect. Notes Comput. Sci.* **6592**, 452–461, (2011) ISBN 978-3-642-20041-0
17. BRIDA, P., MACHAJ, J.: A Novel enhanced positioning trilateration algorithm implemented for medical implant in-body localization. *Int. J. Antennas Propag.* **2013**, 10 (2013) Article ID 819695, ISSN: 1687-5877
18. Krejcar, O., Penhaker, M., Janckulik, D., Motalova, L.: Performance test of multiplatform real time processing of biomedical signals. In: *Proceedings of 8th IEEE International Conference on Industrial Informatics, INDIN 2010*, 13–16 July 2010, pp. 825–839. Osaka, Japan (2010). doi:[10.1109/INDIN.2010.5549635](https://doi.org/10.1109/INDIN.2010.5549635)
19. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic web—a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci. Am.* **284**(5), 35–43 (2001)
20. Ankolekar, A., Krötzsch, M., Tran, T., Vrandecic, D.: The two cultures: mashing up web 2.0 and the semantic web. In: *Proceedings of the 16th International Conference on World Wide Web*, 08–12 May 2007, Banff, Alberta, Canada (2007)
21. Penhaker, M., Cerny, M.: Sensitivity Analysis and Application of Transducers, In: *5th International Summer School and Symposium on Medical Devices and Biosensors*, 01–03 Jun 2008, pp. 103-106. Hong Kong, P.R.China, (2008) doi:[10.1109/ISSMDBS.2008.4575028](https://doi.org/10.1109/ISSMDBS.2008.4575028)
22. Rapoport, A.: *General systems theory*. *Int. Encycl. Soc. Sci.* **15**, 452–458 (1968)
23. Marzano, S., Aarts, E.: *The New Everyday View on Ambient Intelligence*. Uitgeverij 010 Publishers (2003). ISBN 978-9064505027
24. Behan, M., Krejcar, O.: Modern smart device-based concept of sensoric networks. *EURASIP J. Wirel. Commun. Net.* **2013**(155) (2013)

25. Schmidt, A.: Ubiquitous computing—computing in context. Ph.D. thesis, Computing Department, Lancaster University (2002)
26. Frain, B.: Responsive Web Design with HTML5 and CSS3. Packt Publishing, UK (2012). ISBN 978-9350237885
27. Longo, L., Kane, B.: A novel methodology for evaluating user interfaces in health care. In: 24th IEEE International Symposium on Computer-Based Medical Systems CBMS 2011, Bristol, England, June 27–30 (2011)
28. Vossen, G., Hagemann, S.: Unleashing Web 2.0, Morgan Kaufmann, San Francisco, pp. 1–68, ISBN 9780123740342, doi:[10.1016/B978-012374034-2.50002-2](https://doi.org/10.1016/B978-012374034-2.50002-2) (2007)
29. Wahlster, W., Dengel A. (eds.): Web 3.0: Convergence of Web 2.0 and the Semantic Web. Deutsche Telekom Laboratories, Technology Radar Feature Paper, Edition II/2006, June, pp. 1–23 (2006)

Chapter 42

A Novel Distributed Image Steganography Method Based on Block-DCT

Rosemary Koikara, Dip Jyoti Deka, Mitali Gogoi and Rig Das

Abstract Distributed Image Steganography (DIS) is a method of hiding secret information in multiple carrier images, making it more difficult to trace than conventional steganographic techniques, and requiring a collection of affected images for the retrieval of the secret data. In this paper we concentrate on performing DIS on grayscale images using Block-DCT (Discrete Cosine Transformation). Distributed Image Steganography using Block-DCT adds to the security of DIS by embedding the secret data in the Frequency Domain. This makes the carrier images more immune to various steganalysis attacks as the secret data is more evenly distributed amongst the pixels of the carrier images making it more difficult to determine its existence. We use parity check in order to compensate for round-off errors that are typically associated with DCT.

Keywords Steganography · Distributed steganography · Block-DCT · PSNR

R. Koikara · D.J. Deka · M. Gogoi
Department of Computer Science and Engineering and Information Technology, Don Bosco
College of Engineering and Technology, Guwahati 781017, Assam, India
e-mail: rosekoikara@gmail.com

D.J. Deka
e-mail: dipjyotideka123@gmail.com

M. Gogoi
e-mail: mitaligogoi@gmail.com

R. Das (✉)
Department of Computer Science and Engineering, National Institute of Technology,
Rourkela 769008, Orissa, India
e-mail: rig.das@gmail.com

42.1 Introduction

Steganography is the art of hiding information in ways that prevent the detection of hidden messages [1, 2]. It includes a vast array of secret communication methods that conceal the message’s very existence. Some of the more common methods include invisible inks, microdots, character arrangement, digital signatures, covert channels, and spread spectrum communications [3]. Many conventional steganographic schemes hide the secret data in a single host image. These techniques include least significant bit (LSB) insertion or frequency domain embedding using the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT) or Wavelet Transform (WT) [2, 4].

However, a common weakness of these techniques is that the secret data are all in a single information-carrier, and the secret data cannot be revealed completely [5], if the information-carrier is lost or crippled. Use of many duplicates may overcome the weakness but increase the danger of security exposure. Moreover conventional steganographic methods have restricted data hiding capacity.

Distributed Image Steganography overcomes these shortcomings by using a (k, n) threshold based image secret sharing technique for $k \leq n$ and allows large information payload embedding by generating n steganographic images. DIS allows (i) k or more steganographic images to reconstruct the secret image, and (ii) $(k - 1)$ or fewer images cannot reveal the secret image [1]. Figure 42.1 shows

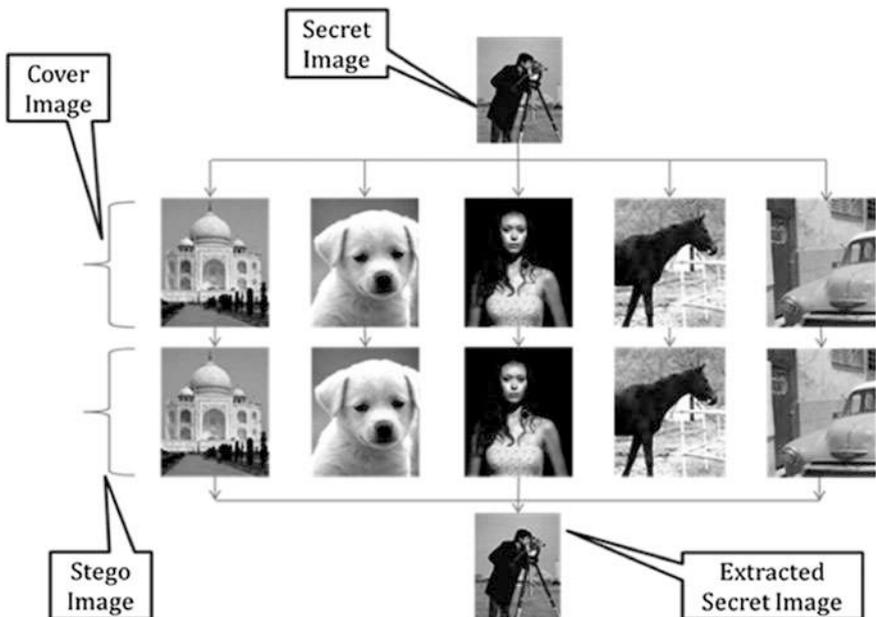


Fig. 42.1 The block diagram of a simple distributed steganographic system with $(3, 5)$ threshold scheme

the block diagram of a simple DIS system with (3, 5) threshold scheme that means the secret image will be distributed within 5 stego images and any 3 stego images are enough to regenerate the secret image.

Distributed Image Steganography in Block-DCT is refinement of the existing DIS technique in order to enhance the security of the secret data against various steganalysis attacks. In this paper, we have implemented Distributed Image Steganography by means of Block-DCT. We take the cover images and embed the shares of the secret data into the transformation domain of the cover images. During extraction, the shares need to be retrieved from the transformation domain of the stego images in order to recreate the original secret data. This whole process comprises of a series of computationally intensive operations and hence is impractical to deploy steganalysis for DIS in a massive scale.

This paper is organized as follows. Section 4.2.2 explains some related works of Distributed Steganography. The proposed novel embedding and extraction algorithms for DIS using Block-DCT are explained in Sect. 4.2.3. All experimental results are shown in Sects. 4.2.4 and 4.2.5 concludes.

4.2.2 Related Work

Different researchers employed different techniques for the purpose of distributing secret image over a set of stego images. Shamir's Secret Sharing Scheme and Thien and Lin's Secret Image Sharing Scheme are the two essential processes to protect secret image in DIS. Following are some of the related works carried out by some of the researchers.

4.2.2.1 Thien and Lin's Secret Image Sharing Scheme [6]

Suppose we want to divide the secret image D into n shadow images (D_1, \dots, D_n) , and the secret image D cannot be revealed without r or more shadow images. In the proposed method, they generate the $r - 1$ degree polynomial, by letting the r coefficients be the gray values of r pixels. Therefore, the major difference between Thien and Lin's method and Shamir's [7] is that they use no random coefficient. Because the gray value of a pixel is between 0 and 255, they took the prime number p be 251 which is the greatest prime number not larger than 255. To apply the method, it must truncate all the gray values 251–255 of the secret image to 250 so that all gray values are in the range 0–250. The image is divided into several sections. Each section has r pixels, and each pixel of the image belongs to one and only one section. For each section j , define the $r - 1$ degree polynomial as: $q_j(x) = (a_0 + a_1x + \dots + a_{r-1}x^{r-1}) \bmod 251$, Where a_0, \dots, a_{r-1} are the r pixels of the section, and then evaluate $q_j(1), q_j(2), \dots, q_j(n)$, The n output pixels $q_j(1) - q_j(n)$ of this section j are sequentially assigned to the n shadow images. Since for each

given section (of r pixels) of the secret image, each shadow image receives one of the generated pixels; the size of each shadow image is $1/r$ of the secret image. The reveal phase uses any r (of the n) shadow images, and the Lagrange's interpolation to extract the secret image.

42.2.2 An Estimation Approach to Extract Multimedia Information in Distributed Steganographic Images [1]

In this paper, a blind steganalysis technique is been proposed to attack DIS in which no host image is required for detecting and extracting hidden information. To develop this counter-measure for DIS they have put two assumptions: (i) One hidden image in a set of unsuspected steganographic images, (ii) Threshold value k is known.

The counter-measure process consists of three modules:

- (i) *Detection Module (DM)* is responsible for detecting possible steganographic images
- (ii) *Estimation Module (EM)* is responsible for extracting image shares embedded in steganographic images
- (iii) *Reconstruction Module (RM)* is responsible for combining quantized image shares to reconstruct the secret image

42.3 Proposed Novel Method for Distributed Image Steganography Based on Block-DCT

As Thien and Lin's method is primarily based on Spatial Domain and there is also a Steganalysis method to counter the LSB based DIS [1], a novel method for DIS based on Block-DCT is been proposed in this paper which will add much more security to the secret image as the secret information is being embedded in frequency domain and extraction of the secret information is much more difficult than spatial domain based techniques.

DIS using Block-DCT is a refinement of DIS. Here we use a polynomial equation to create shares of the secret image and embed these shares into multiple cover images that have been transformed using Block-DCT. The schematic/block diagram of the whole process is given in Figs. 42.2 and 42.3.

Our novel algorithm for DIS based on Block-DCT is based on (k, n) Threshold Scheme and has two parts, one for Embedding the Secret Image inside n -Cover Images and another for Extracting the Secret Image from k -Stego Image. The Secret Image is divided into several sections. Each section has k pixels, and each pixel of the secret image belongs to one and only one section. For each section j , we define the following $k - 1$ degree polynomial:

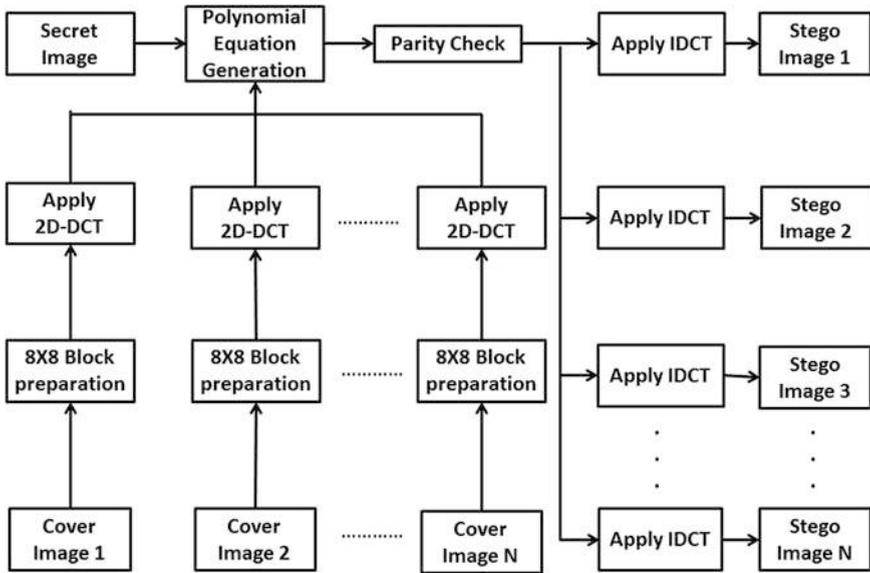


Fig. 42.2 Insertion of a secret image inside n-cover images

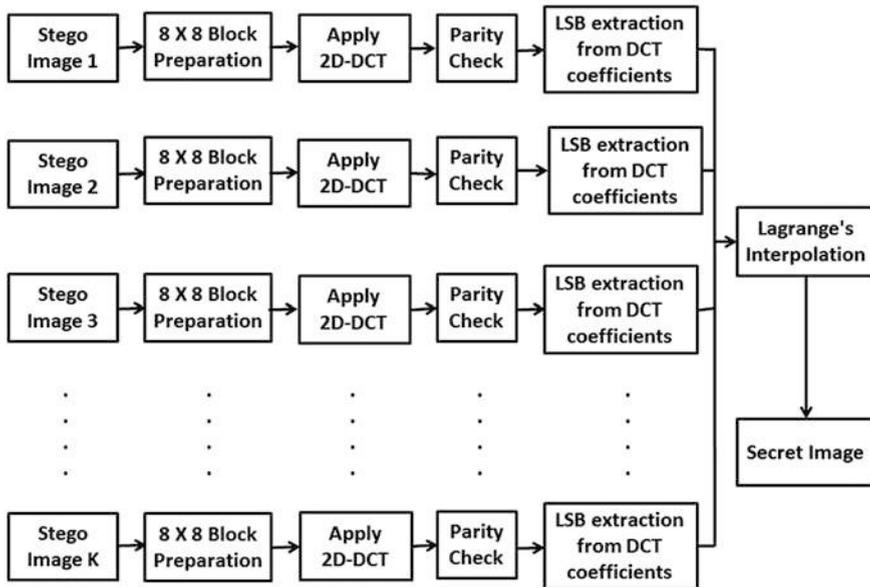


Fig. 42.3 Extraction of secret image from k-stego images

$$p_j(x) = (a_0 + a_1x + \dots + a_{k-1}x^{k-1}) \bmod 256 \quad (42.1)$$

$$q_j(x) = \text{floor} \left[(a_0 + a_1x + \dots + a_{k-1}x^{k-1}) / 256 \right] \quad (42.2)$$

where, value of x ranges from 1 to n and a_0 to a_{k-1} are Secret Image's k number of pixel's intensity values. These intensity values changes sequentially e.g., first we take 1–8 pixels' intensity values then 9–16 pixels' intensity values and so on. Modulus and Floor operations are performed using 256 as divisor because a grey level image has 256 different intensity levels. Equation (42.1) finds the Remainder Value after performing Modulus operation and (42.2) finds the Quotient value after performing Floor operation. This method consists of several phases as explained below.

42.3.1 Block-DCT

We have created blocks of size 8×8 of the cover images and perform DCT operation on each of these blocks. This makes sure that the secret image is evenly distributed amongst all the pixels on the 8×8 blocks.

42.3.2 Polynomial Equation

A polynomial equation is used to divide the secret image into its shares. The polynomial equation used:

$$S_x(i, j) = I(i \times k + 1, j) + I(i \times k + 2, j)x + \dots + I(i \times k + k, j)x^{k-1} \quad (42.3)$$

where, S_x is the share of the secret image I and j denote the pixel positions, I is the secret image, k is the threshold of the secret sharing scheme. We calculate both the remainder and the quotient of this equation as follows:

$$M_x(i, j) = (I(i \times k + 1, j) + I(i \times k + 2, j)x + \dots + I(i \times k + k, j)x^{k-1}) \bmod 256 \quad (42.4)$$

$$Q_x(i, j) = \text{floor} \left((I(i \times k + 1, j) + I(i \times k + 2, j)x + \dots + I(i \times k + k, j)x^{k-1}) / 256 \right) \quad (42.5)$$

where, M_x is the remainder of the polynomial equation, Q_x is the quotient of the polynomial equation, \bmod is the operation that calculates the remainder floor gives the nearest integer that is less than or equal to the real value. We need the values of both the remainder and the quotient to reconstruct the image during extraction.

42.3.3 Lagrange's Interpolation

Lagrange's interpolation is the Nth degree polynomial approximation formula to the function $P(x)$, which is known at discrete points x_i , $i = 0, 1, 2, \dots, N$

$$P(x) = \frac{(x - x_2)(x - x_3) \dots (x - x_n)}{(x_1 - x_2)(x_1 - x_3) \dots (x_1 - x_n)} y_1 + \frac{(x - x_1)(x - x_3) \dots (x - x_n)}{(x_2 - x_1)(x_2 - x_3) \dots (x_2 - x_n)} y_2 + \dots + \frac{(x - x_1)(x - x_2) \dots (x - x_{n-1})}{(x_n - x_1)(x_n - x_2) \dots (x_n - x_{n-1})} y_n \quad (42.6)$$

42.3.4 Parity Check

This is an error detection and correction code used to eliminate the rounded-off errors that may arise due to transformation into the frequency domain. Before performing Block-DCT over the cover image, the cover image is changed into DOUBLE format. After performing Block-DCT the pixels intensity values gets changed into frequency domain which has negative fraction values. Now as the LSB of these negative fraction values can't be modified to insert the Secret Image's data, it should have to be changed into positive integer number. After modifying the LSB of the cover images and performing Inverse 2-D DCT, to generate the stego image it is required to convert these pixels intensity value into UINT8 format from DOUBLE format (as the original cover image was in UINT8 format). Now at the time of change from DOUBLE to UINT8 format the fraction value gets rounded off. But the modification on LSB has been done in those fraction part itself. For example if a pixel's intensity is 169.78 then it gets rounded off into 170 and if it is 152.21 then rounded off to 152. So the value of every LSB gets changed. Thus a small amount of change can change the actual Secret Image into havoc [8]. To reduce this loss of information an algorithm is devised to detect and correct this error. The algorithm we have formulated is a modified form of parity check. This method does not completely eliminate rounded-off error; it just reduces to a certain extent. Following is the devised Parity Checking Algorithm.

42.3.4.1 Algorithm for Parity Check During the Insertion of Secret Image

Input: Stego-image
 Output: Stego-image with parity check information embedded into its 1st and 2nd LSB Position

- Step-1: Calculate even parity and embed it into the 1st LSB position
- Step-2: Calculate odd parity and embed it into the 2nd LSB position

42.3.4.2 Algorithm for Parity Check During the Extraction of Secret Image

Input: Stego Image
 Output: Stego Image with Corrected pixel value

- Step-1: Let x = Pixel value of the Stego Image.
- Step-2: Calculate even parity of x and embed it in x 's 1st LSB
- Step-3: Calculate odd parity of x and embed it in x 's 2nd LSB
- Step-4: If x is equal to the Pixel value then it is correct or else we need to increment or decrement the pixel value respectively.
- Step-5: Repeat Step-2 and Step-3 till x becomes equal to the pixel value

42.3.5 Proposed Novel Algorithm for DIS Based on Block-DCT

Embedding Algorithm

Input: n number of $M \times N$ Carrier Images and a $P \times Q$ Secret message/
 Image
 Output: n number of $M \times N$ Stego-Images

- Step-1: Read Secret Image and Cover Images
- Step-2: Divide the Cover Image into non overlapping blocks of size 8×8 and apply 2-D DCT on each of the blocks of the cover image.
- Step-3: Sequentially take k -number of not-shared-yet pixels of the Secret Image and use (42.1) and (42.2) to find the n number of Remainder and n number of Quotient value.
- Step-4: Change the 3rd LSBs of DCT transformed n Cover Images to insert each set of Remainder and Quotient values found in Step-4 in each of the Cover Images.
- Step-5: Add parity bit information into 1st and 2nd LSBs of every pixel of n -Cover Images using the algorithm proposed in Sect. 42.3.4.1.
- Step-6: Repeat Step-3, Step-4 and Step-5 until all the pixels of the Secret Image are embedded into n -Cover Images.
- Step-7: Write all the n Stego Images into the disk.

Extraction Algorithm

Input: k number of $M \times N$ Stego-Images
 Output: A $P \times Q$ Secret Image

- Step-1: Read k number of *Stego Images*
- Step-2: Divide the Stego Image into non overlapping blocks of size 8×8 and apply 2-D DCT on each of the blocks of the Stego image.
- Step-3: Extract k number of Remainders and k number of Quotients from the k *Stego Images* by extracting the 3rd LSBs of the pixels.
- Step-4: Use Parity bit Checking as described in Sect. 42.3.4.2 to reduce the rounded off error (as described in Sect. 42.3.4).
- Step-5: Use Lagrange's Interpolation to retrieve k number of pixel's intensity.
- Step-6: Repeat Step-3 and Step-4 until the total number of pixels of the *Secret Image* are processed.
- Step-7: Write the Extracted Secret Image into the disk.

42.4 Simulation and Results

In this section, some experiments are carried out on our proposed algorithm for Distributed Image Steganography (DIS). The measurement of the quality between the cover image f and stego-image g of sizes $M \times N$ is done using PSNR (Peak Signal to Noise Ratio) value and the PSNR is defined as:

$$PSNR = 10 \times \log(255^2/MSE) \quad (42.7)$$

where,

$$MSE = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (f(x,y) - g(x,y))^2 / (M \times N) \quad (42.8)$$

$f(x, y)$ and $g(x, y)$ means the pixel intensity value at position (x, y) in the cover-image and the corresponding stego-image respectively. The PSNR is expressed in dB. The larger PSNR indicates the higher the image quality i.e., there is only little difference between the cover-image and the stego-image. On the other hand, a smaller PSNR means there is huge distortion between the cover-image and the stego image.

All the simulation has been done using the MATLAB 7 program on Windows XP platform. Three different sets of 8-bit grayscale TIFF images of size 1024×1024 and 256×256 are used as the cover-images and secret image respectively to form the stego-images. Three different sets of cover images are considered to evaluate our results. Each set consists of five cover images. A single secret image was used to embed into all the three sets. Figure 42.4(1)–(5) shows the first set of original cover (carrier) images, Fig. 42.5(1)–(5) shows second set of cover images. Figure 42.6(1)–(5) shows third set of cover images of the proposed DIS method based on (4, 5) threshold scheme. Figure 42.7 shows the Original Secret Image and Fig. 42.8 shows three extracted secret images from Set 1, Set 2 and Set 3 of the cover images.



Fig. 42.4 Cover images set 1, (1)–(5) five cover images of proposed DIS method



Fig. 42.5 Cover images set 2, (1)–(5) five cover images of proposed DIS method

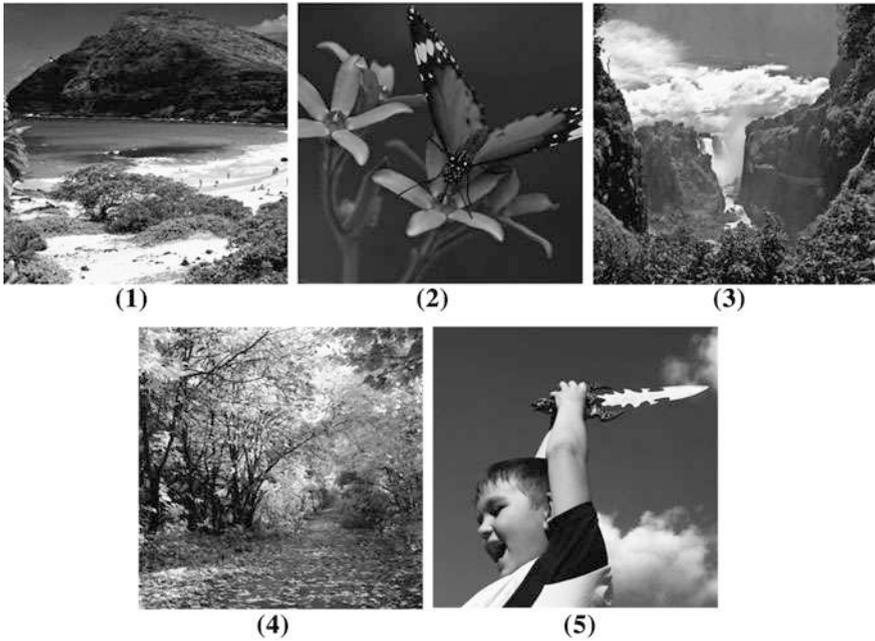


Fig. 42.6 Cover images set 3, (1)–(5) five cover images of proposed DIS method

Fig. 42.7 Original secret images

Steganography is the art of hiding information in ways that prevent the detection of hidden messages. Steganography, derived from Greek, literally means "covered writing". It includes a vast array of secret communications methods that conceal the message's very existence. These methods include invisible inks, microdots, character arrangement, digital signatures, covert channels, and spread spectrum communications.

Table 42.1 exhibit the PSNR comparison of Stego Images with their corresponding Cover Images for (4, 5) threshold scheme for all three sets of images. From Table 42.1 it is observed that for threshold scheme (4, 5) PSNR is greater than 40 dB for all the cases, so the quality of the stego image is quite acceptable and the deterioration in quality due to embedding of secret image cannot be distinguished by naked eye. Best result is achieved in case of Set 2 of Images, as

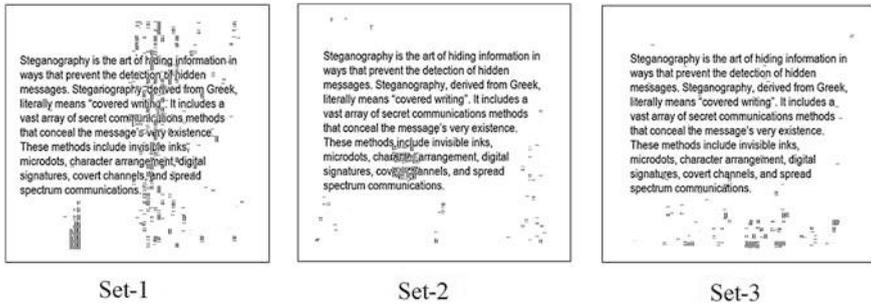


Fig. 42.8 Secret image extracted from cover images set 1, 2 and 3

Table 42.1 PSNR comparison of cover images and stego images for (4, 5) threshold scheme for three different sets of cover images and their corresponding stego images

Cover Image Set	PSNR (DB) between cover image and stego image					
	Threshold scheme (4, 5)					
	Cover image-1 and stego image-1	Cover image-2 and stego image-2	Cover image-3 and stego image-3	Cover image-4 and stego image-4	Cover image-5 and stego image-5	Original secret image and extracted secret image
Set 1	+48.92	+44.86	+50.43	+47.79	+48.50	+40.37
Set 2	+46.64	+48.53	+48.96	+46.01	+47.41	+45.08
Set 3	+47.59	+48.36	+47.27	+47.99	+47.21	+42.97

the PSNR is greater than +45 dB for the extracted secret image. If we pass the extracted secret image through a median filter then there will be further improvement of the extracted secret image quality.

42.5 Conclusion

Our proposed novel Distributed Image Steganographic method based on Block-DCT which uses (k, n) Threshold Scheme improves the security and the quality of the Stego Images. According to the simulation results the Stego Images of our proposed algorithm are very difficult to distinguish from the Cover Images. We have achieved a quite satisfactory quality of the extracted Secret Image for any set of chosen Stego Images for (4, 5) threshold scheme. Distributing the Secret Image among n-number of Cover Images keeps the Secret Image away from stealing; destroying by any unintended users and Block-DCT adds to the security of the secret image as the secret image embedding is done in frequency domain. Hence the proposed method may be more robust against brute force attack.

References

1. Bai, L., Biswas, S., Blasch, P.E.: An estimation approach to extract multimedia information in distributed steganographic images. In: Proceedings of the 10th International Conference on Information Fusion, Quebec, Canada, 9–12 July 2007
2. Jhonson, F.N., Jajodia, S.: Exploring steganography: seeing the unseen. In: Proceedings of the IEEE paper of Feb 1998
3. Cheddad, A., Condell, J., Curran, K., Kevitt, M.P.: Digital image steganography: survey and analysis of current methods. *J. Sign. Proces.* **90**, 727–752 (2010)
4. Li, B., He, J., Huang, J., Shi, Y.Q.: A survey on image steganography and steganalysis. *J. Inf. Hiding Multimedia Sign. Proces.* **2**(2), 142–172 (2011)
5. Provos, N., Honeyman, P.: Hide and seek: an introduction to steganography. *IEEE Secur. Priv.* **1**(3), 32–44 (2003)
6. Thien, C.C., Lin, J.C.: Secret image sharing. *J. Comput. Graph.* **26**(5), 765–770 (2002)
7. Shamir, A.: How to share a secret. *Commun. ACM.* **22**(11), 612–613, (1979)
8. Das, R., Tuithung, T.: A review on “A novel technique for image steganography based on block-DCT and Huffman encoding”. In: Proceedings of the 4th International Conference on Computer Graphics and Image Processing, ICGIP-2012, 6–7 Oct, Singapore, SPIE 2012

Chapter 43

An Improved History-Based Test Prioritization Technique Using Code Coverage

Avinash Gupta, Nayneesh Mishra, Aprna Tripathi, Manu Vardhan
and Dharmender Singh Kushwaha

Abstract Prioritization of test cases provides a way to run test cases with the highest priority earliest. Numerous empirical studies have shown that prioritization can improve a test suite's rate of fault detection. Software testers prioritize test cases, to reduce the cost of regression testing. History Based Approach is one of the methods to prioritize the test cases. This approach takes into account the history of each of the test cases of the test suite such as fault detection, number of executions and other such factors to prioritize the test cases in the coming sessions. This paper extends the above approach to the modified lines. The modified lines are being prioritized first and subsequently followed by the concerned test cases. The proposed approach has been able to detect fault faster than the previous approach with less effort in comparison to the previous approach.

Keywords Prioritization · History based · Fault detection · Test suit

A. Gupta (✉) · N. Mishra · A. Tripathi · D.S. Kushwaha
MNNIT, Allahabad, India
e-mail: rcs1051@mnnit.ac.in

N. Mishra
e-mail: nayaneesh@gmail.com

A. Tripathi
e-mail: aprnatripathi@gmail.com

D.S. Kushwaha
e-mail: dsk@mnnit.ac.in

M. Vardhan
National Institute of Technology, Raipur, India
e-mail: mvardhan.cs@nitrr.ac.in

43.1 Introduction

Since test development is costly, software testers often save the test suites they develop, so that they can reuse those test suites later as software evolves. Such test suite reuse, in the form of regression testing, is pervasive in the software industry and, together with other regression testing activities, has been estimated to account for as much as one-half of the cost of software maintenance. Running all of the test cases in a test suite, however, can require a large amount of effort. Hence, researchers have considered various techniques for reducing the cost of regression testing. Three different techniques have, therefore, been proposed for test suite reduction as—prioritization, selection and minimization of test suite.

A test suite *minimization* technique lowers the cost by reducing a test suite to a minimal subset that maintains equivalent coverage of original set with respect to particular test adequacy criterion [14]. Test suite minimization techniques [5], however, can have some drawbacks. Although one clan of researchers think that, in certain cases there is little or no loss in the ability of a minimized test suite to reveal faults [2] in comparison to its un-minimized original, the other clan thinks otherwise [18]. The fault detection ability of test suites can be severely compromised by minimization.

Test case *prioritization* is the process of scheduling test cases in an order to meet some performance goal [16]. Prioritization gives priority to test cases based on criteria. The criteria may be code coverage etc. However, one of the limitations in this process is that the fault detection efficiency of the test suite may be compromised. The test suite may contain test cases on higher priority which may not be able to detect the errors [8]. Hence, several techniques have been proposed for prioritizing the existing test cases to accelerate the rate of fault detection in regression testing. Some of these approaches are Coverage-based Prioritization [16], Interaction Testing, Distribution-based Approach [10], Requirement-based Approach, and the Probabilistic Approach [9].

All these approaches apart from probabilistic approach referred above consider prioritization as an unordered, independent and one-time model. They do not take into account the performance of test cases in the previous regression test sessions, such as the number of times a test case revealed faults [19]. History Based Approach (HBA) has been applied to increase the fault detection ability of the test suite. Kim and Porter [9] considered the problem of prioritization of test cases as a probabilistic approach and defined the history-based test case prioritization. Alireza et al. [8] proposed an extension of history-based prioritization proposed in [9], and modifies the equation given by Kim and Porter [9], to have dynamic coefficients. The priority is calculated using the mathematical equation by computing the coefficients of the equation from the historical performance data.

In this paper, we propose a new approach which is an extension of the history based approach in [9]. Unlike in [9], where the prioritization equation has been applied on each test case, we apply the approach on each modified line of the code. The application of the prioritization at the code level makes the selected test suite

more effective in terms of fault detection effectiveness in comparison to that obtained by applying the approach in [9].

The rest of the paper is organized as follows. Section 4.3.2 reviews related literature. In Sect. 4.3.3, we present the proposed approach and implementation. Section 4.3.4 describes performance analysis and comparison results. We conclude the paper and discuss future work in Sect. 4.3.5.

4.3.2 Related Research Work

Regression Testing is the process of validating modified software to provide confidence that the changed parts of the software behave as intended and that the un-changed parts of the software have not been adversely affected by the modification [7].

In [15], Rothermel showed that because of time and resource constraints, it is never feasible to re-run the entire test-suite for regression testing, which may be very large in case of software of large size. Hence there is a need for prioritization, selection, minimization of test suite. Horgan and London applied linear programming to the test case minimization problem in their implementation of a data-flow based testing tool, ATAC [6]. Chen and Lau [2] applied GE and GRE heuristics. The GE and GRE heuristics can be thought of as variations of the greedy algorithm that is known to be an effective heuristic for the set cover problem [13]. Offutt et al. [12] also treated the test suite minimization problem as the dual of the minimal hitting set problem, i.e., the set cover problem [13]. Marre and Bertolino formulated test suite minimization as a problem of finding a spanning set over a graph [11]. Tallam and Gupta developed the greedy approach further by introducing the delayed greedy approach, which is based on the Formal Concept Analysis of the relation between test cases and testing requirements [18]. Harder et al. approached test suite minimisation using operational abstraction [4]. Harder et al. use the widely studied Daikon dynamic invariant detector [3] to obtain operational abstractions.

Test case prioritization is the process of scheduling test cases in an order to meet some performance goal [16, 17]. Agrawal et al. [1] in his work considered version specific test case prioritization [1] instead of general test case prioritization [1] so that the final test suite may contain test cases relevant as per the modifications. However, his approach required too many iterations to arrive at the final test suite and hence not very efficient. However, as discussed in [9], most of the prioritizing algorithms are unordered, independent and one-time model. In [14], it was stated that feedback mechanism makes the fault detection effectiveness of the test suite better.

From the discussion above, it can therefore be easily deciphered that those approaches which consider history to prioritize the test cases would certainly be better in terms of fault detection effectiveness than those that did not. HBA approach takes into account the history of test cases while prioritizing them in the

present session to increase the fault detection effectiveness of the test suite. The technique by Alireza et al. directly computes the priority of each test case using the historical information of the test case, such as the number of executions, the number of times it exposed a fault and other relevant data. In this approach, the prioritization has nothing to do with the source code. The test cases are prioritized simply on the basis of their history.

43.3 Proposed Approach

In our proposed approach, we have extended the previous approach [8] by prioritizing the modified lines. The history is kept for each of the modified lines which act as feedback for the next session. In our proposed approach, the test cases are selected for each modified line such that the test case is having the maximum coverage among all the test cases which contain the modified line. By the phrase ‘containing a modified line’ we mean that the test case executes the modified line in its line of execution.

Our proposed approach has been implemented in a ‘C’ program and the history is being stored in text format in text files. The history contains all the test cases and parameters such as number of executions of test case, number of times fault detected by test case, number of times each line has been delayed execution. The test cases contain the number of all the lines traversed along the line of execution of the test case. The parameters have been stored in the form of arrays, where each index represents a line in the code. The proposed approach in this paper includes the steps shown in Fig. 43.1.

Step 1: Extract History from Database

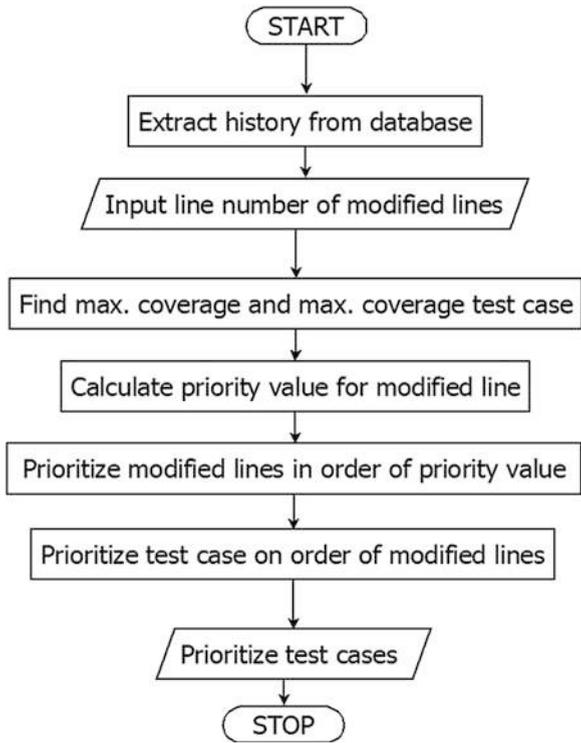
In this phase all the parameter values of the previous session are extracted from the database by the program into arrays for use in the present session. The following parameters are extracted from the database:

$ec_k[]$ is an array to store no. of times a line has executed till the session k . $fc_k[]$ is an array to store no. of times a line has been detected as faulty till the session k . $h_k[]$ is the value is set to 0 if the line is executed in the last session else keeps on increasing from 0 by one in each session till it does not execute. $mod_locode[]$ is an index is set to 1 if the corresponding line has been modified else to 0. $del_locode[]$ is an index is set to 1 if the corresponding line has been deleted else 0. $PR_{k-j}[]$ is the value at an index indicates the priority value of the corresponding line in last session. $PR_k[]$ is the value at an index indicates the priority value of the corresponding line in present session.

Step 2: Input Modified Lines The modified lines are taken as input from the user through a well defined interface of the program. Any new test cases are also entered through this interface.

Step 3: Find max coverage and ax coverage test case If the modified line say, m_i is present in a test case T_i , then max coverage of line $m_i = \max$ (no. of lines in

Fig. 43.1 Flow chart of proposed algorithm



T_i /total number of lines) * 100 for all T_i in which m_i is found. T_i is the max coverage test case for m_i , if T_i has max code coverage.

Step 4: Calculate Priority Value for Modified Value For each modified line, the priority value is calculated using Eq. (43.1) [11].

$$PR_k = (\alpha.h_k + \beta.PR_{k-1})/k \tag{43.1}$$

$$\alpha = \left(1 - ((fc_k + 1)/(ec_k + 1))^2\right)_k^h \tag{43.2}$$

$$\beta = ((fc_k + 1)/(ec_k + 1))^x \tag{43.3}$$

In Eq. (43.1), $0 \leq \alpha, \beta < 1, k \geq 1$.

In Eq. (43.3), $x = 1$ if the test case has revealed some fault in the previous session and $x = 2$ if the test case has not revealed any fault in the previous session. In Eq. (43.1), h_k is the test cases execution history. ec_k denotes the total number of executions done by a test case till the session k . fc_k denotes the total number of faults detected by a test case till the session k . In Eq. (43.1), PR_0 is defined for each test case as the percentage of code coverage of the test case. The presence of PR_0 will be helpful in refining the ordering of the test cases in the first session.

Step 1: Prioritize Modified Lines

Modified lines are prioritized by their corresponding PR_k , in descending order. If the modified lines m_1 , m_2 and m_3 have PR_k values as— $PR_k [m_1] = 10.56$ $PR_k [m_2] = 54.56$ $PR_k [m_3] = 9.64$. Hence ordering would be— m_2, m_1, m_3 .

Step 2: Prioritize Test cases in order of modified lines

Max coverage test case for, say $m_1 = T_2$ $m_2 = T_1$ $m_3 = T_3$

Hence ordering of test cases in order of the respective modified line m_2, m_1 and m_3 would be— T_1, T_2, T_3 .

Step 3: Output prioritized the test cases

The final output for session k is T_1, T_2, T_3 . After prioritizing, the test cases are executed. Let us assume that only 40 % of all the test cases prioritized are able to get executed. Out of all the test cases executed, there are certain test cases which detect fault, and after debugging a fault is detected. Then, the parameters are updated in the following manner:

- For each executed line i , in the present session k , increment the value of parameter ec_k by 1 and set the value of h_k to 0 for the line i .
- For the rest of the lines which did not execute in the present session k , increment the value of h_k by 1.
- For each faulty line i , detected in session k , increment the value of parameter fc_k by 1 for the line i .

The database is updated with all these modifications.

43.4 Implementation

The proposed approach has been implemented using 'C' program. The database to keep the history and all the test cases has been kept in two text files: 'testcases.txt' and 'Parameters.txt'. The file 'testcases.txt' contains the test cases in the form of traces of each of the test case. This means that 'testcases.txt' contains the lines which will be covered by each test cases once they execute. There is another file called 'Parameters.txt' which keeps the history of all the parameter values which will be used to calculate the priority value PR_k for each modified line. The C program, itself has namely two sections:

- **An interface:**

This is the interface provided to insert all the test cases. This interface is also used to insert the line numbers of modified, deleted and added lines. It is also used to insert the line number of faulty line of the previous session.

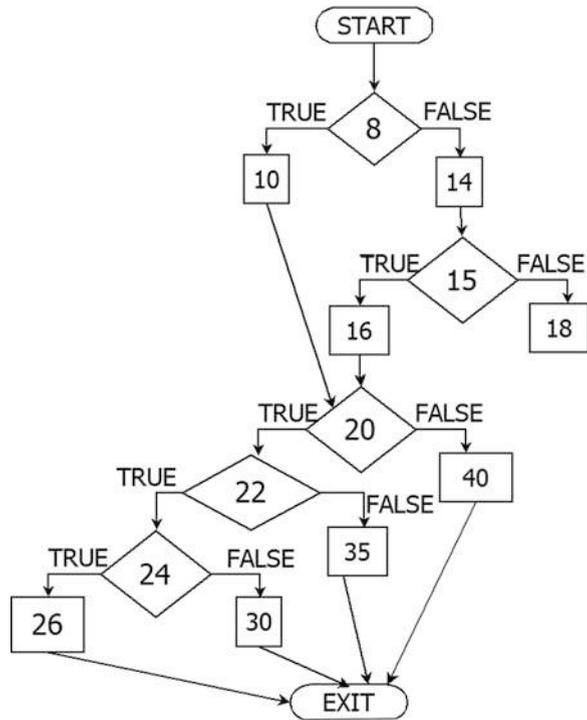
- **Prioritizing section:**

This section calculates the priority of each modified line and outputs them in order of priority. This section also outputs all the test cases in order of priority.

Table 43.1 Changes in the sample program

Line no.	Original line	Modified line
8	$a > 0$	$a < 0$
26	$(1/(b + 2))$	$(1/(b-1))$
30	$(1/(x + 4))$	$(1/(x-4))$
35	$(1/(x + 5))$	$(1/(x-5))$
40	$(1/(x + 5))$	$(1/(x-5))$

Fig. 43.2 Control flow graph for program in Tables 43.2



Case Study

The proposed approach is demonstrated with an example here. We are considering a small ‘C’ program as that have 43 lines and that calculates the value of mathematical equations. The equation consists of variables x and y whose value depends on the value of a, b, c, d and e.

The program is modified at line numbers 8, 26, 30, 35 and 40. The changes in each of the modified lines are shown in Table 43.1. These modifications in the program will introduce divide by zero error in the program.

The Control Flow Graph (CFG) for the sample program and the modified sample program is same since no any change in branch condition and none of the line is added that is given in Fig. 43.2. The CFG has been developed using

SourceCode Visualizer Tool [2] which is a plug-in of Eclipse. Based on branch coverage, we have the following test cases:

T1-8 9 10 11 20 38 39 40 41 42 43; **T2**-8 12 13 14 15 17 18 19 20 21 22 33 34 35 36 37 42 43; **T3**-8 12 13 14 15 16 19 20 21 22 23 24 25 26 27 42 43; **T4**-8 12 13 14 15 17 18 19 20 21 22 23 24 25 26 27 42 43; **T5**-8 9 10 11 38 39 40 41 42 43; **T6**-8 12 13 14 15 16 19 38 39 40 41 42 43; **T7**-8 12 13 14 17 18 19 20 21 22 23 28 29 30 31 32 37 42;

These test cases are kept in the 'testcases.txt' file.

Now, for session $k = 1$:

Step 1: Extract history from database The test cases are extracted from the 'testcases.txt' file into the program which is developed to implement the proposed algorithm. Before the first session, for each modified line, the value of parameters, namely h_k , e_{c_k} , f_{c_k} , pr_k and pr_k 1, is set to 0. For each modified line the value of parameter x is set to 2.

Step 2: Input line numbers of modified lines As is mentioned in Fig. 43.3, the line numbers of modified lines are 8, 26, 30, 35 and 40. The line numbers of modified lines are entered into the program via the input interface of the program.

Step 3: Find max coverage and max coverage test case Line 26 is found in test cases: T3, T4. Since, Code coverage of a test case $T = (\text{no. of lines in the test case} / \text{total number of lines in the program}) * 100$

So, Code coverage of test case T3 = $(17 / 50) * 100 = 34 \%$

Code coverage of test case T4 = $(18 / 50) * 100 = 36 \%$

Hence, max code coverage for line 26 = 36 %

Since, Test case T4 is having the max code coverage value of 36 among all the test cases containing the line 26. So, max code coverage test case for line 26 = T4.

As we calculated the value of max code coverage and found out the max code coverage test case for line number 26, we can similarly compute the value of max code coverage and find out the max code coverage test case for line number 8, 30, 35 and 40. The results are:

Max code coverage test case (MCCTC) for line 8 = T2, and Max Code Coverage (MCC) = $(18 / 50) * 100 = 36 \%$

MCCTC for line 30, 35 and 40 are T7, T2 and T6 respectively. The MCC for line 30, 35 and 40 are 36 %, 36 % and 26 % respectively.

Step 4: Calculate priority values for modified lines For session $k = 1$, the status of all the parameters is as shown in the Table 43.2. Calculating

Now, substituting the values of h_k , e_{c_k} , f_{c_k} , x and PR_{k-1} in Eqs. (43.1), (43.2) and (43.3) and from Table 43.2, to calculate the corresponding value of PR_k for lines 8, 26, 30, 35 and 40. We get: $PR_k [8] = 36$, $PR_k [26] = 36$, $PR_k [30] = 36$, $PR_k [35] = 36$, $PR_k [40] = 26$.

Step 5: Prioritize the modified lines in order of priority value Based on the priority values calculated in step 4, modified lines in order of priority as per PR_k values are: 8, 26, 30, 35, 40.

Table 43.2 Status of parameters before session $k = 1$

Line no.	h_k	ec_k	fc_k	x	PR_k	PR_{k-1}
8	0	0	0	2	0	36
26	0	0	0	2	0	36
30	0	0	0	2	0	36
35	0	0	0	2	0	36
40	0	0	0	2	0	26

Table 43.3 Status of parameters before session $k = 2$

Line No.	h_k	ec_k	fc_k	x	PR_k	PR_{k-1}
8	0	1	1	1	0	36
26	0	1	1	1	0	36
30	0	1	1	1	0	36
35	0	1	1	1	0	36
40	1	0	0	2	0	26

Step 6: Prioritize the test cases in order of modified lines Modified lines in order of priority as per PR_k value are: 8, 26, 30, 35, 40. Max code coverage test case for Line 8 is T2, Line 26 is T4, Line 30 is T7, Line 35 is T2, Line 40 is T6. So ordering the test cases in the same order as their respective modified lines are, we get: T2, T4, T7, T2, T6. Removing the repeated test case T2 from the fourth place, we get T2, T4, T7, T6.

Step 7: Output prioritized the test cases The final output for session $k = 1$ is T2, T4, T7, T6. After the end of session 1, 40 % of all the test cases are executed i.e., test cases T2, T4, T7 are executed in the order of priority as given by the final output of session 1. After the execution of the test cases it is found that all the test cases fail and subsequently debugging is carried out. As a result of debugging, fault is found at line number 8, 26, 30, and 35.

Session 2: (For session $k = 2$) Lines to be prioritized in session $k = 2$ are 26, 30, 35, 40 and 8. The status of parameters shown in Table 43.3 will be used in session 2. All the seven steps of the proposed approach are followed for $k = 2$.

The final output for session $k = 2$ is: T2, T4, T7, T6. We then used 40 % of the total test cases to execute. The test cases T2, T4, T7 are executed in the given order. T6 is left out this time. No new faults found as all the faults have already been revealed in session $k = 1$.

Session 3: (For session $k = 3$) The status of parameters shown in Table 43.4 will be used in session 3.

The final output for session $k = 3$ is: T6, T2, T4, T7. T6, T2, T4, T7 was executed in order. T6 fails and reveals the error at line number 40. Thus all the errors were revealed in 3 sessions. Hence, we can conclude that the proposed approach has a better rate of fault detection than the Alireza approach.

Table 43.4 Status of parameters before session $k = 3$

Line no.	h_k	ec_k	fc_k	x	PR_k	PR_{k-1}
8	0	2	1	2	0	18
26	0	2	1	2	0	18
30	0	2	1	2	0	18
35	0	2	1	2	0	18
40	2	0	0	2	0	13

Table 43.5 Proposed Approach Results

Program	Line no. modified	Session No.	Faulty line detected by Proposed approach	Faulty line detected by Alireza approach	No. of sessions in Proposed approach	No. of sessions in Alireza approach
Branch coverage Sample program	26, 30, 35, 40, 8	S1,S2	26, 30, 35, 40, 8	26, 30, 35, 40	2	2
Bank account	6, 17, 22, 24, 27	S1, S2, S3	17, 22, 24, 27, 6	17, 22, 24, 27, 6	3	3
Library management	199, 172, 223, 143, 126	S1, S2, S3	199, 172, 223, 143, 126	199, 172, 223, 143, 126	2	3
Kruskal algorithm	97, 93, 106, 47, 110	S1, S2	97, 106, 110, 47, 93	97, 106, 110, 47, 93	3	3
Payroll management system	70, 118, 124, 207, 231	S1, S2, S3	124, 118, 231, 207, 70	124, 118, 231, 70, 207	2	3
Heap sort	11, 12, 19, 27, 30	S1, S2, S3	12, 19, 30, 27, 11	12, 19, 30, 27, 11	2	3
Airline reservation system	27,43, 64,97,126	S1, S2, S3	27,43, 64,97,126	27,43, 64,97,126	2	3
Linked list	26, 78, 280, 54, 147	S1, S2, S3	280, 147, 78, 54, 26	280, 147, 78, 54, 26	2	3
Tic-tac toe	159, 164, 191, 195, 261	S1, S2, S3	159, 164, 191, 195, 261	159, 164, 191, 195, 261	2	3
Student records management System	24, 27, 66, 69, 80	S1, S2, S3	24, 27, 66, 69, 80	24, 27, 66, 69, 80	3	3

43.5 Performance Analysis

Proposed approach is illustrated by ten programs of Java, C and C++ based platforms. After that, we compared our proposed approach with Alireza et al. approach. For comparison, we applied our proposed approach as well as Alireza approach on ten example programs and five faults were seeded in each of the programs with multiple sessions of regression test. Results in Table 43.5 shows the faults detected in each session by our proposed approach and Alireza approach. This is followed by a comparison of the total number of sessions to find all the faults in proposed approach and Alireza approach as shown in Table 43.5. After analyzing the results, we find that the number of fault detected per session is more than or equal to in our proposed approach than in comparison to Alireza approach. At the same time the number of sessions to discover all faults in our proposed approach is less than or equal to Alireza approach. Hence we can say that, our approach is more efficient in terms of fault detection effectiveness.

43.6 Conclusion and Future Work

Proposed approach is helpful in early detection of fault. The proposed approach brings out those test cases to the front which are relevant to the modifications made. Also, the proposed approach is able to prioritize test cases in less number of iterations. This is because only those test cases are being processed which are having the modified line in their line of execution, instead of calculating the priority value for each of the test case as is done in Alireza approach. In the proposed approach we have used different factors in the history of each modified line. In future work, we may consider other factors such as severity of fault detected, which may help to refine the process of prioritization.

References

1. Aggrawal, K., Singh, Y., Kaur, A.: Code coverage based technique for prioritizing test cases for regression testing. *ACM SIGSOFT Softw. Eng. Notes* **29**(5), 1–4, (2004)
2. Chen, T.Y., Lau, M.F.: Dividing strategies for the optimization of a test suite. *Inf. Proces. Lett.* **60**(3), 135–141 (1996)
3. Ernst, M.D., Cockrell, J., Griswold, W.G., Notkin, D.: Dynamically discovering likely program invariants to support program evolution. *IEEE Trans. Softw. Eng.* **27**(2), 99–123 (2001)
4. Harder, M., Mellen, J., Ernst, M.D.: Improving test suites via operational abstraction. In: *Proceedings of the 25th IEEE International Conference on Software Engineering*, pp. 60–71 (2003)
5. Harrold, M.J., Gupta, R., Soffa, M.L.: A methodology for controlling the size of a test suite. *ACM Trans. Softw. Eng. Methodol. (TOSEM)* **2**(3), 270–285 (1993)

6. Horgan, J.R., London, S.: A data flow coverage testing tool for C. In: IEEE Proceedings of the Second Symposium on Assessment of Quality Software Development Tools, pp. 2–10 (1992)
7. Hsu, H.-Y., Orso, A. Mints: a general framework and tool for supporting testsuite minimization. In: Proceedings of IEEE 31st International Conference on Software Engineering. ICSE, pp. 419–429 (2009)
8. Khalilian, M., Azgomi, A., Fazlalizadeh, Y.: An improved method for test case prioritization by incorporating historical test case data. *Sci. Comput. Program.* **78**(1), 93–116 (2012)
9. Kim, J.-M., Porter, A.: A history-based test prioritization technique for regression testing in resource constrained environments. In: Proceedings of the 24th IEEE International Conference on Software Engineering. ICSE, pp. 119–129 (2002)
10. Leon, D., Podgurski, A.: A comparison of coverage-based and distribution-based techniques for filtering and prioritizing test cases. In: 14th IEEE International Symposium on Software Reliability Engineering. ISSRE, pp. 442–453 (2003)
11. Marfé, M., Bertolino, A.: Using spanning sets for coverage testing. *IEEE Trans. Softw. Eng.* **9**(11), 974–984 (2003)
12. Offutt, J., Pan, J., Voas, J.M.: Procedures for reducing the size of coverage based test sets. In: Proceedings of the 12th International Conference Testing Computer Software. Citeseer, pp. 111–123 (1995)
13. Papadimitriou, H., Steiglitz, K.: Combinatorial optimization: algorithms and complexity. Courier Dover, New York (1998)
14. Rothermel, G., Harrold, M.J., Ostrin, J., Hong, C.: An empirical study of the effects of minimization on the fault detection capabilities of test suites. In: Proceedings of the IEEE International Conference on Software Maintenance, pp. 34–43 (1998)
15. Rothermel, G., Harrold, M.J., von Ronne, J., Hong, C.: Empirical studies of test-suite reduction. *Softw. Test. Verification Reliab.* **12**(4), 219–249 (2002)
16. Rothermel, G., Untch, R.H., Chu, C., Harrold, M.J.: Prioritizing test cases for regression testing. *IEEE Trans. Softw. Eng.* **27**(10), 929–948 (2001)
17. Tallam, S. Gupta, N.: A concept analysis inspired greedy algorithm for test suite minimization. *ACM SIGSOFT Softw. Eng. Notes.* **31**, 35–42 (2005)
18. Wong, W.E., Horgan, J.R., Mathur, A.P., Pasquini, A.: Test set size minimization and fault detection effectiveness: a case study in a space application. *J. Syst. Softw.* **48**(2), 79–89 (1999)
19. Yoo, S., Harman, M.: Regression testing minimization, selection and prioritization: a survey. *Softw. Test. Verification Reliab.* **22**(2), 67–120 (2012)

Chapter 44

Local Pricewatch Information Solicitation and Sharing Model Using Mobile Crowdsourcing

Hazleen Aris

Abstract Shopping, in particular groceries shopping, is a necessity and has become part and parcel of our life, be it daily, weekly or monthly. Tied to this activity is the concern over the ever increasing price of goods. Ideally, one would want to be able to compare the price of an item between one shop and another before getting the best deal. However, this is impracticable if not impossible due to a number of factors, especially the time constraint. Nevertheless, advances in the internet technology, combined with the widespread use of the internet-ready devices that are often packaged with mobile internet plans have enabled rapid and timely information sharing, hence, creating an opportunity that can overcome the impossibility. People who go to different shops everyday can share information on the prices of items at the shops that they visit, which will be useful for those who are planning to buy similar items. What is missing is the mechanism to enable such information sharing. This paper therefore explains about such mechanism that we call the local pricewatch information solicitation and sharing (LoPrice) model that uses voluntary crowdsourcing technique to solicit the information from the public. Prior to that, justification on the need for the model is presented. Judging from the success of its counterparts in other areas of social networking, it is expected that the model can quickly gain interest from the public, hence helping them to save on their expenditures.

Keywords Mobile crowdsourcing • Price comparison • Crowdsourcing model • Open crowdsourcing • Voluntary crowdsourcing

H. Aris (✉)
College of Information Technology, Universiti Tenaga Nasional,
43000 Kajang, Selangor, Malaysia
e-mail: hazleen@uniten.edu.my

44.1 Introduction

Were we ever be in a situation where the price of good that we had just bought from a shop is found cheaper at another nearby shop? We would have bought the item at the latter shop should we knew it, right? But how could we possibly know about it unless somebody else shares the information with us? And how can we know the information before it is too late? In this modern society, groceries shopping is certainly inevitable and has become the main source of obtaining food and other supplies. It is no longer an option, but a necessity. Gone were the days where raw food can be obtained just by plucking them from the backyard plants or from the livestock reared around the house.

When one thinks about shopping, one cannot help but to also think about the expenses incurred by it. The ever increasing price of goods, that is becoming more and more significant lately, has therefore become the main concern of many people. Latest report on household expenditure survey by the Department of Statistics Malaysia [1], which is done once every 5 years, showed that the average monthly household expenditure had increased by 12.1 % from 2004/05 to 2009/10. An independent survey [2] also showed that increment on prices of items even reached 43 % within a six-month period, which was certainly burdensome as the prices hike was far beyond the yearly income increment. Making life affordable and less burdensome is so important that it has now become the concern of our (Malaysian) government. A number of initiatives are seen being implemented by the government to help easing the citizens with their expenditures such as the cash aid scheme [3].

In our opinion, solution to this problem can be approached from at least three directions. First, is by raising the living standard of low-income household, which happens to be one of the national key result areas (NKRAs) of our government [4]. Second, is by keeping the goods prices low and third, is by being more selective in buying the goods. While the first approach is a long term approach, the second approach may require strict and close monitoring from the authorised bodies. Therefore, it is the third approach that seems most amenable to all of us and hence, the approach that we are interested in.

As we are aware, and probably as part of the marketing strategies, prices for the same goods differ between one shop and another. This difference can be significant, amounting to as much as RM10 per item, if, for example, the other shop is having a sale. Most of the time, we have to depend on luck in order to get the best price as the price changes from time to time and the pattern of change is unknown to us. In an ideal situation, we would want to be able to compare the price between one shop and another for each item that we want to buy. Unfortunately, it is time consuming to move from one shop to another just to compare prices and then to return to the shop that offered the lowest price to buy it. Let alone to do it within a reasonable timeframe. It will also incur additional cost, as we need transportation to move about. Not to mention the difficulty in getting a parking space and et cetera, which render this practice infeasible. To overcome these problems, the

local pricewatch information solicitation and sharing (LoPrice) model is therefore proposed. The aim is to help the users in finding the needed items at the most reasonable, if not lowest price and at nearby stores. The difficulty in comparing prices of items to be bought within an acceptable period of time is therefore the problem statement to be addressed by the research.

44.2 Background

Recent advances in the internet and mobile technology have shed some light towards overcoming the stated problem above. The widespread use of internet-ready mobile devices combined with the availability of more and more affordable mobile data plans and packages has enabled the users to share information on just about anything at anytime. One type of information that they can share, and should be sharing is on the prices of items. People who go to different shops everyday can share information on the prices of items at the shops that they visit with others using their mobile devices. This information will be useful to those who are planning to buy similar items. In this way, the task of visiting each different shop can be distributed amongst them and price comparison can be done by just tapping and scrolling the mobile devices. This idea of distributing the potential solutions to the public in order to obtain the best one is what Surowiecki [5] termed as ‘the wisdom of crowds’ and has become amongst the seminal work in crowdsourcing. The term ‘crowdsourcing’ was later formally coined by Howe [6] to mean the act of outsourcing tasks to an undefined, large group of people or community (the crowd) through an open call [7]. Crowdsourcing process generally comprises the following steps [7].

1. Crowdsourcer proposes tasks and make it known to the public, with clear requirements and reward if any.
2. The public submit potential solutions.
3. The public investigate, or evaluate the solutions and choose the best one.
4. Crowdsourcer determines the best solution and reward the winner.
5. Crowdsourcer gets and owns the best solution.

Two prerequisites of crowdsourcing are (1) the open call format and (2) the large network of potential labourers [6]. Figure 44.1 shows the steps in a crowdsourcing process in a slightly different arrangement, but maintaining the same composition.

Motivations to participate in a crowdsourcing exercise vary. More often than not, the strongest motivation is monetary reward, as discovered by Brabham [8], although a more recent study showed otherwise [8]. As a result, crowdsourcing has been successfully applied in business or other profit making sectors such as Threadless and iStockphoto. The potential of crowdsourcing has also been explored in auction [8].



Fig. 44.1 The eight steps in a crowdsourcing process

However, the presence of crowdsourcing and widespread availability of mobile devices per se are not sufficient to solve the stated problem. A model is needed for the crowdsourcing to be successfully applied in information sharing of items prices and for the benefits to be optimally reaped, which prompts the need for this research. Furthermore, Brabham [9] argued that the benefits of crowdsourcing should also be enjoyed by the non-profit making sectors. Though examples on the use of crowdsourcing for non-profit purposes are already seen [10–12], rooms for research in this area are still plenty.

With regard to price comparison application, a number of applications were found as a result of our exhaustive search, noticeably in the US [13–18] and France [19, 20]. In the context of Malaysia, only three such applications were found as shown in Table 44.1. However, they were all web-based applications, which means that the information sharing was not done using dedicated mobile applications. Furthermore, the information on these websites was also outdated. In one particular instance, we found that the price of an item was last updated in 2011!

44.3 The LoPrice Model

Taking into consideration (1) the need to have a means to properly share the information on prices of items, (2) the widespread use of mobile devices, (3) the research opportunity on the use of crowdsourcing for non-profit making sectors

Table 44.1 Price comparison applications in Malaysia

Name	Website	Source of information	Coverage (items)	Coverage (stores)
Yellavia	http://www.yellavia.com/	Users	21 categories of items, groceries is one of them	Not specified
Ipengguna	http://www.ipengguna.com/	Administrator	Groceries	Not specified
PriceChecker.my	http://www.pricechecker.my/	Inactive – the site is not working		MYDIN, Carrefour, Tesco, Giant, Lazada

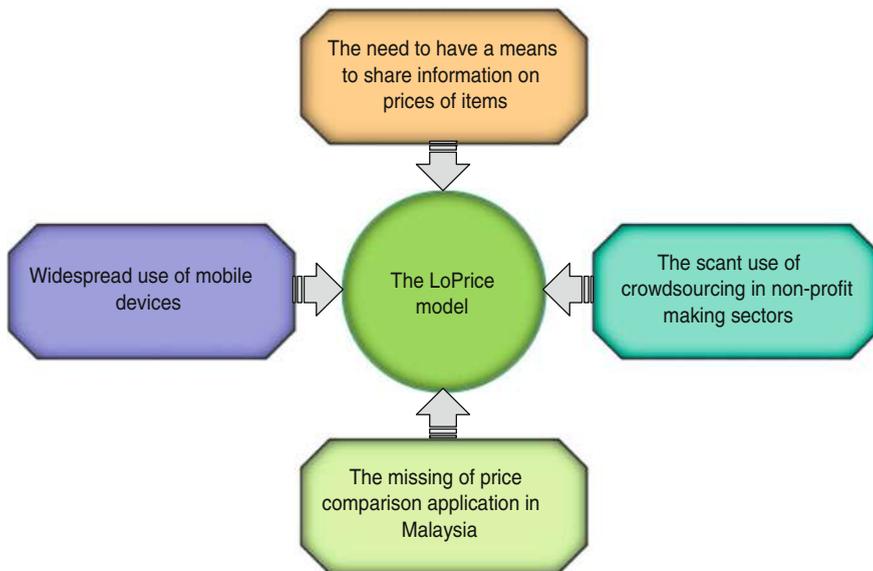


Fig. 44.2 Motivations for the research

that is still wide and (4) the missing of such applications in Malaysia, the LoPrice model is proposed with the four considerations above being the motivations towards the construction of the model as shown in Fig. 44.2.

The conceptual view of the LoPrice model is shown in Fig. 44.3. As can be seen from the figure, information comes from the public users through their mobile devices. Information for each item includes its price, location of the store where it can be bought from, its category, and the validity period of the stated price, if known. Application on the server side captures the information and makes it available to other users whenever requested. The existing crowdsourcing process are modified to suit the non-profit making purpose and subsequently applied in

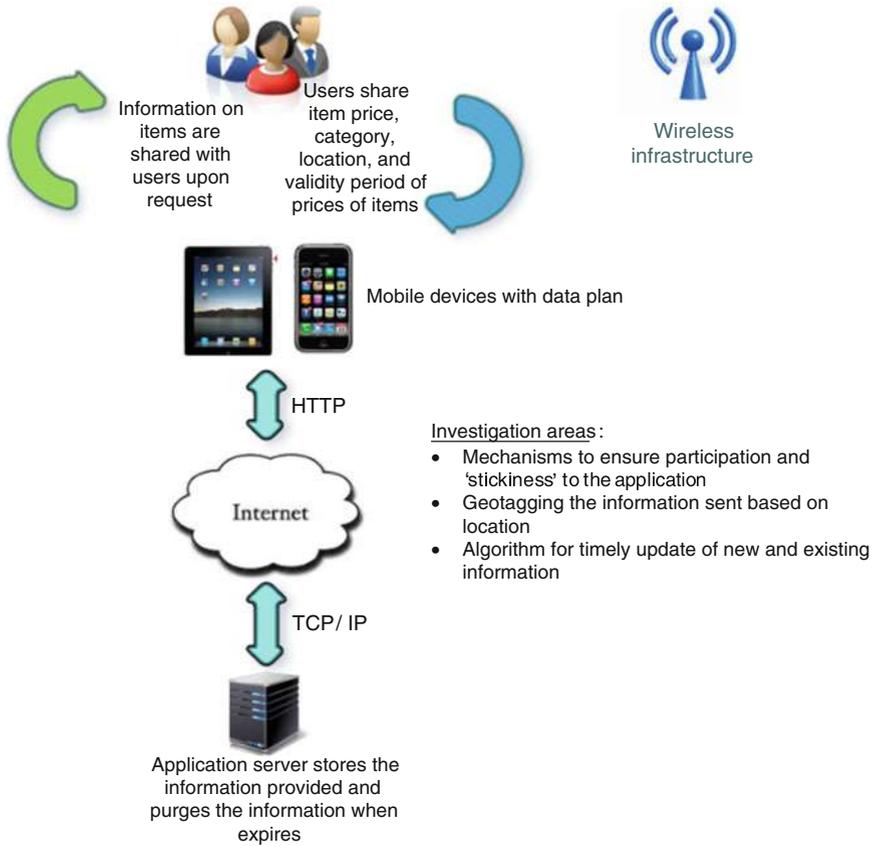


Fig. 44.3 Conceptual view of the LoPrice model and the areas of investigation

obtaining the information on prices of items from the public. The LoPrice model will be developed in line with the characteristics of social perspective based on the taxonomy of network communities proposed by Albors et al. [21]. To realise the LoPrice model, a price information sharing application will be developed, which can be used as a point of reference for users to find the location of the needed items at the best price.

44.3.1 Investigation Areas

As explained above, in the context of Malaysia, applications that can be used to compare prices of items were once developed and in use. However, their use could not be sustained over time. Information was not updated, hence, defeating their

intended purpose. The preliminary part of the construction of the LoPrice model is therefore a study that investigates the factors that influence users participation and their ‘stickiness’ to such application. Findings from this investigation are taken into consideration in constructing the LoPrice model. Secondly, to enable quick updates on items information, users are expected to provide as minimum information as possible. In other words, the LoPrice model should automate as much information as possible and one particular information that can potentially be automatically obtained is the location of the item, assuming that a user provides the information while he or she is still in the shop. For this reason, mechanism to geographically tag (geotag) the information sent based on location sent is investigated. Finally, an algorithm for timely update of new and existing information is also developed. The algorithm, among others, ensures that obsolete information is being purged from the application server. These areas of investigation are also shown in Fig. 44.3.

44.4 Issues and Challenges

Two main issues regarding crowdsourcing, which are the misuse of cheap labour and the reliability of information provided [22] are noted. With regard to cheap labour, it was argued that the crowds get rewarded and they are not forced to do the tasks. With regard to quality, Surowiecki [5] discovered that groups, albeit amateurs are often smarter than the smartest people in them, under the right circumstances. Also, a study reported on the use of crowdsourcing in transcribing showed 74 % accuracy, which is viable compared to 88.5 % with professional transcription [23]. Not to mention that measures to overcome the issues are already in the pipeline [22]. Nevertheless, at this point, we did not anticipate these issues to become hindrances to this research because no direct profit in monetary form will be generated. The benefit of using the LoPrice model will return to its users, hence they are not the labourers in any way. Since they are benefitting from the use of the model, we can account on them to provide accurate information.

44.5 Conclusion and Further Work

This paper provides justification on the need for the LoPrice model and briefly explains about the model. Judging from the success of its counterparts in other areas of social networking, it is expected that the application developed from the LoPrice model can quickly gain interest from the public. The expected benefit gained by the users is still in the form of monetary reward, albeit indirectly, through the reduced amount of expenditures. Future work includes the development of the prototype system to be used to evaluate the LoPrice model. At least

two types of evaluation will be performed; experimental and empirical. Experimental evaluation will assess the efficiency of the LoPrice model such as the timeliness of the information shared and the empirical results obtained will be quantitatively analysed using appropriate statistical analyses to determine its effectiveness. From the analysis results, the research contribution and the scope of applicability of the LoPrice model will be determined. Additionally, further work will also be made to look at the scalability of the LoPrice model, that is, the potential of extending its coverage beyond local community.

References

1. Department of Statistics Malaysia. http://www.statistics.gov.my/portal/index.php?option=com_content&view=article&id=767&Itemid=111&lang=en#2
2. Kenaikan harga barang keperluan isirumah. <http://jamal7478.blogspot.com/2011/08/kenaikan-kos-harga-barang-keperluan.html>
3. BR1M (Bantuan Rakyat 1Malaysia) Cash Aid Scheme Info. <http://www.br1m.info>
4. Enam Bidang Keberhasilan Utama Negara. <http://pmr.penerangan.gov.my/index.php/component/content/article/466-pengenalan-nkra/4808-pointers-6-bidang-keberhasilan-utama-negara-nkra.html>
5. Surowiecki, J.: *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, New York (2004)
6. Howe, J.: *Crowdsourcing: a definition*. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html
7. Zhang, L., Zhang, H.: Research of crowdsourcing model based on case study. In: *Proceedings of the 8th International Conference on Service Systems and Service Management*, pp. 1–5 (2011)
8. Liu, Y., Alexandrova, T., Nakajima, T., Lehdonvirta, V.: Mobile image search via local crowd: a user study. In: *Proceedings of the 17th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*, pp. 109–112, IEEE (2011)
9. Brabham, D.C.: Crowdsourcing as a model for problem solving an introduction and cases, in convergence. *Int. J. Res. New Media Technol.* **14**(1), 75–90 (Sage) (2008)
10. Armstrong, A.W., Harskamp, C.T., Cheeney, S., Wu J., Schupp, C.W.: Power of crowdsourcing: novel methods of data collection in psoriasis and psoriatic arthritis. *J. Am. Acad. Dermatol.* **67**(6), 1273–1281 (Elsevier) (2012)
11. Rovere, A., Raymo, M.E., O’Leary, M.J., Hearty, P.J.: Crowdsourcing in the quaternary sea level community: insights from the Pliocene. *Quaternary Sci. Rev.* **56**, 164–166 (Elsevier) (2012)
12. Heipke, C.: Crowdsourcing geospatial data. *ISPRS J. Photogrammetry Remote Sens.* **65**(6), 550–557 (Elsevier) (2010)
13. HuMuch. <http://www.humuch.com>
14. CatCrunch. <https://play.google.com/store/apps/details?id=com.saverr>
15. ShopSavvy. <http://shopsavvy.mobi>
16. Smoopo Price Checker. <http://techcrunch.com/2012/04/19/smoopa>, Smoopo
17. Price Comparison. http://www.appzoom.com/android_applications/shopping/price-comparison_bhexw.html
18. Grocery King Shopping List. <https://play.google.com/store/apps/details?id=com.groceryking&hl=en>
19. Grocery Price Watch. <https://itunes.apple.com/us/app/grocery-price-watch>

20. Leclerc. <http://shopperculture.integer.com/2011/08/index.html>
21. Albors, J., Ramos, J.C., Hervas, J.L.: New learning network paradigms: communities of objectives, crowdsourcing, wikis and open source. *Int. J. Inf. Manage.* **28**, 194–202 (2008) (Elsevier)
22. Hodson, H.: Crowdsourcing grows up as online workers unite. *NewScientist* **2903**, 22–23 (2013)
23. Aron, J.: Crowdsourcing serves up the subtitles to your life. *NewScientist* **2874**, 21 (2012)
24. Brabham, D.C.: Moving the crowd at threadless: motivations for participation in a crowdsourcing application. *Inf. Commun. Soc.* **13**(8), 1122–1145 (Taylor & Francis Online) (2010)
25. The crowdsourcing process in eight steps. http://en.wikipedia.org/wiki/File:Crowdsourcing_process2.jpg
26. Satzger, B., Psailer, H., Schall, D., Dustdar, S.: Auction-based crowdsourcing supporting skill management. *Inf. Syst.* **38**(4), 547–560 (2012)

Chapter 45

Enhancement of Nurse Scheduling Steps Using Particle Swarm Optimization

Norhayati Mohd Rasip, A.S.H. Basari, Nuzulha Khilwani Ibrahim and Burairah Hussin

Abstract Allocating of working schedule, especially for shift approach is hard to ensure its fairness among them. In the case of nurse scheduling, to set up the time table for each available nurse is time consuming and complicated, which consider many factors including rules, regulation and human factor. Moreover, most nurses are women, which have personnel constraints and maternity leave factors. The undesirable schedule can affect the nurse productivity, social life and the absenteeism can significantly as well affect the patient's life. This paper aimed to enhance the scheduling process by utilizing the particle swarm optimization in order to generate an intelligent nurse schedule. The result shows that the multiple initial schedules can be generated and can be selected with the lowest cost of constraint violation.

45.1 Introduction

Staff scheduling is the process of constructing work timetables encoding for staff in order to satisfy the demand of services. The ability to develop a good staff schedule is a crucial process if the services demand round-the-clock and complicated balancing act between an organization's need and the legal contractual

N. Mohd Rasip (✉) · A.S.H. Basari · N.K. Ibrahim · B. Hussin
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia
Melaka, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: mnrnorhayati@gmail.com

A.S.H. Basari
e-mail: abdsamad@utem.edu.my

N.K. Ibrahim
e-mail: nuzulha@utem.edu.my

B. Hussin
e-mail: burairah@utem.edu.my

obligations to its staff. Thus, all such factors should be considered when designing a staff shift schedule, leading to an extremely complex problem for which finding a better solution and within a reasonable time frame, can be difficult. Additional working shift can have a significant impact on the health and lives of staff while can in turn affect a person's productivity at work. Differences between industries, their goals and restrictions, often mean that specific models and algorithms must be developed for each of them [1].

This paper focuses on the hospital's staff and nurse is the proposed staff to be scheduled. Usually hospital wards must be staffed 24 h for 7 days a week by a limited number of staff, which very difficult to satisfy everybody. However, as well as the usual legal and cost constraints are met, high patient care services are the goal. Consequently, the risk is encountered when the nurse work on an undesirable schedule that may affect the patients' lives and also nurse performance since the job is beyond their expertise. Nevertheless, with shortages in qualified nurses reported regularly, good schedules are important to provide satisfactory patient care and potentially improve nurse retention.

Most hospitals face the nurse scheduling problem in its primary form, which is to allocate shifts and days to nurses so as to satisfy a given set of hard and soft constraints. The unique of combination of constraints depending on the pertinent policies that usually differ from one hospital to another. Once the number of resources defining the success of each schedule is determined, the problem may become complex, which is more constraints occur with feasible schedule to satisfy and more conflicting objectives are identified.

The class of nurse scheduling problem is well known by rapid growth in the number of potential solutions and the modest growth of the resources to be scheduled is known as NP-hard problem. Thus, the importance of finding the potential solution has led to research initiative in industries including airlines, call centers and hospital with a range of solution methods by utilizing such as mathematical programming, heuristic technique and artificial intelligence [2].

45.1.1 Related Work

The advance of information technology is rising to help researchers in developing a number of approaches to solve nurse rostering problem in a reasonable amount of time. As pinpointed in the varies depth literature reviews in [1, 3–5]. These four papers have provided details about problem formulation on scheduling. Basically, nurse scheduling formulated by using mathematical programming, artificial intelligence and meta-heuristics methods. In [6] is presented an integer mathematical programming (IP) to solve the nurse scheduling problem with two phase strategy.

The strategy is to effectively deal with the conflicting requirement based on nurses contracts and preferences. In [7] formulated mathematical programming with the multi objective to construct the schedule of medical resident based on seniority.

They aim to maximize the quality of shift allocation by considering the shift coverage requirements, seniority-based workload rules and resident work preferences. The multi objective model with sequential method and weighted method based on the seniority level has promised the optimal solution and less computational time.

One of the aims of nurse scheduling is to meet the balancing between schedule feasibility and quality. The quality of the timetable is normally measured as a maximize the fulfillment of soft constraints. In [7–9], dealing soft constraints with hierarchical oriented. These approaches are generally achieved the optimal solution based on individual preferences, but hard to solve real world problems that have high numbers of constraints and not relatively static over the time. Then the calculation of penalties is considered as pessimistic because the past penalty value is assumed that a roster for a week before is empty. Glass and Night proposed in [10] a continuous rostering environment for handling constraints and preferences arising. This solution approach is calculated the roster across a number of whole weeks, including the transition week from the current rostering period to the next. The result shows that the penalties are more optimistic.

Recently many researchers turn focus to population-based stochastic search methods such as genetic algorithms, ant colony optimization and particle swarm optimization (PSO), which able to reach near optimal solution within acceptable computation time. The PSO based algorithms have been successful designed and applied in many research area such as in university timetabling [11], job shop scheduling [12], manufacturing scheduling [13]. For nurse scheduling problem, PSO is utilized to solve unfair assignment among the nurses using a new evaluation function [14]. This new evaluation function allows a better discrimination between candidate solutions and is in accordance with the idea of the original function, which searched for assignment fairness. Wu et al. applied in [15] the basic PSO to the nurse scheduling problem. Improving work stretch pattern during initial solution using PSO had proved that using PSO produce optimal solution in a very efficient manner. The advantages of PSO are that this method not needed the calculation of derivatives due to the result of good solution is retained by all particles and that particle in the swarm share information between them. The objective function can be used for stochastic objective function and can escape from local minima. In order to be programmed, PSO has few parameters to adjust, more effectively memory capability than the GA which is every particle to remember their own previous best value as well as the neighborhood and more efficient in maintaining the diversity of the swarm [17].

The main contribution of this paper is to simplify the actual process of allocating shift to nurse by utilizing the advantages of the particle swarm algorithm with the same objective. Although there are plenty of PSO algorithms applied to the scheduling problems in the literature, to the best of our knowledge, applied to the nurse scheduling problem there is still significant room for improvements in this area [16]. This was one of the main motivations in order to design and apply a PSO based algorithm so as to solve effectively and efficiently the nurse scheduling problem with the real world problem case at Malaysia public educational hospital.

This paper is organized as follows: Sect. 45.2 discusses the nurse rostering problem while the particle swarm optimization approaches to the Nurse Rostering Problem is described in Sect. 45.3. Section 45.4 discusses the experimental results in conclusion and possible research directions are provided in Sect. 45.5.

45.2 Nurse Scheduling Problem

Nurse scheduling is the process of allocating, subject to constraints, the available resources into slot in a pattern within scheduling period. The process need a lot of time to produce a good roster. In this work, we consider a real nurse scheduling problem (NRP) one of the public hospital in Malaysia [17].

There have the four shifts a day that are morning shift (M), evening shift (E), night (N) or day off (O). For any holiday is included as days off. There have a two level of nurses either senior or junior. For each shift must be at least one senior.

The challenges of assigning process are considered subject to hard and soft constraints. All the constraint is presented in Sect. 45.2.1. The details step show as below:

- Step 1: Nurse request: Each nurse fill in the log book manually and submit their preference at least 1 day before the head nurse starts to construct the schedule. regarding the nurse's requests, the head nurse will consider some of the criteria before accept or reject the request.
- Step 2: After the nurse has the number of available nurses, the head nurse will construct the initial of schedule. In this stage the head nurse will allocate the night shift which is an unpopular shift until the number of nurses for night shift is satisfied. Then the head nurse continues to allocating the morning and evening shift (popular shift). This process is repeated until all the hard constraint and soft constraint is accepted.
- Step 3: The initial schedule will send to the department of nursing and after it is approved the schedule will print and distribute to all nurses and get their feedback.

The actual scheduling is "person based" process which is based on head nurse's knowledge and relies on her capabilities to create the best schedule for her unit. When assigning shift, the head nurse has to verify if all assignments fulfill to the quota requirement for each shift. The head nurse always changes the spreadsheet to find out whether it is respected or not, and memorize it while moving to the next step. This "searching process" is very time-consuming, not efficient and certainly not optimized. Figure 45.1 illustrates the mapping of the process.

The standard rules are when all the hard constraint is satisfied, then the feasible nurse scheduling is constructed, that can actually be used by the ward it was made for. Nevertheless, the number of soft constraint is satisfied the main factor that affects the quality of a nurse scheduling. The final goal, indeed, is to create feasible

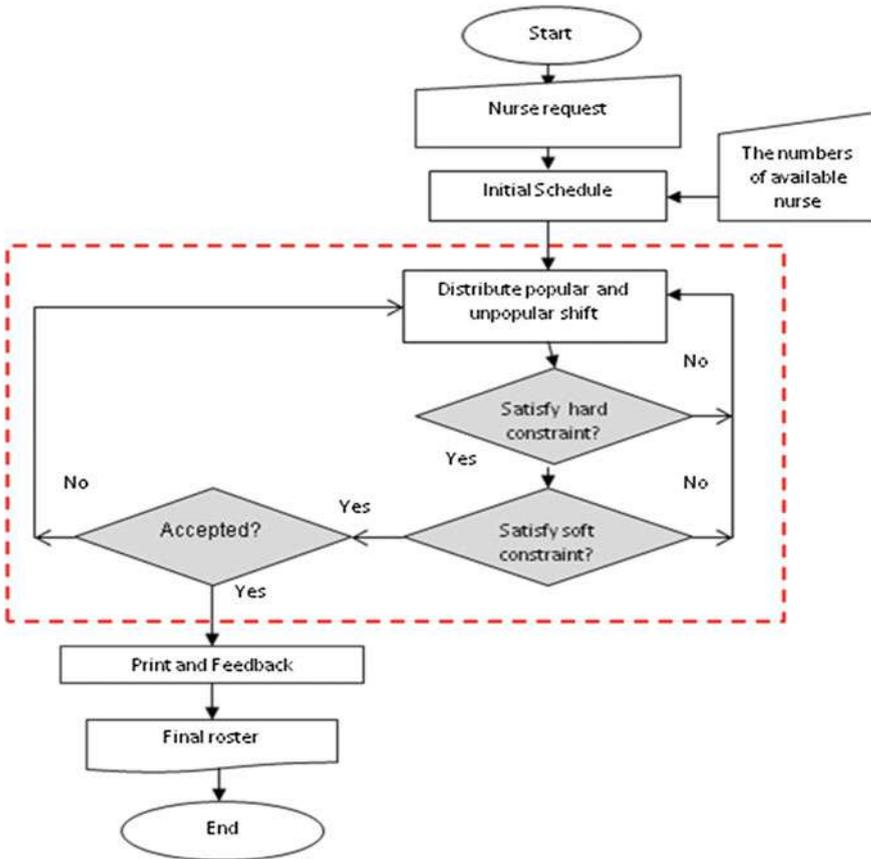


Fig. 45.1 Scheduling process map

schedule while maximizing its quality, means to create nurse scheduling which satisfied all the hard constraints and fulfilled as many as possible number of soft constraints.

45.2.1 Hard Constraints

In public hospital the rules are known as constraints. The hard constraint is the rules which must be fulfilled; otherwise the schedule is considered to be infeasible and unacceptable. The constraints are set based on current practices of the hospital studied, which taking into account for regulations, contractual requirements, operational requirements for all nurses. Most of these rules agree with published hospital policies.

The hard constraints detail as shown below:

- (H1) Each shift has to meet the minimum coverage demanded of nurses
- (H2) Each nurse cannot work more than one shift a day
- (H3) Each nurse must have at least 2 days off during the scheduling period
- (H4) Each shift must have at least one senior nurse
- (H5) Each nurse must not be scheduled for an isolated working day
- (H6) This constraint presents the minimum number of consecutive days that any nurse should work
- (H7) Each nurse must be working at least 10 days and not more than 12 days
- (H8) Each nurse must work in consecutive working days is 4 days
- (H9) Each nurse that has four consecutive night shifts must follow 2 days off.

45.2.2 Soft Constraints

The soft constraints taken into consideration which is not a compulsory to be fulfilled, but satisfying as much as possible represent the degree of a high quality result. There are other soft constraints that are normally created by an agreement between the head nurse and the individual nurses.

The hospital rules for soft constraints are as shown below:

- (S1) The total number of working days and day off should be uniformly distributed to all nurses.
- (S2) The nurse should have at least 1 day off in a weekend during the scheduling period.
- (S3) The nurse should have 1 day off after four consecutive days shift (morning and evening).

45.3 Particle Swarm Optimization for NSP

Particle swarm optimization (PSO) is a population-based optimization algorithm that was originally proposed by Kennedy and Eberhart (1995). PSO is an evolutionary optimization technique that is a population-based search algorithm. Each particle flies through the search space with velocities adjusted dynamically according to their previous behavior to look for the optimal solution. In PSO particles continuously adjust their position on their own and the experience of their neighboring particles. When compared with other evolutionary optimization techniques, PSO perform better in terms of success rate and solution quality. The movement of a particle is influenced by its inertia, its personal best position, and the global best position. Its velocity, its personal best value, that is the best

objective value the particle ever experienced, and its personal best position, that is the position at which the personal best value has been found.

The particle’s data could be anything. In the flocking birds, the data would be the X, Y, Z coordinates of each bird. The individual coordinates of each bird would try to move closer to the coordinates of the bird which is closer to the food’s coordinates (global best value).

To start the PSO approach, the algorithm starts with random initialization of a particle’s position and velocity. In this section, this paper describes all the components of the algorithm. The algorithm works to solve a problem with discrete domains, thus in order to apply PSO to this problem, a careful about the meaning of both the position and the velocity are very crucial.

45.3.1 Solution Representation

Step 1 Initialization: The assignment of shift to each nurse is randomly generated and for each nurse representing as one particle (solution). The position of particle represents the type of shift. The search space, represent as a matrix X of size N*D, where N represents the number of particles and D represents the total number for schedule period. Each particle X_i^t has a position, x_{id}^t is a discrete number represented the type of assignments (M, E, N, O) for each pair (nurse, day). If the $x_{id}^t = 1$ represent the morning shift (M), number 2 represents the evening shift (E), number 3 represents the night shift (N) and the number 4 represent the day off (O). If there are 5 nurses, N and the schedule period is 14 days, the particle position vector, then has the value of $5 \times 14 = 60$ elements. Elements 1–14 stand for the nurse schedule of the first nurse, 15–28 for the second nurse, 29–42 for the third nurse, and so on as shown in Table 45.1. First, each particle is initialized and the range of the initial position is set between 1 and 4.

Step 2 Evaluate the fitness value: All the particle’s fitness value is obtained by the fitness function. It checks all violations of hard and soft constraints in every particle. The fitness value for each particle is obtained by using weighted sum objective function MinWS using formula (45.1). Let $c \in C$ be the set of soft constraints then, w_c is the weight associated with constraint c and n_c the number of violations of c . The penalty weight as assigning by head nurse based on the importance rate as Table 45.2. The MinWS consists of two parts: C_{vio} associated with the costs of coverage nurse to a given shift and H_{vio} with the costs of assigning a nurse to a given shift. Thus the objective is to find a roster with the lowest overall penalty, denoted as MinWS.

$$C_{vio} = \sum_{0 \leq c \leq |C|} w_c n_c \tag{45.1}$$

Table 45.1 A sample matrix of initial position for PSO

Particle in the swarm	(1, 1)	(1, 2)	...	(1, 13)	(1, 14)	(2, 15)	(n, d)	...	(2, 28)	...	(N, N*D)
Particle 1	M	M	...	E	E	E	O	...	M	...	E
:	M	M	...	M	E	E	M	...	P	...	E
Particle 5	E	O	...	M	M	M	N	...	O	...	M
:	E	E	...	M	M	M	M	...	E	...	M
Particle i	M	M	...	O	M	N	M	...	M	...	M
:	N	N	...	E	E	N	E	...	M	...	O
Particle N	O	N	...	E	N	O	E	...	E	...	E

Table 45.2 The constraint weighting

Soft constraints	Penalty
Each shift has to meet the minimum coverage demanded of nurses	100
Assign morning shift following night shift	100
Assign each nurse at least 1 day off on the weekends	100
Assigns four consecutive morning shifts followed by 1 day off (O) or an evening (E) shift after 2 days off that follow the night (N) shift pattern	10
Assign four consecutive evening shifts followed by 1 day off	10
Assigns either a day off (O) or an evening (E) shift after 2 days off that follow the night (N) shift pattern, i.e., (NNNNOOE) or (NNNNOOO)	10

$$\text{MinWS} = \sum_{0 \leq c \leq |C|} C_{vio} + H_{vio} \tag{45.2}$$

Step 3 Updating particle’s position: After the particle position is initialized, each particle is assigned a rate of position change or velocity randomly. Just in the way the particles adjust its flight direction through referring to its own experience and population’s experience. The particles update the velocity and position according to the following two equations:

$$V_i^{t+1} = W \times V_i^t + c1 \times r1 \times (P_{i_best} - x_i^t) + c2 \times r2 \times (G_{best} - x_i^t) \tag{45.3}$$

$$x_i^{t+1} = x_i^t + V_i^{t+1} \tag{45.4}$$

where X_i^t represents the position of the particles, V_i^t represents the velocity of the particles, w represents the inertia weight, which is the inertia weight value setting will have an effect on the velocity of particle position update and also indirectly on the next position the particle will move to. If the w value setting is inappropriate, it is impossible to explore unknown areas and the search for the personal best and global best will also be affected. The $c1$ and $c2$ are the learning rates, which is a positive constant parameter called acceleration coefficients (which control the maximum step size of a particle) and how to approach the new position either for close to the individual experience position or expecting close to the global experience position. The $r1$ and $r2$ are two random numbers independently generated, which are between 0 and 1. P_{i_best} is a particle’s position at which the best fitness so far has been achieved. G_{best} is the population global (or local neighborhood, in the neighborhood version of the algorithm) position at which the best fitness so far has been achieved. The first part of Eq. (45.1) is the “inertia”, determined by a particle’s previous velocity; the second part is the “individual cognition”, represents a particle’s own experience, which guides a particle to evolve into its “ P_{i_best} ”; the third part is the “social cognition”, denotes the cooperation and information sharing among the particles, which guides a particle

Table 45.3 Result of experiment

Particle in the swarm	MinP	MaxP	Avg. hard constraint	Avg. soft constraint	Fitness value (MaxP–MinP)
Particle 1	100	400	180	240	420
Particle 2	100	700	260	370	630
Particle 3	10	600	170	157	327
Particle 4	10	310	180	98	278
Particle 5	110	510	130	303	433

to evolve into the population's " G_{best} ". Just in this way, each particle adjusts its flight (evolution) direction through referring to its own experience and the population's experience.

45.4 Preliminary Results

In this study, the proposed PSO algorithm based technique has been developed in order to make it suitable for solving nurse constrained optimization problems. A procedure is imposed to check the feasibility of the candidate solutions in all stages of the search process. The proposed technique has been implemented on personal computer a Intel Core i5 CPU and 4 GB RAM. To complete the following experiments, the PSO-based intelligent mechanism for scheduling is programmed by using MATLAB. The computational flowchart of the proposed PSO algorithm is depicted in Fig. 45.1.

For this stage the result shows the fitness value of each schedule generated (particles) with hard and soft constraints. The advantages are by using intelligent nature inspired optimization; the head nurse can generate more than one schedule for the initial schedule and can select the best schedule which has the best fitness value. The guided randomized assignment of shift to available nurse can reduce time consuming and be more flexible to use because the user can change the few parameters based on their need. Table 45.3 shows the preliminary result of the experiment.

45.5 Conclusion

The result shows that the effectiveness and efficiency of constructing the nurse schedule can be achieved by using particle swarm optimization. The advantages of particle swarm optimization which are the setting of parameters is reasonably simple and implementation time to find a proper schedule is fast can be able to use for all the Malaysian Hospital. The finding suggests that the proposed solution using particle swarm optimization is capable to enhance the nurse scheduling by utilizing the intelligent process in order to obtain the intelligent nurse schedule.

For further research could be extended with other scheduling scenario and variant particle swarm optimization.

Acknowledgments This paper is part of Master by research in Information and Communication Technology and funded under the Ministry of Education research grant RAGS/2012/FTMK/TK06/1 B00018.

References

1. Ernst, A., Jiang, H., Krishnamoorthy, M., Sier, D.: Staff scheduling and rostering: a review of applications, methods and models. *Eur. J. Oper. Res.* **153**(1), 3–27 (2004)
2. Hussin, B., Basari, A.S.H., Shibghatullah, A.S., Asmai, S.A., Othman, N.S.: Exam timetabling using graph colouring approach. In: *Proceedings of the 2011 IEEE Conference on Open Systems*, pp. 133–138, Sept. 2011
3. Brucker, P., Qu, R., Burke, E.: Personnel scheduling: models and complexity. *Eur. J. Oper. Res.* **210**(3), 467–473 (2011)
4. Burke, E.K., de Causmaecker, P., Berghe, G.V., van Landeghem, H.: The state of the art of nurse rostering. *J. Sched.* **7**(6), 441–499 (2004)
5. Cheang, B., Li, H., Lim, A., Rodrigues, B.: Nurse rostering problems—a bibliographic survey. *Eur. J. Oper. Res.* **151**(3), 447–460 (2003)
6. Valouxis, C., Gogos, C., Goulas, G., Alefragis, P., Housos, E.: A systematic two phase approach for the nurse rostering problem. *Eur. J. Oper. Res.* **219**(2), 425–433 (2012)
7. Topaloglu, S.: A shift scheduling model for employees with different seniority levels and an application in healthcare. *Eur. J. Oper. Res.* **198**(3), 943–957 (2009)
8. Ziarati, K., Akbari, R., Zeighami, V.: On the performance of bee algorithms for resource-constrained project scheduling problem. *Appl. Soft. Comput.* **11**(4), 3720–3733 (2011)
9. Brucker, P., Burke, E.K., Curtois, T., Qu, R., Berghe, G.V.: A shift sequence based approach for nurse scheduling and a new benchmark dataset. *J. Heuristics* **16**(4), 559–573 (2008)
10. Glass, C.A., Knight, R.A.: The nurse rostering problem: a critical appraisal of the problem structure. *Eur. J. Oper. Res.* **202**(2), 379–389 (2010)
11. Tassopoulos, I.X., Beligiannis, G.N.: A hybrid particle swarm optimization based algorithm for high school timetabling problems. *Appl. Soft. Comput.* **12**(11), 3472–3489 (2012)
12. Ping, Y., Minghai, J.: An improved PSO search method for the job shop scheduling problem. In: *Proceedings of the 2011 Chinese Control Decision Conference CCDC*, vol. 2, pp. 1619–1623 (2011)
13. Chou, F.-D.: Particle swarm optimization with cocktail decoding method for hybrid flow shop scheduling problems with multiprocessor tasks. *Int. J. Prod. Econ.* **141**(1), 137–145 (2013)
14. Altamirano, L., Riff, M.-C., Trilling, L.: A PSO algorithm to solve a real anaesthesiology nurse scheduling problem. In: *Proceedings of the 2010 International Conference of Soft Computing and Pattern Recognition*, pp. 139–144 (2010)
15. Wu, T., Yeh, J., Lee, Y.: A particle swarm optimization approach for nurse rostering problem. In: *Proceedings of the International Conference on Business and Information*, pp. 737–752 (2013)
16. Rasip, N.M., Ibrahim, N.K., Basari, A.S.H.: An investigation of intelligent search techniques for nurse scheduling improvement in healthcare organization. In: *Proceedings of the EHealth Symposium 2013*, pp. 1–7 (2013)
17. Abobaker, R., Ayob, M., Hadwan, M.: Greedy constructive heuristic and local search algorithm for solving nurse rostering problems. In: *Proceedings of the 2011 3rd Conference on Data Mining and Optimization*, pp. 28–29, June 2011

Chapter 46

Hardware Implementation of MFCC-Based Feature Extraction for Speaker Recognition

P. Ehkan, F.F. Zakaria, M.N.M. Warip, Z. Sauli and M. Elshaikh

Abstract The most important issues in the field of speech recognition and representative of the speech is a feature extraction. Feature extraction based Mel Frequency Cepstral Coefficient (MFCC) is one the most important features required among various kinds of speech application. In this paper, FPGA-based for speech features extraction MFCC algorithm is proposed. The complexities of computational as well as the requirement of memory usage are characterized, analyzed, and improved. Look-up table (LUT) scheme is used to deal with the elementary function value in the MFCC algorithm and fixed-point arithmetic is implemented to reduce the cost under accuracy study. The final feature extraction design is implemented effectively into the FPGA-Xilinx Virtex2 XC2V6000 FF1157-4 chip.

Keywords Speaker recognition · Mel frequency cepstral coefficients · Field programmable gate array

P. Ehkan (✉) · F.F. Zakaria · M.N.M. Warip · M. Elshaikh
School of Computer and Communication Engineering, Universiti Malaysia Perlis,
Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: phaklen@unimap.edu.my

F.F. Zakaria
e-mail: ffaiz@unimap.edu.my

M.N.M. Warip
e-mail: nazriwarip@unimap.edu.my

M. Elshaikh
e-mail: elshaikh@unimap.edu.my

Z. Sauli
School of Microelectronic Engineering, Universiti Malaysia Perlis, Pauh Putra Campus,
02600 Arau, Perlis, Malaysia
e-mail: zaliman@unimap.edu.my

46.1 Introduction

Biometric systems are the automated method of verifying or recognizing the identity of the person on the basis of some physiological characteristic, such as a finger print, face pattern and human voice [1]. The human voice conveys information about the language being spoken and the emotion, gender and identity of the speaker. Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices [2, 3]. Voice of a person has many prominent characteristics like pitch, tone which can be used to distinguish a person from the other. It has a history dating back some four decades [4] where the output of several analog filters was averaged over time for matching. Speaker recognition uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (size and shape of the throat and mouth) and learned behavioral patterns such as voice pitch and speaking style. This incorporation of learned patterns into the voice templates has earned speaker recognition its classification as a "behavioral biometric" [5]. Such systems extract features from speech, model them and use them to recognize the human voice.

Speaker recognition system basically involves two main phases namely the training stage and the testing stage [6]. These phases involve two main parts called feature extraction and pattern classification. At the time of training, speech sample is acquired in a controlled and supervised manner from the user. The speaker recognition system has to process the speech signal in order to extract speaker discriminatory information from it and this will form the speaker model, which is a process of enrolment speaker data. At the time of testing a speech sample is acquired from the user. The speaker recognition system has to extract the features from this sample and compare it against the models been stored beforehand. This is a pattern matching or classification task.

The improvement in FPGA technology as well as design tools has introduced a new option for Digital Signal Processing (DSP) applications recently. FPGA is a type of semiconductor device that contain programmable logic and interconnections which mostly used in logic or digital electronic circuits. The programmable logic components or logic blocks as they are known may consist of anything from logic gates, through to memory elements or blocks of memories, or almost any element. FPGA supports thousands of gates and popular for prototyping integrated circuit (IC) designs. Once a design is set, hardwired chips will be produced to faster performance. FPGA chip is programmable and reprogrammable which is considered as an advantage of it. In this way, it becomes a large logic circuit that can be configured according to a design, but if changes are required it can be reprogrammed with an update. Thus, if circuit board is manufactured and contains an FPGA as part of the circuit, this is programmed during the manufacturing process, but can be reprogrammed to reflect any changes. The user programmability gives the user access to complex ICs without the high engineering costs associated with ASICs.

FPGA contains many identical logic cells that can be viewed as standard components. Each design is implemented by specifying the simple logic function for each cell and selectivity closing the switches in the interconnect matrix. The array logic cells and interconnects form a basic building blocks for logic circuits. Complex designs are created by combining these basic blocks to create the desired circuit. The logic cell architecture varies between different device families. This logic cells can be configured via high levels design tools or hardware description languages (HDLs) to implement virtually any digital system. Few years ago these chips required programming using HDLs or weak design tools, making implementation of complex DSP operations, such as FFT and DCT very tedious. The newest generation of design tools offers libraries of common DSP functions, enabling developers to implement something as complex as FFT [7]. FPGA have been used in many areas to accelerate algorithms that can make use of massive parallelism, and in every case allowing them to improve the flexibility, cost reduction and time to market. A great promising application area the use of FPGA is that it can exploit the pipeline ability and parallelism with the solution algorithms in a much more thorough way that can be done with parallel computers using general-purpose microprocessors or a single standard processor.

This paper presents results for the speaker recognition front-end processing using MFCC approach on the platform consisting of a Xilinx Virtex-II XC2V6000 FPGA. The paper is organized the sections as the speaker recognition system, front-end processing, implementations, results and conclusion.

46.2 Speaker Recognition System

A block diagram shown in Fig. 46.1 is the basic structure system designed to implement speaker recognition. The input speech is sampled and converted into digital format. The vectors from the speech consisting of the MFCCs are extracted. The system then branches into two separate phases namely training and classification. In the training phase, each registered speaker has to provide samples of their speech so that the system can train a reference models for that speaker while in the classification phase, the input speech is matched with the stored reference models and then a recognition decision is made. The speaker associated with the most likely, or higher scoring model is selected as the recognized speaker. This is simply a maximum likelihood classifier.

The more general problem for speaker recognition is that out of a total population of N speakers, find that speaker whose reference pattern is most similar to the sample pattern of an unknown speaker. Since the sample pattern is compared to each of the N reference patterns and also there is a finite probability of an incorrect decision for each comparison, it is apparent that the overall probability of an incorrect decision must be a monotonically increasing function of N [8].

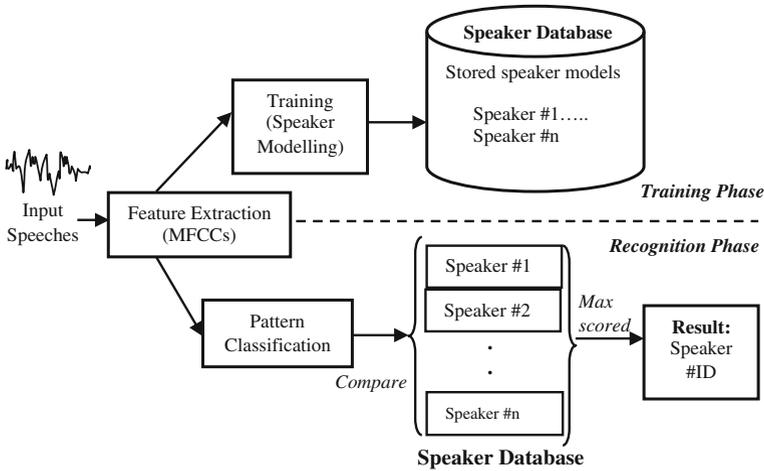


Fig. 46.1 Basic structure of speaker recognition system

46.3 Front-End Processing

This part is the most important issue in the field of speech processing. It is converting the speech waveform to a set of features for further analysis. Speech front-end processing consists of transforming the speech signal to a set of feature vectors [9]. The aim of this process is to obtain a new representation which is more compact, less redundant, and more suitable for statistical modelling [10]. Feature extraction is the key to front-end process.

According to Stolcke et al. [11], the purpose of feature extraction is to convert the speech waveform to some type of parametric representation at a considerably lower information rate. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as LPC [12], MFCC [13], PLP [14] and others. The MFCC and PLP are among the most popular acoustic features used in speaker recognition. It usually depends on the task which of the two methods leads to a better performance. Currently, researchers are focusing on improving these two cepstral features [10] or appending new features on them [15]. In fact, it is commonly believed that the spectrum smoothing done by MFCC and PLP has some sort of speaker normalization effect.

The selection of feature determines the separability of the speaker, and it also has large influence on the classification step, since the classifier must be turned to the given feature space. Thus, the selection of the features should be carefully considered in designing a system. Several analyses have been done for feature extraction technique in order to observe the best technique for transforming the speech signal. Davis and Mermelstein [16] reviewed the literature of a few feature extraction methods and compared them in a syllable-oriented speaker dependent speech recognition system. The experiments were made in a clean environment

and the segmentation was done manually. They found out that features derived using cepstrum analysis outperformed those that do not use it and that filter bank methods outperformed LPC method (PLP methods were not included). This means that the best performance was achieved by using MFCC.

A comparable result of the use of PLP and MFCC was reported by John and Wendy [17] that MFCC-based feature extraction method seems to be performing well in most studies. Besides, a theoretical comparison of MFCC and PLP analysis was given by Milner [18]. The theoretical comparison continues with a practical implementation. The spectral analysis is followed by channel normalization (both RASTA and CMN) and extraction of dynamic features. The best results were reported for MFCC with RASTA filtration. Similarly, Schmidt and Thomas [19] has pointed out the state-of-the-art speaker recognition systems typically employing the MFCC as representative acoustic feature and GMM as pattern classification method has achieved very good performance which even better than recognition by human. Chakroborty et al. [20] listed three reasons why MFCC method has become so dominant for speaker identification system. First, MFCC is less vulnerable to noise perturbation, it gives little session variability and is easier to extract. Moreover, a calculation of MFCC is based on the human auditory system aiming for artificial implementation of the ear physiology assuming that the human ear can be a good speaker recognizer too. Furthermore, computation of MFCC involves averaging the low frequency region of the energy spectrum (approximately demarcated by the upper limit of 1 kHz) by closely spaced overlapping triangular filters while smaller number of less closely spaced filters with similar shape are used to average the high frequency region. Thus, MFCC can represent the low frequency region more accurately than the high frequency region and hence, it can capture formants which lie in the low frequency range and which characterize the vocal tract resonances [21, 22].

All these facts suggest that any speaker recognition system based on MFCC can possibly be improved. Based on above studied, this project has been decided to employ MFCC method as the feature extraction part in order to obtain most advantageous feature vector to be adapted in the propose of hardware implemented pattern classification model.

46.3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC)s are well known features used to describe speech signal. They are based on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. The processes to obtain the MFCC can be summarized as in Fig. 46.2.

The speech input is typically recorded at a sampling rate above a frequency of 10 kHz. This sampling frequency was chosen to minimize the effects of aliasing in

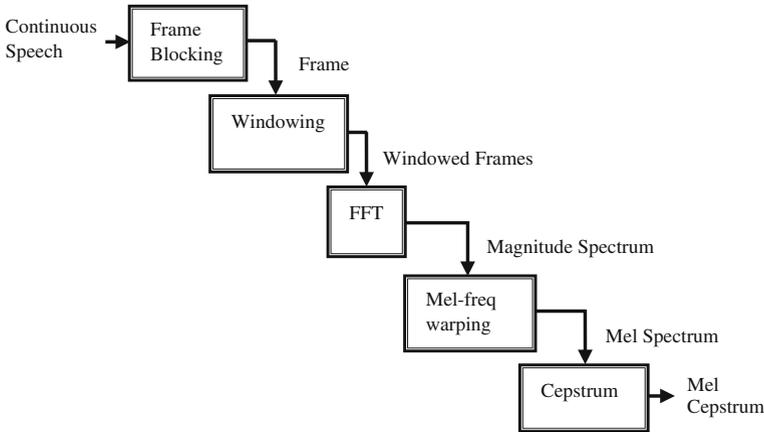


Fig. 46.2 MFCC block module

the analogue-to-digital (ADC) conversion. These sampled signals are able to capture all frequencies up to 5 kHz which cover most energy of sounds that are generated by humans. The main purpose of the MFCC processor is to mimic the behavior of human ears. Besides, rather than the speech waveforms themselves, MFCCs are shown to be less susceptible to the mentioned variations.

Frame Blocking. The continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M ($M < N$). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by $(N-M)$ samples as shown in Fig. 46.3. Similarly, the third frame begins $2M$ samples after the first frame (or M samples after the second frame) and overlaps it by $(N-2M)$ samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are $N = 256$ (which is equivalent to ~ 30 ms windowing and facilitate the fast radix-2 FFT) and $M = 100$, respectively.

Windowing. This step is windowing each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If the window is defined (where N is the number of samples in each frame) then the result of windowing is the signal of

$$y_1(n) = x_1(n)w(n), \quad 0 \leq n \leq N - 1. \quad (46.1)$$

The Hamming window is typically used in a form of

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N - 1 \quad (46.2)$$

where, N is the samples number of each window and n is the sample being evaluated.

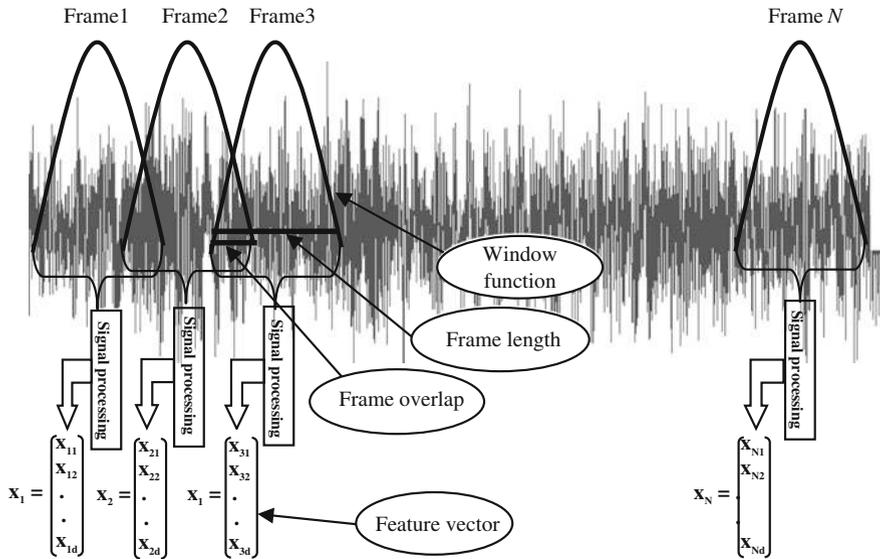


Fig. 46.3 Frame blocking short term analysis

Fast Fourier Transform (FFT). The FFT is a fast algorithm to implement the DFT converts each frame of N samples from time domain into frequency domain. It can be defined on the set of N samples $\{x_n\}$ as shown

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0, 1, 2, \dots, N - 1. \tag{46.3}$$

Generally, the values of X_k are complex numbers and in this case the frequency magnitudes which are their absolute values are considered.

Mel-Frequency Wrapping. Psychophysical studies have shown that the human perception of the frequency consists of sounds for speech signal does not follow the linear scale. Therefore, for each tone with an actual frequency, f , measured in Hertz, a subjective pitch is measured on the scale called the ‘Mel’ scale. The frequency scale is then converted from Hertz to Mel. The Mel-frequency scale is linear frequency spacing below 1 kHz and logarithm spacing above 1 kHz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1,000 mels. Therefore, it is possible to use the following approximate formula to compute the mels for a given frequency f in Hz:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f / 700). \tag{46.4}$$

Cepstrum. The final step is converting the log mel spectrum back to time domain. The result is called the MFCC. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Mel spectrum coefficients are real numbers. Therefore, is possible to convert them to the time domain using the DCT.

$$C_n = \sum_{k=1}^K (\log S_k) \cos[n(k - 1/2)(\pi/K)], \quad n = 1, 2, \dots, K \quad (46.5)$$

where, S_k is the mel scaled signal had after wrapping and C_n is the cepstral coefficient.

46.4 Implementations

The front-end processing features extraction MFCC-based speaker recognition system was implemented into Xilinx-based FPGA Virtex-II XC2V6000 FF1152-4 platform. This device is a mid-range FPGA, member of Virtex-II family has 76,032 logic cells, 2592 k bits block RAM, 144 18 bit multipliers, 824 user I/O and a speed grade of 4. The specifications for the system implemented in the hardware used the first 17 MFCCs and their respective delta values, population size of 20 and 5 s of test utterances. The speech samples were taken from the TIMIT database.

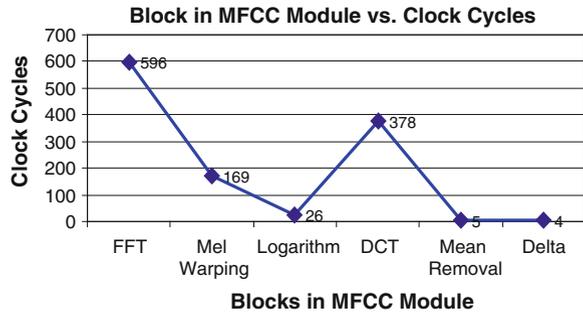
46.5 Results

Hardware Resources Requirement. The breakdowns of hardware utilization for MFCC module are presented in Table 46.1.

Table 46.1 Hardware resources usage

Item	Used	Available	% Utilization
Slices	23251	33792	68.8
Slice FFs	35104	67584	51.9
4 Input LUTs	39452	67584	58.4
Bonded IOBs	295	824	35.8
BRAM	14	144	9.7
MULT18 × 18	33	144	22.9
GCLKs	1	16	6.26

Fig. 46.4 Datapath Breakdown



Simulation Results. The entire datapath taken by the MFCC module is 1,178 clock cycles. Figure 46.4 shows the breakdown of clock cycles used by each block of MFCC module.

The FFT block is the most complexity sub-block in MFCC module and it required the most clock cycles which is 596 followed by the DCT block which occupied 378 clock cycles.

46.6 Conclusion

As the conclusion this paper aims to highlight the implementation of speech signal MFCC features extraction algorithm on hardware-based FPGA. The hardware solution has taken the advantage of parallelism and pipelining throughout the entire signal processing compared to conventional approach which is mostly software-based. The FPGA-based hardware solution can be an optimal and alternative solution to the conventional DSPs for the speech signal processing technology.

References

1. Kung, S.Y., Mak, M.W., Lin, S.H.: Biometric Authentication: a Machine Learning Approach, 1st edn. Prentice Hall, New Jersey, USA (2005)
2. Campbell, J.P.: Speaker recognition: a tutorial. *Proc. IEEE* **85**(9), 1437–1462 (1997)
3. Sadaoki, F.: Fifty years of progress in speech and speaker recognition. *J. Acoust. Soc. Am* **116**(4), 2497–2498 (2004)
4. Atal, B.S.: Automatic recognition of speakers from their voices. *Proc. IEEE* **64**, 460–475 (1976)
5. Furui, S.: An overview of speaker recognition technology, ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, 1–9 (1994)
6. Richard, D.P., Daryl, H.G.: An introduction to speech and speaker recognition. *IEEE Comput. Soc. Press* **23**(8), 26–33 (1990)

7. "System Generator for DSP" Xilinx Inc. (2006). http://www.xilinx.com/ise/optional_prod/system_generator.htm
8. Rosenberg, A.E., Soong, F.K.: Recent research in automatic speaker recognition. In: Sadaoki, F. (ed.) *Advances in Speech Signal Processing*, 701–738 (1992)
9. Moretto, P.: Mapping of speech front-end signal processing to high performance vector architectures. Technical report, International Computer Science Institute (1995)
10. Premakanthan, P., and Mikhad, W. B.: Speaker verification/recognition and the importance of selective feature extraction: review. In: *Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems*, 1(1), 57–61 (2001)
11. Stolcke, A., Shriberg, E., Ferrer, L., Kajarekar, S., Sonmez, K., and Tur, G.: Speech recognition as feature extraction for speaker recognition. In: *IEEE Workshop on Signal Processing Applications for Public Security and Forensics 1–5* (2007)
12. Atal, B.S., Hanauer, L.S.: Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.* **50**, 637–655 (1971)
13. Davis, S. B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust Speech Sig. Process.* **28**(4), 357–366 (1980)
14. Hermansky, H.: Perceptual Linear Predictive Analysis of Speech. *J. Acoust. Soc. Am.* **87**(4), 1738–1752 (1990)
15. Waleed, H.A.: Robust speaker modeling using perceptually motivated feature. *Elsevier Sci. Pattern Recogn. Lett.* **28**(11), 1333–1342 (2007)
16. Davis, S. B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Sig. Process.* **28**(4), 357–366 (1980)
17. John, H., Wendy, H.: *Speech Synthesis and Recognition*, 2nd edn. Taylor & Francis Inc, Bristol, USA (2002)
18. Milner, B.: A Comparison of Front-End Configurations for Robust Speech Recognition. *Proceeding of ICASSP '2002*, 1(1), 797–800 (2002)
19. Schmidt, N.A., Thomas, H.C.: Speaker verification by human listeners: experiments comparing human and machine performance using the NIST1998 speaker evaluation data. *J. Digit. Sig. Process.* **10**(1–3), 249–266 (2000)
20. Chakraborty, S., Roy, A., Saha, G.: Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. *Int. J. Sig. Process.* **4**(2), 114–122 (2008)
21. Rabiner, L., Juang, B.H.: *Fundamentals of Speech Recognition*, 2nd. ed. Pearson Education, USA (2003)
22. Ben, G., Nelson, M.: *Speech and Audio Signal Processing*, 2nd edn. Wiley, USA (2002)

Chapter 47

Parallel ASIP Based Design of Turbo Decoder

F.F. Zakaria, P. Ehkan, M.N.M. Warip and M. Elshaikh

Abstract Application Specific Instruction-set Processor (ASIP) has a general-purpose architecture that can be modified and used in a variety of applications. However, this increases the power and memory utilization and affects the functionality and efficiency of ASIP. This paper is defining the flexibility of ASIP for Turbo decoding in term of its functionality and architecture for specific applications such as DVB-RCS, 3GPP. The proposed architecture has a dedicated SIMD (Single Instruction Set Multiple Data), coupled with distributed memory based ASIP. It has been concluded in this paper that ASIP facilitates parallelism at different levels, thereby, increasing the efficiency, power consumption, and processing time.

Keywords Application-specific instruction-set processor · Bahl-Cocke-Jelinek-Raviv (BCJR) · Turbo decoding recognition

47.1 Introduction

The technology development in field of telecommunication system nowadays had changed it's from analog systems to digital systems exclusively. One of the major reasons is digital communications offer possibility of channel coding (line, and

F.F. Zakaria (✉) · P. Ehkan · M.N.M. Warip · M. Elshaikh
School of Computer and Communication Engineering, Universiti Malaysia Perlis,
Pauh Putra Campus, 02600 Arau, Perlis, Malaysia
e-mail: ffaiz@unimap.edu.my

P. Ehkan
e-mail: phaklen@unimap.edu.my

M.N.M. Warip
e-mail: nazriwarip@unimap.edu.my

M. Elshaikh
e-mail: elshaikh@unimap.edu.my

error control, coding) to minimize effects of noise and interference. Since the introduction of Turbo Codes in 1993 by Berrou and Glavieux [1], several standards and protocols had been adapted to it as it serves astonishing performance. Over the years it has seen the emergence of an increasing number of wireless protocols such as UMTS, LTE and WiMAX. The growth of these protocols is to meet the technologies needs of now-days hand-held wireless devices, which convergence with many disparate features, including wireless communication, real-time multimedia and interactive applications, into a single platform. Thus the flexibility of a decoder becomes a dominant aspect for each wireless device. Today's 3GPP standard limits the throughput to 2 Mbit/s, while future systems are predicted to require above 10 Mbit/s. Scalable and energy efficient architecture are required to catch up with the challenge and to serves time to market constrains. Therefore, insisting on a programmable solution, ASIPs are becoming the architecture of the first choice.

Before the flexibility been taken into account in implementation of turbo decoder, ASIC had been exclusively the choice of turbo decoder implementer. Some of these implementations succeeded in archiving high throughput for specific standards with a fully dedicated architecture for example, in [2] a dedicated high performance turbo decoding for 3GPP implemented on ASIC, in [3], a new class of turbo codes more suitable for high throughput implementation is proposed. However, such implementations do not consider the flexibility issues and implementation of ASICs becomes extremely expensive because their high cost and low design productivity. Only high manufacturing volumes can allow manufacturing. Application specific Instruction-set Processor (ASIPs) solve a part of these problems [4], a higher volume of unit is demanded because there are more applications that fit on it.

This paper presents the architecture of ASIP models for turbo decoding to show their flexibility and performance. The rest of the paper is organized as Sect. 47.2 presents the turbo decoding algorithm and the parallelism technique, Sect. 47.3 presents a decoder requirement regarding the different standards and channel codes, Sect. 47.4 presents the architecture of ASIPs Turbo decoding, Sect. 47.5 results of the proposed architecture has been shown and lastly Sect. 47.6 summarizes the result obtained to conclude the paper.

47.2 Turbo Decoding and Parallelism

Turbo code is an efficient error correction code, using a concatenation of two (or more) simple constituent codes, which are iteratively decoded using soft decision. This work implements Max Log MAP algorithm for the application.

47.2.1 Maximum Log Maximum-a-Posteriori Algorithm

The max log-MAP algorithm is an approximation of the MAP algorithm that operates in the log-domain, allowing multiplications in the original MAP algorithms to be implemented by additions. The max log-MAP algorithm, with a pair of Viterbi algorithm, sweeps through the trellis in forward and reverse direction, hence, the MAP and the logarithmic variants are called the forward-backward algorithm. The branch metric of max log-MAP algorithm with the addition of a priori *log-likelihood ratio* (LLR) is same as that branch metric that is used in Viterbi algorithm.

If σ_k is the trellis state value at time k , then the likelihood values (L_k) at time k , is defined as

$$L_k = \alpha_{k-1}(\sigma_{k-1}) + \gamma_k(\sigma_{k-1}, \sigma_k) + \beta_k(\sigma_k) \quad (47.1)$$

where $\alpha_{k-1}(\sigma_{k-1})$ is the *alpha* metric that calculates the probability of the current state based on the input values before time k , $\gamma_k(\sigma_{k-1}, \sigma_k)$ is the branch metric that calculates the probability of the current state transition, $\beta_k(\sigma_k)$ is the *beta* metric that calculates the probability of the current state given the future state values after time k . Both α and β are called the state metrics and γ is the branch metric. Furthermore the α and β computations are the forward and backward recursive, respectively. They are given as

$$\alpha_k(\sigma_k) = \max(\alpha_{k-1}(\sigma_{k-1}) + \gamma_k(\sigma_{k-1}, \sigma_k)) \quad (47.2)$$

and

$$\beta_k(\sigma_k) = \max(\beta_{k+1}(\sigma_{k+1}) + \gamma_k(\sigma_k, \sigma_{k+1})) \quad (47.3)$$

The branch metric γ is calculated from $y^{\text{systematic}}$, y^{apriori} and $y^{\text{redundancy}}$ as

$$\gamma = \frac{y^{\text{systematic}} + y^{\text{apriori}}}{y^{\text{intrinsic}}} + \frac{y^{\text{redundancy}}}{y^{\text{extrinsic}}} \quad (47.4)$$

If s^1 and s^0 were the 1-branch and 0-branch of the trellis state transitions, the soft output value, *Log-likelihood Calculation* (LLC) at time k , is defined as

$$LLC_k = \max(L_k)_{s^1} - \max(L_k)_{s^0}. \quad (47.5)$$

It requires large computation and storage, forward and backward recursion results in long decoding delay as the drawbacks.

47.2.2 Parallel Processing Levels

In turbo decoding, parallelism is applied at metric level, *soft-in soft-out* (SISO) decoder level and turbo-decoder level. The *Bahl-Cocke-Jelinek-Raviv* (BCJR) metric level parallelism concerns the processing of all metrics involved in the decoding of each received symbol. Trellis transition and BCJR Computation parallelism is exploited by the metric level.

Parallelism of trellis transitions. For the complete frame, each transition pairs has the same trellis computation technique. The trellis computation can be divided into Branch-Metric Calculation (BMC) and Add-Compare-Select (ACS) stages. Since the same operations are repeated in every pair so parallelism can be extracted from trellis structure. This results in equaling to half the number of transitions per trellis section and constitutes the upper bound of the trellis-transition parallelism degree. It facilitates low area overhead as only the ACS units have to be duplicated, eliminates the use of extra memory as all the parallelized operations are executed on the same trellis section, and in consequence on the same data.

Parallelism of BJCR computations. The forward-backward scheme permits parallel execution of backward recursion and A posteriori Probability (APP). The butterfly scheme permits doubling this parallelism degree only by duplicating the BCJR resource without addition of memory. It helps optimizing area efficiency. The parallelism degree is limited by the decoding algorithm and the code structure. Thus, achieving higher parallelism degree implies exploring higher processing levels.

Parallelism of SISO decoder. SISO decoder parallelism consists of the use of multiple SISO decoders, each executing the BCJR algorithm and processing a sub-block of the same frame in one of the two interleaving orders.

Sub-block parallelism. The frames are divided into sub-blocks to be processed on a BCJR-SISO decoder. Duplication of decoders can result in communication conflicts and hence use of conflict management mechanism and finally leads to long and variable communication time. The parallelism, though intended to reach high throughput, has a reduced throughput since resolving the initialization issue implies a computational overhead following an Amdahl's law (due to acquisition length or additional iteration).

Decoder component parallelism. Here, all the decoder's components are executed in parallel and the extrinsic information is exchanged as generated. The shuffled decoding technique performs decoding and interleaving concurrently thereby reducing iteration period but requires techniques to avoid memory conflicts. Shuffled decoding is more efficient than sub-block parallelism for high throughput.

Table 47.1 Selection of standards and channel codes

Standard	Codes	Rates	Throughput (Mbit/s)
LTE	bTC	1/3	...100
UMTS	bTC	1/3	...2
DVB-RCT	dbTC	1/2,3/4	...31
IEEE802.16(WiMAX)	dbTC	1/3...7/8	...20

47.3 Flexible Decoder Requirement

An analysis from [5] has concluded that most of communication standards are using binary convolutional and binary and duo-binary turbo codes. Commonalities between convolutional and turbo decoding can be found although there are differences in block size polynomials, and coding rates.

Table 47.1 summarizes between different communication standards which use turbo code schemes for channel encoding. The differences in parameters such as constraint lengths, generator polynomials and interleaving patterns can be seen. Thus to implement a flexible decoder, ASIP has to support this wide range of coding parameters.

Through investigation of standards in Table 47.1, the following specifications were derived which have to be provided by ASIP to fulfill the flexibility requirements for turbo decoding systems:

- supports of $N = 4 \dots 16$ states for bTC (binary turbo codes),
- supports of $N = 8 \dots 16$ states for dbTC (double-binary turbo codes),
- arbitrary feedback polynomials,
- arbitrary generator polynomials,
- fast reconfigurable trellis structure,
- High throughput and low latency.

47.4 Turbo Decoder Architecture

The top level view of the architecture as Fig. 47.1, is composed of operative, control, communication interface and memories to store data. Inside the operative part consists of two identical BJCR units, corresponding to forward and backward processing in the MAP algorithm. These units will process a window of 64 symbols and produce recursion metrics and extrinsic information. The memory units are divided into two, internal and external. Two recursion metric storage (cross) implements under internal memories, it will be used by each forward and backward unit. Another internal memory (config) contains up to 256 trellis descriptions. The ASIP architecture can be configured for the correspondent standard depending on the information stored in the config. On the other side the external memories consist of `input_data` and `info_ext` to store the incoming data

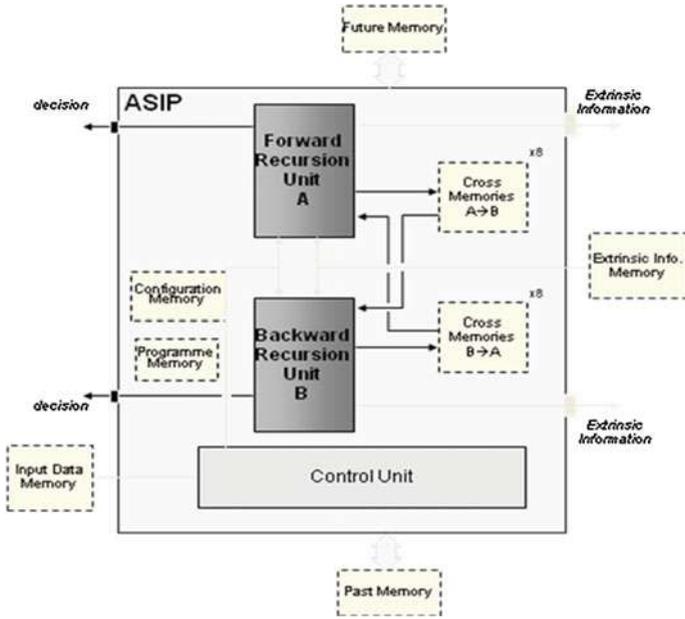


Fig. 47.1 ASIP architecture of turbo decoder

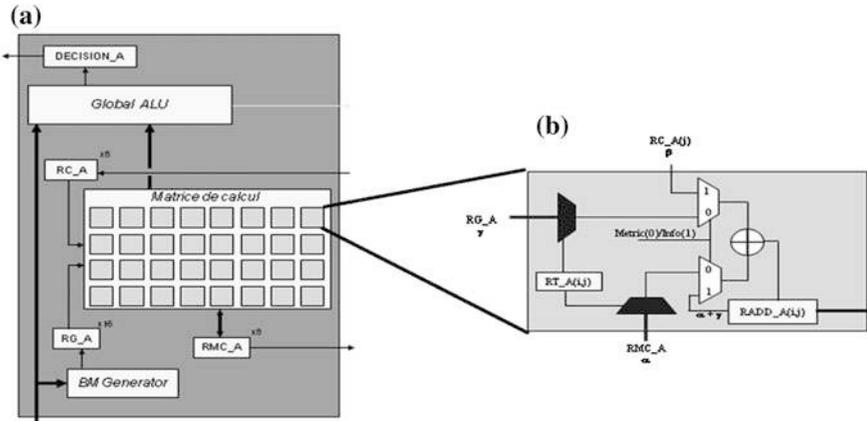


Fig. 47.2 a BCJR computation unit. b Adder node

systematic and redundant, also future and past used to initialize state metric value for beginning and the end of each window.

BCJR computation unit. It also known as recursion unit, in order to support trellis parallelism (up to 32 adder nodes) this unit using Single Instruction Multiple Data (SIMD) architecture. As seen in Fig. 47.2a, this could be decomposed into maximum 8 states incorporated each other by the maximum of 4 decisions per

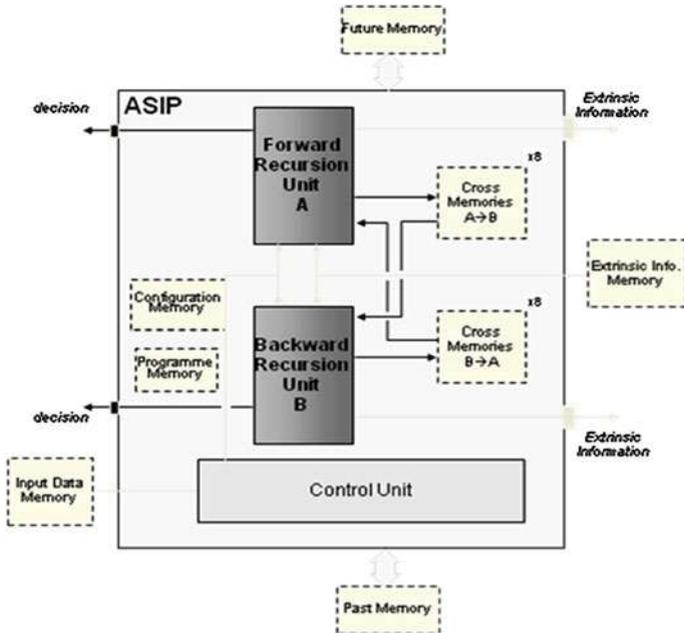


Fig. 47.3 a BCJR computation unit. b Adder node

decoding time. It can be view as a processing matrix. For a 16 state simple binary codes, transition with ending states 0–7 are mapped on matrix nodes of row 0, if the transition bit decision is 0, or matrix node of row 0, or matrix nodes of row 1, if transition bit decision is 1, whereas states 8–15 are mapped on nodes of row 2 and 3.

An adder node (Fig. 47.2b) contains one adder optimized in [7] compare to two adders in [6], multiplexers, one register for configuration (RT) and an output register (RADD). This adder node function to support addition required in recursion between state metric (RMC) and a branch metric (RG) and also the addition required in information generation since it can accumulate the previous result with the state metric of the other recursion coming from register bank RC.

To find the max between recursion a max node contains three max operators is place between four nodes. With this structure it make possible to perform either 4-input maxim or two two-input maximum. Resulted are stored either in first columns or rows of RADD matrix or in RMC bank to archive recursion computation.

Control unit. ASIP design methodologies allow the turbo decoder’s control unit to use the pipeline strategies. These six-stage pipelines consist of *Fetch*, *Decode*, *Operand Fetch*, *Branch Metric* (added stage), *Execute* and *Store*. The pipeline was kept short to preserve some flexibility for further extensions. Branch metric stage was added to reduce the processing cycle in Execute in order to increase the throughput buy increase the clock frequency of the ASIP.

Instruction set. Each instruction set is 16 bit wide, to perform basic turbo decoding, the 30 instruction line shows in Fig. 47.3 are compulsory. It performs

the basic MAP algorithm operation such as control, operative and IO. The configuration and initialization or recursion metric are performed in the first six instructions-line, then it follows by initialize the butterfly loop by ZOLB instruction [7]. Compared to [6], the initialization of the butterfly are done by part, several ZOL instruction are needed. Since the example is to decode duo-binary turbo codes 2 max operator are required (max 2 m in [7] and two max m in [6]). The last 5 instructions are to compute the extrinsic information for 8 states and the program branched to beginning of butterfly.

Concerning of the execution time the enhanced version [7] showed improvement, $7*N/2$ cycles (3.5 cycles roughly) are needed to process N symbols of sub-block while the original version [6], $15*N/2$ (7.5 cycles roughly) cycles are needed to process N symbols of sub-block.

47.5 Results

The architecture was implemented with LISATek tool set, and the generated VHDL model was synthesized with Synopsys Design Compiler in in ST 0.09 μm (resp. 0.18 μm) ASIC technology under worst-case conditions. The synthesized ASIP has a maximum clock frequency of 335 (resp. 180) MHz and occupies about 97 (resp. 93) KGates (equivalent). This means that a single processor, running the two SISO of 6 turbo iterations over a DVB-RCS code, will potentially have a throughput of 7.4 (resp. 4) Mbit/s.

47.6 Conclusion

An ASIP for Turbo decoding has been described in this article. As we can conclude that application specific flexibility is mandatory to meet the flexibility and performance requirement of the emerging communication standards nowadays. It can be archived by Application Specific Instruction set Processor with designed pipeline topology, architecture and memory infrastructure.

References

1. Berrou, C., Glavieux, A., Thitimajhima, P.: Near Shannon limit error-correcting coding and decoding: turbo codes, ICC Geneva, (1993)
2. Prescher, G., Gemmeke, T., Noll, T.: A parameterizable low-power high-throughput turbo decoder. In proceedings Of ICASSP vol. 05, pp. 25–28 Pennsylvania (2005)
3. Gnaëdig, D., Boutillon, E., Jezequel, M., Gaudet, V., Gulak, G.: “On multiple slice turbo code” The 3rd International Symposium on Turbo Codes and Related Topics, pp.343–346 Brest, France (2003)

4. Keutzer, K., Malik, S., Newton, A.R.: "From ASIC to ASIP- next design discontinuity". In: IEEE International Conference on Computer Design: VLSI in computers and processors, pp. 84–90, IEEE Computer Society, Washington (2002)
5. Vogt, T., Wehn, N.: "A reconfigurable application specific instruction set processor for viterbi and Log-MAP decoding". In: IEEE Workshop on Signal Processing Systems Design and Implementation, SIPS '06 pp. 142–147, Banff, Canada (2006)
6. Muller, O., Baghdadi, A., Jezequel, M.: "ASIP-based multiprocessor SoC design for simple and double binary turbo decoding". In: Proceedings Design, Automation And Test In Europe, pp. 1330–1335, Munich (2006)
7. Muller, O., Baghdadi, A., Jezequel, M.: "From application to ASIP-based FPGA prototype: a case study on turbo decoding". In: IEEE/IFIP International Symposium on Rapid System Prototyping, Monterey, (2008)

Chapter 48

A Comparative Study of Web Application Testing and Mobile Application Testing

Maryam Ahmed and Rosziati Ibrahim

Abstract Web application have gained increased acceptance over the years in companies and organization as the world move to a global village. Software developers have also grown interest in developing web applications compared to stand-alone application because of the immense benefits it offers such as ubiquity, platform dependence, low cost of support and maintenance, better speed and performance, piracy proof etc. As mobile application emerged in the last decade, attention has been focused on mobile applications by organizations and businesses in order to maximize their profits as much as possible. There has been a rapid increase of software release in the mobile applications store. As the growth of both web application and mobile application increase, the question of quality assurance remains a concern. A comparative study of software testing techniques can be performed to improve the standard of testing of both web and mobile application. This paper therefore reviews the similarity and difference in the testing mechanism.

Keywords Software testing · Mobile application testing · Web application testing · Mobile applications · Web applications

48.1 Introduction

The idea of quality assurance has been in existence even before the existence of software hence quality of manufactured products including software has been a subject of keen interest from past till present. Quality has been an issue as long as

M. Ahmed (✉) · R. Ibrahim
Computer Science and Informtion Technology, Universiti Tun Hussein Onn Malaysia
(UTHM), 86400 Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: hi120055@siswa.uthm.my

R. Ibrahim
e-mail: rosziati@uthm.edu.my

human has been producing product. Software testing is an important and major area of Software Quality Assurance (SQA). Software testing is the process of executing a program with the purpose of finding faults. Testing include effort to find defects but does not include getting solution to fixing the defects. This is the difference between testing and quality assurance as quality assurance is not limited to developing the test plan but also include testing, preventing and fixing the faults found during the process of testing. However, testing can cover different areas such as Specification testing, design testing and implementation testing. Implementation testing which deals with the working system of the software is most time referred to as software testing [1]. Testing cannot be complete or perfect. Having a flawless testing is not realistic [1], as it could take more than a life time to complete even the simple software. Saleh explained the unrealistic nature of complete testing by using a small program of a user filling a text field of 20 characters. This program test will be complete by testing all possible input values. If an assumption of 80 characters is done, then number of possible combinations is 20^{80} . A computer that takes nanoseconds to process one combination, testing will take 10^{11} years to complete all possible combination which is very impossible and unrealistic. Since it is impossible to guarantee an 100 % error-free software, errors that are not detected at the time of development before deployment may be discovered and reported by the end user or in the process of testing for a subsequent release of the software. Despite the imperfection of software testing, testers need to put their best effort in improving the reliability and efficiency of the software as it can affect the well-being of human and their safety.

The current surge in both number and demand for mobile applications has called the attention of test engineers to the testing of mobile application as users are more concerned about the quality and functionality of the software. However, the mobile applications testing has been an issue given the different platforms the applications work on such as the iOS, Windows mobile, and the Google android; whose market has continued to grow rapidly over the years [2, 3]. Also, Hardware complexity and diversity is another challenge in mobile application testing. Unlike the PC environment that operates with one central processing unit ($\times 86$ -micro-processor) on which applications running on the system must be tested (whether voice or data services), mobile phones are built with different types of processors that run at different speed and with different memory usage. Testing web application involve looking into the application and attention should be paid to the browser being used (considering factors like backward and forward buttons) [4]. Connectivity is another issue with the smart mobile device. A mobile device could use the 3G technology, wifi, GPRS all with different level of signal strength [5]. Therefore, testing mobile application requires a testing technique that will be compatible with the software and hardware complexity of the mobile phone.

48.2 Web Application Testing

The advancement of web application is following an evolution similar to that of software systems. Web applications tend to evolve rapidly and pass through frequent modifications as a result of new technological and business opportunities as well as feedback from users [6]. Web application testing can however be more complex than testing standalone or traditional programs due to their synchronized, dispersed and platform-independent environment on which it is been run. Several researches have been going on since the last decade when web application started becoming common around the globe. In 2000, research was carried out to extend data flow testing technique to web application testing. An approach to supporting data flow testing for web application was presented [7]. The approach is such that the white box test artifacts of an application were captured in a Web Application Test Model (WATM) there by treating each component of the web application as an object, while the elements of an HTML or XML were considered as code variables within an object. However, there is need for review of this technique and other similar approach as this is based on the structural testing.

Qian [8] developed a tool for testing web application and to support analysis of the application. They came up with ReWeb and TestWeb. The ReWeb views could be useful in understanding the system organization in terms of navigation path and variable usage while the TestWeb generator and executor of test cases of an application can be used to explore the system to a satisfactory level. Improvement is needed in reducing the manual activities in the process. Automation is required in the area of state unrolling and merging. Bin Zhu, Huaikou Miao and Lizhi Cai (2009) proposed an approach to generating test cases for web application testing using a navigation tree considering the web browser history mechanisms and the user interface facilities such as the back, forward and refresh tool [9]. Another work on web application testing worth mentioning is the development of a cross-browser web application testing tool (Shauvik Roy, 2010). This tool is to find the difference between corresponding elements of a web page opened in different browsers [10]. The tool was evaluated on nine real world applications and result shows that the tool is effective in finding cross-browser issues while keeping the false positive low. Qian Zhongsheng also came with an approach to reducing and optimizing the test case generation from user request traces [8]. A large volume of meaningful user sessions were obtained after purging their irrelevant information by analyzing user logs on the web server. This approach evaluates test cases considering the test coverage ratio only. Many factors need to be considered aside from the test coverage ratio such as the running cost of each test case (CPU time), loading time, time to save the test case and the influence of different test criteria on a test case. The Table 48.1 summarizes the state of the art of web application testing.

Table 48.1 Web application testing techniques

Author	Topic	Methodology/ Contribution	Comment
Chien-Hung Liu David C. Kung Pei Hsia Chih- Tung Hsu [7]	Structural testing of web applications	Web application test model (WATM)	This approach was based on Structural or white box testing
Fillipo Ricca Pallo Tonella [6]	Analysis and testing of web applications	ReWeb and TestWeb	Involves manual processes, therefore need to review to automate processes
Bin Zhu Huaikou Miao Lizhi Cai [4]	Testing a web application involving web browser interaction	Navigation tree and put into consideration web browser history and user interface	Other factors affect the testing of web aside from these two factors
Shauvik Roy Choudhary Husayn Versee Alessandro Orso [10]	A cross-browser web application testing tool	Tool to find the difference between two different browsers	From the result, it was effective in finding cross browser difference
Qian Zhongsheng [8]	Test case generation and optimization for user session-based web application testing	To reduce and optimize test cases generation from user request traces	More factors need to be considered

48.3 Mobile Application Testing

Mobile application also known as Mobile Applications can be used to define the applications that are developed for handheld smart devices such as the mobile phones, etc. These applications can be pre-installed applications that come with the phone straight from the manufacturer or it can be downloaded from the web or app store to enhance better functionality of the smart device. These devices work with powerful application for user consumption and for this reason increases its complexity. The complexity of the mobile applications comes with a cost [9]. Mobile application testing can be very challenging. Many questions need to be addressed when testing a mobile app. Different framework have been proposed for mobile application testing. A cloud based approach was proposed by Baride and Dutta. This approach was proposed to solve the problem of homogeneous testing of mobile application of different platforms [11]. The success of this work was explained in terms of it automation, adaptability to different mobile environment and actual devices and the complexity of the testing which include performance testing, security testing and synchronization testing. Another mobile testing framework was proposed by Ping et al. This framework was based on the V-model and cloud test mechanism. The criteria to be considered in designing a testing model for mobile application was mentioned to include testing scope in terms of the network connection been used, emulators used in testing and devices used [12].

The developed model is however yet to be implemented. The work of Muhammad Karami (2013) presented an automatic analysis approach for security inspection of android applications. This research was inspired by the scrutiny that some malware examples are only triggered on user behavior or action [13]. Hu Cuixiong proposed a novel technique of bringing android specific bug to light and shows how to construct an effective test automation approach for addressing the bugs detected, hence ensuring the reliability of application running on android platform [5].

Different tools have also been developed to aid mobile application testing. MobileTest (2007) is an automatic testing tool for black box mobile devices. This tool can develop refined, maintainable and reusable test case library for testing system level and application level software on different smart devices [14]. The tool was validated by comparing with the result of the TestQuest. Hermes (2009) is another tool for testing smart mobile applications. The motivation of development of this tool comes from the heterogeneity of the mobile device [15]. Hermes can be used for test writing and a distributed run time for automating test execution and reporting. There is need to look into the cost/benefit ratio to improve the evaluation of this tool. Adaptive Random Testing (ART) 2010: ART was also developed from a black box view of testing mobile application [16]. The motivation was from the fact that mobile application has to deal with user input and constant changing in the device or user environment. MZoltar (2013): Mzoltar offers a vigorous analysis of mobile applications which shows a diagnostic report that makes it easy to comprehend [17]. This approach helps in localizing the bugs in android mobile applications by relying on the Spectrum-based fault localization (SFL). Table 48.2 shows a summary of the state of the art of mobile application testing.

48.4 Difference of Web and Mobile Application Testing

With the high increase of mobile users and the rate at which the internet is being accessed, there is need to bridge the gap in the quality of mobile and web applications. There are different factors to be considered in understanding the adaptability of the two testing techniques. These factors affect the way at which these applications are developed and tested. Such factors include and not limited to: Life conformance or event trigger mode, GUI (GUI of the Input and Output system), Data Synchronization or Network Management, Power Management and Memory management.

48.4.1 Event Context/Life Conformance

In web application running on a desktop, application life cycle is dependent on the operating system. The operating system handles the states of the event the application is passing through and it safeguards the precise comportment of the

Table 48.2 Mobile application testing techniques

Author	Topic	Methodology/ Contribution	Comment
Srikanth Baride Kamlesh Dutta [11]	A cloud based software testing paradigm for mobile applications	Cloud based approach to solve the problem of heterogeneity of mobile devices	This approach did not cover the issue of testing connectivity
Tan Ping Ping Hamizan Sharbini I Wee Bui Lin [12]	Designing a mobile application testing model	V-model and cloud base mechanism	An ongoing research on mobile testing framework
Mohammad Karami Mohamed Elsabagh Parnian Najafborazjani Angelos Stavrou [13]	Behavioral analysis of android applications using automated instrumentation	An automated dynamic analysis approach for Security inspection of android applications	There is need for support for more complex interfaces such as OpenGL and gesture views
Cuixiong Hu Iulian Neamtiu [5]	Automating GUI testing for android applications	An approach to test android application with focus on the GUI	There were exceptions in the result due to bugs that do not fall under activity/event/type categories
Jiang Bo Long Xiang Gao Xiaopeng [14]	MobileTest: A tool supporting automatic black box test for software on smart mobile devices	MobileTest took to event based approach to simplify generation of test cases	This helps in reducing the complexity of smart mobile testing while still being effective
Sakura She Sasindran Sivapalan Ian Warren [15]	Hermes: a tool for testing mobile device applications	Hermes an automated test execution and reporting tool	More research to validate and refine the tool especially in the area of reporting
Zhifang Liu Xiaopeng Gao Xiang Long [16]	Adaptive random testing of mobile application	Adaptive random test (ART)	The experimental result of this tool shows that it reduces time taken to find first defect and reduces number of test cases generated.
Pedro Machado José Campos Rui Abreu [17]	MZoltar: automatic debugging of android applications	MZoltar localizes the bugs in android mobile applications by relying on the Spectrum-based fault localization (SFL)	This approach is unique in identifying potential defects quickly through diagnostic report

application at all statuses. This is not the case in smart mobile operating system such as iOS, Android. Due to limited resources and complexity of the system, the operating systems cannot retain the comprehensive state of an application at any time when there is a change in the state of the life cycle. The application therefore takes care of itself to avoid data loss in case the application is paused or even killed. This makes it a requirement for a mobile application to be life conformance in its design. Testing the life conformance of an application life implies it is responding reasonably to change of state in the operating system such that events like low memory or low battery do not lead to loss of data in the mobile device [18–20].

48.4.2 GUI

Testing GUIs on a general note is a tough and demanding task for so many reasons: Firstly, the input space has a countless and potentially unlimited number of combinations of inputs and states of system outputs, generating test cases then becomes tougher. Secondly, even simple GUIs possess an enormous number of states which are also due to interaction with the inputs. Also, many complex dependencies may hold between different states of the GUI system, and/or between its states and inputs [21]. Web application GUI testing still seems easier. Users of desktop GUI applications might be expected to refer to documentation or lessons to completely understand the usage of the applications. Unlike mobile device, mobile applications are expected to have a simple and intuitive user interface where most, if not all, usage scenarios of an application should be evident to the average user from the GUI [22].

48.4.3 Network Management

The emphasis of present research is tending towards planning and building of network and infrastructures and applications for mobile system. Testing software in relation to network connection now remains a concern. This is becoming a hindrance in the evolution of mobile computing since the development of smart mobile application is challenging owing to the scarcity in computational resources. The change in location and network can also lead to data loss and poor functionality. This implies that a change in network or location can mean a small shift away from the servers that is being used toward a new one [23].

Table 48.3 Adaptability of mobile and web application testing

Features	Mobile	Web	References
GUI	Input: touch screen and key board	Input: key board	Hu [5], She et al [15], Chang et al [17], Belli [20], Yang et al. [26]
	Output: small screen	Output: larger screen	
Life cycle	Life cycle is dependent on application since system resources are limited	Operating system takes care of the life cycle of an application	Amalfitano et al. [8], Bo et al. [14], Franke et al. [21], She et al. [15]
Performance (Memory, storage and network management)	Mobility allow easy change in location and therefore could result in network strip off for short distance network	Network connectivity not frequently cut-off	Zhu et al. [4]
Power management	Most mobile applications tend to consume much power, hence need for power management mobile software	Power management under control	Zhang et al. [24], Ismail et al. [27]

48.4.4 Power Management

As users tend to use vital and complex applications, low power consumption becomes a requirement in building mobile applications. Developers of smart mobile device (hardware and software) have considered power-saving features, such that applications can vigorously regulate their power ingestions depending on the required functionality and performance. Software developers therefore need to understand the implication of building a high power-hungry application and should put into consideration the in-built feature of power management in mobile device [24] (Table 48.3).

48.5 Conclusion

While the web has taken over businesses and individual life, the mobile is however anticipated to surpass as the world platform for local and internet applications in the nearest future. As the functionality of the mobile and web applications increase so is the complexity and hence the complexity of the testing technique. There is need to fill the gap in the testing of these two important applications area. The difference in performance and testing could be viewed from the event context, GUI interface, network management, power and memory management. There are present works that have been looking into these factors as highlighted in this paper;

however, there is a strong need for more research in building more effective testing technique that can handle both the web and mobile application testing.

Acknowledgments This research is supported under the Graduate Research Incentive Grants (GIPS), vote 1256, Universiti Tun Hussein Onn Malaysia.

References

1. Saleh, K.A.: Software engineering. J. Ross Publishing Inc. p. 244
2. Franke, D., Weise, C.: Providing a software quality framework for testing of mobile applications. 2011 fourth IEEE international conference on software testing, verification and validation, 431–434 (2011). doi:[10.1109/ICST.2011.18](https://doi.org/10.1109/ICST.2011.18)
3. Wasserman, A. I.: Software engineering issues for mobile application development. In: Proceedings of the FSE/SDP workshop on future of software engineering research – FoSER-10, p. 397. (2010). doi:[10.1145/1882362.1882443](https://doi.org/10.1145/1882362.1882443)
4. Zhu, B., Miao, H., and Cai, L.: Testing a web application involving web browser interaction. 2009 10th ACIS international conference on software engineering, artificial intelligences, networking and parallel/distributed Computing, 589–594 (2009). doi:[10.1109/SNPD.2009.59](https://doi.org/10.1109/SNPD.2009.59)
5. Hu, C.: Automating GUI testing for android applications, (Section 4), pp. 77–83
6. Ricca, F., Tonella, P.: Analysis and testing of web applications. In: Proceedings of the 23rd international conference on software engineering. ICSE 2001, 25–34 (2001). doi:[10.1109/ICSE.2001.919078](https://doi.org/10.1109/ICSE.2001.919078)
7. Kung, C. L. D. C., Box, P. O.: Structural testing of web applications Chih-Tung Hsu, 84–96
8. Qian, Z.: Test Case Generation and Optimization for User Session-based Web Application Testing. *J. Comput* **5**(11), 1655–1662 (2010). doi:[10.4304/jcp.5.11.1655-1662](https://doi.org/10.4304/jcp.5.11.1655-1662)
9. Zhifang, L., Bin, L.: Test automation on mobile device (2007), 1–7 (2010)
10. Roy Choudhary, S., Versee, H., Orso, A.: A cross-browser web application testing tool. 2010 IEEE international conference on software maintenance, 1–6 (2010). doi:[10.1109/ICSM.2010.5609728](https://doi.org/10.1109/ICSM.2010.5609728)
11. Baride, S., Dutta, K.: A cloud based software testing paradigm for mobile applications. *ACM SIGSOFT Softw. Eng. Notes* **36**(3), 1 (2011). doi:[10.1145/1968587.1968601](https://doi.org/10.1145/1968587.1968601)
12. Ping, T. P., Sharbini, H., Lin, W. B., Tan, V., Mun, W., Julaihi, A. A.: Designing a mobile application testing model, 255–260
13. Karami, M., Elsabagh, M., Najafiborazjani, P., Stavrou, A.: behavioral analysis of android applications using automated instrumentation
14. Bo, J., Xiang, L., Xiaopeng, G.: MobileTest: a tool supporting automatic black box test for software on smart mobile devices. Second international workshop on automation of software test (AST '07), 8–8. (2007). doi:[10.1109/AST.2007.9](https://doi.org/10.1109/AST.2007.9)
15. She, S., Sivapalan, S., Warren, I.: Hermes: a Tool for Testing Mobile Device Applications. *Australian Softw. Eng. Conf.* **2009**, 121–130 (2009). doi:[10.1109/ASWEC.2009.17](https://doi.org/10.1109/ASWEC.2009.17)
16. Liu, Z., Liu, B., Gao, X.: SOA based mobile application software test framework. 2009 8th international conference on reliability, maintainability and safety, 765–769 (2009). doi:[10.1109/ICRMS.2009.5270087](https://doi.org/10.1109/ICRMS.2009.5270087)
17. Machado, P., Campos, J., & Abreu, R.: MZoltar: automatic debugging of Android applications. In: Proceedings of the 2013 international workshop on software development lifecycle for mobile - demobile 2013, 9–16 (2013). doi:[10.1145/2501553.2501556](https://doi.org/10.1145/2501553.2501556)
18. Amalfitano, D., Fasolino, A. R., Tramontana, P.: A GUI crawling-based technique for android mobile application Testing. 2011 IEEE fourth international conference on software testing, verification and validation workshops, 252–261. (2011). doi:[10.1109/ICSTW.2011.77](https://doi.org/10.1109/ICSTW.2011.77)

19. Amalfitano, D., Fasolino, A. R., Tramontana, P., Amatucci, N.: Considering Context Events in Event-Based Testing of Mobile Applications. 2013 IEEE sixth international conference on software testing, verification and validation workshops, 126–133. (2013). doi:[10.1109/ICSTW.2013.22](https://doi.org/10.1109/ICSTW.2013.22)
20. Belli, F.: Finite state testing and analysis of graphical user interfaces. In: Proceedings 12th international symposium on software reliability engineering, 34–43 (2001). doi:[10.1109/ISSRE.2001.989456](https://doi.org/10.1109/ISSRE.2001.989456)
21. Franke, D., Kowalewski, S.: Testing conformance of life cycle dependent properties of mobile applications, (2012). doi:[10.1109/ICST.2012.36](https://doi.org/10.1109/ICST.2012.36)
22. Yang, W., Prasad, M. R., Xie, T.: A grey-box approach for automated gui-model generation of mobile applications
23. Satoh, I., Society, I. C.: A testing framework for mobile computing software, 29(12), 1112–1121 (2003)
24. Zhang, L., Dick, R. P., Mao, Z. M., Wang, Z., Arbor, A.: Accurate online power estimation and automatic battery behavior based power model generation for smartphones, 105–114
25. Chang, T., Yeh, T., Miller, R. C.: GUI testing using computer vision, (Figure 1), 1535–1544 (2010)
26. Yang, Y. J., Kim, S. Y., Choi, G. J., Cho, E. S., Kim, C. J., Kim, S. D.: A UML-based object-oriented framework development methodology. In: Proceedings 1998 asia pacific software engineering conference (Cat. No.98EX240), 211–218. doi:[10.1109/APSEC.1998.733722](https://doi.org/10.1109/APSEC.1998.733722)
27. Ismail, M.N., Ibrahim, R., MdFudzee, M.F.: A survey on content adaptation systems towards energy consumption awareness. *Adv. Multimedia* **2013**, 3 (2013)

Chapter 49

Multi-objective Functions in Grid Scheduling

Zafril Rizal M. Azmi, M.A. Ameen
and Imran Edzereiq Kamarudin

Abstract In order to fully utilize the Grid resources, an implementation of a good scheduling algorithm is greatly important. However, for a complex scheduler that aims to achieve high performance for more than one performance metrics, a suitable objective function should be carefully considered. This paper shows that a different objective function will have different affect to the Grid performance.

49.1 Introduction

Recently, Grid computing have been attracting interest from researchers in the field of mathematics [8], biosciences [10], engineering [17] and others important fields. Furthermore, the increasing need for network, storage and computing resources is projected to double every 9, 12 and 18 months respectively [7]. As the increase in data size, processing complexity as well as communication technology is becoming more challenging in the Grid computing environment, issues related to performance becomes increasingly important. Achieving high-performance Grid computing requires techniques to efficiently and adaptively allocate jobs to available resources in a large scale, highly heterogeneous and dynamic environment. Therefore, it is very important for a Grid system to have a scheduler that meets both administrator and user performance expectation.

Z.R.M. Azmi (✉) · M.A. Ameen · I.E. Kamarudin
Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang,
Lebuhraya Tun Razak, Gambang, 26300 Pahang, Kuantan, Malaysia
e-mail: zafril@ump.edu.my; zafrilrizal@yahoo.com

M.A. Ameen
e-mail: mohamedariff@ump.edu.my

I.E. Kamarudin
e-mail: edzereiq@ump.edu.my

49.2 Single Objective and Multi-objective Function

Although scheduler is a very important element in a Grid based application, a good scheduler always comes paired with an efficient *objective function*. In a Grid environment, a feasible good solution is a must and an optimal solution is a goal. Commonly used Priority Rule (PR) scheduling procedure such as Shortest Job First (SJF), First Come First Serve (FCFS) and Earliest Deadline First (EDF) can quickly create feasible solution. However, as problem complexity increases, the solution may be far from optimal. In order to tackle this problem, more complex scheduling procedure based on approximate and heuristic have been introduced. Schedulers that apply standard PR procedure has no need for mechanisms to filter the solution since the PR generate solution based on the job characteristics such as job length and deadline, and the solution is final, but heuristics based scheduler usually generate a set of solutions that needs a filtering mechanism in order to select the solution that meets the objective. The purpose of objective function is to conduct the filtering process by evaluating the quality of the solutions generated by the scheduler.

From the perspective of this paper, objective function in general can be divided to two categories: direct objective function and indirect objective function. Direct objective function involves certain calculation implemented in the schedulers specifically targeting the main objective function. For example in [11], in order to achieve minimum makespan, the new generated solution by Genetic-Based scheduler must have lower value of makespan compared to previous solution to be accepted. Since the scheduler only accepts much lower makespan solution, the overall makespan for the Grid system can be guaranteed to be minimum. On the other hand, indirect objective function is the outcome or side effect of the direct objective function. For example, by using Shortest Job on Fastest Node (SJFN) [19] which targets to minimize the makespan, the side effect is the increasing of flowtime and machine usage. The indirect objective function always reflects the tradeoffs made by the scheduler to the objective function, for example, minimizing objective A is subject to maximizing objective B. The tradeoff is often a main topic for the multi-objective scheduling which has to be properly considered when developing a new scheduler.

There are two types of direct objective functions implemented in scheduling algorithms. The first one is just using a single objective function and another is combining two or several objective functions to make a decision which also called multi-objective function. Generally, makespan minimization is often used as the criterion of optimization in a single objective function. Pandey in [21] have stated that makespan is the most common objective function used in scheduling. Makespan is the time taken to finish the overall jobs. Since it is very important to minimize the overall time, many researchers used makespan as the objective function in their work [9, 13, 24, 25].

However using just a single objective function to schedule a set of jobs can decrease the overall performance [21, 26]. Moreover, users' objectives such as the need for much faster completion of jobs and resource provider objective such as maximizing resource utilization always conflict with each other [20, 28]. Obviously, it is not realistic to satisfy a single objective, but to search for a tradeoff, which is nondominated solution. In addition, Farzi in [12] have stated that scheduling in Grids computing is a multi-objective optimization problem, hence objectives other than makespan must also be taken into consideration. To address this problem and to increase other performance criteria's, multi-objective function has been introduced. Moreover, using more than single objective function is believed to improve and also balance the performance. Multi-objective function can have two or more objective functions working together to calculate the decision. For a multi-objective function with k sub objectives, the objective can be

defined as $f(x) = \min \sum_{i=1}^k w_i f_i(x) (i = 1, 2, \dots, k)$, and $w_i \geq 0$ is the weight of the i th sub objective, in general, $\sum_{i=1}^k w_i = 1$ [2, 27].

Since multi objective function is very important, many researchers, in their work have combined several single objective functions to make a decision for their scheduler. For example [1] used the combination of makespan, machine usage and acceptable queues to make a decision for a Genetic Algorithms (GA) based scheduler. Although any combination of two criteria's is possible, some combinations are more obvious than the other. Nevertheless, most multi objective based scheduler pairs makespan with other objective functions, but there is no concrete reason why they are paired together. Leung in [18] have stated, a schedule that simultaneously minimizes both makespan and flow time is called an ideal schedule. This statement is agreed by [6] that focusing on minimizing makespan and flow time by using the GA based scheduler.

Table 49.1 summarize the different between single-objective function and multi-objective function. Selecting the best objective function for a specific scheduler will determine the achievement of the scheduler objective. Hence, this paper will further discuss the interaction between the Grid scheduler with objective function in Sect. 49.3 and provide experimentation results that will show the scheduler performance using different objective functions in Sect. 49.4.

49.3 Multi-objective Function in SH-PR-GSA Scheduler

In this paper, various single and multi-objective functions have been considered to be tested. In order to do this, a set of PR schedulers optimize by Backfilling and Simulated Annealing called Smallest Hole-Priority Rule-Guided Simulated

Table 49.1 Comparison of single objective function with multi-objective function

Criteria	Single objective functions	Multi-objective functions
Decision making process	Decision whether to accept or reject a schedule is based on one performance criteria	Decision whether to accept or reject a schedule is based on more than one performance criteria
Target users	A group of user that aim to achieve similar single objective	A group of user that have different objectives to achieve
Advantage	Performance criteria used as an objective function will have higher performance compared to other criteria's	Most of the performance criteria will have average performance, where no criteria is obviously lower than the other criteria's
Disadvantage	Some criteria performance are obviously lower than the criteria used as an objective function	Criteria previously used as single objective function will has a dropdown in performance after combined with other objective functions
Key objective functions analyze in this paper	Makespan	Makespan + nondelayed jobs
	Nondelayed jobs	Makespan + machine usage
	Machine usage	Makespan + tardiness
	Tardiness	Machine usage + nondelayed jobs
		Machine usage + tardiness
	Tardiness + nondelayed jobs	

Annealing (SH-PR-GSA) integrated with the selected objective functions have been used as a test bed. Five PR algorithms [First Come First Serve (FCFS), Shortest Job First (SJF), Longest Job First (LJF), Earliest Deadline First (EDF) and Minimum Time To Deadline (MTTD)] have been used separately. The main objective of SH-PR-GSA schedulers is to produce schedule aimed to achieve minimum makespan, minimum flowtime, minimum total tardiness, minimum delayed jobs and maximum machine utilization. In order to achieve this, objective function is one of the components that should be carefully evaluated. This is because, the decision of whether a new schedule should be accepted or rejected depend on the objective function used in the scheduling algorithms. In SH-PR-GSA, objective function has been integrated to the scheduler in SH-PR and GSA modules separately as shown in grey box in Fig. 49.1. Although separately executed, these two modules are using the same objective function in order to maintain the consistency of the schedule. For example, in order to achieve minimum possible makespan for the Grid, if SH-PR module implements a Makespan Objective Function, GSA module must also implement the same objective function to evaluate the schedule.

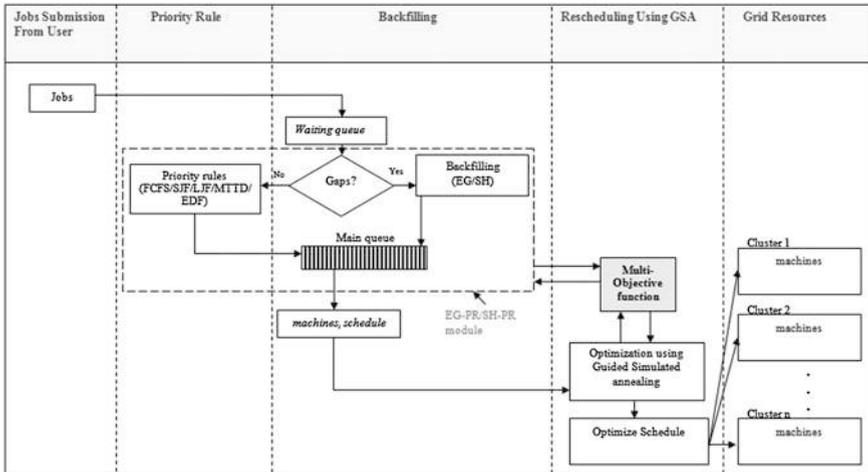


Fig. 49.1 Multi objective function in SH-PR-GSA

In this section objective function referred as “Multi Objective Function” if there is more than one objective function integrated with SH-PR-GSA scheduler. In order to effectively execute SH-PR-GSA to achieve highest possible performance, this section has considered four main objective functions which are total non-delayed jobs (Nondelayed), Makespan, Tardiness and Machine Usage. These objective function have been chosen because each one of them, theoretically representing the targeted performance metrics. For example, by accepting a schedule with lower Nondelayed Objective Function, lowest possible number of delayed jobs can be achieved. On the other hand, Makespan Objective Function aims to achieve the best makespan performance for all SH-PR-GSA schedulers, and as already discussed, higher makespan usually results in lower flowtime. Tardiness Objective Function will ensure that the total tardiness metric will be the lowest metric and finally Machine Usage Objective Function makes sure the scheduler highly utilizes the resources as well as keeping the machine usage metric optimum for SH-PR-GSA. Since objective function play an important part in determining performance of a scheduler, it is very important to conduct a comprehensive study on this matter. During the development process of SH-PR-GSA, apart from the main objective functions explained earlier, there are another six multi objective functions analyzed, each one incorporated by two single objective functions as shown in Table 49.2.

The *weight* is calculated to recognize whether the new schedule is better or worse compared to current schedule by using Eq. 49.1 for objectives to be *minimized* such as makespan and Eq. 49.2 for objective to be *maximized* such as machine usage.

Table 49.2 Objective functions tested with SH-PR-GSA

Number	Type of objective function/multi-objective function
1.	Makespan objective function
2.	Tardiness objective function
3.	Usage objective function
4.	Nondelayed jobs objective function
5.	Makespan + nondelayed jobs multi-objective function
6.	Makespan + usage multi-objective function
7.	Makespan + tardiness multi-objective function
8.	Usage + nondelayed multi-objective function
9.	Usage + tardiness multi-objective function
10.	Tardiness + nondelayed multi-objective function

$$weight_{objective} = (currentObjective - expectedObjective) / currentObjective \quad (49.1)$$

$$weight_{objective} = (expectedObjective - currentObjective) / currentObjective \quad (49.2)$$

Decision based on single objective is done by referring to single $Weight_{objective}$. However, by combining more than one objective functions (multi-objective function) to make a decision over whether to accept or reject a new schedule, a better performance is aimed to be achieved. Instead of focusing on a single objective such as makespan, multi objective such as Makespan + Tardiness Multi-Objective Function decide the solution based on two objectives and in this example they are makespan and total tardiness objective function. Originally, the SH-PR-GSA scheduler applied the combination of makespan and nondelayed jobs weight. The total weight ($totalWeight$) that consists of $weight_{makespan}$ and $weight_{nondelayed}$ are calculated based on the Eq. 49.3 as suggested by Klusacek in [16]. In order to analyze the effects of objective function to the Grid scheduler, this paper has added six multi-objective functions as listed in No. 6–10 in Table 49.2 and defined in Eqs. 49.3–49.8.

Makespan and Nondelayed Jobs (Makespan + Nondelayed Jobs) Multi-Objective Function:

$$\begin{aligned} w_m &= (m_c - m_e) / m_c \\ w_n &= (n_e - n_c) / n_c \\ \sum w &= w_m + w_n \end{aligned} \quad (49.3)$$

where

- w Total weight (*totalWeight*)
- w_m Makespan weight (*weight_{makespan}*)
- w_n Nondelayed jobs weight (*weight_{nondelayed}*)
- m_c Current value of makespan (*currentMakespan*)
- m_e Expected value of makespan (*expectedMakespan*)
- n_e Expected value of nondelayed jobs (*expectedNondelayed*)
- n_c Current value of nondelayed jobs (*currentNondelayed*)

Makespan and Machine Usage (Makespan + Machine Usage) Multi-Objective Function:

$$\begin{aligned}
 w_m &= (m_c - m_e)/m_c \\
 w_u &= (u_e - u_c)/u_c \\
 \sum w &= w_m + w_u
 \end{aligned}
 \tag{49.4}$$

where

- w Total weight (*totalWeight*)
- w_m Makespan weight (*weight_{makespan}*)
- w_u Machine usage weight (*weight_{usage}*)
- m_c Current value of makespan (*currentMakespan*)
- m_e Expected value of makespan (*expectedMakespan*)
- u_e Expected value of machine usage (*expectedUsage*)
- u_c Current value of machine usage (*currentUsage*)

Makespan and Tardiness (Makespan + Tardiness) Multi-Objective Function:

$$\begin{aligned}
 w_m &= (m_c - m_e)/m_c \\
 w_t &= (t_c - t_e)/t_c \\
 \sum w &= w_m + w_t
 \end{aligned}
 \tag{49.5}$$

where

- w Total weight (*totalWeight*)
- w_m Makespan weight (*weight_{makespan}*)
- w_t Tardiness weight (*weight_{tardiness}*)
- m_c Current value of makespan (*currentMakespan*)
- m_e Expected value of makespan (*expectedMakespan*)
- t_e Expected value of tardiness (*expectedTardiness*)
- t_c Current value of tardiness (*currentTardiness*)

Machine Usage and Nondelayed Jobs (Machine Usage + Nondelayed) Multi-Objective Function:

$$\begin{aligned}w_u &= (u_e - u_c)/u_c \\w_n &= (n_e - n_c)/n_c \\ \sum w &= w_u + w_n\end{aligned}\tag{49.6}$$

where

- w Total weight (*totalWeight*)
- w_u Machine usage weight (*weight_{usage}*)
- w_n Nondelayed jobs weight (*weight_{nondelayed}*)
- u_e Expected value of machine usage (*expectedUsage*)
- u_c Current value of machine usage (*currentUsage*)
- n_e Expected value of nondelayed jobs (*expectedNondelayed*)
- n_c Current value of nondelayed jobs (*currentNondelayed*)

Machine Usage and Tardiness (Machine Usage + Tardiness) Multi-Objective Function:

$$\begin{aligned}w_u &= (u_e - u_c)/u_c \\w_t &= (t_c - t_e)/t_c \\ \sum w &= w_u + w_t\end{aligned}\tag{49.7}$$

where

- w Total weight (*totalWeight*)
- w_u Machine Usage weight (*weight_{usage}*)
- w_t Tardiness weight (*weight_{tardiness}*)
- u_e Expected value of machine usage (*expectedUsage*)
- u_c Current value of machine usage (*currentUsage*)
- t_e Expected value of tardiness (*expectedTardiness*)
- t_c Current value of tardiness (*currentTardiness*)

Tardiness and Nondelayed Jobs (Tardiness + Nondelayed Jobs) Multi-Objective Function:

$$\begin{aligned}w_t &= (t_c - t_e)/t_c \\w_n &= (n_e - n_c)/n_c \\ \sum w &= w_t + w_n\end{aligned}\tag{49.8}$$

where

- w Total weight (*totalWeight*)
- w_t Tardiness weight (*weight_{tardiness}*)
- w_n Nondelayed jobs weight (*weight_{nondelayed}*)
- t_e Expected value of tardiness (*expectedTardiness*)
- t_c Current value of tardiness (*currentTardiness*)
- n_e Expected value of nondelayed jobs (*expectedNondelayed*)
- n_c Current value of nondelayed jobs (*currentNondelayed*)

A positive $weight_{makespan}$ indicates the new generated schedule has lower makespan than the current. It goes the same for positive $weight_{nondelayed}$ which means the new schedule expected to have less number of delayed jobs compared to current schedule. Similarly, when $weight_{tardiness}$ and $weight_{usage}$ are positive, it means that the new schedule has lower tardiness and higher machine usage than the current schedule. However, in case of multi-objective function that have clashing values, in example Makespan + Tardiness Multi Objective Function with positive value 0.5 of $weight_{makespan}$ and negative value -0.2 of $weight_{tardiness}$. The value for *totalWeight* will be 0.3 which happen to be positive and also means the new schedule will be accepted with the probability of total tardiness to be slightly high from the original schedule. Suppose the next cycle $weight_{makespan}$ has a negative value -0.2 and $weight_{tardiness}$ carry a positive value 0.5. This time the new schedule is also been accepted since the value of *totalWeight* is positive, with probability of new schedule makespan to be slightly high compared to previous schedule. From this example, the value of final *makespan metric* of Makespan + Tardiness Multi Objective Function might be slightly higher from single Makespan Objective Functions. Similarly, *total tardiness metric* of Makespan + Tardiness Multi-Objective Function might be higher from single Tardiness Objective Function since the first cycle accepted schedule with higher tardiness and second cycle accepted schedule with higher makespan. However, in terms of overall performance, the multi-objective function performance will be much better compared to single objective function because multi-objective consider more than one objective in making the decision and for the example given, the *makespan metric* and *total tardiness metric* for Makespan + Tardiness Multi-objective function is much lower compared to *makespan metric* of single Tardiness Objective Function and *total tardiness metric* of single Makespan Objective Function.

49.4 Performance Metrics

Total Tardiness: One of the main objectives of the scheduling procedure is the completion of all jobs before their agreed due dates. Failure to keep that promise has negative effects on the credibility of the service provider. The lateness of job

j can be defined as the difference between its completion time C_j and the corresponding due date d_j . This metric is known as the tardiness of job and is calculated using the following expression:

$$T_j = \max(0, C_j - d_j) \quad (49.9)$$

On the other hand, for total tardiness involving a set of n jobs, which are to be processed each in a single machine, the formula can be described as follows:

$$T = \sum_{j=1}^n \max(0, C_j - d_j) \quad (49.10)$$

$$C_j = S_j + p_j \quad (49.11)$$

where C_j is the completion time of job j and S_j is the start time of job j in machine m . Each job j has a processing time p_j , and a due date d_j .

Makespan: Makespan is a standard performance metric to evaluate scheduling algorithms. A small value of makespan means the scheduler is providing good and efficient planning of jobs to resources. The makespan of a schedule can be defined as the time it takes from the instance the first task begins execution to the instance at which the last task completes execution [25]. Makespan can be represented by the following equation:

$$C_{\max} = \max_{1 \leq j \leq n} C_j \quad (49.12)$$

In simple terms, makespan is the time it takes to finish the last job.

Flowtime: Flowtime, or also known as response time is the sum of completion times C_j of all the jobs [25]. Mathematically, flowtime can be formulated as:

$$F = \sum_{j=1}^n C_j \quad (49.13)$$

Flowtime and makespan always become two major objectives to be minimized in researches involving scheduling. However, minimization of makespan always results in the maximization of response time [23].

Delayed Jobs: Delayed Jobs stands for jobs that fail to meet their deadline or due date. Deadline is a period of time in which a job must be completed [5]. The goal typically in such problems is to complete the maximum number of jobs by their deadlines [3]. A higher machine usage fulfils resource owner's expectations, while a higher number of non delayed jobs guarantees a higher Quality of Service

(QoS) provided to the users [14, 16]. By reducing the number of delayed jobs, QoS for the system that uses the proposed scheduling technique will also be improved. Delayed jobs D is measured by:

$$D = \sum_{j=1}^n D_j \quad (49.14)$$

$$\text{where } D_j = \begin{cases} 1 & \text{if } (C_j > d_j) \\ 0 & \text{otherwise} \end{cases} \quad (49.15)$$

D_j is equal to 1 if job j is late ($C_j > d_j$). Otherwise D_j is equal to 0.

Machine Usage: Maximizing resource utilization or machine usage in the grid system is another important performance metrics. Utilization is the percentage of time that a resource is actually occupied, as compared with the total time that the resource is available for use. Low utilization means a resource is idle and wasted. Throughout this paper resource utilization will be referred as machine usage. According to [4, 22], Machine usage (MU) is computed as:

$$MU = \frac{CPU_{active}}{\min(CPU_{available}, CPU_{required})} \quad (16)$$

where CPU_{active} denotes the numbers of non idle CPU, $CPU_{available}$ denotes the number of available CPU, and $CPU_{required}$ represent the number of required CPU.

49.5 Results and Discussions

This section will present and discuss results obtained from the experimentation conducted to SH-PR-GSA schedulers based on different objective functions. However this experimentation only focuses on the situation where the Grid system has to manage a heavy load data since optimization is more efficient in this kind of environment. Simulation had been conducted using Alea [15] simulator with total number of 3,000 jobs and 150 machines.

Figures 49.2, 49.3, 49.4, 49.5 and 49.6 provide the performance results of SH-PR-GSA schedulers using different single and multi objective functions. In order to get a clear picture of differentiation between each objective functions, the results have been regrouped based on the objective function, and the percentage of difference for each objective function over makespan objective function

Fig. 49.2 Delayed jobs of SH-PR-GSA using different objective functions

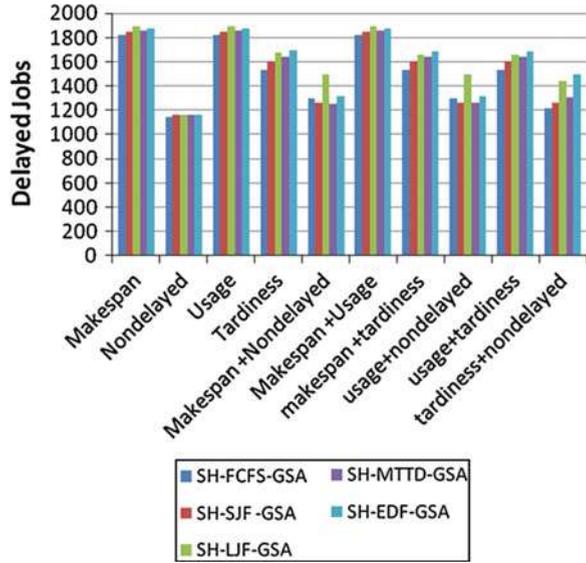


Fig. 49.3 Flowtime of SH-PR-GSA using different objective functions

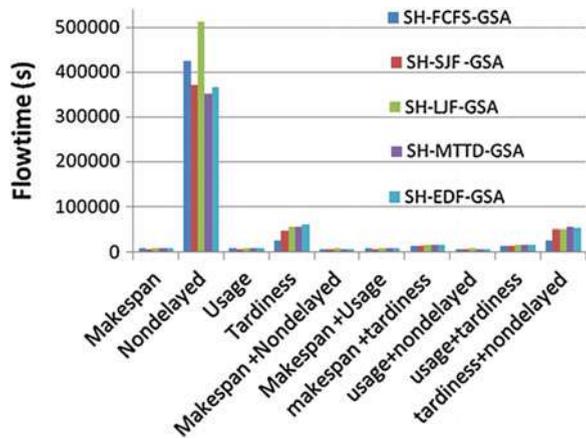


Fig. 49.4 Makespan of SH-PR-GSA using different objective functions

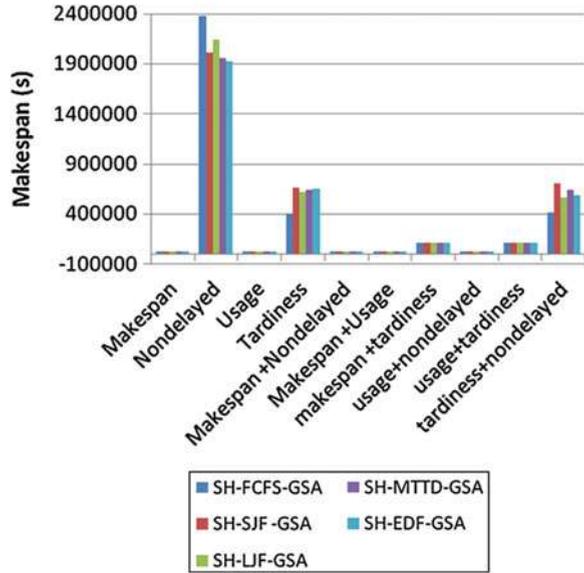


Fig. 49.5 Machine usage of SH-PR-GSA using different objective functions

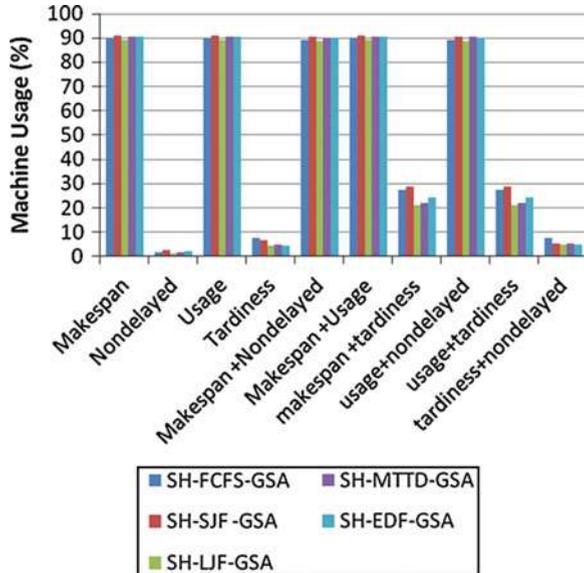


Fig. 49.6 Total tardiness of SH-PR-GSA using different objective functions

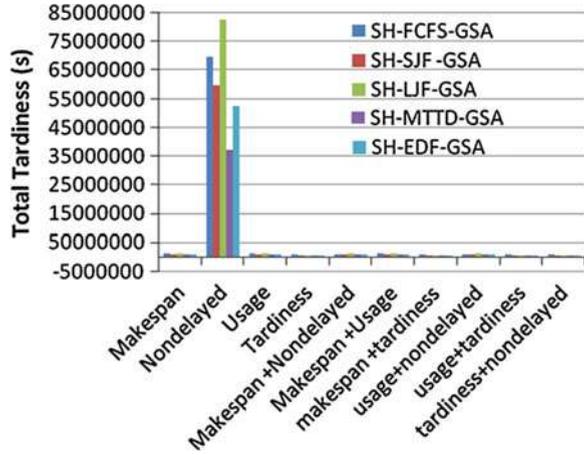


Fig. 49.7 SH-PR-GSA performance using nondelayed objective function: percentage of different compared with makespan objective function

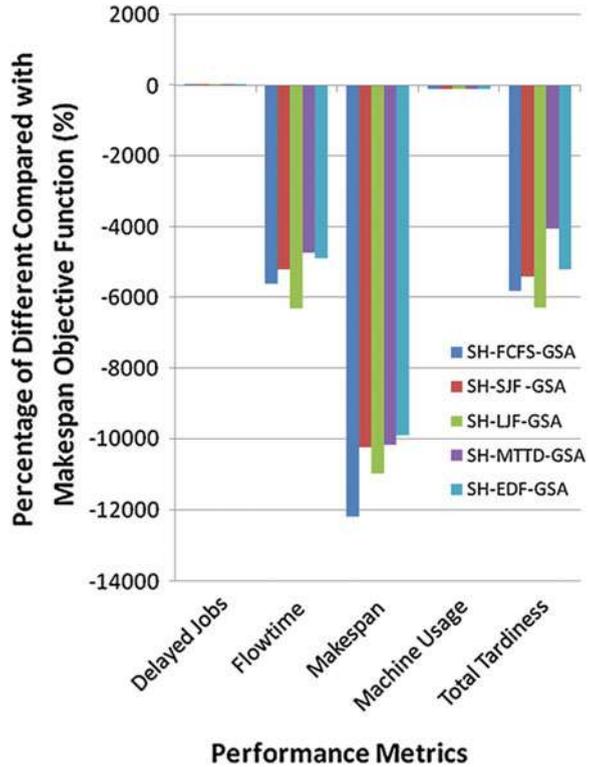
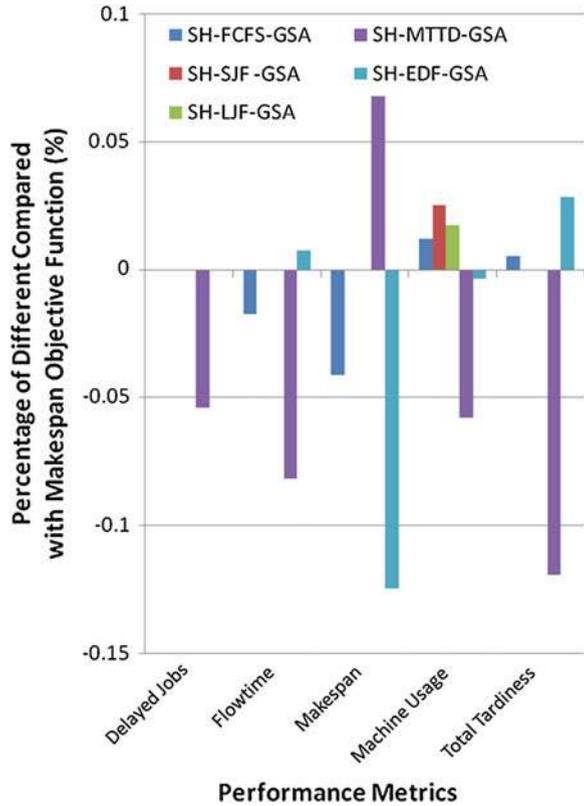
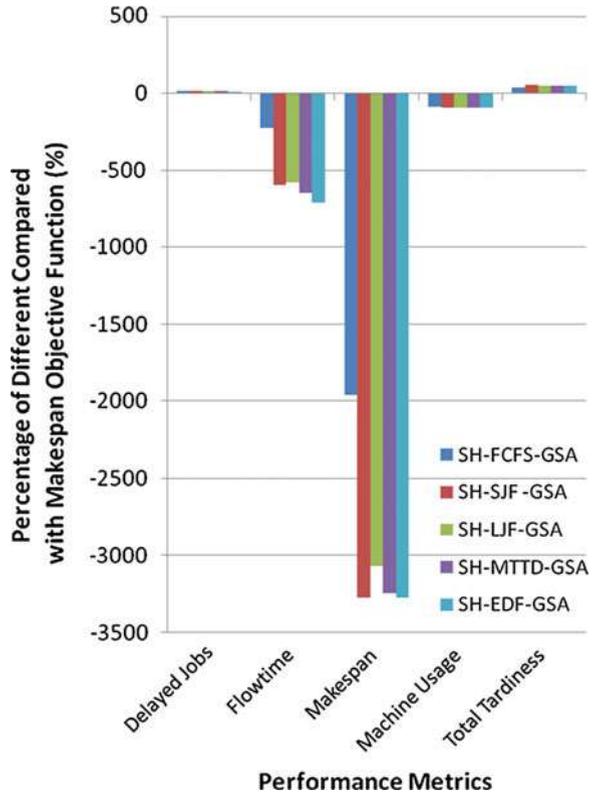


Fig. 49.8 SH-PR-GSA performance using machine usage objective function: percentage of different compared with makespan objective function



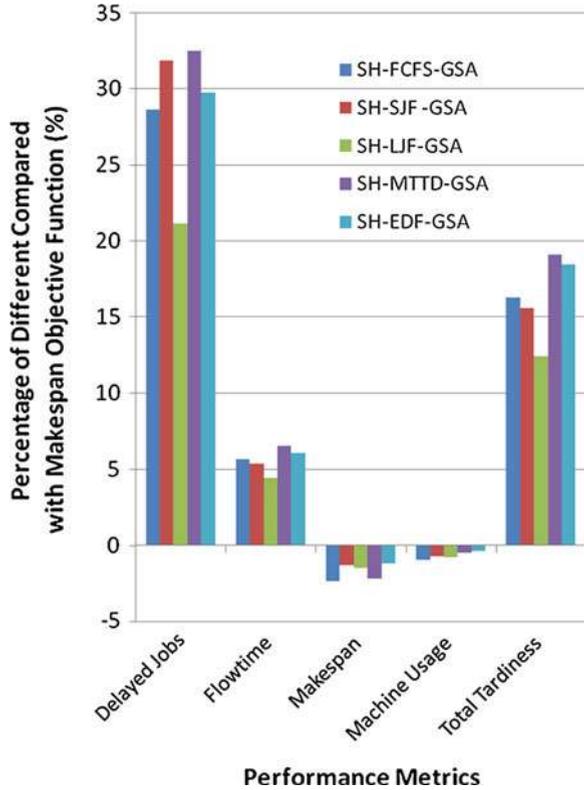
have been calculated. Results for single objective functions can be seen in Figs. 49.7, 49.8 and 49.9. Positive graph means the criteria for that particular objective function is better than makespan objective function while negative means it is worse than makespan objective function. From these results, it can be clearly noticed that if the applied objective function focuses on only one particular metric (i.e. tardiness), the benefit it gives is only to that particular criteria. Other metrics most probably will show significantly bad performances.

Fig. 49.9 SH-PR-GSA performance using tardiness objective function: percentage of different compared with makespan objective function



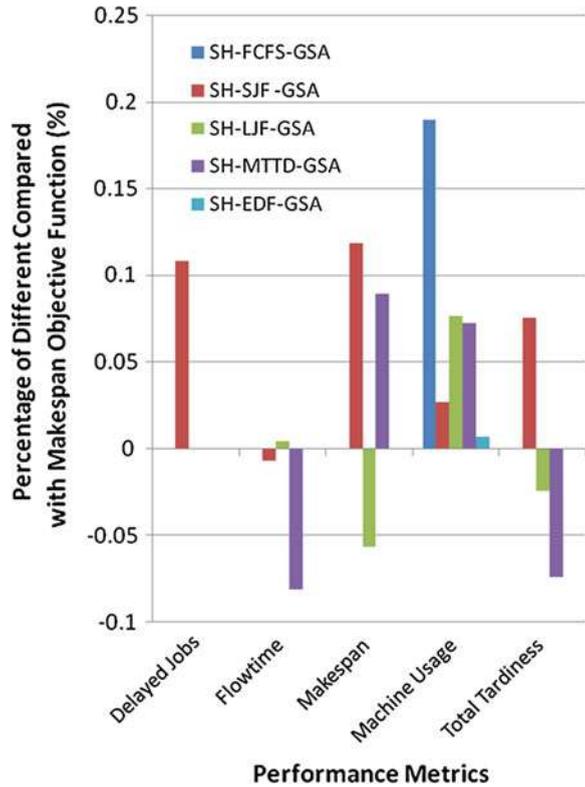
For instant, Fig. 49.9 which applied tardiness objective function in SH-PR-GSA show 36–55 % better than makespan objective functions in term of *total tardiness metric*. However, it suffers from a great disadvantages for other metrics especially *makespan metric* which is –1,900 to –3,300 % lower compared to makespan objective function. This is because by focusing on the tardiness, schedulers only approve new solutions if the new solution appears to be satisfying the tardiness objective without considering other metrics. By doing this, the total tardiness time can be reduced but the scheduler might accept more jobs that have earlier deadline

Fig. 49.10 SH-PR-GSA performance using makespan + nondelayed multi-objective function: percentage of different compared with makespan objective function



and reject jobs with much later deadline. This situation might lead to the blocking of many other jobs especially jobs that does not have specific deadlines (30 % of them in datasets). Hence, the very low percentage in *makespan metrics* justify that it is contributed mainly by jobs that have the longest processing time as well as jobs that does not have deadline. This situation is the same for all single objective functions including makespan objective functions. If makespan objective function

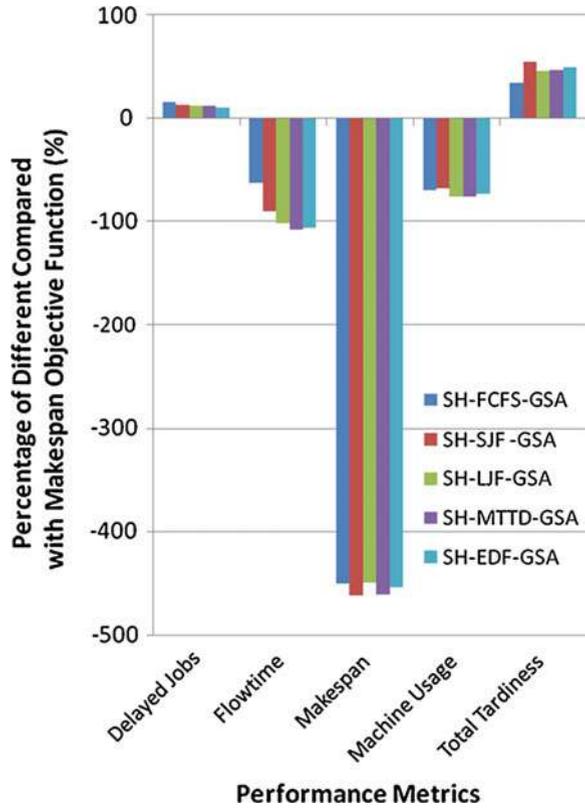
Fig. 49.11 SH-PR-GSA performance using makespan + machine usage multi-objective function: percentage of different compared with makespan objective function



were applied, except for makespan, other matrices might not show their best performance.

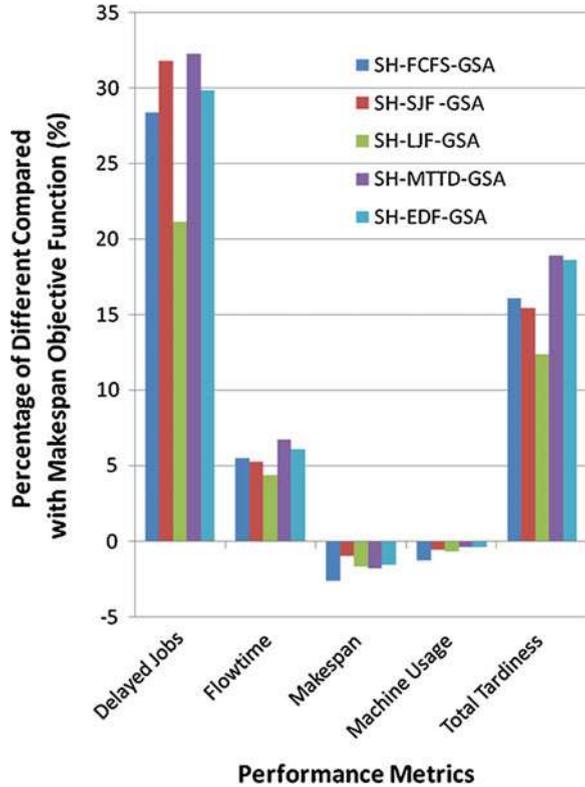
When multi-objective is adapted to the scheduler (Figs. 49.10, 49.11, 49.12, 49.13, 49.14 and 49.15), this situation changes dramatically. The multi-objective function will try to balance the performance between two metrics and give a much better tradeoff to the others compared to single objectives. However an example such as Fig. 49.12 that uses multi objective function does not show as good results as expected. For multi-objective function that combines makespan and tardiness criteria, there are significant drops of performance for flowtime, makespan and machine usage metrics compared to Makespan Objective Function. However, if compared to the results of the single Tardiness Objective Function (previously

Fig. 49.12 SH-PR-GSA performance using makespan + tardiness multi-objective function: percentage of different compared with makespan objective function



discussed Fig. 49.6), actually the performance of Makespan + Tardiness Multi Objective function has recorded significant improvement. For example, with the Tardiness Objective Function, makespan for SH-FCFS-SA is -1,962 % worse than the makespan recorded for makespan objective function. After integrating tardiness with makespan in a multi-objective function (Fig. 49.12), the SH-FCFS-GSA makespan is now -450 %, which is 1,512 % better compared to a single Tardiness Objective Function while at the same time also still preserving the *total tardiness metrics* at 34 %. Comparing between multi objective functions, as for Makespan + Machine Usage Multi Objective Function (Fig. 49.11), some of the performance metrics is slightly better than Makespan + Tardiness Multi Objective

Fig. 49.13 SH-PR-GSA performance using machine usage + nondelayed multi-objective function: percentage of different compared with makespan objective function



Function, but it is not really stable for all the schedulers and the improvement made is too little (below 0.2 %). The most perfect combination is Make-span + Nondelayed Multi Objective Function as shown in Fig. 49.10 which also implemented by EG-EDF-TABU in [16]. Three out of five performance metrics show significant improvement. The tradeoff for the other two, makespan and machine usage is acceptably low. It is also stable compared to the previously explained combinations.

Fig. 49.14 SH-PR-GSA performance using machine usage + tardiness multi-objective function: percentage of different compared with makespan objective function

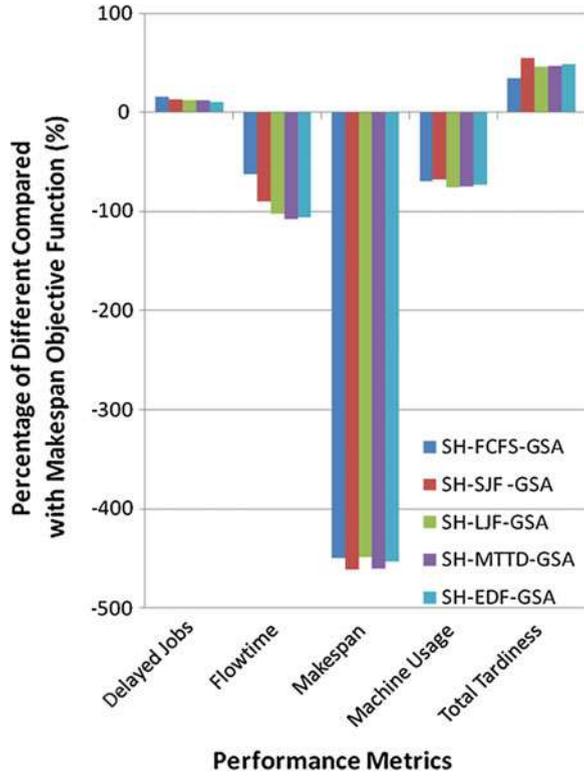
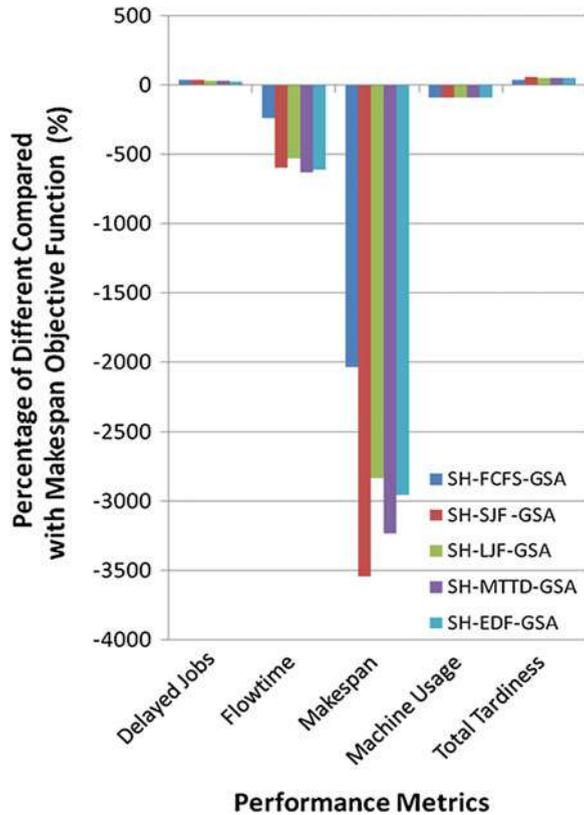


Fig. 49.15 SH-PR-GSA performance using tardiness + nondelayed multi-objective function: percentage of different compared with makespan objective function



49.6 Conclusions

In this paper, analysis on the effects of objective functions has been made by testing the Grid scheduler with various single and multi-objective functions. This analysis is important to show that a bad choice of objective function can extremely affect the scheduler. The best objective function recognized after completing the experimentation is Makespan + Nondelayed Multi Objective Function. This objective function provides the scheduler with a tolerance tradeoff between five criteria's; makespan, flowtime, machine usage, delayed jobs and tardiness.

Acknowledgement This research was supported by Universiti Malaysia Pahang Research Grant (RDU1203116)

References

1. Albert, Y.Z.: Observations on using genetic algorithms for dynamic load-balancing. *IEEE Trans. Parallel Distrib. Syst.* **12**, 899–911 (2001)
2. Altıparmak, F., Gen, M., Lin, L., Paksoy, T.: A genetic algorithm approach for multi-objective optimization of supply chain networks. *Comput. Ind. Eng.* **51**(1), 196–215 (2006)
3. Bansal, N., Chan, H.-L., Lam, T.-W., Lee, L.-K.: Scheduling for speed bounded processors. In: Aceto, L., Damgård, I., Goldberg, L., Halldórsson, M., Ingólfssdóttir, A., Walukiewicz, I. (eds.) *Automata, Languages and Programming*, pp. 409–420. Springer, Berlin (2008)
4. Baraglia, R., Dazzi, P., Capannini, G., Pagano, G.: A multi-criteria job scheduling framework for large computing farms. In: 2010 IEEE 10th International Conference on Computer and Information Technology (CIT) (2010)
5. Brucker, P.: *Scheduling Algorithms*, 5th edn. Springer, Berlin (2007)
6. Carretero, J., Xhafa, F.: Using genetic algorithms for scheduling jobs in large scale grid applications. *J. Technol. Econ. Dev.* **12**, 11–17 (2006)
7. Casanova, H.: Distributed computing research issues in grid computing. *SIGACT News* **33**(3), 50–70 (2002)
8. Collignon, T.P., van Gijzen, M.B.: Minimizing synchronization in IDR (s). *Numer. Linear Algebra Appl.* **18**, 805–825 (2011)
9. Cooper, K., Dasgupta, A., Kennedy, K., Koelbel, C., Mandal, A., Marin, G., Mazina, M., Mellor-Crummey, J., Berman, F., Casanova, H., Chien, A., Dail, H., Liu, X., Olugbile, A., Sievert, O., Xia, H., Johnsson, L., Liu, B., Patel, M., Reed, D., Deng, W., Mendes, C., Shi, Z., YarKhan, A., Dongarra, J.: New grid scheduling and rescheduling methods in the GrADS project. In: *Proceedings of 18th International Parallel and Distributed Processing Symposium* (2004)
10. Dickmann, F., Falkner, J., Gunia, W., Hampe, J., Hausmann, M., Herrmann, A., Kepper, N., Knoch, T.A., Lauterbach, S., Lippert, J., Peter, K., Schmitt, E., Schwardmann, U., Solodenko, J., Sommerfeld, D., Steinke, T., Weisbecker, A., Sax, U.: Solutions for biomedical grid computing—Case studies from the D-Grid project Services@MediGRID. *J. Comput. Sci. In Press*, Corrected Proof (2011)
11. Entezari-Maleki, R., Movaghar, A.: A genetic-based scheduling algorithm to minimize the makespan of the grid applications, in grid and distributed computing, control and automation. In: Kim, T.-h., Yau, S., Gervasi, O., Kang, B.-H., Stoica, A., Ślęzak, D. (eds.), pp. 22–31. Springer, Berlin (2010)
12. Farzi, S.: Efficient job scheduling in grid computing with modified artificial fish swarm algorithm. *Int. J. Comput. Theory Eng.* **1**(1), 1793–8201 (2009)
13. Izakian, H., Abraham, A., Snášel, V.: Metaheuristic based scheduling meta-tasks in distributed heterogeneous computing systems. *Sensors* **9**(7), 5339–5350 (2009)
14. Klusacek, D., Rudova, H.: Improving QoS in computational grids through schedule-based approach. In: *Scheduling and Planning Applications Workshop at the Eighteenth International Conference on Automated Planning and Scheduling (ICAPS 2008)*: Sydney, Australia (2008)
15. Klusacek, D., Rudova, H.: Alea 2: job scheduling simulator. In: *Proceedings of the 3rd International ICST Conference on Simulation Tools and Techniques. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering): Torremolinos, Malaga, Spain*. pp. 1–10 (2010)
16. Klusacek, D., Rudová, H., Baraglia, R., Pasquali, M., Capannini, G.: Comparison of multi-criteria scheduling techniques. In: Gorlatch, S., Fragopoulou, P., Priol, T. (eds.) *Grid Computing*, pp. 173–184. Springer, US (2008)
17. Komisarczuk, P., Welch, I.: Internet sensor grid: experiences with passive and active instruments. In: Pont, A., Pujolle, G., Raghavan, S. (eds.) *Communications: Wireless in Developing Countries and Networks of the Future*, pp. 132–145. Springer, Boston (2010)

18. Leung, J.Y.-T.: *Handbook of Scheduling: Algorithms, Models and Performance Analysis*. CRC Press, Boca Raton (2004)
19. Liu, H., Abraham, A., Hassaniien, A.E.: Scheduling jobs on computational grids using a fuzzy particle swarm optimization algorithm. *Future Gener. Comput. Syst.* **26**(8), 1336–1343 (2010)
20. Oluwatope, A., Iyanda, D., Aderounmu, G., Adagunodo, R.: Computational modeling of collaborative resources sharing in grid system. In: Dua, S., Sahni, S., Goyal, D.P. (eds.) *Information Intelligence, Systems, Technology and Management*, pp. 311–321. Springer, Berlin (2011)
21. Pandey, S., Buyya, R.: Scheduling of scientific workflows on data grids. In: 8th IEEE International Symposium on Cluster Computing and the Grid, 2008. CCGRID '08 (2008)
22. Pasquali, M., Baraglia, R., Capannini, G., Ricci, L., Laforenza, D.: A multi-level scheduler for batch jobs on grids. *J. Supercomput* **57**(1), 81–98 (2011)
23. Subashini, G., Bhuvaneshwar, M.C.: Non dominated particle swarm optimization for scheduling independent tasks on heterogeneous distributed environments. *Int. J. Adv. Soft Comput. Appl.* **3**(1), 1–17 (2011)
24. Vazquez, M., Whitley, D.: A comparison of genetic algorithms for the static job shop scheduling problem. In: Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J., Schwefel, H.-P. (eds.) *Parallel Problem Solving from Nature PPSN VI*, pp. 303–312. Springer, Berlin (2000)
25. Xhafa, F., Abraham, A.: Meta-heuristics for grid scheduling problems. In: Xhafa, F., Abraham, A. (eds.) *Metaheuristics for Scheduling in Distributed Computing Environments*, pp. 1–37. Springer, Berlin (2008)
26. Xiao-Juan, W., Chao-Yong, Z., Liang, G., Pei-Gen, L.: A survey and future trend of study on multi-objective scheduling. In: *Fourth International Conference on Natural Computation, 2008. ICNC '08* (2008)
27. Xue, X.D., Cheng, K.W.E., Ng, T.W., Cheung, N.C.: Multi-objective optimization design of in-wheel switched reluctance motors in electric vehicles. *IEEE Trans. Ind. Electron.* **57**(9), 2980–2987 (2010)
28. Yang, Y., Wu, G., Chen, J., Dai, W.: Multi-objective optimization based on ant colony optimization in grid over optical burst switching networks. *Expert Syst. Appl.* **37**(2), 1769–1775 (2010)

Chapter 50

Experimental Analysis on Available Bandwidth Estimation Tools for Wireless Mesh Network

Imran Edzereiq Kamarudin, M.A. Ameen
and Zafril Rizal M. Azmi

Abstract Measurement of available bandwidth in a wireless mesh network (WMN) environment has always been a topic of great interest. Several active and passive-based tools has been tested and proposed in previous research. However, the performance of these tools was never tested extensively in terms of the condition of the WMN such as varying bandwidth across the network and external traffic factors. In this work, we perform an extensive experimental analysis study on both active and passive available bandwidth tools by looking at the accuracy, failure rate and consistency of each tool. We also investigate the effects of varying the WMN bandwidth and external traffic on the performance of these tools. Our results indicate that all tools performance was affected by the WMN testing environment. In term of accuracy, failure rate and consistency, Pathload was the most favorable tool in these conditions.

Keywords WMN · ASSOLO · Pathchirp · IGI/PTR · Pathload · Pathrate · Wbest

50.1 Introduction

Wireless mesh network (WMN) has been a standard deployment spanning from large enterprise network to small home network. From the beginning with 802.11a which only support up to 2 Mbps of data stream, up to the latest 802.11ac which

I.E. Kamarudin (✉) · M.A. Ameen · Z.R.M. Azmi
Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang,
Malaysia
e-mail: edzereiq@ump.edu.my

M.A. Ameen
e-mail: mohamedariff@ump.edu.my

Z.R.M. Azmi
e-mail: zafril@ump.edu.my; zafrilrizal@yahoo.com

supports up to 1 Gbps, the rapid growth of data stream was achieved in less than 5 years' time frame only. Each new standard defined in the 802.11 protocol show significant improvement in terms of the maximum bandwidth supported. For example, 802.11b supports up to 11 Mbps whereas the later improve 802.11 g supports up to 54 Mbps.

A common mistake done by user is to assume that the actual bandwidth supported by the WMN is based on the protocol itself. For example, in 802.11 g wireless network, it supports constantly at 54 Mbps. As a result, this assumption is wrong and may lead to underestimate the total bandwidth require by the network. In order to correctly measure the actual bandwidth, the concept of available bandwidth is introduced.

As mentioned before, bandwidth capacity refers to the maximum data or throughput that can be transmitted on a link or a medium. It is important to understand and identify the maximum throughput of a link in network planning to cater the needs of the end user or end nodes. The available bandwidth (ABW) at a link is its unused capacity. Since, at any time, a link is either idle or transmitting packets at the maximum speed, the definition of the available bandwidth ought to look at the average unused bandwidth over some time interval T . Thus,

$$A_i(t, T) = \frac{1}{T} \int_t^{T+t} (C_i - \lambda_i(t)) dt \quad (50.1)$$

where $A_i(t; T)$ is the available bandwidth at link i at time t , C_i is the link's capacity, and λ_i is its traffic. The available bandwidth along a path is the minimum available bandwidth of all traversed links.

To measure the AWB in WMN, there are two techniques to estimate—passive and active measurement. Passive measurement is performed by observing the traffic without intruding the network. Active measurement on the other hand, will probe the network by generating packet traffic into the network to perform the measurement. Basically, all available bandwidth tools are created using either one technique. Several recent works was done on both type of AWB but mainly focusing on only one type of technique [1, 2]. As a result, comparison between active and passive tools is not clear. Furthermore, testing was done purely on the network without any external traffic to look at the effect of AWB when other traffic is generated in the WMN [3–5].

The objective of our study is to evaluate existing active and passive AWB tools and study their performance for 802.11-based WMN. We evaluate these tools by varying various parameters such as physical data rate and the existing of other traffic in the network. Our main contributions are highlighted as below:

- We select three AWB tools from both active and passive technique, and evaluate their performance in WMN.
- We perform extensive testing in terms of the WMN physical rate and the condition of the WMN (with and without external traffic other than the AWB tools)

- We evaluate the tools based on: (1) Accuracy: This will measure the accuracy of the tool to estimate the available bandwidth whether it will over estimate or under estimate the available bandwidth. (2) Failure patterns: This attribute will monitor and measure the reliability of the tool's failure or error prone to estimate the bandwidth throughout the testing cycle. (3) Consistency of measurement: This attribute will measuring the consistency of the measurement of the tool as whether it will fluctuate of over estimating or under estimating the bandwidth.

The rest of the paper is organized as follows. Section 50.2 lists the previous works and explains the foundation and motivation behind our work. Section 50.3 outlined the details of our tools, testbed settings and WMN environment for testing and evaluation methodology. In Sect. 50.4, we present the results of the comparison study for both WMN environments. Section 50.5 concludes the paper.

50.2 Foundation

In this chapter, preliminary information regarding the technologies and the testbed environment used to undertake this research is introduce in order for the reader to easily understand the contents of this paper.

50.2.1 Active-Based AWB Tools

Active-based AWB tool is based on the idea of induced congestion, in which probe packets are sent at increasing rates. At the receiver, the delays of the probe packets are measured to determine the point at which they start increasing in a consistent basis. The available bandwidth is then estimated by looking at the probe packet rate utilized when the turning point is found. PathChirp [3], ASSOLO [6], and IGI/PTR [5] are examples of tools utilizing this approach.

PathChirp [3] sends a variable bit-rate stream consisting of exponentially spaced packets. The actual unused capacity is inferred from the rate responsible for increasing delays at the receiver side. During probing process, Pathchirp increases the probing rate within each chirp (having a variable number of packets) in an exponential manner. By doing that, it captures delay correlation information using a smaller number of probing packets. Pathchirp also uses variable size probe packets of minimum 40 bytes.

ASSOLO [3] is a tool based on the same principle, but it features a different probing traffic profile and uses a filter to improve the accuracy and stability of results. It is based on the concept of "self-induced congestion". It features a

replacement inquisitor traffic profile known as REACH (Reflected Exponential Chirp) which tests a large range of rates being a lot of accurate in the center of the inquisitor interval. Moreover, the tool runs inside a real-time operating system and uses some de-noising techniques to improve the activity method.

IGI/PTR [5, 7] obtain the most accurate measurement when the packet-train sending rate at source equals its inbound rate at destination, where the initial packet combine gap that provides a high correlation between the packet gap changes and also the competing traffic throughput on the tight link. Both IGI and PTR share the probing procedure, the main difference between them is that IGI focuses on calculating background traffic load, whereas PTR directly calculates packet transmission rate, to estimate end-to-end available bandwidth.

50.2.2 Passive-Based AWB Tools

In passive-based AWB tool also known as Probe Gap Model (PGM), packet pairs (or packet trains) are sent to the path at a single rate. This probing rate is set to the capacity of the tight link, and so it is larger than (or equal to) the available bandwidth of the path. Pathload [5, 8], Pathrate [9] and WBest [10] are examples of tools utilizing this approach.

Pathload [5, 8] estimate the available bandwidth of an end-to-end path from a host S (sender) to a host R (receiver). The available bandwidth is the maximum IP-layer throughput that a flow can get in the path from S to R, without reducing the rate of the rest of the traffic in the path. The basic idea in Pathload is that the one-way delays of a periodic packet stream show increasing trend, when the stream is larger than the available bandwidth. The measurement algorithm is iterative and it requires the cooperation of both Sender and Receiver. Pathload consists of a process running at S and a process running at R. S sends periodic streams of UDP packets from S to R at a certain rate.

Pathrate [9] uses packet pairs and packet-trains, in conjunction with statistical techniques, to estimate the capacity of the narrow link in the path. It uses UDP for transferring probing packets. Additionally, Pathrate establishes a TCP connection, referred to as control channel, between the sender and the receiver. The control channel acknowledges each correctly received packet pair or train, and it also transfers commands sent from the receiver to the sender regarding the probing packet size, train length, number of measurements, train spacing, etc.

WBest [10] is a two stage algorithm: (1) a packet pair technique estimates the effective capacity over a flow path where the last hop is a wireless LAN (WLAN); and (2) a packet train technique estimates achievable throughput to infer the available bandwidth. WBest parameters are optimized given the tradeoffs of accuracy, intrusiveness and convergence time.

50.3 Comparison of Existing Tools

In this section, we describe the WMN testbed environment used and the validation methodology.

50.3.1 Tools and WMN Testbed Environment

We selected three tools each from active: ASSOLO, IGI/PTR, Pathchirp and passive; Pathload, Pathrate, WBest probe base tools. For evaluation, we created a testbed in a close environment consisting of 802.11 g wireless devices and access point (AP). This was to make sure that no other wireless network signals can interfere with out test environment. Two wireless nodes consist of two Linux-based laptops were setup in order to run the tools. To create the 802.11 g WMN environment, two APs were used. The settings for APs were set as default except for data rate.

In summary, there were two conditions in WMN environments for testing: (1) WMN without any external traffic (2) WMN with external traffic. For the first condition, the tools were tested in an optimum environment where the only traffic was from the tools itself. For the second condition, FTP traffic was generated on top of the tools itself. Both condition were then tested in three different settings (Fig. 50.1).

- One AP with two nodes—Tools was tested across a single AP as shown in Fig. 50.2.
- Two AP with two nodes—Tools was tested across two AP via bridging. Bandwidth between APs remain the same as shown Fig. 50.3.
- Two AP with two nodes and variable bandwidth—Similar with above except that the bandwidth between APs was lower compare to the nodes. This setting helps to look at the impact of changing of bandwidth across the network towards AWB tools.



Fig. 50.1 One AP with two nodes

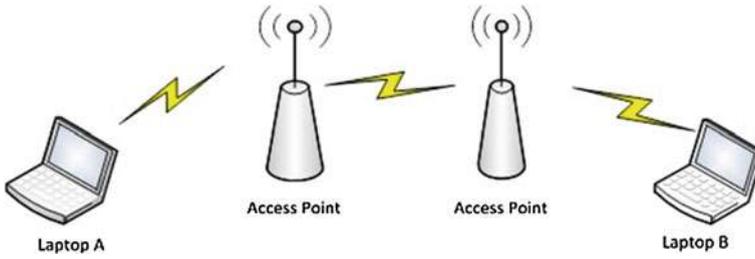


Fig. 50.2 Two APs with two nodes

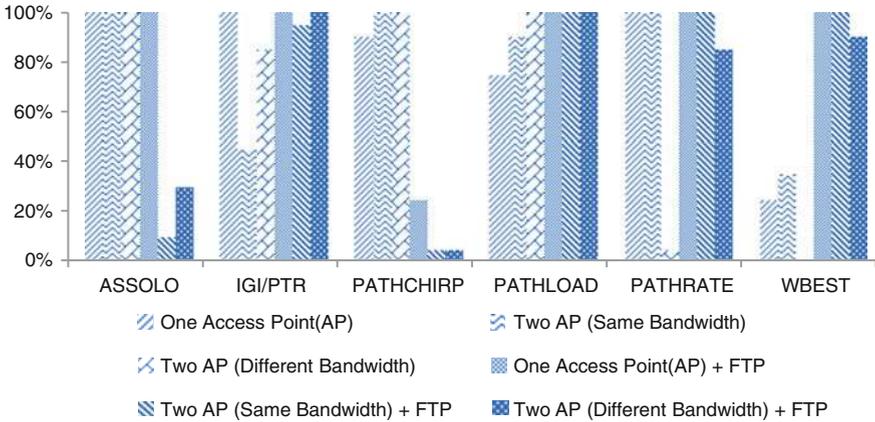


Fig. 50.3 Accuracy comparisons

50.3.2 Testing and Evaluation Methodology

For each of the selected tools, we ran a few experiments just to make sure that all tools has the same default parameter such as probe-size and number of probes. This was an important step to make sure that all tools were tested at the same rate. We observed that some tools where having higher numbers of probe. If some minor changers, we managed to leverage all tools into the same settings in terms of probe size and number of probes.

Each tool was then tested in terms of accuracy, failure pattern and consistency across the two condition WMN environment with three different settings. For sampling purposes, 20 reading were taken for each of the tools. The summary of test conducted as below:

- WMN environment without external traffic: (1) One AP with two nodes (2) Two AP with two nodes (same bandwidth) (3) Two AP with two nodes (variable bandwidth)

- WMN environment with external traffic: (1) One AP with two nodes (2) Two AP with two nodes (same bandwidth) (3) Two AP with two nodes (variable bandwidth)

50.4 Experimental Results

50.4.1 Accuracy

For this experiment, we measured the accuracy of the tool to estimate the available bandwidth whether it will over estimate or under estimate the available bandwidth. For the available bandwidth, although the capacity is 54 Mbps, the achievable available bandwidth will not reach to that level. It is fairly common for 802.11 g connections to run at 36, 24 Mbps, or even lower [11]. Recommended data rate might be close to 45 Mbps [12]. Hence, in our assessment, the benchmark accuracy of available bandwidth was set at the range of 8–45 Mbps. So the reading that falls into this range will be considered as accurate.

However, we took two other factors into consideration: (1) The under and overestimation of the tool to estimate the bandwidth. Overestimation in this case means that the tool overestimated the bandwidth of the 802.11 g bandwidth which is 54 Mbps. (2) For simulation with external traffic, any reading that falls below the average of the same simulation without external traffic were omitted

$$Accuracy [within range 8 - 54 Mbps] = \frac{Number\ of\ times\ within\ range}{20} \times 100\ % \quad (50.2)$$

Based on our findings on WMN environment without external traffic, ASSOLO was the only tool able to produce 100 % accurate reading within the benchmark bandwidth (8–54 Mbps) for all three settings. In contrast, WBest had the worst accuracy rate at below 35 %. We found that WBest had a very high rate of failure in detecting available bandwidth.

When testing on WMN environment with external traffic, Pathload were able to produce 100 % accurate reading within the benchmark for all three settings. In contrast, Pathchirp had the lowest percentage at 12 %.

In summary when looking at the average accuracy rate across all environments and settings, Pathload provided the most consistent accuracy of 94 % while Pathchirp had the worst consistency rate at 54 %. We also found that other than Pathload, the rest of the tools tend to give inaccurate readings especially when tested with external traffic and variable bandwidth settings. Figure 50.4 shows the accuracy rate for all tools with the respective testing environment.

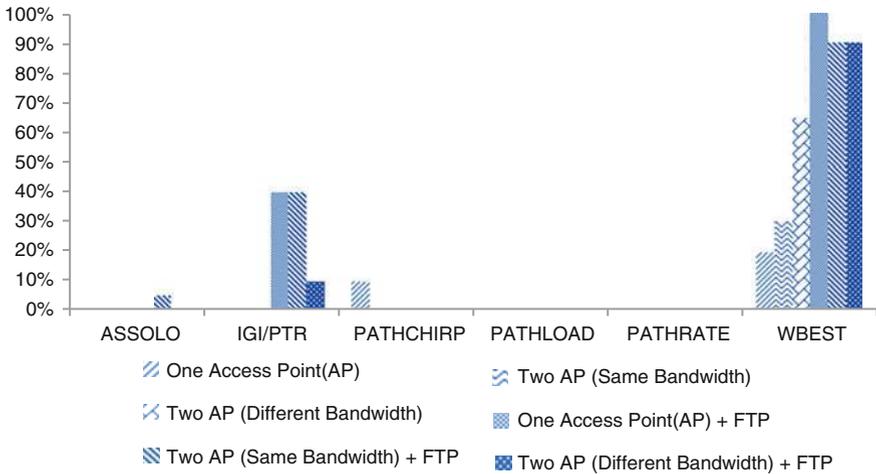


Fig. 50.4 Failure rate comparison

50.4.2 Failure Patterns

The second attributes which is the failure pattern, where we measured and evaluated the reliability of the tool to measure the bandwidth as whether it will succeed to measure the bandwidth under the 802.11 g wireless environment, or otherwise fail to measure. The reading ‘0’ or negative value indicated that the tool failed to estimate the bandwidth. The percentage will be calculated and compared among the tools.

$$Failure\ pattern = \frac{number\ of\ 0\ readings}{20} \times 100\% \tag{50.3}$$

Based on our findings on WMN environment without external traffic, WBest had the highest failure rate at 65 %. ASSOLO, Pathrate and Pathload recorded 0 % failure rate. When we tested on WMN environment with external traffic, WBest failure rate was at 90 %. The rests of the tools were below 40 %. In summary, WBest had the highest average failure rate across all environment and settings with 65 % failure rate while ASSOLO, Pathrate, Patchirp and Pathload recorded less than 2 % average in failure rate. In addition, we found that apart from Pathrate and Pathload, the rest of the tools were affected when tested with WMN environment with external traffic and variable bandwidth settings. Figure 50.4 shows the accuracy rate for all tools with the respective testing environment.

Table 50.1 Standard deviation for all tools

Consistency	ASSOLO	IGI/ PTR	PATHCHIR	PATHLOAI	PATHRATE	WBEST
One access point (AP)	0.0432	1.3982	1.2806	0.6401	0.2421	4.0357
One access point (AP) + FTP	1.2818	0.6821	0.6577	1.3035	0.7891	0.0000
Two AP (Same bandwidth)	1.2187	0.4480	0.3593	0.6621	0.5174	3.8013
Two AP (Same bandwidth) + FTP	0.4510	2.5232	0.6448	0.6654	0.8035	0.9292
Two AP (Different bandwidth)	0.4396	0.4005	0.4656	0.4859	0.7891	0.6540
Two AP (Different bandwidth + FTP)	1.4857	0.5861	0.3656	0.3677	1.0867	1.1005

50.4.3 Consistency

In this assessment, a calculation of mean and standard deviation was done. The calculation was done using the Eqs. 50.4 and 50.5 below. The lower the value of standard deviation indicates that the tool’s consistency in measuring the bandwidth is better. In order to calculate the standard deviation, the mean value needs to be obtained first.

$$\mu = \frac{\sum_{k=0}^n xk}{n} \tag{50.4}$$

$$\sigma = \sqrt{\frac{\sum_{k=1}^n (Xk - \mu)^2}{n}} \tag{50.5}$$

Based on our findings on WMN environment without external traffic, we found that ASSOLO recorded the lowest value at 0.0432. However, when comparing the average value for all three settings, Pathrate were able to produce the lowest value at 0.5162. Wbest has the highest value at an average of 2.8. When tested on WMN environment with external traffic, we found that Pathchirp had the lowest value at 0.3656. Pathchirp also gave the lowest average value at 0.566 for all three settings. ASSOLO and Wbest produce the highest value at 1.072 and 1.012 respectively.

In summary, we found that the effect of external traffic was great especially on ASSOLO, IGI/PTR and Pathrate. Their readings were mostly overestimated. The worst effect was on WBest, which recorded majority with zero readings. In the other hand, Pathload and Pathrate were able to give a more consistence reading throughout the testing as shown in Table 50.1. Figure 50.5 shows the consistency comparison. The space constraint allows us to display only one graph for comparison of consistency between the tools. For a more complete set of results, the readers may contact the lead author.

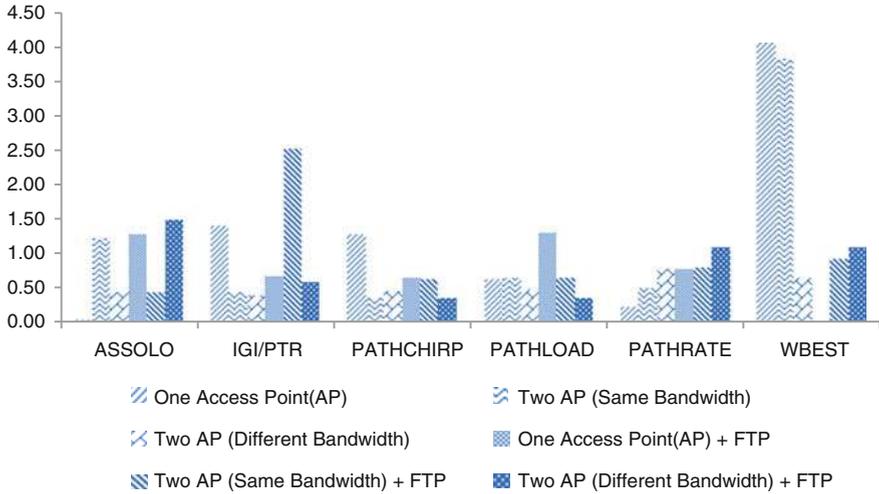


Fig. 50.5 Consistency comparison

50.5 Conclusion

In this paper, we have tackled the problem of estimating available bandwidth in WMN environment. Several available bandwidth tools have been proposed previously but were lacking in certain testing environment. Some of the recent works have proposed of using passive technique for estimation of bandwidth but what not tested particularly on WMN network with the presents of external traffic and varying bandwidth rate across the WMN. We implement our testing approach and look at the performance in terms of accuracy, failure rate and consistency.

The results from our experiments suggest that Pathload is the most favorable tool to estimate available bandwidth across WMN. This is because even when external traffic is present in the WMN, Pathload was able to provide higher accuracy, lower failure rate and consistence reading throughout our testing. It was also observed that using any tools when tested over WMN with the presence of external traffic and variable bandwidth causes further degradation in their performance.

Acknowledgments This research was supported by Universiti Malaysia Pahang Research Grant (RDU120388).

References

1. Gupta, D., et al.: Experimental comparison of bandwidth estimation tools for wireless mesh networks. IEEE (2009)
2. Montesino-Pouzols, F.: Comparative analysis of active bandwidth estimation tools passive and active network measurement. In: Barakat, C., Pratt, I. (eds.) pp. 175–184. Springer Berlin, Heidelberg (2004)
3. Guerrero, C.D., Labrador, M.A.: On the applicability of available bandwidth estimation techniques and tools. *Comput. Commun.* **33**(1), 11–22 (2010)
4. Botta, A., Pescape, A., Ventre, G.: On the performance of bandwidth estimation tools. IEEE (2005)
5. Johnsson, A., Bjorkman, M., Melander, B.: An analysis of active end-to-end bandwidth measurements in wireless networks. IEEE (2006)
6. Goldoni, E., Rossi, G., Torelli, A.: Assolo, a new method for available bandwidth estimation. pp. 130–136 (2009)
7. Kayange, Daniel S., Ramadhani Sinde, A.S.: available bandwidth estimation techniques (ABETS) For An Efficient Telemedicine Content Transport Network. *Int. J. Eng. Res. Technol (IJERT)*, **2**(7) (2013)
8. Kayange, Daniel S., Ramadhani Sinde, A.S.: Pathload for available bandwidth estimation techniques (ABETS) for an efficient telemedicine content transport network. *Int. J. Advancements in Res. Technol.* **2**(8), 6 (2013)
9. Prasad, R., Dovrolis, C., G.I.o.T.: Bandwidth estimation: metrics, measurement techniques, and tools. *Network IEEE.* **17**(6), 27–35 (2003)
10. Mingzhe, L., Claypool, M., Kinicki, R.: WBest: a bandwidth estimation tool for IEEE 802.11 wireless networks. IEEE (2008)
11. Seide, R.: Capacity, coverage, and deployment considerations for IEEE 802.11 g. cisco Systems white paper, San Jose (2003)
12. Prasad, R., et al.: Bandwidth estimation: metrics, measurement techniques, and tools. *IEEE Network* **17**(6), 27–35 (2003)

Chapter 51

A Survey of Petri Net Tools

Weng Jie Thong and M.A. Ameen

Abstract Petri net is a mathematical modeling language used to describe a system graphically. It is a strong language that can be used to represent parallel or concurrent activities in a system. With a Petri net tool, users can view the overall system graphically and edit it with the editor. A Petri net tool can also be used to analyze the performance of the system, generate code, simulate the system and perform model checking on it. A review on twenty Petri net tools in this paper will give the readers an idea on what is a Petri net tool and the main functions of a Petri net tool. This paper can serve as an introduction of twenty Petri net tools to the reader. However, to date, there are many Petri net tools available to be downloaded online. This survey paper aims to compare twenty Petri net tools in different aspect. This is important as users will not have to read the reviews of Petri net tools online one by one. Just by having a look at the discussion provided, readers can determine the best recommended Petri net tools to be used based on their operating systems and the types of Petri net tool to be analyzed. The main purpose of this survey paper is to recommend Petri net tools based on the operating system and the types of Petri net to be analyzed.

Keywords Petri net · Petri net tools

51.1 Introduction

Petri net were introduced by Petri in 1962 [1]. Petri designed a sequence of modules, with each module containing a single data element and communicating with its two neighbors [2]. Petri net can be applied informally to any system that

W.J. Thong (✉) · M.A. Ameen
Fakulti Komputer Sistem and Kejuruteraan Perisian, Universiti Malaysia Pahang,
Kuantan, Pahang, Malaysia
e-mail: briantwj@gmail.com

M.A. Ameen
e-mail: mohamedariff@ump.edu.my

can be described graphically like flow charts and that needs some means of representing parallel or concurrent activities [3]. Since Petri net can be applied in most system to characterize it graphically, a lot of Petri net tools had been developed for this purpose. Using Petri net tools, users can represent their system in details and analyses the performance of the system. Users can also use the Petri net tools as a graphical editor and code generator. Some Petri net tools can also be used to simulate the system and provide model checking for it. To date, there are many different types of Petri net tools for different environments and purposes. However, there are no published papers on the recommendation of Petri net tools to the users based on the users' operating system and Petri net to be analyzed. This survey paper aims to compare twenty different types of Petri net tools in different aspects. At the end of the paper, a discussion will be drawn to recommend different types of Petri net tools to users with different operating systems and depending on what type of Petri net the users wanted to analyze.

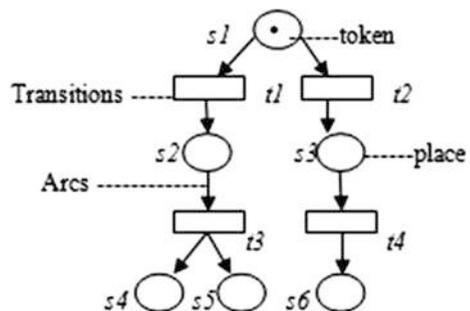
51.2 Preliminaries

Petri net can be defined as a formal modeling language that can be represented graphically with a strong mathematical foundation [4]. It is represented graphically in the sense that it serves as a visual communication aid to model the system behavior. It is based on a mathematical foundation in the sense that it represents the equations, algebraic equations and algorithms in the system. Petri net are used to model control flow in a system and is capable of modeling concurrency and synchronization in distributed systems.

A Petri net consists of three types of components: *places* (circles), *transitions* (rectangles) and *arcs* (arrows). Places represent different states of the system. Transitions represent events or actions which cause the change of a state. An arc connects a place with a transition or a transition with a place. Another element in Petri net is the token. The movement of a token from place to place indicates a change of state. The movement of a token is also known as firing.

Figure 51.1 represents an example of a Petri net with six places (states), four transitions (change of states) and nine connecting arcs.

Fig. 51.1 Example of petri net



The token is at $s1$ for starters. A token travels to the next state via transition. The direction of the token's movement is represented by the arrow head of the arcs. The token can be fired into $s2$ via $t1$. The token is now present in $s2$ after leaving $s1$. The firing continues with the token being fired to $s4$ and $s5$ via $t3$. This is also known as a concurrency or parallel relationship where one token is split into two (depending on the number of concurrent nodes). After this process, both the $s4$ and $s5$ states contain a token each. The $s1$ and $s2$ states are now without any tokens.

For another example, the token in $s1$ can also be fired into $s3$ state via $t2$. This results in $s3$ containing a token and $s1$ being empty. The process continues with $t4$ firing the token from $s3$ into $s6$.

In 1995, Monika Trompedeller proposed a classification of Petri net based on a survey by Bernardinello and De Cindio from 1992 [5]. The classifications have not been updated since then but still it is useful for getting a quick overview of the main differences between various kinds of Petri net. Petri net can be classified into 3 levels; in level 1, Petri net is characterized by places (states) which can represent Boolean value, for example, a place is marked by at most one unstructured token. Examples of level 1 Petri net are Condition/Event (C/E) Systems, Elementary Net (EN) Systems and 1-safe systems. An example of Petri net tool that supports analysis of level 1 Petri net is the Environment for Action and State based Equivalences (EASE).

For level 2, Petri net is characterized by places (states) which can represent integer values, for example, a place is marked by a number of unstructured tokens. An example of level 2 Petri net is the Place/Transition (P/T) Nets. Petri net tools that support analysis of level 2 Petri net include CPN/AMI, MOBY, POSES, PNS and WINPETRI.

In level 3, Petri net is characterized by places (states) which can represent high-level values, for example, a place is marked by a multi-set of structured tokens. Examples of level 3 Petri net include High-Level Petri net with Abstract Data Types (HL + ADT), Environment Relationship (ER) Nets, Traditional High-Level Petri net and Well-Formed (Colored) Nets (WN). An example of Petri net tools that support level 3 Petri net analyzing is the Cabernet which is used to analyze real time systems based on Petri net augmented with data, predicates, actions, and temporal information.

Judging by the classifications of Petri net and their complexity, different Petri net analyzing tools can only support different level of Petri net. There is not any best or worst tool in analyzing Petri net. Different categories of Petri net require different Petri net tools to analyze. For example a level 2 P/T Petri net will require tools such as CPN/AMI to analyze it instead of a level 3 Petri net tool such as Cabernet.

51.3 Survey

In this paper, several Petri net tools are compared.

51.3.1 ALPiNA

ALPiNA (Algebraic Petri net Analyzer) is a model checker for Algebraic Petri net created by the SMV (Software Modeling and Verification) Group at the University of Geneva [6]. ALPiNA is fully written in Java and it is available under the terms of the GNU general public license. ALPiNA provides a user-friendly user interface that was built with the latest metamodeling techniques on the eclipse platform. An Algebraic Petri net is categorized as a level 3 High-Level Petri net with Abstract Data Types (HL + ADT). Hence, ALPiNA is a Petri net tool which is capable of analyzing a level 3 Petri net.

51.3.2 CoopnBuilder

CO-OPN stands for Concurrent Object-Oriented Petri net. CoopnBuilder is an environment composed of a set of tools destined to the support of concurrent software development based on the CO-OPN language [7]. CoopnBuilder is also a research project from the SMV Group. CO-OPN is an object-oriented specification language based on synchronized Algebraic Petri net. CO-OPN allows the definition of active concurrent objects, and includes facilities for sub-typing, subclassing and genericity.

51.3.3 GreatSPN

GreatSPN2.0 is a software package for modeling, validation and performance evaluation of a distributed system using Generalized Stochastic Petri net and their colored extension: Stochastic Well-formed Nets [8]. It provides a friendly framework to experiment with timed Petri net based modeling techniques. GreatSPN implements analysis algorithms to allow its use on complex application.

51.3.4 LoLA

LoLA is a low level Petri net analyzer [9]. It is able to verify a broad variety of behavioral properties on a given Petri net. It is based on explicit state space verification, for example, an exhaustive enumeration of reachability state. In the context of service-technology.org, LoLA can be used to verify compatibility and for the validation of formal semantics.

51.3.5 PEP

PEP (Programming Environment based on Petri net) is a comprehensive set of modeling, compilation, simulation and verification components, linked together within a Tcl/Tk-based graphical user interface [10]. PEP's modeling components facilitate the design of parallel systems by parallel program, the compiler then generate Petri net from such models. The PEP tool is considered as an open platform. Further algorithms can be integrated in the user interface.

51.3.6 Snoopy

Snoopy is software tool to design and animate hierarchical graphs, among others Petri net [11]. The tool has been developed and is still under development at the University of Technology in Cottbus, Dep. of Computer Science, "Data Structures and Software Dependability". The tool is used for the verification of technical systems, typically software-based systems, as well as for the validation of natural systems such as biochemical networks as metabolic, signal transduction etc.

51.3.7 Marcie

Marcie (Model Checking and Reachability analysis done efficiently) is a tool for qualitative and quantitative analysis of a Generalized Stochastic Petri net with extended arcs [12]. The tool consists of four engines with each engine carrying their own unique function. It is possible to use Marcie with a graphical user interface provided by the tool Charlie.

51.3.8 Charlie

Charlie is a software tool used to analyze a level 2 Petri net in particularly a Place/Transition Net [13]. The tool has been developed and is still under development at the University of Technology in Cottbus, Dep. of Computer Science, "Data Structures and Software Dependability". The main features of the Petri net tool include structural properties analysis, invariant based analysis, reachability graph based analysis, reachability/coverability graph visualization using the JUNG library and plugin support. Charlie is able to read Place/Transition nets which have been created by the tool Snoopy, or P/T nets which are given in the Abstract Petri net Notation and also P/T nets that are given in the INA file format.

51.3.9 JSARP

JSARP (Simulator and Analyzer Petri net in Java) is a Petri net tool that describes and verifies Petri net with the support of a graphical interface [14]. JSARP is developed in Java and employs modern object oriented techniques and design patterns. JSARP is able to edit the Petri net with graphics and also works as a simulator.

51.3.10 MIST

MIST is a safety checker for Petri net and other extensions [15]. It implements several algorithms that solve the coverability problem for monotonic extensions of Petri net. The tool implements several algorithms: forward/backward search of the state using a symbolic data structure (the IST library). The tool also implements abstraction-refinement techniques or efficient traversal techniques to tackle the state explosion problem.

51.3.11 Petruchio

Petruchio is a tool used for computing Petri net translations of dynamic networks modeled in terms of Pi-calculus processes [16]. It provides a mean to further analyze nets. Petruchio is able to simulate GSPN (Generalized Stochastic Petri net) which is a high level Petri net, check coverability for low-level Petri net and application of reduction techniques for low-level Petri net.

51.3.12 PNEditor

PNEditor is an open-source Petri net editor [17]. It offers usual features of a graphical editor for the design of Place/Transition nets. In addition, the tool offers features like saving the net to a file, definition of roles, definition of subnets (nested nets), saving of predefined subnets to files and their reuse as reusable components, replacement of subnets, definition of static places etc. It requires JAVA SE 6+ to operate.

51.3.13 Jasper

Jasper (Yet Another Smart Process Editor) is a tool for modeling and simulating stepwise process [18]. It uses extended Petri net as its modeling technique. Jasper offers easy editing, token gameplay and performance analysis with randomized

automatic simulation for basic Place/Transition nets with some extension, including case-specific vs. inter-case token flow, reset and inhibitor arcs, decision nodes with parameterized probabilities of alternatives etc.

51.3.14 PAPETRI

PAPETRI is a general and integrated environment for editing and analyzing Petri net [19]. It allows users to work with difference classes of nets. Several analysis tools are available for each of these classes. PAPETRI aims to provide a friendly editing environment and to afford a greater deal of analysis tools for different classes of Petri net.

51.3.15 Xpetri

XPetri is a graphical simulator of Petri net [20]. XPetri is an Xwindows application designed to be portable across UNIX platform. It supports Place/Transition nets with weighted arcs and a strict firing rule. It also allows a minimum and maximum number of tokens to be specified for each place. Xpetri also supports temporary disabling of transitions for a single fire or until re-enabled.

51.3.16 PROD

PROD is a reachability analysis tool for Predicate Transition Nets [21]. The stubborn set method for reduced state space generation has been implemented in PROD. PROD also has a rich query language for inspecting the generated state space. It is available for download and free of charge.

51.3.17 ARP

ARP (Petri net Analyzer) is a software tool for Petri net analysis and simulation developed by Maziero [22]. The interface is simple and the ARP tool accepts Place/Transition Nets, Timed Nets and Extended Timed Nets. The features of ARP include accessibility analysis, invariant analysis, equivalence analysis, performance evaluation and manual simulation.

51.3.18 JPetriNet

JPetriNet is software that is used to model, analysis conventional Petri net and to simulate Timed Petri net [23]. The project is a Petri net modeling, analysis and simulation tool made in Java Programming Language. The tool is created for educational purpose and also to be used in any other purpose involving concurrent system.

51.3.19 Petri .NET Simulator

Petri .NET Simulator is a tool used for modeling and simulation of Petri net and analysis of their behavior [24]. It can be used to simulate flexible manufacturing systems and also be used for discrete event system. The tool is free for download.

51.3.20 QPME

QPME (Queuing Petri net Modeling Environment) is an open-source tool for stochastic modeling and analysis based on the QPN (Queuing Petri net) modeling formalism [25]. Queuing Petri net is a combination of conventional queuing networks and stochastic Petri net which provides improved expressiveness and thus making it possible to model systems at a higher degree of accuracy. QPME is made of two components, which is QPE (QPN Editor) and SimQPN (Simulator for QPN).

51.4 Discussion and Conclusion

In this section, the survey is tabulated in Table 51.1. Analysis is done based on five criteria, which are Petri net supported, Component, Analysis, Environments and whether it is Free of Charge.

The table above is the comparison between the Petri net tools surveyed in Sect. 51.3 based on five main criteria. The first group of criteria is the Petri net supported. In this category, the Petri net tools are compared in terms of what type of Petri net is supported. Majority of the tools support Place/Transition Petri net with some supporting high-level Petri net (i.e. AIPiNA, CoopnBuilder, PROD, and QPME). However, QPME stands out in this category as it supports Queuing Petri net (a combination of Queuing Network and Petri net). For Continuous Petri net, only Snoopy and Charlie support it.

Table 51.1 Comparison between petri net tools

Petri net tool	Petri net supported							Component			
	High-level petri nets	Object-oriented petri nets	Stochastic petri nets	Petri nets with time	Place/transition nets	Continuous petri nets	Transfer petri nets	Queueing petri nets	Graphical editor	State spaces	Condensed state spaces
AIPINA	x								x	x	x
CoopnBuilder	x	x							x		
GreatSPN	x		x	x					x	x	x
LoLA	x				x					x	x
PEP	x			x	x				x	x	x
SNOOPY			x	x	x	x			x		
MARCIE			x							x	
CHARLIE			x		x	x			x		
JSARP		x							x		
MIST					x					x	x
PETRUCHIO	x		x	x	x		x		x	x	
PNEditor					x				x		
Yasper			x	x	x				x		
PAPETRI	x				x				x		
Xpetri			x		x				x		
PROD	x				x					x	x
ARP				x	x					x	
JPetriNet				x	x				x		

(continued)

Table 51.1 (continued)

Petri net tool		Petri net supported							Component			
		High-level petri nets	Object-oriented petri nets	Stochastic petri nets	Petri nets with time	Place/transition nets	Continuous petri nets	Transfer petri nets	Queueing petri nets	Graphical editor	State spaces	Condensed state spaces
	Petri .NET Simulator				×	×				×		
	QPME	×		×		×			×			
Component												
Code generation	Token game animation	Fast simulation	Place invariants	Transition invariants	Net reduction	Model checking	Petri net generator	Interchange file format	Analysis			
									Simple performance analysis	Structural analysis	Advance performance analysis	
×	×	×	×	×	×			×				
	×	×	×	×	×					×		×
	×	×	×	×	×			×				
	×	×	×	×	×					×		
	×	×	×	×	×					×		
	×	×	×	×	×							
	×	×	×	×	×							
	×	×	×	×	×							
	×	×	×	×	×							

(continued)

Table 51.1 (continued)

Component														
Code generation	Token game animation	Fast simulation	Place invariants	Transition invariants	Net reduction	Model checking	Petri net generator	Interchange file format	Analysis				Advance performance analysis	
									Simple performance analysis	Structural analysis	Macintosh	UNIX		
	x	x			x			x						
	x									x				
	x										x			
						x								
		x	x	x							x			
	x	x											x	
		x						x						
														x

Analysis	Environments										Free of charge		
	Reachability graph based analysis	Invariant based analysis	Java	Linux	Sun	HP, HP-UX	Silicon IRIX	MS DOS	Windows	Macintosh		UNIX	
			x										x
			x										x
				x	x								x
				x	x	x		x					x
				x	x								x
				x					x				x
				x						x			x

(continued)

The second criteria of comparison are the components in each tool. Most of the tools provide a graphic editor and a fast simulation on Petri net. Tools that provide graphic editor and fast simulation on Petri net can be good teaching materials. Users will be able to edit the Petri net and simulate different Petri net to aid them in understanding Petri net. PEP has the highest number of components which includes graphical editor, state spaces, condensed state spaces, token game animation, place invariants, transition invariants, net reduction, model checking, Petri net generator and interchange file format. Users will be able to experience more on PEP compared to the other tools.

The next category of comparison is the analysis of Petri net. Some of the tools surveyed provide simple performance analysis, while tools such as GreatSPN, PEP, Charlie, JSARP, Xpetri, ARP and JPetriNet provide structural analysis. GreatSPN and QPME are also able to carry out advance performance analysis. The reachability graph based analysis is however only able to be carried out by MARCIE.

The next criterion used to compare between the Petri net tools is the environment. LoLA, PETRUCHIO and QPME have the highest amount of environment supported for their tools with six environments for each of them. Tools like AIPiNA, CoopnBuilder, MIST, Yasper, PAPETRI, ARP and Petri .NET Simulator is very environment specific with each of them only supporting one specific environment to be run on.

The final criterion of comparison is the pricing of the Petri net tools. All of the Petri net tools are either free of charge or free of charge for academic purpose to be downloaded.

In this section, Table 51.2 is presented to recommend Petri net tool for a few scenarios which includes the users' operating systems and the types of Petri net to be analyzed.

Table 51.2 summarizes on the recommended Petri net tools for different environments and different types of Petri net to be analyzed. Overall, users running on the MAC OS X are recommended to use the tool PETRUCHIO which supports most of the different types of Petri net. For Continuous Petri net, users will need to use the tool SNOOPY; while for Queuing Petri net, users are recommended to use the tool QPME. For Windows users, PETRUCHIO is recommended too as it supports most of the Petri net types. However, analysis of Continuous Petri net and Queuing Petri net requires users to use the tool CHARLIE and QPME respectively. For those running on Linux, there isn't any specific tool overall that can analyze most of the different types of Petri net. For High-level Petri net, Petri net with time and Place/Transition Petri net, users are recommended to use the tool PEP. For Stochastic Petri net, users are recommended to use GreatSPN or PETRUCHIO. CHARLIE is recommended for the analysis of Continuous Petri net while QPME is recommended for the analysis of Queuing Petri net.

Overall, the PEP tool offers the most components and analysis types amongst the twenty Petri net tools compared. However, due to the lack of supporting environment, PEP is not user friendly. Users will need to have Linux or Sun operating system to support the PEP tool. Where else PETRUCHIO supports most

Table 51.2 Recommended petri net tools for different usage

Operating system	Type of petri nets	Recommended tool
MAC OS X	High-level petri nets	PETRUCHIO
	Object-oriented petri nets	Not supported
	Stochastic petri nets	PETRUCHIO
	Petri nets with time	PETRUCHIO
	Place/transition petri nets	PETRUCHIO
	Continuous petri nets	SNOOPY
	Transfer petri nets	PETRUCHIO
	Queueing petri nets	QPME
Windows	High-level petri nets	PETRUCHIO
	Object-oriented petri nets	Not supported
	Stochastic petri nets	PETRUCHIO
	Petri nets with time	PETRUCHIO
	Place/transition petri nets	PETRUCHIO
	Continuous petri nets	CHARLIE
	Transfer petri nets	PETRUCHIO
	Queueing petri nets	QPME
Linux	High-level petri nets	PEP
	Object-oriented petri nets	Not supported
	Stochastic petri nets	GreatSPN/PETRUCHIO
	Petri nets with time	PEP
	Place/transition petri nets	PEP
	Continuous petri nets	CHARLIE
	Transfer petri nets	PETRUCHIO
	Queueing petri nets	QPME

of the major environments (MAC OS X, Windows, Linux), it is also the tool with the second highest component and analysis types. Based on Table 51.2, each operating system has at least two types of Petri net that is able to be analyzed by PETRUCHIO. From this deduction, we can conclude that PETRUCHIO is the best tool as it supports most of the operating system and it has an adequate amount of features in it.

References

1. Petri, C.A.: Communication with automata (1966)
2. Brauer, W., Reisig, W.: Carl adam petri and "petri nets". Fundam. Concepts Comput. Sci. **3**(5), 129 (2009)
3. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE **77**(4), 541–580 (1989)

4. Ameedeen, M.A.: A model driven approach to analysis and synthesis of sequence diagrams. Diss. University of Birmingham (2012)
5. Trompedeller, M.: A Classification of Petri Nets (1995)
6. Hostettler, S., Marechal, A., Linard, A., Risoldi, M., Buchs, D.: High-level petri net model checking with ALPiNA. *Fundam. Inf.* **113**(3), 229–264 (2011)
7. Al-Shabibi, A., Buchs, D., Buffo, M., Chachkov, S., Chen, A., Hurzeler, D.: Prototyping object oriented specifications. In: *Applications and Theory of Petri Nets 2003*, pp. 473–482. Springer, Heidelberg (2003)
8. Chiola, G., Franceschinis, G., Gaeta, R., Ribaudo, M.: GreatSPN 1.7: graphical editor and analyzer for timed and stochastic petri nets. *Perform. Eval.* **24**(1), 47–68 (1995)
9. Schmidt, K.: Lola a low level analyser. In: *Application and Theory of Petri Nets 2000*, pp. 465–474. Springer, Berlin (2000)
10. Grahlmann, B., Best, E.: PEP—more than a petri net tool. In: *Tools and Algorithms for the Construction and Analysis of Systems*, pp. 397–401. Springer, Berlin (1996)
11. Heiner, M., Herajy, M., Liu, F., Rohr, C., Schwarick, M.: Snoopy—a unifying petri net tool. In *Application and Theory of Petri Nets*, pp. 398–407. Springer, Berlin (2012)
12. Schwarick, M., Heiner, M., Rohr, C.: Marcie-model checking and reachability analysis done efficiently. In: *Proceedings of the IEEE, 8th International Conference on Quantitative Evaluation of Systems (QEST)*, pp. 91–100 (2011)
13. Wegener, J., Schwarick, M., Heiner, M.: A plugin system for charlie. In: *Proceedings of the CS&P*, pp. 531–554 (2011)
14. Oliveira Lino, F.G., Sztajnberg, A.: JSARP: Simulator and analyzer petri net in java, a final project of the undergraduate course of computer and information technology, University of Rio de Janeiro (2006) (Unpublished manuscript)
15. Ganty, P., Van Begin, L., Delzanno, G., Raskin, J-F.: Coverability checkers included in mist. Online, URL: <https://github.com/pierreganty/mist/wiki>
16. Meyer, R., Strazny, T.: Petruchio: From dynamic networks to nets. In: *Computer Aided Verification*, pp. 175–179. Springer, Berlin (2010)
17. Riesz, M., Seckár, M., Juhás, G.: PetriFlow: A petri net based framework for modelling and control of workflow processes. In: *Proceedings of the ACSD/Petri Nets Workshops*, pp. 191–205 (2010)
18. van Hee, K., Oanea, O., Post, R., Somers, L., van der Werf, J.M.: Yasper: a tool for workflow modeling and analysis. In: *Proceedings of the IEEE, 6th International Conference on Application of Concurrency to System Design (ACSD 2006)*, pp. 279–282 (2006)
19. Berthelot, G., Johnen, C., Petrucci, L.: PAPETRI: environment for the analysis of petri nets. In: *Computer-Aided Verification*, pp. 13–22. Springer, Berlin (1991)
20. Geist, R., Crane, D., Daniel, S., Suggs, D.: Systems modeling with xpetri. In: *Proceedings of the 26th Conference on Winter simulation*, Society for Computer Simulation International, pp. 611–618 (1994)
21. Varpaaniemi, K., Heljanko, K., Lilius, J.: Prod 3.2: an advanced tool for efficient reachability analysis. In: *Computer Aided Verification*, pp. 472–475. Springer, Berlin (1997)
22. Maziero, C.A.: ARP: Petri net analyzer. Control and Microinformatic Lab, Federal University, Santa Catarina (1990)
23. Azevedo, M.: JPetriNet. Sapucaí Valley University, Brazil (2004) Online, URL: <http://jpetrinet.sourceforge.net/>
24. Genter, G.: Petri .NET simulator. Online, URL: <http://www.petrinetsimulator.com/>
25. Kounev, S., Dutz, C.: QPME: a performance modeling tool based on queueing petri nets. *ACM SIGMETRICS Perform. Eval. Rev.* **36**(4), 46–51 (2009)

Chapter 52

Towards a Exceptional Distributed Database Model for Multi DBMS

Mohammad Hasan Ali and Mohd Azlishah Othman

Abstract This paper reported on the current issues distributed transaction faces in the multi database management system. We looked at how the distributed transaction in database differs from other distributed processing. After highlighting the major issues in distributing transition in the multi DBMS, concurrency control and recovery such as (Site failure, network failure, time failure, and so on) were considered in the design of our model. We proposed an exceptional distributed transaction model for processing the transaction queries in the multi DBMS by keeping in mind the point-to-point transaction processes. The expected results were described in this paper.

Keywords Distributed transaction · DBMS · Transaction network

52.1 Introduction

Generally, distributed transaction defined as a transaction that updates data on two or more networked computer systems. However, different conceptual structure for processing the transactions elements are extend for fitting a certain needs for the applications that must update distributed data in term of peer to peer connection. Developing a transaction model is difficult because most of these techniques are customized to be occurring in a single transaction point attach with a multiple failures such as (Client's failure, server failure, and the network connection).

M.H. Ali (✉) · M.A. Othman
Centre for Telecommunication Research and Innovation, Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
e-mail: eng.mohammed12@yahoo.com

M.A. Othman
e-mail: azlishah@utem.edu.my

Furthermore, connecting different distributed points require an advance network security technology to achieve the best performance for detecting and recovering these failures over a network connection [1].

Recent transaction applications are optimizing various security integrations for securing the transaction between client and server that detect and manage the change action (such as a database update) before the transaction can occur. The transaction process over the network can be determined a complicated process because of the multi security issues in these networks. The performance of transaction between single and multi database management systems is coordinate to other network parts for the components involved and acts as a transaction manager for each computer that manages transactions.

Algorithms were developed to serve a certain needs, which begin used in banking, investment environments, enterprise applications, etc. [1]. During the communication between computers a transaction manager sends prepare, commit, and abort messages to all its subordinate transaction managers [2]. In addition, the common steps for obtaining a distributed transaction are:

- The client starts a transaction by calling the transaction manager. This application allows for the transaction to be deployed.
- While the transaction manager processing the client request, an update will be saved for the deployed transaction to commit the process based on the client details.
- After processing the client request, the transaction manager keeps a sequential transaction log so that it's committing or abort decisions will be durable.
 - If all components are prepared,
 - If any component cannot prepare,
 - While a component remains prepared but not committed or aborted.

52.1.1 Transaction Problems

During the database transaction over multi database management systems, problems were addressed while dealing in the concurrency control and recovery, most of these problems occurred in the distributed databases [3]. However, the main problems can be simply classified into the following:

- **Site's failure:** this action happened when one or more sites in a DDBMS fail. This type of failure will corroborate with the application manager for restoring the transaction updates,
- **Network Problems:** this action occurred during the communication network fails for unclear transaction, this type of network failure indicates one or more sites to be cut off from the rest of the site in the distributed database,

- **Data Duplication:** this problem occurred when multiple copies of the database are unable to be tracked accordingly for maintaining consistency,
- **Distributed Transaction:** this problem arises when a distributed transaction happened in multi sites,
- **Distributed Deadlocks:** during the database transaction deadlocks could be launched in any single or multi accessing for the site contents.

An additional various problems been reported in the transaction management among the distributed environments is ensuring the consistency of the data in the presence of site and network failure. However, various transaction protocols were developed for providing a stable transaction over a network. Moreover, different problems were addressed in securing the transaction elements over a network

52.1.2 Related Works

The importance of database transaction has brought the needs to develop and adopt new transaction technologies. A study by Sami and Habib [4] explored and developed a new transaction mechanism based on the using of agent systems for providing a high transferring reliable among different distributed algorithms.

However, the study developed a transaction algorithm for solving DCSP issues based on the agent systems by classifying it into different groups: Variables' and Controller Agents, these two groups helps to reformulating of the detected agent communication algorithm. The proposed model can be used also in order to treat non-binary constraints and managing the transaction according to its original destination. The instantiation of these variables can be done by negotiation in order to separate the sub problems into totally independent ones. Figure 52.1 presents the proposed classification based on the agent systems [4].

However, another study by Qiming and Umesh [5] described the transaction issues in the E-Commerce applications. The study customized the computing environment for distributing a limit transaction through the database management systems within a large number of the autonomous service requester. The study suggested the using of transaction agents that cooperate to perform business transaction activities. The study also reported the transaction requirements among these e-commerce applications that require to provide a peer-to-peer protocols based on distributed communication.

Hence, Qiming and Umesh developed cooperative multi-agent transaction model that includes peer-to-peer protocols for committing control and failure recovery. This model could help to facilitate the urgent tasks for transferring the data elements through channel query. Figure 52.2 shows the proposed model structure with the database components [5].

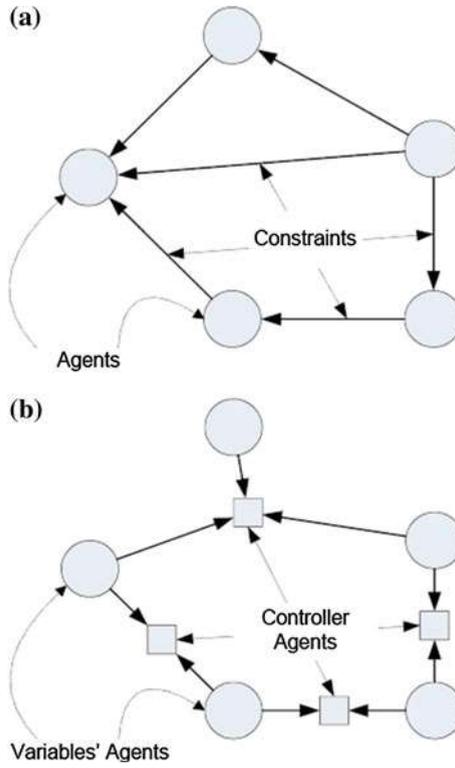


Fig. 52.1 Constraint networks based agent [4]

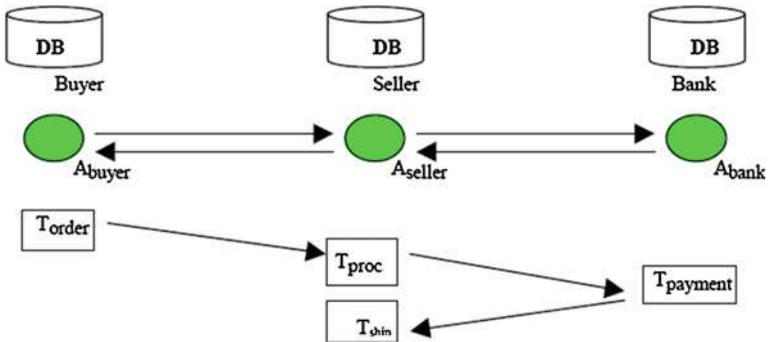


Fig. 52.2 The proposed model structure with the database components [5]

52.1.3 SQL Distributed Transaction Example

The current distributed computing development has generated a new retrieving and representing techniques for the data elements over the grid. SQL presents one of these techniques which obtain a high level processing of client queries for a set of processes that seek to achieve some form of cooperation in the database. Moreover, executing a different conceptual representation simplify the processes of a distributed system for performing some actions by crashing or being disconnected.

SQL server provides a sustainable database processing which starts with the first SQL statement from beginning the transaction. This transaction carries out a different meaning of the grid contents when it is committed or rolled back, the main statements that SQL starts with are COMMIT or ROLLBACK statement.

The next step is to extract the transaction concept by considering a certain procedure for obtaining a successful transaction along with the database elements. Example, when one or many customers transfer a certain amount into other accounts, it will be directed to the saving section to the checking account.

There are important operations for obtaining that transaction such as:

- Decrement the amount of the savings account.
- Add in other elements of the checking account.
- Record the transaction steps by saving its actions.

SQL database allows a huge number of transactions to be performed when all the SQL operations can be applied to maintain the accounts with the adopted balance. Meanwhile, in case of any leaks of the transaction such as transaction invalid number, software and hardware failure, which leads to end the transaction without processing the amounts into another account?

Figure 52.3 describes the transaction mode in the SQL database.

This kind of transferring can be provided by the internet services or wireless technology (Fig. 52.4).

It is desirable to start a new development for adopting a new mechanism to facilitate the distributed transaction process over database management systems which require specific requirements. Below are presented some recommendation for achieving a successful transaction over SQL database components.

- To allow a stable transaction for a certain elements over SQL statements, SQL transaction statements must initially agree to the client request by classifying the transaction details into several points with common formats for reprocessing the transactions.
- Obtaining a successful data exchanging can be easily achieved by determining the require actions for the unknown transactions into the grid with several alternative plans of analyzing.
- Providing a collaborative transaction can also be accepted for starting a new distributed transaction based on the SQL statements for the cooperating processes.

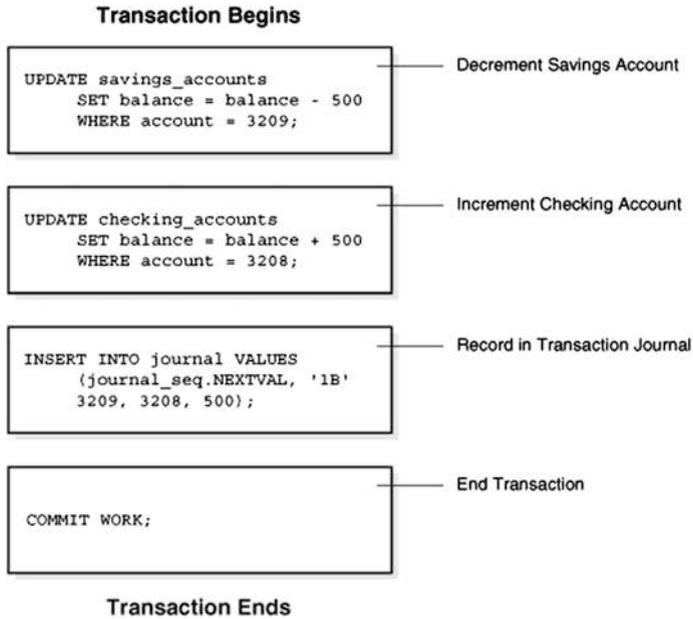


Fig. 52.3 The transaction mode in the SQL database

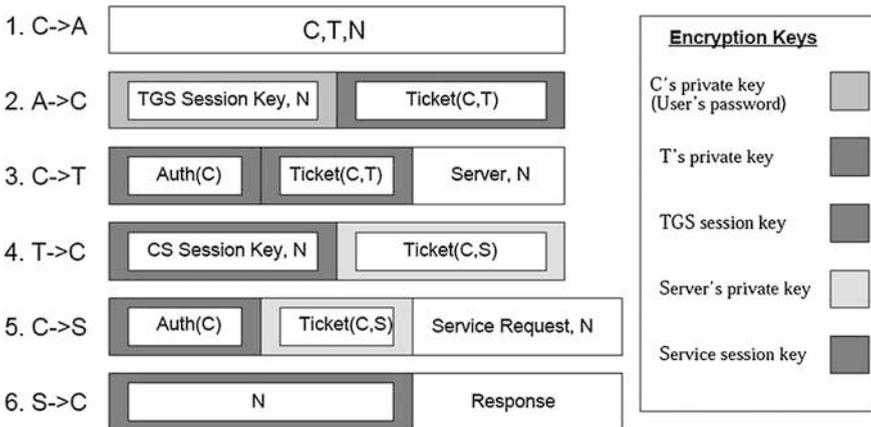


Fig. 52.4 SQL transaction statement's performance

- This kind of transactions may require not only to agree on the requested actions, but also its requiring to execute on the transaction query by which these actions need to be executed among the database.

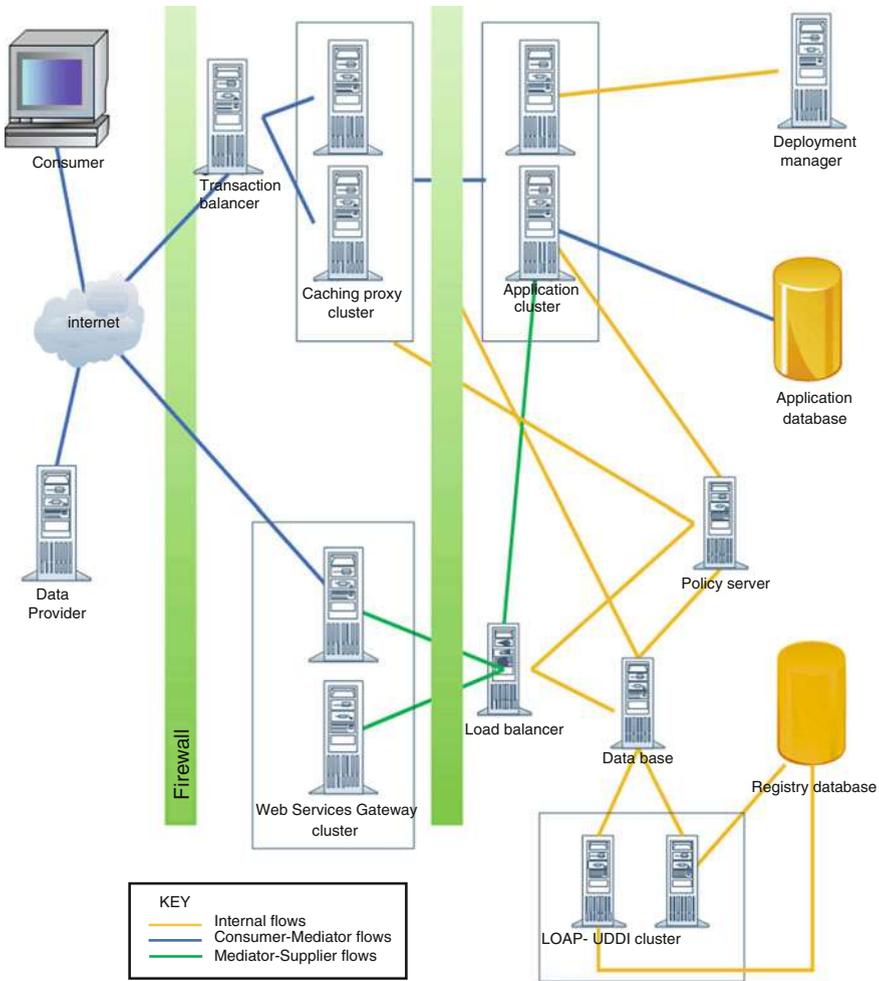


Fig. 52.5 Proposed exceptional distributed database model

52.1.4 Proposed Technique

Based on the problems that have been discussed above in the distributed transaction applications, and based on the related and previous works, we proposed an exceptional distributed transaction over multi database management systems. This model provides a sustainable communication between client and server to perform the client queries in more secure and reliable way. However, the proposed model could reduce the number of errors occurred during the transaction process, which optimize a multi transaction point among servers. Additionally, this model adopted an advance network communication for reducing the network failure which occurs

during the transaction. From the Fig. 52.5 can see that the distributed communication indirectly depends on the database query which gives high privileges for the transactions.

52.1.5 Expected Result

The expected result of performing the proposed model can be summarized into the following:

- Addressed the problem of transaction directly by looking into the history of transactions,
- Discovering dependencies among services from historical transactions,
- Execution time period data can be determined for multi transactions among DBMS,
- Monitoring instrumentation over multi DBMS.

52.2 Conclusion

This study illustrated the major issues in processing transactions in multi database management systems along with the techniques for processing databases in different fragments. Moreover, this study adopted a new way for resolving the site and network failure occurred during the database transaction in a multi DBMS. The study proposed an exceptional distributed transaction model for performing and managing the transaction details more efficiently. We have considered the related studies in developing the proposed model.

References

1. Gupta, M., Neogi, A., Agarwal, M.K., Kar, G.: Discovering dynamic dependencies in enterprise environments for problem determination. In: Proceedings of the 14th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management. Lecture Notes in Computer Science, vol. 2867, pp. 221–232. Springer, Oct 2003
2. Bessiere, C., Brito, I.: Asynchronous Backtracking without Adding Links: A New Member in the ABT Family. *Artif. Intell.* **161**(1–2), 7–24 (2005)
3. Al-Maqtari, S., Abdulrab, H., Nosary, A.: Constraint programming and multi-agent system mixing approach for agricultural decision support system. In: Emergent Properties in Natural and Artificial Dynamical Systems, pp. 199–213 (2006)
4. Sami, M., Habib, A.: Controller-agent based approach for solving distributed constraint problem. LITIS Lab—INSA of Rouen, France (2006)
5. Qiming, C., Umesh, D.: Multi-Agent Cooperative Transactions for E-Commerce, vol 2, pp. 33–37, USA (2007)

Chapter 53

Semantic Search Engine Using Natural Language Processing

Sudhakar Pandiarajan, V.M. Yazhmozhi and P. Praveen kumar

Abstract The World Wide Web has become colossal and its growth is also dynamic. Most of the people rely on the search engines to retrieve and share information from various resources. All the results returned by search engines are not always relevant as it is retrieved from heterogeneous data sources. Moreover a naive user finds it difficult to confirm that the retrieved results are significant to the user query. Therefore semantic web plays a major role in interpreting the relevancy of search results. In this work, a novel algorithm is proposed for retrieving relevant documents using semantic web based on the concept of Natural Language processing (NLP). In this proposed system, NLP is used to analyse the user query in terms of Parts of Speech. The extracted terms are compared to the domain dictionary to identify the relevant domain of the user interest. On the other hand, the retrieved documents of the user query are investigated with the help Natural language processing to identify the relevant domain. Now the documents are ranked as per the relevancy of the contents against user query. The experimental result of the proposed algorithm indicates that the accuracy of the retrieved document is 97 %.

Keywords Semantic search engine · Natural language processing

S. Pandiarajan (✉) · V.M. Yazhmozhi · P. Praveen kumar
Kamaraj College of Engineering and Technology, Virudhunagar 626101, India
e-mail: sudhakarcse@kcetvnr.org

V.M. Yazhmozhi
e-mail: yazh.technovate@gmail.com

P. Praveen kumar
e-mail: praveenkumargen@kcetvnr.org

53.1 Introduction

With the increasing use of World Wide Web as an essential source of information, there is a need to work with the Semantic Web [1], in order to reduce the irrelevant data obtained during the search. The existing search engines [2], [3] do not provide domain specific search and they simply perform keyword matching. Those search engines cannot understand the negative senses, Example: “I do not want clustering”. The result set of the existing search engines for such a query would be related to clustering as they merely perform keyword matching and don’t analyse the actual meaning of the user query. Other example of queries with negative senses includes “I want clustering algorithms except cobweb”. The pro-posed approach using semantic search algorithm overcomes all the above stated pitfalls.

Semantic Web provides the users a comfort zone and reduces the wastage of time. The proposed work on Semantic search is accomplished by POS (Parts Of Speech) tagging using Natural Language Processing. Using POS Tagging, the proposed semantic search algorithm can understand what the user query conveys and hence it provides more relevant results to the user. The query entered by the user is POS tagged using Stanford Parser, and the tags for each word in the user query are obtained. For each tag obtained, if it is a noun it is added to the noun list (NL), and all other tags correspondingly. Each word in the NL is now compared with the word dictionary of each document that is obtained during pre-processing the document. If a match is found, then the word weight is incremented, and is added to the word weight list (WWL). The documents are sorted based on the WWL and added to the result set. Similarly, each sentence in the document is detected using OpenNLP, and its weight is incremented if all the nouns in the NL are matched. This weight of each sentence is added to the sentence weight list (SWL). The documents are then sorted based on the SWL and added to the result set. If there is any occurrence of negative word (e.g., not, except, NEITHER-NOR), then all the nouns in the NL are skipped and compared with the word dictionary of each document if verbs of possession occurs before those negative words. These documents are also added to the result set. The ranked documents from the result set are retrieved to the user. By this approach of Semantic search [4], [5], [6], the search results that are more relevant to the user’s interest are provided. In the proposed work, pdf, word and html documents have been considered for analysis.

53.1.1 Outline of the Paper

Section 53.2 presents the various works on semantic web and natural language processing supportive to the proposed research. Section 53.3 describes the Architectural design of the proposed scheme. Section 53.4 illustrates the experimental results and Performance Evaluation. Section 53.5 depicts the conclusion and future work.

53.2 Related Works

In [7, 8], Mukhopadhyay et.al has ogy which is made effective by mapping the instances and the classes. Even though the results of this research prove that the performance are good than the regular search engine, the results are proved within the domain only. In [9], Cafarella et al. developed a search engine using natural language processing in which it out performs well in terms of producing relevant information using natural language processing. In [10], Karpagam had proposed a framework which is based on ontology to build the semantic search engine. The author finds the relevant document for the user query using the techniques like word stemming, ontology matching, weight assignment, rank calculation. If the approach is extended to a larger data set, the weight assignment and rank calculation will become tedious. In [11], Jiang et al. developed a semantic search engine that overcomes the problem of knowledge overhead by a query interface. In [12], Lei et al. proposed a semantic web portal that has been designed to ensure the quality of the extracted metadata and it also facilitates for data querying. In [13], Kruse et al. had used WordNet, a lexical database to find the word senses in order to achieve semantic. But WordNet does not provide a classification of word senses in technical terms. In [14], Lara et al. had proposed a hybrid searching technology which is the combination of ontology and traditional keyword based matching in which the drawbacks of keyword based search like stop words removal in the user query was reflected. In [15], Madhu had presented a survey on the search engine's generation and role of the search engines in the intelligent web in which it insists the necessity to build semantic search engines. In [16], Kalaivani had proposed a question answering system based on the semantic searching terminology and natural language processing technique. In [11], Jiang et al. developed a full text search engine has been designed to exploit ontological knowledge for document retrieval. In [17], Kerschberg et al. recommended a methodology has been developed to capture the semantics of the users search intent into target search queries of the existing search engine.

53.3 Architectural Design

In the proposed work, Structured and Unstructured documents are maintained separately. Whenever a user query arises, it is given to Stanford Parser for POS tagging. Based on POS Tagging [18] NN (Noun, Singular), NNS (Noun, Plural), NNP (Proper Noun, Singular), JJ (Adjective) and IN (Preposition) are extracted from user query and stored in a term list table. Decisions are taken based on the accompanying diagram. If the user query belongs to only one domain then all

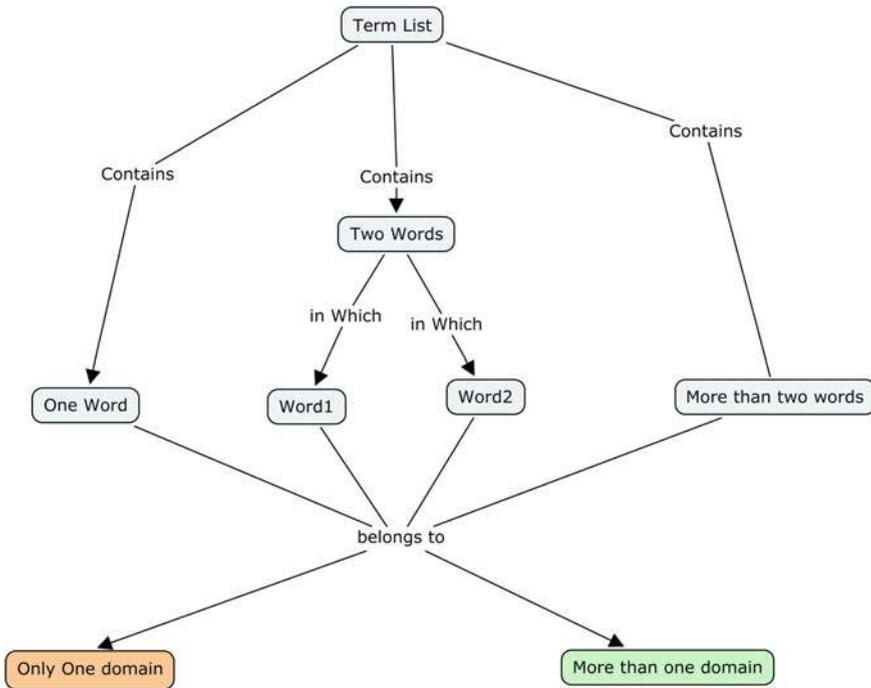


Fig. 53.1 Domain classification based on term list

the documents in the repository are compared against the domain dictionary and the results are computed. On the other hand, if the user query belongs to more than one domain then the dominant domain is computed based on Fig. 53.1. There are 2 possible cases can occur in selecting dominant domain

- Case 1: If Most of the Words belong to one domain then the same domain is taken as dominant domain.
- Case 2: If Equal number of Matches found with two domains then the choice are given to the user to confirm the dominant domain.

Once the dominant domain is selected, each document is split into sentences and each sentence is further divided into words. Each word is compared with term list and domain dictionary for matching. If a match is found then the “sentence weight” is incremented. The same process is continued for the whole document and cumulative document weight is calculated. If the user query contains negative words, then a negative ag is set to the user query and the process will request the

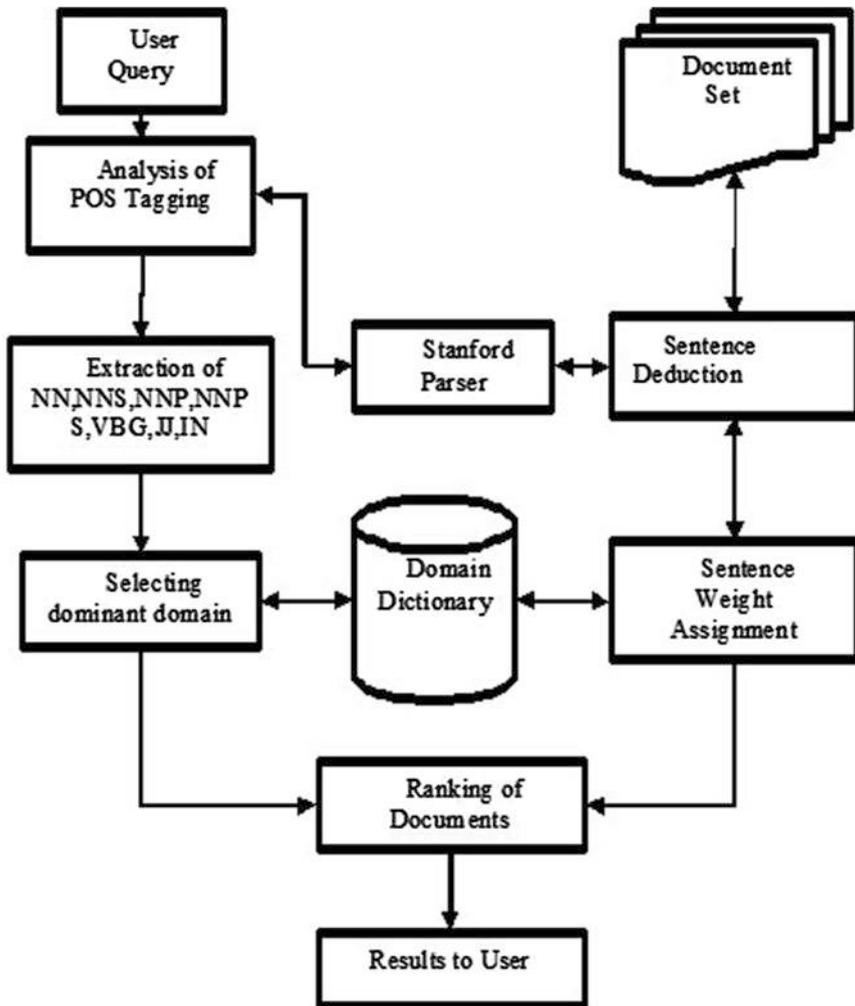


Fig. 53.2 Architectural design of the proposed approach

user to re-enter the direct query for better retrieval of results. In this way the accuracy of the retrieval is brought out. Once all the documents weight is calculated the highest weight of the document is ranked first and given to user as a good match in the retrieved results (Fig. 53.2).

Algorithm 1 - POS Tagging using NLP

INPUT : User query

OUTPUT : Resultant Document Set

METHOD : POS Tagging IN NLP

Step 1 : Initialize the Resultant document set RES = { }

Step 2 : Initialize nounlist NL= { }

Step 3 : Initialize sentencelist SL={ }

Step 4 : Initialize negativelist NGL= { }

Step 5 : Initialize NEGLIST= { not, no,neither, nor, except... }

Step 6 : Initialize neg ag=0, word weight=0, sentence weight=0;

Step 7 : POS Tag the user query using stanford parser.

Step 8 : foreach tag obtained

Step 8a : If tag is NN, NNS, NNP, NNPS, JJ, VBG then add the appropriate word in the query to NL

Step 8b: If tag is RB,DT,CC and the word is in NEGLIST then set neg ag=1

Step 9 : If NL.count=1 and PL.count=0 and neg ag=0 then nd the concept under which the single noun in NL occurs in Domain Dictionary.

Step 10 : If the single noun occurs under more than one concept, display all concepts.

Step 11 : As per the user selection, add that selected concept too in NL.

Step 12 : Detect sentences in the web documents using OpenNLP and store in SL

Step 13 : foreach sentence sent in SL

Step 14 : If all nouns in NL occur within a sentence

Step 15 : Increment weight for each match.

Step 16 : Store sentence weight of each document into SentenceWeightList SWL Step 17 : Sort the documents according to sentence weight and store it in the RES.

Step 18 : Compare each entry in the NL with word dictionary of each document

Step 18a: If a match is found then increment word weight with the count attribute's value.

Step 19 : Store word weight of each document into WordWeightList WWL.

Step 20 : Sort the documents according to word weight and if that document is not in RES store it in RES

Step 21 : If neg ag=1 then skip all nouns in NL and compare with the word dictionary of each document

Step 21a : Store such documents into RES

Step 22 : Retrieve the ranked documents fD1,..Dng in RES to the user

53.4 Experimental Results and Discussion

The experiment was conducted using American national corpus (anc) [19, 20] by in-creasing the number of documents in the repository gradually. The main factor considered for evaluation is accuracy rather than the speed. Three different scenarios are considered in evaluating the performance of the proposed system.

A confusion matrix is created with various combinations of Relevant documents (RD) and Conflicting documents (CD) to analyse the system. Performance evaluation of the proposed approach is done based on the classification context scenario. Precision, Recall, Accuracy and F-measure are the major measures used for classification based performance. Precision is the probability that measure a retrieved document is relevant to the context. Precision is calculated based on the formula

$$\text{Precision} = \frac{TP}{TP + FP}$$

where

TP is true positive (Correctly retrieved)

TN is true negative (Correctly rejected)

FP is false positive (Incorrectly retrieved)

FN is false negative (Incorrectly rejected).

Recall is the probability that calculates a relevant document is retrieved in a search process. Recall is calculated based on the formula

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy is calculated based on the formula

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

First scenario creates a large number of Relevant Documents against the small number of Conflicting documents. The Second scenario creates an equal number of Relevant and Conflicting documents. The Third scenario creates a small number of Relevant Documents against with large number of Conflicting documents. All the three cases, the document set contents are increased gradually (Tables 53.1, 53.2, 53.3).

The accuracy and recall of the system with respect to the above stated 3 cases is plotted in Figs. 53.3 and 53.4 respectively.

All the performance measures of the system clearly highlights that the results produced by the system in terms of all aspects are better compared with existing systems.

Table 53.1 Relevant documents are higher than the conflicting documents

No. of documents	Accuracy	F-measure	Recall	Precision
100	0.978	0.965	0.974	0.962
200	0.975	0.963	0.972	0.96
300	0.974	0.964	0.973	0.961
400	0.974	0.962	0.971	0.957
500	0.97	0.963	0.973	0.961
600	0.967	0.962	0.974	0.964
700	0.966	0.96	0.972	0.961
800	0.968	0.962	0.973	0.962
900	0.97	0.96	0.972	0.961
1000	0.968	0.952	0.965	0.95
1500	0.964	0.958	0.962	0.951
2000	0.962	0.956	0.96	0.95
3000	0.963	0.956	0.961	0.951
5000	0.962	0.955	0.959	0.95

Table 53.2 Relevant documents are equal to the conflicting documents

No. of documents	Accuracy	F-measure	Recall	Precision
100	0.975	0.963	0.972	0.96
200	0.974	0.961	0.97	0.958
300	0.972	0.963	0.967	0.951
400	0.974	0.96	0.962	0.955
500	0.972	0.962	0.963	0.952
600	0.97	0.963	0.965	0.952
700	0.965	0.961	0.962	0.949
800	0.968	0.962	0.964	0.95
900	0.964	0.96	0.96	0.948
1000	0.964	0.952	0.962	0.95
1500	0.962	0.95	0.96	0.949
2000	0.96	0.955	0.958	0.948
3000	0.96	0.952	0.955	0.95
5000	0.961	0.951	0.954	0.95

Table 53.3 Relevant documents are smaller than the conflicting documents

No. of documents	Accuracy	F-measure	Recall	Precision
100	0.972	0.958	0.965	0.954
200	0.97	0.957	0.967	0.953
300	0.968	0.953	0.962	0.95
400	0.973	0.952	0.961	0.952
500	0.97	0.954	0.964	0.952
600	0.965	0.951	0.96	0.95
700	0.964	0.948	0.961	0.951
800	0.965	0.95	0.962	0.953
900	0.965	0.948	0.962	0.951
1000	0.96	0.952	0.958	0.95
1500	0.961	0.947	0.956	0.945
2000	0.958	0.952	0.957	0.946
3000	0.959	0.955	0.958	0.95
5000	0.96	0.954	0.958	0.951

Fig. 53.3 Shows that the accuracy of the proposed systems is maintained between 96 and 98% irrespective of the number of documents

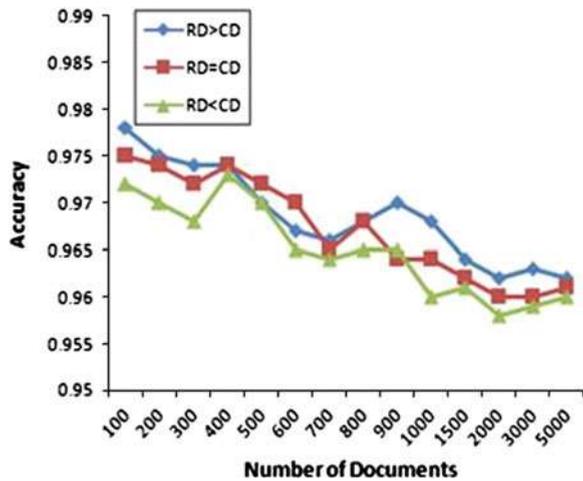
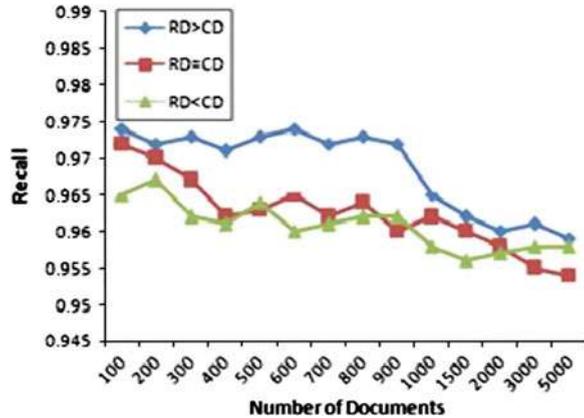


Fig. 53.4 Presents the recall measure of the proposed system against various numbers of document sets. Irrespective of the documents combinations the proposed system outperforms well in between 95.5 and 97.5%



53.5 Conclusion

In this study a new approach is proposed based on Natural language processing to understand and classify the documents based on the user query. The experimental results of the proposed systems point out that the accuracy of this system is vary in between 95 and 97 % in terms of relevancy. However the time taken to classify a document is little high compared with existing search engines. In future, our system will address the existing drawbacks to compete with other search engines in terms of the time factor.

References

1. <http://swoogle.umbc.edu/>
2. <http://hakia.com/>
3. <https://duckduckgo.com/>
4. Yu, L.: A Developers Guide to the Semantic Web. Springer, Berlin (2011). ISBN: 9783642159695
5. Antoniou, G., Groth, P., van Harmelen, F., Hoekstra, R.: A Semantic Web Primer, 3rd edn. ISBN: 0262018284
6. Segaran, T., Taylor, J., Evans, C., O'Reilly (2009) Programming the Semantic Web (2009). ISBN: 9780596802066
7. Mukhopadhyay, D., Banik, A., Mukherjee, S., Bhat-tacharya, J., Kim, Y.-C.: A Domain Specific Ontology Based Semantic Web Search Engine. In: Proceedings of the 7th International Workshop (MSPT), p. 8189, Feb 5 2007. ISSN 1975-5635, 89-8801-90-0
8. FinKelstein, L., Gabrilovich, E., Matias, Y.: Placing search in context: the concept revisited. In: Proceedings of the 10th International Conference on World Wide Web, pp. 406–414 (2001)
9. Cafarella, M.J., Etzioni, O.: A search engine for natural language processing. In: Proceedings of the International World Wide Web Conference Committee (IW3C2), ACM 1595930469/05/0005 (2005)

10. Karpagam, G.R., Uma Maheswari, J.: A conceptual framework for ontology based information retrieval. *Int. J. Eng. Sci. Technol.* **2**(10), 5679–5688 (2010)
11. Jiang X., Tan, A.-H.: OntoSearch: a full-text search engine for the semantic web
12. Lei, Y., Uren, V., Motta, E.: SemSearch: a search engine for the semantic web. In: *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks*, pp. 238–245. Springer, Berlin (2006)
13. Kruse, P.M., Naujoks, A., Rsner, D., Kunze, M.: Clever search: a WordNet based wrapper for internet search engines, computing research repository CORR, vol. Abs/cs/050 (2005)
14. Lara, R., Han, S.-K., Lausen, H., Stollberg, M., Ding, Y., Fensel, D.: An evaluation of semantic web portals. In: *Proceedings of the IADIS Applied Computing International Conference 2004, Lisabon, Mar 23–26 2004*
15. Madhu, G., Govardhan, A., Rajinikanth, T.V.: Intelligent semantic web search engines: a brief survey. *Int. J. Web Semant. Technol.* **2**(1), 34–42 (2011)
16. Kalaivani, S., Duraiswamy, K.: Personalized semantic search based intelligent question answering system using semantic web and domain ontology. In: *Proceedings of the International Conference on Advanced Computer Technology (IJCA)*, pp. 15–17 (2011)
17. Kerschberg, L., Kim, W., Scime, A.: Intelligent web search via personalizable Meta—search agents
18. Stanford Parser: <http://nlp.stanford.edu/software/index.shtml>
19. <http://www.anc.org/data/masc/downloads/data-download/>
20. Lei, Y., Lopez, V., Motta, E.: An infrastructure for building se-mantic web portals. In: *Proceedings of the International Workshop on Web Information Systems Modeling (WISM)*, pp. 283–308 (2006)

Chapter 54

Integration of Mobile Based Learning Model Through Augmented Reality Book by Incorporating Students Attention Elements

Zarwina Yusoff, Halina Mohamed Dahlan and Norris Syed Abdullah

Abstract The limitation of current e-learning technology has caused a lack of student attention in educational environment. Therefore, this study describes the integration of mobile based learning through Augmented Reality Environment to incorporate the student attention elements by computer-generated content. To incorporate student attention element, this study propose an integration model of mobile learning by utilizing Augmented Reality Environment. To validate the integration model, this study has developed the AR prototype called AF-LAR (Animal Fun Learning—Augmented Reality) through smart phone technology as mobile based learning device. (AF-LAR) is an AR Book that adapts learning concept via mobile devices and enables student to bring and access the learning content anywhere and anyplace. AF-LAR has been developed using Metaio and Junaio channel while the design is incorporated with visual learner styles from Visual Auditory Kinesthetic (VAK) model with information visualization approach. The result of this study was evaluated by using attention element in Keller’s Motivation ARCS Model to prove the student attention attribute from the prototype based on the five experts. Descriptive statistics was chosen as a technique to evaluate the mean and reliability of attention attribute; perceptual arousal (PA), inquiry arousal (IA) and variability (V) inside the proposed integration model to achieve the student attention in education environment.

Z. Yusoff (✉) · H.M. Dahlan · N.S. Abdullah
Department of Information System, Faculty of Computing, Universiti Teknologi Malaysia,
Skudai, 81310 Johor Bahru, Malaysia
e-mail: zarwina.utm@gmail.com

H.M. Dahlan
e-mail: halina@utm.my

N.S. Abdullah
e-mail: norris@utm.my

Keywords Augmented reality · Education · Mobile learning · Student attention

54.1 Introduction

As far as the progress of e-learning technology as a medium in learning process, many researchers try to improve the learning process by utilizing technology in particular, towards in enhancing students attention. Today, many mobile devices such as cell phones, notebooks and tablet computer are gaining more relevance to learning environments and education. The recent progress of Augmented Reality (AR) has brought this technology to the mobile computing area to generate more user experience on it. Hence, further facilitates the growth of user experience with AR environment. Furthermore, the emerging of technological concepts in education was generating the rapid development of education where it was bringing it into mobile based learning technology. AR is an effective ways to visualize the learning content, researchers was proven that AR supported particular learning activities such as problem solving, in a highly interactive and memorable fashion [1]. This emerging technology can also be used in any fields to facilitate user in visualizing the object or information in any particular situations.

The main objective of this paper is to embed AR into mobile learning to improve the learning technology in classroom using the AR. In order to actualize this aim, there are other objectives that need to be achieved which are:

- i. To propose the model integration of mobile learning through Augmented Reality Book by incorporating student attention elements.
- ii. To evaluate the prototypes of mobile based learning application through Augmented Reality Book.

Apart from this objective, we were chosen the concept of AR Book to implementing the model integration to be tested in enhancing the student attention.

54.2 Augmented Reality Book for Education

The implementation of AR for mobile learning purposes has been discussed and also presented in a number of studies. In order to achieve the objective, this paper is searching on current model integration of mobile learning to enhance the limitation through achieving the objective in this study. The visualization in AR must be embedded in mobile learning architecture to ensure the mobile learning through AR are fluently can use in learning environment to improve the learning process.

In this research, the process of embedding of educational elements in the mobile AR for learning system derived from AR Magic Books that was developed by Billingham et al. [2] in enhance the traditional normal text books through Augmented Reality Concept [2].

The integration of AR through mobile also successfully was done by Ramdas et al. [3] which was developed the AR Eco System where the concepts of modules are based on four categories; player, context awareness, type of learning and mode of interaction [3]. To get the effective ways in enhancing learning experience among learners; we proposed that, the integration of AR in mobile learning will be implementing through AR Book where this will enhance the concept of Magic Book that proposed by [2].

In order to actualize the objective, image based tracking techniques is chosen to make the mobile phones camera can track the virtual object through the AR book. Besides that, Oh and Woo [4] was developing the AR Gardening System that implements the personalized pedagogical agent through the animated pedagogical agents to improve student's learning experience in educational system [4]. The gardening system that consist the one garden and include flower, blue bird, scoop and sprinkle water to learn the process of growing flower in garden. Each of the objects is representing by a marker to make the tracking process.

54.3 Methodologies

To facilitate the evaluation process, the model was realized in a form of mobile based learning application. The mobile based learning application incorporates all the components of the models. In ensure this research successfully conducted, this section briefly explain the participants, data collection and data analysis that support the evaluation process.

54.3.1 Participants

Five participants of experts from five criteria were chosen to get the analysis in the limitation of current e-learning system in educational environment which conducted through questionnaire. The questionnaire is design to identify the limitation of current e learning technology in educational environment to achieve student attention during learning process which the criteria in the questionnaire will be map with the model integration and will be evaluated through the prototypes implementation. This questionnaire is distributed into five experts as a targeted respondent to identify the problem and limitation in the usage of technological element in learning environment in order to improve student attention.

Table 54.1 Finding for data collection

Limitation of element in e-learning technology [5]	Problem analysis	Factors to achieve student attention through mobile learning model [6]
Technological and hardware limitation	1. The limitation of traditional method in learning process	Technology
	2. How the technological tools can help educational environment	
	3. The current limitation of e-learning courseware technology	
	4. The limitation of current educational application software	
Design limitation	1. The appropriate combination color to attract student attention	Pedagogy
	2. The effective and efficient ways in conveying learning content through visualization approach	
	3. The visualization approach to maintain the student attention in classroom	
	4. The suitable subject to be adapting in implementing learning through mobile	
Personal issues	1. Factor that causes student lack of attention in learning and teaching process.	People
	2. The suitable approach to improve student attention in a classroom	
	3. The factor of mobile based learning should to be implemented	

54.3.2 Data Collection

In this research study, the data collection about the current limitation of e-learning technology was get from the five experts through the questionnaire. The Question for the data collection is based on the three main limitations in e-learning technology that identify Mahanta [5] which consist by technology and hardware limitation, design issues and personal issues [5]. Based on that, eleven question was derives from the factor to analysis the current limitation of e-learning technology. Table 54.1 listed the data collection and analysis.

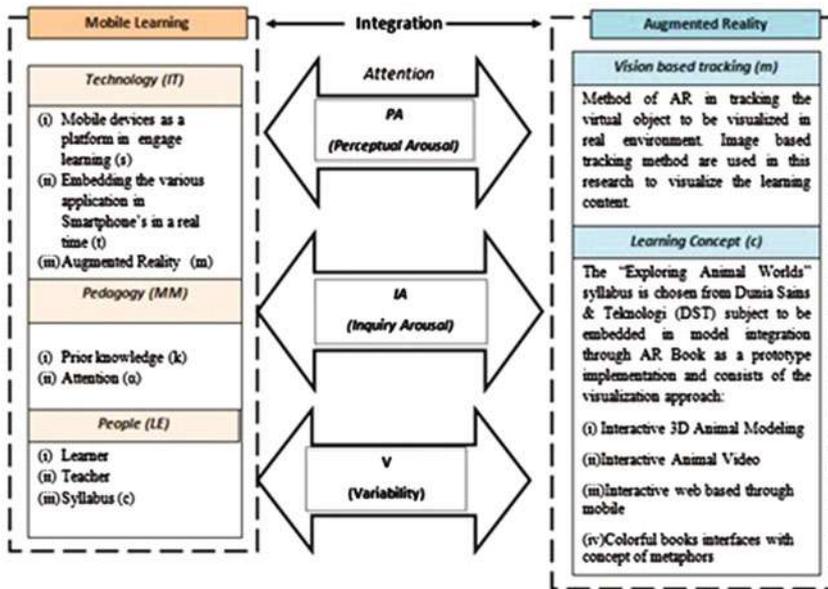


Fig. 54.1 Integration model of mobile learning based on augmented reality

54.3.3 Data Analysis

Statistical Package for Social Science (SPSS) was used to measure the mean resulted from the evaluation. The descriptive statistics was used to measure the prototypes in improving the student attention.

The process development of this prototype is start with designing the system flowchart where the design consists of two parts, the Book interface and the AR system. Analyzing the user is the first step in the development of AF-LAR Book (Fig. 54.1).

54.4 Findings

The conceptual of mobile learning model is derived from Prasertsilp [6] where mobile learning was based on two factors, mobile learning environment and learning outcomes [6]. This research identified the model to be integrated with mobile learning system so that it can contribute in improving the limitation of current mobile learning system because it covers the three aspects of mobile learning environment; users, technology and pedagogy to be implementing through AR. Figure 54.2 presented the model integration.

Mobile Learning dimension consists of three main attributes; technology, people and pedagogical elements while educational dimension consists of type of

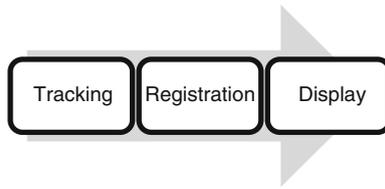
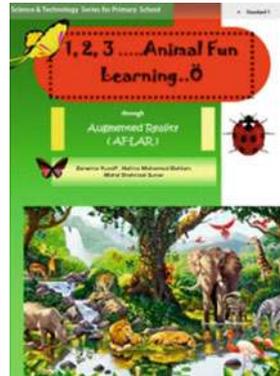


Fig. 54.2 Processes in augmented reality technology

Fig. 54.3 The cover of AF-LAR



learning and mode of interaction. The model of mobile learning is derives from Pratersilp [6], where it consist of from three main factors; technology, pedagogy and people and the attributes inside of mobile learning are derives from Laouris and Eteokleous [7] which consists of space (s), time (t), method (m), prior knowledge (k), attention and learning environment (LE) [6, 7].

People are the target of user who will use the mobile learning application which consists by student in any related subject to enhance their learning experiences during class. The learning environment (LE) is a people which conducting the learning session successfully. In these factors, it consists by learner as a student who wants to gain their knowledge, using the AR Books to adding their information. Besides that, people in learning environment (LE) is also consists by teacher in give instruction to students for completing their task. Based on the figure, the attribute of student attention is support by element attention in ARCS model to evaluate the model integration.

Variability is used to evaluate the model integration through the AR Book prototypes. In variability aspects, it emphasized the variety tactics to maintain the attention through the mobile technology, teacher’s instruction, and integrated of learning content.

In order to emphasize the model integration, Vision based tracking was used to visualize the AR object. The main of Augmented Reality process are consists of three main phases; tracking, registration and display the AR content.



Fig. 54.4 Image based marker



Fig. 54.5 Interactive animation and video



Fig. 54.6 Interactive website about animation

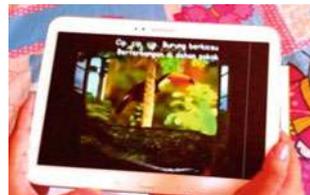


Fig. 54.7 Animal song through video

Table 54.2 AF-LAR design

Animal fun learning book	Results
(a) Book cover that shows the adaptation of visual learner concept to generate student interest in explore the animal world through the features below:	Fig. 54.3
i. The colorful graphics	
ii. The enjoyable concept through the images and pictures	
iii. The text styles is suitable for primary schools student	

Based on the model integration, the mobile learning attributes are strongly support the student attention when realizing it through prototypes implementation. The embedding of educational elements in the mobile AR for learning system derived from Ramdas et al. [3] developed the AR Eco System where the concepts of modules are based on four categories; player, context awareness, type of learning and mode of interaction [3]. To get the effective ways in enhancing learning experience among learners; we proposed that, the integration of AR in mobile learning will be implementing through AR Book. The book shows it displaying information visualization through the AR Book, where its focus on visualization and interactivity to improve student learning experience in educational systems. In order to actualize this research objective, the concept of AR Book was chosen to embed the model integration and to validate the conceptual framework of mobile through AR in support student attention to perform successfully. Based on that, this study was proposed the AR Book through mobile display that called AF-LAR (Animal Fun Learning through AR) that embedding with visual learner styles for designing and visualization approach through Augmented Reality Technology in order to improve student attention. Table 54.2 shows the results of AF-LAR design.

Based on the Table 54.2, the concept of designing AF-LAR is very interesting to get student attention in order to provide platform to tracking the AR object based on image based marker in this book through mobile. In aligning with primary schools characteristics, the concept of thematic stories and short stories are used to be adapt with the level of student in reading the information about the animal, in this case the concept of people in mobile learning model are emphasize which is, student must have their teacher or facilitator to help them in learning session. In order to construct the AR element to be integrated with the mobile, Table 54.3 illustrates the concept of Mobile AR and the explanation results in AF-LAR in achieve student attention.

The process of evaluation is implemented in objective to achieve the model integration that supports student attention through AF-LAR. This evaluation process has implemented in targeted respondent, which is consist by five experts that have some criteria using Samsung Galaxy Tab 3 10.1 and AF-LAR Tangible book as a marker based tracking. The evaluation process is based on the Attention element that derives from ARCS Motivation Model that support by Keller 2000. Table 54.4 presented the explanation of attention criteria that used to be evaluated in AF-LAR.

Table 54.3 Displaying learning content through mobile AR

Mobile AR	Results
Image based marker to display an AR object using mobile camera detection	Fig. 54.4
(a) The picture background is based on jungle situations to make student feel more real with the animal world	
(b) During image based marker displays 3D object, the video about the animal presented	Fig. 54.5
(c) Visual mode presents the interactive 3D graphics and auditory mode present the video	
(d) Learning through interactive web based also provides in showing learner the animal website with the real environment	Fig. 54.6
(e) Learning through animal song, more interactive and attract student attention to sing together through the lyrics	Fig. 54.7

Table 54.4 Explanation about attention criteria

Attention	Explanation
Capture interest (Perceptual arousal)	What I can do to capture their interest?
Stimulate inquiry (Inquiry arousal)	How I can stimulate attitude of inquiry?
Maintain attention (Variability)	How I can use a variety of tactics to maintain the attention?

Based on the perceptual arousal, we can capture the student interest through the colorful images, picture, graphic element in the book, and usage of mobile and AR as a part of technology. To stimulate of an attitude of inquiry by learners experiences, it need the basic knowledge about the learning information and to achieve the variability elements, it can implemented through the mobile technology, teachers instructions and integrated of learning content. In order to full fill the element of attention in Table 54.4, we identify eleven questions that representative as Perceptual Arousal (PA), Inquiry Arousal (IA) and Variability (V) to evaluate the prototypes. In achieve the research contribution; Table 54.5 below indicates the details of descriptive statistic for the prototype testing results. Based on that table, descriptive statistics on mean was performed to describe the level of student attention on five experts' participation.

Perceptual Arousal (PA1) in Table 54.5 represents “visualization based approach is the best method in enhance student attention” has a spread of 0.447 over 4.80 mean with the maximum score is 5. Perceptual Arousal (PA2) in Table 54.5 represents “Smartphone’s is the flexible ways to be used in educational environments” has a spread of 0.548 over 4.40 mean with the maximum score is 5. Perceptual Arousal (PA3) in Table 54.5 represents “The features of AR that allows multimedia element such as graphic, audio, video, and interactivity can able to attract student attention during learning session” has a spread of 0.548 over 4.60 mean with the maximum score is 5. Perceptual Arousal (PA4) in Table 54.5 represents “The usage of multimedia element can help the student attention in

Table 54.5 Results of prototype testing

Item	Likert scale points						Mean	Standard deviation
	Strongly disagree	Disagree	Neutral	Agree	Strongly agree			
PA 1				1	4	4.80	0.447	
PA 2				3	2	4.40	0.548	
PA 3				2	3	4.60	0.548	
PA 4				2	3	4.60	0.548	
IA 1				1	4	4.80	0.447	
IA 2				2	3	4.80	0.447	
IA 3			1	2	2	4.20	0.837	
V 1				2	3	4.60	0.548	
V 2				2	3	4.60	0.548	
V 3			1	1	3	4.40	0.894	
V 4				1	4	4.80	0.447	

mobile learning” has a spread of 0.548 over 4.60 mean with the maximum score is 5. Based on the results of Perceptual Arousal, all the mean is >2.5 that indicates the element of perceptual arousal in the proposed model is strongly emphasized the technology element is support student attention attributes.

The attributes; are spaces (s), real time (t) and AR (m) a strong combination technology that used in enhances student attention. Spaces are provided through mobile devices and AR is a method that applying objects visualization in the learning content. Apart from that, real time (t) is represent by technology of AR and the prototypes is mobile is run in a real time to providing the spaces for students to conduct their learning activity. In AR, the element of technology also integrates with vision based tracking technique that conducted to produce AR content in the mobile devices. Based on the results, it proves the model integration of mobile learning and AR through technological element is strongly emphasized can improve student attention.

Inquiry Arousal (IA1) in Table 54.5 represents “The lack of student attention in mobile learning can be avoid through visualization approach” has a spread of 0.447 over 4.80 mean with the maximum score is 5. Inquiry Arousal (IA2) in Table 54.5 represents “The concept in Augmented Reality Technology is very appropriate to be adapted in Science Subject” has a spread of 0.447 over 4.80 mean with the maximum score is 5. Inquiry Arousal (IA3) in Table 54.5 represents “Visualization can give easier to student for remember the learning content because the conveying of object and information graphically to student” has a spread of 0.837 over 4.20 mean with the maximum score is 5.

Based on the results of Inquiry Arousal (IA), all the mean is >2.5 that indicates the entire element in inquiry arousal are support of attention element in pedagogical of mobile learning model. The attributes of student attention in pedagogical element are prior knowledge (p) and attention (α) is achieved to support the proposed model integration that proposed in Chap. 5. Apart from that, the result proves the element of attention attributes is embedding in the prototypes successfully to achieve student attention.

Variety of tactics 1 (V1) in Table 54.5 represents “The features of Smartphone’s that ubiquitous, easy to bring anywhere and can be access the internet is the main factor to encourage learning through mobile” has a spread of 0.548 over 4.60 mean with the maximum score is 5. Variety of tactics 2 (V2) in Table 54.5 represents “Learning through mobile is the new methods which must to introduce and monitored by teachers”. Has a spread of 0.548 over 4.60 mean with the maximum score is 5. Variety of tactics 3 (V3) in Table 54.5 represents “Mobile technologies can help teachers in get student attention” has a spread of 0.894 over 4.40 mean with the maximum score is 5.

Variety of tactics 4 (V4) in Table 54.5 represents “Augmented Reality technology is very appropriate to be integrated with mobile technologies” has a spread of 0.447 over 4.80 mean with the maximum score is 5.

Based on the variability results, it strongly supports the people element in the model integration because all the variability analysis is >2.5 . The people element in model integration are consists by teacher, learner and syllabus (c) as a learning environment (LE). Variability result was proves the people element has a strong relation with learning concept through visualization approach in Augmented Reality.

Learner and teacher is interdependencies attributes to ensure the learning session is fluency. Besides that, syllabus is needed to be embedding into prototypes development. Prototypes evaluation is implemented as the ends phase in the research methodology and to measure the model validation by incorporate student attention elements. To validate the accuracy of data, the calculations of mean and standard deviation are used to extract the results in order to evaluate the prototypes and the Cronbach’s Alpha also is implemented to found the reliability statistics among the question. As a conclusion, the mean of the prototypes is >2.50 , that means the prototypes can improve student attention through the model integration of mobile learning through Augmented Reality.

54.5 Conclusion

This paper solved limitation of current e-learning technology by improving student attention in educational environment. This paper proposed the integration of mobile based learning through Augmented Reality Environment by incorporating the student attention elements. The objective of this paper was achieved when the elements of student attention that embedded in the AR Book can improve student

attention based on expert perspective. The content of AR Book that display through mobile devices was proves the new model integration can be used in order to improve student attention in educational environment. Based on the finding, it was proven that the utilizing of technology can enhance the student attention through the AR Book concept. However, the mobile devices used must have the internet connection to ensure the right channel can be loaded to visualize the learning content. For the future work, a robust AR application is proposed in displaying the AR content during the tracking process.

References

1. Luckin, R., Stanton Fraser, D.: Limitless or pointless? an evaluation of augmented reality technology in the school and home. *Int. J. Technol. Enhanced Learn.* **3**(5), 510–524 (2011)
2. Billinghamurst, M., Kato, H., Poupyrev, I.: The magic book: a transitional AR interface. *Comput. Graph.* **25**(5), 745–753 (2001)
3. Ramdas, C.V., Parimal, N., Utkrash, M., Sumit, S., Ramya, K., Smitha, B.P.: Application of sensors in augmented reality based interactive learning environments. In: *Proceedings of the IEEE , 6th International Conference on Sensing Technology (ICST)* (2012)
4. Seijin, O.H. Woontack, W.O.O, Augmented Gardnering System with Personalized Pedagogical Agents. *International Symposium on Ubiquitous Virtual Reality (ISUVR), CEUR-WS Proceedings*, **62**, 17–18, (2007)
5. Mahanta, D., Ahmed, M.: E-learning objectives, methodologies, tools and its limitation. **1**, 46–51 (2012)
6. Prasertsilp, P.: Mobile learning: designing a socio-technical model to empower learning in higher education mobile learning: designing a socio-technical model to empower learning in higher education. **2**(1) (2013)
7. Laouris Y.: Eteokleous, N.: We need an educationally relevant definition of mobile learning. In: *Proceedings of the 4th World Conference on Mobile Learning, mLearn, Cape Town, South Africa, Oct 25–28 2005*
8. Keller J.M.: How to Integrate learner motivation planning into lesson planning. The ARCS model approach. Paper presented at VIISemanario, Santiago, Cuba (2000)

Chapter 55

Law Reckoner for Indian Judiciary: An Android Application for Retrieving Law Information Using Data Mining Methods

S. Poonkuzhali, R. Kishore Kumar and Ciddarth Viswanathan

Abstract Data mining is an emanating area not only in the field of computer science but also contributes more to knowledge mining and information management. This analysis mainly focuses on organizing the Indian judicial system which enables the lawyers and law students to access the constitutions details, code for civil and criminal procedure in an easier and faster way. In this work android is used as a medium for this analysis because smart phones are being used by 9 persons out of 10 in the legal society. The user of this application can search using keywords and all sections from the constitution relating to the keyword are displayed along with review of related procedures either civil procedure or criminal procedure as mentioned above. The major advantage of this application is that it reduces the time for the lawyers and law students to search using android phones rather than to do survey on books and materials related to their case. In this proposed work all the legal data are classified using ID3 classification algorithm which results in efficient searching.

Keywords Classification · Civil procedure · Criminal procedure · Indian constitution · Keyword search

S. Poonkuzhali (✉) · R. Kishore Kumar · C. Viswanathan
Rajalakshmi Engineering College, 602105 Chennai, India
e-mail: Kuzhal_s@yahoo.co.in

R. Kishore Kumar
e-mail: rskishorekumar@yahoo.co.in

C. Viswanathan
e-mail: ciddarthviswanathan.2009.it@rajalakshmi.edu.in

55.1 Introduction

Data mining, the extraction of concealed analytical information from large databases, is a powerful new technology with great potential to focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors which help to make proactive, knowledge-driven decisions. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. Knowledge management (KM) is an effort to increase useful knowledge within the organization.

Though vast researches are being carried out in the area of knowledge management, yet it requires a wider acceptance from people from various fields for data/pattern analysis. As most of the judicial data contains large number of attributes, data mining techniques such as feature relevance analysis and classification are required to perform efficient searches. Among the data mining methods, classification plays a major role in analyzing the clinical data. Some of the frequently used classifiers in this domain are ID3, Naive Bayes, C4.5, CART, Multi layer perceptron, KNN, J48, Random forest and Support Vector Machines.

As the Indian judiciary is the largest democracy of the world, this research is carried on the attributes of the Indian constitution and more significantly in the civil and criminal procedures. Moreover it carries out many critical functions such as classifying the Indian judiciary based on civil and criminal domains. The most common judicial terms are the organization of the state, how a state and unions are formed, what is the legal procedure for civil cases and criminal cases.

55.2 Related Works

Silwattananusarn and Tuamsuk [6] presented about data mining and knowledge management process. The Knowledge management rationale and major knowledge management tools integrated in the knowledge management cycle are described. Finally the applications of the data mining techniques are summarized. The limitations in this paper are it just provides generalization of the knowledge management process and does not provide any specific details and moreover data integration is not addressed. Kumar Varma [4] presented the use of data mining concepts to match data from multiple sources in order to enrich the data and improve its quality on the whole. It is also used for the Integration of the databases on-line negotiations and interactions. The major drawback on this system is that it causes bottlenecks in distributed environment. Srivatsan et al. [7] introduced how to use the contents of a shopping cart to predict what else the customer is likely to

buy. Association rule mining, boolean vectors and prediction techniques are used in this paper. The major drawbacks of this system are a lot of complex algorithms are used which may lead to loss in efficiency and speed. And also it predicts based on the existing items on the cart and does not adapt dynamically. Mohamed et al. [5] explained, record Linkage is the computation of the associations among records of multiple databases. In a framework where entities are unwilling to share data with each other, the problem of carrying linkage computation without full data exchange is called private record linkages. This paper provides improvements when compared with previous techniques such that (i) they make no use of a third party and (ii) they achieve much better performance in terms of execution time and quality of output. The major drawbacks of this system are it cannot work in distributed systems and it does not use creative techniques for evaluation. Kishore Kumar et al. [3] uses various data mining techniques to classify spam dataset.

55.3 Methods and Materials

In this proposed system an android application named Law Reckoner was developed to classify and search Law Information quickly. An Android mobile service is made available for the user to log into the server. The information provided by the end user travels through the network to the server. Then the server validates the request and sends back the response to the android mobile. If the person is a registered user it allows the person to search through the constitutions or else it shows a error message and ask the user to register. In registration phase the user needs to submit user name, password and other details to the server. All information is stored in the database via server for future purpose. Then the user will start to search for the law details via choosing the category that they want to search. Once they choose the category, they have to provide the other information like chapter number, article number and part number and submit the request to the server. The general architecture of the proposed work is shown in Fig. 55.1.

This is implemented by creating an application using android and installing the application in the user Smart phone. By using this application the user can communicate with the server via GPRS connectivity as in Fig. 55.2. Once the user sends the request to the server, the server will validates the request and display the law details in the users Android Smart phone as shown in Fig. 55.3. Once the server receives the user's request, it will search for the requested data in the law database and classify them using decision tree induction (ID3) algorithm and display the relevant data to the user's Smart phone itself. ID3 algorithm builds the tree in a top down manner starting from a set of objects and a specification of

properties. At each node of the tree, a property is tested and the results used to partition the object set using information gain as a splitting criteria based on entropy measure. This process is recursively done till it contains objects belonging to the same category. ID3 can handle high-cardinality predictor variables. ID3 doesn't apply any pruning procedure. It doesn't handle numeric attributes or missing values. Figures 55.4, 55.5, 55.6, 55.7 and 55.8 shows the results of this application.

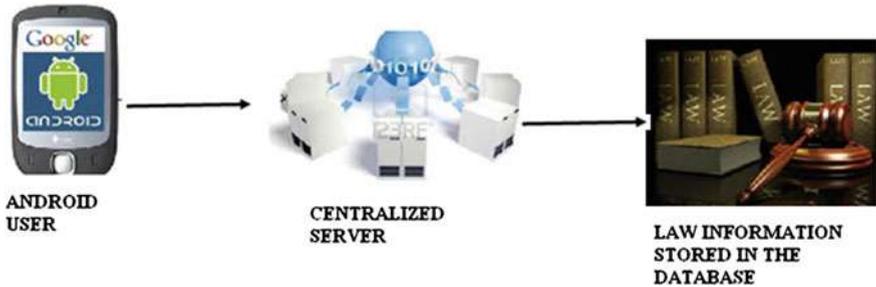


Fig. 55.1 Architecture of the proposed system



Fig. 55.2 User registration

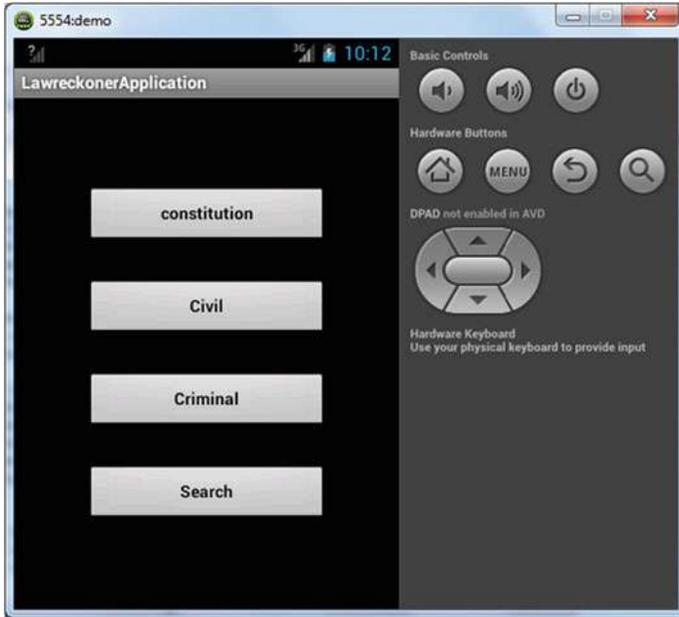


Fig. 55.3 Main menu to choose the category

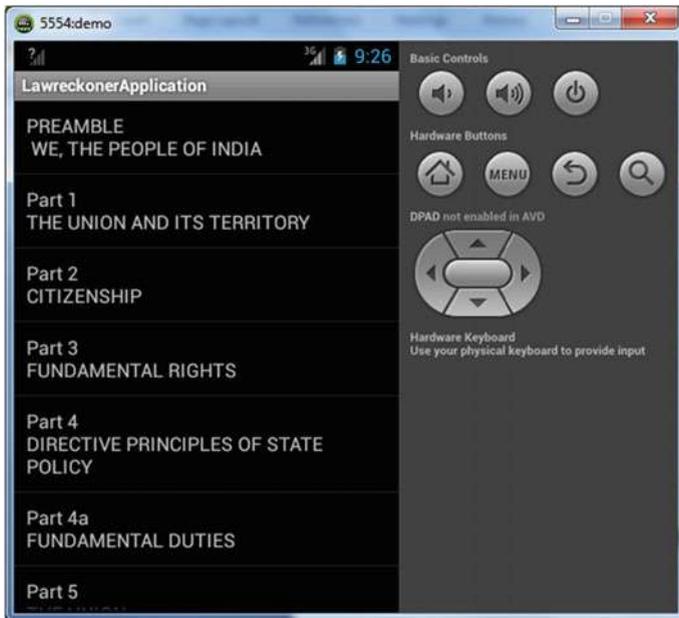


Fig. 55.4 Sub menu to constitution

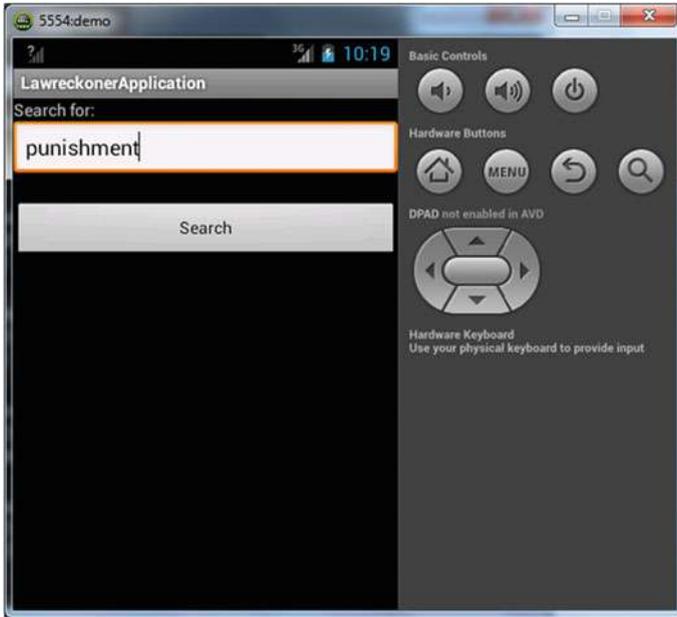


Fig. 55.5 Search menu of the proposed system

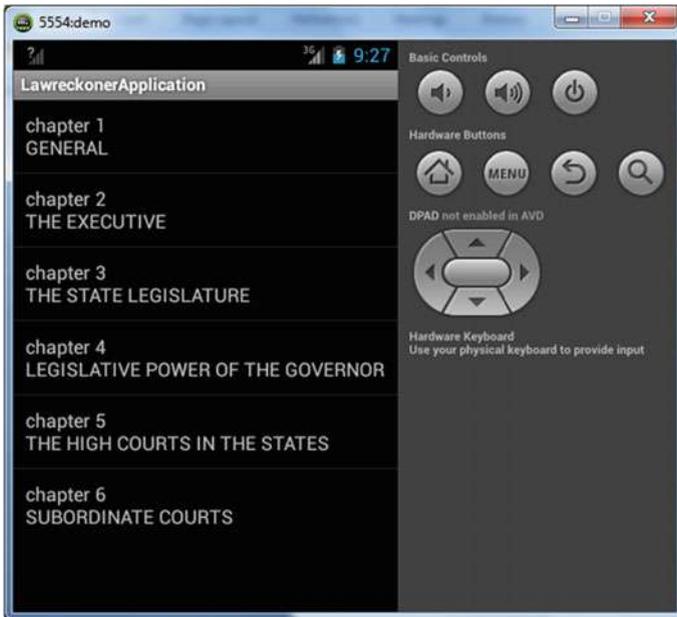


Fig. 55.6 On click the main menu chapters will be listed

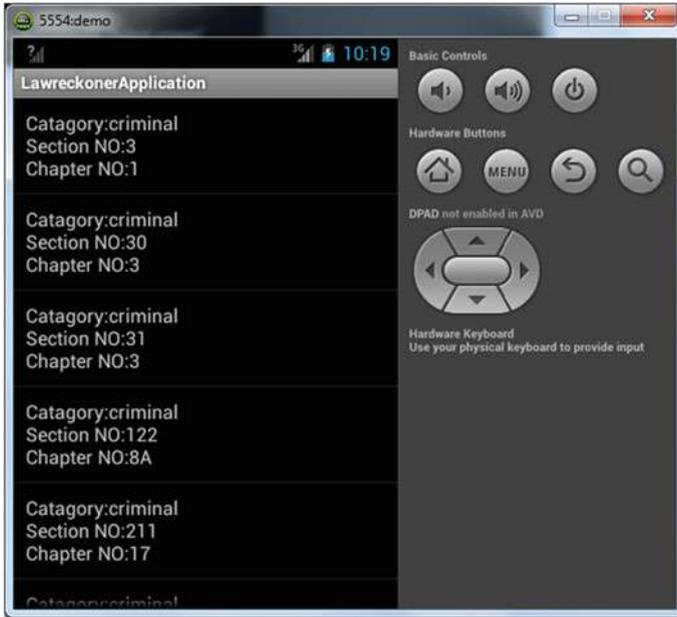


Fig. 55.7 List of Punishment under criminal category

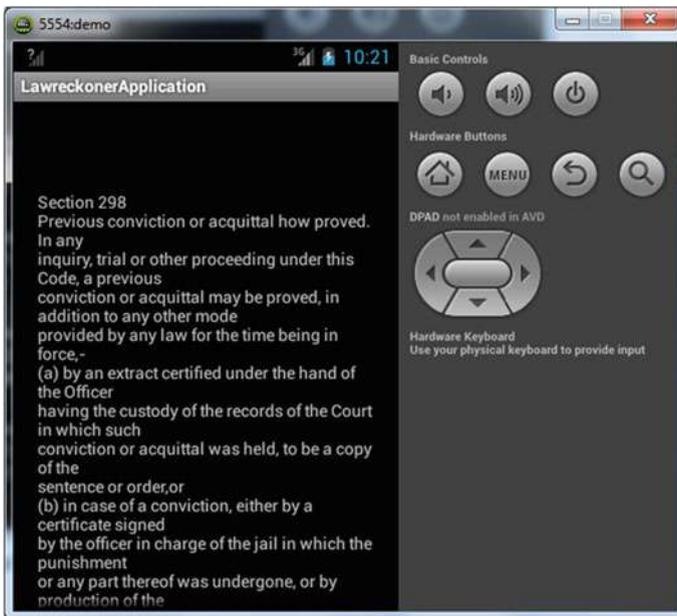


Fig. 55.8 Punishment under criminal category

55.4 Pseudo for ID3 Classification Algorithm

Input: Law data.

Output: Decision Rules.

Step 1: Extract the input training dataset (TD).

Step 2: Let IA be input attributes & TA be the Target Attributes.

Step 3: For a base class construct the tree as follows:

Step 3A: If TD is empty then

Return a single node with the value Failure

Step 3B: If all instance from TD have the same class value for TA then

Return the single node with that class value.

Step 3C: If IA is empty then

Return a single node with the most frequent value of TA in TD.

Else proceed to step 4

Step 4: Compute the information gain for all IA relative to TD based on Entropy measure.

$$Info(D) = - \sum_{i=1}^m (p_i \log(p_i)) \quad (55.1)$$

$$E(A) = - \sum_{j=1}^v (|D_j|/|D| * Info(D)) \quad (55.2)$$

Where $|D_j|/|D|$ is weight of the jth partition $Gain(A) = Info(D) - E(A)$

p_i is the probability that arbitrary instance in TD

$E(A)$ is the entropy of the attribute A

$Gain(A)$ is the information gain of attribute A

$Info(D)$ expected information needed to classify an instance in TD

Step 5: Let X be the split attribute with largest information gain among IA

Step 5A: let $X_j | j = 1, 2, \dots, m$ be the value of X.

Step 5B: Let $T_j | j = 1, 2, \dots, m$ be the subsets of T when T is partitioned.

Step 6: Partition the TD according to the value of X

Step 7: Return trees with the root node labeled X and arcs labeled X_1, X_2, \dots, X_m

Step 8: Repeat step 4 to 6 until it contains instants belonging to the same category.

55.5 Conclusion

This research work presents an efficient application for lawyers and law students across the country which gives them mobile access for law related data such as Indian constitution, code of civil procedure and code of criminal procedure. In addition to information retrieval, an efficient search mechanism is incorporated with which they can perform precise and accurate searches using article number, section number, chapter number, part number etc. Future works can be carried out for:

- To provide more content like the Indian penal code and a database of Indian origin.
- To create a law dictionary to search the meanings of certain legal terms
- To reveal a case history for similar occurrences of the articles and its actions.

References

1. Gupta, G.K.: Introduction to Data Mining with Case Studies. Prentice-Hall India, New Delhi (2006)
2. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Jose (2001)
3. Kishore Kumar, R., Poonkuzhali, G., Sudhakar, P.: Comparative study on email spam classifier using data mining techniques. In: Proceeding of the International Multiconference on Engineers and computer scientists, Vol I (2012)
4. Kumar Varma, M.V.K.: Efficient techniques for online record linkage. *Int. J. Comput. Trends Technol.* **3**(3), 321–324 (2012)
5. Mohamed, Y., Mikhail, J.A., Ahmed, E.: Efficient Private Record Linkages. Proceedings of the 2009 IEEE International Conference on Data Engineering, IEEE Computer Society, Vol. 4, pp. 1283–1286 (2009)
6. Silwattananusarn, T., Tuamsuk, K.: Data mining and its applications for knowledge management: a literature review from 2007 to 2012. *Int. J. Data Min. Knowl. Management Process.* **2**(5), 13–24 (2012)
7. Srivatsan, M., Sunil Kumar, M., Vijaya Shankar, V., Leela Rani, P.: Predicting missing items in shopping carts using fast algorithm. *Int. J. Comput. Appl.* **21**(5), 35–41 (2011)

Chapter 56

Enhancing the Efficiency of Software Reliability by Detection and Elimination of Software Failures Through Univariate Outlier Mining

S. Poonkuzhali, R. Kishore Kumar and R. Kumar

Abstract In this competitive market, there is always a growing demand for providing high quality software in all aspects of societal services: education, industry, defense, travel, health, retail, telecommunications and so on. As the size and complexity of software grows drastically, the demand for highly complex software systems also increases rapidly which in turn raises the number of software failures. These software failures not only degrade the business performance but also cause more economic damages. Therefore, detection and elimination of these software failures becomes a vital importance for better software reliability which is a measure of quality. In this proposed work, identification of software failures is done through outlier mining which mainly focuses on rare and infrequent patterns like faults, noise, irrelevant and redundant data. After the removal of these software defects, classification algorithms are applied to provide a better and more reliable result.

Keywords Classifiers · Error rate · Defects · Outlier mining · Quality · Software reliability

56.1 Introduction

In recent years, more number of software was used for decision making process in various fields. As software becomes voluminous, sophisticated in complexity, and originated by integration of multiple components, it is an increasingly challenging

S. Poonkuzhali (✉) · R. Kishore Kumar · R. Kumar
Rajalakshmi Engineering College, Chennai 602105, India
e-mail: Kuzhal_s@yahoo.co.in

R. Kishore Kumar
e-mail: rskishorekumar@yahoo.co.in

R. Kumar
e-mail: kumar.r@rajalakshmi.edu.in

task to ensure software reliability. Software Reliability is the application of statistical techniques to data collected during system development and operation to specify, predict, estimate, and assess the reliability of software-based systems. From a knowledge perspective, the analysis of executions of a buggy program is essentially a data mining process which traces the data generated during program executions containing both relevant patterns and outliers that may help the discovery of software bugs. In general, the design for reliability techniques includes: fault avoidance, fault detection, masking redundancy, and dynamic redundancy. An efficient data mining technique can improve the software reliability by analyzing, predicting and removing software faults in advance.

Data mining is a process of extracting hidden and useful information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, valid and of high quality. Finding outliers is an important task in data mining. Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community [1]. Early detection of outliers not only reduces the risk of making poor decisions based on erroneous data but also aids in identifying, preventing, and repairing the effects of malicious or faulty behavior. Outliers are observations that deviate so much from other observations to arouse suspicions that they might have been generated using a different mechanism. Outliers may also reflect the true properties of data from rare and interesting events which may contain more valuable information than normal data. Outlier mining is dedicated to find data objects which differ significantly from the rest of data.

In the proposed system Univariate outlier detection technique is used to detect and remove the sample that contains software defects. The classification algorithms are applied to the original dataset as well as after applying outlier detection techniques. Then the error rate of various classifiers is compared. From the accuracy of various classifiers, it is proved that the classification results after applying outlier detection technique gives quality software products which in turn improves software reliability.

56.2 Related Works

56.2.1 Software Reliability

Go et al. [2] have proposed that there is a need to model an approach that is capable of considering the architecture and evaluating the reliability by taking into account the interaction between components, reliability of components and reliability of interfaces with other components. Xie et al. [3] have proposed practical method to predict the software reliability using test data. Two major parameters are used to evaluate reliability while testing one is total number of initial faults and fault detection rate in testing. Hence the software reliability of the new system

could be predicted from the test data of similar systems. Huang and Lin [4] classified faults into two types, dependent faults and independent faults. Mutual independent faults are those that can be removed directly. Dependent faults could be removed only if leading faults are removed. Higher proportion of dependent faults affects the reliability of the system. As the reliability requirement increases, time taken to achieve the reliability also increases hence increases the cost involved to achieve reliability. Robert et al. [5] stated that reliability increases as the software is improved. The concept of visit counts is used to represent the criticality of the component based on the combination of depth of the component, number of user functions called and number of test cases written for that component. System reliability is calculated by considering the individual reliability of the component along with the number of visits to each component. Podgorelec [6] applied on a real-world software reliability analysis dataset based on evolutionary induced decision trees to identify the outliers.

56.2.2 *Outlier Mining*

The distribution based methods use a statistics or probability model (e.g. a normal distribution or Poisson distribution) for the given data set and then identifies outliers using a discordancy test. A discordancy test verifies whether an object O_i is significantly large (or small) in relation to the standard distribution model [1]. Depth based techniques represent every data in a k-d space, and assign depth to each object. Here the data objects with smaller depths are declared as outliers [7]. Roousseeuw [8] propose robust regression to provide resistant results in the presence of outliers. ISODEPTH and Fast Depth Contours (FDC) are two depth based algorithms based on depth contours. Knorr and Raymond [9] proposed a method based on distance calculation to identify outliers. Ramaswamy et al. [10] presented new definition for outliers and propose a novel formulation for distance-based outliers that is based on the distance of a point from its kth nearest neighbour and developed a highly efficient partition-based algorithm for mining outliers. Peng and Biao [11] defined the correlation matrix considering the importance and relationship of attributes to detect outlier in large dataset. Ali et al. [12] presented an overview of the major developments in the area of detection of Outliers in numerical datasets. These include projection pursuit approaches as well as Mahalanobis distance-based procedures. They also discuss principal component-based methods, which is applicable for high dimensional data. Aleksandar and Vipin [13] presented a novel feature bagging approach for detecting outliers in high dimensional and noisy databases by combining multiple outlier detection algorithms. Breunig et al. [14] introduced a new method for finding outliers in a multidimensional dataset through density based which uses a local outlier factor (LOF) for each object in the dataset, indicating its degree of outliers. Podgorelec et al. [15] introduced the class confusion score metric based on the classification results for predicting outlier in medical dataset. Gongxian Cheng [16] presents a

novel attempt in automating the use of domain knowledge in helping distinguish between different types of outliers. Jin et al. [17] proposed cluster based algorithm that restricts LOF value computation to selected clusters that constitute the candidate outlier set. It computes the lower and upper limits for each cluster based on their local reachability densities for detecting outliers. The replicator neural network (RNN) is employed to detect outliers by Hawkins et al. [18].

56.3 Framework of the Proposed System

The system design proposed in this paper comprises of two data mining techniques and it's framework is presented in Fig. 56.1. The first design focuses on outlier mining of the KC1 dataset. The second design focuses on the classification of the dataset. This phase also includes the comparison of classifier results before and after Outlier detection and removal. Outlier detection is performed based on Sigma Rule. After the error rates of each classification algorithm are recorded, the algorithm with the least error-rate is considered as the efficient classifier. A test data is used to test and validate the accuracy of this efficient classifier.

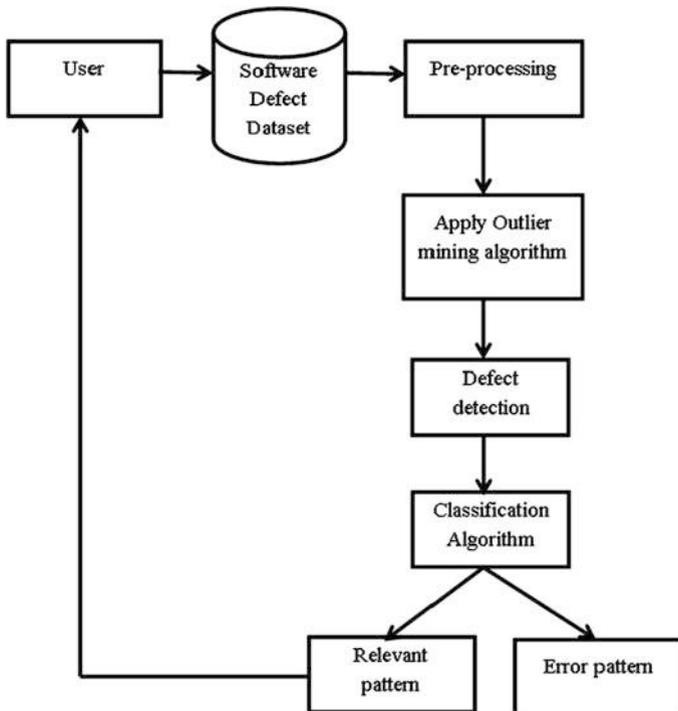


Fig. 56.1 Architecture of the proposed system

56.3.1 Training Dataset

The training dataset is downloaded from the NASA Promise software defect data sets [19] and is described by 2096 instances of KC1 data. This data comprises of 21 input attributes and one target attribute. The target class has two distinct values yes, no whether it is defective or not. These attributes are categorized as: 5 different lines of code measure, 3 McCabe metrics, 4 base Halstead measures, 8 derived Halstead measures, 1 branch-count, and 1 class field. The detailed description of the KC1 data is given in Table 56.1. After the data is pre-processed through outlier detection algorithm 633 outliers are removed. Then 16 classification algorithms are executed on the original dataset as well as on filtered dataset.

56.3.2 Outlier Detection

Outlier mining has been studied extensively by statistics community. Data objects that show significantly different characteristics from the remaining data are referred as outliers [20]. Outliers may occur as a result of mechanical faults, changes in system behavior, fraudulent behavior or through natural deviations in population. Outlier detection approaches focus on discovering patterns that are

Table 56.1 Error rate of classification algorithms on KCI dataset

S.No	Classification algorithm	Before outlier detection	After outlier detection
1	C4.5	0.0692	0.0560
2	C-PLS	0.3216	0.2871
3	C-RT	0.1551	0.0868
4	CS-CRT	0.1551	0.0868
5	CS-MC4	0.1345	0.0868
6	C-SV4	0.1498	0.0868
7	ID3	0.1446	0.0868
8	K-NN	0.1126	0.0752
9	Log-reg TRIRLS	0.1527	0.0882
10	Multilayer perceptron	0.1360	0.0868
11	Multinomial logistic regression	0.1403	0.0848
12	Naive Bayes continuous	0.1937	0.1962
13	PLS-DA	0.1455	0.0868
14	PLS-LDA	0.1441	0.0957
15	Rnd tree	0.0148	0.0157
16	SVM	0.1512	0.0868

rare or infrequent, while remaining data mining techniques attempt to find patterns that are frequent. As the removal of outliers from a dataset leads to specific model or algorithm to succeed, the outliers detection strategies are used for data cleaning before any traditional mining algorithm is applied to the data. In this proposed work, Univariate Statistical Outlier Detection technique is applied using TANAGRA tool [21] to detect and remove outliers in the KC1 dataset.

56.3.3 Univariate Outlier Detection

Most of the earliest univariate methods for outlier detection rely on the assumption of an underlying known distribution of the data, which is assumed to be identically and independently distributed. Moreover, many discordance tests for detecting univariate outliers further assume that the distribution parameters and the type of expected outliers are also known [22]. Needless to say, in real world data-mining applications these assumptions are often violated.

Given a data set of n observations of a variable x , let μ be the mean and let σ be standard deviation of the data distribution. One observation is declared as an outlier if lies outside of the interval.

$$(\mu - k\sigma, \mu + k\sigma) \quad (56.1)$$

where the value of k is usually taken as 2 or 3. The justification of these values relies on the fact that assuming normal distribution one expects to have a 95 % (99 %, respectively) of the data on the interval centered in the mean with a semi-length equal to two (three, respectively) standard deviation.

A total of 633 instances in the KC1 dataset are detected as outliers by this univariate outlier detection method. Then these outliers were removed from further analysis. Next various classification algorithms are applied on these filtered dataset to enhance the accuracy of the classifier.

56.3.4 Classification

The sixteen classification algorithms were executed and its misclassification rate was recorded for the KC1 data from NASA Promise software defect data sets. Classification is a supervised learning method which predicts a class label to a set of data values whose class label is unknown. Except Rand Tree and Naive Bayes Continuous algorithm all other classifiers have improved accuracy after the outliers are removed from the KC1 dataset. The classifier error-rates before and after outlier detection on the KC1 data are compared and the best performing classifier is based upon the accuracy in classification. Here, the Rand Tree classifier gives 98.4 % accuracy and is considered as the best classifier for this dataset.

The accuracy of this best classifier is verified by testing the rules generated with a KC1 test data. The classifier reported the class label for each test data with high level of precision.

56.4 Experimental Results

The experimental results are dealt with two sections. In the first section, the results of Univariate Outlier detection algorithm are provided. The second section gives the comparative analysis of the error rate of the classification algorithms before and after outlier detection. Finally, performance analysis of various classifiers before and after outlier detection in terms of accuracy is represented graphically.

56.4.1 Outlier Analysis

The results obtained using the data mining tool, TANAGRA on execution of the Univariate Outlier Detection Algorithm is presented in Fig. 56.2. Nearly 633 instances are removed from further analysis on the KC1 dataset using Sigma rule.

56.4.2 Classification Results of KC1 Dataset

The performance of the 16 classification algorithms was executed on the KC1 data of NASA Promise software defects dataset. The error rate is reduced for all the classifiers listed in Table 56.1. The accuracy of Rand tree classifier is 98.4 %, C4.5 is 94.0 % and the remaining classifiers except C-PLS and Naive Bayes continuous all other classifiers gives above 90 % accuracy after performing outlier detection. The accuracy of all the classifiers before as well as after outlier detection is presented in Fig. 56.3. Since almost all the classifiers shows improved performance after outlier detection, it is proved that the quality of the results gets improved which in turn gives better reliability.

56.5 Conclusion

The proposed research work mainly aims at highlighting the impact of outlier mining prior to classification of software defect dataset will leads to the overall improvement of software reliability. The comparison of classification techniques on predicting the defect will certainly assist in providing better quality of software products. As the presence of outliers degrades the performance of the software

Variable	Grubbs Stat.	Sigma rule		
		L.B	U.B	Detected
-	Cut : 4.2169			
A1	14.5426	-9.8309	13.3690	46
A2	10.7902	-18.7684	28.1205	57
A3	16.8409	-1.9820	2.2463	31
A4	13.9140	-8.3316	10.2323	41
A5	10.7760	-8.8933	14.5794	55
A6	12.5376	-7.6079	12.7071	58
A7	11.0208	-4.9422	8.2991	52
A8	10.2070	-58.0966	87.3247	51
A9	7.9846	-43.1372	85.8786	30
A10	5.9636	-16.8000	30.4243	45
A11	18.2649	-47207.4582	57757.2581	42
A12	14.7807	-0.4322	0.6045	51
A13	12.6048	-201.1636	301.4385	54
A14	5.3134	-0.6278	1.2687	21
A15	18.2649	-2622.6358	3208.7365	42
A16	14.7987	-1292.2420	1812.8423	51
A17	12.7286	-77.5181	115.3225	53
A18	12.4668	-124.4021	186.8725	53
A19	9.0414	-27.0377	46.2286	43
A20	5.1283	-9.4751	24.8310	27
A21	8.9710	-69.1283	109.8688	53

Fig. 56.2 Results of outlier detection algorithm on KC1 data using sigma rule

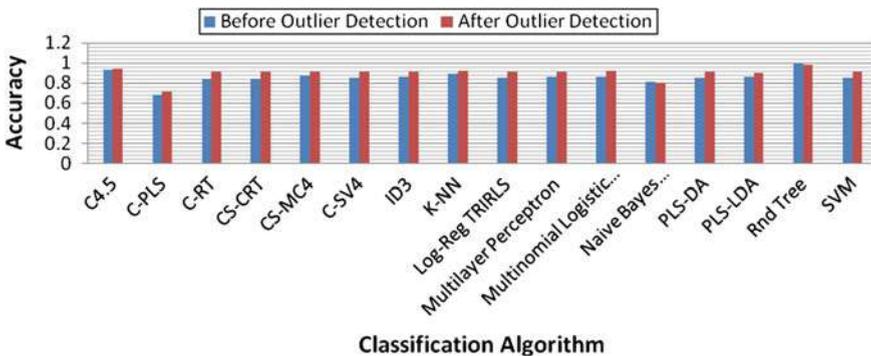


Fig. 56.3 Accuracy of classification algorithm

products, it needs special attention. In this work, the impact of a Univariate outlier detection method is analyzed on KC1 dataset to show its dependency on improving the software reliability.

References

1. Han, J., Kamber, M.: *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Jose (2001)
2. Go, K., seva-Popstojanova, Trivedi, K.S.: Architecture-based approach to reliability assessment of software systems. *Perform. Eval.* **45**(2/3), 179204 (2001)
3. Xie, M., Hong, G.Y., Wohlin, C.: Software reliability prediction incorporating information from similar projects. *J. Softw. Syst.* **49**(1), 43–48 (1999)
4. Huang, C.Y., Lin, C.T.: Software reliability analysis by considering fault dependency and debugging time lag. *IEEE Trans. Reliab.* **55** (3), 436450 (2006)
5. Roberto, R.S., Trivedi, K. S.: Software reliability and testing time allocation: an architecture based approach. *IEEE Trans. Reliab.* **36**(3), 322–337 (2010)
6. Podgorelec, V.: Improved mining of software complexity data on evolutionary filtered training sets. *WSEAS Trans. Inf. Sci. Appli.* **6**(11), 1751–1760 (2009)
7. Ruts, I., Rousseeuw, P.: Computing depth contours of bivariate points cloud. *Comput. Stat. Data Anal.* **23**, 153–160 (1996)
8. Rousseeuw, P., Leroy, A.: *Robust Regression and Outlier Detection*. Wiley, New York (2003)
9. Knorr, E.M., Ng, R.T.: A unified notion of outliers: properties and computation. In: *Proceedings of KDD 97*, pp. 219–222 (1997)
10. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithm for mining outliers from large datasets. In: *Proceedings of ACM SIGMOD*, pp. 127–138 (2000)
11. Peng, Y., Biao, H.: An efficient outlier mining algorithm for large dataset. In: *International Conference on Information Management, Innovation Management and Industrial Engineering*, IEEE Computer Society, pp. 199–202 (2008)
12. Ali, H.S., Rahmatullah Imon, A.H.M., Mark, W.: Detection of outliers overview. *Wiley Interdisc. Rev. Comput. Stat.* **1**(1), 57–70 (2009)
13. Aleksandar, L., Vipin, K.: *Feature Bagging for Outlier Detection*, KDD05, Chicago, Illinois, USA, August 2124 (2005)
14. Breunig, V., Kriegel, M.M., Ng, R.T., Sander, J.: LOF: identifying density based local outliers. *Proc. of ACM SIGMOD* (2000) (Dallas, TX)
15. Podgorelec, V., Hericko, M., Rozman, I.: Improving mining of medical data by outliers prediction. In: *Proceedings of the 18th IEEE Symposium on Computer-based Medical Systems CBMS'2005* (2005)
16. Gongxian Cheng, J.: *Outlier Management in Intelligent Data Analysis*. Department of Computer Science, Birkbeck College, University of London, London (2000)
17. Jin Tung, W.A.K.H., Han, J.: Mining top-n local outliers in large databases. In: *Proceedings of KDD*, San Francisco, CA, USA, (2001)
18. Hawkins, S., He Willams, H.G.J., Baster, R.A.: Outlier detection using replicator neural networks. In: *Proceedings of the DaWaK02*, pp. 170–180 (2002)
19. Promise <http://promise.site.uottawa.ca/SERepository/datasets-page.htm>
20. Hawkins, D.: *Identification of Outliers*. Chapman and Hall, London (1980)
21. Tanagra-Data Mining tutorials <http://data-mining-tutorials.blogspot.com>
22. Barnett, V., Lewis, T.: *Outliers in Statistical Data*. Wiley, New York (1998)

Author Biographies

Dr. S. Poonkuzhali received B.E degree in Computer Science and Engineering from University of Madras, Chennai, India, in 1998, and the M.E degree in Computer Science and Engineering from Sathyabama University, Chennai, India, in 2005, and PhD in the Faculty of Information and Communication Engineering at Anna University Chennai, in 2012. Currently she is working as a Professor and Head in the Department of Information Technology in Rajalakshmi Engineering College, Chennai. Her research interests include Data Mining, Neural Networks, Web Mining and Knowledge Management. She has organized and chaired many Workshops, Seminars and Conferences in national and International level. She has presented and published more than 25 research papers in international conferences & journals and authored 5 books. She is a life member of ISTE (Indian Society for Technical Education), IAENG (International Association of Engineers), ISCSIT and CSI (Computer Society of India).

R. Kishore Kumar received B.E degree in Computer Science and Engineering from Rajalakshmi Engineering College, Anna University, Chennai, India in 2011 and M.E degree in Computer Science and Engineering in SSN College of Engineering, Anna University Chennai, India in 2013. Currently he is working as Assistant Professor in Department of Information Technology in Rajalakshmi Engineering College, Chennai. He has presented 11 papers in International conferences and published 6 research papers in international journals and 3 papers in national journals. One of his papers has been selected as the Best Paper. He also received the Certificate of Merit Award for the paper presented in IAENG Conference. He is also the member of Computer Society of India and IAENG.

R. Kumar is an Associate Professor in Rajalakshmi Engineering College, Chennai. He holds M. Tech degree in Computer Science and Engineering and is pursuing his PhD in The Anna University, Chennai as a part time scholar. His research interests include Knowledge Engineering and Management, Data Mining and Software Engineering. He is a member of Computer Society of India (CSI).

Chapter 57

A Survey on the Application of Robotic Teacher in Malaysia

Noraidah Blar and Fairul Azni Jafar

Abstract Robot application in Malaysia is just a couple of number these days contrast with different nations particularly in education. Robotic teacher application is likewise not extremely commonplace around Malaysian students. The survey is intended to distinguish the sentiment around Malaysian technical institutes about utilization of robotic teacher in their organization. An overview was developed and dispersed by utilizing web interface that is Google Form application. The result demonstrates that the greater part of students who completed the survey do not consent to utilize a robot as a teacher. Numerous Malaysian individuals finished not think about the genuine proficiency of a robotic teacher. A further study about this theme will be led after this investigation.

57.1 Introduction

As a human, there are likewise various types of teaching method. There are always questions highlighted that it is, is adequate or not to make students completely comprehend what they attempt to educate. What's more yes, a few students are neglected to comprehend what they take in because of numerous factors, one of them is teaching method conveyed by their teacher. Living in a world that undeniably propelled, constrained human to find some other elective to make the instruction more viable. One of the finding is utilizing robots as a part of teaching session.

N. Blar (✉)

Center for Graduate Student, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: noraidahblar@student.utem.edu.my

F.A. Jafar

Faculty of Manufacturing Engineering, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: fairul@utem.edu.my

According to Berk [1], there are twelve potential sources of evidence to measure teaching effectiveness that are critically reviewed; student ratings, peer ratings, self-evaluation, videos, student interviews, alumni ratings, employer ratings, administrator ratings, teaching scholarship, teaching awards, learning outcome measures, and teaching portfolios. From [2], it is stated that the strategies to convey learning are; practical examples, show and tell, case studies, guided design projects, open-ended labs, the flowchart technique, open ended quizzes, brainstorming, question-and-answer method, software, teaching improvement, and fast feedback form.

In order to understand about robot teacher, basic information about robot must be learned first. A robot is a mechanical device that can perform preprogrammed physical tasks. It may act under the direct control of a human or automatically under the control of a pre-programmed computer. Robot can be classified by several types. There is industrial, mobile, service, military, humanoid and other type of robot.

In a research conducted by Chang et al. [3], the research is aim to find the possibilities of using humanoid robots as instructional tools for teaching second language in primary school. They found that the absolute majority of the students actively participated in learning activities throughout the lesson and interacted with the robot-teachers with great interest.

Numerous researches about robot teacher have been carried out in other nation particularly at the advancing countries, for example, Japan and United State. It is accepted that the level of competencies and thinking in instruction for Malaysian's students is much distinctive contrast with different countries. Consequently, a considerable measure of exploration is required to know the suitability of utilizing robots within teaching Malaysian individuals. The focus of this paper is to know the feedback of Malaysian technical students about the application of robotic teacher.

57.2 Survey

Main subject of this survey is the feedback from the respondents. Respondents are consisting of Malaysian technical students from technical universities (Universiti Teknikal Malaysia Melaka, Universiti Malaysia Pahang, Universiti Tun Hussein Onn Malaysia, and Universiti Malaysia Perlis). The survey's questions was created and it then distributed by using Google Form. The link of this form was shared at social webpages of the technical universities that listed. The survey's question is constructed as follow;

1. Is there any robot used in your university?
2. If yes, what they are used for?
3. What is your level of exposed with robot?
4. Have you ever experience a teaching delivered by a ROBOT?
5. Just imagine, what will you feel when a robot is teaching your subject in class?

A Robot As A Teacher



Aim of this form is to observe responses from technical students about usage of a robot as a teacher. The result of this form will be used as a data collection for my master degree thesis. Thank you for your time and response. :)

* Required

Gender *

Male

Female

Occupation *

Student

Lecturer

Other:

Your current university *

Universiti Teknikal Malaysia Melaka (UTeM)

Universiti Tun Hussein Onn Malaysia (UTHM)

Universiti Malaysia Pahang (UMP)

Universiti Malaysia Perlis (UniMaP)

Other:

1. Is there any robot used in your university? *

Yes

No

2. If yes, what they are used for? *

You can tick more than one

Teaching tool

Laboratory equipment

Student's project

Display

Contest

Other:

3. What is your level of exposed with robot? *

1 2 3 4 5 6 7 8 9 10

None Most frequent



4. Have you ever experience a teaching delivered by a ROBOT

Yes

No

5. Just imagine, what will you feel when a robot is teaching your subject in class? *

You can tick more than one

Anxious/Resah

Apathy/Tidak Peduli

Bored/Bosan

Cautious/Waspada

Confident/Yakin

Distracted/Teganggu

Excited/Tenja

Fatigue/Kalu

Fear/Takut

Hesitant/Ragu

Impressed/Kagum

Inspired/Inspirasi

Panic/Panik

Peaceful/Amam

Pressured/Tertekan

Relaxed/Tenang

Satisfied/Puas Hati

Other:

6. In your experience, is teaching method delivered by a human teacher is effective enough? *

Yes

No

7. In your opinion, what will affect the effectiveness of delivering lesson? *

Interaction

Teaching method

Teaching skills

Communication

Facilities

Other:

8. In your prediction, will robot teach students more effective than human teacher did? *

Yes

No

Please give your opinion

Fig. 57.1 Survey's question using google form (minimized)

6. In your experience, is teaching method delivered by a human teacher is effective enough?
7. In your opinion, what will affect the effectiveness of delivering lesson?
8. In your prediction, will robot teach students more effective than human teacher did?
9. Opinion.

The purpose of the first and the second question is to know the alertness of respondent on the existing robot at their place. The third and the fourth question are constructed to know the exposure condition among the respondents. While, the other questions are asked in order to know about respondent's feedback on the application of the robot teacher (Fig. 57.1).

57.3 Results

There are a total of 67 respondents. 45 out of them are male and 22 persons are female. Total of respondents according to each universities are 30 persons (45 %) from Universiti Teknikal Malaysia Melaka, 12 people (18 %) from Universiti Tun Hussein Onn Malaysia, 15 people (23 %) from Universiti Malaysia Pahang, 8 people (12 %) from Universiti Malaysia Perlis, and a person (2 %) from other institution. According to the survey, there are 59 students, 3 lecturers, and 4 others. Below are the graphs of results of each question.

57.4 Discussion

Figure 57.2 shows the result for the first question. The question asks about the existing robot in their university. It is shown that 54 respondents state that there is robot(s) in used in their university. While the other 12 respondents stated that robot is not in used in their institute. It is shown that some of them are not alert enough about the robot. This situation may due to they did not use robot in their study. The second question (Fig. 57.3) asked about the function of robot in their university if

Fig. 57.2 Result of question 1

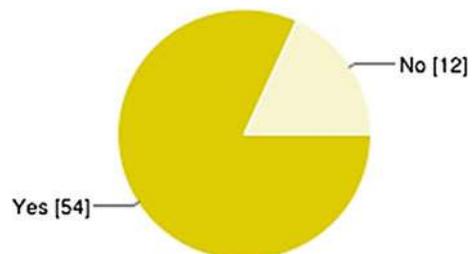


Fig. 57.3 Result of question 2

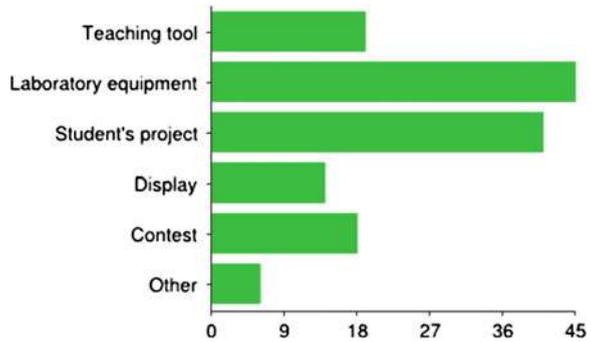
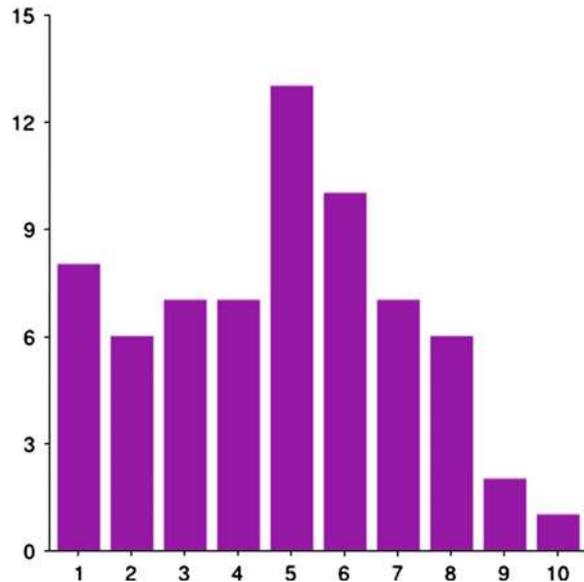


Fig. 57.4 Result of question 3



there is any robot in used. From the list of function suggested, the highest number is 45 which is laboratory equipment. The second highest usage of robot is for student’s project with 41 respondents, followed by teaching tool, contest, display, and other with values of 18, 18, 14, and 6 subsequently. The next graph (Fig. 57.4) shows the result for the third question. The question is asking about the level of exposure of respondents with robot. In the question, there are 10 level of exposure. Which is Level 1 is the fewest, while Level 10 is the most frequent. From the result, it is shown that most of the respondents are level 5 for the question. The highest value is 13 respondents with 19 % of the graph, followed by the second highest with value of 10 respondents (15 %) which is level 6. The third highest value of respondent is level 1 which has the number of respondent of 8 (12 %). Level 3, 4, and 7 have the same number of respondent that is 7 (10 %).

Meanwhile, level 2 and 8 share the same value of 6 respondents (9 %). Level 9 and 10 has the smallest value with 2 (3 %) and 1 (1 %) respectively. It is shows that most of the respondents have average level of experience.

The next question (Fig. 57.5) is about the experience of respondents in teaching by robot. Most of them say no with the number of 62 respondents, while the other 5 respondents have experience learning from robot. Respondents do not have experience yet in having a lesson delivered by a robot. The graph in Fig. 57.6 shows the statistical result of feelings by respondents. There are 17 feelings that listed in the question. Respondents can choose more than one feeling to answer the question. Most of the respondents feel excited when they imagine getting lesson from a robot. There are 56 people that feel that way with the highest percentage of 16 %. The second highest is impressed feeling with 50 people, followed by inspired expression with 35 people. The lowest number of results in the graph is

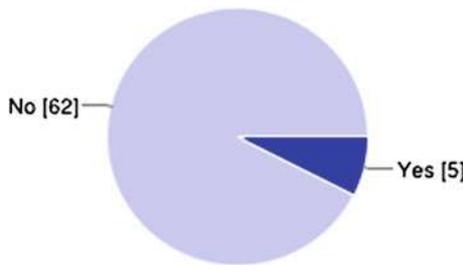


Fig. 57.5 Result of question 4

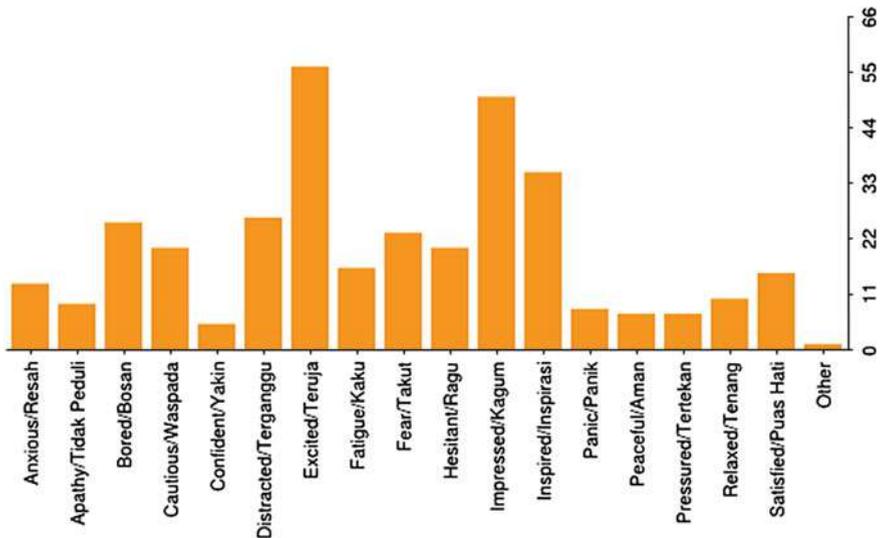


Fig. 57.6 Result of question 5

Fig. 57.7 Result of question 6

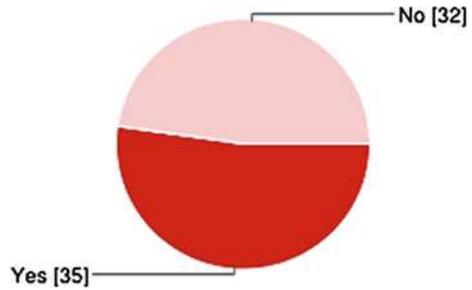
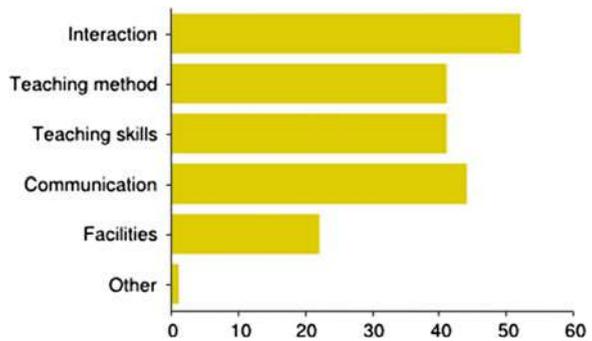


Fig. 57.8 Result of question 7



confident with 5 respondents. Most of the respondents responded to be excited as because they never imagine robot cannot do such things.

The graph in Fig. 57.7 shows the opinion from the respondents on the effectiveness of teaching by human teacher. It is shown that 35 people are agreed that human teacher already teach in an effective way, while the other 32 feel that it is not good enough. In Fig. 57.8, there are 52 people choose on interaction factor that will affect the effectiveness on delivering lesson. This is the highest number of factor selected among the other factor. Meanwhile, communication factor become the second highest of option from the respondents with value of 44. This is followed by teaching method and teaching skills that shared the same value of number of choice from the respondents. There are least people that choose facilities as the factor. Interaction factor become the highest choice from the respondents because the might be known that, to achieve a full understanding from each other, interaction is important to each other.

The graph of Fig. 57.9 shows the result on the prediction of respondents about comparison of effectiveness of teaching by both human and robot. Majority of them, with 45 number of people, disagree with the statement that robot will teach more effective than human teacher did. The rest of the respondent, with 22 numbers of respondent, have a positive feedback that believe robot will teach better than the human teacher did.

Fig. 57.9 Result of question 8

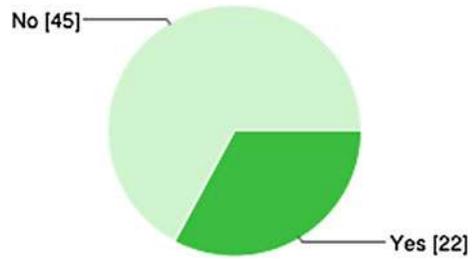


Table 57.1 List and summary of respondent’s opinion

Positive opinion	Negative opinion
Both robot and human can be a good teacher	Robot did not have feeling like human
Robot is fun	A teacher is a human that know how their student perform. They can approach that kind of student
Robot can attract people to learn better	Teacher will lose their job
Useful for future study	Beware of 3 Asimov law of robot
High technology robot can replace human in the future	Robot is dangerous to human being
Everything have their own weakness, both human and robot	Robot is heartless, human is close with human heart

Below is the list and summary of the last survey’s question, that asking about the respondent’s opinion about robotic teacher. Overall of the result shows that mostly human still did not have believed in robot in delivering lesson (Table 57.1).

57.5 Conclusion

Through the result that we get by this survey, we can see that the implementation of robot teacher have many challenges. This is because there are many respondents that still not believe about the effectiveness of robot teacher (Fig. 57.8). Although when we ask about their feeling when robot give them a teaching lesson, most of them feel excited and impressed (Fig. 57.5).

So, we have decided to further this study in order to identify the effectiveness of robot teacher in delivering technical lesson to human. The study will involve a robot who will teach the respondents and then the respondents will have to fill up a survey form.

References

1. Berk, R.A.: Survey of 12 strategies to measure teaching effectiveness. *Int. J. Teach. Learn. High. Educ.* **17**, 48–62 (2005) (John Hopkins University, USA)
2. Couter, S., Balaraman, P., Lacey, J., Hochgraf, C.: *Strategies for Effective Teaching: a Handbook for Teaching Assistants*. Teaching Assistant Fellows, University of Wisconsin, Madison, College of Engineering, Madison (1995)
3. Chang, C.W., Lee, J.H., Chao, P.Y., Wang, C.Y., Chen, G.D.: Possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Educ. Technol. Soc.* **13**(2), 13–24 (2010)

Chapter 58

A Novel Method for Distributed Image Steganography

Bismita Choudhury, Rig Das and Themrichon Tuithung

Abstract Distributed Image Steganography (DIS) is a new method of concealing secret information in several host images, leaving smaller traces than conventional steganographic techniques and requiring a collection of affected images for secret information retrieval. DIS, compared to other conventional steganographic techniques, can improve security and information hiding capacity because DIS leaves reduced signatures of hidden information in host images. This paper presents a novel technique for Distributed Image Steganography based on Shamir's (k, n) threshold based secret sharing scheme. An 8-bit, gray level, Secret Image of size $P \times Q$ is divided into several sections and embedded inside n -Cover Images of size $M \times N$ and only k -Stego Images ($k \leq n$) are required to retrieve the secret image. The size of the secret image is also embedded inside every cover images, so that the stego images become standalone information to the receiver. Experimental result shows that the proposed novel method has high capacity and good invisibility.

Keywords Steganography · Distributed image steganography · PSNR · Shamir's secret sharing scheme

B. Choudhury (✉)

Department of Computer Science and Engineering and Information Technology, Don Bosco College of Engineering and Technology, Guwahati 781017, Assam, India
e-mail: bismi.choudhury@gmail.com

R. Das

Department of Computer Science and Engineering, National Institute of Technology, Rourkela 769008, Orissa, India
e-mail: rig.das@gmail.com

T. Tuithung

Department of Computer Science and Engineering, North Eastern Regional Institute of Science and Technology (Deemed University), Itanagar 791109, Arunachal Pradesh, India
e-mail: t_tuithung@yahoo.com

58.1 Introduction

Image Steganography is the art of hiding secret data or images into a Cover Image [1]. In recent years, many techniques to increase the security of the secret image were proposed. However a common weakness of these techniques is that the secret data are all in a single information-carrier and the secret data cannot be revealed completely if the information-carrier is lost or crippled [2, 3]. If we use many duplicates to overcome the weakness, the danger of security exposure will also increase [2]. To solve this dilemma, Distributed Image Steganography (DIS) might be one of the possible solutions. DIS is a new method of concealing secret information in several host images, leaving smaller traces than conventional steganographic techniques and requiring a collection of affected images for secret information retrieval [4]. DIS is a secret image sharing method derived from the Shamir's Secret Sharing Scheme [5]. Shamir developed the idea of a (k, n) threshold-based secret sharing technique where $k \leq n$ and for this we will require:

1. A secret image that will be used to generate n stego images.
2. Any k or more stego images can be used to reconstruct the secret image.
3. Any $k - 1$ or less stego images cannot get sufficient information to reveal the secret image [6].

By using a (k, n) threshold scheme with $n = 2k - 1$ we get a very robust key management scheme: We can recover the original key even when $\lfloor n/2 \rfloor = k - 1$ of the n pieces are destroyed, but our opponents cannot reconstruct the key even when security breaches expose $\lfloor n/2 \rfloor = k - 1$ of the remaining k pieces [5]. The main objective of Distributed Image Steganography is to communicate securely in such a way that the true message is not visible to the observer i.e. any unwanted parties should not be able to distinguish between cover-images (images not containing any secret message) and stego-images (modified cover-images that containing secret message). Thus the stego-images should not deviate much from original cover-images. Figure 58.1 shows the block diagram of a simple DIS system with $(3, 5)$ threshold scheme that means the secret image will be distributed within 5 stego images and any 3 stego images are enough to regenerate the secret image.

This paper organized as follows. Section 58.2 reviews some of the related works performed by the researchers. Proposed novel method is presented in Sect. 58.3 along with the block diagram and the algorithm. Experimental results are shown in Sect. 58.4 and finally the conclusion is being stated in Sect. 58.5.

58.2 Related Work

Different researchers employed different techniques for the purpose of distributing secret image over a set of stego images. Shamir's Secret Sharing Scheme and Thien and Lin's Secret Image Sharing Scheme are the two essential processes to protect secret image in DIS. Following are the related works carried out by these groups:

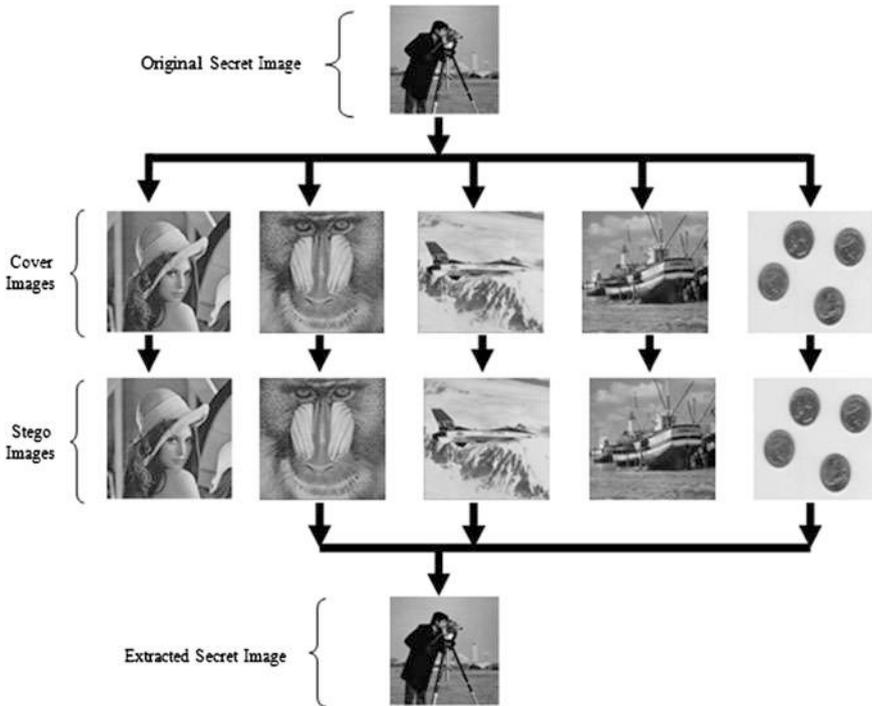


Fig. 58.1 The block diagram of a simple distributed steganographic system with (3, 5)

58.2.1 Shamir's Secret Sharing Scheme [5]

Shamir's scheme is based on polynomial interpolation: given k points in the 2-dimensional plane $(x_1, y_1), \dots, (x_k, y_k)$ with distinct x_i 's, there is one and only one polynomial $q(x)$ of degree $k - 1$ such that $q(x_i) = y_i$ for all i . It is assumed that the data D is (or can be made) a number. To divide it into pieces D_i , pick a random $k - 1$ degree polynomial $q(x) = a_0 + a_1x + \dots + a_{k-1}x^{k-1}$ in which $a_0 = D$, and evaluate: $D_1 = q(1), \dots, D_i = q(i), \dots, D_n = q(n)$. Given any subset of k of these D_i values (together with their identifying indices), the coefficients of $q(x)$ can be found by interpolation, and then evaluating $D = q(0)$. Knowledge of just $k - 1$ of these values, on the other hand, does not suffice in order to calculate D .

The set of integers modulo a prime number p forms a field in which interpolation is possible. Given an integer valued data D , we pick a prime p which is bigger than both D and n . The coefficients a_1, \dots, a_{k-1} in $q(x)$ are randomly chosen from a uniform distribution over the integers in $[0, p)$, and the values D_1, \dots, D_n are computed modulo p . Now assuming that $k - 1$ of these n pieces are revealed to an opponent. For each candidate value D' in $[0, p)$ he can construct one and only one polynomial $q'(x)$ of degree $k - 1$ such that $q'(0) = D'$ and $q'(i) = D_i$ for the $k - 1$

given arguments. By construction, these p possible polynomials are equally likely, and thus there is absolutely nothing the opponent can deduce about the real value of D .

58.2.2 Thien and Lin's Secret Image Sharing Scheme [6]

Suppose we want to divide the secret image D into n shadow images (D_1, \dots, D_n) , and the secret image D cannot be revealed without r or more shadow images. In the proposed method, they generate the $r - 1$ degree polynomial, by letting the r coefficients be the gray values of r pixels. Therefore, the major difference between Thien and Lin's method and Shamir's is that they use no random coefficient. Because the gray value of a pixel is between 0 and 255, they took the prime number p be 251 which is the greatest prime number not larger than 255. To apply the method, it must truncate all the gray values 251–255 of the secret image to 250 so that all gray values are in the range 0–250. The image is divided into several sections. Each section has r pixels, and each pixel of the image belongs to one and only one section. For each section j , define the following $r - 1$ degree polynomial as $q_j(x) = (a_0 + a_1x + \dots + a_{r-1}x^{r-1}) \bmod 251$, Where a_0, \dots, a_{r-1} are the r pixels of the section, and then evaluate $q_j(1), q_j(2), \dots, q_j(n)$. The n output pixels $q_j(1) - q_j(n)$ of this section j are sequentially assigned to the n shadow images. Since for each given section (of r pixels) of the secret image, each shadow image receives one of the generated pixels; the size of each shadow image is $1 / r$ of the secret image. The reveal phase uses any r (of the n) shadow images, and the Lagrange's interpolation to extract the secret image.

58.3 Proposed Novel Method for Distributed Image Steganography

In this paper we have proposed a spatial distributed steganographic technique based on Shamir's secret sharing scheme for hiding a large amount of data with high security, a good invisibility and no loss of secret message. The schematic/block diagram of the whole process is given in Figs. 58.2 and 58.3.

Our proposed algorithm for DIS is based on (k, n) Threshold Scheme and has two parts, one for Embedding the Secret Image inside n -Cover Images and another for Extracting the Secret Image from k -Stego Image. The Secret Image is divided into several sections. Each section has k pixels, and each pixel of the secret image belongs to one and only section. For each section j , we define the following $k - 1$ degree polynomial:

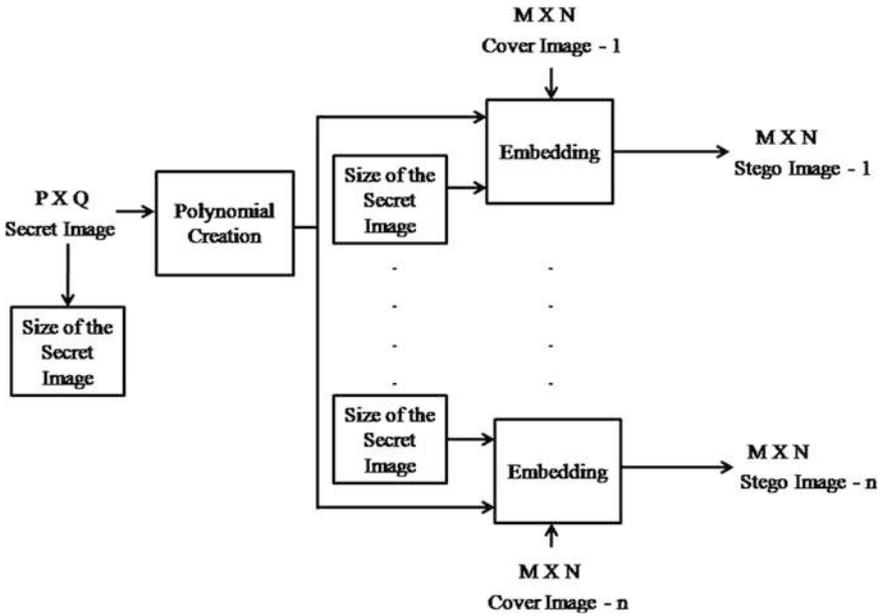


Fig. 58.2 Insertion of a secret image inside n-Cover images

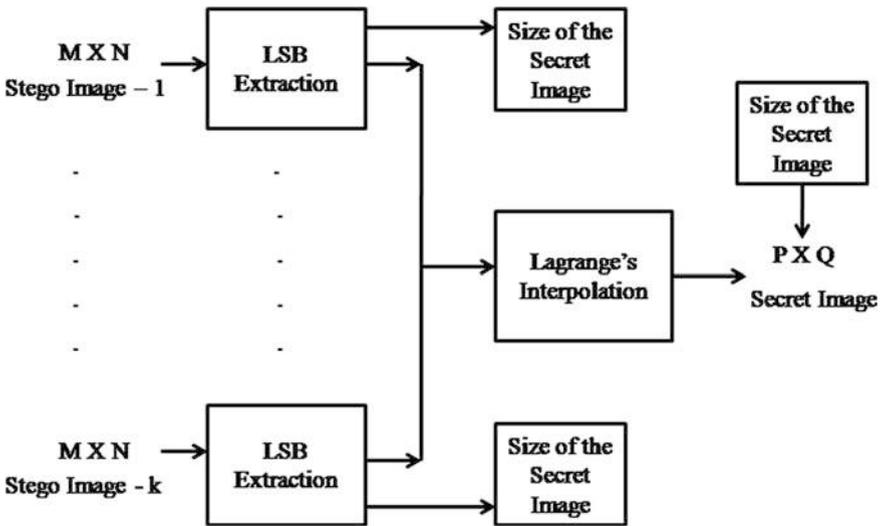


Fig. 58.3 Retrieval of secret image from k-Stego images

$$p_j(x) = (a_0 + a_1x + \dots + a_{k-1}x^{k-1}) \bmod 256 \quad (58.1)$$

$$q_j(x) = \text{floor} [(a_0 + a_1x + \dots + a_{k-1}x^{k-1}) / 256] \quad (58.2)$$

Where, value of x ranges from 1 to n and a_0 to a_{k-1} are Secret Image's k number of pixel's intensity values. These intensity values changes sequentially e.g. first we take 1–8 pixels' intensity values then 9–16 pixels' intensity values and so on. Modulus and Floor operations are performed using 256 as divisor because a grey level image has 256 different intensity levels. Equation (58.1) finds the Remainder Value after performing Modulus operation and (58.2) finds the Quotient value after performing Floor operation.

The main objective in here is to develop a procedure which will provide a better security to the secret image without compromising on the quality of the stego image. Our algorithm has two main parts. First, it embeds the size of the secret image in first 8×8 block of every cover image. Second, it embeds each set of $p_j(x)$ and $q_j(x)$ that means the remainder and the quotient value sequentially to every cover image.

58.3.1 Four Tier Storage Procedure of “Size of Secret Image” [7]

The size of the Secret Image multiplied by 8 (for 8 bit images) should be less than the size of the Cover Image. Because every pixel of secret image has 8 bit (for 8 bit images) and it requires 8 pixels of the cover image to get embedded. E.g. if the size of the Secret Image is $256 \times 256 = 65536$ then after multiplying it by 8 it becomes 523288. This is lesser than the size of Cover Image i.e. $1024 \times 1024 = 1048576$. So this secret image can be embedded. But if the size of the secret image is $512 \times 512 = 262144$ then after multiplying it by 8 it becomes 2097152 which is more than the number of pixels in the cover image of size 1024×1024 . So a secret image of size 512×512 can't be embedded inside the cover image of size 1024×1024 .

The size of the Secret Image needs to be embedded inside the cover image to let the decoder know which pixel's LSB holds the Secret Image's information. Now the question is how will we get the size of the “Size of Secret Image”? If we store this size of the “Size of Secret Image” then also the same question comes recurrently like what is the size of this size? As a permanent answer to this question we have used a four tier storage procedure to store the “Size of Secret Image” which had been described in [7].

58.3.2 Proposed Novel Algorithm for Distributed Image Steganography

Embedding Algorithm

Input: n number of $M \times N$ Carrier Images and a $P \times Q$ Secret message/Image.

Output: n number of $M \times N$ Stego-Images.

- Step-1: Read all the n -Cover Images and the Secret Image and store their intensity values of different pixels in different arrays.
- Step-2: Calculate size of Secret Image. The size of the Secret Image multiplied by 8 (for 8 bit images) should be less than the size of the Cover Image. E.g. if the size of the Secret Image is $256 \times 256 = 65536$ then after multiplying it by 8 it becomes 523288. This is lesser than the size of Cover Image i.e. $1024 \times 1024 = 1048576$.
- Step-3: Store the size found in Step-2 using the four tier storage procedure described above by modifying the LSB of first 8×8 block of pixels of all the cover images.
- Step-4: Sequentially take k -number of not-shared-yet pixels of the Secret Image and use (58.1) and (58.2) to find the n number of Remainder and n number of Quotient value.
- Step-5: Change the LSBs of n Cover Images to insert each set of Remainder and Quotient values found in Step-4 in each of the Cover Images excluding the first 8×8 block.
- Step-6: Repeat Step-4 and Step-5 until all the pixels of the Secret Image are processed.
- Step-7: Write all the n Stego Images into the disk.

Extraction Algorithm

Input: k number of $M \times N$ Stego-Images.

Output: A $P \times Q$ Secret Image.

- Step-1: Read k number of Stego Images
- Step-2: Read the Stego Image and extract the size of the Secret Image from the first 8×8 block of any Cover Image by extracting the LSB of the pixels using the procedure described in the four tier storage method like variable D to C to B to A.
- Step-3: Extract k number of Remainders and k number of Quotients from the k Stego Images by extracting the LSBs of the pixels.
- Step-4: Use Lagrange's Interpolation to retrieve k number of pixel's intensity.
- Step-5: Repeat Step-3 and Step-4 until the total number of pixels of the Secret Image are processed.
- Step-6: Regenerate the original Secret Image from the pixel intensities found in Step-4 and Step-5.

58.4 Simulation Results

In this section, some experiments are carried out on our proposed algorithm for (DIS). The measurement of the quality between the cover image f and stego-image g of sizes $M \times N$ is done using PSNR (*Peak Signal to Noise Ratio*) value and the PSNR [8, 9] is defined as:

$$PSNR = 10 \times \log(255^2/MSE) \quad (58.3)$$

$$\text{Where, } MSE = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (f(x,y) - g(x,y))^2 / N^2 \quad (58.4)$$

$f(x, y)$ and $g(x, y)$ means the pixel intensity value at position (x, y) in the cover-image and the corresponding stego-image respectively. The PSNR is expressed in dB. The larger PSNR indicates the higher the image quality i.e. there is only little difference between the cover-image and the stego-image. On the other hand, a smaller PSNR means there is huge distortion between the cover-image and the stego image.

All the simulation has been done using the MATLAB 7 program on Windows XP platform. A set of 8-bit grayscale TIFF images of size 1024×1024 and 256×256 are used as the cover-image and secret image respectively to form the stego-image. Figure 58.4a–e shows the original cover (carrier) images and Fig. 58.4f shows the original secret message. Figure 58.5a–e shows the stego images of proposed DIS method based on (4, 5) threshold scheme and Fig. 58.5f shows the Secret Image retrieved from the any four Stego Images. Figure 58.6a–e shows another sets of the original cover (carrier) images and Fig. 58.6f shows the original secret message. We have tested for both the sets of cover images and

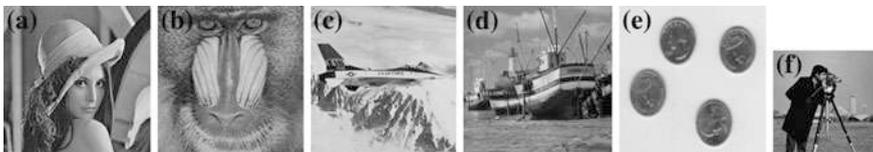


Fig. 58.4 a–e Five cover images for simulations, f original secret image

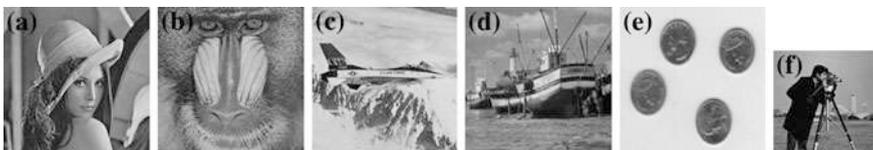


Fig. 58.5 a–e Five stego images of proposed DIS method, f extracted secret image

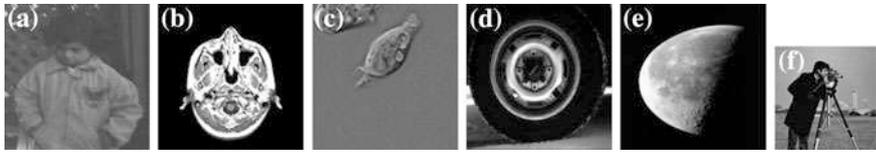


Fig. 58.6 a–e Another five sets of cover images for simulations, f original secret image

Table 58.1 PSNR comparison of cover images and stego images for different threshold scheme

Cover image	PSNR (dB) between cover image and stego image				
	<i>Threshold scheme</i>				
	(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)
Lenna	+56.88	+55.49	+57.41	+58.69	+55.85
Baboon	+56.88	+55.65	+57.55	+59.05	+56.06
Airplane	+56.89	+55.67	+57.73	+59.41	+56.23
Boat	+56.89	+55.77	+58.02	+59.76	+56.42
Coin	+56.42	+55.40	+57.52	+59.43	+55.88

Table 58.1 exhibit the PSNR comparison of first set of Ste-go Images with their corresponding Cover Images for different threshold scheme.

From Table 58.1 it is observed that for threshold schemes (1, 5) to (5, 5) PSNR is greater than 55 dB, so the quality of the stego image is very high and the deterioration in quality due to embedding of secret image cannot be distinguished by naked eye. Best result is achieved in case of (4, 5) threshold scheme, as the PSNR is greater than +58 dB for all the stego images.

In case of threshold scheme (1, 5), Secret Image is extracted from any one of the five Stego Images, which is identical with the original Secret Image. Although (1, 5) threshold scheme is nothing but alike normal LSB based steganography, as the Secret Image is not distributed over the five Cover Images and any of the 5(five) stego images can be used to extract the secret image. Similarly for threshold scheme (2, 5), (3, 5), (4, 5) and (5, 5), the Secret Image is extracted from any chosen 2(two), 3(three), 4(four) or 5(five) (respectively) Stego Images out of five stego images. Extracted Secret Image is identical to the original Secret Image as the PSNR between original secret image and extracted secret image is infinite for all the threshold schemes from (1, 5) to (5, 5). So 100 % recovery of the secret image is being achieved for all the cases.

Shamir’s (k, n) threshold scheme says $k \leq n$ so (2, 5), (3, 5), (4, 5) should be the ideal threshold based secret sharing scheme. But to justify proposed novel method, simulation is also done for (1, 5) and (5, 5) threshold scheme and for all possible threshold scheme proposed novel method extracts exactly identical secret Image. According to the simulation result our proposed method gives best result when threshold scheme (4, 5) is used, as the PSNR values are mostly over +59 dB.

Table 58.2 PSNR comparison of second set of cover images and stego images for (4, 5) threshold scheme

Cover image	PSNR (dB) between cover image and stego image
	<i>Threshold scheme (4, 5)</i>
Pout	+58.68
MRI	+74.07
Cell	+59.41
Tire	+59.65
Moon	+59.85

Table 58.3 PSNR comparison between steganography based on Huffman encoding [7] and proposed DIS algorithm

Cover image	PSNR (dB) between cover image and stego image	
	Steganography algorithm	
	Based on Huffman encoding	DIS with (4, 5) threshold scheme
Lenna	+57.43	+58.69
Baboon	+57.46	+59.05
Airplane	+57.46	+59.41
Boat	+57.46	+59.76

Table 58.2 provides the PSNR value comparison for second set of cover images (shown in Fig. 58.6) with their corresponding stego images for the threshold scheme (4, 5) only. As it is seen from Table 58.1, best results are achieved in case of threshold scheme (4, 5) we have only chosen (4, 5) threshold scheme for Table 58.2. The PSNR values are mostly over +59 dB. For these set of images also the PSNR value is infinite between the extracted secret image and original secret image. That means Identical Secret Image is being recovered. Same experiments also being performed over some more sets of cover images and for each one of those sets identical secret image is being recovered. So it can be claimed that 100 % recovery is possible by our proposed method.

Table 58.3 shows the PSNR and Stego Image Quality of proposed DIS algorithm is better than the one proposed in [7].

58.5 Conclusion

Our proposed novel Distributed Image Steganographic method based on (k, n) Threshold Scheme improves the security and the quality of the stego images. According to the simulation results the Stego Images of our proposed algorithm are almost identical to the Cover Images and it is very difficult to differentiate between them. We have achieved 100 % recovery of the Secret Image for any set

of chosen Stego Images. That means original and extracted Secret Images are identical. Distributing the Secret Image among n -number of Cover Images keeps the Secret Image away from stealing; destroying by any unintended users hence the proposed method may be more robust against brute force attack. Apart from images the proposed DIS method can also be used for textual information hiding, audio information hiding etc. We think that the proposed method can become an excellent intelligence embedding and transferring mechanism for law enforcements and military analysts.

References

1. Jhonson, N.F., Jajodia, S.: Exploring steganography: seeing the unseen. In: IEEE Paper of February (1998)
2. Provos, N., Honeyman, P.: Hide and seek: an introduction to steganography. *IEEE Secur. Priv.* **1**(3), 32–44 (2003)
3. Filler, T., Fridrich, J.: Gibbs construction in steganography. In: *IEEE Transactions on Information Forensics and Security* (2010)
4. Bai, L., Biswas, S., Blasch, E.P.: An estimation approach to extract multimedia information in distributed steganographic images. In: *10th International Conference on Information Fusion*, 9–12 July 2007, Quebec, Canada (2007)
5. Shamir, A.: How to share a secret. *Commun. ACM* **22**(11), 612–613 (1979)
6. Thien, C.C., Lin, J.C.: Secret Image Sharing. *J. Comput. Graph.* **26**(5), 765–770 (2002)
7. Das, R., Tuithung, T.: A novel steganography method for image based on Huffman encoding. In: *3rd IEEE National Conference on Emerging Trends and Applications in Computer Science (NCETACS)*, Shillong, India (2012)
8. Cheddad, A., Condell, J., Curran, K., Mc Kevitt, P.: Digital image steganography: survey and analysis of current methods. *J. Sign. Process.* **90**, 727–752 (2010)
9. Li, B., He, J., Huang, J., Shi, Y.Q.: A survey on image steganography and steganalysis. *J. Inf. Hiding Multimed. Sign. Process.* **2**(2), 142–172 (2011)

Chapter 59

An Efficient Beam Scanning Algorithm for Hidden Node Collision Avoidance in Wireless Sensor Networks

Moorthy Sujatha and Raghuvél Subramaniam Bhuvanéswaran

Abstract Wireless sensor networks are characterized by an assembly of low power nodes that collect information about the environmental conditions and report to a base station in general. The hidden node collision problem is one major problem faced by the wireless sensor networks (WSNs). Many solutions have been proposed and implemented with an aim to mitigate to the effect of the same. H-NAME is one such scheme that fulfills the quality-of-service (QoS) requirements imposed by the applications of the WSNs. It relies on a grouping strategy that splits each cluster of a WSN into disjoint groups of non-hidden nodes that scales to multiple clusters that guarantees no interference between overlapping clusters. A design weakness identified in this scheme, energy consumption of nodes, when a new node tries to join a group, has been eliminated by the proposed work. In this paper, an efficient Intra cluster grouping scheme (IC-GS) for a new node to be added into the WSN with and without beam scanning is proposed and simulated using network simulator. The IC-GS with beam scanning is different from the IC-GS scheme as it proposes a beam scanning process at every fixed angle to determine a cluster head using directional antennas before it communicated with the determined cluster head. The simulation results are provided to prove that IC-GS with beam scanning as an energy efficient method for hidden node collision avoidance in WSNs.

59.1 Introduction

Wireless Sensor Networks have found their place in various industrial applications and intelligent systems, where intensive sensing operations take place and the sensed information is continuously reported to either a base station or a server. In a

M. Sujatha (✉)
Sathyabama University, Jeppiaar Nagar, Rajiv Gandhi Salai,
Chennai 600119, TamilNadu, India
e-mail: suja10_pec@yahoo.co.in

R.S. Bhuvanéswaran
Anna University, Chennai, India

WSN, collisions may happen when a receiver is within the transmission range of two transmitters that are transmitting simultaneously, so that the receiver captures neither frame [1]. As each collision represents unnecessary energy dissipation, reducing collisions should be the main design objective in any method. Characteristics of the physical environment lead to a major source of QoS degradation in WSNs—the “hidden node problem”. This problem greatly impacts network throughput, energy-efficiency and message transfer delays, and the problem dramatically increases with the number of nodes.

A lot of research has been done to avoid hidden node collision problems and many schemes have emerged. Many research works have proposed the solutions for eliminating or reducing the impact of the hidden-node problem in wireless networks, roughly categorized as: (1) Request-To-Send/Clear-To-Send (RTS/CTS) mechanisms; and (2) node grouping mechanisms [2]. Grouping mechanisms are effective compared to the other mechanisms mentioned above. In this work, we propose two grouping mechanisms, Intra-cluster grouping scheme (IC-GS) and IC-GS with beam scanning, for the efficient group joining process of a new node in a WSN. The IC-GS with beam scanning is different from the other in the fact that a beam scanning process takes place at every fixed angle to determine a cluster head and communicate with it. These two schemes are energy efficient and produce lesser delay than previous methods. The remainder of this paper is organized as follows. Section 59.2 gives the existing works relevant to this paper. Section 59.3 describes the proposed work followed by the simulation and analysis in Sect. 59.4.

59.2 Related Work

The existing methods have various design issues. The Carrier Sense Multiple Access (CSMA) was first introduced by Kleinrock and Tobagi [3] in which the hidden node problem was later identified in [4]. The Busy Tone Mechanisms as in [4–7] consists of a node that is currently hearing an ongoing transmission sends a busy tone to its neighbors (on a narrow-band radio channel) for preventing them from transmitting during channel use. The reasons why these mechanisms are not energy efficient are that they need additional hardware to be able to provide separate radio channels and they are not cost-effective. The RTS-CTS handshake mechanisms were also proposed [8–14] with an aim to avoid hidden node collision problems. But clearly these mechanisms do not completely solve the problem as stated in [15] because the energy consumption is still high and the RTS/CTS handshake signals are smaller frames and the probability of the hidden node problem reduction is not optimal.

Carrier Sense Tuning Mechanism consists of tuning the receiver sensitivity threshold of the transceiver, which represents the minimum energy level that indicates channel activity, to have extended radio coverage. Higher receiver sensitivities enable a node to detect the transmissions of nodes farther away, thus allowing it to defer its transmission (to avoid overlapping). The works that use this mechanism were proposed in [16–19]. The limitation of this mechanism is that

increasing receiver sensitivity directly leads to more energy consumption, which might not be acceptable for most WSN applications. Even in situations where energy consumption may not be a major concern, it is not possible to indefinitely increase the carrier sense range due to hardware physical limitations [2].

A Grouping Strategy for Solving Hidden Node Problem in IEEE 802.15.4 LR-WPAN was proposed in [20] that solves the hidden node problem quite well. The grouping strategy enhances the IEEE 802.15.4 protocol to solve the HNP by precisely finding HNC and collect the hidden relationships among nodes using a four-phase process. After grouping, the coordinator periodically allocates each group the bandwidth according to the group size among all groups. The reserved channel period for grouping access is named as hidden avoidance guaranteed time slots (HA-GTS) and nodes access their own HA-GTS by applying the standard modified CSMA/CA [20]. By controlling the transmissions of nodes, the grouping strategy can significantly relieve the contentions and improve the network performance. The limitations of this grouping process are: Complexity is high because it needs $O(n^2)$ messages to collect information; grouping is performed at each time when a new node arrives and it does not take number of nodes in each group into consideration.

H-NAME [2] proposed an efficient, practical and scalable approach for synchronized cluster-based WSNs—H-NAME. Importantly, it showed how to integrate the approach in the IEEE 802.15.4/ZigBee protocols with only minor additions and fully respecting backward compatibility which is better than [11]. First, H-NAME requires no hidden-node detection since it relies on a *proactive* approach (grouping strategy is node-initiated) rather than a *reactive* approach to the hidden node problem. Second, the complexity of the group join process was drastically reduced. In this approach, for each group assignment, only the requesting node and its neighbors will be subject to the *group join* procedure and not all cluster nodes, resulting in a simpler, more energy-efficient and scalable mechanism, especially appealing for more densely deployed clusters.

However, a design issue has been identified in the design of H-NAME that arouses the following drawbacks: Hidden node relationship table construction is unreliable due to the fact that the neighbor nodes not able to hear group join request from the new node because they would be in the sleep mode. Therefore HNC table would not be reliable and this leads to a wrong group formation and results in collision. Complexity is high because it needs $O(n^2)$ messages to collect information, grouping is performed at each time when a new node arrives and it does not take number of nodes in each group into consideration.

59.3 Proposed Work

Two methods of grouping mechanisms are proposed to address the design issues in H-NAME. Intra-Cluster Grouping Scheme (IC-GS) aims at eliminating the unreliability grouping in H-NAME and the IC-GS with beam scanning proposes a novel beam scanning algorithm that proves to be better than IC-GS.

59.3.1 Intra-Cluster Grouping Scheme

The Intra-cluster grouping scheme is an efficient and reliable grouping scheme initiated by node. This is a proactive approach where, the process does not wait for a hidden node collision problem to occur, but tries to avoid such a problem before occurrence. Coordinator sends beacon signal to all the nodes in coverage area in order to wakeup nodes. Thereby it meets a reliable information collection process. It consists of four phases:

Hidden node information collection phase. In this phase all the nodes collect information of the nodes hidden and non-hidden to it. The incoming node sends a group joining request to the coordinator as a first step. Coordinator sends Beacon_2 to all nodes. Beacon_2 contains two important messages: (1) Beacon broadcaster role will be performed by new node; and (2) The time, when new node would broadcast Beacon_3 is also a part of the information in Beacon_2. Then the coordinator sends Beacon_2_data information signal to new node, the new node needs to acknowledge it. Beacon_2_data information carries necessary information to be broadcasted and the time when new node can transmit Beacon_3. After new node receives Beacon_2_data information without error it sends Ack signal back to the coordinator. The third step in the hidden node collection phase is the Neighbor notification phase. New Node sends the Beacon_3 and triggers the timer which is used to wait for coordinator collecting responses (Beacon_2 ack) from nodes. Beacon_3 is like the normal beacon but broadcaster here is the new node and the receiving nodes send ack signal to coordinator (Fig. 59.1).

Grouping Phase. The grouping phase is the stage at which the coordinator assigns a new group to the newly joining node based on the hidden node

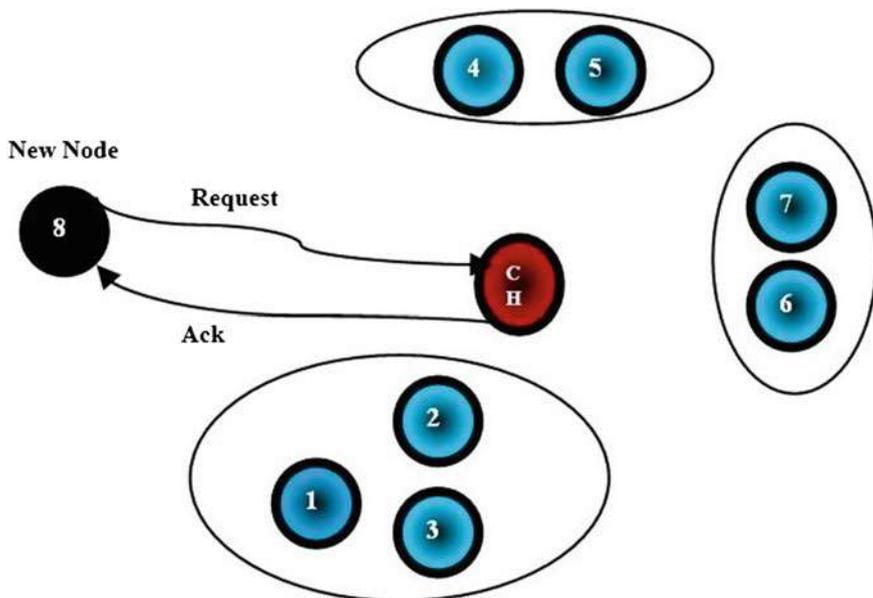


Fig. 59.1 An example scenario where node 8 is the new node and CH is the coordinator

information table. Coordinator allocates a group to a new node, if all nodes in group are not hidden to the new node. It gives priority to the group having minimum of nodes. If none of a group existed with non hidden nodes, then a new group is assigned to the new node.

The grouping algorithm used in this grouping phase is given in the following section. A hidden node information table of the new coming nodes (T) and the number of the existed group (K) are the inputs given to the system. The expected output is that the new node is allocated to one of the group sets G_i .

```

group set  $G = \{G_1, G_2, \dots, G_g\}$ ,  $g \leq K$ ;
while  $|T| > 0$  do
    Pick a node  $X$  having the maximal number of
    hidden nodes from  $T$ ;
        if no existed group then
             $K = 1$ ;
            construct group set  $G_k = \{X\}$ ;
             $T = T - \{X\}$ ;
            return  $G_k$ ;
        else
            while  $|G| > 0$  do
                pick a group  $G'$  having the minimal
                number of nodes from  $G$ ;
                if node  $X$  has no edge to any nodes in  $G'$ 
then
                     $G' = G' + \{X\}$ ; // join into group  $G'$ 
                     $T = T - \{X\}$ ;
                    return  $G'$ ;
                else
                     $G = G - G'$ ;
                end if
            end
            end
             $K++$ ;
             $T = T - \{X\}$ ;
            construct group set  $G_k = \{X\}$ ;
            return  $G_k$ ;
        end
end

```

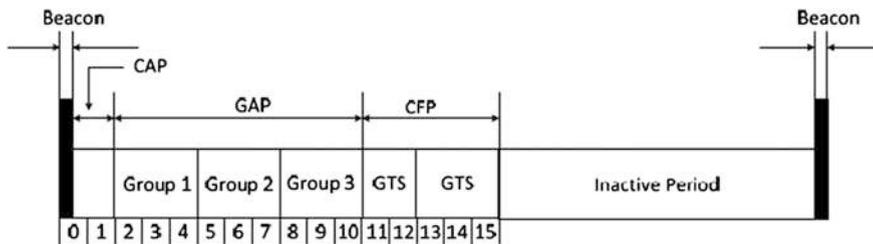


Fig. 59.2 Group slots in the IC-GS scheme

Group access period allocation. The group access period allocation has to be determined for the efficient functioning of this algorithm which is given by the following formula.

$$T_{Gi}(N_i) = \frac{I_{bi}(N_i) \times N_i}{\sum_{i=0}^k I_{bi}(N_i) \times N_i} \times L \tag{59.1}$$

$$I_{bi}(N_i) = \sum_{j=0}^3 P_i \{1 - P_i\}^j \left[I_{j,s} + \sum_{x=0}^{j-1} I_{x,f} \right]$$

Assume that the active period excluding the minimal contention access period (CAP) and the Contention Free Period (CFP) is L in Unit Backoff Period (UBP) and there are total M nodes which can be separated into K groups. Group_i has N_i devices, access period is TG_i, and the probability of a frame successfully transmitted is P_i, for i = 1, 2, ..., k. Let w_j be the size of the contention window (CW) for a frame transmission in the jth backoff retries, and the basic CW unit is UBP.

$P_i = (e^{-\lambda})^{N_i}$, where N_i is the total number of nodes in group i and λ is the frame arrival rate in the system. Since we want to balance the transmission opportunity of all nodes, we let each group have the same transmission times and let each node has the same probability of transmission. And the sum of group access period is, (Fig. 59.2)

$$TG1 + TG2 + TG3 + \dots TGk = L. \tag{59.2}$$

In this phase new node is informed about its group by the coordinator. Therefore, new node can contend to transmit its data in a particular group access period. This is the final phase of the IC-GS scheme and the node is admitted as part of the WSN under a particular coordinator in a suitable group. There exists a limitation that has been identified in this method. When a new node requests to join a WSN by sending a request signal to all the coordinators present in the network, it is possible to still encounter a collision there. And there is a good amount of energy wastage because of more than one cluster head communicating to the new node.

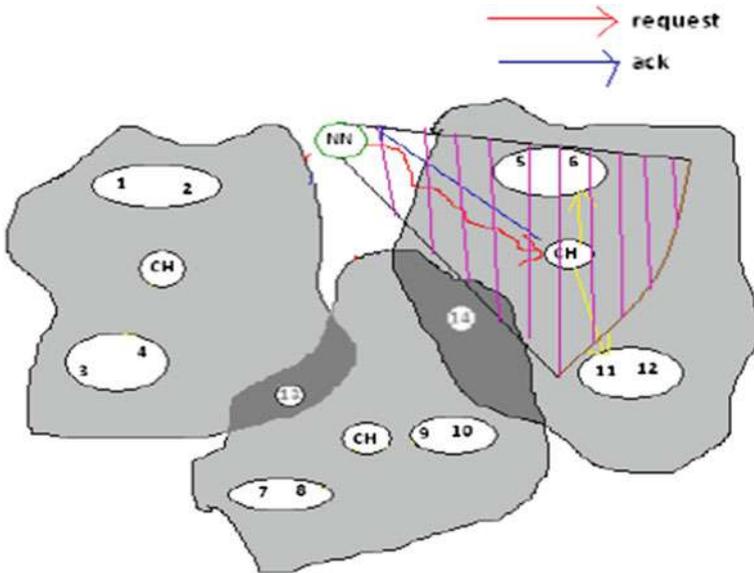


Fig. 59.3 Beam scanning in the IC-GS scheme

59.3.2 IC-GS with Beam Scanning

To eliminate the design issue occurred in the IC-GS scheme, a beam scanning method is used to scan the area at a fixed angle of 60° to find a coordinator that allocates a group in the cluster as a member in its groups (Fig. 59.3).

The new node first broadcasts (360°) the beacon signal to collect the information about available cluster heads and their respective weight (based on minimum number of nodes and maximum energy). After that it creates a cluster head list. It then selects a cluster head from the cluster head list. Next it steers antenna to 60° (beam scanning) and checks the location of selected cluster head by transmitting a beacon signal. If the new node finds the selected cluster head, then it starts the group joining process with the selected cluster head.

59.4 Simulation and Analysis

Simulation is performed by using the simulator NS2. Network simulator is a discrete event time driven simulator that has been extensively used for research purposes, especially for simulation of network scenarios and structures.

The IC-GS scheme and IC-GS with beam scanning are simulated using NS-2.28 with the parameters shown in Table 59.1. The existing H-NAME scheme serves as a standard for comparison of the two schemes proposed here. Figure 59.4 is a

Table 59.1 Simulation parameters for the simulation of IC-GS-with beam scanning

S.No	Parameter	Value
1	Simulation area	1000 × 1000
2	Number of nodes	50
3	Min. no. of nodes in a cluster	10
4	Simulation time	60 ms
5	Radio propagation model	Two ray ground type
6	Antenna type	Omni antenna
7	Initial energy	100 KJ

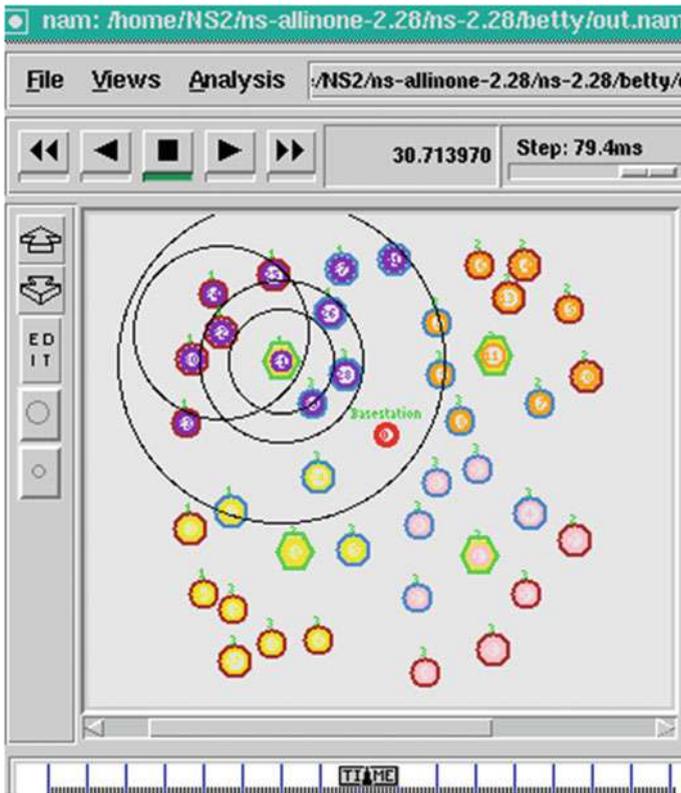


Fig. 59.4 New node joins into cluster group

screenshot of the Network animator (NAM) output obtained during simulation of the proposed schemes. It shows that a new node enters into a wireless sensor network, where all nodes of the same color form a cluster, with the cluster heads highlighted using a hexagon mark. Grouping has been done according to the

grouping algorithm proposed in the earlier sections. Then it sends the Beacon_2 signal to all cluster heads in that network. An acknowledgement is received from all the nodes in response to the beacon signal. It shows that the new node joins the cluster group according to the proposed scheme. Then it will start functioning as a normal member of the cluster belonging to one of its subgroups.

59.4.1 Packet Receive Ratio

See Fig. 59.5.

59.4.2 Throughput

See Fig. 59.6.

59.4.3 End-to-End Delay

In Fig. 59.7, the total time consumed in the process by which a node admitted into the group is calculated by the delay occurred in the system. The graph that follows shows that the overall delay caused by IC-GS with beam scanning is lesser than the IC-GS scheme. IC-GS with beam scanning is the better among all the three

Fig. 59.5 Packet receive ratio

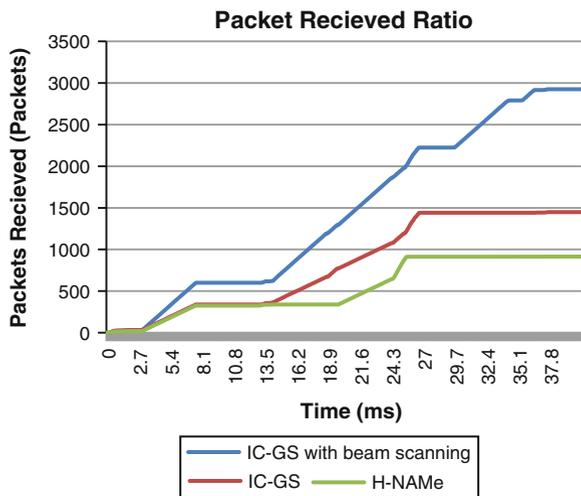


Fig. 59.6 Throughput

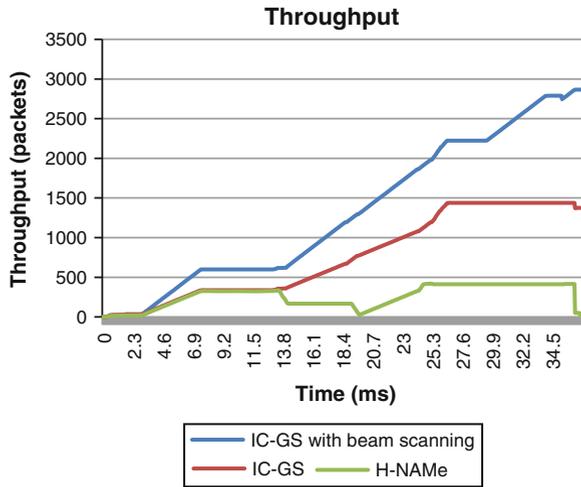
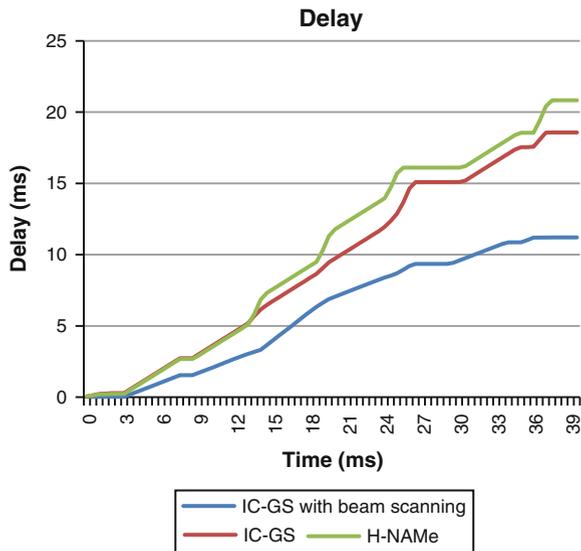


Fig. 59.7 End-to-end delay



schemes for delay reduction. Also, it is significant to mention that the IC-GS scheme finds a solution for a new node to join a group faster than the other schemes. Hence the delay is comparatively less for the IG-CS scheme with beam scanning.

59.4.4 Energy Analysis

The energy remaining in a node during network operations is called as residual energy. The graph in Fig. 59.8 shows that the residual energy of the IC-GS scheme remains greater compared to the other schemes. This is due to the fact that energy consumption is less during the beam scanning phase of the new node. Energy reduction takes place for the nodes that perform sensing, transmit and receive operations including those that perform routing operations. Even though the experiment was run for around 50 runs, for clarity, the screenshot of the initial run is presented in Fig. 59.8.

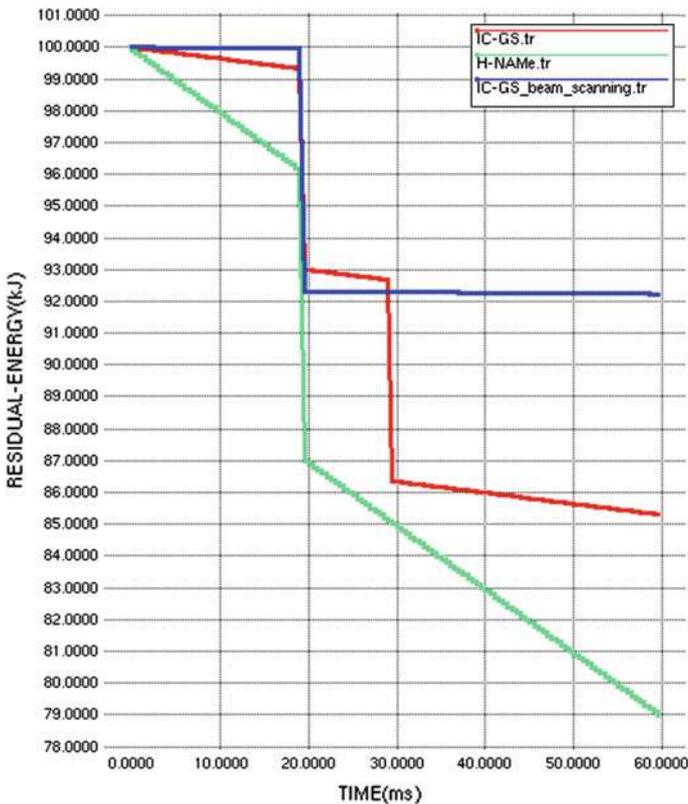


Fig. 59.8 Residual Energy

59.5 Conclusion

Two schemes, IC-GS and IC-GS with beam scanning have been proposed, simulated and analyzed. To minimize the energy consumption during the new node joining process, a beam scanning mechanism was introduced in the IC-GS scheme, which outperforms the IC-GS scheme proposed and also the existing H-NAME method. The simulation results have proved that IC-GS with beam scanning offers improved and better Quality of service.

In future, mobility will be taken into consideration for the IC-GS and the IC-GS with beam scanning schemes.

References

1. Bachir, A., Barthel, D., Heusse, M., Duda, A.: Hidden nodes avoidance in wireless sensor networks. In: Proceedings of the International Conference on Wireless Networks, Communications and Mobile Computing, 2005
2. Koubâa, A., Severino, R., Alves, M., Tovar, E.: Improving quality-of-service in wireless sensor networks by mitigating hidden-node collisions. *IEEE Trans. Ind. Inform.* **5**(3), 299–313
3. Kleinrock, L., Tobagi, F. A.: Packet switching in radio channels: part I—carrier sense multiple-access modes and their throughput-delay characteristics. *IEEE Trans. Commun.* **COM-23**(12), 1400–1416 (1975)
4. Tobagi, F.A., Kleinrock, L.: Packet switching in radio channels: Part II—the hidden terminal problem in carrier sense multiple-access and the busy-tone solution. *IEEE Trans. Commun.* **23**, 1417–1433 (1975)
5. Michalewicz, Z.: *Genetic algorithms + Data structures = Evolution programs*, 3rd edn. Springer, Berlin (1996)
6. Haas, Z.J., Deng, J.: Dual busy tone multiple access (DBTMA)—a multiple access control scheme for ad hoc networks. *IEEE Trans. Commun.* **50**, 975–985 (2002)
7. Chandra, A., Gummalla, V., Limb, J.O.: Wireless collision detect (WCD): multiple access with receiver initiated feedback and carrier detect signal. In: Proceedings of the IEEE ICC, pp. 397–401 (2000)
8. Ji, B.: Asynchronous wireless collision detection with acknowledgement for wireless mesh networks. In: Proceedings of the IEEE Vehicular Technology Conference, vol. 2, pp. 700–704, Sept 2005
9. Tobagi, F.A., Kleinrock, L.: Packet switching in radio channels: Part III—polling and (dynamic) split channel reservation multiple access. *IEEE Trans. Comput.* **24**(7), 832–845 (1976)
10. Karn, P.: MACA—A new channel access method for packet radio. In: Proceedings of the 9th ARRL/CRRL Amateur Radio Computer Network Conference, pp. 134–140 (1990)
11. Bharghavan, V., Demers, A., Shenker, S., Zhang, L.: MACAW: a media access protocol for wireless LAN's. In: Proceedings of the ACM SIGCOMM, pp. 212–225, London, U.K., Aug 1994
12. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, ISO/IEC IEEE-802-11, IEEE Standard for Information Technology (1999)
13. Fullmer, C.L., Garcia-Luna-Aceves, J.J.: Solutions to hidden terminal problems in wireless networks. In: Proceedings of the ACM SIGCOMM, pp. 39–49, Cannes, France, Sept 1997

14. Yang, Y., Huang, F., Ge, X., Zhang, X., Gu, X., Guizani, M., Chen, H.: Double sense multiple access for wireless ad hoc networks. *Int. J. Comput. Telecomm. Netw.* **51**(14), 3978–3988 (2007)
15. Xu, K., Gerla, M., Bae, S.: How effective is the IEEE 802.11 RTS/CTS handshake in ad hoc networks. In: *Proceedings of the Global Telecommunication Conference (GLOBECOM'02)*, vol. 1, pp. 72–76 (2002)
16. Deng, J., Liang, B., Varshney, P.K.: Tuning the carrier sensing range of IEEE 802.11 MAC. In: *Proceedings of the IEEE Global Telecommunication Conference (GLOBECOM)*, vol. 5, pp. 2987–2991 (2004)
17. Ye, F., Yi, S., Sikdar, B.: Improving spatial reuse of IEEE 802.11 based ad hoc networks. In: *Proceedings of the IEEE Global Telecommunication Conference (GLOBECOM'03)*, vol. 2, pp. 1013–1017, San Francisco, CA, Dec 2003
18. Zhai, H., Fang, Y.: Physical carrier sensing and spatial reuse in multirate and multihop wireless ad hoc networks. In: *Proceedings of the IEEE INFOCOM*, pp. 1–12, Apr 2006
19. Ho, I., Liew, S.: Impact of power control on performance of IEEE 802.11 wireless networks. *IEEE Trans. Mob. Comput.* **6**(11), 1245–1258 (2007)
20. Hwang, L.-J., Sheu, S.-T., Shih, Y.-Y., Cheng, Y.-C.: Grouping strategy for solving hidden node problem in IEEE 802.15.4 LR-WPAN. In: *IEEE, Proceedings of the First International Conference on Wireless Internet (WICON'05)*, 2005

Chapter 60

Evaluation of Stereo Matching Algorithms and Dynamic Programming for 3D Triangulation

Teo Chee Huat and N.A. Manap

Abstract A good result of triangulation or known as Three-Dimensional (3D) is depending on the smoothness of the disparity depth map that obtained from the stereo matching algorithms. The smoother the disparity depth map, the better results of triangulation can be achieved. This paper presents the evaluation of the existing stereo matching algorithms in the aspects of the speed of computational on depth map obtained. The stereo matching algorithms that we applied for experimental purpose are basic block matching, sub-pixel accuracy and dynamic programming. The dataset of stereo images that used for the experimental purpose are obtained from Middlebury Stereo Datasets. This research is to provide an idea on choosing the better stereo matching algorithms to work on the disparity depth map for the purpose of 3D triangulation applications, as the good result of 3D triangulation is depending on how smooth is the disparity depth map can be obtained.

Keywords Dynamic programming · 3D triangulation · Stereo matching algorithms · Depth map

60.1 Introduction

Stereo matching algorithms are the procedure to obtain a disparity depth map from the corresponding pixels from the pair of stereo images. Many existing stereo matching algorithms developed by researchers on image processing field as it is an important function in computer vision technology to analyze the two dimensions, (2D) and three dimensions, (3D) of output based on stereo images. In developing

T.C. Huat (✉) · N.A. Manap
Faculty of Electronics and Computer Engineering, Department of Computer Engineering,
Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: nictco@student.utem.edu.my

N.A. Manap
e-mail: nurulfajar@utem.edu.my

an algorithm of stereo matching, the accuracy and speed are conditions that need to be concerned to produce a precise output of computer vision. As to get better results of disparity on stereo images, there are many stereo matching methods evaluated on the stereo datasets for this research such as dynamic programming, basic block matching, sub-pixel estimation, and image pyramiding. The datasets used for the experimental on the stereo matching algorithms are from the Middlebury Stereo Vision Page which is an open source of datasets and evaluation on stereo vision algorithms.

The existing stereo matching algorithms mostly will go through the standard steps like the matching cost computation, cost aggregation, disparity optimization and disparity refinement. There are some equations can be included to find the pixel-based matching cost, which are the Sum of Absolute Differences (SAD), Sum of Squared Difference (SSD), Sum of Truncated Absolute Differences (STAD), Normalized Cross Correlation (NCC) and Zero Mean Normalized Cross Correlation (ZMNCC). In this particular research, the SAD equation is used to get the corresponding points between the original and target of stereo images. The cost of the correspondence between the original and target data set can be represented by Disparity Space Image (DSI) [1–3]. The DSI is formed from the matching cost values from the disparities and pixels when summing up the cost values on each matching image data set [4].

In order to find the disparity depth map, the basic blocking matching is used as it is the initial step to find the absolute difference between the pixel intensities by using the Sum of Absolute Differences (SAD) to compare the correspondence pixels on the left and right of stereo images [1, 2, 4]. One of the examples of dataset used for stereo matching experimental is the Tsukuba stereo pairs, taken from Middlebury Stereo Datasets. The disparity map is calculated using three sizes of window which are 3 by 3, 5 by 5 and 7 by 7 pixel of block around the original of image data. In basic block matching, it locates the pixels in the range of ± 15 over columns sequence [5]. On the step of cost aggregation, there are two types of support region which are the two dimensional, 2D fixed disparity and the three dimensional, 3D with the variables of x , y and d as in Eq. 60.1. The 2D aggregation is by using square windows, shiftable windows and windows with adaptive sizes.

Furthermore, the disparity optimization can be categorized into two approaches which are the local approaches which use the Winner Takes All (WTA) by picking each pixel where the disparity is correlated to the minimum cost value in order to increase the signal to noise ratio (SNR) while to reduce the ambiguity. Another approach which is the global approach, it is a frame work to find for disparity function, d that minimize an energy function or global energy over the disparity computation phase like the pixel-based matching cost [1],

$$E(d) = E_{data}(d) + E_{smooth} \quad (60.1)$$

where $E_{data}(d)$ represent the disparity function with the input of data set which minimized through the pixels on the correspondence of disparity map, d if there is similarity in intensities and maximized when the disparity map putting the pixels

in correspondence which slightly differ in intensities. The E_{smooth} represent the conjecture of the smoothness made from the algorithm which measured from the disparity between the pixels on pixel grid [1, 2, 4].

On disparity refinement, there are many approaches to improve the smoothness of the disparity map. In this research, the method of dynamic programming is chosen as the approach to improve the output obtained from the basic block matching stereo correspondence algorithm. Dynamic programming is able to find the global minimum of independent scanlines when in polynomial time for smoothness functions [1]. It is chosen based on its accuracy when dealing at the areas of depth borders and uniform regions [6, 7].

60.2 Algorithm Outlines

60.2.1 Basic Block Matching

Block matching is a technique that used to find the corresponding pixels in stereo matching datasets. The basic block matching is used in this research as the initial step to find corresponding pixels as of the dataset that used throughout the experimental of stereo matching. During the experimental phase, the pixel value of the target image is predicted as the corresponding pixel in the reference image where the displacement of the corresponding pixels or as the motion vector to be approximated using the block matching [8]. The block matching is used to reduce the matching errors between the block at position of (x, y) in the target image, I_t while for the position of the reference image, I_{t-1} will be $(x + u, y + v)$ where u and v is the motion vector. These variable defined can be reviewed as sum of absolute difference (SAD) [8],

$$(u, v) \equiv \sum_{j=0}^{S-1} \sum_{i=0}^{S-1} |I_t(x + i, y + j) - I_{t-1}(x + u + i, y + v + j)| \quad (60.2)$$

where S is the block size, $S \times S$, i and j representing the pixels. As to minimize the $SAD_{(x, y)}(u, v)$, (a, b) is defined as the motion vector estimation to compare and obtain the SAD of each position, $(x + u, y + v)$ for the dataset. The equation shows as [8],

$$(a, b) \equiv \arg \min_{(u,v) \in Z} SAD_{(x,y)}(u, v) \quad (60.3)$$

where $Z = \{(u, v) | -B \leq u, v \leq B \text{ and } (x + u, y + v) \text{ represent the suitable position of pixel in the reference image, } I_{t-1}\}$ while B is an integer to seek for range. From the SAD equation, the global minimum of matching error can be attained. Block matching algorithm able to save more computations to obtain a depth map but it still have its disadvantage which is its disability to guarantee the

global matching error. Besides that, the main issue with block matching algorithm is that the prediction on the position of the images sequence is not accurate enough and this may cause on the minimum matching error will be larger. In addition, the most critical issue is that the matching errors are decreasing directly proportionally to the coming order of the search positions.

60.2.2 Sub-Pixel Accuracy

Most of the stereo matching algorithms apply sub-pixel refinement or the sub-pixel accuracy after attaining the correspondence points of the stereo datasets. Sub-pixel accuracy is a technique that goes through isolated disparity levels that from the matching cost and also the iterative gradient descent [9]. The major point of the sub-pixel accuracy method is to enhance the resolution of the stereo matching algorithm output from the stereo datasets. This method is able to smooth the alteration between the regions from different disparity that cause contouring effect on the images on a depth map. Through the progression of sub-pixel accuracy, it will focus on the minimum cost and the neighboring cost values to acquire the sub-pixel alteration. In applying the technique of sub-pixel accuracy, the Normalized Cross Correlation (NCC) is used to work out the essential stereo images where the cross correlations at the sub-pixel location of the stereo images can be figured proficiently and the equation used for computation is shown as following where the NCC (x, y, u, v) is equal to [9],

$$\frac{\sum_{(i,j) \in w} I_1(x+i, y+j) \cdot I_2(x+u+i, y+v+j)}{\sqrt{[2]I_{1^2}(x, y) \cdot I_{2^2}(x+u, y+v)}} \quad (60.4)$$

where the NCC can be defined as left image window at the position (x, y) while for the right image window position at $(x+u, y+v)$. The I_1 and I_2 are represented as target image and reference image. From the NCC equation, sub-pixel accuracy can be computed with NCC through integral images as it can be substituted by using integral stereo images which with squared of pixel values. There are limitations can be found from sub-pixel accuracy algorithm as it unable to determine the local ambiguities successfully. Another issue on sub-pixel accuracy is it causes the right matching unrecognized to be count in for local maximum and this bring failures for the usage of local maxima [9].

60.2.3 Dynamic Programming

For the disparity optimization phase, the dynamic programming is selected as global optimization, which is optimizes energy function to be non-deterministic polynomial-time hard (NP-hard) for smoothness purpose [1, 6]. There are two

categories of global optimization such as one dimension and two-dimension optimization methods. One dimension optimization is traditional technique, where its evaluation on the disparity is focusing on a pixel that depending on other pixels on the same scanline, but independent on disparity that focus on other scanlines. One dimension is not considered as a truly global optimization as its smoothness technique is only focus on horizontal direction. However, one dimension optimization is still being used by some of the researchers due to its simple implementation and its usefulness on the disparity maps outputs.

Two-dimension optimization approach is smoothing the stereo images in the vertical and horizontal direction by using simulated annealing, continuation methods and mean-field annealing [10–12]. However, these methods are not well-organized enough to optimize the equation in (60.1). There are two methods which companionable in optimizing the Eq. (60.1), the graph-cuts and belief propagation as these two methods able to attain better results accordingly to ground truth data from stereo matching algorithm [13–15] In this paper, the dynamic programming used for the experimental results is the dynamic programming on tree due to its competence as the one dimension optimization. The tree graph of the dynamic programming can represent as $T(V, E)$ where V as vertices and E as edges. The effectiveness of the dynamic programming on tree begins with its optimization on the energy function [6],

$$E(d) = \sum_{a \in V} m(d_a) + \lambda \sum_{(a,b) \in E} s(d_a, d_b) \tag{60.5}$$

where a is the pixel in the left image and the d_a as the value of disparity map, d at the pixel of a . Assuming $m(d_a)$ is the matching consequence of relating d_a to the pixel of a where it is the absolute difference between the pixel, a in the left image and the a pixel which shifted to the right image and can be summarized as $\sum_{a \in V} m(d_a)$. Meanwhile, assume $s(d_a, d_b)$ as the smoothness consequence for the disparity of d_a and d_b to the pixel a and b and the variables can be summarized as $\sum_{(a,b) \in E} s(d_a, d_b)$.

As to get the minimum energy of Eq. (60.5), let h as the root vertex of tree, $h \in V$ and assume $z \in V$ as the number of edges the root of distance between h and z . Each node of z belongs a parent as $a(z)$ and the depth is equally to the depth of $z-1$ while if it is not a root, the minimum value of the energy in Eq. (60.5) have a sub-tree rooted at z and the edge in the middle of z and $a(z)$ can be summarized as $d_{a(z)}$ [6],

$$E_z(d_{a(z)}) = \min_{d_z \in D} (m(d_z) + s(d_z, d_{a(z)}) + \sum_{w \in C_z} E_w(d_z)) \tag{60.6}$$

where C_z as the children set of z . while for the optimal disparity for the root node h can be represented as [6],

$$L_{h^*} = \arg \min_{d_h \in D} (m(d_h) + \sum_{w \in C_h} E_w(d_h)) \quad (60.7)$$

where if z is a node that without children then C_z is empty and the function of E_z and L_z can be estimated directly. h^* shows that the variable can be replaced with z and w . Let take J as the maximum depth in the tree, the energy function of Eq. (60.5) is optimized by appraising the functions E_z and L_z for each node z at the depth, J . After estimation on the functions, proceed with the estimation on the same functions for all the nodes at depth of $J - 1$ due to any child w has the depth of J , this is the evaluation on E_w and L_w . Next step is to keep appraising the function of E_z and L_z in decreasing order for the depth till it reach to the root for disparity assignment optimal computation use. Dynamic programming on a tree is an algorithm that simple to be implemented and efficient as well on the results obtained compared to most of the stereo matching algorithms but there still a disadvantage of it which is its speed is slower due to its tree traversal is less capable than its array traversal [6]. Therefore, based on the characteristics of the three stereo matching algorithms described, dynamic programming has been selected as the main stereo matching algorithm due to its accuracy and its capability in smoothing most of noise in the datasets of stereo images compared to the other two algorithms, basic block matching and sub-pixel accuracy.

60.3 Experimental Results

There are three existing stereo matching algorithms evaluated in this paper which are the basic block matching, sub-pixel accuracy, and dynamic programming. From the three stereo matching algorithms, dynamic programming is the well-organized algorithm among the other algorithms in smoothing the disparity depth map. The helpfulness of dynamic programming in smoothing the depth map is also depending on the appropriate disparity range (DR) of the stereo pair of images used. The stereo pairs of images that used in this research are Tsukuba, Teddy, Sawtooth and Venus which these stereo pairs of images are chosen from Middlebury Stereo Pages. The results obtained are shown as Table 60.1, Figs. 60.1 and 60.2.

From Table 60.1, the disparity range of each stereo images pair is determined based on the smoothness of the basic depth map that obtained from the stereo matching algorithms. The biggest number of the disparity range is most preferable to be used for stereo matching algorithms process, for example Sawtooth stereo pairs sample from Table 60.1, the suitable and the most preferable disparity range to obtain a good result is by using 30, while for Tsukuba is by using 16. The interval range that provided as results in Table 60.1 are to show that few number of disparity range on the range can obtain slightly good results.

Table 60.1 Disparity range

Stereo images	Tsukuba	Teddy	Sawtooth	Venus
Disparity range	15–16	45	27–30	30

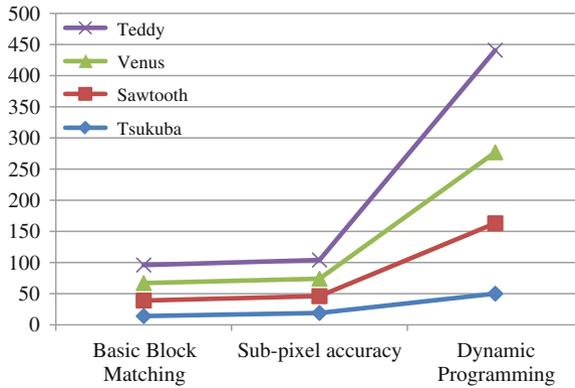


Fig. 60.1 Time taken (in second) for different sets of stereo image applied on each type of stereo matching algorithm. For the dataset of Teddy used the disparity range of 45, Venus used disparity range of 30, Sawtooth used 29 and Tsukuba used 16 as the disparity range

From the evaluation on the results of three stereo matching algorithms, the precision which depends on smoothness in rising order is starting from sub-pixel accuracy followed by basic block matching and the smoothest results are obtained by the dynamic programming algorithm as revealed in Fig. 60.2. The hardware used to run the simulation of the three stereo matching algorithms is the portable computer with integrated of processor of 2.5 gigahertz (GHz) and three gigabytes (GB) of installed memory. Table 60.1 shows on the disparity range for each dataset of stereo image where the disparity range are achieved from the experiment by applying stereo matching algorithms and observing on the output accuracy for different disparity range applied on every dataset coding. Figure 60.1 shows the results of the time taken in second for each stereo matching algorithm on different sets of stereo images, where the computation competence of the dynamic programming is the lowest among the stereo matching algorithms. Meanwhile the highest computation effectiveness of stereo matching algorithms is the basic block matching where its average time taken for all the stereo images datasets is more rapidly than the average time taken of sub-pixel accuracy algorithm. Besides that, from the results acquired in Fig. 60.1, it also shown that the higher disparity ranges of stereo images datasets, the longer time taken to run on the stereo matching algorithms.

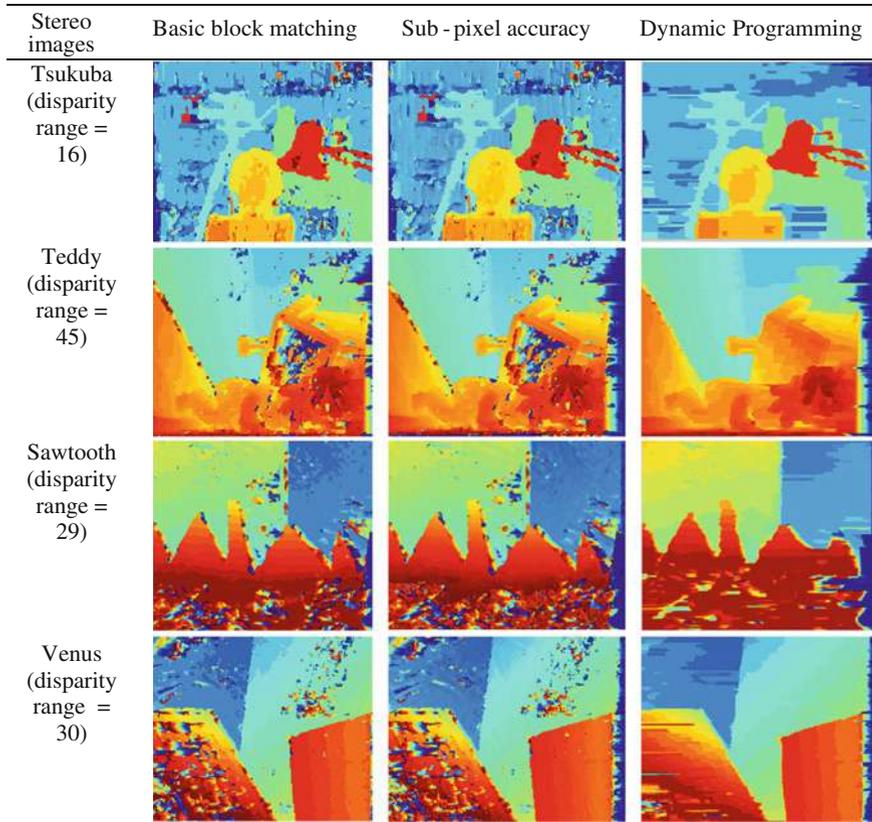


Fig. 60.2 Results achieved from stereo matching algorithm

60.4 Discussion

From observation on each stereo matching algorithms, it is found that each of it have the issues faced or the imperfect of the algorithms developed. Therefore, for each issue that been through, we will try to overcome it by modifying the equations used on each stereo matching algorithm. For example, in block matching algorithm we can attempt to combine some other equations such as the two dimension optimization to search for the corresponding pixels in horizontal and vertical so that the prediction on the position of images sequence are much more accurate than before. Besides that, for sub-pixel accuracy as it is used to improve the resolution, it can be added some part of algorithm in improving the contouring effect by applying the suitable disparity range. Therefore, by pairing with the algorithm with sub-pixel accuracy, the dataset can be doubled up in smoothing the ambiguities. For dynamic programming, as the results obtained are the most

accurate among the other stereo matching algorithms, the only issue that we observed is the speed is slower than the other stereo matching techniques. However, the time taken is a minor issue for the scope of this research.

60.5 Conclusion

As a conclusion, the evaluation between the three stereo matching algorithms can clearly shows that dynamic programming is the most capable technique between the others in smoothing the depth map. It is depending on the appropriate disparity for diverse content of stereo pair of images used. Besides that, from the experimental outcomes by using the Middlebury datasets shows that dynamic programming algorithm able to reduce the matching errors and acquire an improved stereo matching result compare to the basic block matching and sub-pixel accuracy algorithm. Therefore, based on the comparison among the three stereo matching algorithms, it is suggested that dynamic programming could helps to obtain more agreeable effect of depth map especially in removing the apparent defective stripes. The depth map that obtained from the dynamic programming algorithm is effective for most applications compared to basic block matching and sub-pixel algorithm such as appliances on 3D intentions.

Acknowledgements This research is funded by the grant from Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM) with project number of PJP/2013/FKEKK (16C)/S01203.

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Technical Report MSR-TR-2001-81, Microsoft Research (2001)
2. Bleyer, M., Gelautz, M.: A layered stereo matching algorithm using image segmentation and global visibility constraints. *J. Photogram. Remote Sens.* **59**(3), 128–150 (2005)
3. Di Stefano, L., Mattoccia, S., Mola, M.: An efficient algorithm for exhaustive template matching based on normalized cross correlation, *IEEE Comput. Soc.* **2**, 322–327 (2003)
4. Tombari, F., et al.: A 3D Reconstruction System Based on Improved Spacetime Stereo. In: *IEEE*, pp. 7–10 (2010)
5. Birchfield, S., Tomasi, C.: Depth discontinuities by pixel- to- pixel stereo. In: *ICCV*, pp 270–293 (1998)
6. Veksler, O.: Stereo correspondence by dynamic programming on a tree. *CVPR* **2**, 384–390 (2005)
7. Olofsson, A.: Modern stereo correspondence algorithms: investigation and evaluation. PhD thesis, Linköping University, Sweden, 5–86, June 2010
8. Chen, Y., Hung, Y., et al.: Fast block matching algorithm based on the winner update strategy. *IEEE* **10**(8), 1212–1222 (2001)
9. Donate, A., Wang, Y., et al.: Efficient and accurate subpixel path based stereo matching. *IEEE* **8**(6), 1–4 (2008)

10. Geiger, D., Giosi, F.: Parallel and deterministic algorithms from MRF's: surface reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**(5), 401–412 (1991)
11. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984)
12. Blake, A., Zisserman, A.: *Visual Reconstruction*. MIT Press, Cambridge (1987)
13. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: *ECCV02*, p. III: 82 ff. (2002)
14. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: *CVPR04*, p. I: 261–268
15. Sundstr, O., Guzzella, L.: A generic dynamic programming matlab function. In: *IEEE International Conference on Control Applications*.**1**(7), 1625–1630 (2009)

Chapter 61

Image Enhancement Filter Evaluation on Corrosion Visual Inspection

Syahril Anuar Idris and Fairul Azni Jafar

Abstract This project is focusing on corrosion inspection using image. Inspection which have particularly challenging environmental conditions and characteristics, increase the complexity of the inspection operation. By using software image filter to enhance the image data, it is believe that the object recognition technique will be able to analyse the image data accurately. A few software filters have been identified in this works based on textural feature and colour progression factor that are the characteristic of image corrosion. Therefore, in order to obtain suitable software image filter, neural network is use for optimization. The experiment result shows among those identified image enhancement filters for visual corrosion inspection, Wavelet De-noising gives desirable result in terms of Mean Square Error, Peak Signal to Noise Ratio and Neural Network optimization.

61.1 Introduction

These days, utilization of camera as inspection tools has been expanding. The flexibility functions of camera, fit to get different sorts of information, for example, position, speed, rate of development and others from single apparatus. However, in visual inspection the quality of raw image affect the accuracy of inspection result. The quality of raw image can be enhanced externally or internally, by externally means to control the inspection environment. Unfortunately for corrosion inspection, inspection environment is hard to be controlled. Therefore to obtain

S.A. Idris (✉)

Center of Graduate Studies, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
e-mail: syahrilidris@gmail.com; syahril.idris.my@ieee.org

F.A. Jafar

Faculty of Manufacturing Engineering, Universiti Teknikal Malaysia Melaka,
Melaka, Malaysia
e-mail: fairul@utem.edu.my

good quality image for visual inspection, the inspection system need to be able to enhance raw image internally, after the image acquisition process.

There are lot of methods for image enhancement process. One of the popular enhancement method used for inspection is filtering [1]. In selecting suitable filters, an understanding of target inspection is required. Common question arise for the selection of suitable image enhancement filter is “what kind of data needed to be extract out from target?” By knowing the type of data required in inspection, one can clustered the possible filter to be used for enhancement process.

This research work is focusing on selecting best image enhancement filter use for visual corrosion inspection. The end-result should present the comparison of performance between each recognized image enhancement filter, and identified the best filter for enhancement corrosion image.

61.2 Visual Corrosion Inspection

One of the primary requisitions of vision system is application in inspection system. By utilizing vision system as part of inspection system, the accuracy and reliability of value could be setup during the predetermine range. Visual inspection systems are utilized to check parts for dimensional accuracy and geometrical integrity. Regardless of the fact that human able to perform image inspection using their own eyes, the detail repetition of some inspection tasks is simply beyond the capability of humans.

Remote Visual Inspection or RVI is one method of non-destructive testing (NDT) use in corrosion inspection [2]. RVI is a visual examination method that aids in acquisition of visual information by utilizing visual equipment but not limited to video pan/tilt/zoom cameras, borescopes, push cameras, or automated crawlers. It is frequently used where distance, angle of view and restricted lighting may hinder direct visual examination or where access is constrained by time, financial constraints or atmospheric hazards.

There are lot of different experimental and analysis methods used to identify corrosion damage for inspection and monitoring purposes. One of the methods is mechanical measurements (weight loss, chemical analysis, and visual inspections). In visual inspections, the corrosion level identification requires expert who can clearly determine the corrosion based on experience as well as types of corrosion, with red rust as a common experience. Usually, the corrosion process produces rough surfaces, and image analysis based on textural features can be used for quantification and discriminate corrosion extent and type [3, 4]. Additionally to textural features, colour progressions of metallic surfaces are also used for the detection of corrosion because of different metal oxides and other corrosion products [5].

61.3 Image Enhancement Filter

As explain in previous topic, image analysis for corrosion can be identified by textural feature and colour progression. Thus in enhancement image process, selected filter must be able to preserve both characteristics of the image. The image enhancement filters identified for this study emphasizes on preserving these characteristics. Table 61.1 lists the image enhancement filter based on “How they do it?”, and “What do they enhance?”

There are several methods in selecting best filter for image enhancement, and of them is by calculating the image error measurement. Image error can be measure by calculating Mean Square Error (MSE) and Peak Signal Noise-to-Ratio (PSNR).

However, for corrosion visual inspection, selecting one suitable filter only based on image error measurement is not accurate. This is because, image error measurement does not consider on textural feature and colour progression of corrosion image. Therefore, neural network is used to determine the best filter for corrosion visual inspection by taking consideration on preserving textural feature and colour progression of corrosion image.

61.3.1 Image Error Measurement

The Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are the two error metrics used to compare image compression quality. The MSE represents the cumulative squared error between the compressed and the original image, whereas PSNR represents a measure of the peak error. The lower value of MSE shows that the image error is low [12].

The PSNR block computes the peak signal-to-noise ratio, in decibels, between two images. This ratio is often used as a quality measurement between the original and a compressed image. The higher the PSNR value, the better the quality of the compressed or reconstructed image [12].

To compute the PSNR, the block first calculates the mean-squared error using the following equation:

$$MSE = \frac{\sum_{M,N}[I_1(m,n) - I_2(m,n)]^2}{M * N} \quad (61.1)$$

In the Eq. (61.1), M and N are the number of rows and columns in the input images, respectively. Then the block computes the PSNR using the following equation:

$$PSNR = 10\log_{10}\left(\frac{R^2}{MSE}\right) \quad (61.2)$$

Table 61.1 Image enhancement filter characteristics

Homomorphic [6]	Bayer [7]	Wavelet de-noising [8]	Gaussian [9]	Linear [10]	Anisotropic diffusion [11]
Image enhancement and correction	Filter mosaic, colour filter array (CFA)	Reduces high frequency noise	Give no overshoot to a step function input while minimizing the rise and fall time	Improve images in many ways: sharpening edges, reducing random noise, correcting unequal illumination, etc.	Reducing image noise without removing significant parts of the image content
Simultaneously normalizes the brightness across an image and increases contrast	Arrange RGB colour filters on a square grid of photo sensors	Direct wavelet transform is computed from the original image	Impulse response on Gaussian function	Based on the same two techniques as conventional DSP: convolution and Fourier analysis	Image generates a parameterized family of successively more and more blurred images based on a diffusion process

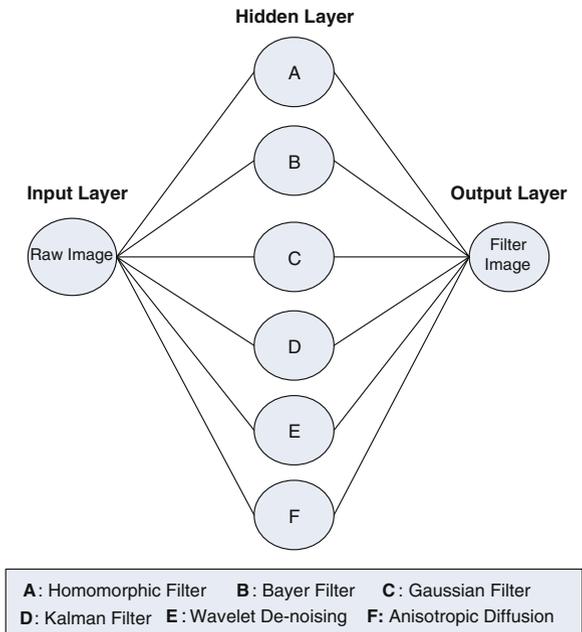
In the Eq. (61.2), R is the maximum fluctuation in the input image data type. For example, if the input image has a double-precision floating-point data type, then R is 1. If it has an 8-bit unsigned integer data type, R is 255.

61.3.2 Optimization: Neural Network

In this research, artificial neural network will be used for optimizing image enhancement filter algorithm. As network representation provides such powerful visual and conceptual aid for portraying the relationship between the components or tools of systems that it is used in virtually every field of scientific, social, and economic endeavor [13]. In artificial neural network, the elements called as neurons or node, process the information. The signals are transmitted by means of connection links, and these links possess an associated weight which is multiplied along with the incoming signal (input) for any typical neural network. Finally the output signal is obtained by applying activations to the network input [14].

Figure 61.1 represent neural network used for the optimization of the enhancement filter selection. Each filter will be denoted by each single node on hidden layer. Raw image will be the input and final selected filter image would be the output for this network. The weight for each node is determined as shown in Table 61.2.

Fig. 61.1 Neural network map for optimizing image enhancement filter used on visual corrosion inspection



From Table 61.2, the weight for Linear Filter is 0.29, Bayer Filter is 0.07, Homomorphic Filter is 0.14, Gaussian Filter is 0.07, Wavelet De-Noising is 0.29 and lastly Anisotropic Diffusion is 0.14. The PSNR values obtain from each filter will be multiplied with the weight of each filter. Highest value will be taken as best filter for corrosion visual inspection

61.4 Experiment Result

The experiment of finding suitable image enhancement filter for corrosion visual inspection has been tested on 3 corrosion images (corrosion 01, corrosion 02, and corrosion 03) as in Fig. 61.2. The sample image use in the experiment is RGB image and the size is converted to become square image for ease of filter instalation on the image. Figure 61.3 shows the result of filtered image for each sample image. In addition Table 61.3 shows the RMSE and PSNR value for each image enhancement filter. Blue highlighted block in the table, show the maximum value of PSNR compared to other filters. While red highlighted blocks show the highest value for each PSNR value multiplied with weight that determined earlier. From the Table 61.3, the suitable image enhancement filter for corrosion visual inspection is Wavelet De-noising.

Table 61.2 Weight determination for image enhancement filters

Filters	Edge preserving	Colour restoration	De-noising	Fix unequal illumination	Sharpen image	Total	Weight, w_i
Linear	-	1	1	1	1	4	0.29
Bayer	-	1	-	-	-	1	0.07
Homomorphic	1	-	-	1	-	2	0.14
Gaussian	-	-	1	-	-	1	0.07
Wavelet De-noising	1	1	1	-	1	4	0.29
Anisotropic Diffusion	1	-	1	-	-	2	0.14

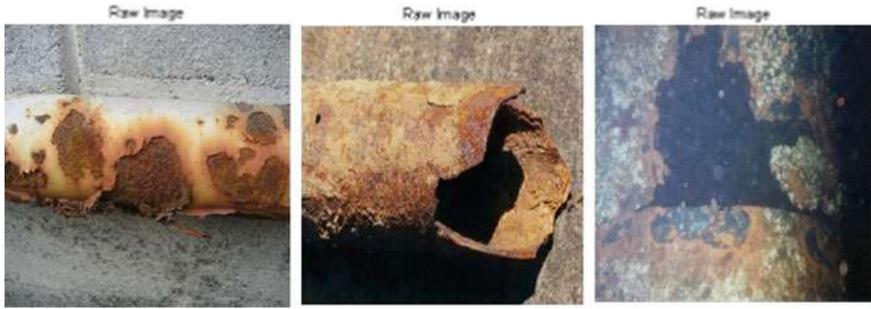


Fig. 61.2 Raw images (input)

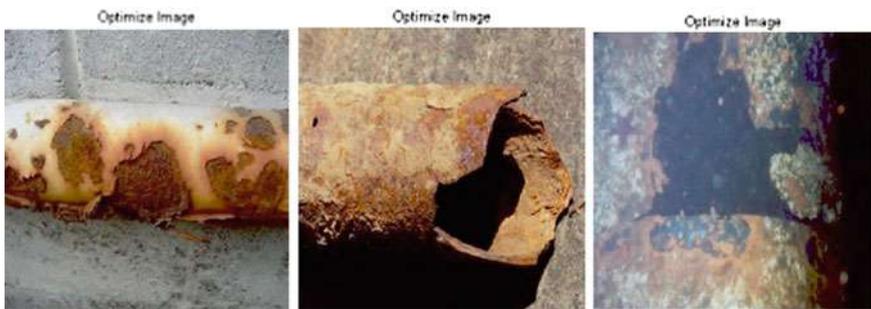


Fig. 61.3 Filtered images (output)

61.5 Discussion

From the result, the selected suitable image enhancement filter for corrosion visual inspection is Wavelet De-noising. In Table 61.3, each corrosion image tested in the experiment, Wavelet De-noising gives the highest value compare to other filters. Although in image corrosion 03, the highest PSNR value obtains by Gaussian filter. However, due to the weight identified earlier for each filter, Wavelet De-noising has the highest value and selected as best filter for corrosion visual inspection.

Figure 61.4 shows the result for filtered images on image corrosion 03. Even though the Gaussian filter image obtains highest PSNR, the textural feature of corrosion is tempered, thus affect the corrosion extraction data, resulting for possible un-accurate inspection result.

Table 61.3 Result of image enhancement filter optimization on corrosion image

Filter	Corrosion 01					Corrosion 02					Corrosion 03				
	PSNR	RMSE	Weight, w_i	Total	Total	PSNR	RMSE	Weight, w_i	Total	Total	PSNR	RMSE	Weight, w_i	Total	Total
Linear	61.232	0.22	0.29	17.757	17.757	57.435	0.34	0.29	16.656	16.656	60.438	0.24	0.29	17.527	17.527
Gaussian	62.197	0.20	0.07	4.354	4.354	58.158	0.32	0.07	4.071	4.071	60.757	0.23	0.07	4.253	4.253
Wavelet DN	62.217	0.20	0.29	18.043	18.043	58.295	0.31	0.29	16.905	16.905	60.443	0.24	0.29	17.529	17.529
Bayer	60.440	0.24	0.07	4.231	4.231	57.897	0.32	0.07	4.053	4.053	60.296	0.25	0.07	4.221	4.221
Homomorphic	60.672	0.24	0.14	8.494	8.494	57.195	0.35	0.14	8.007	8.007	60.690	0.24	0.14	8.497	8.497
Anisotropic DF	61.555	0.21	0.14	8.618	8.618	56.631	0.38	0.14	7.928	7.928	60.745	0.23	0.14	8.504	8.504
	Max PSNR	Min RMSE				Max PSNR	Min RMSE				Max PSNR	Min RMSE			
	62.217	0.20				58.295	0.31				60.757	0.23			

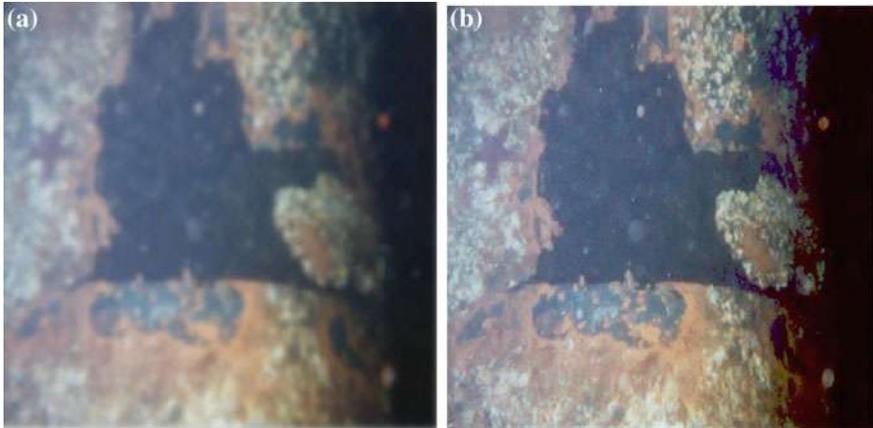


Fig. 61.4 Filtered image for image corrosion 03

61.6 Conclusion

This paper identified several image enhancement filters, applicable in corrosion inspection. The image enhancement filters emphasize on textural features and colour progression that are the characteristics of surface defects created due to corrosion. The performances of the filter are compared and analysed by PSNR and MSE values for corrosion images.

Using Neural Network to incorporate textural features and colour progression that are characteristic of corrosion, from the result, the most suitable image enhancement filter to be use on corrosion visual inspection is Wavelet De-noising. Though the Gaussian filter shows higher PSNR value in images corrosion 03, due to the neural network weight identified based on image corrosion characteristic, wavelet De-noising was selected as the best suitable filter for corrosion visual inspection. The selected filters are to be analysed in term of the possibility for any combination, which may produce a better image enhancement quality.

References

1. He, K., Sun, J., Tang, X.: Guided image filtering. In: Computer Vision–ECCV 2010. pp. 1–14, Springer, Berlin, Heidelberg (2010)
2. Remote Visual Inspection. Advantech Alliance Sdn Bhd. [http:// www.advantech.net.my](http://www.advantech.net.my). Accessed 16 Dec 2013
3. Livens, S., et al.: A texture analysis approach to corrosion image classification. *Microscopy microanalysis microstructures* 7.2, p. 143 (1996)
4. Pidaparti, R.M., Hinderliter, B., Maskey, D.: Evaluation of corrosion growth on SS304 based on textural and color features from image analysis. In: *ISRN Corrosion*, vol. 2013, Article ID 376823, 7 p (2013). doi:[10.1155/2013/376823](https://doi.org/10.1155/2013/376823)

5. Medeiros, F.N.S., et al.: On the evaluation of texture and color features for nondestructive corrosion detection. In: EURASIP Journal on Advances in Signal Processing 2010, p. 7 (2010)
6. Hamblin, J.D.: Oceanographers and the Cold War: Disciples of Marine Science, University of Washington Press, Seattle (2005)
7. Compton, J., Hamilton, J.: Color Filter Array 2.0., A Thousand Nerds: A Kodak blog. <http://archive.today/vqZt4> (2013). Accessed Dec 2013
8. Larson, D.R.: Unitary systems and wavelet sets. In: Wavelet Analysis and Applications. Appl. Numer. Harmon. Anal. Birkhäuser, pp. 143–171 (2007)
9. Williams, D.B., Vijay, M.: The Digital Signal Processing Handbook, Second Edition. CRC Press, p. 438. ISBN: 978-1-4200-4606-9 (2009)
10. Gonzalez, C., Woods, E., Eddins, L.: Digital Image Processing Using MATLAB®, Pearson Education Inc., p. 155. ISBN 81-7758-898-2 (2007)
11. Sapiro, G.: Geometric Partial Differential Equations and Image Analysis. Cambridge University Press, p. 223 (2001). ISBN 978-0-521-79075-8
12. Padmavathi, G., Subashini, P., Muthu, M., Suresh, T.: Comparison of filters used for underwater image pre-processing. IJCSNS Int. J. Comput. Sci. Netw. Secur. **10**(1):58 (2010)
13. Hillier, S., Lieberman, J.: Introduction to Operation Research. Tata McGraw Hill Education PLT. p. 368 (2010). ISBN 978-007-126767-0
14. Sivanandam, S.N., Sumathi, S., Deepa S.N.: Introduction to neural networks using MATLAB 6.0. Tata McGraw Hill Education PLT. p. 11 (2011). ISBN-10: 0-07-059112-1

Chapter 62

A Framework for Sharing Communication Media in Supporting Creative Task in Collaborative Workspace

Norzilah Musa, Siti Z.Z. Abidin and Nasiroh Omar

Abstract In networked collaborative working environment, users are involved with either repeatable or exclusive task. These variety tasks required the collaborative system to support it dynamic and creative solutions especially in sharing the communication media. Most of the related research are objectively focusing on the alignment of user behavior in shared activities. There is not much research conducted concerning on sharing communication media components in creative collaborative activity in details. In this paper, we analyze collaborative work application systems focusing on the communication module in four common areas; business, health, education and manufacturing. Based on the analysis, communication media sharing elements are formed into three layers. In each layer, media sharing element and its functions are identified. Then, we proposed a general software framework for communication media sharing for creative collaborative activities. The framework is functioning as a platform for users to manage their dynamic creative solutions to accomplish complex task.

62.1 Introduction

Globalization and changed in market trends emerged from the evolving of internet and communication technology. These trends contribute to the creating of new organizational structure and business sector to expand in remote areas. Currently,

N. Musa (✉) · S.Z.Z. Abidin · N. Omar
Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia
e-mail: norzi105@salam.uitm.edu.my

S.Z.Z. Abidin
e-mail: sitizaleha533@salam.uitm.edu.my

N. Omar
e-mail: nasiroh@tmsk.uitm.edu.my

workers in Collaborative Working Environments (CWE) are composed of dynamic assemble groups from diverse professional skills that work together within shared Collaborative Working Environments (CWE) [1]. This new structure makes possible for dispersed experts matters and workers be connected to collaborate and accomplish business projects [2]. Hence, collaborative work breaks through the common concept of computer application by providing collective cooperative workspace for users.

Collaborative workspaces are also designed for unpredictable problems that need current workflow systems adapt to the changing environment [3]. In this situation, users need a system that can support their creative solution activities such as ad hoc communications and allow them to react accordingly. This kind of activities needs the system efficiently control the services between sustaining the space for user creativity and coordinating the resources [4]. Moreover, the content of resources involved in such activities might be changed. Hence, provisioning of new knowledge and data transformation need to be coordinated efficiently. Furthermore, dispersed users in such collaborative activities are not sharing the same physical work environment. They are outside of sensory range of each other. Thus, communication technology plays a major role in supporting the information exchange and interaction between remote group of users by using available communication media (text, video, voice and etc.).

Since the communication is a central activity in any collaborative work, issues in communication became one of the ongoing challenges encounter by the group of users [5]. Moreover, most of the problems highlighted by researchers are associated to project-related task coordination, such as workflow management, access control and conflict management [6, 7]. Moreover, the solutions proposed by the researchers are to adhere the need of user behavior alignment on shared activities and resources. In this context, communication media became utility tools with the predefined functions and fixed parameters. Hence, it is important for the systems to provide users with flexibilities to facilitate the sharing of communication media to support creative activities.

62.2 Collaborative Workspace System Architecture

In designing the high-level system application architecture, most of the collaborative application systems embraced the layered design method. The ideology of this process is adopted from the network reference model; The Open Systems Interconnection (OSI) Model [8]. This model consists of seven layers. From the top to bottom are; application, presentation, session, transport, network, data-link and physical layers. Each layer has its own functions that categorized by a set of standard protocols. It receives information from the above layer and pass it to the next layer based on their hierarchy.

62.2.1 Classification of Layered Architecture Reference Model

There is three reference model that used by the researcher in designing the high-level collaborative application system architecture; *client/server model*, *N-level of distributed computing* and *web based application model* [9]. These reference models are the backbone of any communication between networked computers; mainly in collaborative work application systems. The approach in designing collaborative frameworks that adopt these reference models can be classified into three categories: *service composition oriented*, *rule-based* [10] and *network technology driven*.

Service Composition Oriented This approach utilizes services provided by the web services. The system allows multi-user to access central information and communicate each other through web services. The communication media between users are through chat tools, whiteboards, audio conferences and webcams. The web services language, *Service Oriented Architecture Protocol (SOAP)* and *Web Services Description Language (WSDL)* make the configuration of the communication capabilities more flexible. In order to support variety and rigid collaborative activities, services such as service abstraction, discovery and selection, placement and binding services [11] based on *Service Oriented Architecture (SOA)* is applied. Most of these services are associated in the middleware of the framework [12, 13]. However, these useful services have to comply with the well-defined rules in a specific domain problem.

Rule-based This approach defines rules to control the behavior of services, users, resources and application in a collaborative environment [10]. This technique allows the system environment to provide flexible features for configuration, operation and creation of supporting services. In the context of communication media, synchronization rules are enforced to maintain the harmony of the collaborative environment.

Network Technology Driven This approach deals with the components and logical entities of the network. It utilizes the advantages features and services provided by the technology. The approach is suitable for problem with well-defined problem scope and knowledge about tools supported by the network. This enables developers to reuse current techniques for dissimilar problems. Activity such as interaction among users is supported by the pre-defined and established communication techniques of the chosen technology.

In summary, most of the reviewed architectures and framework of collaborative working application systems have four common elements in their communication module. The four elements are user interface, collaborative activities, software design method and network technology. Hence, based on the layered architecture design method, Fig. 62.1 shows the general view of communication layers in the collaborative system. The communication is the main element in creating collaborative culture. Its allow user to create, update, change and delete shared artefact. As users are not sharing the same physical work, exchanging and sharing resources and information are infeasible without sharing the communication media.

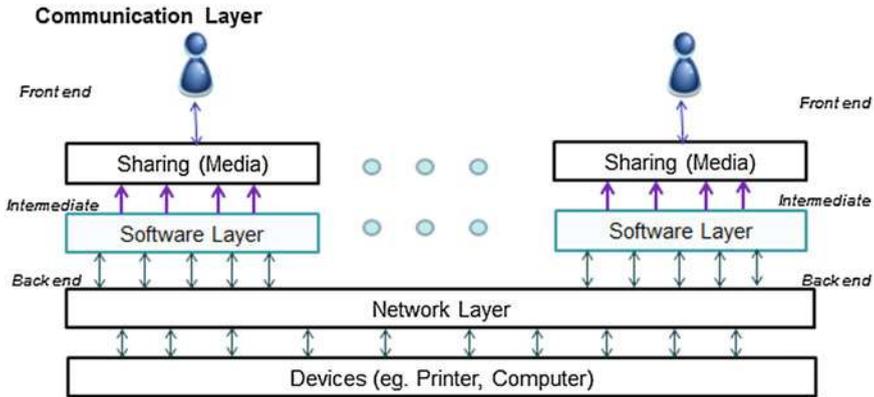


Fig. 62.1 Summarization of communication layers [14]

In the next section, sharing communication media in three contextual domains will be explore in order to define the sharing components.

62.3 Communication Media Sharing Entities

There are three uses of communication in collaborative work activities [1]. First, to assists groups of users in integrating and coordinating their work activities by deliberating their past, current and future work activities. Secondly, serves as a platform for user to share and exchange information. Finally, it is one of the tools to simulate and encourage new knowledge among the users.

Communication is not only about interaction among a group of users. It is also include several types of objects that need to be shared through various platforms and user interfaces. Hence, user interaction and object sharing in a networked collaborative working system are analyzed. Based on the four common areas (business, education, health and manufacturing), seven components involve in communication in collaborative work activities have been studied. The components are *user*, *communication channel*, *resources*, *software component*, *work activity*, *coordination method* and *shared services*. Table 62.1 shows the sharing communication media components in the existing system.

Each domain has its own approach of disseminating information to its end-users. This dislocated and dispersed users establish various kinds of communication relationships during collaborative activities. The relationship can be one-to-one, one-to-many and many-to-many [14] that can determine the level of communication between the users. Furthermore, user's profile; such as role, privacy setting and own objects to share are important features in coordinating rules on handling the sharing process.

Table 62.1 Sharing communication media components in selected domains

Domain	Comm channel	Type of interaction	Co ordination	Shared resources	Shared services	Context profile	System component
Business [10], (Belen Pelegrina et al. 2010; Cheaitb et al. 2009)	Audio, video, chat, email	One to many, many to many, one to one	Event-based mechanism, agent-interaction protocol (agreed-on, request—response), rule-based	Documents, multimedia contents	Paint	Roles, actions (context awareness), coordinate	Mediator—service control, composable service, reusable and extended component
Education (El Saddik et al. 2008; Bijlani et al. 2011)	Whiteboard, chat, audio, video, message	One to many, many to many, one to one	JXTA protocol suit, distributed client- server module	Web browser, telepointer, video, documents, graphics	Web browser (IE, Mozilla)	Role, name, access level	JXTA architecture, adaptive bitrate streaming methodology
Health (Ciampi et al. 2010; Dube et al. 2005)	Text, video	One to one	Rule paradigm, interaction protocol	Multimedia objects, patients information, imaging	Nil	Name, medical record	Triggers mechanism, agent-based infrastructure
Manufacturing (Sadeghi et al. 2010), (Mourtzis 2010)	Text, chat, email	One to one, one to many	Product Process Organisation Module and meta rules, Model-view- controller pattern	Designer drawing model, Planning Information, company schedule	Nil	Name, design copies, company-name, production status, products availability	API, Web based technology

The medium of communication in any work activities will determine the scope of the task supported by the system. The user interface and sharing method would become more complex when the activity used various type of communication media. Most of the systems pre-defined their communication media into standard operation. The credibility of the coordination techniques affects by this condition. In order to collaborate effectively, the systems should dynamically adapt changes in task-related activity environment. This is due to the nature of works in certain domains that required changes of communication media during the interaction. Furthermore, not all work activities in the domains are repeatable workflow (create, edit, delete, etc.). Some activities need user creativity in using the communication media to solve ad hoc and important problems such as exception handling problem. This can be well supported if appropriate communication media can match with the user requirements dynamically.

Most of the work activities in the studied domains are supported by various kinds of shared resources. The resources can be report document, slides presentation, proposal, architectural design, image, audio or video files. Moreover, some of the objects are associated with special shared features such as special access control that tagged with selected users. The features are set by different methods that tailored to the solutions strategy. In fact, the success of the sharing process is depending on the interaction and synchronization of user, resources and media communication elements.

Different organizations will have different objectives and activities with different input devices and network communication model. Therefore, in dealing with this heterogeneous communication architecture; software component act as an intermediary between the user and network devices play a vital role in giving seamless communication in the collaboration process. Application program interface (API), embedded software components, special purpose program or custom made, and toolkits are common software components that used by shared media applications. This component helps developers to create tools to ease users in their tasks especially group formation along with sharing activities that include managing access and security.

In summary, the formation of communication media sharing is made possible by five related entities; *context profile*, *sharing method*, *algorithm strategy*, *interaction protocol* and *shared objects*. These entities will embed in the software framework of the systems, particularly in the communication module. General software based frameworks for collaborative work system focusing on the communication module is depicted in Fig. 62.2. The framework was mapped with common communication structure as shown in Fig. 62.1 which divide the structure into three layers: *front-end*, *intermediate* and *back-end*.

The *front-end layer* act as an interface for user to interact with the systems; and a place where communication relationship establishes among the group of users. It also provides an information about the user profile, communication media and work activity. In contrary, the *intermediate layer* is a place where the coordination method, interaction protocol and shared services work together to create the seamless sharing communication process. It is the place where the request of the

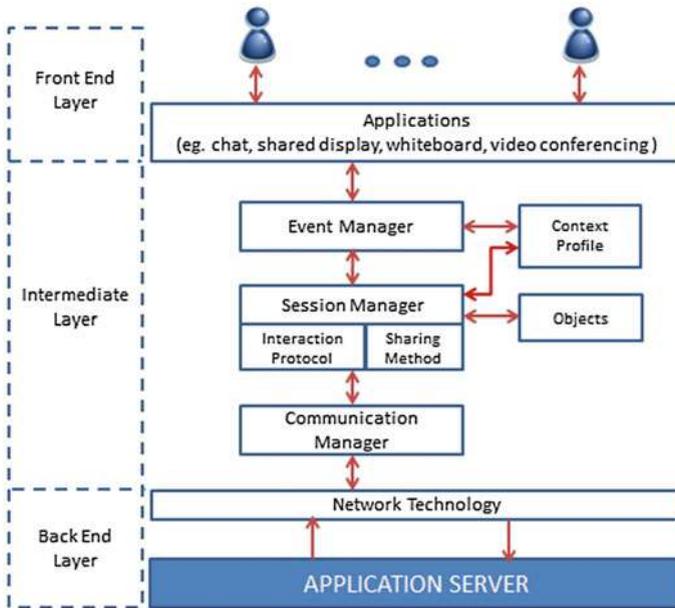


Fig. 62.2 General communication framework in collaborative work

user or systems are entertained and served harmoniously by the system. Then, the *back-end layer* is the place where the communication channel transforms the data technically to be transported to the assigned users.

62.4 Conceptual Framework

62.4.1 Dynamic and Creative Collaborative Activities

In any software development, developers programmed solutions to all foresee exception that could happen during the process. This is where the testing phase is crucial in software development methodology. All the possible errors are captured and user requirements are hard coded into the system after the user acceptance test phase is done. In collaborative work, it is crucial for the systems to provide a space for creative and knowledge intense processes like online meeting and report documentation [15]. The work activities always involve ad hoc and short-based unique project that needs a different approach which cannot be mapped with a repeatable workflow solution. All the dynamic changes required during the runtime should be supported and adopt by the system. Therefore, the users should allow to communicate and shared data in creative manners to support various kinds of activities by the groups, subgroups or individual in the collaborative work environments.

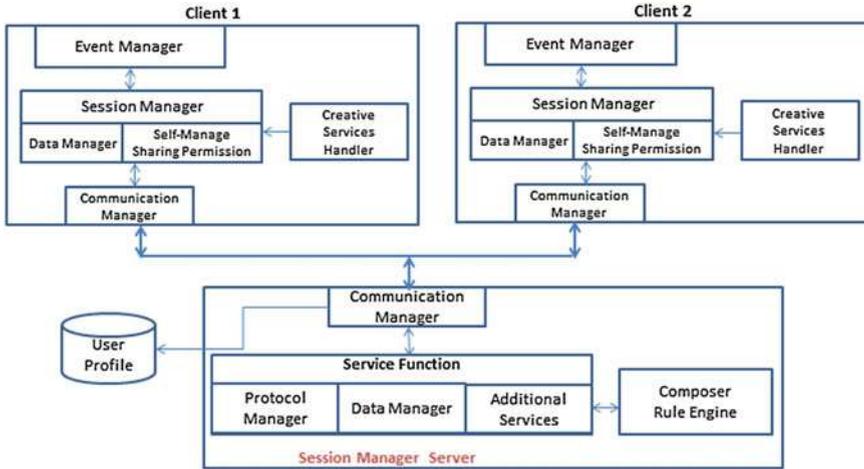


Fig. 62.3 Communication media sharing framework for creative collaborative activity

62.4.2 Proposed Framework

The proposed framework (Fig. 62.3) was conformed to the general communication framework for collaborative work structure (Fig. 62.2) for creative activities with media sharing elements. Each user (client) has three logical layers: front-end layer, immediate layer and back-end layer that adapts to client/server reference model for the layered architecture. The front-end layer host the collaborative applications communication media such as chat, shared browser and whiteboard. It provides awareness and inform intercepting events (local and remote) to the next layer. It sends messages to the next layer once any work activity is initiated by the users.

The *intermediate layer* is the cornerstone of any collaborative work application, as it is a central of many important functions such as controlling consistency among users, detecting and overcome any arise conflict and controlling information exchange. The *event manager* is in charge of intercepting events including the objects and users involved. It passed the information related to the event to the session module to initiate and administer the session.

A *session manager* handles sessions, which represent shared workspace for users to collaborate. It responsible to create, maintain and terminate of any sessions among users. It maintains the user context profile and identified the session owner or moderator. The moderator invoked appropriate tools to coordinate activities and users in the session. All used objects in the session are handled by the data manager. The creative trigger mechanism will invoke special rules to allow users override administrator-defined policy especially regarding sharing policy. The *communication manager* will receive requests and information about the activities and transport it to the server for permission and further process.

Since the proposed framework is based on the client/server reference model; the server is acting like a central processor while clients are dealing with the users request. Hence, components in the server are mirrored for clients intermediate layer components except the event manager that only have in the clients. It contains communication manager that responsible to receive any request from clients and in charge of the application's context profile. All messages received will be translated and pass to server session manager to take further action. The session manager will control and facilitate all services related to the activities include the shared objects, protocol and invoke composer engine to cater any creative collaborative activities. The composer engine will process user self-manager sharing permission policy and prepare a platform for users to manage the shared media and objects flexibly.

62.5 Conclusion

In this paper, we give detailed explanations about the high-level design architecture for collaborative work application systems that embrace the layered-design method which adopted from the OSI model. We also discussed the three approaches used in designing communication module in the system framework: service composition oriented, rule-based and network technology. Three main domains have been selected based on their active researcher contribution in this area. Based on these domains, seven important components have been identified: user, communication channel, resources, software component, work activity, coordination method and shared services. We also explain the importance of handling creative activity in the workspace and highlight the issues when dealing with the scattered users. Currently, most research proposed new approaches and techniques in their collaborative working environment. However, not much research interested in detail on communication media sharing components in creative collaborative activity. Hence, based on the analysis, we proposed a framework which represents the communication media sharing activity in the collaborative work. The software framework is functioning as a platform for users to dynamically manage the creative work activities. We left the media sharing compose engine functionality attributes and its constraints as one of future work for this research.

Acknowledgments The authors would like to thank Universiti Teknologi MARA and Ministry of Education, Malaysia for the financial support.

References

1. Pallot, M., Bergmann, U., Kühnle, H., Pawar, K.S., Riedel, J.C.K.H.: Collaborative working environments: distance factors affecting collaboration. In: Proceedings of the 16th International Conference on Concurrent Enterprising, ICE'2010 (2010)
2. Xue, X., Shen, Q., Fan, H., Li, H., Fan, S.: IT supported collaborative work in A/E/C projects: A ten-year review. *Autom. Constr.* **21**, 1–9 (2012)
3. Kammer P.J., Taylor, R.N.: Techniques for supporting dynamic and adaptive workflow. pp. 1–19 (2002)
4. Schuster, N., Zirpins, C., Tai, S., Battle, S., Heuer, N.: A service-oriented approach to document-centric situational collaboration processes. In: 2009 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, pp. 221–226 (2009)
5. Bassanino, M., Fernando, T., Wu, K.-C.: Can virtual workspaces enhance team communication and collaboration in design review meetings? *Architect. Eng. Des. Manag.* (May 2013) pp. 1–18, Mar 2013
6. Sun, D., Xia, S., Sun, C., Chen, D.: Operational transformation for collaborative word processing. In: Proceedings of the 2004 ACM conference on Computer supported cooperative work, vol. 6, no. 3, pp. 437–446 (2004)
7. Jeffery, C., Dabholkar, A., Tachtevrenidis, K., Kim, Y.: A framework for prototyping collaborative virtual environments. In: *Groupware: Design, Implementation, and Use*, pp. 17–32 (2005)
8. Bielstein, B.: Computer networks and protocol—the OSI reference model. http://nsgn.net/osi_reference_model/the_osi_reference_model.htm (2006). Accessed 16 Jul 2013
9. Sun, Q., Qiu, Y., Ma, W., Gu, Y.: Collaborative development of network application system based on layered conceptual model. In: 2010 2nd International Work Database Technology Applications, pp. 1–4, Nov 2010
10. Han, S.W., Kim, J.: A service composition oriented framework for configuring SMeet multiparty collaboration environments. *Multimedia Tools Appl.* (2012)
11. Jerstad, I., Dustdar, S., Thanh, D.V.: A service oriented architecture framework for collaborative service. In: 14th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise 2005, pp. 121–125 (2005)
12. Galatopoulos, D.G., Kalofonos, D.N., Manolagos, E.S.: A P2P SOA enabling group collaboration through service composition. In: Proceedings of 5th International Conference on Pervasive Services—*ICPS '08*, p. 111 (2008)
13. Kim, N., Kim, J., Uram, T.: A hybrid multicast connectivity solution for multi-party collaborative environments. *Multimedia Tools Appl.* **44**(1), 17–37 (2009)
14. Musa, N., Abidin, S.Z.Z., Omar, N.: Towards flexible media sharing: control and coordination issues in network collaborative virtual environment. In: 2012 IEEE Colloquium on Humanities, Science and Engineering, no. Chuser, pp. 154–158, Dec 2012
15. Schuster, N., Zirpins, C., Scholten, U.: How to balance flexibility and coordination? Service-oriented model and architecture for document-based collaboration on the Web. In: IEEE International Conference on Service-Oriented Computing and Applications (SOCA), pp. 1–9 (2011)

Chapter 63

Joint Torque Estimation Model of sEMG Signal for Arm Rehabilitation Device Using Artificial Neural Network Techniques

M.H. Jali, T.A. Izzuddin, Z.H. Bohari, H. Sarkawi, M.F. Sulaima, M.F. Baharom and W.M. Bukhari

Abstract Rehabilitation device is used as an exoskeleton for peoples who had failure of their limb. Arm rehabilitation device may help the rehab program to whom suffered with arm disability. The device is used to facilitate the tasks of the program and minimize the mental effort of the user. Electromyography (EMG) is the techniques to analyze the presence of electrical activity in musculoskeletal systems. The electrical activity in muscles of disable person is failed to contract the muscle for movements. To minimize the used of mental forced for disable patients, the rehabilitation device can be utilize by analyzing the surface EMG signal of normal people that can be implemented to the device. The objective of this work is to model the muscle EMG signal to torque for a motor control of the arm rehabilitation device using Artificial Neural Network (ANN) technique.

M.H. Jali (✉) · T.A. Izzuddin · Z.H. Bohari · M.F. Sulaima · M.F. Baharom · W.M. Bukhari
Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Malacca, Malaysia
e-mail: mohd.hafiz@utem.edu.my

T.A. Izzuddin
e-mail: tarmizi@utem.edu.my

Z.H. Bohari
e-mail: zulhasrizal@utem.edu.my

M.F. Sulaima
e-mail: fani@utem.edu.my

M.F. Baharom
e-mail: mohamad.faizal@utem.edu.my

W.M. Bukhari
e-mail: bukhari@utem.edu.my

H. Sarkawi
Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka,
Malacca, Malaysia
e-mail: hafez@utem.edu.my

The EMG signal is collected from Biceps Brachii muscles to estimate the elbow joint torque. A two layer feed-forward network is trained using Back Propagation Neural Network (BPNN) to model the EMG signal to torque value. The performance result of the network is measured based on the Mean Squared Error (MSE) of the training data and Regression (R) between the target outputs and the network outputs. The experimental results show that ANN can well represent EMG-torque relationship for arm rehabilitation device control.

63.1 Introduction

Human support system is endoskeleton. Endoskeleton plays a role as a framework of the body which is bone. Our daily movements are fully depends on the functionality of our complex systems in the body. The disability one or more of the systems in our body will reduce our physical movements. The assistive device is a need for rehab as an exoskeleton. The functionality of the rehabilitation device has to smooth as the physical movement of normal human.

The rehabilitation programs provide the suitable program for conducting the nerve and stimulate the muscles. People who have temporary physical disability have the chances to recover. Nowadays, rehabilitation program are using exoskeleton device in their tasks. The functionality of exoskeleton depends on muscle contraction. Electromyogram studies help to facilitate the effectiveness of the rehabilitation device by analysing the signal transmitted from the muscle.

The rehabilitation device is a tool that used to help the movements for daily life activities of the patients who suffer from the failure of muscle contractions, due to the failure of the muscles contractions the movements is limited. The ability of the patients to do the tasks in the rehabilitation programs need to be measured. The rehabilitation programs have to assure whether the tasks will cause effective or bring harm to the patients [1].

Historically, the rehabilitation tasks have been avoided due to a belief that it would increase spasticity [2]. In this research, the analysis of the data will be focusing on upper limb muscles contraction consisting of biceps muscles only. The experiment is limited to the certain of upper limb movements that use in training. EMG is a division of bio signal; the bio signal analysis is the most complex analysis. Thus, the signal analysis is a complicated process that has to be through many phases of analysis [3].

EMG signal function as a control signal for the arm rehabilitation device. A system needs a model to estimate relationship between EMG and torque [4]. EMG signal based control could increase the social acceptance of the disabled and aged people by improving their quality of life. The joint torque is estimated from EMG signals using Artificial Neural Network [5, 6]. The Back Propagation Neural Network (BPNN) is used to find a solution for EMG-joint torque mapping. The EMG signal of the biceps brachii muscle act as the input of the ANN model whiles

the desired torque act as the ideal output of the model. Hence the EMG signals considered the 'intent' of the system while the joint torque is the 'controlled' variable for the arm rehabilitation device [7]. The network is evaluated based on the best linear regression between the actual joint torque and the estimated joint torque [4]. The experiment results shows that the model can well represent the relationship between EMG signals and elbow joint torque by producing MSE of 0.13807 and average regression of 0.999.

This paper is organized as follows. Section 63.1 explains brief introduction about this research work. Section 63.2 describes all the related works of this study. Section 63.3 demonstrates the method implemented for this work that covered the experimental setup, EMG data processing, desired torque determination and ANN technique. Section 63.4 presents the experimental results as well as the discussion of each result obtained. Finally the paper ends with conclusion and recommendations.

63.2 Related Works

There have been several studies that have applied ANN for modeling the muscle activity to joint relationship. Kent et al. [7] proposed the ANN model to measure the ankle EMG-joint torque relationship at a full range of torque under isometric, supine conditions by inserting EMG signal from 6 muscle sites into the model as in the input, while the measured torque is entered into the model as the ideal output. The learning process occur approximately 16,000 iterations resulting error that is less than 6 %.

More recent study has been conducted by Li et al. [4] to predict the elbow joint angle based on EMG signals using ANN. The three layer BPNN was constructed by using the RMS of the raw EMG signal from biceps and triceps. The result from 40 group EMG signals when subjects do bowing and extending elbow joint action reveal that the prediction output from the trained network was very close with the target angle. According to Morita et al. [6], it is quite difficult to know the elbow shoulder joint torque on the natural condition. The learning method which is based on the feedback error learning schema is proposed by modifying the ANN with the torque error which calculated from the desired angle and measured angle.

Studies of muscle force models have been carried out by Naeem et al. [8]. The model was estimated based on a rectified smoothed EMG signal using the BPANN method to predict the muscle force. The proposed model can efficiently extract muscle force features from (EMG) signals in a fast and easy method. The results showed that the regression of the ANN model exceeded 99 %. However among all the previous studies related to EMG based ANN model been conducted, none of them really emphasize on the data processing phase which involve EMG data processing and desired torque determination. Therefore this paper will explain in detail this phase.

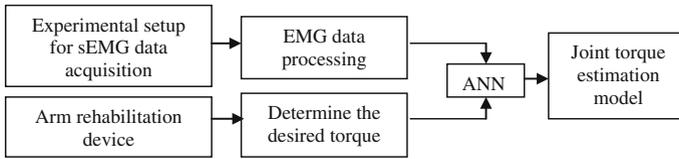


Fig. 63.1 Research methods

63.3 Methods

Figure 63.1 shows a block diagram of our research methodology that consists of two major phases. First phase is EMG data processing and desired torque determination. Second phase is the ANN construction and testing. The data collection from the first phase is used to validate and teach the ANN algorithm in second phase.

63.3.1 Experimental Setup

Implementation of arm rehabilitation device based on movement is recorded from the EMG signal of healthy subjects. From the human anatomy studies, different angle movements of upper limb with elbow as the reference is depends on relation of agonist and antagonist. In this study is focusing on the behaviour of biceps muscle as agonist and the triceps as the antagonist respectively. Muscle that involved in this movement is biceps and triceps, however in this study to understand the electrical activity during muscle contraction, the biceps is the only muscle that taking into account. The movements' ranges in between position of arm flexion until arm fully extend.

The environment is in a room with low lighting especially the fluorescent light, any electromagnetic devices is away from the experiment equipment and the environment is in silent room. Then, the experimental is set up with the subject sit on the chair while the hand is on the table. The subject has to complete the task of lift up the dumbbell with 2.268 kg of weight as shown in Fig. 63.2b. Normally, the appearance of EMG signal is chaos and noisy depends on the type of electrodes also the noise factor. To simplify the difference of amplitude response for the motion, the dumbbell is functioned to amplify the amplitude in analysing the electrical activity during rest and contract. The rehabilitation devices (white in color on Fig. 63.2a) helps to keep the position of the elbow joint and the wrist joint in line Mostly, the EMG signal is obtained after several trials of the movements. These movements are specified from angle of 0° (arm in rest position), up to 120° (arm is fully flexion). Data was collected from two subjects by 5 repetitions of each flexion movements [9].

Prior of data collection process, the skin needs a preparation. The preparation of skin is ruled by the Surface Electromyography for the Non-Invasive Assessment of

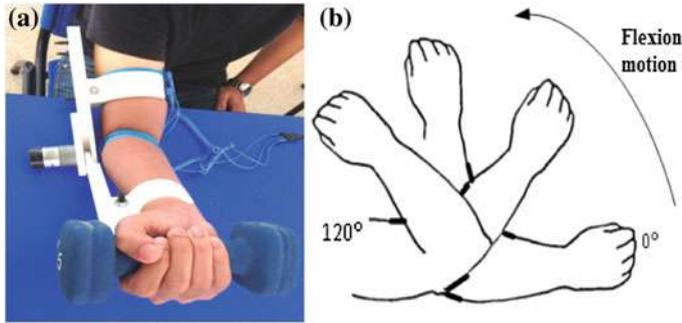
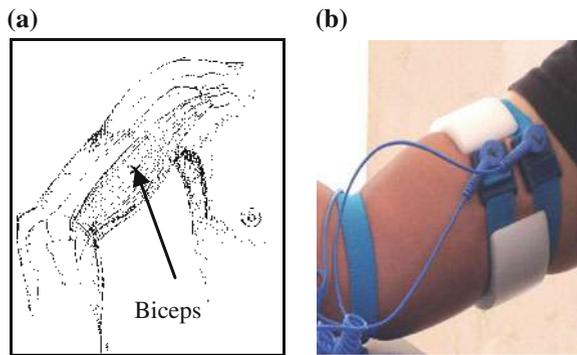


Fig. 63.2 Subject is set-up with arm rehabilitation assistive device for experiment (a). Simulation of subject's to lift up the dumbbell 2.268 kg of weight (b)

Fig. 63.3 The biceps brachii muscles for electrode positions (a), The electrode placements on subject skin (b)



Muscles (SENIAM) procedure for non-invasive methods. The subject's skin has to be shaved by using small electrical shaver and cleaned with sterile alcohol swabs saturated with 70 % Isopropyl Alcohol. This step is to be taken for minimizing the noise and to have a good contact with the electrodes of the skin by decreasing the impedance of the skin. The skin has to be clean from any contamination of body oil, body salt, hair and the dead cells. The preparation of skin can be done by wiping the alcohol swab into the area of skin that electrode placement to be applied. The placements of the electrode have to be at the belly of the muscles not in the tendon or motor unit. This ensured the detecting surface intersects most of the same muscle on subject as in Fig. 63.3a at the biceps brachii, and as a result, an improved superimposed signal is observed. Reference electrode has to be at the bone as the ground, for this experiment it placed at elbow joint as shown in Fig. 63.3b. These electrodes are connected to the combination of hardware Olimex EKG-EMG-PA and Arduino Mega for data collection.

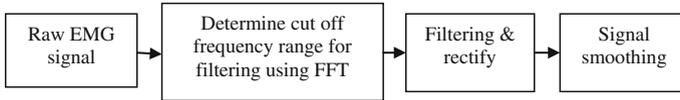


Fig. 63.4 EMG data processing

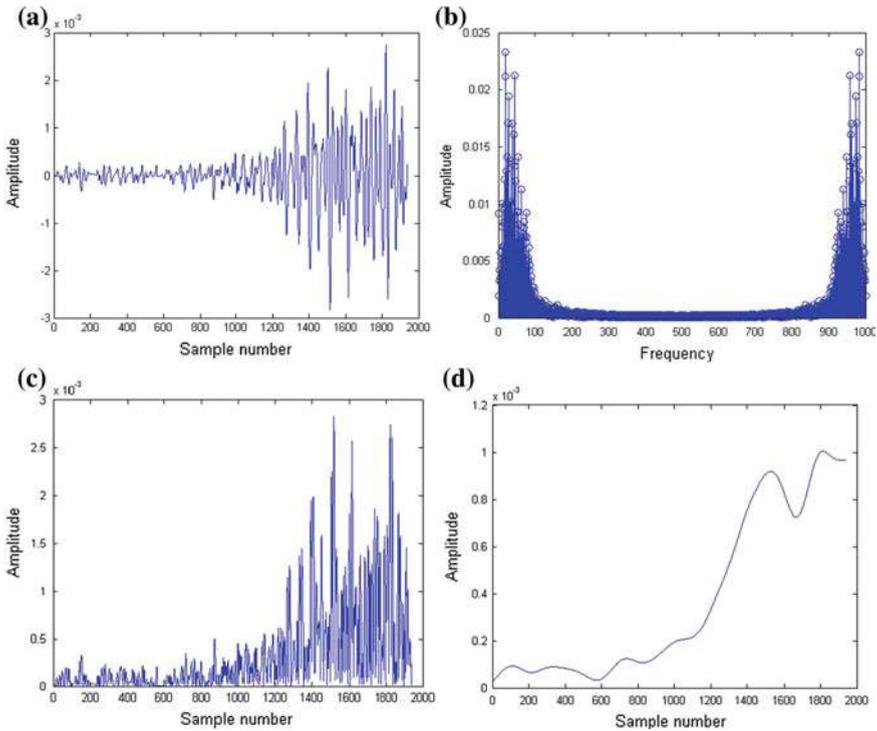


Fig. 63.5 Raw EMG signal (a), Power Spectral Density (b), Rectified signal (c), Smooth signal (d)

63.3.2 EMG Data Processing

Figure 63.4 shows the EMG data processing block diagram. After obtained satisfactory EMG signal as shown in Fig. 63.5a, Fast Fourier Transform (FFT) is performed to the signal to analyses the frequency content of the signal. The EMG signal is break into its frequency component and it is presented as function of probability of their occurrence. In order to observe the variation of signal in different frequency components, the FFT signal is represented by Power Spectral Density (PSD). From the PSD we can describes how the signal energy or power is distributed across frequency. Figure 63.5b shows that most of the power is in the range of below 10 Hz, therefore the EMG signal should be filtered in the range of above 10 Hz as a

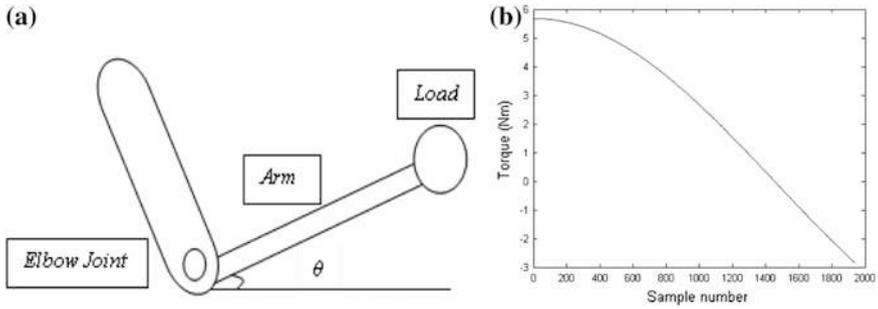


Fig. 63.6 Arm rehabilitation device position (a), Desired torque characteristic (b)

cut off frequency for low pass filtered. After decide the cut off frequency for filtering, the DC offset of the EMG signal is removed and is rectified to obtain its absolute value as shown in Fig. 63.5c. Finally, the signal was smoothed and normalized passing it through a 5th order Butterworth type low-pass filter with cut off frequency 10 Hz and the smooth signal is illustrate in Fig. 63.5d [8, 9].

63.3.3 Desired Torque

Desired torque of the elbow joint is used as target data for our ANN techniques as well as act as output signal for muscle [10]. The data is collected throughout the angle from 0° to 120° angle with increment of 0.0619° each step to align with the sample number of EMG signal. Figure 63.6a shows the arm rehabilitation device position for torque calculation. The desired torque for elbow joint is determined by applying standard torque equation:

$$\tau = (r_{load}F_{load} + r_{arm}F_{arm})\cos \theta \tag{63.1}$$

where r_{load} is distance from the elbow joint to the load, F_{load} is force due to load, F_{arm} is the force due to the mass of the lever arm and r_{arm} is evenly distributed distance of mass of arm distance which is half of r_{load} . The angle θ between r and F is drawn from the same origin. A applied load is 2.268 kg dumbbell and the distance from the elbow joint is 0.25 m while the mass of the lever arm is 0.1 kg. Figure 63.6b shows the desired torque characteristic for the elbow joint.

63.3.4 Artificial Neural Network Techniques

ANN is a computing paradigm that is loosely modeled after cortical structures of the brain. It consists of interconnected processing elements called nodes or neuron

that work together to produce an output function. It capable to map a data set of numeric inputs with a set of numeric outputs. It is also the most widely applied training network which has input layer, hidden layer and output layer. The neurons on each layer need to be considered carefully to produce high accuracy network. The number of hidden neurons could affect the performance of the network. The network performance not always been improved if the hidden layer and its neurons is increased [4]. Therefore the number of hidden neurons is tested to achieve the optimized network. However there is constraint in determining the number of neurons. If the numbers of hidden neurons is too large, the network requires more memory and the network become more complicated while if the number of hidden neurons is too small, the network would face difficulty to adjust the weigh properly and could cause over fitting which is problem where the network cannot be generalized with slightly different inputs [11].

The input signal is propagated forward through network layer using back propagation algorithm. An array of predetermined input is compared with the desired output response to compute the value of error function. This error is propagated back through the network in opposite direction of synaptic connections. This will adjust the synaptic weight so that the actual response value of the network moved closer to the desired response [12]. BPNN has two-layer feed-forward network with hidden neurons and linear output neurons. The function used in the hidden layer of network is sigmoid function that generates values in range of -1 to 1 [9]. There are layers of hidden processing units in between the input and output neurons. For each epoch of data presented to the neural network, the weights (connections between the neurons) and biases are updated in the connections to the output, and the learned error between the predicted and expected output, the deltas, is propagated back through the network [13].

A Lavenberg-Marquardt training back propagation algorithm is implemented for this work to model the EMG to torque signal. This algorithm is chosen due to the facts that it outperforms simple gradient descent and other conjugate gradient methods in a wide variety problem such as convergence problems. Single input EMG signal that act as a training data while single output which is the desired torque act as a target data. The network is optimized for 20 hidden neurons as shown in Fig. 63.7 [4]. The network was trained using 1839 sets of EMG data for arm flexion motion from 0° angles to 120° angle. It also has output data which is torque of correspondent arm motion. The training process was iteratively adjusted to minimize the error and increased the rate of network performance [11]. The training has been done by dividing the input data of 70 % for training, 15 % for validation and 15 % for testing [8].

The performance evaluation of the network is based on the Mean Squared Error (MSE) of the training data and Regression (R) between the target outputs and the network outputs as well as the characteristics of the training, validation, and testing errors. The network is consider has the best performance if it has lowest MSE and highest R while exhibit similar error characteristics among the training, validation and testing. However even if the MSE shows very good result but the validation and testing vary greatly during the training process, the network

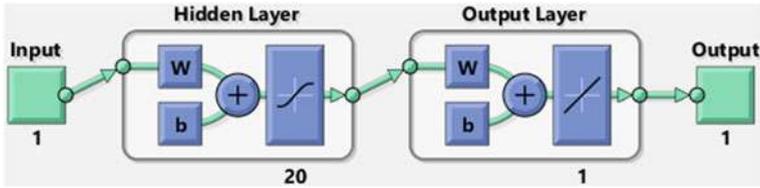
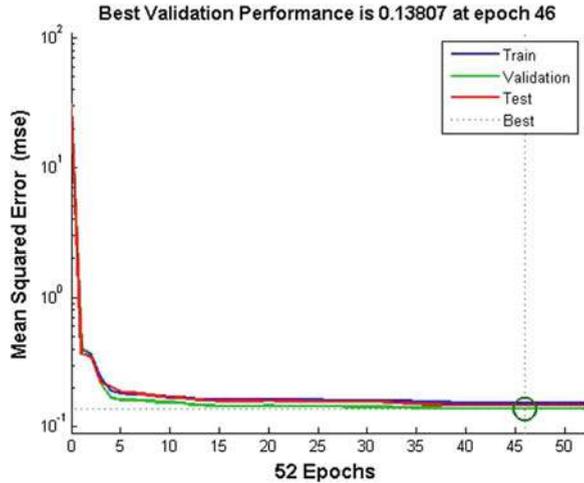


Fig. 63.7 BPNN model

Fig. 63.8 Best validation performance



structure is still considered unsatisfactory because the network is not generalized. Therefore further tuning and training need to be conducted in order to improve the network performance [11].

63.4 Experimental Results

In order to optimize the network performance, different number of hidden neurons is simulated for several times until achieved the satisfactory results [11]. The network is trained using Lavemberg-Marquardt algorithm and the performance of the network is measured using MSE and R. The best validation performance of the network is 0.13807 at epoch 46 as shown in Fig. 63.8. It is an acceptable result since the test set error and the validations set error have similar behavior. It shows that the MSE has decreased rapidly along the epochs during training. The regression also produces a good curve fitness for training, test and validation data around 0.999 which give an optimal value for our model as shown in Fig. 63.9 [8, 14].

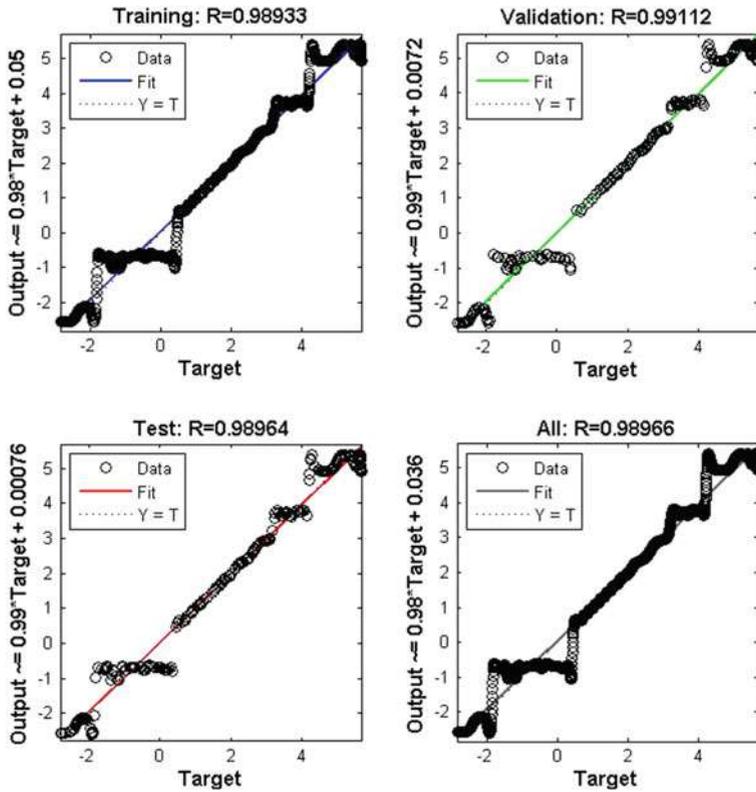


Fig. 63.9 Regression of the trained model

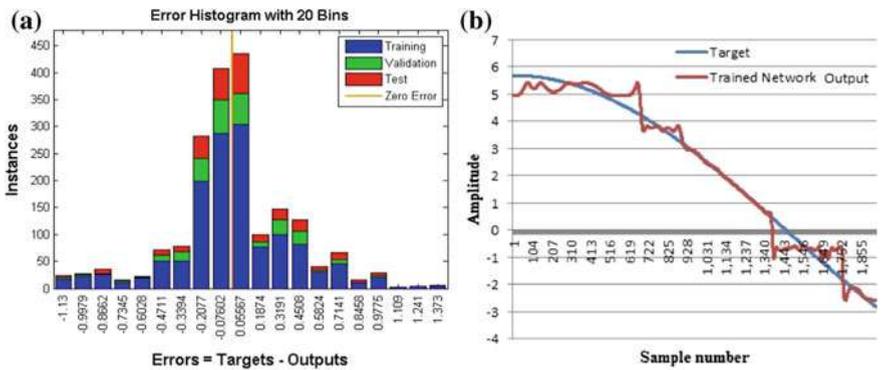


Fig. 63.10 Error histogram of the trained model (a), Target versus trained network output (b)

Error sizes are well distributed since most error approaching zero values that make the trained model perform better as shown in Fig. 63.10a. The network has maximum instance around 450 of MSE distributed around the zero line of the error histogram [14]. Figure 63.10b shows a comparison between prediction output from the trained network and target torque. The prediction output has fairly good agreement with the characteristics of the target data.

63.5 Conclusions

Based on the result, it can be concluded that the ANN model with 20 hidden neurons produce MSE of 0.13807 and average regression of 0.999. It is consider a good performance as it shows that this neural network model can well represent the relationship between EMG signals and elbow joint torque. Hence this model can be used for motor torque control of the arm rehabilitation devices. The model can be further improved by applying other artificial intelligence training algorithm such as genetic algorithm and particle swarm optimization to produce better mean squared error and regression performance result.

Acknowledgements The authors would like to thanks Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Education, Malaysia for the financial supports given through Research Grant.

References

1. Louise Ada, S.D.C.G.C.: Strengthening interventions increase strength and improve activity after stroke: a systematic review. *Aust. J. Physiotherapy* **52**, 241–248 (2006)
2. Adult Hemiplegia, B.B.: Evaluation and treatment. Butterworth-Heinemann, Oxford (1990)
3. Muthuswamy, J.: Biomedical Signal Analysis in Standard Handbook of Biomedical Engineering and Design, pp. 18.1–18.30. McGraw-Hill, New York (2004)
4. Li, D., Zhang, Y.: Artificial neural network prediction of angle based on surface electromyography. In: International Conference on Control, Automation and Systems Engineering (CASE), pp. 1–3 (2011)
5. Reaz, M.B.I., Hussain, M.S., Mohd-Yasin, F.: Techniques of EMG signal analysis: detection, processing, classification and applications. *Biol. Proced. Online* **8**, 11–35 (2006)
6. Morita, S., Kondo, T., Ito, K.: Estimation of forearm movement from emg signal and application to prosthetic hand control. In: IEEE International Conference on Robotics and Automation (ICRA), vol. 4, pp. 3692–36972 (2001)
7. Kent, L.M., Siegler, S., Guez, A., Freedman, W.: Modelling of muscle EMG to torque by the neural network model of back propagation. In: Proceedings of the Twelfth Annual IEEE International Conference of the Engineering in Medicine and Biology Society, pp. 1477–1478 (1990)
8. Naeem, U.J., Abdullah, A.A., Caihua X.: Estimating human arm’s muscle force using artificial neural network. In: Proceedings of IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–6 (2012)

9. Favieiro, G.W., Balbinot, A., Barreto, M.M.G.: decoding arm movements by myoelectric signals and artificial neural networks. In: Conference of Biosignals and Biorobotics (BRC), pp. 1–6 (2011)
10. Jali, M.H., Sulaima, M.F., Izzuddin, T.A., Bukhari, W.M., Baharom, M.F.: Comparative study of EMG based joint torque estimation ANN models for arm rehabilitation device. *Int. J. Appl. Eng. Res (IJAER)* **9**(10), 1289–1301 (2014)
11. Ahsan, M.R., Ibrahimy, M.I., Khalifa, O.O.: EMG motion pattern classification through design and optimization of neural network. In: International Conference on Biomedical Engineering (ICoBE), pp 175–179 (2012)
12. Hudgins, B., Parker, P., Scott, R.: A new strategy for multifunction myoelectric control. *IEEE Trans. Biomed. Eng.* **40**(1), 82–94 (1993)
13. Mars, P., Chen J.R., Nambiar, R.: *Learning Algorithms: Theory and Applications in Signal Processing, Control and Communications*. CRC Press , Boca Raton (1996)
14. Supeni, E.E., Epaarachchi, J.A., Islam, M.M., Lau, K.T.: Development of artificial neural network model in predicting performance of the smart wind turbine blade. In: 3rd Malaysian Postgraduate Conference (MPC 2013), pp. 4–5. Sydney, Australia (2013)

Chapter 64

Enhancement of RSA Key Generation Using Identity

Norhidayah Muhammad, Jasni Mohamad Zain,
M.Y.M. Saman and Mohd Fadhil Ramle

Abstract The purpose of this paper is to enhance previous algorithm called Tripathi algorithm. The Tripathi algorithm proposes an RSA based algorithm to generate cryptographic keys using user identity such as email address of a person. This algorithm used user's identity to replace the numbers that are used as a public key in the RSA algorithm. However, the Tripathi algorithm cannot use all of the users' email addresses as a public key. This is because, there are two reasons why it is unable to use all email addresses: (i) this algorithm use the same modulo value for every email, if the email is not related prime to modulo value, the new email should be entered. (ii) Entered email is composed of odd and even number. If the email is even number, then it cannot be the public key. Therefore the Tripathi algorithm needs to be improved. Proposed algorithm called CLB-RSA has been implemented. This algorithm can used all user emails as a public key, and this achievement is after two experiments are done on this study.

64.1 Introduction

Information security or also known as computer security is an approach to protect information from unauthorized access, disruption, modification or manipulation of information [1]. Sensitive data need to protect from unauthorized by any unrelated

N. Muhammad (✉) · J.M. Zain
University Malaysia Pahang, 26600 Pahang, Malaysia
e-mail: mrs.hidayah@yahoo.com.my

J.M. Zain
e-mail: jasni@ump.edu.my

M.Y.M. Saman
University Malaysia Terengganu, 21300 Terengganu, Malaysia
e-mail: yazid@umt.edu.my

M.F. Ramle
Kolej Komuniti Kuala Terengganu, Terengganu, Malaysia
e-mail: fadhil@kkktu.edu.my

person. As done by [2, 3] to protect medical image. Cryptographic system or cryptosystem is a form of encryption, decryption algorithm, and the key generation [4]. Key generation become the biggest problem for cryptography. This is because the key generation process to determine the public key and private key for use during encryption and decryption process. The safety of a cryptography algorithm depends on the complexity of a cryptography keys. RSA algorithm security is commonly known can be a tool for good security. The actual RSA cryptographic process is generally a complicated mathematical formulation, the more complex of keys, the more difficult to break the cipher text, and more secure, the disadvantage of RSA algorithm is the RSA algorithm using a key consisting of a row of numbers and also requires large storage space and is only suitable for use on large device memory.

Several techniques had already been proposed for distribution of public keys. Out of them one was public key certificates. The basic idea was to use trusted third party called Certificate Authority (CA) to provide trusted public key to the various participants on demand. To setup the hierarchical infrastructure for numerous CA's extra overhead was required. In 1984, Shamir [5] proposed public key encryption scheme in which the public key can be an arbitrary string. Shamir's original motivation for identity based encryption was to simplify the certificate management. Several proposals for IBE schemes [6–9]. Hence the Tripathi algorithm [10] presents an RSA based algorithm to generate the cryptographic keys required by participants for secure communication using their identities which is similar to identity based encryption scheme (IBE).

However the main problem in Tripathi algorithm is cannot use all of the user email addresses as a public key. This is because, there are two reasons why it is unable to use all email addresses: (i) this algorithm use the same modulo value for every email, if the email is not related prime to modulo value, the new email should be entered. (ii) Entered email is composed of odd and even number. If email users cannot be used as a public key, then the user must enter a new email and the process is repeated until the user's email can be used as a public key. Therefore the Tripathi algorithm needs to be improved. In this study, two times of experiments that were carried out to produce satisfactory results and make sure all emails entered by user can be a public key. In a first experiment, the algorithm that has been improved from the Tripathi algorithm and it's called LB-RSA.

Looping process has been added to this algorithm to produce a new modulo value and helps to produce more email addresses that can be used as a public key and LB-RSA have shown good results, that is, 50 % of the total number of emails, can be used as a public key. This number was increased compared to the amount generated by Tripathi algorithm that is 25 %. The resulting decision issued by LB-RSA algorithms in experiment 1 do not reach 100 %, the second experiment carried out to provide a better algorithm and can achieve 100 % of emails that can be used as a public key. The classification process added in LB-RSA algorithm is to determine whether the decimal is even or odd number, this process helps to make all email can be a public key. Before emails are tested in these algorithms, emails will be converting to decimal value.

64.2 Related Work

64.2.1 Tripathi Algorithm

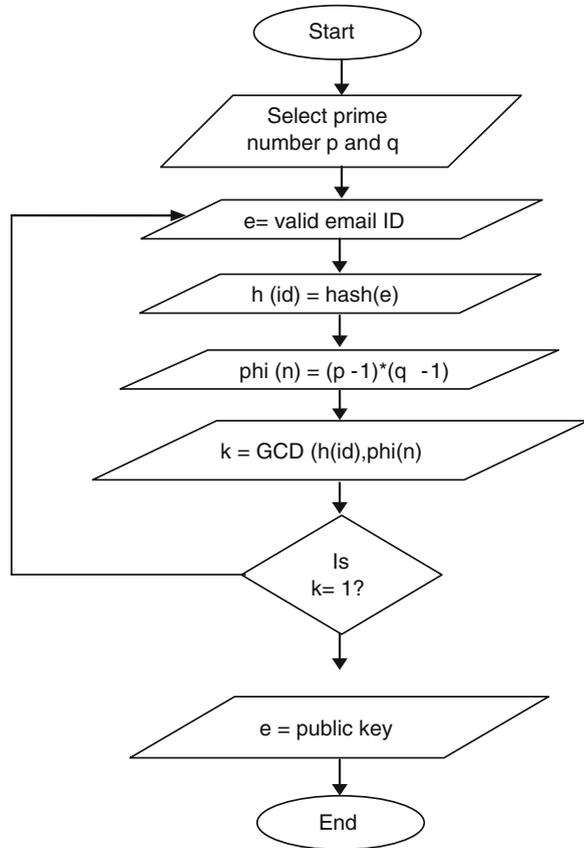
RSA algorithm was gone through several phases of change towards, a variety of improvements have been made in the original RSA algorithm, including algorithms that have been developed by Sachin Tripathi [11]. This algorithm also has several advantages compared with the original RSA algorithm. (1) Tripathi algorithm used user string identity as a public key such as an email address. (2) Need small device memory to store the public key. The algorithm presents the RSA based algorithm to generate the cryptographic keys required by participants for secure communication using their identities, which is similar to identity based encryption (IBE) scheme. Hence the authentication of the public key in IBE is a big challenge and for which public key certificates, provided by a Certification Authority (CA), are used. At large scale communication to set up numerous CA's is a major overhead of public key cryptography.

Hence this algorithm attempts to avoid the use of public key certificates and proposes an RSA based algorithm to generate the cryptographic keys using identity such as an email identity of a person. Figure 64.1 shows the key generation algorithm of the Tripathi flow algorithm. The email address entered will be converted into a fixed—size string using CRC32 hash function and then converted to a decimal value before test in this algorithm. After that the decimal value $h(id)$ will be tested whether it is relative prime to modulo $\phi(n)$ or not, $h(id)$ is relative prime to the modulo $\phi(n)$, then email address can be used as a public key, but if the email is not relative prime to the modulo, then the user must enter a new email as a public key. This is an example of email address convert to form of decimal using CRC32 hash function. However the problem in Tripathi algorithm is unable to use all user email as a public key. There are two reasons why it is unable to use all email addresses: (i) this algorithm use the same modulo value for every email, if the email is not relative prime to modulo value, the new email should be entered. (ii) Entered email is composed of odd and even number.

64.3 Proposed Algorithm (CLB-RSA)

The methodology proposed will be named CLB-RSA algorithm. The character of “CLB” is standing for classification and loop based process. A representation of “C” is standing for classification and “LB” is standing for loop based. The advantage of this algorithm is able to use all user emails address as a public key. There are two enhancement process added in this proposes algorithms are: looping and classification process. Advantages of using user identity abridged in order to facilitate the user to remember public key. Example of user identity is including user name, user email, nickname and so on. Improvements done on this algorithm has

Fig. 64.1 Tripathi algorithm flowchart



improved the performance of CLB-RSA, the performance on the number of emails that can be used as a public key is increase and reach of hundred percent results.

The improvement was made to ensure that all email address can be used as a public key. As discussed in related work, Tripathi algorithm cannot use all the emails as a public key, and only certain email that can be used as a public key. Comparison of the performance can be seen between the Tripathi algorithms and CLB-RSA algorithms. Comparison between these algorithms is in terms of the number of emails that can be made as a public key email based on twenty samples listed in Table 64.1. CLB-RSA is second algorithm evaluate from first experiment by Muhammad [12].

64.3.1 LB-RSA Algorithm

In this study, two times of experiments that was carried out to produce an algorithm that can produce a satisfied result. In a first experiment, the algorithm that

Table 64.1 Result LB-RSA and CLB-RSA

No	Email	Decimal	LB-RSA	CLB-RSA
1	yaron4329@coolmail.co.il	786745081	Yes	Yes
2	odedny@012.net.il	2730798235	Yes	Yes
3	sismal@t2.technion.ac.il	1812161612	No	Yes
4	knarik2000@mail.ru	3238888356	No	Yes
5	simon.goldman@intel.com	795509258	No	Yes
6	kristal@netvision.net.il	3700340025	Yes	Yes
7	daniel227@bezequnt.net	4067366447	Yes	Yes
8	nitzan@tzel.net	1199321065	Yes	Yes
9	tall_g@telepark.co.il	3074048744	No	Yes
10	zhan_t@hotmail.com	2205045334	No	Yes
11	Jacklml@consultant.com	3009834837	Yes	Yes
12	alayan1@012.net.il	2654073728	No	Yes
13	romangr@matrix.co.il	3765641505	Yes	Yes
14	rbarash@elta.co.il	296517678	No	Yes
15	motyy@isa.gov.il	2386521674	No	Yes
16	mrs.hidayah@yahoo.com.my	2962995952	No	Yes
17	fatem_alyahya@yahoo.com	3843922542	No	Yes
18	igor@bizportal.co.il	2489219047	Yes	Yes
19	max@usermail.com	3211800885	Yes	Yes
20	sharon_m@ifat.com	2388137343	Yes	Yes

has been improved from the Tripathi algorithm and it's called LB-RSA [12]. Looping process is added in LB-RSA to generate the new parameters of modulo phi (n). In Tripathi algorithm, looping process will be run on email address, it's mean, if email cannot be a public key, so users need to enter new email address, so the loop process is happen on email address. However in LB-RSA algorithm, the looping process is running modulo value phi (n), if the email address cannot be a public key, so new modulo value will be created. When the new parameters of phi (n) are produced, then the probability of the value of h (id) and phi (n) is relative prime is higher, and it is usually referred to equation $k = \text{GCD}(h(id), \text{phi}(n))$. If k is equal to 1, so this it meant h (id) is relative prime to phi (n), and the email's address can be used as public key. In Tripathi algorithm, modulo phi (n) generated only once and if the value of k is not equal to 1, then the email entered is declared as a non-public key, and the user should enter a new email address. These processes write:

```

If GCD (h(id), phi(n))=1;
    h(id)=public key;
    else
Looping process until k=1;
    
```

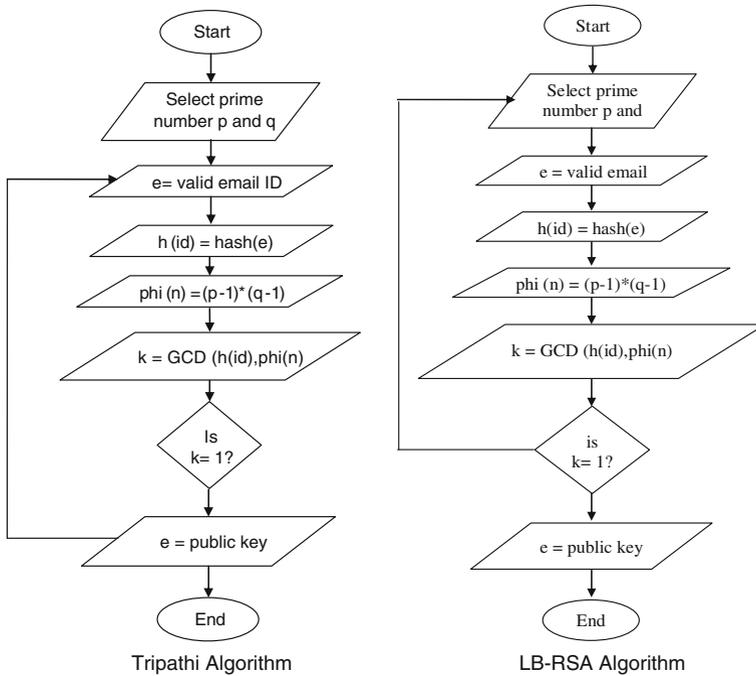


Fig. 64.2 Comparison between Tripathi and LB-RSA Key Generation Algorithm

Figure 64.2 shows the flowchart of LB-RSA and Tripathi key generation algorithm. This comparison between two algorithms: Tripathi algorithm and LB-RSA algorithm. As shown in Fig. 64.2, the main differences between these algorithms are, in Tripathi algorithm, looping process is running on email address. When the email entered by the users, can't be a public key, so users need to enter the new email address until email address can be used as a public key. So it's mean reenter the new emails is the looping process in Tripathi algorithm. However in LB-RSA algorithm, the looping process occurred on modulo value, if the email address can't be a public key, thus a new modulo value phi (n) will be continuously created until the email can be a public key. So it's mean the looping process is done in modulo value and not on email address. This process will be fully effective because the email addresses that can be a public key are increasing in LB-RSA algorithm.

64.3.2 CLB-RSA Algorithm

The mathematical calculations that are used in CLB-RSA key generation algorithm describe to show detail how it's calculated in table 64.2. Steps 6 and 7 are

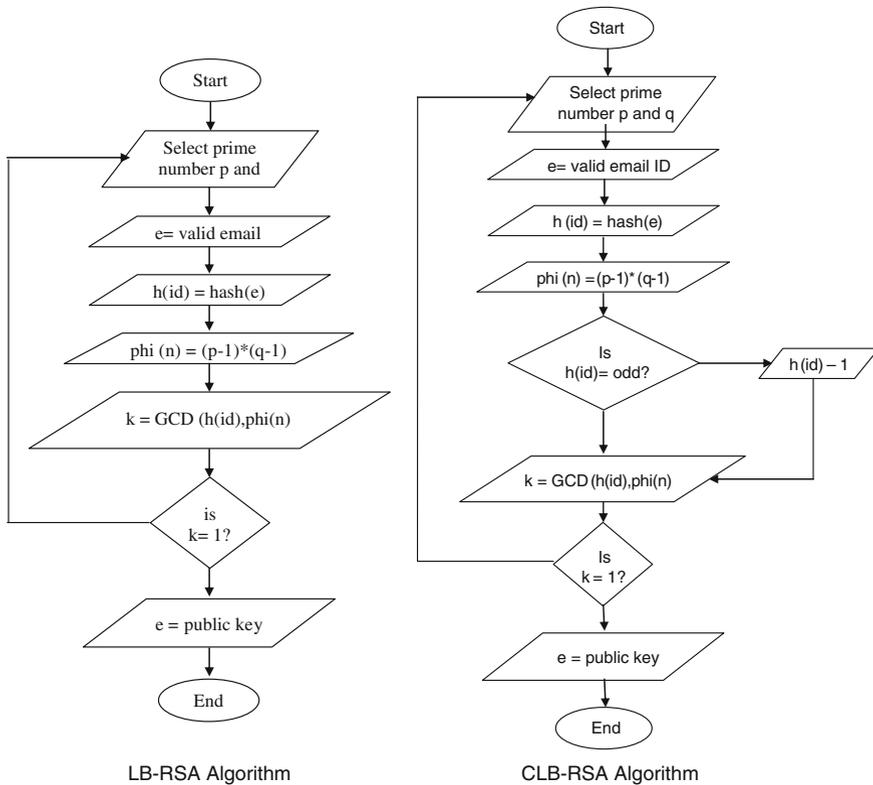


Fig. 64.3 Comparison between CLB-RSA Algorithm and LB-RSA Algorithm

the additional steps of this algorithm different with Tripathi algorithm. Small number used in example in table 64.2 to give simple calculation, and also used even decimal value as a public key to show how this algorithm make a classifying process to determine either the decimal entered is even or odd number. The even number of decimal value is converted to odd number by reducing the value as shown in step 6. After emails tested, and the result for entered email is an odd number, then the next process can be carried out and an email address can be used as a public key, even maybe the looping process requires more than a one-time loop. However, if the decimal value is even, then the value should be converted to the odd before proceed to the nest process. Therefore, when the value is odd umber, then the email can be used as a public key. This classifying process will be solving the second problem (even decimal value can't be a public key) in Tripathi algorithm that causes this algorithm cannot used all email as a public key. Therefore LB-RSA can solve the first problem that causes the Tripathi algorithm

cannot use all email as a public key using looping process hence create the new modulo value until email modulo value is related prime.

To convert an even decimal value to odd decimal value, decimal value should be minus 1 or plus 1. But for this algorithm, the decimal value will be minus 1 to produce odd decimal value. Once the email is turned to the odd number, and then the next process can be implemented. This is the difference process between LB-RSA and CLB-RSA, and this improvement step can complement this algorithm because CLB-RSA can make all email entered can be used as a public key. The addition process of this algorithm is to make tests on the decimal to classify the decimal value, whether it is an odd number, or even number. This equation is used for this process. If $h(id) = \text{odd}$, then continue step 7, else $h(id)-1$, continue step 7. Figure 64.3 shows the flowchart of the steps in CLB-RSA key generation algorithm, the process begins with the declaration of two random numbers as a parameter p and q and lastly public key and private key are determined.

Figure 64.3 shows the comparison between two key generation algorithms, LB-RSA and CLB-RSA. As shown in Fig. 64.3, the main difference is the classification process. This process is added in this algorithm to make classifying on decimal value, this process is very important because if the decimal value is even number, so the email address entered cannot be a public key. Classification process added in this algorithm is to make a classifying of decimal value to odd or even number. If the decimal value is an odd number, so there is nothing happened to the decimal value and can continue to the next step and make it as a public key. However, if the decimal value is even, so this decimal value must be converted to odd number first before proceeds as a public key. However the looping process is still available in this algorithm because it will be useful when the decimal value is an even number. If a decimal value is an odd number, but decimal value and modulo value is not a prime number, so new modulo value will be created and this process is fully working when the email that can be a public key is increasing.

Table 64.1 shows the result produced by LB-RSA and CLB-RSA. LB-RSA algorithm is tested in experiment 1 and CLB-RSA algorithm is tested in algorithm 2. Both of these tests using the same twenty sample email listed in Table 64.1. In LB-RSA, half of data can be used as a public key, and a half of the data cannot be a public key, but in CLB-RSA all data can be used as a public key. This proves the effectiveness of CLB-RSA algorithm and the addition step is performed on this algorithm have a significant impact on this algorithm, successfully making all the emails as a public key.

Table 64.2 CLB-RSA Key Generation Calculation

1. Select two parameters, namely p and q, these two parameters must be a prime number.
 $p=3, q=7$
2. Compute n by using the following formula.
 $n = p * q$
 $n = 21$
3. Compute phi (n).
 $\phi(n) = (p - 1) * (q - 1)$
 $\phi(n) = 12$
4. Input the user valid email Id.
 $e = \text{valid email id}$
 $e = \text{odedny@012.net.il}$
5. Convert email id to hash function.
 $h(id) = \text{hash}(e)$
 $h(id) = 1812161612$
6. Classified decimal value.
if h(id) = odd number?
Continue step 7
else
 $h(id) - 1$
Continue step 7
 $1812161612 == \text{odd?} = \text{NO}$
else
 $1812161612 - 1 = 1812161611$
 $h(id) == 1812161611$
Continue step 7
7. Choose the exponent k if :
If GCD (h(id), phi (n)) = 1
Continue step 8
Else, loop step 1 the process until k=1
 $\text{GCD}(1812161611, 12) = 1$
8. Compute private key exponent d through following formula
 $d = e^{-1} \text{ mod } (\phi(n))$
 $d = 7$
9. Thus the public key consists of public key exponent e and n. And the private key consists of private key exponent d and n.
 Public key: (n, e)
 Public key: (n, d)
 Public key: (21, 1812161611)
 Public key: (21, 7)

64.4 Conclusion

The algorithm has been improved and called CLB-RSA has achieved the objectives of improving the RSA algorithm using the email address as the public key, and make sure all the email entered by the user can be used as a public key. The email addresses used in this study consisted of a variety of different domains. The results produced by CLB-RSA are very gratified that 100 % of the email can be public key. These results are better than the previous algorithm (Tripathi algorithm). Problems faced by Tripathi algorithm can be solve by CLB-RSA algorithm using loop based and classification process. However in future, this algorithm can be improved better than CLB-RSA such as private key can be replaced with private user identity.

Acknowledgments This study is funded by Skim Latihan Akademik IPTA (SLAI) under the Malaysia Ministry of Higher Education (MOHE) and Universiti Sultan Zainal Abidin (UniSZA) Malaysia. The authors would like to acknowledge all contributors, and others who have helped and greatly assisted in the completion of the study.

References

1. Diffie, W., Hellman, M.E.: New directions in cryptography. *IEEE Trans. Inf. Theory* **22**, 644–654 (1976)
2. Zain, J.M., Fauzi, A.R.: Medical image watermarking with tamper detection and recovery. In: 28th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBS'06) (2006)
3. Zain, J.M., Fauzi, A.R.: Evaluation of medical image watermarking with tamper detection and recovery (AW-TDR). In: 29th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBS 2007) (2007)
4. Wenbo, M.: *Modern cryptography: theory and practice*. Publisher: Prentice Hall PTR, Hewlett Packard (2004)
5. Shamir, A.: Identity-based cryptosystems and signature schemes. In: *Advances in Cryptology* (1985)
6. Desmedt, Y., Quisquater, J.-J.: Public-key systems based on the difficulty of tampering (Is there a difference between DES and RSA?). In: *Advances in Cryptology—CRYPTO'86* (1987)
7. Maurer, U.M., Yacobi, Y.: Non-interactive public-key cryptography. In: *Advances in Cryptology—EUROCRYPT'91* (1991)
8. Hühnlein, D., Michael, Jr J., Weber, D.: Towards practical non-interactive public-key cryptosystems using non-maximal imaginary quadratic orders. *Des. Codes Crypt.* **30**, 281–299 (2003)
9. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. *J. Comput. SIAM* **32**, 586–615 (2003)
10. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **26**, 96–99 (1983)
11. Tripathi, S., Biswas, G.P., Kisan, S.: Cryptographic keys generation using identity. In: 3rd International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2011) (2011)
12. Muhammad, N., Zain, J.M., Mohd Saman, M.Y.: Loop-based RSA key generation algorithm using string identity. In: 13th International Conference on Control, Automation and Systems (ICCAS) (2013)

Chapter 65

Rules Mining Based on Clustering of Inbound Tourists in Thailand

Wirot Yotsawat and Anongnart Srivihok

Abstract Tourism industries are growing up rapidly with more competition. So, travel agencies or tourism organizations must have a good planning and provide campaign for tourist's needs. This study proposes the usage of data mining for tourism industries in Thailand. Data clustering and association rule mining were chosen as the data mining methods in order to discover useful knowledge. Two-level clustering with decision tree bagging was applied to construct the segments of tourist. Apriori algorithm was then used to find the rules on each cluster. The experimental results indicated that the tourists data was separated into eleven differently segments and decision tree bagging for attributes weighting can enhance the quality of clusters. The eleven segments were analyzed in order to identify tourists' behavior patterns and their preferences. Association rule mining was applied to each segment in order to find the relationship among the features of tourist data. The rules were filtered again by experts. The clustering and association rule results can be served to tourism organization in order to support their strategic and market planning.

65.1 Introduction

Tourism industries are growing up rapidly in many countries. There are many supporting policy and developing plans from their governments. Activities and destinations are promoted for the attraction of tourists and tourism investors from foreign countries. Tourists receive more choices for selecting the best interesting places. So, travel agencies or tourism organizations must have a good planning and provide campaigns for tourist's needs. They have to know tourists' behavior

W. Yotsawat (✉) · A. Srivihok
Department of Computer Science, Faculty of Science, Kasetsart University,
Bangkok, Thailand
e-mail: g5314401258@ku.ac.th

A. Srivihok
e-mail: fsciang@ku.ac.th

patterns and their preferences. This paper proposes data mining techniques on inbound tourists in Thailand by using segmentation and association rule techniques. Two-level clustering with factors weighting by Decision Tree bagging, was a methodology for applying cluster analysis to explore the tourists patterns for market planning, promotion and package design for each group. Association rule mining was then applied to each segment. The results of association rule can be applied for tourists' recommendation system. Thus, this study can serve as useful knowledge for travel agencies and other tourism organizations.

65.2 Literature Review

Data mining is the process of automatically discovering useful information in large data base [1]. Since data mining was introduced, it has developed by many researchers. Data mining technique has been applied in many fields such as accounting, medicine, law, and so forth. Some researchers focused on the implementation of data mining for tourism such as Wong, Chen, Chung, and Kao [2], they proposed the usage of three data mining techniques to analyze the travel patterns of Northern Taiwan tourists. RFM (Recency, Frequency and Monetary Value) was applied to identify valuable travelers, C4.5 decision tree and association rule were then applied to discover the traveling pattern and rule. Gul Gokay Emel, Cagatan Taskin and Omer Akat [3] applied Apriori rule mining to profile the domestic tourists of Bursa, Turkey and provided suggestions for relevant tourism enterprises. Moreover, Brida et al. [4] implemented two-level approach to conduct cluster analysis based on Italian Christmas Market visitors.

Some researchers used the Decision Tree weighting for improving the accuracy of classifier such as Kaewchinporn et al. [5] used Decision Tree bagging to weight features for improve the predictive performance of K-Means. Hall [6] applied Decision Tree bagging to weight attributes for improve the performance of Naive Bayes classifier. This study, the researchers enhanced the quality of cluster by applying the Decision Tree bagging weighted to the features of data. After clusters were constructed, tree bagging and attributes weighting were applied. K-Means and Self Organizing Map (SOM) algorithm were then applied to refine the clusters. Useful knowledge can be served to tourism enterprises.

65.3 Related Algorithms

65.3.1 *Two-Level Clustering*

There were some researchers who used two-level clustering for market segmentation. Conventional approach of two-level clustering is the combination of Hierarchical and non-Hierarchical techniques. The advantages of first level are

solved the limitation of algorithm on the second level. For example, Punj and Steward [7] presented two-step clustering by binding of Hierarchical Clustering (HAC) and K-Means approach. The HAC was used to find the number of cluster and initial seeds for the input of K-Means algorithm. However, HAC cannot handle the large data set. Moreover, once a decision is made to combine two clusters, it cannot be undone. Kuo [8] proposed two-stage clustering by combining of SOM and K-Means algorithms. In the first stage, the researcher replaced HAC with SOM for solve the limitation of HAC. His proposed method is slightly better than the conventional two-stage method.

In this present, we implement SOM and K-Means approach to segment tourist’s dataset. SOM algorithm was discovered by Kohonen. The SOM algorithm is used to find the optimum number of cluster by calculating two criterions consist of RMSSTD and RS [9] which defined as.

$$RMSSTD = \sqrt{\frac{\sum_{j=1..d} \sum_{i=1..n_c} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{j=1..d} (n_{ij} - 1)}} \tag{65.1}$$

$$RS = \frac{\sum_{j=1}^d \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 - \sum_{j=1..c} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2}{\sum_{j=1}^d \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2} \tag{65.2}$$

- c is the number of cluster.
- d is the number of dimension.
- \bar{x}_j is the mean value of j th dimension.
- n_{ij} is the number of sample in i th cluster, j th dimension.

After the optimum number of cluster was found, the initial seeds were determined. They were used to input to K-Means method in the second step. For further calculation of SOM and basic K-Means algorithm are available in Tan [1].

65.3.2 Decision Tree Bagging and Features Weighting

This study, the researchers used Decision Tree bagging for features weighting because of the factors may have different importance. This method may improve the quality of cluster. Tree bagging uses a Decision Tree algorithm to construct n models based on a different diversity of training data. The bagging algorithm was shown as following [5].

```

Algorithm: Bagging
Input: D: data set;
       n: the number of models;
       - a learning scheme (e.g., decision tree)
Output: A composite model, M*.
Method:
(1) For i=1 to n do
(2)   create bootstrap sample, Di, by sampling D with
replacement
(3)   use Di to derive a model, Mi;
(4) End for

```

The attributes weighting technique uses the features which appear in the trees. After decision trees were constructed, attributes which appeared in each trees were selected and computed the weight. The weight of attribute was varying on a size of tree and the position of that attribute appearing in that tree. The computation of weight for each attribute was defined as [5].

$$w_{(k,i)} = (\text{height_M}_i - j + 1) / (\text{height_M}_i + 1) \quad (65.3)$$

where $w_{k,i}$ is the weight of attribute k in tree i , height_M_i is the height of tree i and j is the level of attribute node k in tree i .

All attribute weights from each tree models were calculated for an average weight by the following equation.

$$w_k = (w_{k,1} + w_{k,2} + \dots + w_{k,n}) / n \quad (65.4)$$

where w_k is the total weight of attribute k and n is the number of tree models.

65.3.3 Apriori Algorithm

Association rule was discovered by Agrawal et al. [10]. It is used in the recommendation systems such as www.amazon.com. Apriori was well-known algorithm and popular usage in market basket data analysis. It was used to find the relationship between two or more attributes in the large database. There were two standard measurements such as minimum support (Minsup) and minimum confidence (Minconf). Support was used to evaluate the statistical importance of a set of transactions in database such as $\text{Sup}(X, D)$ represented the rate of transactions in D containing the item set X . Confidence represented the rate of transactions in D that contain item set X and also item set Y . It was defined as $\text{Conf}(X \rightarrow Y) = \text{Sup}(X \cap Y) / \text{Sup}(X, D)$. The first step of Apriori algorithm was to detect a large item set with greater than minimum support and the second step was to generate association rules with greater than minimum confidence. Moreover, lift value was used to reduce the possible biases when used the support and confidence values. Lift was defined as $\text{Lift} = \text{Conf}(X \rightarrow Y) / \text{Sup}(Y)$ [11]. The association rules were

useful for many applications such as tourism marketing [3], tourism recommendation systems [12, 13], new product development and customer relationship management [14].

65.4 Methodology

65.4.1 Data Collection

This study, the researchers used secondary data accumulated from the Department of Tourism, Ministry of Tourism and Sports, Thailand. The total number of inbound tourist in the data files is 83,402 who arrived to Thailand in a period from 2008 through 2010. Data pre-processing were applied by removing unreliable, missing values and outliers. So, the total number of observation in the sample was 79,473. The attributes of data are the length of stay (days), age (years), gender, occupations, annual incomes (US Dollars), average expenditures (Baht per day), purpose of visit, types of accommodation, tourist origins, places of residence and types of transportation.

65.4.2 Study Framework

The specification of this study required the applying of data mining technique to partition the data into segments and to discovered the relationship among the features of tourist in each segment, respectively. Using cluster analysis and association rule, this study analyzed tourist behaviors and then extracted knowledge to explore useful information for tourism businesses. The useful knowledge can be helpful for tourism organizations in order to understand the tourist needs and their preferences.

So, the research design was divided into two phases included data clustering phase and association rule mining phase. Preprocessing methods were used to clean the data. Clustering phase was then conducted by two-level clustering and weighting features by decision tree bagging. SOM algorithm was performed to find the optimum number of cluster. The appropriate number of cluster was defined by the computation of RMSSTD and RS. The cluster labels were assigned to build decision tree bagging for attributes weighting. Attributes weighting by decision tree bagging included four cases. Case one, case two and case three consisted of ten trees which each tree was constructed by 25, 50 and 75 % of random with replacement tourist data, respectively. Case four was built by overall data for one tree. The features set were weighted and the data set was refined by the clustering algorithms. Each cluster was analyzed when the clustering phase was finished. After some attribute was preprocessed in order to qualify the requirement of

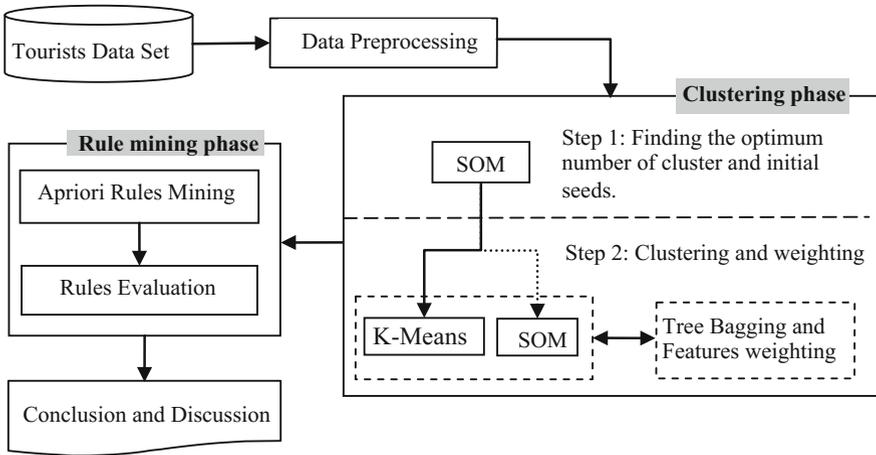


Fig. 65.1 Study framework

Apriori algorithm, association rule mining was performed for each segment. The study framework was illustrated in Fig. 65.1.

65.5 Experimental Results

65.5.1 Clustering Results

Finding the optimum number of cluster, Table 65.1, eleven clusters are obtained from SOM method by the computation of RMSSTD and RS. After cluster labels were assigned to each record, decision tree bagging was constructed for calculate the weight of each attribute. After that, tourists' data was recomputed and compared the RMSSTD and RS between data with weighted attributes and data without weighted attributes. The experimental result indicated that data with weighted attributes give a better quality. It was illustrated on Table 65.2.

The result of clustering phase was shown on Table 65.3. The differences of attribute among segments were illustrated on Table 65.3. The result showed no significant difference on gender, age and occupation. Segment 1 was relatively dominant among the eleven clusters. It comprised over 31 % (n = 24,777) of overall tourists (N = 79,473) and can be considers as a homogeneous cluster. The majority of visitor in Cluster 9 traveled to Thailand for business purpose. The expenditure per day of Cluster 9 was highest around 4,747 Baht by average. The tourists of cluster 11 stayed in Thailand for a longest time when compare with the tourists in other clusters. Thus, they used a variety of transportation' types. Cluster 3 chose domestic airplane for transportation. Moreover, clustering results

Table 65.1 Finding the optimum number of cluster by SOM clustering

#Cluster	RMSSTD	RS	#Cluster	RMSSTD	RS
2	2.7653	0.1092	9	2.4865	0.2798
3	2.6827	0.1616	10	2.4487	0.3015
4	2.6279	0.1955	11	2.4279	0.3134
5	2.5873	0.2202	12	2.4320	0.3111
6	2.5678	0.2319	13	2.4523	0.2995
7	2.5173	0.2618	14	2.4359	0.3089
8	2.4880	0.2790			

Table 65.2 The comparison of RMSSTD and RS between SOM and K-Means for data with weighted and not weighted attributes (number of cluster = 11)

Algorithms	Weighting	RMSSTD	RS
SOM	No	2.4279	0.3134
K-Means	No	2.4277	0.3135
SOM	Yes	1.6933	0.4395
K-Means	Yes	1.7001	0.4350

suggested that the expenditure per day was inverse variation with the length of stay. In other word, the expenditure of cluster 11 was the least but cluster 11 stayed in Thailand for the longest term, the expenditure of cluster 9 was the highest but cluster 9 stayed in Thailand for a short time. Cluster analysis results shown that the tourism organizations can apply the useful knowledge for market planning, promotion design and other related tourism developments.

65.5.2 Association Rule Mining Results

Clustering methods provided the distinct characteristics of segment but cannot show the association among the features of data. Thus, Apriori rule mining was applied to each tourist cluster. The rules with lift value ranging from 1.00 to 1.34 are obtained with minimum rule support (Minsup) of 25–35 % and minimum rule confidence (Minconf) of 80–85 %. Table 65.4 described a part of association rules which discovered from each tourist segment. Four significant rules were demonstrated on Table 65.4.

The experimental results indicated that the tourist segments provided association rules which were related to the characteristics of each segment. The rules could be transformed to if-then clause. Finally, the rules were evaluated by tourism

Table 65.3 The significant characteristics of cluster

Factors	Clusters										
	1	2	3	4	5	6	7	8	9	10	11
Sample size (%)	31.18	2.50	5.6	15.47	5.65	3.59	4.41	7.5	10.33	10.14	3.61
Stay' length (days)	6	12	13	8	9	9	11	11	6	8	17
Income \$US (%)											
<20 k	43	40	25	44	50	34	36	32	32	32	42
20-40 k	36	29	34	33	30	31	34	34	31	36	33
40-60 k	12	14	24	14	12	18	17	18	17	22	14
Expenditure (THB)	4,439	3,242	4,551	4,174	3,322	4,268	3,568	4,147	4,747	4,215	2,815
Purpose (%)											
Holiday	100	85	91	92	94	85	93	90		92	92
Business		3	2	3	1	6	2	2	55	2	1
Other		12	7	5	5	9	5	8	45	6	7
Accommodation (%)											
Hotel	100		94	100		100			91	99	
Resort		35			60		32		1		89
Guesthouse		38			18		57	100	1		
Apartment		23	5		12		7		4	1	9
Zone (%)											
America	5	14	11	5	10	12	10	9	5	7	16
East Asia	28	21	11	30	30	22	29	20	21	19	8
Europe	13	27	46	20	24	25	34	38	10	30	55
Oceania	4	7	11	3	6	9	4	12	4	6	8

(continued)

Table 65.3 (continued)

Factors	Clusters										
	1	2	3	4	5	6	7	8	9	10	11
South Asia	1	5	4	12	6	3	4	3	17	3	1
SEA	33	17	6	21	19	20	12	11	34	26	5
Destination (%)											
Central	100				100				97		
Southern			96							91	94
Northern		88				100		95			
Eastern				100			100				
Transportation (%)											
Plane	6	36	100	4	15	29	10	41	5	0	53
Bus	11	37	23	15	45	14	38	25	10	12	68
Train	6	16	10	3	21	6	8	10	6	4	30
Ferry	7		36	9	26		28	30	5	9	62
Other	65	60	72	45	76	48	63	54	80	42	71

n = 79,473

Table 65.4 Four significant rules for each tourist cluster with Minsup = 25–35 %, Minconf = 80–85 % and lift = 1.00–1.34

#	Antecedents	Consequent
R ₁₁	P ^d = Holiday, A = Hotel, E ^c = 4,514–6,832 Baht/day	D ^b = Central
R ₁₂	P = Holiday, Occupation = Professional	A ^a = Hotel, D = Central
R ₁₃	A = Hotel, Age = 25–34 years	P = Holiday, D = Central
R ₁₄	P = Holiday, Zone = East Asia	A = Hotel, D = Central
R ₂₁	A = Guesthouse	D = Northern, P = Holiday
R ₂₂	D = Northern, Income = less than 20,000 \$US	P = Holiday
R ₂₃	Income = less than 20,000 \$US	D = Northern
R ₂₄	T ^c = Plane	D = Northern
R ₃₁	D = Southern, Zone = Europe	T = Plane, P = Holiday
R ₃₂	T = Plane, Age = 25–34 years	D = Southern
R ₃₃	T = Plane, P = Holiday, Gender = Male	D = Southern, A = Hotel
R ₃₄	D = Southern, Age = 25–34 years	A = Hotel
R ₄₁	D = Eastern, Gender = Male, T = Other	A = Hotel
R ₄₂	D = Eastern, Income = 20,000–39,999 \$US	A = Hotel, P = Holiday
R ₄₃	D = Eastern, A = Hotel, Age = 25–34 years	P = Holiday
R ₄₄	P = Holiday, Income = less than 20,000 \$US	D = Eastern, A = Hotel
R ₅₁	P = Holiday, Age = 15–24 years	D = Central
R ₅₂	Income = less than 20,000 \$US	D = Central, P = Holiday
R ₅₃	D = Central, P = Holiday	A = Resort
R ₅₄	P = Holiday, T = Other, Income = less than 20,000 \$US	D = Central
R ₆₁	D = Northern, Gender = Female	A = Hotel, P = Holiday
R ₆₂	A = Hotel, P = Holiday, Gender = Female	D = Northern
R ₆₃	D = Northern, Income = 20,000–39,999 \$US	A = Hotel
R ₆₄	Occupation = Professional	D = Northern, A = Hotel
R ₇₁	Zone = Europe	D = Eastern, P = Holiday
R ₇₂	D = Eastern, T = Bus	P = Holiday
R ₇₃	D = Eastern, A = Guesthouse	P = Holiday
R ₇₄	Age = 25–34 years	D = Eastern
R ₈₁	T = Ferry	A = Guesthouse, D = Southern
R ₈₂	Occupation = Professional	A = Guesthouse
R ₈₃	A = Guesthouse, Income = 20,000–39,999 \$US	P = Holiday
R ₈₄	A = Guesthouse, Age = 25–34 years	D = Southern
R ₉₁	A = Hotel, P = Business	D = Central
R ₉₂	T = Other, Gender = Male, P = Business	D = Central, A = Hotel
R ₉₃	D = Central, Occupation = Professional	A = Hotel
R ₉₄	D = Central, Zone = South East Asia	A = Hotel
R ₁₀₁	P = Holiday, D = Southern, Gender = Female	A = Hotel

(continued)

Table 65.4 (continued)

#	Antecedents	Consequent
R ₁₀₂	Income = 20,000–39,999 \$US	A = Hotel
R ₁₀₃	P = Holiday, Gender = Female	D = Southern
R ₁₀₄	D = Southern, Gender = Male	A = Hotel, P = Holiday
R ₁₁₁	D = Southern, T = Plane	P = Holiday, A = Resort
R ₁₁₂	P = Holiday, Zone = Europe	D = Southern, A = Resort
R ₁₁₃	D = Southern, P = Holiday, T = Ferry	A = Resort
R ₁₁₄	D = Southern, Income = less than 20,000 \$US	A = Resort

- ^a A = Accommodation
- ^b D = Destination
- ^c E = Expenditure
- ^d P = Purpose
- ^e T = Transportation

Table 65.5 Example of implementation of clustering and association rules mining results

Knowledge founding	Actionable activities which should be focused on
Most of the extracted rules on cluster 1 belong to “Holiday” purpose, “Hotel” accommodation and “Central” destination	Tour agencies design package tours as a base for segmentation., Tourism recommendation systems
Cluster 11 is a “price-conscious” segment	Marketing managers should be focused on pricing strategies and planning
Cluster 9 is a “hi-end” segment	Marketing managers should be focused on the quality of services and products
For cluster 3, if “Destination = Southern” and “Zone = Europe” then “Transportation = Plane” and “Purpose = Holiday”	Information center should be focused on tourist attraction and transportation for holiday on Southern part of Thailand
For cluster 9, if “Destination = Central” and “Purpose = Business” then “Accommodation = Hotel”	Hotel at Central part of Thailand should be prepared their location and relationship between business activities

professionals. Some rule was accepted by minimum criterions but it was rejected by experts filtering. However, the most difficult of this study was the translation of the segmentation and association rules to the suggestion of tourism management. Table 65.5 shown the activity which tourism organizations should be focused.

65.6 Conclusion

Data mining can be extracted hidden information and patterns on the inbound tourist data. This study focused on the role of data mining for tourism industries in Thailand. Data clustering and association rule mining were chose as the data

mining methods in order to discover hidden knowledge. The clustering results indicated that inbound tourists in Thailand consisted of various segments with different profiles. Tourist data was segmented into eleven clusters by two-level clustering. Decision tree bagging method was then applied in order to enhance the quality of cluster. The eleven segments were analyzed to identify tourists' behavior patterns and their preferences. The experimental results indicated that the distinct characteristics among clusters can be helpful for tourism organizations in order to define the market planning or strategic making such as tourism management or package tour designing. After eleven segments were analyzed, association rule mining was applied in order to discover the relationship among the features of tourist in each segment. Minimum support and minimum confidence were set to filter the rules which generated by Apriori algorithm. Finally, the rules were evaluated by experts. Association rules mining results indicated that the tourist segments provide association rules which related to the characteristics of each segment. The rules can be implemented on recommendation systems or marketing action which should be focused on the rules found.

Future studies of data mining on tourism can focus on other related features such as cultural, socio-economic variables and values added from incompletely data such as RFM analysis. Finding a good way to weight features is very important for optimizing clustering results of many real world dataset. Moreover, the researchers should be suggested the data collector to collect more information for data analysis and researching.

Acknowledgments This work was funded by Faculty of Science and the Graduate School of Kasetsart University.

References

1. Tan, P.N., Steinbach, M., Kumar, V.: *Introduction to Data Mining*. Pearson Education Inc, San Francisco (2006)
2. Wong, J., Chen, H., Chung, P., Kao, N.: Identifying valuable travelers and their next foreign destination by the application of data mining techniques. *Asia Pac. J. Tourism Res.* **11**(4), 355–373 (2006)
3. Emel, G.G., Taskin, C., Akat, O.: Profiling a domestic tourism market by means of association rule mining. *Anatolia: Int. J. Tourism Hospitality Res.* **18**(2), 334–342 (2007)
4. Brida, J.G., Disegna, M., Osti, L.: Segmenting visitors of cultural events by motivation: a sequential non-linear clustering analysis of Italian Christmas market visitors. *Expert Syst. Appl.* **39**(13), 11349–11356 (2012)
5. Kaewchinporn, C., Nattakan, V., Vongsuchoto, S.: A combination of decision tree learning and clustering for data classification. In: *Proceedings of 2011 Eight International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 11–13 (2011)
6. Hall, M.: A decision tree-based attribute weighting filter for Naïve Bayes. *Knowl.-Based Syst.* **20**(2), 120–126 (2007)
7. Punj, G., Steward, D.W.: Cluster analysis in marketing research: review and suggestions for applications. *J. Mark. Res.* **20**(2), 134–148 (1983)

8. Kuo, R.J., Ho, L.M., Hu, C.M.: Integration of self-organizing feature map and K-means algorithm for market segmentation. *Comput. Oper. Res.* **29**(11), 1475–1493 (2002)
9. Kovacs, F., Legany, C., Babos, A.: Cluster Validity Measurement Techniques. In: *World Scientific and Engineering Academy and Society (WSEAS)*, pp. 388–393 (2006)
10. Agrawal, R., Imilienski, T., Swami, A.: Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216. ACM Press, New York, NY, USA (1993)
11. Wang, Y.F., Chuang, Y.L., Hsu, M.H., Keh, H.C.: A personalized recommender system for the cosmetic business. *Expert Syst. Appl.* **26**(3), 427–434 (2004)
12. Aghdam A. R., Mostafa, K., Dong, C. et al.: Finding interesting places at Malaysia: a data mining perspective. In: *Math. Comput. Contemp. Sci.*, pp. 89–93 (2013)
13. Juwattanasamrn, P., Supattranuwong, S., Sinthupinyo, S.: Applying data mining to analyze travel pattern in searching travel destination choices. *Int. J. Eng. Sci. (IJES)* **2**(4), 38–43 (2013)
14. Liao, S., Chen, Y., Deng, M.: Mining customer knowledge for tourism new product development and customer relationship management. *Expert Syst. Appl.* **37**(6), 4212–4223 (2010)

Chapter 66

Designing a New Model for Worm Response Using Security Metrics

Madiah Mohd Saudi and Bachok M. Taib

Abstract Nowadays, worms are becoming more sophisticated, intelligent and hard to be detected and responded than before and it becomes as one of the main issues in cyber security. It caused loss millions of money and productivities in many organizations and users all over the world. Currently, there are many works related with worm detection techniques but not much research is focusing on worm response. Therefore, in this research paper, a new model to respond to the worms attack efficiently is built. This worm response model is called as eZSiber, inspired by apoptosis or also known as cell-programmed death. It is a concept borrowed from human immunology system (HIS), where it has been mapped into network security environment. Once the user's computer detects any indication of the worm attacks, the apoptosis is triggered. In order to trigger the apoptosis, security metrics plays a very important role in identifying the weight and the severity of the worm attacks. In this model, the static and dynamic analyses were conducted and the machine learning algorithms were applied to optimize the performance. Based on the experiment conducted, it produced an overall accuracy rate of 99.38 % using Sequential Minimal Optimization (SMO) algorithm. This performance criteria result indicated that this model is an efficient worm response model.

Keywords Worm response · Security metrics · Apoptosis · Static analysis · Dynamic analysis · Sequential minimal optimization (SMO)

M.M. Saudi (✉) · B.M. Taib
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia
e-mail: madiah@usim.edu.my

B.M. Taib
e-mail: bachok@usim.edu.my

66.1 Introduction

Australia Computer Emergency Response Centre (AuSCERT) reported that 7,962 cases of the compromised Australia web sites were serving malwares in the year 2013 [1]. While statistics from Malaysia Computer Emergency Response Team (MyCERT) show that malwares contributes as one of the major cyber threats in Malaysia [2]. In this research, a worm is defined as a malicious program that can replicate itself, moving from one computer to another or can propagate via a network without human intervention or an owner's consent. One of the examples of worm that has caused chaos all over world is called as the Stuxnet worm. It has infected many computers at Iran's nuclear station and moreover it has also infected computers all over the world primarily at Indonesia, India and United States of America [3]. Furthermore, earlier in the year 2012, Ramnit worm has succeeded stole 45,000 Facebook passwords and after a year the worm characteristics have been improved and it is capable to steal bank password and account details [4, 5]. If a comparison between current trends and those of 10 years ago is made, these historical worm attacks and infections ensured the reputation of the attacker and thus gaining respect from other attackers or hackers was paramount [6]. In contrast, the motivation for cyber-attacks in the past 5 years is based on profit.

When dealing with worm attacks, one of the hardest jobs to do is to identify either the attack is genuine or not. Then once the attack has been identified, the next step is to decide the severity level. If the attack caused damage to the infrastructure and stole confidential information from the organization or user, it can be categorized as high severity. Prior assigning the severity level, a proper procedure or mechanism should be used as guidance. As for this research, security metrics has been identified as one of the efficient way how to assign severity level for worm attacks which then leads to the worm response mechanism. Therefore, the objectives of this research paper are to produce a new model called as eZSiber for worm response by applying security metrics and apoptosis concept, where the apoptosis is mapped from human immunology system (HIS) into computer security perspective and then to evaluate the proposed model. Prior to that, the existing worm response techniques are investigated and evaluated. As for security metrics, it is very important method on how to assign the worm attacks severity level for the apoptosis to be triggered.

The apoptosis that is also known as the cell-programmed death is applied to avoid the worm from propagating to other computers in the same network or via the Internet. The apoptosis will only be triggered based on five (66.5) main characteristics of a worm, which are the payload, operating algorithm, infection, activation and propagation. These five main characteristics of a worm are further refined and reused by assigning it with a weight and severity, where the security metrics has been applied to trigger the apoptosis.

As mentioned earlier, security metrics is very important method to assign the worm severity level. Security metrics is a method that helps to quantify, classify and measure information on security operations. In security metrics, the studied threats are defined, and then the threats are transformed into metrics or representations that can easily be measured. Security metrics can also be measured based on the perimeter defence, control and coverage, availability and reliability and application risks. All these measurements were already taken into consideration when the worm analysis was conducted. Therefore, as a result, the weight and severity performance and value are tested based on data criticality level, infrastructure availability and loss of productivity. The main reason why security metrics method has been chosen in this research is due to its capabilities to make the job of defining, understanding, identifying and measuring information security efficient, accurate, measurable and reliable. This is also supported by Atzeni and Lioy [7], where they state that work can be more profitable if it is enhanced using the security metrics and is more efficient if it is measurable.

For eZSiber worm response model, the security metrics is applied based on data criticality, infrastructure availability and loss of productivity. Then these 3 factors were used as the basis for assigning a weight and severity value. The above 5 main characteristics of a worm were extracted from STAKCERT worm classification, which were used as the input for this worm response model. The STAKCERT worm classification is not discussed in this research paper and can be referred in paper by Saudi et al. [8].

This paper is organised as follows. Section 66.2 presents the related works and the challenges in applying weight, severity and security metrics into the worm response model. Section 66.3 explains the methodology used in this research paper which consists of static and dynamic analyses and the architecture of the controlled laboratory environment. Section 66.4 presents the research findings which consists the results of the testing and evaluation of the proposed worm response model. Section 66.5 concludes and summarises the future work of this research paper.

66.2 Related Works

In applying apoptosis in worm response model, the main challenges, which should be considered thoroughly is the method of assigning the apoptosis. Therefore, based on the experiments and analysis conducted in this research, weight and severity (assigned using security metrics) are identified as two important factors that trigger apoptosis. There are a few studies, which have considered weight as part of their work. Examples are those of Su [9], who built a real time anomaly detection system for denial of service (DoS) attacks using weighted k-nearest neighbour classifiers, Siddique and Maqbol [10], who used weighting in software clustering, Kim et al. [11], who used weight as part of the log analysis of incident response in a DSS system, Fisch et al. [12], who used weight to optimise radial basis function neural networks for an intrusion detection system and Middlemiss

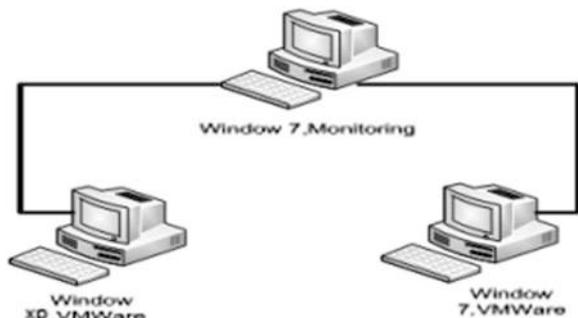
and Dick [13], who used weighted feature extraction using a genetic algorithm for an intrusion detection system. Based on these works, it can be concluded that there is no standard way of assigning weight, which has been seen as an important feature in increasing the accuracy or optimizing the performance of different works in different fields. Therefore, weight has been integrated in the eZSiber model, where security metrics and frequency analysis have been used to retrieve the rank and the value of the weights. Then, the weights are used for assigning the level of severity which triggers the apoptosis.

In a study conducted by Miles [14], he assigned a severity incident into three categories, which are high, medium and low. The high severity involves incidents with long term effects to the business or critical system for examples root access, denial of service (DoS) and it also involves with unauthorized privilege (root), limited access (user), unsuccessful attempt, utilisation of services and probe, poor security practices, malicious logic, hardware, software or infrastructure failure and espionage. Medium severity involves non-critical system and detection of initial attack and low severity involves detection on reconnaissance, threats of future attacks and rumours of security incidents. On the other hand, Reese [15] defined high severity as posing a threat to an entire autonomous system, such as a university network; that is a threat to the operation of critical network systems that threatens one or more applications that are integral to daily university functions. Medium severity involves a risk to isolated and non-production university systems and low severity involves minimal exposure of threats. By referring to the previous studies conducted in assigning severity, it can be concluded that severity must consider the data criticality, infrastructure availability and loss of productivity where these have been mapped in security metrics.

66.3 Methodology

In order to produce a new worm response technique, the researchers' had conducted few experiments and researches. A controlled laboratory environment is created to conduct the experiment as illustrated in Fig. 66.1. It is a controlled

Fig. 66.1 Controlled laboratory architecture



laboratory environment and almost 80 % of the software used in this testing is an open source or available on a free basis. No outgoing network connection is allowed for this architecture. The static and dynamic analyses were conducted.

The dataset used for this research are from VX Heaven [16] and Offensive Computing [17]. Many studies have used these dataset in their experiment [18, 19]. In these datasets there are many variant of malwares and benign files. The Knowledge Discovery and Data Mining (KDD) Processes as displayed in Fig. 66.2, were applied to these datasets. As a result, 160 datasets which consist of variants of the Windows worm and benign executable have been used for this research.

For this research, enhancements have been made to the KDD data pre-processing and pattern extraction process. Under the data pre-processing process, the static and dynamic analyses are implemented using the incident response standard operating procedures (SOP). While under the pattern extraction process, statistical methods comprising Chi-square and symmetric measure and security metrics are also introduced, as illustrated in Fig. 66.2. The details how the security metrics is applied in eZSiber worm response model can be referred in Table 66.1.

Furthermore, to retrieve the exact number of values for each of the worm’s attributes, relative frequency is used. Then these values were further tested with different algorithms to identify the best overall accuracy value. The equation used for relative frequency is shown in Eq. 66.1.

$$rf_n(E) = \frac{r}{n} \tag{66.1}$$

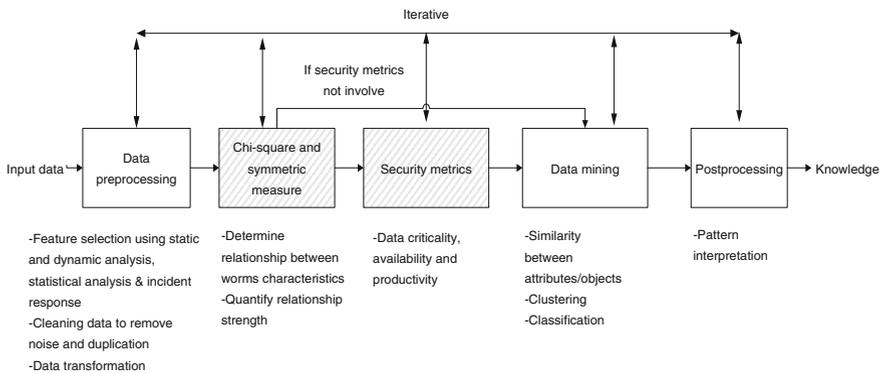


Fig. 66.2 KDD processes integrated with security metrics

Table 66.1 Security Metrics in eZSiber worm response model

Security metrics processes	Applying security metrics in eZSiber
Define worm threats	Yes
Represents worm threats into metrics	Yes. Worm data is represented based on payload, infection, activation, propagation and operating algorithm
Understand and identify the vulnerability, flaw, problem, weakness and damage to security infrastructure	Yes <ul style="list-style-type: none"> • Run the static and dynamic analysis • Identify the need to assign weight and severity value to assign the countermeasure process
Check the performance of the existing countermeasures	Yes <ul style="list-style-type: none"> • Integrate and run data mining using JAVA-WEKA to check the accuracy rate of weight and severity assigned
Recommend any technology or countermeasure process for improvement	Yes Apoptosis to isolate the most severe worm attacks

where: $rf_n = \text{relative frequency}$, $E = \text{number of events}$, $n = \text{total number of experiments conducted}$, $r = \text{number of times an event occurs}$.

Relative frequency is another term for proportion. It is the value calculated by dividing the number of times an event occurs by the total number of times an experiment is carried out. Since the cases involved a long run relative frequency, probability was seen as the best way to calculate the weight. It was in the range of 0–1. The equation is simplified in Eq. 66.2. Based on the frequency analysis, the worm’s attributes are ranked.

$$P(E) = \lim_{n \rightarrow \infty} rf_n(E) \tag{66.2}$$

where, $P(E) = \text{number of outcomes corresponding to event } E/\text{total number of outcomes}$, $rf_n = \text{relative frequency}$.

There are several performance parameters that need to be measured during the experiment which are as in the Eqs. 66.3–66.5. TPR represents the true positive rate, FPR represents as false positive rate, TP represent as true positive, FN represents as false negative, FP represents as false positive and TN represents as true negative.

$$TPR = TP/(TP + FN) \tag{66.3}$$

Table 66.3 Severity Results Using Different Algorithms

Classifier	Multilayer Perceptron			SMO			Naïve Bayes			J48			IBk		
	H	M	L	H	M	L	H	M	L	H	M	L	H	M	L
Severity in (%)	100	83.3	0	100	100	0	99.4	100	0	98.7	83.3	0	100	83.3	0
TPR	16.7	0	0	0	0	0	0	0.6	0	16.7	1.3	0	16.7	0	0
FPR	0	16.7	0	0	0	0	0.6	0	0	1.3	16.7	0	0	16.7	0
FNR	99.38	99.38	0	0	0	0	99.38	99.38	0	98.1	98.13	0	99.38	99.38	0

TPR = true positive rate, FPR = false positive rate, FNR = false negative rate, OA = overall accuracy, H = high, M = medium, L = low

```

Given:
- Set security metrics.
- Set worm attributes: {payload, infection, activation, operating algo-
  rithm and propagation}.
- Set frequency analysis.

Output:
- Weight ranks.
- Severity ranks.
- Weight values.
- Severity values.
- Triggers or halts Apoptosis.

Algorithms:
  1) Apply security metrics to worm attributes.
     a) Go to Weight_cases to determine the weight ranks.
     b) Go to Severity_cases to determine the severity ranks.
  2) Apply frequency analysis to worm attributes.
     a) Go to Frequency_cases to compute the weight and severity values.
  3) Apply apoptosis to Severity_cases.
     a) Go to Apoptosis_cases to trigger the apoptosis.

```

Fig. 66.3 eZSiber worm response model pseudocodes integrated with security metrics

66.5 Conclusions and Future Works

As conclusion, using the security metrics, the severity level of the worm attacks can be easily assigned by using weight. Once the severity level is high, the apoptosis is triggered to avoid the worm from further propagate. The eZSiber worm response model was developed by using the security metrics to assign the weight and the severity of the worm attacks. This model can be used as a reference and comparison by other researchers with the same interests. For future work, different human immunology system (HIS) concept will be tested to the dataset produced from this research. This paper is part of a larger project to build up an automated malware response model. Ongoing research will include other malware classification and the development of software to automate the malware dataset cleanup.

Acknowledgments The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Universiti Sains Islam Malaysia (USIM) grant [PPP/GP/FST/SKTS/30/11912], [PPP/FST/SKTS/30/12812], [PPP/FST/SKTS/30/12812] and Research Management Centre, Universiti Sains Islam Malaysia (USIM).

References

1. Australia Computer Emergency Response Team (AusCERT).: AusCERT Incident Metrics. <https://www.auscert.org.au/render.html?it=17856> (2013). June 2013
2. Malaysia Computer Emergency Response Team (MyCERT).: MyCERT Incident Statistics. <http://www.mycert.org.my/en/services/statistic/mycert/2013/main/detail/914/index.html> (2013)

3. Telegraph Media Group Limited.: Computer worm infects Iran's nuclear station. <http://www.telegraph.co.uk/news/worldnews/middleeast/iran/8026284/Computer-worm-infects-Irans-nuclear-station.html> (2010)
4. Warwick, A.: Ramnit worm steals 45,000 Facebook passwords. <http://www.computerweekly.com/news/2240113383/Ramnit-worm-steals-45000-Facebook-passwords> (2012)
5. Press Trust of India.: Bank details stealing Ramnit malware hits Indian cyberspace. <http://gadgets.ndtv.com/internet/news/bank-details-stealing-ramnit-malware-hits-indian-cyberspace-358719> (2013)
6. Whitty, B.: Why do people create computer Viruses?. Technibble.com. <http://www.technibble.com/why-do-people-create-computer-viruses/> (2007) Accessed 27 Feb 2014
7. Atzeni, A., Lioy, A.: Why to adopt a security metric? A brief survey. In: Gollmann, D., Massacci, F., Yautsiukhin, A. (eds.) *Quality of Protection Security Measurements and Metrics*, pp. 1–12. Springer, Berlin (2006)
8. Saudi, M.M., Cullen, A.J. Woodward, M.: Efficient STAKCERT KDD processes in worm detection, world academy of science. *Eng. Technol. J.* **79**, 453–457 (2011) (pISSN 2010-376X, eISSN 2010-3778)
9. Su, M.-Y.: Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. *Expert Syst. Appl.* **38**(4), 3492–3498 (2011)
10. Siddique, F., Maqbool, O.: Analyzing term weighting schemes for labeling software clusters. In: *Proceedings of 15th European Conference on Software Maintenance and Reengineering*, pp. 85–88 (2011)
11. Kim, H.K., Im, K.H., Park, S.C.: DSS for computer security incident response applying CBR and collaborative response. *Expert Syst. Appl.* **37**, 852–870 (2010)
12. Fisch, D., Hofmann, A., Sick, B.: On the versatility of radial basis function neural networks: a case study in the field of intrusion detection. *Inf. Sci.* **180**(12), 2421–2439 (2010)
13. Middlemiss, M.J., Dick, G.: Weighted feature extraction using a genetic algorithm for intrusion detection. *Proc. Evol. Comput.* **3**, 1669–1675 (2003)
14. Miles, S.G.: Incident response part #2: identification. <http://www.securityhorizon.com/whitepapersTechnical/IncidentResponsepart2.pdf> (2001)
15. Reese, R.L.R.: Incident handling: an orderly response to unexpected events. In: *Proceedings of the 31st Annual ACM SIGUCCS Conference on User Services (SIGUCCS 03)*, pp. 21–24. Texas, USA, Sept 2003, ACM, New York, pp. 97–102 (2003)
16. Heaven, V.X.: Computer virus collection. <http://vxheaven.org/vl.php> (2014)
17. Offensive Computing.: Malware search. <http://www.offensivecomputing.net> (2014)
18. Schultz, M.G., Eskin, E., Zadok, E., Stolfo, S.J.: Data mining methods for detection of new malicious Executables. In: *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, IEEE Computer Society, pp. 38 (2001)
19. Henchiri, O., Japkowicz, N.: A feature Selection and evaluation scheme for computer virus detection. In: *Proceedings of the Sixth International Conference on Data Mining, 2006 (ICDM '06)*, pp. 891–895. IEEE Xplore, Hong Kong (2006)
20. Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., Ian, H.W.: The WEKA data mining software: an update. *SIGKDD Explor.* **11**(1), 10–18 (2009)

Chapter 67

Neural Network Training Algorithm for Carbon Dioxide Emissions Forecast: A Performance Comparison

Herrini Mohd Pauzi and Lazim Abdullah

Abstract Artificial neural network with many types of algorithms is known as an efficient tool in forecasting as it is able to handle nonlinearity behaviour of data. This paper investigates the performances of Levenberg-Marquardt and gradient descent algorithms of back propagation neural networks carbon dioxide emissions forecast. The inputs for the model were selected and the ANNs were trained using the Malaysian data of energy use, gross domestic product per capita, population density, combustible renewable and waste and carbon dioxide intensity. The forecasting performances were measured using coefficient of determination, root means square error, mean absolute error, mean absolute percentage error, number of epoch and elapsed time. Comparison between these algorithms show that the Levenberg-Marquardt was outperformed the gradient descent in carbon dioxide emissions forecast.

67.1 Introduction

It is a well-known fact that carbon dioxide (CO₂) emissions are largely responsible for global warming problems. CO₂ emissions happen when the gas is released into the atmosphere over a specific area and of time either through natural processes or human activities [1]. Due to overzealous of the growing of development over the world, continuing growth of the world population and acceleration of industrialization and urbanization, there are continual increment of CO₂ emissions over these few decades. According to Jana et al. [2], there are two major processes that could reduce the CO₂ emissions: by reducing anthropogenic emissions of CO₂ and by creating and/or enhancing carbon sink in the biosphere.

A considerable amount of literatures have been published to explain the relationship between CO₂ emissions and socio economic variables. One of the possible

H.M. Pauzi · L. Abdullah (✉)

School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia
e-mail: lazim_m@umt.edu.my

methods in estimating the relationship is through forecasting modeling. For example, Brondfield et al. [3] applied a linear model regression downscaling to model on-road CO₂ emissions at Boston, Massachusetts and tested the approach with surface-level CO₂ observations. Another study applied linear regression analysis to determine the strength of correlation between methane (CH₄) and CO₂ concentrations and barometric pressure [4]. On the other hand, Shih and Tsokos [5] have developed two statistical models for predicting the CO₂ emissions and the atmosphere in the United States. A statistical modeling of CO₂ emissions in Malaysia and Thailand constructed by Hui et al. [6], where a few statistical tests were carried out to ensure that the best regression model will be selected for further analysis. However, a major problem with forecasting model is the linearity of data. It has been admitted that the forecasting of the air quality such as CO₂ emissions are sometimes experienced the problems of non-linearity of the selected data.

Apart from linear models, another category of exploring the relationship between CO₂ emissions and causal variables is non linear model. Artificial neural network (ANN) is one of the most widely used models in this category. ANN is a modelling technique that can determine non-linear relationships between variables in input datasets and variables in output datasets [7]. By its learning capability, the required information could be extracted directly from the data [8]. It is an artificial intelligence method which facilitates nonlinear state space through function approximations in constructing a dynamic model. Its ability to handle nonlinearity data could provide better accuracy of forecasting. Numerous studies have attempted to explain the success of ANN with nonlinearity data. The networks have been applied in several areas such as speech recognition [9–11], dynamic modelling [12–14], forecasting in financial field [15–17], monitoring water quality [18], just to name a few. The networks also have been applied in several of fields including gas emissions. Since data of air pollution emissions usually show its non-linearity behaviour, many studies have used the model to forecast air pollutant emissions [7, 19–21].

Literature reviews shown that the gradient descent learning algorithm is the most frequently used as the ANN training algorithm [22–25]. The algorithm lets the network learns the relationship between stipulated input-output data pairs in supervised manner. Although many different types of training algorithms appear in recent neural network literature, it is difficult to know which algorithm works best, in terms of convergence speed and accuracy. There many influential factors of ANN that determine an accuracy of the network. The factors including complexity of problem, number of datasets used in training, number of weights and biases in network, error goal seem to have influence [26]. There are many substantial research have been carried out to alleviate this problem and to date, comparison study is still one of the possible mechanisms for ascertaining which algorithm works best under certain specific cases or problems.

Tongyu et al. [25] conducted a study to recognize transient power based on back propagation (BP) neural network theory. They had applied three different algorithms, the gradient descent, gradient descent with momentum and the LM algorithm to train the BP network. Obviously, they found that the LM outperformed the other two algorithms.

Another researchers, presented multi-layer perceptron model with a novel hybrid training method to perform the forecasting of ozone layer [27]. The training method synergistically couples a stochastic particle swarm optimization algorithm and a deterministic Levenberg–Marquardt (LM) algorithm, which aimed at exploiting the advantage of both. From their comparison of error analysis, they concluded that the multi-layer perceptron with hybrid algorithm was the best training method followed by LM algorithm and particle swarm optimization algorithm.

On the other hand, Ghaffari et al. [28] carried out a study where five different training algorithms belonging to three classes; gradient descent, LM and genetic algorithm were used to train ANN containing a single hidden layer of four nodes. Besides of trying to model the effect of two causal factors for bimodal drug delivery, the researchers were trying to compare the performance of the aforementioned neural network training algorithms. The result of the study showed that back propagation (gradient descent) outperformed the others followed by LM, quick propagation and genetic algorithm.

A study by Piotrowski and Napiorkowski [29], investigated on eight different Evolutionary Computation optimization methods, mostly from Differential Evolution family, to multilayer perceptron neural network training for daily rainfall-runoff forecasting. The overall performance of the LM algorithm and the Differential Evolution with Global and Local Neighbours method for neural networks training turns out to be superior to other Evolutionary Computation-based algorithms.

Based on the mentioned comparison studies, it is observed that the LM optimization must be considered as one of the most efficient algorithms due to its convergence speed. On the other hand, the gradient descent is the most popular and basic training algorithm discussed by researchers. It has been suggested that LM algorithm has better convergence properties than the gradient descent algorithm based on an investigation case of emissions from biodiesel fuelled transit buses [30]. The investigation relies too heavily on coefficient of determinations to compare the performance of conjugate gradient, gradient descent and LM in prediction of biodiesel emissions.

However, the performances of the LM algorithm against the gradient descent algorithm specifically in a case of relationship between CO₂ emissions data and its socio economic and demographic variables are not fully explored. Owing to the facts of the LM and gradient descent algorithm above, this paper seeks to address the issue of forecasting performance between these two ANN algorithms. Performances between these two algorithms will be compared as to decide which algorithm could provide a good forecasting tool of ANN and prevent loss of predictive power for the forecasting.

67.2 Simulation Data

ANNs need a considerable amount of historical data to be trained; upon the satisfactory training an ANN should be able to provide output for previously “unseen” inputs [31]. The selection of input variables for an ANN forecasting

model is a key issue, since irrelevant or noisy variables may have negative effects on the training process, resulting to unnecessarily complex model structure [32].

This study collects annual historical data for Malaysia from 1980 to 2008. The data obtained from the website of World Bank's World Development Indicators. Gross domestic product per capita (GDP) with the unit constant 2000 US\$, energy use in kg of oil equivalent per capita, population density (people per sq. km of land area), combustible renewable and waste (% of total energy), CO₂ intensity (kg per kg of oil equivalent energy use) were selected as the input variables for CO₂ emission model [33].

Construction of the neural networks (NNs) required three processes that are training, validation and testing. Collected data were divided into three datasets. These datasets are defined as training dataset, validation data set and test dataset. The data from 1980 to 2003 was used as the training and validation datasets and the data from 2003 to 2008 was used as testing dataset.

67.3 Design of the Artificial Neural Network Architecture

Two models of ANN were developed in this study. Generally, the architecture of the ANNs are designed based on the input, output, hidden layer and number of node the respective layer. Both of the two models were constructed and designed with the same architecture as to perform a better comparison of performances. A tool for numerical computing known as MATLAB was used to implement the networks. Matlab command line operations were applied to develop the model instead of using Matlab graphic user interfaces (GUIs). Figure 67.1 illustrates the architectures of the two designed models.

67.3.1 Input Layer

The influential variables of CO₂ emissions that have been chosen in this study are fed to the input layer of the ANNs. Hence, the present study was constructed with five nodes in the input layer represented the five variables.

67.3.2 Hidden Layer

According to Ayat et al. [34] higher number of neurons in the hidden layer may derive better learning capability by the network. However, in this research, the authors decided to implement only one hidden layer for the constructed ANNs. The decision was made in order to prevent high computational cost of the networks.

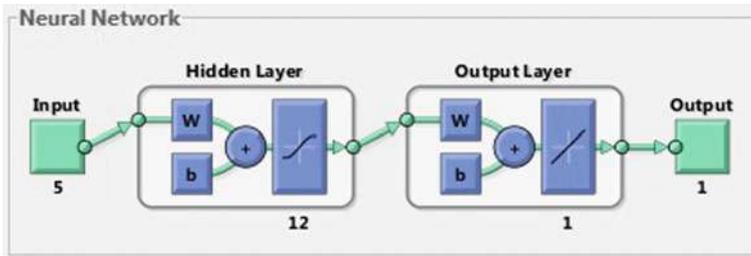


Fig. 67.1 Architecture of neural network models

The authors decided to choose twelve nodes in the hidden layer. The decision is made after considering three different numbers of nodes, 8, 12 and 15. The three different numbers of nodes were randomly picked by the authors and eight numbers of nodes shown the best result among them.

Another important element in this layer is activation function. The function is well known as a transfer function as it transmitted information from one layer to another layer. Tan-sigmoid function was chosen in this study, in which can be expressed as below:

$$a = \text{Tan - sig}(n) = \frac{2}{(1 + \exp(-2 * n))} - 1 \quad (67.1)$$

67.3.3 Output Forms

Output layer consists of one neuron represents our main goal. This study aims to search the best learning algorithm for neural network in CO₂ emissions forecast. Therefore, the CO₂ emissions were considered as the single variable in this layer.

67.4 Results

Results of the performance comparison of testing dataset are presented into two subsections.

67.4.1 Error Analyses

Precision of the forecasting model is examined by calculating three different error measures. The measures are root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). All these measures are expressed as follows:

$$\begin{aligned}
 RMSE &= \sqrt{\sum_{i=1}^n \frac{(x - \hat{x})^2}{n}} \\
 MAE &= \sum_{i=1}^n \frac{|x - \hat{x}|}{n} \\
 MAPE &= \frac{100}{n} * \left| \frac{x - \hat{x}}{\hat{x}} \right|.
 \end{aligned}
 \tag{67.2}$$

where x = actual value, \hat{x} = predicted value of x , n = total number of points.

By using the three error measures and two efficiency indices for each training algorithm, the forecasting performance of LM algorithm against the gradient descent algorithm of NN can be measured. Table 67.1 summarises the performances based on the five measures.

As can be seen in Table 67.1, better precision of prediction were seen with LM as compared with gradient descent. As mentioned previously, the training was terminated when the minimum error of the data was attained. Small values of RMSE, MAE, and MAPE indicate the accuracy of the ANN.

It was found that these values from LM algorithm were about two to three-fold smaller than the gradient descent algorithm (training set: 0.3901 vs. 10.9289; testing set 1.7296 vs. 4.0176). Therefore, the LM algorithm outperformed the gradient descent algorithm.

The number of training epochs and time elapsed for total epochs will reflect the terminated point of training dataset. Number of epochs and CPU time elapsed at the end of the training for gradient descent was obtained as 9 epochs within 30 s. Meanwhile, for LM algorithm, it took 10 epochs within a very short time, 0 s. As can be seen in Figs. 67.2 and 67.3, the gradient descent algorithm needed 6 validation checks and the LM needed only 3 validation checks to stop the training. Based on these values, it is worth mentioning that LM algorithm showed better performance than the gradient descent algorithm.

67.4.2 Regression Analyses

Regression analysis of outputs and targets for LM and gradient descent algorithms during training and testing phase are shown in Fig. 67.4.

The best fit lines in Fig. 67.4 illustrate the relationship between the predicted value (calculated) and observed value (CO₂ emissions). Generally, for evaluation of robustness of a modelling technique, R² of a test data set should be computed and R² greater than 0.9 can be regarded as a good overall fit [35]. It can be interpreted that the value of R² that approaches to 1 shows that the actual values and the output value from ANN model have very high correlation between each other while 0 shows random relationship.

Table 67.1 Error measures and comparison efficiency index of the training algorithm

Performance index	Training algorithms	
	Gradient descent	Levenberg-Marquadt
Training dataset		
RMSE	0.4063	0.0185
MAE	0.3070	0.0102
MAPE (%)	10.9289	0.3901
Testing dataset		
RMSE	0.1182	0.0532
MAE	0.1139	0.0492
MAPE (%)	4.0176	1.7296
Number of epoch at the end of training	9	9
CPU time elapsed at the end of training (s)	30	1

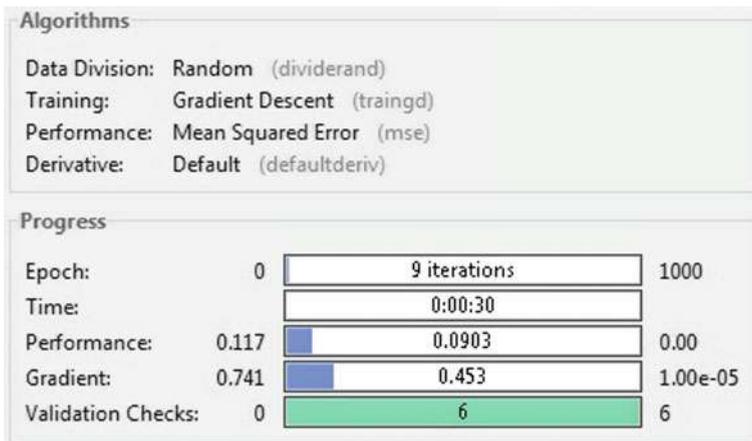


Fig. 67.2 Results of NN model with gradient descent algorithm

From the Fig. 67.4 it can be seen that the R^2 value of NN model with LM algorithm are fitter compared to gradient descent algorithm (0.8921 vs. 0.9783). The gradient descent algorithm achieved lower value of R^2 which is means that this algorithm is lack of ability to train data for the model compared to LM algorithm. Thus higher value of R^2 by LM algorithm indicated that it was successfully trained the data until it can fit the network that has been built. Overall, this suggested that forecasting ability of the NN using LM algorithm for training the data was much better than NN with gradient descent training algorithm in the case of CO₂ emissions in Malaysia.

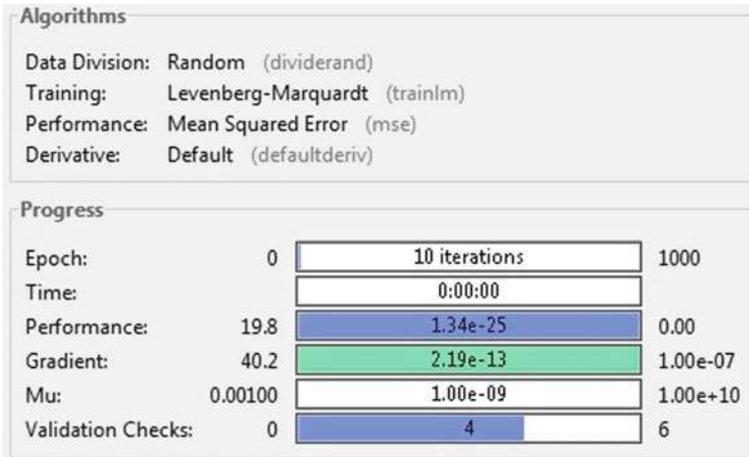


Fig. 67.3 Results of NN model with LM algorithm

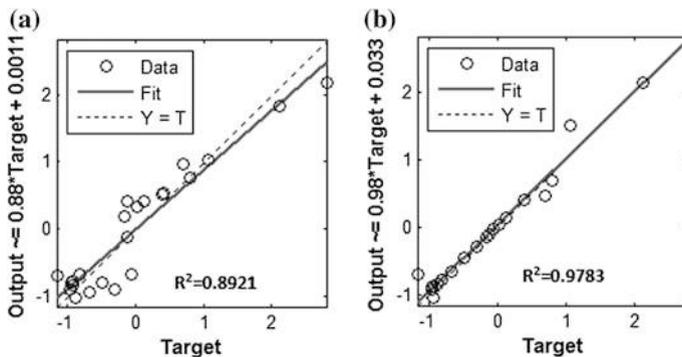


Fig. 67.4 Regression analyses of outputs and targets during training and testing for **a** gradient descent algorithm, **b** LM algorithm

67.5 Conclusions

This paper presented the development of two NN models in application of forecasting annual CO₂ emissions using the socio-economic and demographic parameters. MATLAB tools were implemented to construct the models. The performance of the model using the Levenberg-Marquardt algorithm shows that it can provide better agreement with the collected data of CO₂ emissions than those trained by gradient descent algorithm. This can be concluded based on the observation of error analysis. The RMSE, MAE, MAPE from Levenberg-Marquardt algorithm are smaller than the values from gradient descent algorithm.

The presented ANN models not only can be used for CO₂ emissions forecasting, but also for simulating various scenarios of CO₂ emissions by changing the parameters for the input variables. The results obtained from this research could be useful for the governments and authorities to take any initiatives or preventions in order to keep the sustainability of air quality. Deciding on the appropriate algorithm with higher efficiency is a worth effort to better dealt with environmental assessment. Another featured environmental quality indicators could be considered in further researches. Improvement of the forecasting models using other computational intelligence tools such as adaptive neuro-fuzzy inference system could be explored in future research. Malaysia.

Acknowledgments The authors are grateful to the Malaysian Ministry of Higher Education and University Malaysia Terengganu for financial support under the FRGS grant number 59243.

References

1. Hui, T.S., Rahman, S.A., Labadin, J.: An empirical study on CO₂ emissions in ASEAN countries. In: International Conference on Statistics in Science, Business, and Engineering (ICSSBE), pp. 1–6 (2012)
2. Jana, B.K., Biswas, S., Majumder, M., Roy, P., Mazumdar, M.: Estimation of Carbon Dioxide Emission Contributing GHG Level in Ambient Air of a Metro City: A Case Study for Kolkata, Impact of Climate Change on Natural Resource Management, (pp. 3–18). Springer, Netherlands (2010)
3. Brondfield, M.N., Hutyra, L.R., Gately, C.K., Raciti, S.M., Peterson, A.: Modeling and validation of on-road CO₂ emissions inventories at the urban regional scale. *Environ. Pollut.* **70**, 123–133 (2012)
4. Nwachukwu, A.N., Anonye, D.: The effect of atmospheric pressure on CH₄ and CO₂ emission from a closed landfill site in Manchester, UK. *Environ. Monit. Assess.* **185**, 5729–5735 (2012)
5. Shih, S.H., Tsokos, C.P.: Prediction models for carbon dioxide emissions and atmosphere. *J. Neural Parallel Sci. Comput.* **16**, 165–178 (2008)
6. Hui, T.S., Rahman, S.A., Labadin, J.: Statistical modeling of CO₂ emissions in Malaysia and Thailand. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2**, 10–15 (2012)
7. Antanasijević, D.Z., Pocajt, V.V., Povrenović, D.S., Ristić, M.Đ., Perić-Grujić, A.A.: PM¹⁰ emission forecasting using artificial neural networks and genetic algorithm input variable optimization. *Sci. Total Environ.* **443**, 511–519 (2012)
8. Manjunatha, R., Narayana, P.B., Redy, K.H.C.: Application of artificial neural networks for emission modeling of biodiesels for a C.I engine under varying operating conditions. *Mod. Appl. Sci.* **4**, 77–89 (2010)
9. Dede, G., Sazli, M.H.: Speech recognition with artificial neural networks. *Digit. Signal Proc.* **20**, 763–768 (2010)
10. Sabato, M.S., Yu, D., Deng, L., Lee, C.H.: Exploiting deep neural networks for detection-based speech recognition. *Neurocomputing* **106**, 148–157 (2013)
11. Ting, H.N., Yong, B.F., Mirhassani, S.M.: Self-adjustable neural network for speech recognition. *Eng. Appl. Artif. Intell.* **26**(9), 2022–2027 (2013)
12. Lamrini, B., Valle, G.D., Trelea, I.C., Perrot, N., Trystram, G.: A new method for dynamic modelling of bread dough kneading based on artificial neural network. *Food Control* **26**(2), 512–524 (2012)

13. Li, N., Xia, L., Shiming, D., Xu, X., Chan, M.Y.: Dynamic modeling and control of a direct expansion air conditioning system using artificial neural network. *Appl. Energy* **91**, 290–300 (2012)
14. Timothy Hong, Y.S.: Dynamic nonlinear state-space model with a neural network via improved sequential learning algorithm for an online real-time hydrological modeling. *J. Hydrol.* **468**(469), 11–21 (2012)
15. Wang, J.J., Wang, J.Z., Zhang, Z.G., Guo, S.P.: Stock index forecasting based on a hybrid model. *Omega* **40**, 758–766 (2012)
16. Venkatesh, K., Ravi, V., Prinzie, A., Van den Poel, D.: Cash demand forecasting in ATMs by clustering and neural networks. *Eur. J. Oper. Res.* **232**(2), 383–392 (2014)
17. Yu, L., Wang, S., Lai, K.K.: A multiscale neural network learning paradigm for financial crisis forecasting. *Neurocomputing* **73**, 716–725 (2010)
18. Xu, L., Liu, S.: Study of short-term water quality prediction model based on wavelet neural network. *Math. Comput. Model.* **58**(3–4), 807–813 (2013)
19. Kurt, A., Oktay, A.B.: Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Syst. Appl.* **37**, 7986–7992 (2010)
20. Şahin, Ü.A., Bayat, C., Uçan, O.N.: Application of cellular neural network (CNN) to the prediction of missing air pollutant data. *Atmos. Res.* **101**(1–2), 314–326 (2011)
21. Fernando, H.J.S., Mammarella, M.C., Grandoni, G., Fedele, P., Di Marco, R., Dimitrova, R., Hyde, P.: Forecasting PM¹⁰ in metropolitan areas: efficacy of neural networks. *Environ. Pollut.* **163**, 62–67 (2012)
22. Yaghini, M., Khoshraftar, M.M., Fallahi, M.: A hybrid algorithm for artificial neural network training. *Eng. Appl. Artif. Intell.* **26**, 293–301 (2013)
23. Zazoun, R.S.: Fracture density estimation from core and conventional well logs data using artificial neural networks: the Cambro-Ordovician reservoir of Mesdar oil field, Algeria. *J. Afr. Earth Sci.* **83**, 55–73 (2013)
24. Nakama, T.: Theoretical analysis of batch and on-line training for gradient descent learning in neural networks. *Neurocomputing* **73**(1–3), 151–159 (2009)
25. Tongyu, X., Wei, Z., Peng, S., Qin, Z.: Transient power quality recognition based on BP neural network theory. *Energy Procedia* **16**, Part B, 1386–1392 (2012)
26. Coskun, N., Yildirim, T.: The effects of training algorithm in MLP network on image classification. In: Proceedings of International Joint Conference on Neural Network
27. Wang, D., Lu, W.Z.: Forecasting of ozone level in time series using MLP model with a novel hybrid training algorithm. *Atmos. Environ.* **40**, 913–924 (2006)
28. Ghaffari, A., Abdollahi, H., Khoshayand, M.R., Bozchalooi, I.S., Dadgar, A., Tehrani, M.R.: Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *Int. J. Pharm.* **327**(1–2), 126–138 (2006)
29. Piotrowski, A.P., Napiorkowski, J.J.: Optimizing neural networks for river flow forecasting—Evolutionary Computation methods versus the Levenberg–Marquardt approach. *J. Hydrol.* **407**, 12–2 (2011)
30. Mudgal, A., Gopalakrishnan, K., Hallmark, S.: Prediction of emissions from biodiesel fueled transit buses using artificial neural networks. *Int. J. Traffic Transp. Eng.* **1**, 115–131 (2011)
31. Palani, S., Liong, S.Y., Tkalich, P.: An ANN application for water quality forecasting. *Mar. Pollut. Bull.* **56**, 1586–1597 (2008)
32. Voukantsis, D., Karatzas, K., Kukkonen, J., Rasinen, T., Karppinen, A., Kolehmainen, M.: Intercomparison of air quality data using principal component analysis and forecasting of PM₁₀ and PM_{2.5} concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* **409**, 1266–1276 (2011)
33. World Bank: World Development Indicators. <http://data.worldbank.org/country/malaysia>
34. Ayat, S., Farahani, H.A., Aghamohamadi, M., Alian, M., Aghamohamadi, S., Kazemi, Z.: A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. *Neural Comput. Appl.* **23**, 1381–1386 (2013)
35. Bourquin, J., Schmidli, H., Hoogevest, P., Leuenberger, H.: Application of artificial neural networks (ANNs) in the development of solid dosage forms. *Pharm. Dev. Technol.* **2**, 111–121 (1997)

Chapter 68

Theorem Prover Based Static Analyzer: Comparison Analysis Between ESC/Java2 and KeY

Aneesa Saeed and S.H.A. Hamid

Abstract Software developers utilize static analyzers to discover the defects in the software source code. One of the static analyzer categories is based on theorem prover. Due to the strength of the theorem prover in proving the programs correctness and soundness without producing false warnings, analysis on the verification tools is important to assess the performance. The objective of this paper is to analyze the performance of the open source theorem prover based static analyzers for Java. The analysis is done by comparing two static analyzers namely KeY and ESC/Java2 using four evaluation metrics for 20 test cases developed based on Common Weakness Enumeration (CWE). The result shows the performance of KeY is better than ESC/Java2 especially on Detection Coverage. The analysis also presents that the performances of theorem prover based static analyzers are effective on small developed benchmark only.

Keywords Static analysis · Theorem prover · Verification · KeY · ESC/java2 · Software testing

68.1 Introduction

Testing becomes an important step in software engineering lifecycle because it typically consumes 40–50 % from development efforts [1]. Static analysis is one of testing strategies that detects defects automatically for assessing code quality without running the code. Static analysis tools inspect automatically software

A. Saeed (✉) · S.H.A. Hamid
Faculty of Computer Science and Information Technology, University of Malaya,
Kuala Lumpur, Malaysia
e-mail: aneesa.saeed@siswa.um.edu.my

S.H.A. Hamid
e-mail: sitihaifah@um.edu.my

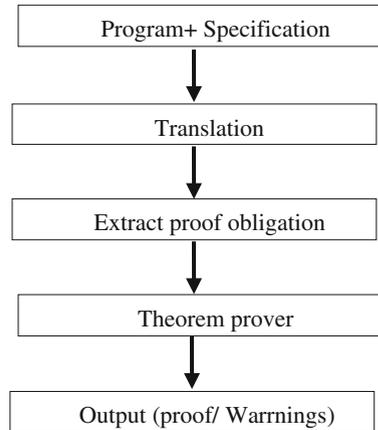
vulnerabilities such as redundant code, unsuitable use of variables, uncaught runtime exceptions and division by zero. Static analyzers automatically identify the software defects by parsing and analyzing the source code. The purpose is to search for a fixed number of patterns in the code. Static analyzers analyze the source code using control flow analysis, interface analysis, information flow analysis, data flow analysis and path analysis. The use of the static analyzers are increased currently in the software development process as a fault detector [2]. The powerful of the static analyzers come from their ability to detect the defects in the code which they are hardly discovered during execution [3]. There are some programming errors that hardly caught by static analyzers [4], for example, Basic cross-site scripting (XSS) vulnerability [2]. Static analyzers have six categories based on internal design schemes as discussed more in Sect. 68.3. The categories are type checking, style checking, program understanding, program verification (theorem prover) and property checking, bug finder and security review. Verification tools have progressed noticeably during the last years due to the advances in verification aspects which are: the specification, program verification and theorem prover methods.

Theorem prover based verification tools take the programs with their specifications as an input and prove the codes are matched with the specification. The general framework is described in Fig. 68.1. Several verifiers have been developed rely on theorem prover for Java, C, and C#. Examples for these static analyzers are: for Java; ESC/Java2 [5], KeY [6], Jive [7], LOOP [8] and Jstar [9], for C; veriFast [10] and Frama-C [11], and for C#.Net; Boogie [12]. Due to the strength of the theorem prover, analysis on the verification tools is important to assess the performance. The ESC/Java2 and KeY are common open source static analyzers based on theorem prover that have reported successful results in [13, 14]. KeY was used to verify the correctness of such resource guarantees. The results show that the KeY caught all the bugs except one bug [13]. Results of using ESC/Java2 to verify tally subsystem of the Dutch Internet voting system were presented well in [12]. In few comparison studies [15–17], they focused on the KeY and ESC/Java2 separately with insufficient results, only searching on proofing time with KeY and on reduced small set of vulnerabilities with ESC/Java2.

The objective of this study is to analyze the performance of the open source theorem prover based static analyzers for Java. The analysis is done by comparing two static analyzers namely KeY and ESC/Java2. We developed a small benchmark based on the CWE. The objective evaluation is used to assess the performance of the selected static analyzers. The objective evaluation is measures the Detection Coverage, the number of defects that are correctly and incorrectly detected (True and False positives respectively) and the F-measure.

The rest of this paper is organized as follows: Sect. 68.2 presents the previous related work. Section 68.3 provides a background on static analyzers with their classification. The methodology that we follow in this research is discussed in Sect. 68.4. Section 68.5 discusses our experiments and finding. Finally, conclusion and future work are presented in Sect. 68.6.

Fig. 68.1 Framework of verification tools based on theorem prover



68.2 Related Work

This section presents the previous works related to the comparison studies on the performance of the static analyzers. Beckert et al. [18] studied two verification tools which are VCC and KeY. The evaluation on the two static analyzers is conducted by using test suites provided by each tool developers. A new coverage criterion has been investigated to test the tool correctness depends on the axiomatization. Using the test suites that are written by the tool developers themselves is not enough to the evaluation.

Study reported by Collavizza et al. [15] presented a comparison between CPBVP with the ESC/Java, CBMC, Blast, EUREKA and Why verification tools. The analysis was done against six small programs. However, the small program is not efficient to evaluate the tools performance especially current systems are scalable.

Rutar et al. [17] analyzed five open source tools (Bandera, ESC/Java2, FindBugs, JLint, and PMD) against small test suite that involved common for Java. The authors investigated a meta-tool to allow developers to recognize the different classes of vulnerabilities. The experimental results showed that the tools find non-overlapping vulnerabilities.

In educational view, Feinerer and Salzer [16] compared four tools (namely the Frege Program Prover, the KeY system, Perfect Developer, and the Prototype Verification System (Pvs)) to check their suitability for teaching formal software verification. They evaluated the tools against a suite of small programs, which are typical of courses dealing with Hoare-style verification, weakest preconditions, or dynamic logic. Finally they report their experiences by Perfect Developer in class.

In other work in [2], they have compared the performance of nine static analyzers in term of detecting security vulnerabilities. The tools are CBMC, K8-Insight, PC-lint, Prevent, Satabs, SCA, Goanna, Cx-enterprise and Codesonar. Seven of them are commercial tools but having different designs. The tools are based on model checking and syntactic analysis. They conducted the experiments

against two dataset test suites (45, and 46) from SAMATE Reference Dataset for C language. One includes test cases (TC) with known vulnerabilities and the other one is designed with specific vulnerabilities fixed. The results are objectively assessed by using a set of well known evaluation metrics (Recall, Precision and F-measure). Some recommendations for improving the reliability and usefulness of static analyzers and the process of benchmarking are recommended.

In [19] 11 software contract technologies were explored namely JML, IContract, Contract Java, Handshake, Jass, JContractor, JMsassert, Spec#, Code Contracts, Eiffel and OCL. They presented the similarities with the areas of significant disagreement and highlight the shortcomings of existing technologies. They briefly introduced PACT, a software contract tool under development, explaining its approach to various aspects of software contracts.

The previous related works concerned on the tools based on verification, lexical and syntactic analysis against small programs except the study by Díaz and Bermejo [2]. They analyzed static analyzers for C against standard benchmark. To our best knowledge, this research is the first comparison studies that focusing on KeY theorem prover static analyzer as fault detector. Moreover, well-known evaluation metrics in terms of Recall, Precision and F-measure are used for the true performance assessment.

68.3 Static Analyzers

68.3.1 Background

All static analyzers follow the same steps for analyzing source code as the following: (1) transforming the code into set of data structures called program model, (2) analyzing the model using different rules and/or properties, (3) showing the results to the static analyzer. Code transforming is done via a mixture of various techniques as lexical and semantic analysis, abstract syntax and parsing. Intraprocedural (local) and interprocedural (global) analysis are used for analyzing the program model to assess the individual functions and their interactions by tracing control flow and data flow, pointer aliasing, etc... The rules can be fixed or extended based on the selected tool. Finally the results of the selected tools are concise, without more information about the possible existed defects, or more detailed. The differences in the information of the output lead to hard comparison among the different tools.

68.3.2 Static Analyzers Categories

The static analyzers are classified based on different schemes. According to [20], the categorization can be based on the language or the defects types. The schema presented by Chess and West [3] utilized general purpose of the tools. This schema is the most relevant one to our work. In next paragraphs, we discuss each category in the schema.

Type checking tools Type checking is used by programmers to remove whole categories of programming mistakes. For example, it prevents assigning integral values into other variables types. Because of detecting errors at compile stage, the type checking based tools avoid runtime errors. However, these tools suffer from the false positive (FP) and false negative (FN).

Style checking tools This kind of tools achieves checking based on lexical and syntactic analysis to detect the code problems as inconsistencies in function calls such as calling the function with changeable number of attributes. Although, the produced errors often affect the readability and the maintainability of the code, these tools do not refer that the specific defect will happen in the execution time. Therefore, these tools have limitations compared with others due to they perform the analysis without simulating what happens in runtime. One Example of style checking based static analyzers is the PMD tool for Java [21].

Program Understanding Tools These tools are involved in several Integrated Development Environments and are designed to support programmers to gain vision about how the program works. The reviewers use these tools to understand the large code and detect the security vulnerabilities. Nevertheless, using these methods is time consuming. One example of such tools is Fujaba tool [22].

Bug Finding Tools Bug Finding tools are not concern on formatting issues as style checker nor are doing matching between the code and its specification. These tools basically notify about places in the code where the program will act in a different way that it is not desired by the programmer. Bug Finding tools have two major features. The first feature is easy to use due to these tools have a set of rules describing the patterns in the code that usually specify security vulnerabilities. The second feature is scalability where they are usually used for analyzing large applications code. Less number of is produced by using bug finders. FindBugs [23] and Coverity [24] are examples of Find Bug tools.

Security Review Tools These tools designed as a hybrid of property checkers and bug finding technique. However, they concern on identifying specific security vulnerabilities. Due to the many security properties can be expressed as program properties. The thoughtful side of the balance between the number of FP and FN motivates the researchers to use these tools. Furthermore, many points in the code are shown that will be manually reviewed after applying the tools. These tools producing more FP compared with bug finder. Ounce 6 and SCA are the two most relevant examples of this category.

Theorem Proven Verification and Property Checking Tools Using these kinds of tools requires the code and the specification. The verification tools (based on theorem prover) attempt to prove that the code meets the specification. Verification tools sometimes go by the name property checking when they check the code against a fractional specification that details only part of the program behavior. The majority of property checking tools work either by applying logical inference or by performing model checking. The programmers rarely insert the specification that is detailed enough to be used for equivalence checking. Writing specification needs more time and effort than writing the code, this is the main limitation. These

tools never suffer from FP. Numbers of verifiers are available, for examples CBMC, KeY, ESC/Java2, or Satabs tools.

Unfortunately, there are some FP or FN produced by all static analyzers. Fortunately, some of theorem prover verification tools do not suffer from false positives.

68.4 Methodology

In this section, we present the methodology that used to conduct the analysis study.

68.4.1 Select Tools

In this study, we selected KeY and ESC/Java2 which are well-known verification tools for program static analysis. The performance evaluation is important to give overall insights to programmers to choose suitable verification tools. Both tools are open source that their complete code and documentation are available. This can make us understand better their limitations. They need specifications written in Java Modeling language (JML).

KeY ([25]): KeY is a formal software development tool that aims to integrate design, implementation, formal specification, and formal verification of object-oriented software as seamlessly as possible. At the core of the system is a novel theorem prover for the first-order dynamic logic for Java with a user-friendly graphical interface. KeY is sound and complete. The framework of KeY is shown in Fig. 68.2.

ESC/Java2 ([26]): Extended Static Checker for Java (ESC/Java2) is a programming tool that aims to detect common run-time errors in JML-annotated Java programs by static analysis of the program code and its formal specifications. The main limitation of this tool is it is neither sound nor complete. Therefore, false positive and negatives may occur. Figure 68.3 shows the frame work of the ESC/Java2.

68.4.2 Select Test Suites

The best approach to compare and evaluate the static analyzers is by analyzing the same standard benchmark. The input of the theorem based tools must be the source code and the specification. Finding suitable benchmark for conducting experiments was complicated. This is because there are no standard benchmark test suites containing the code in Java with the JML specification. In this study, we developed our benchmark depending on the common code defects that described in

Fig. 68.2 KeY framework [6]

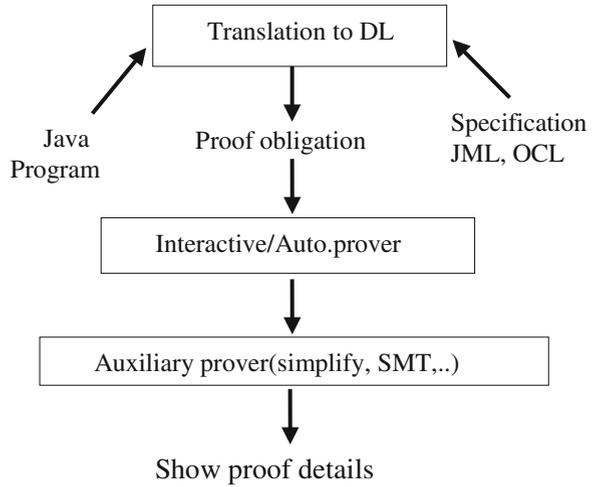


Fig. 68.3 ESC/Java2 framework [27]

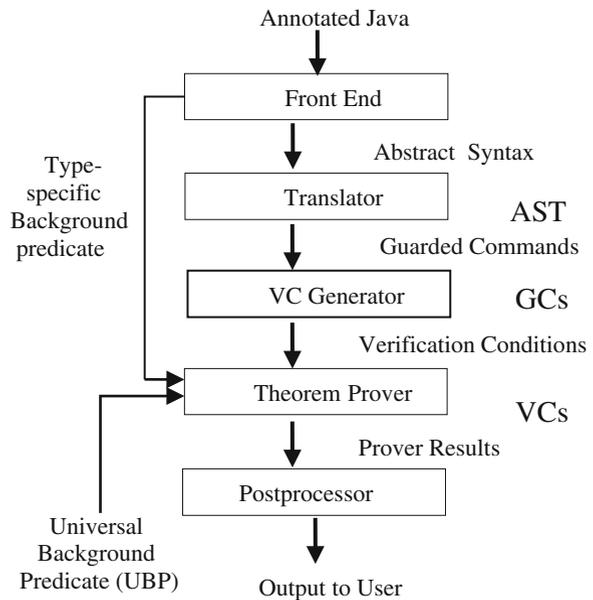


Table 68.1. We used and modified some TC (test cases) that developed by KeY developers, Samate Reference Dataset and the TC used in [15] study. Table 68.1 illustrates the code defects based on CWE from MITRE corporation [20, 28]. The defects names, number of the TC under each defect and small descriptions on the defects are presented in Table 68.1.

Table 68.1 Common defects in the source code [20, 28]

Common defects	TC	Description
Divide by zero	2	An input divides by zero can lead undefined behavior
Heap overflow	2	Input is used in an argument to the creation or copying of blocks of data beyond the fixed memory boundary of a buffer in the heap portion of memory
Unchecked loop condition	2	No checking for loop conditions that potentially leading to a denial of service because of excessive looping
Infinite loop	2	The number of loop repetition is infinite where the exit condition are not be reached
Integer overflow	2	A numeric value that is too large to be represented within the available storage space
Null dereferences	2	A pointer with a value of NULL is used as though it pointed to a valid memory area
Array index out of bound	2	Try to read or write data into an index of the array that doesn't exist
Unused variable	2	A variable is created without using it. It is subsequently referenced in the program, causing potential undefined behavior or denial of service
Uninitialized variable	2	A variable is created without assigning it a value. It is subsequently referenced in the program, causing potential undefined behavior or denial of service
Unchecked error condition	2	No action is taken after an error or exception occur

68.4.3 Evaluation Metrics

Using only the defects Detection Coverage percentage is not enough for real assessment for the tools performance. Due to the static analyzers suffer from FP and FN. FN and FP are not covered in Detection Coverage percentage. Therefore, our evaluation covers the F-measure [29]. The F-measure represents the harmonic mean of Recall and Precision measurements which can be calculated as the following:

Precision is the ratio of the number of correct discovered defects to the number of all detected defects and can be calculated as:

$$P = \frac{TP}{TP + FP} \quad (68.1)$$

Recall is the ratio of the number of correct detected defects to the number of the known defects and can be calculated as:

$$R = \frac{TP}{TP + FN} \quad (68.2)$$

F-measure is the harmonic mean of Recall and Precision and can be calculated as

$$\text{F-measure} = \frac{2 * P * R}{P + R} \quad (68.3)$$

Defects Detection Coverage is the ratio of the number of detected defects to the number of all defects in the benchmark and can be calculated as:

$$\text{Defects Detection Coverage} = \frac{\text{TP} + \text{FP}}{\text{Number of all defects}} \quad (68.4)$$

where TP (true positive): is the number of true detected defects, FP is the number of detected defects that in fact do not exist and FN is the total number of defects not detected in the code.

68.5 Experiments

In this experiment, we have run the two static analyzers as Eclipse plug-in in Intel Pentium 4 CPU 3.20 GHz 3.19 GHz, 4.0 GB ram and 32-bit operating system. We used our benchmark with 20 TC written in Java language. The complexity of the TC ranges from small Java programs to more complex programs (included many complex properties like inheritance). Each kind of defects in the benchmark has two TC. KeY 2.0.1 is the latest stable version released on June 20th, 2013 and ESC/Java2 2.0.5 is the latest version released on 8 November 2008 were used to conduct the experiments.

We calculated the detection percentage of each static analyzer. Table 68.2 illustrates the summary of the execution results of the two static analyzers against 20 TC. TP means the analyzer correctly detect this number of defects. Fails (FN) means the analyser could not detect the existing defects in the benchmark. FP refers to the number of detected defects but they are not real defects. The static analyzers suffer from FP so depending on the Detection Coverage is not enough to assess correctly the static analyzers performance. Furthermore, we calculate the Recall, Precision and F-measure for each static analyzer. The recall rate refers to the percentage of all the known defects in the benchmark are detected and the percentage calculated by $(1 - \text{recall}) * 100$ from the defects are uncovered. Precision expresses the chance of the defects to be detected correctly.

We conclude from the Table 68.2 that the performance of the KeY is more effective than the ESC/Java2 in terms of all evaluation metrics. The effective performance of KeY means it catches the known defects without producing any FP. All the defects are detected by KeY except divide by zero and unused variable. The KeY obtains value of 1 compare to ESC/Java2 which only receives 0.4. The difference of 0.6 result value shows that all defects are detected and correctly covered in KeY.

Table 68.2 Results of KeY and ESC/Java2 execution

Static analyzers	Result summary				Precision	Detection coverage	F-measure
	Good (TP)	FP	Fails (FN)	Recall			
ESC/Java2	6	9	5	0.546	0.4	30	0.4617
KeY	15	0	5	0.75	1	75	0.8571

The variation between the obtained recall values is 0.204 that indicates less number of defects is missed by KeY compared to ESC/Java2. The KeY does not produce FP in all the TC. Comparing the number of true positives (6) is less than the number of false warnings (9) by ESC/Java2. ESC/Java2 detects less number of defects with producing many FP. The ESC/Java2 needs less time and easy to trace for the warning signs compared to KeY which needs more time for tracing the errors and warnings. Using the KeY for verification complex programs will prove correctly the programs correctness and soundness but it takes long time that consumes the developing cost (in term of time). In our opinion, using one static analyzer is not enough to analyze the code and discover all the defects. Analyzing the code with many static analyzers guarantees the most defects in the code will be discovered.

68.6 Conclusions and Future Work

In this study, we presented a comparison study between KeY and ESC/Java2 as fault detectors on our benchmark. An objective evaluation has been conducted in terms of Recall, Precision and F-measure to assess correctly the static analyzers performance. Using KeY verification tool as fault detector obtained better Detection Coverage without producing FP rather than ESC/Java2. We believe that using theorem prover based static analyzers are a promising way in checking statically the code vulnerabilities. For future, we plan to study the performance of all the theorem prover based static analyzers against big benchmarks that covered more common defects.

Acknowledgments The authors acknowledge the support provided by University of Malaya Research Grant, RG10612ICT.

References

1. Luo, L.: Software Testing Techniques. Institute for software research international Carnegie Mellon University, Pittsburgh, 19 p (2001)
2. Díaz, G., Bermejo, J.R.: Static analysis of source code security: assessment of tools against SAMATE tests. *Inf.Softw. Technol.* **55**(8), 1462–1476 (2013)
3. Chess, B., West, J.: *Secure Programming with Static Analysis*. Pearson Education, Boston (2007)

4. Young, M., Taylor, R.N.: Rethinking the taxonomy of fault detection techniques. In: Proceedings of 11th international conference on Software engineering, pp. 53–62. ACM (1989)
5. Cok, D.R., Kiniry, J.R.: Esc/java2: Uniting esc/java and jml. Construction and Analysis of Safe, Secure, and Interoperable Smart Devices, **3362**, pp. 108–128. Springer (2005)
6. Beckert, B., Hähnle, R., Schmitt, P.H.: Verification of Object-Oriented Software: The KeY Approach. Springer, Berlin, Heidelberg (2007)
7. Meyer, J., Poetzsch-Heffter, A.: An architecture for interactive program provers. Tools and Algorithms for the Construction and Analysis of Systems, **1785**, pp. 63–77. Springer (2000)
8. Jacobs, B., Poll, E.: Java program verification at Nijmegen: Developments and perspective. Software Security-Theories and Systems, **3233**, pp. 134–153. Springer (2004)
9. Distefano, D., Parkinson, J., jStar, M.J.: Towards practical verification for Java. ACM SIGPLAN Not. **43**, 213–226 (2008)
10. Jacobs, B., Smans, J., Piessens, F.: VeriFast: Imperative programs as proofs. In: VSTTE Workshop on Tools and Experiments. (2010)
11. Cuoq, P., Kirchner, F., Kosmatov, N., Prevosto, V., Signoles, J., Yakobowski, B.: Frama-C. Software Engineering and Formal Methods, pp. 233–247. Springer (2012)
12. Barnett, M., Chang, B.-Y.E., DeLine, R., Jacobs, B., Leino, K.R.M.: Boogie: a modular reusable verifier for object-oriented programs. In: Formal Methods for Components and Objects, **7504**, pp. 364–387. Springer (2006)
13. Cok, D., Kiniry, J.: ESC/Java2: Uniting ESC/Java and JML. In: Barthe, G., Burdy, L., Huisman, M., Lanet, J.-L., Muntean, T. (eds.) Construction and Analysis of Safe, Secure, and Interoperable Smart Devices, vol. 3362, pp. 108–128. Springer, Berlin Heidelberg (2005)
14. Albert, E., Bubel, R., Genaim, S., Hähnle, R., Puebla, G., Román-Díez, G.: Verified resource guarantees using COSTA and KeY. In: Proceedings of the 20th ACM SIGPLAN Workshop on Partial Evaluation and Program Manipulation. ACM, pp. 73–76 (2011)
15. Collavizza, H., Rueher, M., Van Hentenryck, P.: Comparison between CPBPV, ESC/JAVA, CBMC, BLAST, EUREKA and WHY for bounded program verification. arXiv preprint arXiv:0808.1508 (2008)
16. Feinerer, I., Salzer, G.: A comparison of tools for teaching formal software verification. Formal Aspects Comput. **21**, 293–301 (2009)
17. Rutar, N., Almazan, C.B., Foster, J.S.: A comparison of bug finding tools for Java. In: Software Reliability Engineering. 5th IEEE International Symposium on ISSRE, pp. 245–256 (2004)
18. Beckert, B., Borner, T., Wagner, M.: A Metric for Testing Program Verification Systems, **7942**, pp. 56–75. Tests and Proofs. Springer (2013)
19. Voigt, J., Irwin, W., Churcher, N.: Comparing and Evaluating Existing Software Contract Tools. Evaluation of Novel Approaches to Software Engineering, **275**, pp. 49–63. Springer (2013)
20. Zheng, J., Williams, L., Nagappan, N., Snipes, W., Hudepohl, J.P., Vouk, M.A.: On the value of static analysis for fault detection in software. IEEE Trans. Softw. Eng. **32**, 240–253 (2006)
21. Copeland, T.: PMD Applied. Centennial Books, San Francisco (2005)
22. Burmester, S., Giese, H., Niere, J., Tichy, M., Wadsack, J.P., Wagner, R., Wendehals, L., Zündorf, A.: Tool integration at the meta-model level: the Fujaba approach. Int. J. Softw. Tools Technol. Transfer **6**, 203–218 (2004)
23. Hovemeyer, D., Pugh, W.: Finding bugs is easy. ACM SIGPLAN Not. **39**, 92–106 (2004)
24. Almassawi, A., Lim, K., Sinha, T.: Analysis Tool Evaluation: Coverity Prevent. Carnegie Mellon University, Pittsburgh (2006)
25. <http://www.key-project.org/>
26. <http://www.kindsoftware.com/products/opensource/ESCJava2/download.html>
27. Flanagan, C., Leino, K.R.M., Lillibridge, M., Nelson, G., Saxe, J.B., Stata, R.: Extended static checking for Java, **37**(5), pp. 234–245. ACM SIGPLAN Notices. ACM (2002)
28. <http://cwe.mitre.org>
29. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)

Chapter 69

Designing a New Model for Trojan Horse Detection Using Sequential Minimal Optimization

Madihah Mohd Saudi, Areej Mustafa Abuzaid, Bachok M. Taib and Zul Hilmi Abdullah

Abstract Malwares attack such as by the worm, virus, trojan horse and botnet have caused lots of troublesome for many organisations and users which lead to the cybercrime. Living in a cyber world, being infected by these malwares becoming more common. Nowadays the malwares attack especially by the trojan horse is becoming more sophisticated and intelligent, makes it is harder to be detected than before. Therefore, in this research paper, a new model called Efficient Trojan Detection Model (ETDMo) is built to detect trojan horse attacks more efficiently. In this model, the static, dynamic and automated analyses were conducted and the machine learning algorithms were applied to optimize the performance. Based on the experiment conducted, the Sequential Minimal Optimization (SMO) algorithm has outperformed other machine learning algorithms with 98.2 % of true positive rate and with 1.7 % of false positive rate.

Keywords Malwares · Trojan horse · Detection · Automated analysis · Sequential minimal optimization (SMO) · True positive rate · False positive rate · Machine learning

M. Mohd Saudi (✉) · A.M. Abuzaid · B.M. Taib · Z.H. Abdullah
Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
71800 Bandar Baru Nilai, Negeri Sembilan, Malaysia
e-mail: madihah@usim.edu.my

A.M. Abuzaid
e-mail: abuzaid.areeg88@gmail.com

B.M. Taib
e-mail: bachok@usim.edu.my

Z.H. Abdullah
e-mail: zulhimi.a@usim.edu.my

69.1 Introduction

Trojans have been improved and ranges of sophisticated techniques have been integrated, which make the detection processes much harder and longer than in the past. Lack of understanding and knowledge and proper procedures of the trojan analysis have led to money loss, reduced productivity and tarnishing of organization's reputations [1]. Nowadays there are so much information about vulnerabilities and exploitation in operating systems and applications, attacks and defense technologies. Though information regarding these widely distributed and different defense technologies has been implemented, many organizations constantly being attacked and exploited by malwares such as worm and trojan horse. Nowadays, the computer network systems are more threatened by trojan horse. This is mostly true for Windows system because there are large numbers of trojan horses designed to enable attacks upon running Windows platforms. Furthermore, new trojan horses emerge almost every day. Trojan horse is defined as a malicious program, that must be executed on a victim's computer and once it is installed, it can control the victim's computer remotely and steal any confidential information from it. It is different compared to worm and virus, since it has the capability to control the victim's computer remotely and it does not replicate itself [2].

On May 2012, the Flame trojan has infected thousands of computers all over the world and it has been described as one of the most complex threats ever discovered. It has the capabilities to take screenshots secretly, record audio and sends this information to its creator via encrypted channel. It caused chaos and loss of money and productivity [3]. In the year of 2013, the Australia Computer Emergency Response Centre (AuSCERT) reported that 7,962 cases of the compromised Australia web sites were serving malwares [4]. The implications of the trojan horse show that a good detection technique is crucial in detecting the trojan horse.

The objectives of this research paper are to investigate and evaluate existing trojan horse detection techniques, to develop a new model trojan horse detection technique and to evaluate the proposed model. Knowledge Discovery and Data Mining (KDD) has been used as part of the methods in this research. Based on the gaps and challenges found in the existing works, the pre-processing in the KDD processes has been improved. This includes the improvement of the feature selection where the static, dynamic and automated analyses have been integrated. As a result, prior forming the trojan horse detection technique, a new trojan horse classification has been developed. This classification is used as the input for the trojan horse detection technique. This trojan horse classification is called as an Efficient Trojan Classification (ETC) and is not discussed in this research paper and can be referred in paper by Saudi and her colleagues [5]. Furthermore, bigger dataset were used in the experiment and different machine learning algorithms have been integrated with the proposed model to optimise the trojan horse detection technique. Moreover, from the experiment conducted, SMO algorithm has been proven as the best machine learning algorithm to optimise the trojan horse detection performance.

This paper is organised as follows. Section 69.2 presents the related works with trojan horse detection techniques. Section 69.3 explains the methodology used in this research paper which consists of static, dynamic and automated analyses and the architecture of the controlled laboratory environment. Section 69.4 presents the research findings which consists the results of the testing and evaluation of the proposed trojan horse detection. Section 69.5 concludes and summarises the future work of this research paper.

69.2 Related Works

Though the trojan horse study was started by Thimbleby et al. [6], only after 10 years later, more studies were carried out such as by [7–10]. However, this works more focusing on trojan horse hardware taxonomy and hardware detection techniques instead. Each of these works has it owns strengths and gaps that can be further improved. Zhang and colleagues used timestamp-based data stream clustering algorithm to detect trojan horse theft activity [11]. The researchers used clusters to compress trojan horse communication data stream information and extracted clusters characteristics for the detection processes. Based on the experiment conducted, it produced 90 % an accuracy rate and lower false negative rate. However this work is only focusing on trojan horse with theft capability.

Apart from that, Tang presented a new trojan horse detecting method, based on Portable Executable (PE) file static attributes [12]. An intelligent information processing technique is used to analyze those static attributes in the PE files. The experiment result showed the test pass rate is 63.90 %. The result can be further improved if the experiment involves bigger volume of dataset.

While Liu and colleagues used data mining to detect the trojan horse in Windows environment [13]. This study shows that the accuracy of classification can be increased when the more relevant features are used in the data mining processes and reduces the consumption of time space. However, the more features are selected, the more time building classification cost, it responds slower in real time and it needs bigger dataset from real network environment. As for work by Dai and colleagues, they presented a novel malicious code detection approach by mining dynamic instruction sequences [14]. Their result showed that their approach is accurate, reliable and efficient. But they used dynamic analyses only and when conducting their experiments, the method was not able to detect any malicious code hooked in the remaining part of the executable code. Improvement can be done if their experiment combining both static and dynamic analysis.

Based on all the previous works discussed above, the main challenges which should be considered thoroughly are the dataset types and volume, analysis and detection techniques and feature selection to detect the trojan horse efficiently. Therefore, in this research, a new trojan detection technique is developed by integrating static, dynamic and automated analyses and by using bigger and standard dataset and an improved feature selection, which is further explained in Sects. 69.3 and 69.4.

69.3 Methodology

In order to produce a new trojan horse detection technique, the researchers' had conducted few experiments and researches. A controlled laboratory environment is created to conduct the experiment as illustrated in Fig. 69.1. It is a controlled laboratory environment and almost 80 % of the software used in this testing is an open source or available on a free basis. No outgoing network connection is allowed for this architecture. The static, dynamic and automatic analyses were conducted. The cuckoo sandbox (for automated analyses) has been integrated in the laboratory architecture as well.

The dataset used for this research are from VX Heaven [15] and Offensive Computing [16]. Many studies have used these dataset in their experiment [14, 17–20]. In these datasets there are many variant of malwares and benign files. A total amount of 1640 trojan horse datasets have been tested in this lab. These dataset were categorized into different types which are the Clicker, DDOS, BAT, AOL, BOOT, PHP and PWS. For this research, Knowledge Discovery and Data Mining (KDD) is used as a technique to identify the trojan horse patterns in the datasets. KDD is defined as several stages of process that are experienced by the data to determine the patterns, which must be important, potentially useful and understandable. Enhancements have been made to the KDD data pre-processing and pattern extraction process. Under the data pre-processing process, the static, dynamic and automated analyses are implemented as illustrated in Fig. 69.2.

The whole processes involved in the experiment are summarized in the flow-chart in Fig. 69.3. After the static, dynamic and automated analyses are conducted successfully (refer Fig. 69.3), a pattern of useful Trojan horse characteristics is identified and is transformed into nominal data to be mined in the WEKA software. It is an open source JAVA software [21]. In the WEKA, five different classification algorithms were executed in order to study the type of classifier suitable to deal with huge amount of features extracted from the static, dynamic and automated analyses processes. The different classifiers that were executed are the Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), IBk, Naïve Bayes and J48 tree algorithm.

After the above data analyses are conducted successfully, a pattern or features of useful trojan horse characteristics is identified, where the ETC Classification has been formed where it consists of the trojan horse infection, activation, payload and operating algorithm [5]. Later, these four trojan horse characteristics are used to represent all the dataset used for the experiments. Then, these dataset were used in data mining (using WEKA), the trojan horse characteristics are transformed into nominal data with a certain number representation.

As for the evaluation, 10-fold cross validation is used to divide the dataset into testing and training dataset. The reasons why the 10-fold cross validation is used are firstly, it uses as much data as possible for training and testing and secondly, the better accuracy of its findings. All analysis of the trojan horse, were documented and recorded properly. This record is useful in understanding on how the

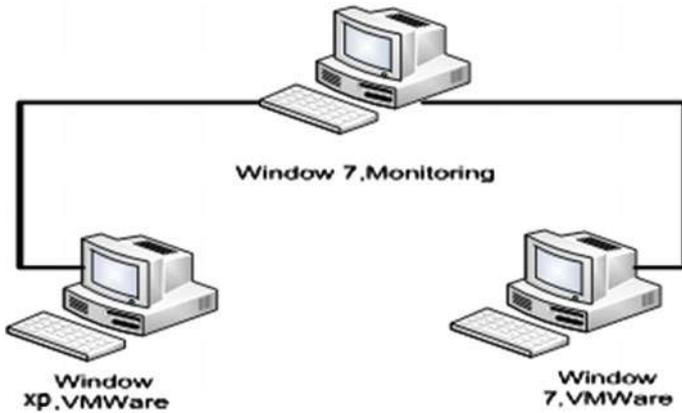


Fig. 69.1 Controlled laboratory architecture

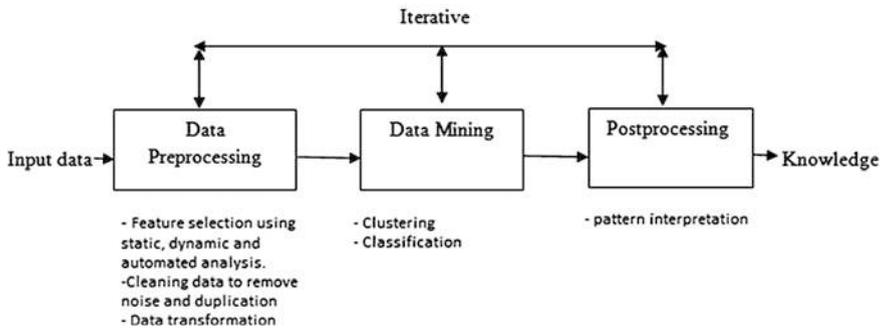


Fig. 69.2 ETDMo KDD processes

trojan horse works. There are several performance parameters that need to be measured during the experiment which are as in the Eqs. 69.1 and 69.2. TPR represents the true positive rate, FPR represents as false positive rate, TP represent as true positive, FN represents as false negative, FP represents as false positive and TN represents as true negative.

$$TPR = TP / (TP + FN) \tag{69.1}$$

$$FPR = FP / (FP + TN) \tag{69.2}$$

69.4 Findings

This section presents the finding results of the machine learning algorithms. The dataset have been classified using the WEKA. The MLP, SMO, IBk, Naïve Bayes and J48 algorithms were chosen to classify the dataset and the result as displays in

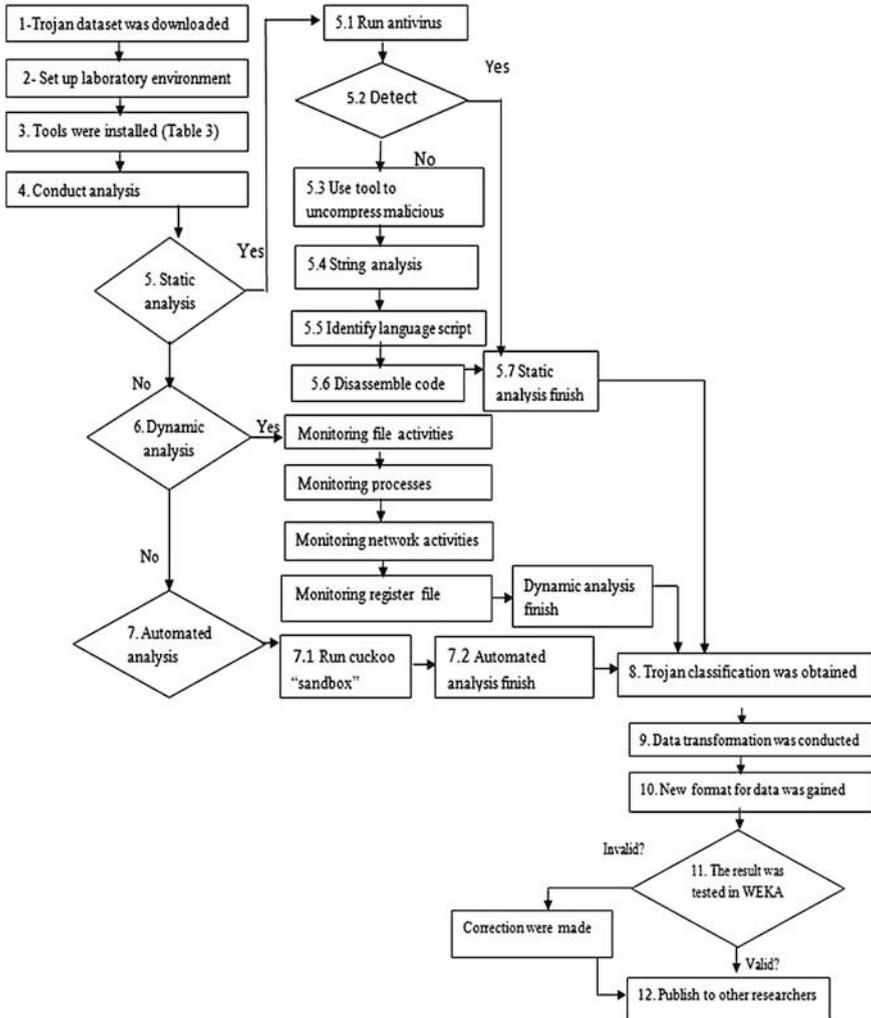


Fig. 69.3 Summarization of the ETDMO research processes

Table 69.1. Table 69.1 illustrates the number of true positive rates (TPR) for five different algorithms that are used in the analysis were conducted. The SMO algorithm got the highest true positive rate with 98.2 % accuracy rate, followed by MLP algorithm with 98.0 % accuracy rate, J48 algorithm with 94 % accuracy rate and IBK algorithm with 90.7 % accuracy rate. The lowest TPR was 86.2 % accuracy rate for Naive Bayes algorithm. The False positive Rate are (FPR): the Naïve Bayes with 13.8 % accuracy rate has the highest FP accuracy, followed by IBk with 9.3 %, J48 with 6.0, and 2.0 % for Multilayer Perceptron and SMO with 1.7 % accuracy rate. Based on the experiment conducted, SMO algorithm has the

Table 69.1 Machine learning algorithm results

Classifier	ETDMO results (%)	
	TPR	FPR
SMO	98.2	1.7
MLP	98.0	2.0
J48	94.0	6.0
IBk	90.7	9.3
Naïve Bayes	86.2	13.8

TPR represents: True Positive Rate, *FPR* represents: False Positive Rate

Table 69.2 Comparison with the existing work

Classifier	ETDMO results (%)		Existing work	
	TPR	FPR	TPR	FPR
SMO	98.2	1.7	93.2	9.6

highest of the True Positive Rate (TPR) with 98.2 % and the lowest of the False Positive Rate (FPR) of 1.7 %.

A comparison with similar work by Dai and colleagues [14] is carried out (refer Table 69.2). This work used the same source of the dataset. True positive rate (TPR) and false positive rate (FPR) are used during the experiment. The experiment was conducted using the WEKA software. The experiment result after the compared with related existing work shows that, the ETDMo model has the higher accuracy rate for trojan horse detection. For this research the TP rate was 98.2 % with FP rate 1.7 %, compared with Dai and colleagues work [14], the TP rate was 91.9 and 9.6 % for FP rate. After the comparison with the existing work found that, this research results got a better performance, which could be due to the development made by the integration of static, dynamic and automated analysis analyses which are part of whole ETDMo KDD process.

69.5 Conclusions and Future Works

As a conclusion, this research has managed to provide a better TPR, which is 98.2 % that outperformed the existing work. This result can be used as a reference and comparison by other researchers with the same interests. For future work, different machine learning algorithms will be tested to the dataset produced from this research. This paper is part of a larger project to build up an automated malware clean up model. Ongoing research will include other malware classification and the development of software to automate the malware dataset cleanup.

Acknowledgments The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) for the support and facilities provided. This research paper is supported by Universiti Sains Islam Malaysia (USIM) grant [PPP/GP/FST/SKTS/30/11912], [PPP/FST/SKTS/30/12812] and Research Management Centre, Universiti Sains Islam Malaysia (USIM).

References

1. Mitropoulos, S., Patsos, D., Douligeris, C.: On incident handling and response: a state-of-the-art approach. *Comput. Secur.* **25**(5), 351–370 (2006)
2. Saudi, M.M.: User awareness in handling computer viruses incident for windows platform. *Jurnal Teknologi Maklumat dan Multimedia* **4**, 53–72 (2007). ISSN: 1823–0113
3. Kaspersky Lab ZAO: What is flame? <http://www.kaspersky.com/flame> (2012)
4. Australia Computer Emergency Response Team (AusCERT): AusCERT incident metrics June 2013. <https://www.auscert.org.au/render.html?it=17856> (2013)
5. Saudi, M.M., Abuzaid, A.M., Taib, B.M., Abdullah, Z.H.: An efficient trojan horse classification (ETC). *IJCSI Int. J. Comput. Sci. Issues* **10**(2), 3. March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
6. Thimbleby, H., Anderson, S., Cairns, P.: A framework for modelling trojan horse s and computer virus infection. *Comput. J.* **41**(7), 444–458 (1998)
7. Chakraborty, R.S., Narasimhan, S., Bhunia, S.: Hardware trojan horse : threats and emerging solutions. In: *Proceedings IEEE International High Level Design Validation and Test Workshop*, San Francisco, CA, pp. 166–171 (2009)
8. Karri, R., Rajendran, J., Rosenfeld, K., Tehranipoor, M.: Trustworthy hardware: identifying and classifying hardware trojan horse s. *Computer* **43**(10), 39–46 (2010)
9. Tehranipoor, M., Koushanfar, F.: A survey of hardware trojan horse taxonomy and detection. *Des. Test of Comput. IEEE* **27**(1), 10–25 (2010)
10. Karri, R., Rajendran, J., Rosenfeld, K.: Trojan horse taxonomy. In: Tehranipoor, M., Wang, C. (eds.) *Introduction to Hardware and Security Trust*, pp. 325–338. Springer, New York (2011)
11. Zhang, X., Liu, S., Meng, L., Shi, Y.: Trojan horse detection based on network flow clustering. In: *Proceedings of the Multimedia Information Networking and Security Conference*. IEEE, pp. 947–950 (2012)
12. Tang, S.: The detection of Trojan horse based on the data mining. In: *Fuzzy Systems and Knowledge Discovery International Conference*. IEEE, vol. 1, pp. 311–314 (2009)
13. Liu, Y., Zhang, L., Liang, J., Qu, S., Ni, Z.: Detecting Trojan horses based on system behavior using machine learning method. In: *2010 Machine Learning and Cybernetics Conference*. IEEE, vol. 2, pp. 855–860 (2010)
14. Dai, J., Guha, R., Lee, J.: Efficient virus detection using dynamic instruction sequences. *J. Comput.* **4**(5), 405–414 (2009)
15. VX Heaven: Computer virus collection. <http://vxheaven.org/vl.php> (2014)
16. Offensive Computing: Malware search. <http://www.offensivecomputing.net> (2014)
17. Schultz, E.E., Shumway, Russell: *Incident Response: a Strategic Guide to Handling System and Network Security Breaches*, 1st edn. New Riders Publishing, USA (2001)
18. Henchiri, O., Japkowicz, N.: A feature selection and evaluation scheme for computer virus detection. In: *Proceedings of the Sixth International Conference on Data Mining*, 2006. ICDM '06. IEEE Xplore, Hong Kong, p. 891 (2006)
19. Moskovitch, R., Stopel, D., Feher, C., Nissim, N., Japkowicz, N., Elovici, Y.: Unknown malware detection and the imbalance problem. *J. Comput. Virol.* **5**(4), 295–308 (2008). doi:10.1007/s11416-009-0122-8
20. Khan, H., Mirza, F., Khayam, S.A.: Determining malicious executable distinguishing attributes and low-complexity detection. *J. Comput. Virol.* **7**(2), 95–105 (2010)
21. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *SIGKDD Explor.* **11**(1), 10–18 (2009)

Chapter 70

An Access Control Framework in an Ad Hoc Network Infrastructure

Tanya Koohpayeh Araghi, Mazdak Zamani, A.A. Manaf
and Sagheb Kohpayeh Araghi

Abstract The union of an arbitrary topology constitutes of self-configuring mobile routers and associated hosts connecting by wireless links is called a mobile ad hoc network (MANET). In this paper, we propose an access control framework for an ad hoc network infrastructure. Compared with the wired and semi wireless network infrastructures in which users should be authenticated and given specific rules for access to the network resources, in an ad hoc network infrastructure some trusted nodes will be required to play the roles of authentication servers and policy enforcement rule servers when the previous servers leave the network. In order to reach to this purpose firstly, a model for Dynamic Monitoring in Ad hoc Networks is described which is going to be used as the trust based part of our proposed model, after that an access control Framework for semi-infrastructure ad hoc networks will be explained. Then, we will introduce our proposed framework to develop the mentioned semi ad hoc framework in a fully ad hoc network environment.

T.K. Araghi (✉) · M. Zamani · A.A. Manaf
Advanced Informatics School, Universiti Teknologi Malaysia,
54100 Kuala Lumpur, Malaysia
e-mail: Tanya.koohpayeh@gmail.com

M. Zamani
e-mail: mazdak@utm.my; zamani.mazdak@gmail.com

A.A. Manaf
e-mail: azizah07@citycampus.utm.my

S.K. Araghi
Faculty of Engineering, Multimedia University Malaysia,
63100 Cyberjaya, Selangor, Malaysia
e-mail: s_koohpaeh@yahoo.com

70.1 Introduction

Setting up a MANET is usually in case of emergency for temporary operations or when there are no resources to set up an elaborate network. Routing the packets between nodes in ad hoc networks relies on implicit trust relationships in one hand, and on the other hand, the ad hoc routing protocols must address either general security objectives like authentication, confidentiality, integrity, availability and non-repudiation or, location confidentiality, cooperation fairness and absence of traffic diversion [1–5].

Besides of security threats and vulnerabilities including lack of proper authentication, insecure routing and complex mobility and identity management, the most important challenges in trust management and security caused by self organizing in mobile ad hoc networks. Trust establishment becomes complicated while nodes participating in the network do not have any prior knowledge from each other [6–9].

Protecting the network resources from unauthorized accesses makes policy based security as a major need for such networks. Although, this purpose is being very complex because of the diversity in security loop holes in different network layers and uncontrolled media access to change topology specifications [10–13].

Managing permissions and roles are usually manual tasks that impose a huge administrative burden and it is error-prone if a large number of permissions and roles are frequently changed. Enforcing policies for security is in the primary stages in MANETs. In general, to facilitate retrieval of data management, data access in a wireless ad hoc network requires a service channel to boost cache recovery, application logging and role base access logging [14–17].

In this paper we developed an access control framework in a fully ad hoc environment which had been introduced for semi-ad hoc environment by [6]. In the proposed framework firstly, based on the dynamic trust monitoring the key nodes which are going to play the roles of authentication servers and role servers will be specified. Then, considering the proposed framework, the policies will be defined to let the routing protocol for every routing decision.

70.2 Literature Review and Related Works

Memon summarized the database access concerns in order to highlight access constraints and the pertinent issues surface after an ad hoc network of devices is formed. These constraints are divided to organizational policy and access criteria including policies defined for users, groups and roles in which the users access depends on their organizational placement, application contents that enables identification of applications regardless of used ports, standard ACLs like those worked in layers 2, 3, 4 which defined based on the source MAC address, source and destination IP Address and protocol and either type, periodic and absolute time

such as policies defined for a specific time or during particular time of days, location in which Access policy can be defined based on the direction of traffic that is a function of the source and destination locations and finally access control based on the user's roles in which once a session is established, a subset of roles is activated, which contains functional roles include permissions that a user need during a session based on his role in organization [15, 18, 19].

Alicherry et al. and Zhao et al. proposed frameworks to implement the mechanisms of access control similar to a service on the resources of the network, but the effect of these access control mechanisms is not specified on routing schemes [20, 21, 13].

In the following some related works about the concept of trust in the MANET and also a framework for implementing access control policies which are the basis of our proposed model will be investigated.

70.2.1 Dynamic Monitoring for Trust in Ad Hoc Networks

Shohreh et al. proposed a trust based clustering model setting up based on a computed trust value of the nodes participating in the network with regards to previous experience of the nodes in forwarding or receiving the packets in order to supervise the behavior of the nodes in each cluster. For this purpose a head cluster and an agent for it will be selected in every cluster. The cluster head monitors and supervises intra cluster transactions considering interactions among its cluster. If the cluster head leaves the network its agent compensates its role in the network, so cluster head and its agent maintain the information of all nodes in their cluster. The highest trust value will be assigned to master head which supervises the inter cluster transactions and monitors all cluster heads in the network. Direct trust value is computed based on the rate of successful packet and total packet which transmitted with nodes. T is total packets, s is success packets and n is the number of nodes, p is the number of recorded previous interactions and $p \leq (n/3)$.

$$T \cdot V = \left(\frac{S}{T} + \sum_{N-1}^P TV \right) \frac{1}{3}. \quad (70.1)$$

Based on this formula, last one-third experiences of each node are preserved in a table. For example when the number of nodes is 15, 5 last experiences will be maintained in the table as trust levels. The range of trust value is from 0 to 1, which represents uncertain, distrust and trustable nodes, so three levels as A, B and C considered as Table 70.1 shows. Nodes within a range of 0.7 to 1 are called as the most trustable nodes and can be selected for particular roles and responsibilities [22, 23].

Table 70.1 The different trust levels

Trust level	Min range	Max range
Level A	0.7	1
Level B	0.3	0.7
Level C	0	0.3

70.2.2 An Access Control Framework in Semi-Infrastructures Ad Hoc Network Model

This framework designed for an implemented semi ad hoc network connected to a backbone network. The functionality of backbone depends on the global security policies enforced on LAN or WAN, but in ad hoc section these policies cannot be implemented because of the mobility of nodes. Hence, the framework is implemented to consider the security policies in both backbone and ad hoc sections. Backbone network includes one or more authentication server(s) and a Global Policy Management Server (GPMS). The GPMS is used in the framework for managing the global policies appropriately. On the other hand, several mobile nodes will be selected in MANET by the backbone network as Policy Enforcing Node (PEN). These nodes are responsible for running security rules in MANET. The security rules will be distributed through a network message named Policy Enforcing Message (PEM). A network connectivity matrix is also needed in each time interval in order to distribute global policies among nodes connected to the network. The connectivity matrix represents connectivity between different nodes in ad hoc network. The process of selecting PEN and distributing the global policy rules will be done by the framework after every time interval using the network information given by the authentication and role servers. When PEN is selected, GPMS sends the global policy rules to the policy enforcing node. Receiving the global policy rules, the PEN distributes the rules to different MANET nodes. The GPMS selects the PENs. The policy rules will be bound by each node to its own Ad hoc interface and the routing decisions will be made by the nodes based on the policy rules. For the new nodes joining to the network, they should be authenticated with the backbone. Message Authentication Code (MAC) is used in order to authenticate PEMs. These policies can be specified by the role of the MANET nodes, which let enforcing the Role Based Access Control (RBAC) policies in MANET. The message exchanged among servers, PENs and nodes include the sequence and TTL fields [6, 24, 25].

70.3 Proposed Model

The mentioned model for semi ad hoc network can be implemented for the whole of ad hoc network environment, if the role of authentication and role servers and also global policy management servers (GPMS) are distributed among nodes in the

network. For this purpose, trusted nodes within the radio range of each zone is required to have a vast capacity using as repositories to maintain the role information and global policy rules. Also a master head will be selected based on the Table 70.1, which is a node with the highest value in the range A as an authentication server.

Selecting nodes as GPMS is based on the proposed formula in the literature, so the nodes with the highest trust level (placed in level A) will be chosen as GPMS, also there is an agent for each GPMS.

Regarding the mobility of nodes, when a GPMS node leaves its zone its agent selected as a new GPMS in a short delay time, and then the supervising among zone is preserved dynamically.

Since authenticity of the policy enforcing message is vital, nodes can confirm the source of the PEM by Message Authentication Code (MAC) to overcome the threat of impersonation. In MAC, the hash value of the whole message body is calculated and sent with the header. The hash key will be negotiated at the time of authentication of the node [26–28].

Policy enforcing messages are transmitted between MANET nodes and GPMS.

In comparison to the previous mentioned framework, since the new framework has developed in an ad hoc environment, the two parts of “GPMS at backbone” and “PENS” will be eliminated. GPMS will be selected based on the highest trust level and also the agent node is prepared to compensate the role of GPMS in case of leaving the network. GPMS constitutes of two parts, role assigning and enforcing the policy rules to the nodes.

70.3.1 Functional Steps

- Step 1: When a node wants to join to the MANET, it will be authenticated by the master head as an authentication server.
- Step 2: A role request will be sent to GPMS. The roles will be assigned and the access level for the specified role will be granted to the node.
- Step 3: Afterwards, policy rules will be sent through policy enforcing messages (PEM) to the nodes. Each PEM will be enforced based on the roles that each node acquired on step 2.
- Step 4: Based on the enforced rules from step 3, the decision will be made for routing that whether a route request is allowed to be performed based on the policies or not. The flow of operational framework is shown in Fig. 70.1.

The rules contain a five tuple. For example, Deny A * B80 says, node A cannot access B via port 80. These rules are acted like the Access Lists (ACL) in order to enforce the filtering on input and output traffics and protect the resources of networks from unauthorized access.

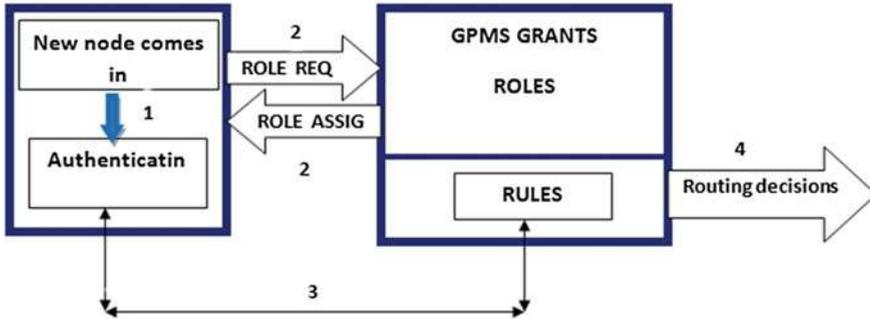


Fig. 70.1 Ad hoc flow of operational framework

70.4 Conclusion and Future Work

This framework is designed for small organizations with a limited number of nodes. So, the future work is to develop this model for large organizations.

References

1. Bhatia, R., et al.: Security issues pertaining to ad hoc networks-a survey, *IJCSMS Int. J. Comput. Sci. Manage. Stud.* **12** (2012)
2. Araghi, T.K., Zamani, M., Manaf, A.A., Abdullah, S.M., Bojnord, H.S., Araghi, S.K.: A survey for prevention of black hole attacks in wireless mobile adhoc networks using IDS agents. 12th International Conference on Applied Computer and Applied Computational Science (ACACOS '13), Kuala Lumpur, Malaysia, 2–4 April 2013
3. Ghazizadeh, E., Zamani, M.: Cloud computing and single sign on: a literature review. *Sci. World J.* ISSN: 1537-744X. IF = 1.73. Q1 (Hindawi Publishing Corporation (UR)) (2013)
4. Janbeglou, M., Zamani, M., Ibrahim, S.: Redirecting network traffic toward a fake DNS server on a LAN. 3rd IEEE International Conference on Computer Science and Information Technology, pp. 429–433, Chengdu, China, 9–11 July 2010
5. Janbeglou, M., Zamani, M., Ibrahim, S.: Redirecting Outgoing DNS requests toward a fake DNS server in a LAN. IEEE International Conference on Software Engineering and Service Science, pp. 29–32. Beijing, China, 16–18 July 2010
6. S. Maity, et al., An access control framework for semi-structured ad hoc networks, in *Computer Technology and Development (ICCTD)*, 2010. 2nd International Conference on, 2010, pp. 708–712
7. Araghi, T.K., Zamani, M., Manaf, A.A., Abdullah, S.M., Bojnord, H.S., Araghi, S.K.: A secure model for prevention of black hole attack in wireless mobile ad hoc networks. 12th International Conference on Applied Computer and Applied Computational Science (ACACOS '13), Kuala Lumpur, Malaysia, 2–4 April 2013
8. Ghazizadeh, E., Zamani, M., Manan, J.A., Alizadeh, M.: Trusted Computing Strengthens Cloud Authentication. *Sci. World J.* **2014**, Article ID 260187, 17 (2014). doi:[10.1155/2014/260187](https://doi.org/10.1155/2014/260187) (IF: 1.730)

9. Araghi, T.K., Zamani, M., Manaf, A.A.: Performance analysis in reactive routing protocols in wireless mobile ad hoc networks using DSR, AODV and AOMDV. *International Conference on Informatics and Creative Multimedia 2013 (ICICM2013)*, pp. 76–79. Kuala Lumpur, 3–6 September 2013
10. Savola, R.M.: Node level security management and authentication in mobile ad hoc networks, in *mobile data management: systems, services and middleware, 2009. MDM'09. Tenth International Conference on*, 2009, pp. 449–458 (2009)
11. Araghi, T.K., Zamani, M., Manaf, A.A., Abdullah, S.M., Bojnord, H.S., Araghi, S.K.: A survey for prevention of black hole attack in wireless mobile ad hoc networks using cryptographic techniques. *12th International Conference on Applied Computer and Applied Computational Science (ACACOS '13)*, Kuala Lumpur, Malaysia, 2–4 April 2013
12. Mohebbi, K., Ibrahim, S., Zamani, M.: *UltiMatch-NL: A web service matchmaker based on multiple semantic filters. PLOS ONE J. Publ. Libr. Sci.* eISSN: 1932-6203. (UR). Impact factor (2012) = 3.73. Q1 (2013)
13. Beigzadeh, S., Zamani, M., Ibrahim, S.: Development of a web-based community management information system. *The Fourth International Conference on Information and Computing (ICIC2011)*. pp. 3–6. Phuket, Thailand, 25–27 April 2011
14. Araghi, T.K., Zamani, M., Manaf, A.A., Abdullah, S.M., Bojnord, H.S., Araghi, S.K.: A survey for prevention of black hole attack in wireless mobile ad hoc networks using trusted neighbor nodes. *12th International Conference on Applied Computer and Applied Computational Science (ACACOS '13)*, Kuala Lumpur, Malaysia, 2–4 April 2013
15. Memon, Q.A.: Implementing role based access in healthcare ad hoc networks. *J. Netw.* **4**, 192–199 (2009)
16. Sadeghian, A., Zamani, M.: Detecting and preventing DDoS attacks in botnets by the help of self triggered black holes. *Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*. Bali, Indonesia, 10–12 Feb 2014
17. Yazdanpanah, S., Chaeikar, S.S., Zamani, M., Kourdi, R.: Security features comparison of master key and ikm cryptographic key management for researchers and developers. *2011 3rd International Conference on Software Technology and Engineering (ICSTE 2011)*, pp. 365–369. Kuala Lumpur, Malaysia, 12–13 Aug 2011
18. Janbeglou, M., Zamani, M., Ibrahim, S.: Improving the security of protected wireless internet access from insider attacks. *AISS: Adv. Inform. Sci. Serv. Sci.* **4**(12), 170–181. ISSN: 2233-9345 (2012)
19. Ghazizadeh, E., Zamani, M., Manan, J.A., Pashang, A.: A survey on security issues of federated identity in the cloud computing. *IEEE International Conference on Cloud Computing Technology and Science*, pp. 562–565. Taiwan, 3–6 Dec 2012
20. Alicherry, M., et al.: Deny-by-default distributed security policy enforcement in mobile ad hoc networks, In: *Security and Privacy in Communication Networks*, pp. 41–50 (2009)
21. Zhang, H., et al.: Bootstrapping deny-by-default access control for mobile ad hoc networks, In: *Military Communications Conference, 2008. MILCOM 2008.* IEEE, pp. 1–7 (2008)
22. Shohreh, H., et al.: Dynamic monitoring in ad hoc network. *Appl. Mech. Mater.* **229**, 1481–1486 (2012)
23. Honarbakhsh, S., Zamani, M., Honarbakhsh, R.: Dynamic Monitoring in Ad Hoc Network. *Applied mechanics and materials*, vols. 22–231, pp. 1481–1486. Trans Tech Publications, Switzerland. ISSN: 1660-9336 (2012)
24. Zamani, M. Manaf, A.A., Daruis, R.: Azizah technique for efficiency measurement in steganography. *ICIDT 2012, 8th International Conference on Information Science and Digital Content Technology 3*, pp. 480–484. Jeju, Korea, 26–28 June 2012
25. Ghazizadeh, E., Zamani, M., Manan, J.A., Khaleghparast, R., Taherian, A.: A trust based model for federated identity architecture to mitigate identity theft. *International Conference for Internet Technology and Secured Transactions*. London. 10–12 Dec 2012
26. Min, Z., Jiliu, Z.: Cooperative black hole attack prevention for mobile ad hoc networks, pp. 26–30 (2009)

27. Chaeikar, S.S., Razak, S.A., Honarbakhsh, S., Zeidanloo, H.R., Zamani, M., Jaryani, F.: Interpretative key management (IKM), a novel framework. 2010 International Conference on Computer Research and Development, pp. 265–269. Kuala Lumpur, Malaysia 7–9 May 2010
28. Nikbakhsh, S., Zamani, M., Manaf, A.A., Janbeglou, M.: A novel approach for rogue access point detection on the client-side. 26th IEEE International Conference on Advanced Information Networking and Applications. Japan, 26–29 March 2012

Chapter 71

Enhancement of Medical Image Compression by Using Threshold Predicting Wavelet-Based Algorithm

N.S.A.M. Taujuddin and Rosziati Ibrahim

Abstract In recent decades with the rapid development in biomedical engineering, digital medical images have been becoming increasingly important in hospitals and clinical environment. Apparently, traversing medical images between hospitals need a complicated process. Many techniques have been developed to resolve these problems. Compressing an image will reduce the amount of redundant data with the good quality of the reproduced image sufficiently high, depending on the application. In the case of medical images, it is important to reproduce the image close to the original image so that even the smallest details are readable. The aim of this paper is to propose a new compression algorithm by using the threshold values. It started by segmenting the image area into Region of Interest (ROI) and Region of Background (ROB) and use the special features provide by wavelet algorithm to produce efficient coefficients. These coefficients are then will be used as threshold value in our new proposed thresholding predicting for compression algorithm. The new compression algorithm is expected to produce a fast compression algorithm besides decreasing the image size without tolerating with the precision of image quality.

Keywords Wavelet · Hard threshold · Soft threshold · Image compression

N.S.A.M. Taujuddin (✉)

Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: shahidah@uthm.edu.my; nikshahidah712@gmail.com

R. Ibrahim

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn
Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
e-mail: rosziati@uthm.edu.my

71.1 Introduction

Advances over the past decade in many aspects of digital technology especially devices for image acquisition, data storage, and bitmapped printing and display have brought about many applications of digital imaging. However, problems involving storage space and network bandwidth requirements arise when large volumes of images are to be stored or transmitted, as is the case with medical images.

From the diagnostic imaging point of view, the challenge is how to deliver clinically critical information in the shortest time possible. A solution to this problem is through image compression. The main objective of this compression is to reduce redundancy of the data image in order to be able to stored or transmit data in an efficient form [1–5].

Among various algorithms proposed by researchers, wavelet gains a high popularities in compression domain because of its distinctive features. Wavelet is well known because of its energy compactness in the frequency domain. Thresholding is one of the processes in wavelet. Tay et al. says that the selection of threshold(s) is/are the key performance to an effective compression [6].

The rest of the paper is organized as follows: the principles of wavelet is detailed in Sect. 71.2. The project's methodology is presented in Sect. 71.3. Section 71.4 contains the experimental result and analysis. While Sect. 71.5 conclude the entire paper.

71.2 The Wavelet

Wavelet is a flexible tool with rich mathematic content and has enormous potential in many applications and greatly being used in the field of digital images. Wavelet algorithm work as signal processing in such a way like the human vision do. It provides a much more precisely in digital image, movies, color image and signal.

It also has widely used in data compression, fingerprint encoding and also image. There are three properties of wavelets, (a) separability, scalability and translatability (b) multiresolution compatibility and (c) orthogonality [7].

One of the popular wavelet is Discrete Wavelet Transform (DWT). The term discrete in DWT refer to the separation of transformation kernel as well as separation in fundamental nature and function.

Basically, DWT is a transformation process that produces the minimum number of coefficients that sufficient enough for reconstruction of the transform data accurately to the original signal. DWT is usually presented in term of its recovery transformation:

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} d(k, \ell) 2^{-\frac{k}{2}} \Psi(2^{-k}t - \ell) \quad (71.1)$$

$d(k, \ell)$ is sampling of $W(a, b)$ at discrete points k and ℓ . While k is referring to a as $a = 2^k$ and b is referring to ℓ as $b = 2^{\ell}$.

The DWT introduce the scaling function or sometimes referred as smoothing function. It use dilation or two-scale difference equation:

$$\phi(t) = \sum_{n=-\infty}^{\infty} \sqrt{2}c(n)\phi(2t - n) \quad (71.2)$$

In this equation, $c(n)$ is a series of scalars describing specific scaling function. Wavelet in DWT itself can be define from these scaling function:

$$\Psi(t) = \sum_{n=-\infty}^{\infty} \sqrt{2}d(n)\phi(2t - n) \quad (71.3)$$

Here, $d(n)$ is a series of scalar that define the discrete wavelet in terms of scaling function. DWT can well implemented in the above equation as well as using filter bank technique. Typically, wavelet use two filters, namely analysis filter and synthesis filter. The analysis filter is used to split the original signal to several spectral components called *subband*.

Firstly, the signal will pass a low pass filter for approximation coefficients outputs. Then, it will pass through the high pass filter resulting the detail coefficients.

In the analysis filter, some points need to be eliminated. This operation is called downsampling or quantization process and usually illustrated as $\downarrow 2$. The process is done to maximizing the amount of necessitate detail and ignoring 'not-so-wanted' details.

Here, some coefficient values for pixel in image are thrown out or set to zero. This is called as the thresholding process and it will give some smoothing effect to the image.

In order to compress the image, Wavelet analysis can be used to divide the information of an image into approximation and detail sub-signals. The approximation sub-signal shows the common trend of pixel values, and three detail sub-signals show the horizontal, vertical and diagonal details or changes in the image.

If these details are very small then they can be set to zero without significantly changing the image. The details under the fixed threshold represent a small enough detail and it can be set to zero. The greater the number of zeros leads to the greater compression.

In inverse of analysis bank, the synthesis bank will do the upsampling ($\uparrow 2$) to reconstruct the original fine scale coefficient by combining the scale and wavelet coefficients at lower coarser scale. During upsampling the value of zero will be inserted between 2 coefficients because during the downsampling, the every second coefficient is thrown away.

The process of upsampling and downsampling is illustrated as in Fig. 71.1.

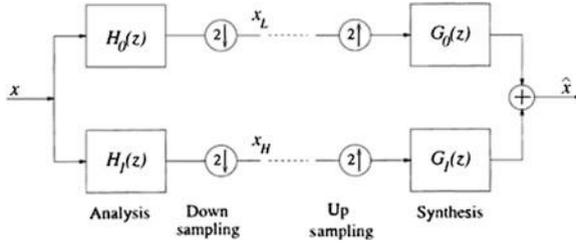


Fig. 71.1 Two-band multirate analysis/synthesis system [3]

71.2.1 Concept of Thresholding

In the wavelet transform, the noise energy is distribute in all wavelet coefficients, while the original signal energy is found in some of the coefficients. Therefore, the signal energy is found much larger than noise energy. So, small coefficients can be considered as caused by noise while large coefficients are triggered by significant signal features.

Based on this idea, thresholding process is proposed. Thresholding is a process of shrinking the small absolute coefficients value while retaining the large absolute coefficient value. It will produce finer reconstruct signal. Threshold also can be define as the Peak Absolute Error (PAE) accepted for image reconstruction [8].

Hard and soft threshold are the common operator used in conjunction with DWT. Donoho is the person who first introducing the word ‘de-noising’ to explain the process of noise reduction in threshold [9].

The wavelet coefficient for hard threshold, h_h , is performed as follows:

$$h_h(i) = \begin{cases} y(i), & |y(i)| > \lambda \\ 0, & \text{others} \end{cases} \tag{71.4}$$

where $y(i)$ is the wavelet coefficients, λ is the specified threshold. While, for soft threshold, h_s , the coefficients is expressed as below:

$$h_s(i) = \begin{cases} \text{sgn}(y(i))[|y(i)| - \lambda], & |y(i)| \geq \lambda \\ 0, & \text{others} \end{cases} \tag{71.5}$$

The elements with absolute value is lower than the threshold value will be set to zero and then the other coefficient will be shrunk. $\text{sgn}(*)$ is a sign function.

$$\text{sgn}(n) = \begin{cases} 1 & n > 0 \\ -1 & n < 0 \end{cases} \tag{71.6}$$

The hard threshold zeros all the coefficients valued below than threshold value and retain the rest unchanged. Whereas, the soft threshold scaled the coefficients in continuous form with the center of zero [10]. These two techniques have their own

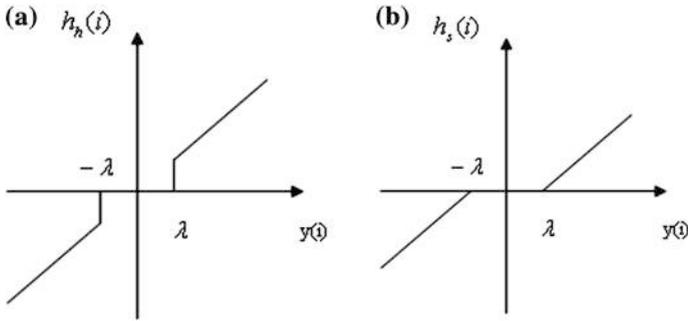


Fig. 71.2 a Hard threshold function. b Soft threshold function

strength and weakness. Hard threshold known as good in preserving edges but bad in de-noising while soft threshold is contradict as can be seen in Fig. 71.2.

There are two types of thresholding; global and level dependent threshold. Global threshold imply single threshold value globally to all wavelet coefficient while level dependent threshold use different threshold value at different level.

Threshold value estimation is very crucial. If the threshold value set too small, it will adopt noise into the signal. While, if the threshold value is too high, the important coefficients value will be screened out leading to deviation condition.

There are some threshold estimation techniques exists such as [11]:

(a) VisuShrink

Donoho is first introducing the wavelet threshold shrinkage method where he suggest that the coefficients have a tendency be to set as zero when the coefficients are greater than threshold. The threshold must meet the specification of $\lambda = \sigma_n \cdot \sqrt{2 \ln N}$ where σ_n is the noise variance and N is the length of the signal.

(b) SureShrink

SureShrink is a soft estimator, where for the given λ , the probability is collected while reducing the non-probability.

(c) HeurSure

HeurSure will choose the best predictor variable threshold.

(d) Minmax

By using the fixed threshold value, it will generate a minimum mean-square error based on minimax criterion.

As an example for wavelet threshold de-noise calculation, the signal $x(t)$, contain of impulse and noise can be expressed as:

$$x(n) = p(t) + n(t) \tag{71.7}$$

where $n(t)$ refer to Gaussian noise with mean zero and standard deviation σ while $p(t)$ refer to impulse. The signal $x(t)$ is then being transferred to time-scale plane.

The threshold, λ , is measured by using the existence rules to shrink the wavelet coefficients. Here, universal threshold is applied:

$$\lambda = \sigma\sqrt{2 \ln N} \quad (71.8)$$

For unknown σ , one can be replace by $MAD/0.6745$, where MAD is median absolute value of the finest scale wavelet coefficients and N is the number of data samples in measured signal. Inverse wavelet transform use the shrunken coefficient, and the series retrieve represent the estimation of impulse $p(t)$ [12].

71.2.2 Recent Research on Wavelet Threshold Predicting Algorithm in Image Compression

The wavelet threshold uses the ordinary multiresolution analysis where the discrete detail coefficients and discrete approximation coefficients are attained by multilevel wavelet decomposition.

To find the best threshold is hard because it requires the knowledge of original data. Beside the contemporary hard threshold, the soft-threshold can be used since it is closer to the optimum minimal rate and protect signal regularity beside reducing the gap between preserve and discarded coefficient for a better recovery [13]. Another solution is by using adaptive soft-threshold or fix the threshold for each wavelet sub-band.

Chen et al. [14] are using adaptive prediction technique in solving multicollinearity problem. Besides, he suggest to adjust predictor variable based on image properties so that more accurate prediction is archived.

Tree-Structured Edge-Directed Orthogonal Wavelet Packet Transform (TS-E-DOWPT) proposed in [15] decomposes image using edge-directed orthogonal and calculate the cost function. This technique improves the PSNR value and visual quality. Besides, trimming the structure into quad-tree may reduce the stain.

To get a better compression ratio, Hosseini et al. [16] recommending a technique that separating the Region of Interest (ROI) and Background (BG) using growing segment and then encode both of the segment using Contextual Quantization. Different weightage is used for different region, where higher rate for ROI and lower rate for BG.

Pogam et al. in [17] give different view which combining the wavelet transform with curvelet transform and incorporating it with the local adaptive analysis thresholding. This technique contributes to an efficient denoising while pre-serving as much as possible the original quantitative and structural information and ROI of image.

71.3 Methodology

Although the forgoing techniques can make an effective approximation of threshold value, but they got no idea on the relationship between threshold value and quantization step [18]. Besides, the usage of hard threshold in quantizing the coefficients will lead to blocky artifact on medical image [19].

Targeting on this problems, this research is done to develop an efficient threshold prediction algorithm by using the wavelet features to produce a fast compression algorithm besides decreasing the image size without tolerating with the precision of image quality.

Medical community also raise a high intention to produce a low computational cost algorithm with high speed compression and decompression to assist the existence network bandwidth capability while reducing the image size to upkeep the limited storage size.

As in the literature, the wavelet coefficient is predicted based on fix location and variables. But, medical images have its own statistical distribution and have different properties on different subbands. So, to get more precise prediction, the amount of predictor variable must be adjusted based on the image's properties.

Figure 71.3 shows the general image compression system with examples of algorithms used in each process. The compression process starts with transforming the image into coefficient where it usually done by using Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) or Fast Fourier Transform (FFT).

Then, the retaining coefficients are quantized by using hard, soft or semi-soft threshold resulting the stream of symbol. This is where information lost occurs.

The entropy coding or compression process will use efficient lossy or lossless compression algorithm for example the EZW, SPHIT, Huffman, Bitplane Encoding and many more to produce the bit streams representing the compressed image.

By using the same general image compression system, we propose a new compression algorithm with some extended features. Below are the proposed algorithm steps and it is illustrated in Fig. 71.4:

1. The original image is segmented to Region of Interest (ROI) and Region of Background (ROB).
2. DWT is used to produce sequence of wavelet coefficient and separate it to low frequency and high frequency subband.
3. The correlation between adjacent wavelet coefficients are analyzed to get the best suit coefficient relationship.
4. Resulting wavelet coefficient are thresholded by using efficient prediction scheme to get the best truncated threshold. Then the prediction equation is applied for thresholding process to get the significant predicted wavelet coefficient.



Fig. 71.3 General image compression system

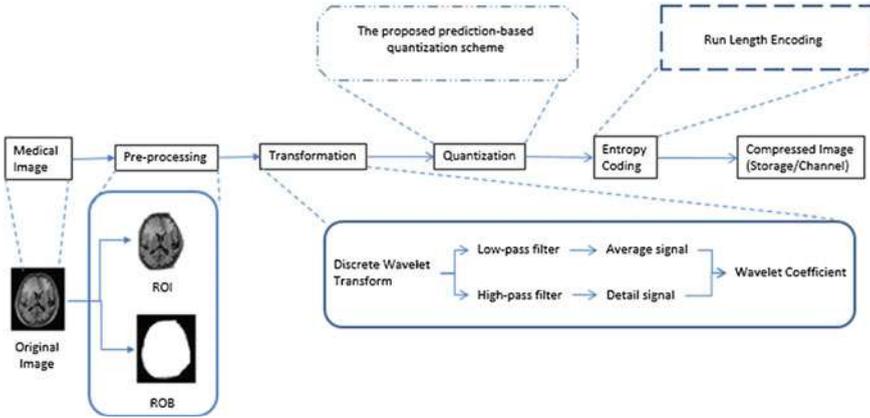


Fig. 71.4 The proposed algorithm steps

71.4 Experimental Result and Analysis

The standard gray-scale 512×512 sized Barbara image is used in this testing to evaluate the popular existing threshold algorithms; the hard and soft threshold. The image is added with Gaussian white noise to facilitate the comparison.

To evaluate the efficiency of the compression algorithm, Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) are used to evaluate the quality of compression. Higher the PSNR value representing a higher compression quality and vice versa [16]. While lower MSE value representing better image quality vice versa.

The PSNR can be defined as:

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \text{dB}$$

While MSE is define as:

$$MSE = \frac{1}{M \cdot N} \sum_{x=1}^M \sum_{y=1}^N |f(x, y) - \hat{f}(x, y)|^2$$

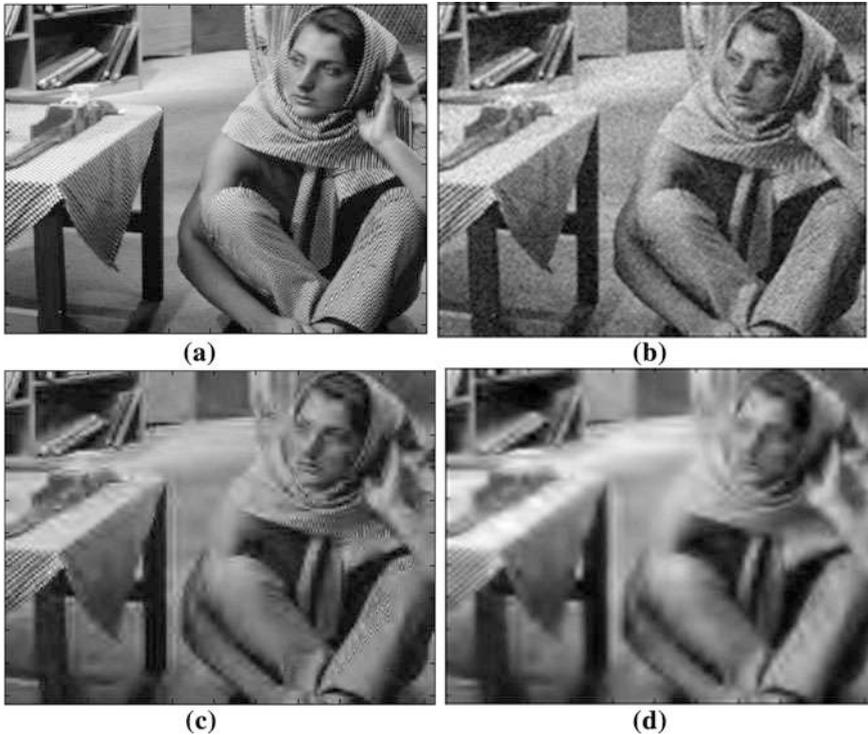


Fig. 71.5 a Original image. b Noisy image. c Hard thresholded image. d Soft thresholded image

M, N is the size of the image while $f(x, y)$ is the pixel value of original image and $\hat{f}(x, y)$ is the pixel value of the reconstructed image.

Figure 71.5 shows the resulting Barbara original, noisy, hard thresholded and soft thresholded image. It is clearly can be seen that the quality of image is significantly affected with different threshold algorithm. The soft thresholded image is better in terms of noise overwhelmed while protecting the image edge. But soft threshold method shows the degradation value of PSNR as can be seen in Table 71.1.

71.5 Conclusion

This paper has discussed a new compression algorithm by putting an efficient threshold prediction algorithm using the wavelet features to produce a fast compression algorithm. The purpose of this new compression algorithm is to decrease the image size without tolerating with the precision of image quality. As the project is still under go, the preliminary analysis done in Sect. 71.4 shows the need

Table 71.1 The PSNR value for different de-noising method

De-noising method	PSNR		
	$\sigma = 25$	$\sigma = 50$	$\sigma = 100$
Noisy image	28.8167	27.8777	27.4628
Hard threshold	23.0161	21.3947	19.9306
Soft threshold	21.7055	20.3142	19.1190

to develop an effective thresholding algorithm that can provide the improvement preservation details at low bit rates while increasing the PSNR value. Protecting details at edges is very crucial especially for sensitive data such as the medical images. Therefore, modification on prediction procedure in threshold step as proposed in our algorithm should be performed in order to abolish the blocking and edge effect while increasing the effectiveness and reliability of the compressed image.

Acknowledgment The authors would like to thanks the Universiti Tun Hussein Onn Malaysia (UTHM), Office of Research, Innovation, Commercialization and Consultancy Management (ORICC) and Malaysian Ministry of Education for facilitating this research activity.

References

1. Strintzis, M.G.: A review of compression methods for medical images in PACS. *Int. J. Med. Inform.* **52**(1–3), 159–165 (1998)
2. Burak, S., Carlo, G., Bernd, T., Chris, G.: Medical image compression based on region of interest, with application to colon CT Images. In: 23rd Annual EMBS International Conference, 2001, pp. 2453–2456 (2001)
3. Kofidis, E., Kolokotronis, N., Vassilarakou, A., Theodoridis, S., Cavouras, D.: Wavelet-based medical image compression. *Futur. Gener. Comput. Syst.* **15**(2), 223–243 (1999)
4. Janaki, R.: Enhanced ROI (region of interest algorithms) for medical image compression. *Int. J. Comput. Appl.* **38**(2), 38–44 (2012)
5. Celik, M.U., Sharma, G., Tekalp, A.M.: Gray-level-embedded lossless image compression. *Signal Process. Image Commun.* **18**(6), 443–454 (2003)
6. Tay, P.C., Acton, S.T., Hossack, J.A.: Computerized medical imaging and graphics a wavelet thresholding method to reduce ultrasound artifacts. *Comput. Med. Imaging Graph.* **35**(1), 42–50 (2011)
7. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, Second, p. 43. Prentice Hall, Upper Saddle River (2002)
8. Baligar, V.P.: Low complexity, and high fidelity image compression using fixed threshold method. *J. Inf. Sci.* **176**, 664–675 (2006)
9. Donoho, D.L.: De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **41**(3), 613–627 (1995)
10. Li Jianmin, X.S.: Analysis and application of modified methods of wavelet threshold functions. In: *Asia-Pacific Conference on Computational Intelligence and Industrial Application*, 2009, pp. 150–153 (2009)
11. Zhen, C., Su, Y.: Research on wavelet image threshold de-noising. In: *2010 International Conference on Future Power Energy Engineering*, pp. 3–6, Jun 2010

12. Zang, H., Wang, Z., Zheng, Y.: Analysis of signal de-noising method based on an improved wavelet thresholding. In: 2009 9th International Conference on Electronic Measurement and Instruments, vol. 1, pp. 1–987, 1–990, Aug 2009
13. Bruni, V., Vitulano, D.A.: Combined image compression and denoising using wavelets. *Signal Process. Image Commun.* **22**, 86–101 (2007)
14. Chen, Y.T., Tseng, D.C.: Wavelet-based medical image compression with adaptive prediction. *Comput. Med. Imaging Graph.* **31**(1), 1–8 (2007)
15. Huang, J., Cheng, G., Liu, Z., Zhu, C., Xiu, B.: Synthetic aperture radar image compression using tree-structured edge-directed orthogonal wavelet packet transform. *AEUE—Int. J. Electron. Commun.* **66**(3), 195–203 (2012)
16. Hosseini, S.M., Naghsh-Nilchi, A.R.: Medical ultrasound image compression using contextual vector quantization. *Comput. Biol. Med.* **42**(7), 743–750 (2012)
17. Le Pogam, A., Hanzouli, H., Hatt, M., Le, C.C., Visvikis, D.: Denoising of PET images by combining wavelets and curvelets for improved preservation of resolution and quantitation. *Med. Image Anal.* **17**(8), 877–891 (2013)
18. Manikandan, M.S., Dandapat, S.: Wavelet-threshold based ECG compression with smooth retrieved quality for telecardiology. In: 2006 Fourth International Conference on Intelligent Sensing and Information Processing, pp. 138–143, Dec 2006
19. Prudhvi, N.V., Venkateswarlu, T.: Denoising of medical images using dual tree complex wavelet transform. *Procedia Technol.* **4**, 238–244 (2012)

Chapter 72

Detection and Revocation of Misbehaving Vehicles from VANET

Atanu Mondal and Sulata Mitra

Abstract The present work is the detection and revocation of misbehaving vehicles in vehicular ad hoc network. In the present work vehicles are within the coverage area of base stations and the base stations are within the coverage area of certifying authority. Each vehicle detects misbehaving vehicles from its neighbors, creates a certificate revocation list by mentioning the identification of the misbehaving vehicles and sends this list to its parent base station. Each base station creates a certificate revocation list after receiving the certificate revocation lists from the vehicles within its coverage area and sends it to the certifying authority. The certifying authority creates a final certificate revocation list after receiving the certificate revocation lists from the base stations within its coverage area and broadcasts it among the vehicles within its coverage area. The qualitative and quantitative performance of the proposed scheme outperforms the existing schemes.

72.1 Introduction

The vehicular ad hoc network (VANET) consists of a group of independent vehicles which are moving throughout the wireless network freely. The potential threat and road accident are increasing due to high velocity of vehicles in VANET. Several types of messages are exchanged among vehicles such as traffic information, emergency incident notifications and road conditions to avoid road accidents and congestion. It is important to forward message correctly in VANET.

A. Mondal (✉)
Department of Computer Science and Engineering,
Camellia Institute of Technology, Kolkata, India
e-mail: atanumondal@hotmail.com

S. Mitra
Department of Computer Science and Technology,
Bengal Engineering and Science University, Shibpur, India
e-mail: sulata@cs.beecs.ac.in; bulu456@yahoo.co.in

However, attacker nodes or misbehaving vehicles may damage the messages. The misbehaving vehicles are authentic vehicles but their behavior deviates from the required standard behavior. These vehicles may jeopardize the safety of other vehicles, drivers, passengers as well as the efficiency of the transportation system. Hence the critical part of any security mechanism in VANET is to identify and revoke misbehaving vehicles.

Several revocation schemes have been proposed so far. In [1] the service provider issues certificates for the vehicles with a limited temporal/spatial scope. These certificates are usable by the vehicles within a particular geographic area or within a certain time or both. The certificates are not tied to the vehicle's registration and can be changed periodically during one service period. But it has a lot of overhead for the creation and revocation of certificates for each new coming and leaving vehicle.

The certificate revocation techniques and privacy-protection techniques are proposed in [2, 3]. A security architecture for vehicular communication systems is developed in [2]. The authors identify threats and models of adversarial behavior as well as security and privacy requirements that are relevant to the vehicular communication context. A SeVeCom baseline architecture is presented in [3]. The various implementation and deployment-specific aspects such as flexible integration in existing communication-stacks, use of a hardware security module and secure connections of vehicular communication on board units to in-vehicle bus systems are also highlighted in [3]. But the authors discussed the concept of certificate revocation list (CRL) distribution without considering high vehicle traffic densities.

In [4] the scalable methods of distributing CRL and other large files over VANET using vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications while taking advantage of the multi-channel operations in IEEE 1609.4 is proposed. The size of the CRL increases with vehicle density. Hence it is difficult to distribute the CRL over the entire network with minimum delay and network congestions.

In [5] the authors consider the distribution of CRL across a large-scale and multi-domain vehicular communication system in a timely manner. A collaboration scheme between regional certificate authorities (CAs) that allows CRLs to contain only regional revocation information is proposed in [5]. Moreover it uses erasure codes to enhance the robustness of the CRL distribution. The scheme does not require any communication and cooperation between road side units (RSUs) during CRL distribution. It minimizes CA-RSU and vehicle-CA-RSU interactions thus limiting congestion in the network. The CRL size is reduced by using regional CA and short lived certificates for travelling vehicles. But the distribution time of CRL is tens of minutes which is too long for a high and dense network like VANET.

A flexible, simple and scalable design for VANET certificates and new methods for efficient certificate management is proposed in [6]. It reduces channel overhead by eliminating the use of CRL. It also protects the system from adversary vehicles by distributing information about adversary vehicles among the whole network

and by revoking the certificates of malicious nodes. But the performance of the scheme is not evaluated on the basis of overhead and delay.

An adaptable method to detect packet forwarding misbehavior based on the principal of flow conservation of messages and the application of policy-based management [7] is proposed in [8]. The message is forwarded to those vehicles which are moving towards the location of the event. In the proposed scheme the different tasks are assigned to a vehicle at different instant of time depending upon the certain policy for identifying the misbehaving vehicles which increases the complexity of each vehicle.

The present work is the detection and revocation of misbehaving vehicles in VANET. The proposed VANET is a hierarchy having CA at the root level, base stations (BSs) at the intermediate level and vehicles at the leaf level.

Each vehicle has an electronic license plate (ELP) in which its vehicle identification number (VIN) is embedded in encrypted (E_VIN) form. The ELP of a vehicle broadcasts the E_VIN after entering into the coverage area of a new BS. The BS verifies the authentication of the vehicle after receiving its E_VIN and assigns a digital signature (D_Sig) to the vehicle if it is authentic [9]. The CA maintains a CRL (CA_CRL) to store the E_VINs of the misbehaving vehicles and broadcasts it for the vehicles within its coverage area.

Each vehicle receives beacon message periodically and service message (S_MSG) after the occurrence of an event from the vehicles which are within its coverage area. The steganography method is used to protect S_MSGs from the access of intruder. The S_MSG consists of D_Sig of the sender vehicle, hidden message (H_MSG), E_VIN of the sender vehicle and the operational part (O_MSG) to retrieve the message (MESS) from H_MSG.

Each vehicle switches on a timer. It receives a S_MSG from a vehicle within its coverage area, and searches CA_CRL for the value of E_VIN field of S_MSG. If found it discards S_MSG. Otherwise it stores the S_MSG in a FIFO queue, triggers ALGO_MESS to retrieve MESS from H_MSG field of S_MSG and inserts a record in the form (E_VIN, MESS, REMARK) in a data table (DT). It repeats the same steps of operation for the other received S_MSGs till the timer expires. Initially the value of the REMARK field of all the records in DT is null. Each vehicle triggers ALGO_V algorithm to detect the misbehaving vehicles using the records in the DT after the expiry of the timer and inserts the E_VIN(s) of the misbehaving vehicle(s) in a CRL. The algorithm repeats the same steps of operation for all the records in the DT and finally sends the CRL to its parent BS.

Each BS switches on a timer and triggers ALGO_B algorithm. The algorithm receives CRLs from the vehicles within its coverage area and stores them in a FIFO queue. It performs union operation among the received CRLs in the queue for generating a new CRL. It repeats the same steps of operation till the timer expires and sends the new CRL to CA after the expiry of the timer.

The CA switches on a timer and triggers ALGO_C algorithm. The algorithm receives CRLs from the BSs under it and stores them in a FIFO queue. It performs union operation among the received CRLs in the queue for generating CA_CRL.

It repeats the same steps of operation till the timer expires and broadcasts CA_CRL among the vehicles within its coverage area after the expiry of the timer.

Unlike [1, 4, 6] the present work uses steganography to protect S_MSGs from intruder in VANET. The use of encryption/decryption algorithm for secure message transmission consumes a considerable amount of time due to their computational hardness. Moreover the length of message may increase during encryption which increases storage overhead along with communication overhead. In the present work CA_CRL consists of E_VIN of the misbehaving vehicles instead of their IP address like [5]. So only the E_VIN of a vehicle is sufficient to identify, detect and revoke misbehaving vehicles from VANET. Unlike [5] the present work performs well in high density of vehicles. In [5] CA distributes pieces of CRL among vehicles via RSUs. So in the worst case a vehicle needs to encounter all the RSUs for collecting all the pieces of CRL. Hence the performance of CRL generation degrades if congestion occurs, if vehicles move slowly and if the distance between RSUs increase. The misbehaving vehicles are identified by their E_VINs during V2V communication in the present work. So no extra task is required to assign to a vehicle for identifying misbehaving vehicle like [8]. There is no requirement of reporting the reason of revocation to CA in the present work like [10] which helps to reduce the delay in revocation.

72.2 Present Work

In this section the function of vehicle, BS and CA are elaborated for v th vehicle (V_v) within the coverage area of B th BS (BS_B) under CA. The number of BSs under CA and the number of vehicles under BS_B are assumed as NO_OF_BS and NO_OF_V_B respectively.

72.2.1 Function of V_v

The function of V_v is elaborated for j th S_MSG (S_MSG_j) that it receives from j th vehicle within its coverage area. S_MSG_j is of the form of (D_Sig_j , E_VIN_j , H_MSG_j , O_MSG_j). D_Sig_j is the digital signature of the j th vehicle, E_VIN_j is the E_VIN of the j th vehicle, H_MSG_j is the hudden message and O_MSG_j is the operational part to retrieve the message (MESS _{j}) from H_MSG_j . O_MSG_j field contains the bit pattern corresponding to the different bitwise operators (AND, OR, XOR). Each bitwise operator is represented by Size_BWO number of bits and the size of O_MSG_j (Size_O_MSG _{j}) is exact multiple of Size_BWO. The ALGO_MESS algorithm is used by V_v to retrieve MESS _{j} from H_MSG_j by performing (Size_O_MSG _{j} /Size_BWO) number of bitwise operations (Number_BWO _{j}) among H_MSG_j and E_VIN_j .

V_v switches on a timer and initializes it to τ_v . It receives S_MSG_j , increases a counter (NS_MSG_v) by 1 and searches CA_CRL for E_VIN_j . If E_VIN_j is in CA_CRL or if $(Size_O_MSG_j \% Size_BWO \neq 0)$ it discards S_MSG_j . Otherwise it stores S_MSG_j in a FIFO queue (Q_v), triggers $ALGO_MESS$ algorithm to retrieve $MESS_j$ from H_MSG_j and inserts a record (R_j) in the form (E_VIN_j , $MESS_j$, $REMARK_j$) in a data table (DT_v) for S_MSG_j .

<pre> /*Retrieval of MESS_j from H_MSG_j*/ { i ← 1 k ← 1 while (k < Number_BWO_j) { if (O_MSG_j (i, i + Size_BWO-1) represents AND bit wise operator) {MESS_j ← H_MSG_j ∧ E_VIN_j Go to L1} else if (O_MSG_j (i, i + Size_BWO-1) represents OR bit wise operator) </pre>	<pre> {MESS_j ← H_MSG_j ∨ E_VIN_j Go to L1} else if (O_MSG_j (i, i + Size_BWO-1) represents XOR bit wise operator) {MESS_j ← H_MSG_j ⊕ E_VIN_j Go to L1} else Go to L1 L1: {i ← i + Size_BWO k ← k + 1}} </pre>
---	---

It repeats the same steps of operation for the other received S_MSGs till the timer expires. The maximum number of S_MSGs in Q_v and the maximum number of records in DT_v is NS_MSG_v .

It triggers $ALGO_V$ algorithm to detect the misbehaving vehicle after the expiry of the timer. The function of this algorithm is elaborated for S_MSG_j in Q_v . It compares $MESS_j$ of R_j with the value of the $MESS$ attribute field of all other records in DT_v . In case of match it increases $Match_j$ counter and in case of mismatch it increases $Mismatch_j$ counter by 1 after each comparison. If $Match_j = Mismatch_j$ it ignores R_j . If $Match_j > Mismatch_j$ the $Algo_V$ inserts $TRUE$ in the $REMARK_j$ attribute field of R_j . Otherwise it inserts $FALSE$ in the $REMARK_j$ attribute field of R_j , inserts E_VIN_j in a CRL (CRL_v) and increases a counter ($NEVIN_CRL_v$) by 1. The $Algo_V$ repeats the same steps of operation for all the records in DT_v to generate CRL_v . V_v sends CRL_v to its parent BS . The maximum number of E_VINs in CRL_v is $NEVIN_CRL_v$.

72.2.2 Function of BS_B

BS_B triggers $ALGO_B$ algorithm for generating a CRL (CRL_B) and for sending CRL_B to CA . The algorithm switches on a timer and initializes it to τ_B . It receives $CRLs$ from the vehicles within its coverage area of BS_B , stores them in a FIFO queue (Q_B), increases a counter (NB_CRL) after receiving each CRL by 1 and assigns the first received CRL to CRL_B . It starts to update CRL_B by performing union operation among the existing CRL_B and the received $CRLs$ in Q_B as soon as NB_CRL becomes equal to 2 till the timer expires. The $ALGO_B$ algorithm sends CRL_B to CA after the expiry of τ_B . The maximum value of NB_CRL is $NO_OF_V_B$ and the maximum number of E_VINs in CRL_B is assumed as $NEVIN_CRL_B$.

72.2.3 Function of CA

CA triggers ALGO_C algorithm for generating CA_CRL and for broadcasting CA_CRL among the vehicles within its coverage area. The algorithm switches on a timer and initializes it to τ_C . It receives CRLs from the BSs within the coverage area of CA, stores them in a FIFO queue (Q_C), increases a counter (NC_CRL) after receiving each CRL by 1 and assigns the first received CRL to CA_CRL. It starts to update CA_CRL by performing union operation among the existing CA_CRL and the received CRLs in Q_C as soon as NC_CRL becomes equal to 2 till the timer expires. The ALGO_C algorithm broadcasts CA_CRL among the vehicles within the coverage area of CA after the expiry of τ_C . The maximum value of NC_CRL is NO_OF_BS and the maximum number of E_VINs in CA_CRL is assumed as NEVIN_CA_CRL.

72.3 Simulation

The performance of the proposed scheme is evaluated qualitatively and quantitatively. In this section the simulation parameters, qualitative analysis and quantitative analysis of the proposed scheme are considered for discussion.

72.3.1 Simulation Parameters

The size of E_VIN (Size_E_VIN) is 17 characters [9] and the size of each character is assumed as 8 bits (extended ASCII format) in the proposed scheme. Hence Size_E_VIN is 136 bits. The bit wise operation is performed among H_MSG and E_VIN to retrieve MESS from H_MSG. Hence the size of H_MSG (Size_H_MSG) is 136 bits. The size of D_Sig (Size_D_Sig) is 160 bits [9]. The size of S_MSG (Size_SMSG) is (Size_D_Sig + Size_E_VIN + Size_H_MSG + Size_O_MSG) bits. The value of τ_v , τ_B , τ_C , NO_OF_BS, NEVIN_CRL_v, NEVIN_CRL_B, NEVIN_CA_CRL, NO_OF_V_B, NS_MSG_v and Size_SMSG are assumed as 30, 20, 20 s, 3, 400, 400, 1200, 144, 400, and 3,232 bits. The data transmission rate (Data_TR) is assumed as 6 Mb/s [11].

72.3.2 Qualitative Analysis

The qualitative analysis on the basis of communication overhead (COMM_OH), storage overhead (STO_OH) and computation overhead (COMP_OH) is reported in this section. The qualitative performance is evaluated by considering the

maximum length of Q_v , Q_B , and Q_C i.e. Q_v has NS_MSG_v number of S_MSGs , Q_B has $NO_OF_V_B$ number of CRLs and Q_C has NO_OF_BS number of CRLs. The qualitative performance of the proposed scheme is compared with [12] on the basis of STO_OH .

Communication Overhead. The $COMM_OH$ of the proposed scheme is $\sum_{B=1}^{NO_OF_BS} (COMM_OH_B)/Data\ sec$ where $COMM_OH_B$ is the communication overhead of BS_B in bits.

Computation of $COMM_OH_B$. BS_B receives CRLs from $NO_OF_V_B$ number of vehicles within its coverage area. The size of CRL_v is $(NEVIN_CRL_v \times Size_E_VIN)$ bits. Hence $COMM_OH_B$ due to the reception of $NO_OF_V_B$ number of CRLs is $\sum_{v=1}^{NO_OF_V_B} (NEVIN_CRL_v \times Size_E_VIN)$ bits.

BS_B sends CRL_B to CA. The size of CRL_B is $NEVIN_CRL_B \times Size_E_VIN$ bits. Hence $COMM_OH_B$ due to the transmission of CRL_B is $NEVIN_CRL_B \times Size_E_VIN$ bits.

BS_B receives CA_CRL from CA. The size of CA_CRL is $NEVIN_CA_CRL \times Size_E_VIN$ bits. Hence the $COMM_OH_B$ due to the reception of CA_CRL is $NEVIN_CA_CRL \times Size_E_VIN$ bits.

BS_B sends CA_CRL to $NO_OF_V_B$ number of vehicles within its coverage area. Hence $COMM_OH_B$ due to the transmission of CA_CRL is $NO_OF_V_B \times NEVIN_CA_CRL \times Size_E_VIN$ bits.

Hence $COMM_OH_B = \sum_{v=1}^{NO_OF_V_B} (NEVIN_CRL_v \times Size_E_VIN) + Size_E_VIN \times (NEVIN_CRL_B + NEVIN_CA_CRL (1 + NO_OF_V_B))$ bits.

Storage Overhead. The STO_OH of the proposed scheme is the sum of STO_OH of CA (STO_OH_CA) and STO_OH of NO_OF_BS number of BSs under CA (STO_OH_BS).

STO_OH_CA is due to the maintenance of NO_OF_BS number of CRL_B in Q_C . Hence $STO_OH_CA = \sum_{B=1}^{NO_OF_BS} NEVIN_CRL_B \times Size_E_VIN$ bits.

STO_OH_BS is $\sum_{B=1}^{NO_OF_BS} (STO_OH_B)$ bits where STO_OH_B is STO_OH of BS_B .

STO_OH_B is due to the maintenance of Q_B and $NO_OF_V_B$ number of Q_v .

Now Q_B has $NO_OF_V_B$ number of CRLs and hence size of Q_B ($Size_Q_B$) is $\sum_{v=1}^{NO_OF_V_B} (NEVIN_CRL_v) \times Size_E_VIN$ bits.

Q_v has NS_MSG_v number of S_MSGs and hence the size of Q_v ($Size_Q_v$) is $\sum_{j=1}^{NS_MSG_v} Size_SMSG_j$ bits, where $Size_SMSG_j$ is the size of S_MSG_j .

Hence $STO_OH_B = Size_Q_B + \sum_{v=1}^{NO_OF_V_B} (Size_Q_v)$ bits.

Computation Overhead. The $COMP_OH$ of the proposed scheme is the sum of $COMP_OH$ of CA ($COMP_OH_CA$) and $COMP_OH$ of NO_OF_BS number of BSs under CA ($COMP_OH_BS$).

$COMP_OH_CA$ is for updating NC_CRL and for performing union operation among the received CRLs from NO_OF_BS number of BSs.

Now $COMP_OH$ of updating NC_CRL for NO_OF_BS times is $O(NO_OF_BS)$.

The COMP_OH of performing union operation among NO_OF_BS number of CRLs is $O(\lceil \log(\text{NO_OF_BS}) \rceil)$.

COMP_OH_BS is $O\left(\sum_{B=1}^{\text{NO_OF_BS}} (\text{COMP_OH}_B)\right)$ where COMP_OH_B is the sum of COMP_OH of BS_B and NO_OF_V_B number of vehicles under it.

COMP_OH of BS_B is for updating NB_CRL and for performing union operation among the received CRLs from NO_OF_V_B number of vehicles.

Now COMP_OH of updating NB_CRL for NO_OF_V_B times is $O(\text{NO_OF_V}_B)$ and of performing union operation among NO_OF_V_B number of CRLs is $O(\lceil \log(\text{NO_OF_V}_B) \rceil)$.

The COMP_OH of NO_OF_V_B number of vehicles is the sum of searching overhead of CA_CRL for E_VIN, execution overhead of ALGO_MESS and ALGO_V.

Now COMP_OH of V_v for searching CA_CRL for NS_MSG_v number of E_VINs corresponding to NS_MSG_v number of received S_MSGs in Q_v is $O(\text{NS_MSG}_v \times \text{NEVIN_CA_CRL})$, for executing ALGO_MESS algorithm for NS_MSG_v number of S_MSGs is $O(\text{Number_BWO} \times \text{Size_H_MSG} \times \text{NS_MSG}_v)$ and for executing ALGO_V algorithm for NS_MSG_v number of records in DT_v is $O(\text{NS_MSG}_v^2)$.

Hence COMP_OH of NO_OF_V_B number of vehicles under BS_B is $O\left(\sum_{v=1}^{\text{NO_OF_V}_B} ((\text{NS_MSG}_v) \times \text{NEVIN_CA_CRL}) + (\text{Number_BWO} \times \text{Size_H_MSG} \times \text{NS_MSG}_v) + (\text{NS_MSG}_v^2)\right)$.

Figures 72.1, 72.2 and 72.3 show the plot of COMM_OH, STO_OH and COMP_OH versus number of vehicles in VANET. STO_OH of the present scheme is less than that in [12] as observed from Fig. 72.2. The number of vehicles per BS increases with the number of vehicles in VANET which in turn increases COMM_OH, STO_OH and COMP_OH as observed from Figs. 72.1, 72.2 and 72.3 respectively.

72.3.3 Quantitative Analysis

The performance of the proposed scheme is studied quantitatively on the basis of CA_CRL distribution time (CA_CRL_DT) and it is compared with [3, 4]. The CA_CRL_DT is the time which is required to broadcast CA_CRL by CA among the vehicles within its coverage area. It is determined during simulation. The quantitative performance of the proposed scheme is also studied on the basis of delay in detection of misbehaving vehicles (Delay_MV). Delay_MV for V_v (Delay_MV_v) is computed as the sum of waiting time of S_MSGs in Q_v, searching time of CA_CRL for E_VINs corresponding to the received S_MSGs in Q_v, time to execute ALGO_MESS and ALGO_V. So, $\text{Delay_MV} = \sum_{v=1}^{\text{NO_OF_V}_B} \text{Delay_MV}_v$.

Fig. 72.1 COMM_OH versus number of vehicles

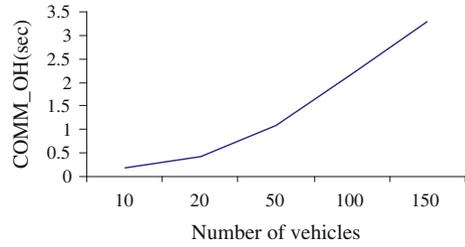


Figure 72.4 shows the plot of CA_CRL_DT versus Size_CA_CRL. The Size_CA_CRL is the size of CA_CRL and it is computed as NEVIN_CA_CRL \times Size_E_VIN bits. It can be observed from Fig. 72.4 that CA_CRL_DT increases with Size_CA_CRL both in the present scheme and in [4]. In [4] Size_CA_CRL is 1 Mbyte and CA_CRL_DT is 300 s. The curve is a straight line as Size_CA_CRL is constant to 1 Mbyte. In the present work Size_CA_CRL varies dynamically with NEVIN_CA_CRL and hence CA_CRL_DT increases slowly with Size_CA_CRL. In the present work CA_CRL_DT is about 1 s.

Figure 72.5 shows the plot of CA_CRL_DT versus number of vehicles in VANET. CA_CRL_DT depends upon the size of CA_CRL which in turn depends upon the number of misbehaving vehicles in VANET. Hence CA_CRL_DT

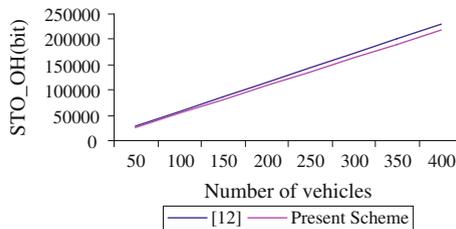


Fig. 72.2 STO_OH versus number of vehicles

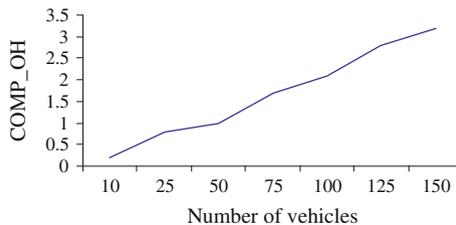


Fig. 72.3 COMP_OH versus number of vehicles

Fig. 72.4 CA_CRL_DT versus Size_CA_CRL

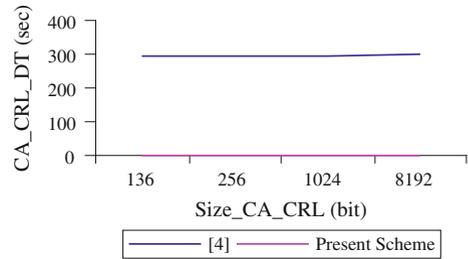


Fig. 72.5 CA_CRL_DT versus number of vehicles

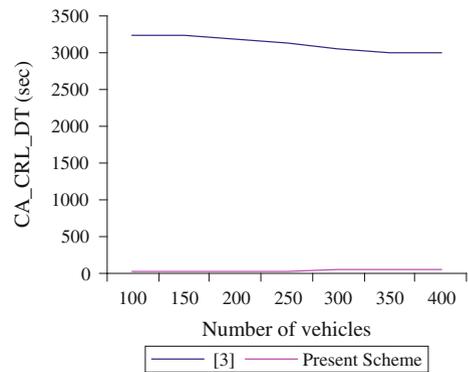
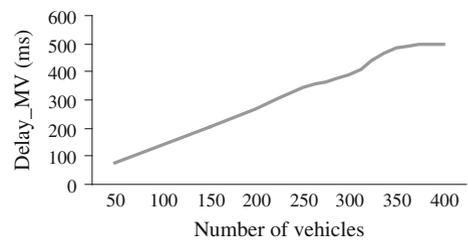


Fig. 72.6 Delay_MV versus number of vehicles



increases slowly with the number of vehicles as observed from Fig. 72.5. In [3] CA_CRL_DT is approximately 3,000 s whereas in the present work CA_CRL_DT is almost 3 s.

Figure 72.6 shows the plot of Delay_MV versus number of vehicles in VANET. It can be observed from Fig. 72.6 that Delay_MV increases with the number of vehicles as per its definition.

72.4 Conclusion

Each vehicle detects and revokes misbehaving vehicles from VANET during V2V communication. The CA_CRL is created by CA after collecting the CRLs from the BSs within its coverage area. Each BS creates the CRL after collecting the CRLs from the vehicles within its coverage area.

The CA may verify the revocation decision of the vehicles before creating CA_CRL. The performance of the proposed scheme may be studied by varying the value of the other parameters and by incorporating the fuzzy logic concept.

References

1. Aslam, B., Zou, C.C.: Distributed certificate architecture for VANET. *Sigcomm* (2009)
2. Papadimitratos, P., Buttyan, L., Holczer, T., Schoch, E., Freudiger, J.: Secure vehicular communication systems: design and architecture. *IEEE Commun. Mag.* **46**, 100–109 (2008)
3. Kargl, F., Papadimitratos, P., Buttyan, L., Muter, M., Wiedersheim, B., Schoch, E., Thong, T.V., Calandriello, G., Held, A., Kung, A., Hubaux, J.P.: Secure vehicular communication systems: implementation, performance and research challenges. *IEEE Commun. Mag.* **46**, 110–118 (2008)
4. Nowatkowski, M.E., Owen, H.L.: Scalable certificate revocation list distribution in vehicular ad hoc networks. *IEEE Globecom, Workshop on Seamless Wireless Mobility* (2010)
5. Papadimitratos, P., Mezzour, G., Hubaux, J.P.: Certificate revocation list distribution in vehicular communication systems. *Proceedings of the Fifth ACM International Workshop on Vehicular Internetworking*, pp. 86–87 (2008)
6. Samara, G., Al-Salihy, W.A.H., Sures, R.: Efficient certificate management in VANET. *Second International Conference on Future Computer and Communication*, vol. 3, pp. 750–754 (2010)
7. Boutaba, R., Aib, I.: Policy-based management: a historical perspective. *J. Netw. Syst. Manage.* **15**, 447–480 (2007)
8. Duque, O.F.G., Hadjiantonis, A.M., Pavlou, G., Howarth M.: adaptable misbehavior detection and isolation in wireless ad hoc networks using policies. *IFIP/IEEE International Symposium on Integrated Network Management* (2009)
9. Mondal, A., Mitra, S.: Identification, authentication and tracking algorithm for vehicles using VIN in centralized VANET. *International Conference on Advances in Communication, Network, and Computing*, Springer LNICST, vol. 108 (2012)
10. Kherani, A., Rao, A.: Performance of node-eviction schemes in vehicular networks. *IEEE Trans. Veh. Technol.* **59**, 550–558 (2010)
11. Towards Effective Vehicle Identification: The NMVTRC's strategic framework for improving the identification of vehicles and components (2004)
12. Huang, J.L., Yeh, L.Y., Chien, H.Y.: ABAKA: an anonymous batch authenticated and key agreement scheme for value-added services in vehicular ad hoc networks. *IEEE Trans. Veh. Technol.* **60**, 248–262 (2011)

Chapter 73

A Novel Steganalysis Method Based on Histogram Analysis

Bismita Choudhury, Rig Das and Arup Baruah

Abstract Steganalysis is the art of detecting hidden messages embedded inside Steganographic Images. Steganalysis involves detection of steganography, estimation of message length and its extraction. Recently Steganalysis receives great deal of attention from the researchers due to the evolution of new, advanced and much secured steganographic methods for communicating secret information. This paper presents a universal steganalysis method for blocking recent steganographic techniques in spatial domain. The novel method analyses histograms of both the cover and suspicious image and based on the histogram difference it gives decision on the suspicious image of being stego or normal image. This method for steganalysis extracts a special pattern from the histogram difference of the cover and . By finding that specific pattern from the histogram difference of the suspicious and cover image it detects the presence of hidden message. The proposed steganalysis method has been experimented on a set of stego images where different steganographic techniques are used and it successfully detects all those stego images.

Keywords Steganalysis · Steganography · Histogram · PSNR

B. Choudhury (✉) · A. Baruah
Department of Computer Science and Engineering and Information Technology,
Don Bosco College of Engineering and Technology, Guwahati 781017, Assam, India
e-mail: bismi.choudhury@gmail.com

A. Baruah
e-mail: arup.baruah@gmail.com

R. Das
Department of Computer Science and Engineering, National Institute of Technology,
Rourkela 769008, Orissa, India
e-mail: rig.das@gmail.com

73.1 Introduction

The battle between Steganography and Steganalysis never ends. For hiding secret message or information, Steganography provides a very secure way by embedding them in unsuspecting cover media such as image, text or video [1]. As a counter action Steganalysis is emerging out as a process of detection of steganography. Steganalysis refers to the science of discrimination between stego-object and cover-object. Steganalysis detects the presence of hidden information without having any knowledge of secret key or algorithm used for embedding the secret message into the cover image [2].

In the general process of steganalysis, steganalyzer simply blocks the stego image and sometimes try to extract the hidden message. Figure 73.1 shows the block diagram of the generic steganalysis process. Generally, Steganalysis techniques are classified into two broad categories: specific and universal blind steganalysis. The targeted steganalysis process is designed for some specific steganographic methods where all features of that particular steganographic method are well known. On the other hand, universal blind steganalysis process uses combination of features to detect arbitrary steganographic methods [3, 4].

Steganalysis can be achieved by applying various image processing techniques like image filtering, rotating, cropping etc. Also it can be achieved by coding a program that examines the stego-image structure and measures its statistical properties, e.g., first order statistics (histograms) or second order statistics (correlations between pixels, distance, direction) [4].

This paper, presents a novel steganalysis method which uses histogram difference for detection of steganography in spatial domain. Here a special pattern in the histogram difference of suspicious image and cover image is utilized for the detection purpose.

This paper is organized as follows. Section 73.2 reviews some previous work done in steganalysis. The proposed novel steganalysis method is explained in Sect. 73.3. Simulation and results are shown in Sects. 73.4 and 73.5 concludes.

73.2 Related Work

Many research works have been carried out on steganalysis till now. Based on the domain of message embedding (Spatial or Frequency domain) different methods are employed to detect presence of steganography. Some of them are as follows.

73.2.1 *RS Steganalysis*

Fridrich et al. described a reliable and accurate method for detecting Least Significant Bit (LSB) based steganography [5]. For performing RS Steganalysis they

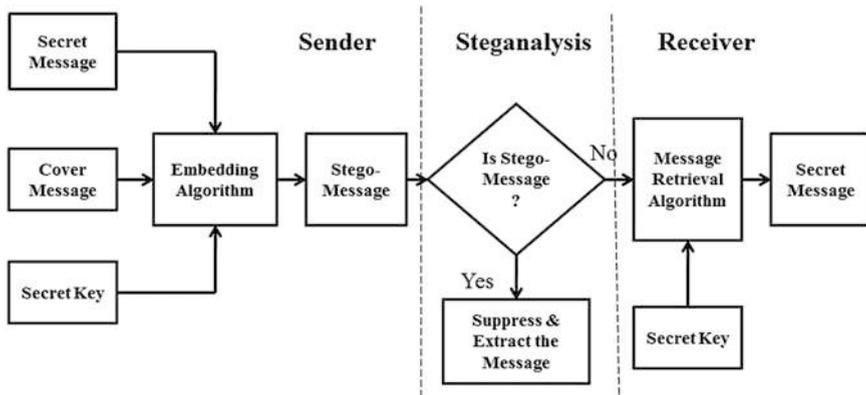


Fig. 73.1 Block diagram of Steganalysis

divided the image pixels into three groups—Regular, Singular and Unchanged group. In normal image number of regular groups is greater than that of singular group. But after embedding any data in the image, Regular and Singular group of pixels have a tendency of becoming equal. Based on this characteristic they proposed RS steganalysis technique for attacking steganography. Here detection is more accurate for messages that are randomly scattered in the stego-image than for messages concentrated in a localized area of the image.

73.2.2 Breaking F5 Algorithm

Fridrich et al. presented a steganalysis method to reliably detect messages (and estimate their size) hidden in JPEG images using the steganographic algorithm F5 [6]. The estimation of the cover-image histogram from the stego-image is the key point. This is done by decompressing the stego-image, cropping it by four pixels in both directions to remove the quantization in the frequency domain, and recompressing it using the same quality factor as the stego-image. The number of relative changes introduced by F5 is determined using the least square fit by comparing the estimated histograms of selected DCT coefficients with those of the stego-image.

73.2.3 Histogram Estimation Scheme for Defeating Pixel Value Differencing Steganography Using Modulus Function

In this paper Jeong-Chun Joo Kyung-Su Kim and Heung-Kyu Lee presented a specific steganalysis method to defeat the modulus Pixel Value Differencing

(PVD) steganography [7]. By analyzing the embedding process they provided three blind Support Machines (SMs) for the steganalysis and each are used for checking three different features. SM1: the fluctuations around the border of the sub range, SM2: the asymmetry of the stego PVD histogram, and SM3: the abnormal increase of the histogram value. The Support Vector Machine (SVM) classifier is applied for the classification of the cover and stego images. Here Original histogram is estimated from the suspicious image using two novel histogram estimation schemes (HES): a curve-fitting method and a histogram reverse-tracing method those work without the cover image.

73.2.4 Steganalysis by Subtractive Pixel Adjacency Matrix

Tomas Pevny and Patrick Bas and Jessica Fridrich presented a method for detection of steganographic method LSB matching [8]. By modeling the differences between adjacent pixels in natural images, the method identifies some deviations those occur due to steganographic embedding. For steganalysis a filter is used for suppressing the image content and exposing the stego noise. Dependences between neighboring pixels of the filtered image are modeled as a higher-order Markov chain. The sample transition probability matrix is then used as a vector feature for a feature-based steganalyzer implemented using machine learning algorithms.

73.3 A Novel Method for Steganalysis Using Histogram Analysis

In this paper we proposed a novel steganalysis technique for detection of steganography in spatial domain based on the histogram analysis of the cover and the suspicious image. The schematic diagram of the whole process is given in Fig. 73.2.

The main goal in here is to develop a steganalysis method which is able to block most of the recently developed steganographic algorithms with a good accuracy. The novel algorithm first finds the histograms of both the cover and suspicious image. Then it uses difference values of both the histograms to detect the stego-image.

73.3.1 Histogram Difference

Image histogram proves to be one of a good feature for analyzing the difference between cover image and stego image. In general, histograms of cover image and stego image have some significant differences that help in discriminating between

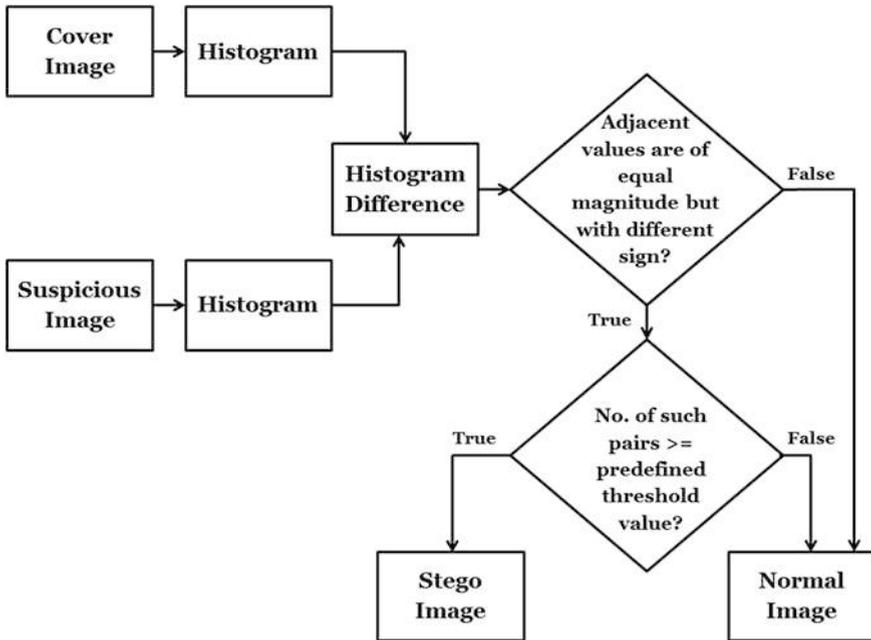


Fig. 73.2 Block diagram of proposed steganalysis method

cover and stego image. In steganography, while embedding secret data in a cover image by modifying the Least Significant Bits (LSBs) of the cover image, some of the pixel values of the cover image get changed and thereby the histogram of the stego image acquires some variations from that of the cover image. If we find the histogram difference of both the cover and stego image we can observe that some of the difference values possess same magnitude to their adjacent values but of different signs (For e.g. 2, -2; -35, 35; ... etc.). But this kind of pattern is not found in the histogram difference between cover and noisy image or any processed image.

The Table 73.1 shows the histogram difference values of the cover image with stego image (LSB embedding) and noisy image introduced with Gaussian noise tested on the Lenna image. From the table we can see that the most of the adjacent difference values are having same magnitude but with different sign only in case of stego image (For e.g. -2, 2; -48, 48; -132, 132), but not in case of noisy image. In this way the steganalysis method tries to find out such pairs in the histogram difference of the cover and the suspicious image and based on this characteristic stego images are detected.

Table 73.1 Histogram difference of cover image with stego image and noisy image

Histogram difference of cover and stego image	Histogram difference of cover and noisy (Gaussian noise) image
-2	-37,079
2	-2,101
-9	-2,180
9	-2,204
-48	-2,179
48	-2,048
-58	-1,747
58	-1,662
-152	-1,454
152	-1,088
-132	-711
132	-383
-266	120
266	601

73.3.2 Proposed Novel Algorithm for Steganalysis

Algorithm

Input: $M \times N$ Suspicious Image and $M \times N$ Cover Image.

Output: Decision whether the Suspicious Image is a Stego Image or not.

- Step-1: Read both the Cover and Suspicious Image and store their intensity values of different pixels in two different arrays.
- Step-2: Find histograms of both the Cover and Suspicious Image.
- Step-3: Plot both the histograms in a single plot and find the difference.
- Step-4: In the different values, if there are adjacent values those are same in magnitude but different in sign then increment a counter.
- Step-5: Repeat Step 4 until all the difference values are checked and the counter incremented accordingly.
- Step-6: Set a threshold value of the counter and if the counter value goes beyond the threshold value then detect the Suspicious Image as the Stego Image else as the Normal Image.
- Step-7: End.

73.4 Simulation and Results

Some experiments are carried out to check the capability and efficiency of the novel steganalysis process. This method is capable of detecting stego image where most of the newly developed steganographic algorithms are used. The proposed

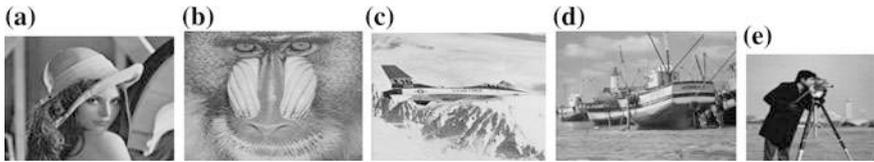


Fig. 73.3 a–d Four cover images for training, e Secret image/message. a Lenna, b Baboon, c Airplane, d Boat, e Cameraman

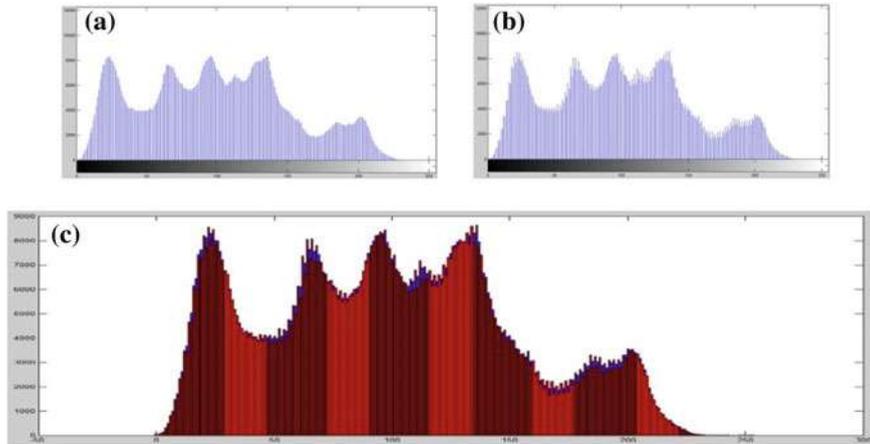


Fig. 73.4 a Histogram of cover image of Lenna, b Histogram of stego image using LSB replacement, c Histogram difference of cover and stego image

steganalysis algorithm is tested on six steganographic algorithms in spatial domain, viz. Least Significant Bit (LSB) replacement, LSB matching, Steganography based on Huffman Encoding, Wavelet Obtained Weight (WOW), Universal Wavelet Relative Distortion for spatial domain (S_UNIWARD) and HUGO.

For the testing purpose, all the simulation has been done in MATLAB 2012 on Windows 7 platform. A set of 8-bit grayscale images of size 1024×1024 are used as cover-image and image of size 256×256 are used as the secret image to form the stego-image. Figure 73.3a–d shows the four original cover images (Here test results are shown only for Lenna Image) and Fig. 73.3e shows the secret image used to embed using LSB replacement [8], LSB matching [8] and Steganography based on Huffman Encoding [9]. For the steganographic algorithms S_UNIWARD [10], WOW [10] and HUGO [10] randomly generated message bits are used to create stego-image. The histogram of the cover image is used to compare with the histogram of the stego image created for testing the proposed steganalysis method. The novel steganalysis algorithm successfully detects the stego-image by analyzing the histogram difference of both suspicious and cover image.

Figure 73.4a shows the histogram of Lenna image, Fig. 73.4b shows histogram of Lenna image after using LSB replacement steganography in which LSBs of

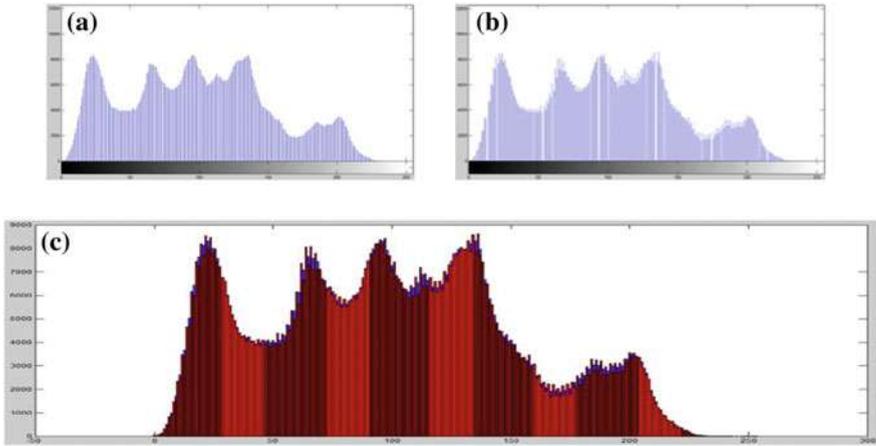


Fig. 73.5 a Histogram of cover image of Lenna, b Histogram of stego image using LSB matching, c Histogram difference of cover and stego image

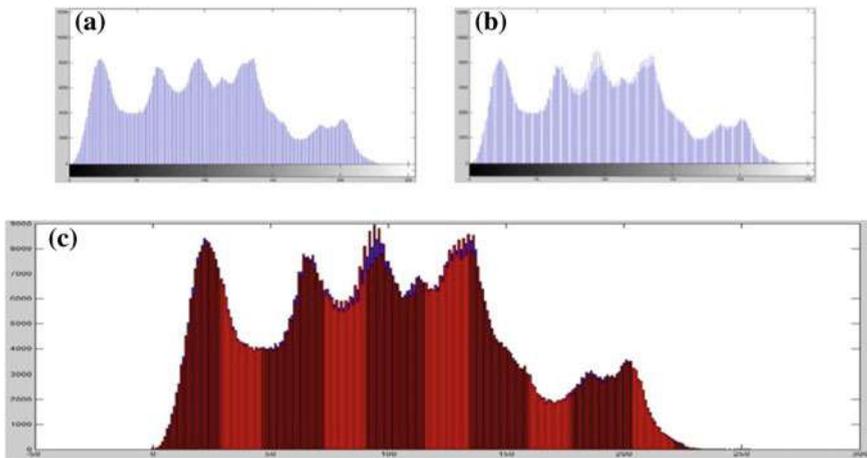


Fig. 73.6 a Histogram of cover image of Lenna, b Histogram of stego image created by steganography based on Huffman encoding, c Histogram difference of cover and stego image

individual cover elements are replaced with message bits [8], Fig. 73.4c shows histogram difference of the cover and the stego image.

Figure 73.5a shows the histogram of Lenna image, Fig. 73.5b the histogram of Lenna image after using LSB matching steganography which randomly increases or decreases pixel values by one to match the LSBs with the communicated message bits [8], Fig. 73.5c shows histogram difference of cover and stego image. The recent Steganographic method based on Huffman encoding proposed by Das and Tuithung [9] is also a very much secured method and very few specific

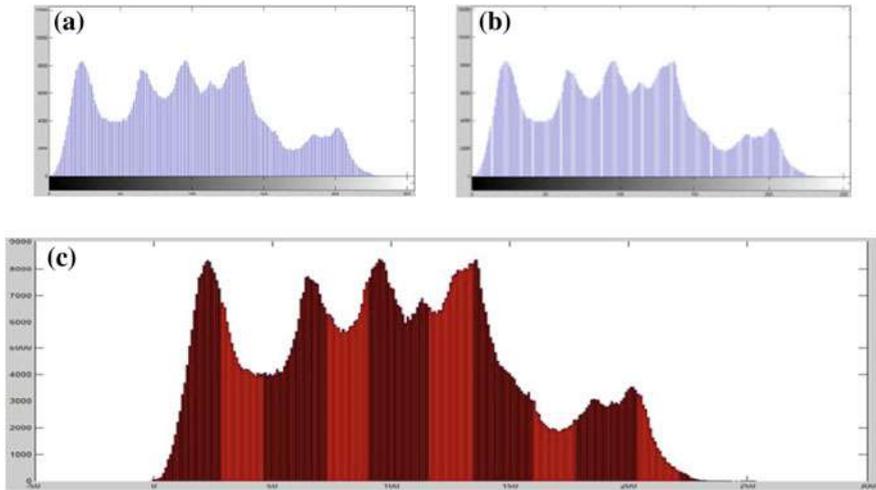


Fig. 73.7 **a** Histogram of cover image of Lenna, **b** Histogram of stego image created using S_UNIWARD method, **c** Histogram difference of cover and stego image

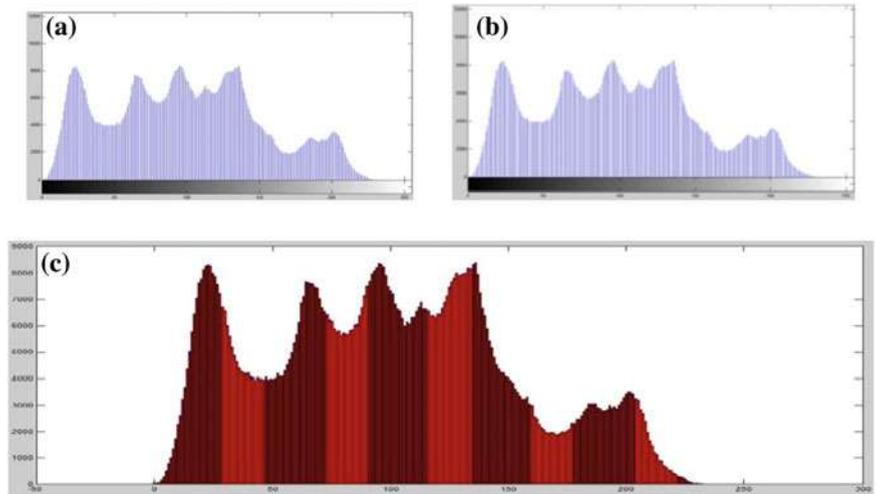


Fig. 73.8 **a** Histogram of cover image of Lenna, **b** Histogram of stego image using steganographic method WOW, **c** Histogram difference of cover and stego image

patterns can be observed in the histogram difference. However, our proposed steganalysis algorithm is able to block it (Fig. 73.6a-c).

Three very recent and secure steganographic algorithms S_UNIWARD [11] (Fig. 73.7a-c), WOW [12] (Fig. 73.8a-c) and HUGO [13] (Fig. 73.9a-c), proposed by Fridrich et al., make a few modifications in the cover image to embed randomly generated message bits. The novel steganalysis method successfully detects those stego images even though they possess few artifacts.

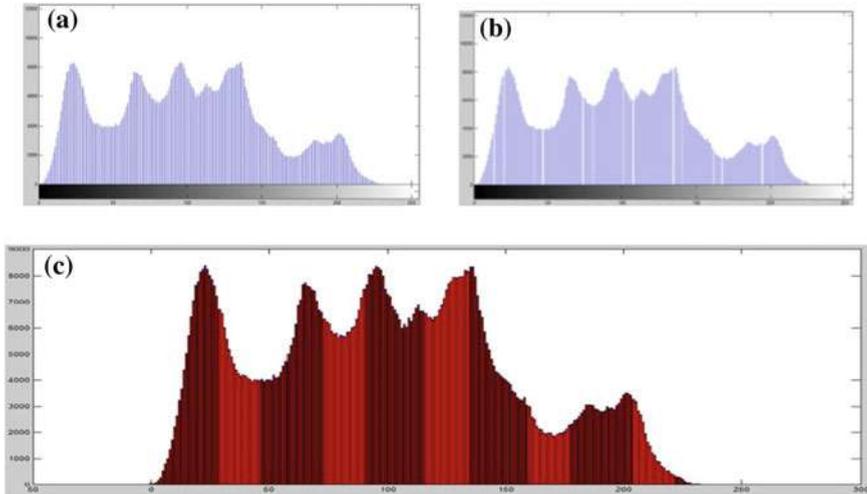


Fig. 73.9 **a** Histogram of cover image of Lenna, **b** Histogram of stego image using steganographic method HUGO, **c** Histogram difference of cover and stego image

Table 73.2 PSNR between the cover and the stego image

Steganographic algorithms	PSNR value between the cover and the stego image (dB)
LSB embedding	+56.88
LSB matching	+56.88
Steganography based on Huffman encoding	+57.43
WOW	+62.69
S_UNIWARD	+62.21
HUGO	+61.92

From the Peak Signal to Noise Ratio (PSNR) values, shown in Table 73.2, it can be seen that the most of the used steganographic methods have done less modification to the cover image which is very difficult to get noticed. However, the proposed steganalysis method successfully blocks the stego images where these steganographic techniques are applied.

73.5 Conclusion

In this paper, we have proposed a universal steganalysis method that checks the histogram difference of the suspicious image with that of the cover image to get adjacent difference values having same magnitude but of different sign. This method

has a great capability of detecting stego images even though very small changes are done in the cover image. Experimental results show that it can block from generic LSB modification techniques to much secured recent steganographic methods. The PSNR values, shown in the Table 73.2, for tested stego images using different steganographic methods depicts that the tested steganographic methods are efficient methods.

Most of the steganalysis algorithms are targeted methods to attack specific steganographic techniques. So in the small group of the universal blind steganalysis this novel algorithm provides a new addition. In future we will work on the steganalysis of the steganography in frequency domain. Then we would like to develop a universal steganalysis method to detect stego images irrespective of the data embedding domain.

References

1. Johnson, F.N., Jajodia, S.: Exploring steganography: seeing the unseen. IEEE Computer Society Press. **31**(2), 26–34 (1998)
2. Fridrich, J., Goljan, M.: Practical steganalysis of digital images—state of the art. In: Proceedings of Electronic Imaging, SPIE, vol. 4675, pp. 1–13 (2002)
3. Lou, D.C., Hu, C.H., Chiu, C.C.: Steganalysis of histogram modification reversible data hiding scheme by histogram feature Coding. Int. J. Innov. Comput. Inf. Control **7**, 11 (2011)
4. Cheddad, A., Condell, J., Curran, K., Kevitt, M.P.: Digital image steganography: survey and analysis of current methods. Elsevier Signal Process. **90**, 727–752 (2010)
5. Fridrich, J., Goljan, M., Du, R.: Reliable detection of LSB steganography in grayscale and color images. In: Proceedings of ACM, Special Session on Multimedia Security and Watermarking, Ottawa, Canada, October 5 (2001)
6. Fridrich, J., Goljan, M., Hoge, D.: Steganalysis of jpeg images: breaking the F5 algorithm. In: Proceedings of the 5th Information Hiding Workshop, Springer, vol. 2578, pp. 310–323 (2002)
7. Joo, C.J., Kim, S.K., Lee, K.H.: Histogram estimation-scheme-based steganalysis defeating the steganography using pixel-value differencing and modulus function. Opt. Eng. **49**, 077001 (2010)
8. Pevny, T., Ba, P., Fridrich, J.: Steganalysis by subtractive pixel adjacency matrix. In: ACM Multimedia and Security Workshop, Princeton, NJ, September 7–8, pp. 75–84 (2009)
9. Das, R., Tuithung, T.: A novel steganography method for image based on huffman encoding. In: 3rd IEEE National Conference on Emerging Trends and Applications in Computer Science (NCETACS—2012), pp. 14–18 (2012)
10. Steganography codes for windows. http://dde.binghamton.edu/download/stego_algorithms/
11. Holub, V., Fridrich, J.: Digital image steganography using universal distortion. In: ACM Workshop on Information Hiding and Multimedia Security, June (2013)
12. Holub, V., Fridrich, J.: Designing steganographic distortion using directional filters. In: IEEE Workshop on Information Forensic and Security (WIFS), Tenerife, Spain, December (2012)
13. Filler, T., Fridrich, J.: Gibbs construction in steganography. IEEE Trans. Inf. Forensics and Security. **5**(4), 705–720 (2010)

Chapter 74

Pattern Recognition Techniques: Studies on Appropriate Classifications

Sasan Karamizadeh, Shahidan M. Abdullah, Mazdak Zamani
and Atabak Kherikhah

Abstract Pattern recognition techniques are divided into categories of supervised, unsupervised and semi supervised. Supervised pattern recognition methods are utilized in the examination of various sources' chemical data such as sensor measurements, spectroscopy, and chromatography. The unsupervised classification techniques use algorithms to classify and analyze huge amounts of raster cells. Semi-Supervised Learning is an approach that is in the middle ground between supervised and unsupervised learning and guarantees to be better at classification by involving data that is unlabeled. In this paper, we tried to categories pattern recognition methods and explain about each of them and we compared supervised method with unsupervised method in terms of types and location of features.

74.1 Introduction

Pattern recognition techniques are divided into categories of supervised, unsupervised and semi supervised. This is dependent on the analyst's intention of the information that needs to be utilized or that is available regarding the samples comprising of the data matrix. In the supervised methods, or the classification method, prior description is made on the classes as the concept or the attribute used

S. Karamizadeh · S.M. Abdullah (✉) · M. Zamani · A. Kherikhah
Advanced Informatics School (AIS), Universiti Teknologi Malaysia,
54100 Kuala Lumpur, Malaysia
e-mail: mshahidan@ic.utm.my

S. Karamizadeh
e-mail: ksasan2@live.utm.my

M. Zamani
e-mail: mazdak@utm.my

A. Kherikhah
e-mail: katabak2@live.utm.my

to classify the samples into subsets are already known [1]. In the unsupervised method, the classification is removed by considering only the variations and resemblances among the samples, without utilizing any of their details. The semi-supervised method is in the middle ground between the supervised and unsupervised analysis and assures to be a better classification using the non-labeled details [2]. Figure 74.1 Classification of Pattern-Recognition Techniques.

74.2 Supervised Methods

Supervised pattern recognition methods are utilized in the examination of various sources' chemical data such as sensor measurements, spectroscopy, and chromatography. Various supervised techniques exist which have been widely utilized in the analytical chemistry [2]. In all the cases, the most suitable method is reliant on the problem that needs to be addressed since the criteria and bases of the techniques differ significantly according to the problems faced. As revealed in Fig. 74.1, different criteria can be utilized to apply the supervised methods. Several of the common methods are elaborated below.

74.2.1 *Parametric and Non-parametric Methods*

Metric methods utilize the mathematical models that have parameters that can be adjusted to perform classification of samples. These methods involve Soft Independent Modeling of Class Analogy (SIMCA), Linear Discriminate Analysis (LDA), Discriminate Analysis (DA), and Support Vector Machine (SVM) [3].

Non-parametric techniques do not utilize the parameters according to the mathematical model for sample classifications. Some of the popularly utilized non-parametric techniques include Artificial Neural Networks (ANN), and k-Nearest Neighbors (kNN) [3].

74.2.2 *Discriminate and Class Modeling Assessment*

Supervised pattern recognition methods differentiate the variables' hyperspace that distinguish the samples into various classifications. Utilizing the discriminant methods, when a new sample is put into the hyperspace classifications, it is identified with that classification, however, when it is put outside, this does not happen. There is a lack of an in-between or middle ground [4]. The techniques in use here are KNN, LDA, ANN, and DA. The analysis on class modeling considers the samples that fit the model as part of the class, whereas rejected non-members are the objects that do not fit. In the event of modeling more than one class, three

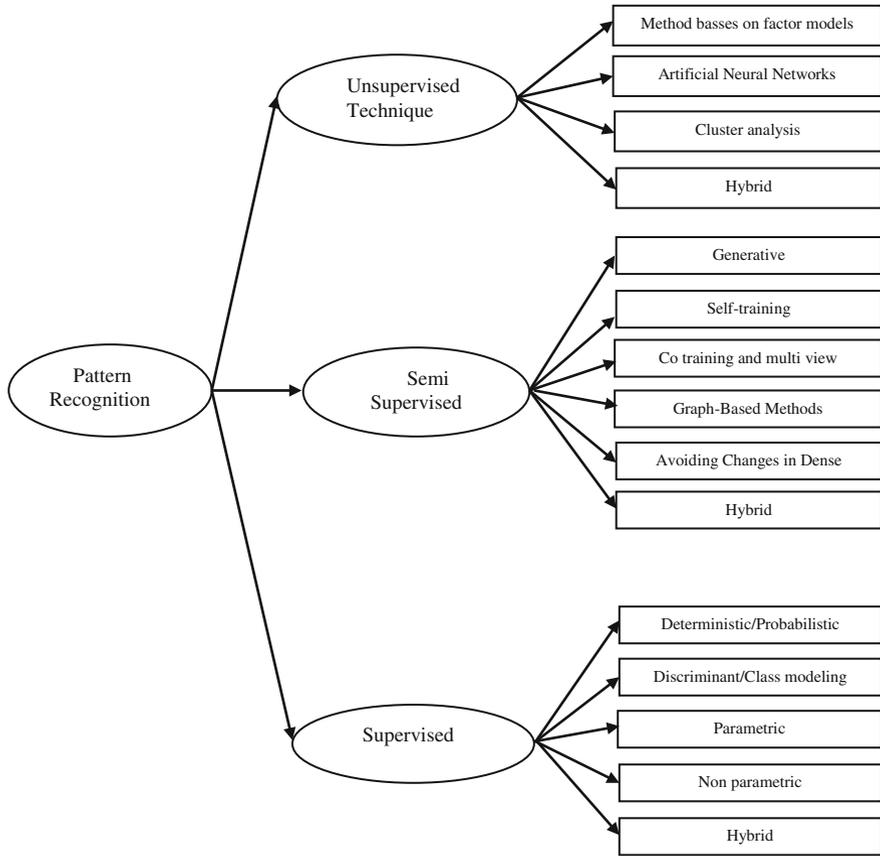


Fig. 74.1 Classification of pattern-recognition techniques

various circumstances can be identified; for example every sample can be designated into a single classification, or more than one classification or none of the classifications [1].

74.2.3 Deterministic/Probabilistic Methods

No statement is made regarding the reliability of the decision when a deterministic system is utilized to designate each sample’s class. Probabilistic methods, however, do measure the classification’s reliability. Deterministic methods are namely KNN, and the probabilistic methods are namely DA, LDA and ANN [1].

74.3 Unsupervised Methods

The unsupervised classification techniques use algorithms to classify and analyze huge amounts of raster cells. These procedures need set values for several of the operating parameters; however, the classifying method goes on without any intervention from the users. The efficacy of the unsupervised techniques is dependent on the basis that the input raster dataset has natural statistical groups of spectral patterns that consist of specific forms of physical characteristics [1]. The entire unsupervised classification techniques, aside from the Simple One-Pass Clustering, utilize the interactive procedure to examine a set of sample input cells and decide on the set of class centers and other related statistical features [5]. All the input raster set is then analyzed, and a classification rule is utilized to designate each raster cell to a defined class. The techniques that are normally utilized can be classified into four significant groups as revealed in Fig. 74.1.

74.3.1 Cluster Analysis

Cluster Analysis or CA has been the most commonly utilized used method of pattern recognition until several years ago. This particular method designates samples to similar clusters based on the level of similarity in the variables (characteristics) which have been utilized to identify the objects, and, simultaneously designate samples that are dissimilar to various other clusters. It is commonly utilized to design a new category of samples being studied study, however, it can also be used to confirm an existing group. Researchers have introduced a complete monograph of cluster analysis based on analytical chemical data [6].

74.3.2 Artificial Neural Networks

ANNs are mathematical methods that follow the workings of the nervous system in humans, by making up pattern recognition models. They are normally very successful in addressing the challenges often faced during the process of classification. ANNs begin from a data training set, that contains characteristics such as spectra or concentration levels that measure samples that are indifferent, to measure probabilities of samples that are a class member (output variables) [3]. ANNs are utilized in both supervised and unsupervised pattern identification, however, since their usage not as simple as CA, their usages are limited somewhat [4].

74.3.3 Techniques According to Factor Models

The aim of these techniques are to constrain the n-dimensional information about objects to a limited and more inclusive aspect. In this way, all the samples can be depicted graphically in a two or three-dimensional (2D or 3D) arena, simplifying identification of the major characteristics. Principal Component Analysis (PCA) are some of the commonly used methods that operate the 2D data tables and multi-set methods [5].

74.4 Semi Supervised

Semi-Supervised Learning (SSL) is an approach that is in the middle ground between supervised and unsupervised learning and guarantees to be better at classification by involving data that is unlabeled. Since getting labeled data is costly and complicated, by causing unlabeled data to be less expensive to get in many applications [6], SSL tries to gain better classification function by utilizing both unlabeled and labeled data. The self-training technique is one of the first algorithms suggested for utilizing the unlabeled data. Another two significant methods include the transductive S3VM and co-training.

74.4.1 Generative

One of the oldest semi-supervised learning techniques is the generative models. The model's assumption is that $p(x, y) = p(y) p(x|y)$ where $p(x|y)$ is a distribution of identifiable mixture; for instance the Gaussian mixture models. Having a huge sum of unlabeled data, the mixture elements can be recognized; and typically, only one labeled example per element is needed to completely decide on the distribution of the mixture [7].

74.4.2 Self-Training

Self-training or decision-directed or self-labeling learning, is the easiest and frequently used SSL approach. This wrapper algorithm utilizes the forecasting of a supervised learning technique to label the unlabeled data. The classifier utilizes its own forecasts to teach itself. Initially, it begins by training a separated hyper plane with only the labeled data. At every stage, the algorithm chooses a portion of the unlabeled samples for labeling, based on the target or a decision task. After that, the technique adds on these objects to the set of training. Lastly, the classifier retrains itself and the procedure is repeats once more [8].

The self-learning algorithm is easy and can be utilized as an algorithm for meta-learning. However, it depends on the goodness- of-fit of the obtained classifier, taking into consideration that errors tend to strengthen themselves. One other drawback of self-learning is the complexity of examining it generally, but there have been several researches on the convergence of particular base learners [9]. Self-training will be used as one of the semi-supervised tactics to develop the models on credit scoring.

74.4.3 Co-training and Multi View Learning

Co-training techniques depend on three assumptions. Firstly, it is stated that must be a natural variables split in two of the subsets. Secondly, every subset must be large enough to train a good classifier. Lastly, the technique presumes that both of the subsets are conditionally independent considering the class. This technique trains two various classifiers; one for each subset and uses just the labeled data. After that, each of the classification tasks categorizes a portion of the unlabeled data and trains the other classifier. Both classifiers will be retrained using this new labeled data handed out by the other classifier (cross information) in an iterative method [10].

74.4.4 Graph-Based Methods

Semi-supervised techniques that are graph-based refer to a graph where the nodes consist of both the labeled and unlabeled samples in the dataset, and edges (may be weighted) show the samples' similarity. These techniques normally assume that there is smoothness of label across the graph. Graph techniques are non-parametric, transductive and discriminative [11, 12].

74.4.5 Avoiding Changes in Dense Areas

Discriminative techniques function directly on $p(y|x)$. This causes the danger of leaving $p(x)$ outside the parameter's estimation loop, if $p(x)$ and $p(y|x)$ do not have similar parameters. Normally, $p(x)$ is all that can be retrieved from the unlabeled data. It is suggested that if $p(x)$ and $p(y|x)$ do not have similar parameters, semi-supervised learning is unable to assist emphasizes this fact [13].

A method of binary classification that locates the optimal linear decision surface between two classifications is known as the Support Vector Machine. The decision surface is a weighted mixture of the supported vectors. The SVM in these utilizations need to be imputed with an individual's images, which will contain one class and the other class will contain images of other individuals besides the first individual. The SVM will then create a linear decision surface [14–16].

74.5 Hybrid

The hybrid models are defined as the models for credit scoring that have been developed by integrating two or more existing models. The benefit of these models is that the creditor can gain from having two or more models aside from reducing the weakness of the model by combining them with other models. However, these techniques are difficult to plan and execute in comparison to other methods that are easier [14] claim that the hybrid method faces faster compared to the traditional concept of neural networks. Several successful credit scoring prototypes of hybrid techniques have also been developed in current years [13]. Examine the hybrid model empirically by implementing two real groups of domain information.

74.6 Comparison Technique Between Supervised, Unsupervised and Semi Supervised

Unsupervised classification techniques use algorithms to classify and analyze huge amounts of raster cells. These procedures need set values for several of the operating parameters; however, the classifying method goes on without any intervention from the users. The efficacy of the unsupervised techniques is dependent on the basis that the input raster dataset has natural statistical groups of spectral patterns that consist of specific forms of physical characteristics. The entire unsupervised classification techniques, aside from the Simple One-Pass Clustering, utilize the interactive procedure to examine a set of sample input cells and decide on the set of class centers and other related statistical features. All the input raster set is then analyzed, and a classification rule is utilized to designate each raster cell to a defined class [17].

The supervised methods of classification are carried out according to the user-defined classes and subsequent representative sample sets. The training raster data sets specify the sample sets, which must be developed before imputing the Automatic Classification procedure. The activation of the Training Data button is carried out when a supervised classification technique is selected, which shows that selection of training is needed and to set the raster. The Feature Mapping procedure offers the tools required to develop a training raster, as shown in the segment known as designing the Training Set Raster. The training sectors are first examined to decide on the statistical characteristics of each classification. In the last stage of classification, every cell in the input raster set is designated to each of the training classes by utilizing a suitable decision rule [17].

Supervised techniques result in superior outcomes when the classification idea is translated into specific groupings that are represented well by training sections and suitable for input raster's. In the Table 74.1 shows advantage and disadvantage of techniques in pattern recognition.

Table 74.1 Advantage and disadvantage of pattern recognition techniques

	Advantage	Disadvantage	Reference
PCA	<ul style="list-style-type: none"> • It is used to reduce the dimension of the data • It gives high accuracy and low computational cost • PCA gave better results for varying poses 	<ul style="list-style-type: none"> • It is very time consuming • High order dependencies still exist in PCA analysis 	[18, 19]
BTC	<ul style="list-style-type: none"> • The algorithm is independent of the size of a face image • Simple image coding technique 	<ul style="list-style-type: none"> • Larger size of the feature vector at BTC level 4 compare with other levels 	[18]
DCT	<ul style="list-style-type: none"> • DCT is used to reduce image information redundancy • DCT has been implemented in a single integrated circuit because of input independency • DCT packing the most information into the fewest coefficients for most natural images, and Minimizing block like appearance 	<ul style="list-style-type: none"> • DCT based features are sensitive to changes in the illumination direction 	[18]
LDA	<ul style="list-style-type: none"> • More efficient if model correct, borrows strength from $p(x)$ 	<ul style="list-style-type: none"> • Bias if model is incorrect 	[20]
SVM	<ul style="list-style-type: none"> • Produce very accurate classifiers • Less over fitting, robust to noise • SVM is defined by a convex optimization problems (no local minima) for which there are efficient methods 	<ul style="list-style-type: none"> • SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes) • Computationally expensive, thus runs slow 	[21]

74.7 Conclusion

In this paper, we have divided pattern recognition techniques in three categories in order to supervised, unsupervised and semi supervised. We have elaborated each category and finally we compare supervised and unsupervised and comparing of methods show that unsupervised method is better than supervised when we do not have good knowledge of the surface, set of training classes are not involved all significantly distinctive types of surface materials, and each training area is not representative of its intended class.

Acknowledgments The work we presented in this paper has been supported by the Universiti Teknologi Malaysia.

References

1. Karamizadeh, S., Abdullah, S.M., Manaf, A.A., Zamani, M., Hooman, A.: An overview of principal component analysis. *J. Sig. Inf. Process.* **4**, 173 (2013)
2. Karamizadeh, S., Abdullah, S.M., Zamani, M.: An overview of holistic face recognition. *IJRCCCT* **2**, 738–741 (2013)
3. Pandya, J.M., Rathod, D., Jadav, J.J.: A survey of face recognition approach. *Int. J. Eng. Res. Appl. (IJERA)* **3**, 632–635 (2013)
4. Delac, K., Grgic, M.: A survey of biometric recognition methods. In: *Proceedings of 46th International Symposium Electronics in Marine 2004, Elmar 2004*, pp. 184–193 (2004)
5. Price, J.R., Gee, T.F.: Face recognition using direct, weighted linear discriminant analysis and modular subspaces. *Pattern Recogn.* **38**, 209–219 (2005)
6. Zhu, X.: Semi-supervised learning. In: Sammut, C., Webb, G.I. (eds.) *Encyclopedia of Machine Learning*, pp. 892–897. Springer, Berlin (2010)
7. Vandewalle, V., Biernacki, C., Celeux, G., Govaert, G.: A predictive deviance criterion for selecting a generative model in semi-supervised classification. *Computat. Stat. Data Anal.* **24**, 220–236 (2013)
8. Carlson, A., Betteridge, J., Wang, R.C., Hruschka Jr, E.R., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining, 2010*, pp. 101–110 (2010)
9. Rosset, S., Zhu, J., Zou, H., Hastie, T.J.: A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances in Neural Information Processing Systems, 2004*, pp. 1161–1168 (2004)
10. Zhou, Z.-H., Li, M.: Semi-supervised regression with co-training. In: *IJCAI, 2005*, pp. 908–916 (2005)
11. Narayanan, H., Belkin, M., Niyogi, P.: On the relation between low density separation, spectral clustering and graph cuts. In: *Advances in Neural Information Processing Systems, 2006*, pp. 1025–1032 (2006)
12. Abokhdair, N.O., Manaf, A.A., Zamani, M.: Integration of chaotic map and confusion technique for color medical image encryption. In: *International Conference on Digital Content, Multimedia Technology and Its Applications, 16 August 2010*, pp. 20–23. Korea (2010)
13. Seeger, M.: Learning with labeled and unlabeled data. Technical report, University of Edinburgh (2001)
14. Pradhan, A.: SUPPORT VECTOR MACHINE-A Survey (2012)
15. Abdullah, S.M., Manaf, A.A., Zamani, M.: Capacity and quality improvement in reversible image watermarking approach. In: *IEEE Networked Computing and Advanced Information Management (NCM)*, pp. 81–85 (2010)
16. Abdullah, S.M., Manaf, A.A.: Multiple layer reversible images watermarking using enhancement of difference expansion techniques. In: *Networked Digital Technologies*, pp. 333–342. Springer, Berlin, Heidelberg (2010)
17. Long, F., Xu, H.: Data mining technique of Acoustic Emission signals under supervised and unsupervised mode. In: *2011 Seventh International Conference on Natural Computation (ICNC), 2011*, pp. 752–755 (2011)
18. Sonal, G., Tanuja, S.S.: A review of feature extraction techniques BTC, DCT, Walsh and PCA with FDM and BDM for face recognition (2013)
19. Bhatt, B.G., Shah, Z.H.: Face feature extraction techniques: a survey. In: *National Conference on Recent Trends in Engineering & Technology, 2011*, pp. 13–14. B.V.M. Engineering College, V.V.Nagar, Gujarat, India (2011)
20. Hastie, T., Tibshirani, R., Friedman, J.: Linear methods for classification. In: *The Elements of Statistical Learning*, pp. 101–137. Springer, New York (2009)
21. Bhavsar, H., Ganatra, A.: Variations of support vector machine classification technique: a survey. *Int. J. Adv. Comput. Res.* **2** (2012)

Chapter 75

Environmental Noise Analysis for Robust Automatic Speech Recognition

N. Sai Bala Kishore, M. Rao Venkata and M. Nagamani

Abstract Most of the speech communication applications viz. telephony, hands-free communication, voice recording, automatic speech recognition, interactive voice response system, human-machine interfaces, etc. that require at least one microphone, desired speech signal is usually contaminated by background noise and reverberation. As a result, the speech signal has to be “cleaned” with digital signal processing tools before it is played out, transmitted, or stored. The noise estimation and reduction techniques will help to clean and attenuate the noise component in speech data, known as Speech Enhancement. In this paper, we recorded the speech in different environmental conditions and estimated the noise signal/background noise distribution in speech. Now the speech is enhanced by using the compliment of Weiner-Hopf optimal filter. And this enhanced speech signal is given for training and testing the Automatic Speech Recognition (ASR) system, which will improve the word accuracy. The analysis of speech and results presented in this paper are produced using MATLAB.

Keywords Speech recognition · Speech enhancement · Noise filtering · Automatic speech recognition (ASR)

N.S.B. Kishore · M.R. Venkata (✉)
Vignan University, Vadlamudi, India
e-mail: mvrao239@gmail.com

N.S.B. Kishore
e-mail: saiit1289@gmail.com

M. Nagamani
University of Hyderabad, Hyderabad, India
e-mail: manidcis@gmail.com

75.1 Introduction

Speech Enhancement [1] means the improvement in the value or quality of speech signal. When applied to speech, this simply means the improvement in intelligibility and/or quality of a degraded speech signal [2] by using speech processing technique. Speech enhancement is a very difficult problem for two reasons. First the nature and characteristics of the noise signals can change dramatically in time and application to application. It is therefore laborious to find versatile algorithms that really work in different practical environments. Second the performance measure can also be defined differently for each application. Two perceptual criteria are widely used to measure the performance: quality and intelligibility. While the former is subjective (it reflects individual preferences of listeners), the latter is objective (it gives the percentage of words that could be correctly identified by listeners). It is very hard to satisfy both at the same time. In fact, it can easily be shown that in the single-channel (one microphone) case and when the degradation is due to the uncorrelated additive noise, noise reduction (quality improvement) is possible at the expense of speech distortion (intelligibility reduction).

75.2 Noise Analysis and Estimation

Environmental robustness [3] is an important area of research in speech recognition. Mismatch between trained speech models and actual speech to be recognized is due to factors like background noise mingled [4] with unvoiced sound. It can cause severe degradation of performance in the context of speech recognition. So analysis and estimating of noise plays a vital in speech recognition performance. Here we estimate noise to signal [5] to the speech recorded wave and do noise reduction from speech wave and this noise reduced wave can be given as input to the speech recognition system. And this will improvise the accuracy of speech recognition of recorded speech. In the application perspective we use this in mobile transmission channel to hear speech without any noise.

The optimal estimate of the clean speech can be achieved by optimizing some criterion, such as the mean-squared error (MSE) [6] between the clean speech and its estimate, the signal-to-noise ratio (SNR), the a posteriori probability of the clean speech given its noisy observations, etc.

75.3 Noise Reduction

The below diagram is the general block diagram of noise reduction system (Fig. 75.1).

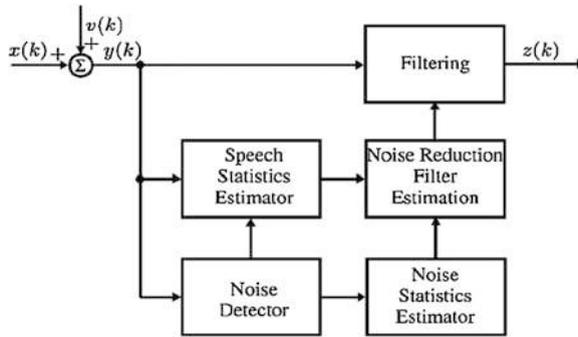


Fig. 75.1 General block diagram of noise reduction system

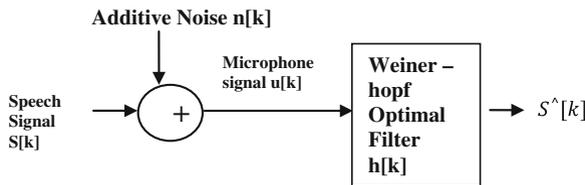


Fig. 75.2 Architecture of proposed algorithm

The model in the above figure begins with a clean speech signal (from a desired speaker), $x(k)$, being corrupted with an unwanted additive noise, $v(k)$. The noisy signal, $y(k)$, which is a superposition of $x(k)$ and $v(k)$ is first processed to determine whether the desired speech is currently present or absent. The noise and speech statistics such as the covariance matrices and power spectral densities are estimated based on the detection results and the input noisy speech. These statistics will be used to estimate a noise reduction filter. This filter can be optimal in the sense that it optimizes some error criterion (e.g., MSE) [7]. It may be suboptimal, where parameters are introduced to better control the quality of the output speech. The estimated filter is applied to the noisy speech to filter out the noise signal [8], thereby producing an output signal, $z(k)$, which is supposed to be an estimate of the clean speech, $x(k)$ (Fig. 75.2).

Coming to the proposed algorithm and we are doing compliment of weiner_Hopf optimal filter algorithm [9]. The below is famous weiner-hopf equation has been used for the cross-correlation and auto-correlation.

$$C_k^{xy} = \sum_{j=M1}^{M2} h_j C_{k-j}^{xx} \quad k = M1, \dots, M2$$

Equation: Weiner-Hopf Equation
Where the shorthand notation

$$C_k^{xy} = \sum_i x_i y_i + k \quad C_l^{xx} = \sum_i x_i x_{i+l}$$

C_k^{xy} is the cross correlation of x, y and C_l^{xx} is the auto correlation of x.

Algorithm

1. Read speech utterance wave.
2. Calculate length of speech utterance denoted by M.
3. Next calculate cross correlation of signal denoted by C_{ss}.
4. Similarly I do cross correlation of white noise denoted by C_{ww}.
5. function [h, e] = WH(C_{ss}, C_{ww}, M).


```

      while abs(e0 - e1) > 1e - 6
          N = N + 1;
          e0 = e1;
          Cxs = Css(M:M + N - 1);
          Cxx = Cww(M:M + N - 1) + Css(M:M + N - 1);
          C_xx = zeros(N);
          for j = 1:N
              for n = 1:N
                  C_xx(j, n) = Cxx(abs(j - n) + 1);
              end
          end
          h = inv(C_xx)*Cxs';
          e1 = Css(M) - h'*Cxs';
      end, N
      
```
6. e = e1;

It returns estimation of noise between filtered signal [10] and original signal.

75.4 The Proposed System Architecture

Here First we take input speech utterance can be given to proposed algorithm and here it removes the noise from the speech signal and this clean speech can be given as input to ASR system and from the ASR system can get recognized text as output (Fig. 75.3).

75.5 Analysis and Results

The implementation process of proposed Environmental noise analysis for robust ASR system with the different data conditions, recognition accuracy is tested. The data is collected in the classrooms and laboratory of U.G and P.G students of

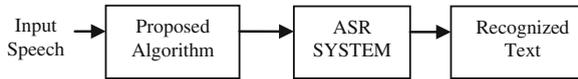


Fig. 75.3 The proposed total system architecture

Department of Computer science engineering with A/C noise, and fan noise conditions, Open space environment in live mode and command modes of ASR system. Here we are specifying the signal which contains background noise, which is considered as a silence and represented with a symbol(*) as shown in REF. and corresponding HYP.

REF: ***** ***** ***** ***** ***** ***** *****

***** NAXMAXSKAARAXMUH (s10)

HYP: PIXTTAX PAXTNAXM IXCCIX DHAXSAXLUHAMDUHNUH

VAXKXAXDIA KAXNNA PAALUH UHAMTAARUH (s10)

SENTENCE 1 (s10)

Correct = 0.0 % 0 (0)

Errors = 800.0 % 8 (8)

SC NAXMAXSKAARAXMUH ==> PAALUH UHAMTAARUH

=====

REF: ***** ***** ***** THXAILUHGUHVAARAXNDHAXRAXMUH

***** THXAILUHGXIA MAATLAADUHDHAAM (s11)

HYP: PUHLIX PAXDHAX KAXNCAI THXAAGAXDAXM

AIKKAXDAX UHAMTAARUH PUHLIX (s11)

SENTENCE 2 (s11)

Correct = 0.0 % 0 (0)

Errors = 233.3 % 7 (15)

SC THXAILUHGUHVAARAXNDHAXRAXMUH ==> KAXNCAI

THXAAGAXDAXM

SC THXAILUHGUHVAARAXNDHAXRAXMUH ==> THXAAGAXDAXM

AIKKAXDAX

SC THXAILUHGXIA ==> AIKKAXDAX UHAMTAARUH

Experiment 1:

Here in this first we observe manually how the speech signal vary to noise [11] and how speech signal parameters can vary from utterance to utterance and for this purpose we recorded telugu aksharas and observe speech utterances manually by using praat and the observed results can be shown in below table. In the Below table we analyzing the background noise (indicated by silence) in telugu (Table 75.1).

Table 75.1 Telugu Aksharas silence duration, pitch, and standard deviation values

Telugu literals	Sound	Silence before (s)	Duration (s)	Silence after (s)	Mean pitch(Hz)	Standard deviation
అ	A	0.333.712	0.287773	0.40275	111.659	7.27
ఆ	Aa	0.40275	0.43363	0.438229	112.543	13.394
ఇ	I	0.438229	0.335078	0.294343	124.86	12.231
ఈ	Ii	0.294343	0.385668	0.476336	110.601	8.641
ఉ	U	0.476336	0.306826	0.356102	118.031	5.351
ఊ	uu	0.356102	0.513786	0.519699	108.09	6.528
ఋ	rx	0.519699	0.269376	0.279888	122.599	3.58
ౠ	rx ~	0.279888	0.466481	0.486848	108.196	8.062
లు	lx	0.486848	0.342305	0.291058	123.191	6.398
ల్లు	lx ~	0.291058	0.574231	0.443485	110.263	7.949
ఎ	e	0.443485	0.312082	0.536124	117.085	7.55
ఏ	ei	0.536124	0.468452	0.487505	106.711	8.831
ఐ	ai	0.487505	0.312739	0.542037	119.433	6.033
ఓ	o	0.542037	0.43363	0.545322	112.89	9.917
ఔ	oo	0.545322	0.501959	0.590656	110.334	9.969
ఌ	ah	0.590656	0.327193	0.487505	121.822	7.891
అం	n'	0.487505	0.237839	0.335078	123.072	8.314
అః	:	0.335078	0.354131	0.579487	102.986	4.553
క	k	0.579487	0.236525	0.603797	112.467	2.623
ఖ	kh	0.603797	0.296314	0.472394	112.155	4.516
గ	g	0.472394	0.37647	0.545322	111.659	7.987
ఘ	gh	0.545322	0.27069	0.495389	109.946	5.172
ఙ	ng	0.495389	0.507215	0.505901	116.89	12.802
చ	ch	0.505901	0.319966	0.442828	108.303	3.898
ఛ	chh	0.442828	0.35216	0.429688	107.088	3.571
జ	j	0.429688	0.496046	0.489476	107.88	5.873
ఝ	jh	0.489476	0.438229	0.52824	106.362	5.193
ఞ	nj	0.52824	0.412605	0.492104	112.888	7.395
ట	t'	0.492104	0.255579	0.577516	108.029	3.065
ఠ	th'	0.577516	0.287116	0.451369	106.922	5.512
డ	d'	0.451369	0.378441	0.488819	109.762	7.249
ఢ	dh'	0.488819	0.362015	0.394209	105.849	5.119
ణ	nd'	0.394209	0.403407	0.50853	110.177	5.202

(continued)

Table 75.1 (continued)

Telugu literals	Sound	Silence before (s)	Duration (s)	Silence after (s)	Mean pitch(Hz)	Standard deviation
త	t	0.50853	0.28843	0.496703	109.322	6.188
థ	th	0.496703	0.385668	0.545979	108.185	7.091
ద	d	0.545979	0.388296	0.429688	108.86	6.628
ధ	dh	0.429688	0.434287	0.429031	108.35	7.549
న	n	0.429031	0.408006	0.568318	109.335	4.356
ప	p	0.568318	0.27726	0.557149	107.476	2.524
ఫ	ph	0.557149	0.346247	0.499331	109.387	7.081
బ	b	0.499331	0.412605	0.484877	108.325	4.649
భ	bh	0.484877	0.423775	0.489476	108.34	5.189
మ	m	0.489476	0.353474	0.467138	112.049	3.002
య	y	0.467138	0.394209	0.462538	106.83	7.472
ర	R	0.462538	0.367271	0.539409	110.831	8.079
ల	L	0.539409	0.427717	0.513786	108.453	3.137
వ	V	0.513786	0.3653	0.470423	107.352	4.041
శ	Sh	0.470423	0.475022	0.502616	109.004	7.41
ష	shh	0.502616	0.423118	0.422461	108.921	7.035
స	S	0.422461	0.478307	0.448084	106.463	5.282
హ	H	0.448084	0.417861	0.461881	107.729	4.934
ఱ	l'	0.461881	0.394209	0.505901	106.295	5.918
క	kshh	0.505901	0.331136	0.542037	110.547	10.407
ఱ	r'	0.542037	0.310768	0.48811	112.801	10.438

Experiment 2:

Here in this experiment how the utterance can be formed from the literals and the formants of these literals can vary and based on these analysis we understand what differentiates noise from speech utterance. And these analyzed results can be shown in the following table (Table 75.2).

Table 75.2 Telugu words analysis based on formants

Telugu	Words	Literals	Start (s)	End (s)	Duration (s)	Pitch (Hz)	F1 (Hz)	F2 (Hz)
అది	Adi							
		a	103.312087	103.471702	0.159614	154.668	624.422568	1685.96567
		di	103.487438	103.716744	0.229305	152.011	328.070656	2254.1126
ఇల్లు	Illu							
		i	172.380741	172.524619	0.143878	173.63	373.177863	2172.44666
		llu	172.533611	172.774157	0.240546	176.3	585.846423	1800.05142
అమ్మ	Aame							
		aa	551.546091	551.721442	0.175351	147.424	844.84012	1295.31013
		me	551.757411	551.959739	0.202328	164.44	558.923428	2023.92283
డాక్టర్	Daakt'ar							
		Daa	504.453758	504.631357	0.177599	141.164	558.482715	1649.39412
		kt'ar	504.862911	505.049502	0.186592	134.192	389.985309	1554.98185
ప్రొఫెసర్	profesar							
		pro	1256.661	1256.8296	0.168607	135.143	714.400342	1533.15624
		fe	1256.83588	1256.97076	0.134885	192.012	760.921128	1926.15001
		sar	1256.991	1257.19557	0.204576	150.149	652.577183	1747.85237

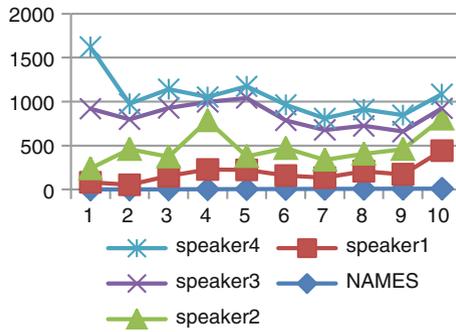


Fig. 75.4 Comparative results diagram of different speakers

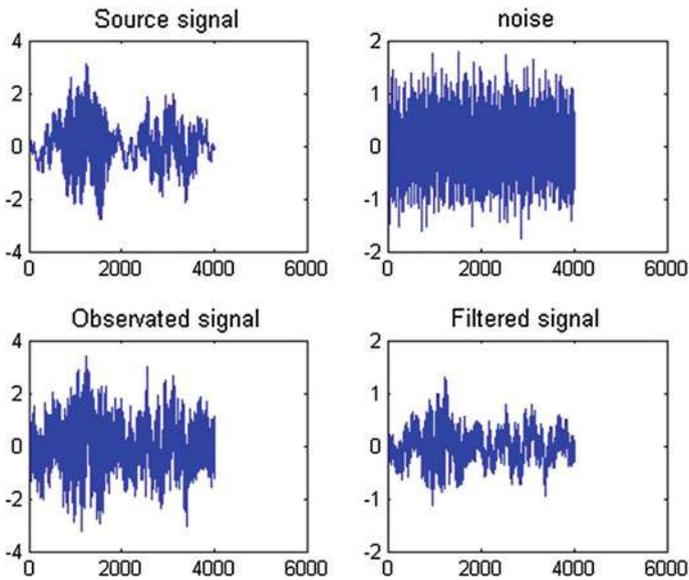


Fig. 75.5 Source signal conversion to filter signal

Experiment 3:

In this experiment, we recorded utterances with different speakers and estimate the noise from the original signal, filtered signal [12] and filter the noise from original signal and we specify this analyzed result in the following table (Figs. 75.4, 75.5 and 75.6) (Table 75.3).

Fig. 75.6 The Noise estimation to source signal

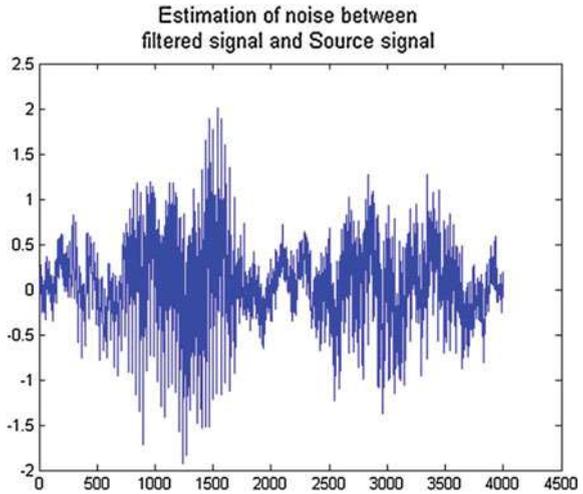


Table 75.3 Noise estimation values of different speakers with accuracy before and after noise reduction (NR)

Estimation of noise between filtered signal and original signal				
Recorded utterances	Speaker1 (s)	Speaker2 (s)	Speaker3 (s)	Speaker4 (s)
1	79.8852	157.8153	681.2529	700.0783
2	50.953	405.217	338.6157	179.7575
3	150.4635	219.2007	555.8127	213.5948
4	223.981	557.311	210.1028	57.0407
5	217.9731	155.5096	660.6022	134.9966
6	151.8165	311.5184	315.4281	178.0364
7	128.1514	202.1541	340.905	131.0829
8	201.2992	196.0372	317.9495	185.5135
9	163.5398	285.7568	200.374	186.6847
10	431.5037	362.5521	120.1197	158.4786
Accuracy before NR (%)	50	40	20	60
Accuracy after NR (%)	70	80	30	90

75.6 Conclusion

The compliment of Weiner-Hopf optimum filter algorithm is giving better results for lab environmental conditions. When it is complex environmental situation like outdoor environments the performance of the system is not up to the level. The system works well for small amount of data.

When the environmental condition is complex situation, the system has to train in such a way that the complex noise structure should be identified and need to be removed from speech utterance. The different environment noise conditions are to be estimated and appropriate filter design analysis has to be studied to clean the speech signal in real time ASR applications.

References

1. Veisi, H., Sameti, H.: Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement. *IET Sig. Process.* **6**(1), 54–63 (2012)
2. Kashiwagi, Y., Suzuki, M., Minematsu, N., Hirose, K.: Audio-Visual feature interaction based on piecewise linear transformation for noise robust automatic speech recognition. In: *ICASSP*, pp. 149–152. *IEEE* (2010)
3. Ghaemmaghami, H., Dean, D., Sridharan, S., McCown, I.: Noise robust voice activity detection using normal testing and time-domain histogram analysis. In: *ICASSP*, pp. 4470–4473. *IEEE* (2010)
4. Dhananjaya, N., Yegnanarayana, B., Senior Member, *IEEE*: Voiced/nonvoiced detection based on robustness of voiced epochs. *IEEE Sig. Process. Lett.* **17**(3) (2010)
5. Benesty, J., Chen, J., Huang, Y., Cohen, I.: *Noise reduction in speech processing*. Springer, Heidelberg (2009)
6. Kim, D.Y., Un, C.K., Kim, N.S.: Speech recognition in noisy environments using first-order vector Taylor series. *Speech Commun.* **24**, 39–49 (1998)
7. Ramirez, J., Yelamos, P., Gorriz, J.M., Segura, J.C.: SVM-based speech endpoint detection using contextual speech features. *Electron. Lett.* **42**(7), 426–428 (2006)
8. Sebastian Seung, H.: Wiener-Hopf equations. Convolution and correlation in continuous time, 9.29 Lecture 3: February 11, 2003 (2003)
9. Verma, A.R., Singh, R.K., Kumar, A., Ranjeet, K.: An improved method for speech enhancement based on 2D-DWT using hybrid weiner filtering. In: *2012 IEEE International Conference on Computational Intelligence and Computing Research* (2012)
10. Samudravijaya, K., Barot, M.: A comparison of public domain software tools for speech recognition. *Workshop on spoken language processing*, pp. 125–131 (2003)
11. Fukane, A.R., Sahare, S.L.: Role of noise estimation in enhancement of noisy speech signals for hearing aids. In: *Computational Intelligence and Communication Networks (CICN)*, pp.648–652. *IEEE* (2011)
12. Ling, G., Yamada, T., Makino, S., Kitawaki, N.: Performance estimation of noisy speech recognition using spectral distortion and snr of noise-reduced speech. In: *TENCON 2013*, *IEEE* (2013)

Chapter 76

Performance Comparison of Selected Classification Algorithms Based on Fuzzy Soft Set for Medical Data

Saima Anwar Lashari and Rosziati Ibrahim

Abstract Medical data is heterogeneous in nature and associated with uncertainties. For that reason, data mining has been assisting physicians in decision making and to cope with the information overload. A considerable amount of literature has been available on medical data classification based on data mining techniques to automate or facilitating the delineation of images. However, from image formation to the final analysis, medical imaging is still facing challenges. New imaging procedures for classification could overcome the inefficiencies and provide more reliable information to the medical experts. Therefore, this paper assesses the performance of selected classification algorithms based on fuzzy soft set for classification of medical data. There are two concepts that underlie the classification in the fuzzy soft set theory namely: classification based on decision making problem and classification based on similarity between two fuzzy soft set. The selected algorithms are evaluated based on two criteria: accuracy and computational time. Moreover, the conducted experiments demonstrated the effectiveness of fuzzy soft set for medical data categorization.

Keywords Medical data · Classification · Data mining · Soft set · Fuzzy soft set

76.1 Introduction

The primary objective of medical data classification is not only to achieve good accuracy but to understand which parts of anatomy are affected by the disease to help clinicians in early diagnosis of the pathology and in learning the progression

S.A. Lashari (✉) · R. Ibrahim

Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, 86400 Batu Pahat, Johor, Malaysia
e-mail: hi120040@siswa.uthm.edu.my; hi100008@siswa.uthm.edu.my

R. Ibrahim
e-mail: rosziati@uthm.edu.my

of a disease. Thus, the concept of data mining was created and to present rising levels of computerization in the information creation process, subsisting a vast amount of time burly human activity with regular techniques that improves accuracy [1]. Most of recent efforts reviewed are more related to the development where ideas are motivated from concepts of pattern recognition, image processing, and computer vision. However, with this in mind, it is important to realize that medical data analysis efforts are heavily influenced by the particular image datasets being utilized and the clinical or biological tasks that underlie the need for medical data analysis.

Thus, for medical data various classification algorithms have been put forward in articles and scientific writings, including bayesian rule [2], nearest- neighbor methods [3, 4], decision tree induction [5], error back propagation [6]. Every one of them has contributed to address problems in data mining. However, no single method has been found to be superior over all others for all datasets [7].

In 1999 Molodtsov [8] introduced a new mathematical tool named soft set theory to deal with uncertain problems. The initial description of the object has an approximate nature and there is no need to introduce the notation of exact solution. The applications of this theory boom in recent years and is extended to, among others, data analysis [9], soft decision making [10], texture classification [11], musical instrument classification [12].

Soft set can work well on the parameters that having binary numbers but still difficult to work with parameters that have a real number. There are many issues in the classification involving real numbers. To overcome this problem Maji et al. [13] offered a more general concept namely fuzzy soft set which can handle parameters in the form of real numbers. Since medical data is associated with uncertainty and most of data is available in the form of real numbers. This motivated us to see the viability of fuzzy soft set theory to perform medical data classification and to see how effective this theory is to handle uncertainty nature of medical data.

The rest of the paper is organized as follows: theoretical background of classification algorithms is detailed in Sect. 76.2. The modeling process presented in Sect. 76.3. Section 76.4 contains the conclusion of this study.

76.2 Classification Algorithms

Classification algorithm normally known as classifier is a method of finding a classification model. The role of classifier is significant as it has to analyse and extract information from numerical vectors into implicit form. However, despite the fact that there are many classification algorithms, in certain circumstances, one classification algorithm may perform better than the other [14]. Many different classification strategies were applied such as nearest neighbor-based approach [4], decision trees [15] as well as support vector machine [16]. Analogous to feature combination, classifier combination has also been a popular way to improve classification performance [4]. The entire algorithm array belongs to the category of supervised learning methods.

Therefore, this seems to indicate that data mining techniques is beginning to be applied widely in the detection and differential diagnosis of many different types of abnormalities in medical data obtained in various examinations by use of different imaging modalities [15].

76.2.1 Fuzzy Soft Set

The concept of fuzzy set was introduced by Zadeh in 1965 [17] to allow elements to belong to a set in a gradual rather than an abrupt way (i.e. permitting memberships valued in the interval $[0, 1]$) instead of in the set $\{0, 1\}$. Ever since then, applications and developments based on this simple concept have evolved to such an extent that it is practically impossible nowadays to encounter any area or problem where applications, developments, products etc. are not based on fuzzy sets [18]. Later, Maji et al. [12] have studied a general concept, namely theory of fuzzy soft set.

There are two important concepts underlying the application of the theory of fuzzy soft set in numerical classification problems.

- concept of decision making problems
- concept of measuring similarity

Based on an application of soft set in a decision making problem presented by Maji et al. [19], Mushrif et al. [10] presented a novel method for classification of natural textures using the notions of soft set theory, all features on the natural textures consist of a numeric (real) data type, have a value between $[0, 1]$ and the algorithm used to classify the natural texture is very similar to the algorithm used by Roy and Maji [13] in the decision making problems. The algorithm was successfully classify natural texture with very high accuracy when compared with conventional classification methods such as Bayes classifier and a minimum distance classifier based on Euclidean distance. He has also proved that the computation time for classification is much less as compared to with ayes classification method.

Later, Lashari et al. [11] applied soft set theory to classify musical instruments sounds and their results revealed that soft set theory can be successfully used for the classification of musical instruments.

Measuring similarity between two entities is a key step for several data mining tasks, such as classification and clustering. Similarity measures quantify the extent to which different patterns, signals, images or sets are alike. The studies on measuring the similarity between soft set have been carried out. They extended their research to measure the similarity of fuzzy soft set and describe how it can be applied to medical diagnosis to detect whether a person is suffering from a certain disease [8].

This paper investigates two existing classification approaches based on fuzzy soft set theory, one is fuzzy soft set based on decision making problems (comprises of comparison table [13]) and other is based on similarity measurement between two fuzzy soft set [20].

76.2.2 Preliminaries

In this subsection, the basic definition and results of soft set theory and fuzzy soft set which would be useful for subsequent discussion. Most of the definitions and results presented in this section may be found in [8, 21].

Definition 1 Let U be an initial universe set and E be a set of parameters. Let $\tilde{P}(U)$ denote the power set of U and $A \subset E$

A pair (\tilde{F}, E) a fuzzy soft set over U where F is a mapping given by $\tilde{F} \rightarrow \tilde{P}(U)$.

In the above definition, fuzzy subsets in the universe U are used as substitutes for the crisp subsets of U . Therefore, it is well known that notion of fuzzy sets provides convenient tool for representing vague concepts by allowing partial memberships. Every fuzzy soft set can be viewed as fuzzy soft set information system and be represented by a data table with entries belongs to the interval $[0, 1]$.

Table 76.1 represents both approaches based on fuzzy soft set theory. The comparison table classifier works by calculating the average value of each parameter from all objects with same class/label. Then construct a comparison table in the manner as the preparation of comparison table in the case of decision making problem where optimal decision is taken from maximum score computed from the comparison table. The next step is to calculate score to determine class label for the test data. Whereas, the approach based on similarity measure have the same learning phase with comparison table only the classification method is different.

76.3 Modeling Process

In this section, the modeling process of this study is presented. The modeling process consists of three phases which are data collection, data partitioning and validation measure. Each of the steps is discussed in detail in the following subsections.

76.3.1 Data Collection

Dataset is one of the crucial elements for designing and developing successful classification algorithms. Data collection has been done using University of California at Irvine (UCI) machine learning repository. Since UCI is a public repository that makes easier for data collection. Dataset includes breast-cancer-wisconsin (wdbc & wpc), breast tissue, heart, dermatology, liver disorder, hepatitis, pima Indians diabetes and indian liver dataset.

Table 76.1 Fuzzy soft set based classification

FSSCT: Fuzzy soft set based on decision making problems [13]	FSSSM: Fuzzy soft set based on similarity measure [20]
Training phase	Training phase
1. Input fuzzy soft sets $(\tilde{F}, A), (\tilde{G}, B)$ and (\tilde{H}, C)	1. Given N samples obtained from the data class w
2. Input parameters set P as observed by the observer	2. Calculate the cluster center vector $E_{wi}, i = 1, 2, \dots, N. E_w = \frac{1}{N} \sum_{i=1}^N E_{wi}$ Obtain Fuzzy soft set model (F, E) , which is $W \times D$ table of cluster centers in which an element of the table is g_{wd} , where $w = 1, 2, \dots, W$ and $d = 1, 2, \dots, D$ and a row g_{wd} is a cluster centre vector for every class w having D features
3. Compute corresponding resultant fuzzy soft set (\tilde{S}, P) from fuzzy soft sets $(\tilde{F}, A), (\tilde{G}, B), (\tilde{H}, C)$ and place it tabular form	3. Repeat the process for all W classes
Classification phase	Classification phase
1. Obtain the unknown class data	1. Obtain the unknown class data
2. Compute comparison table of fuzzy soft (\tilde{S}, P) and compute r_i using equation $r_i = \sum_{i=1}^n c_{ij}$ and t_j using equation $t_j = \sum_{j=1}^n c_{ij}$	2. Obtain a fuzzy soft sets model for unknown class data (\tilde{G}, E) and compute similarity between (\tilde{G}, E) and (\tilde{F}_w, E) for each w using equation $S(F_\rho, G_\delta) = M_i(\tilde{F}, \tilde{G}) = 1 - \frac{\sum_{j=1}^n \tilde{F}_j - \tilde{G}_j }{\sum_{j=1}^n (\tilde{F}_j + \tilde{G}_j)}$
3. Compute the score vector S using equation $s_i = r_i - t_i$ If k has more than one value then any one of o_k may be chosen	3. Assign the unknown data to class w if similarity is maximum $w = \arg[\max_{w=1}^W S(\tilde{G}, \tilde{F}_w)]$

Datasets undergo for pre-processing treatments. Data pre-processing phase involve fuzzification technique to make sure that data range lies between $[0, 1]$.

To classify numerical data with these algorithms, the second step is replaced with fuzzification process (refer to [10]) having similar algorithm with [13], which is like counting the normalization so that all parameters have a value between $[0, 1]$. For example, if the classification algorithm is applied to Breast Tissue dataset. Fuzzification [20] can be done by dividing each attributes value with the largest value at each attributes, $e_{fi} = \frac{e_i}{\max(e_i)}$ where $e_i, i = 1, 2, \dots, n$ is the old attribute and e_{fi} is attribute with new value between $[0, 1]$.

Table 76.2 provides description of all dataset. Most of the datasets having real numerical features and some contains multiclass labels.

Table 76.2 Dataset description

No.	Dataset	Description
1.	Breast tissue	i:106, f:10, c:6
2.	Statlog (Heart)	i:270, f:13, c:2
3.	Dermatology	i:366, f:33, c:6
4.	Liver disorder	i:345, f:7, c:2
5.	Hepatitis	i:155, f:19, c:2
6.	Pima Indians diabetes	i:768, f:8, c:2
7.	Lung cancer	i:32, f:56, c:3
8.	Lymphography	i:148, f:18, c:4
9.	Indian liver patient dataset (ILPD)	i:583, f:10, c:2

i instance, *f* features, *c* class

76.3.2 Data Partitioning

For data partition, a general course in data mining is to split into training and testing sets. Each dataset divided into two parts: 70 % for training and 30 % for testing and data were selected randomly for every experiment.

76.3.3 Validation Measure

For classification problems, the major source of performance measurement is coincidence matrix. However, when the classification problem is not binary, performance evolution becomes limited to overall classifier accuracy. Therefore, in order to quantify the performance of classification method, the performance metrics: overall classifier accuracy (OCA) and computational time have been used to access the performance of both classifiers:

76.4 Results and Discussion

Table 76.3 provides the results obtained from different datasets. In general both classifiers can do numerical classification however, highest achievement occurs in lung cancer dataset where FSSSM accuracy 97.05 and FSSCT accuracy 93.67. The reason that FSSSM gives better results because it does not need to build comparison table therefore it can work faster.

Table 76.3 Performance analysis of classification approaches for medical data

Classification algorithms based on fuzzy soft set				
Accuracy measures				
Classification algorithms	FSSCT		FSSSM	
	Accuracy (%)	Cpu time (s)	Accuracy (%)	Cpu time (s)
UCI datasets				
Breast tissue	56.13	0.0542	63.87	0.0033
Statlog (Heart)	82.72	0.0050	77.04	0.0049
Dermatology	82.97	0.0114	97.03	0.0098
Liver disorder	51.17	0.0065	53.01	0.0057
Hepatitis	81.91	0.0036	83.62	0.0034
Pima Indians diabetes	70.35	0.0154	70.22	0.0098
Indian liver patient dataset (ILPD)	64.19	0.0033	78.52	0.0050
Lung cancer	93.67	0.0096	97.05	0.0102
Lymphography	58.06	0.0031	78.89	0.0050

76.5 Conclusion

Current research in medical data classification mainly focuses on the use of efficient data mining algorithms and visualization techniques. Meanwhile, the major objective of current studies strives towards improving the classification accuracy, precision and computational speeds of classification algorithms, as well as reducing the amount of manual interaction.

Therefore, in this paper, we investigated existing classification algorithms based on fuzzy soft set for medical data. Nine datasets from UCI were used to test the accuracy and computational time of both classifiers. In general, both can do the classification of numerical datasets. It is experimentally demonstrated that both classification algorithms yields better accuracy. From these evidences on medical data classification, it can be seen that there is still much room for further improvement over current medical data classification task. More research, however, is needed to identify and reduce uncertainties in medical data classification to improve classification accuracy. For future work we will design a new classification algorithm based on fuzzy soft set using medical images.

Acknowledgments The authors would like to thank office for Research, Innovation, Commercialization and Consultancy Management (ORICC) and Universiti Tun Hussein Onn Malaysia for supporting this research under vote no 1255.

References

1. Antonie, M.L., Zaiane, O.R., Coman, A.: Application of data mining techniques for medical image classification. In: MDM/KDD, pp. 94–101 (2001)
2. Fesharaki, N.J., Pourghassem, H.: Medical X-ray images classification based on shape features and Bayesian rule. In: 2012 Fourth International Conference on Computational Intelligence and Communication Networks (CICN), pp. 369–373. IEEE (2012)
3. Latifoglu, F., Polat, K., Kara, S., Gunes, S.: Medical diagnosis of atherosclerosis from carotid artery Doppler signals using principal component analysis (PCA), k-NN based weighting pre-processing and Artificial Immune Recognition System (AIRS). *J. Biomed. Inform.* **41**, 15–23 (2008)
4. Suguna, N., Thanushkodi, K.: An improved k-nearest neighbour classification using genetic algorithm. *Int. J. Comput. Sci. Issues (IJCSI)* **7**(4, 2), 18 (2010)
5. Rajendran, P., Madheswaran, M., Naganandhini, K.: An improved pre-processing technique with image mining approach for the medical image classification. In: 2010 International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2010)
6. Hadidi, M.R.A.A., Gawagzeh, M.Y., Alsaaidah, B.: Solving mammography problems of breast cancer detection using artificial neural networks and image processing techniques. *Indian J. Sci. Technol.* **5**(4), 2520–2528 (2012)
7. Ali, S., Smith, K.A.: On learning algorithms selection for classification. *Appl. Soft Comput.* **6**, 119–138 (2006)
8. Majumdar, P., Samantra, S.K.: Generalized fuzzy soft set. *J. Comput. Appl. Math.* **58**, 1279–1286 (2010)
9. Ali, M.I.: A note on soft sets, rough sets and fuzzy soft sets. *Appl. Soft Comput.* **11**, 3329–3332 (2011)
10. Mushrif, M.M., Sengupta, S., Ray, A.K.: Texture classification using a novel soft set theory based classification algorithm. In: LNCS, vol. 3851, pp.246–254. Springer, Heidelberg (2006)
11. Lashari, S.A., Ibrahim, R., Senan, N.: Soft set theory for automatic classification of traditional Pakistani musical instruments sounds. In: 2012 International Conference on Computer & Information Science (ICCIS), vol. 1, pp. 94–99. IEEE (2012)
12. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. *J. Fuzzy Math.* **9**(3), 589–602 (2001)
13. Roy, A.R., Maji, P.K.: A fuzzy soft set theoretic approach to decision making problems. *J. Comput. Appl. Math.* **203**(2), 412–418 (2007)
14. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. *Informatica* **31**, 249–268 (2007)
15. Tu, M.C., Shin, D., Shin, D.: A comparative study of medical data classification methods based on decision tree and bagging algorithms. In: Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, DASC'09, pp. 183–187. IEEE (2009)
16. Xing, Y., Wang, J., Zhao, Z., Gao, Y.: Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In: International Conference on Convergence Information Technology, 2007, pp. 868–872. IEEE (2007)
17. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**(3), 338–353 (1965)
18. Zimmermann, H.J.: Fuzzy set theory-and its applications. Springer, New York (2001)
19. Maji, P.K., Roy, A.R., Biswas, R.: An application of soft sets in decision making problem. *Comput. Math. Appl.* **44**, 1077–1083 (2002)
20. Handaga, B., Deris, M.M.: Similarity approach on fuzzy soft set based numerical data classification. *Commun. Comput. Inf. Sci.* **180**(6), 575–589 (2011)
21. Molodtsov, D.: Soft set theory—first results. *Comput. Math. Appl.* **37**(4–5), 19–31 (1999)

Chapter 77

A Hybrid Selection Method Based on HCELFS and SVM for the Diagnosis of Oral Cancer Staging

Fatihah Mohd, Zainab Abu Bakar, Noor Maizura Mohamad Noor, Zainul Ahmad Rajion and Norkhafizah Saddki

Abstract A diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method is developed to diagnose the stage of oral cancer in patients. The hybrid feature selection method, named Hybrid Correlation Evaluator and Linear Forward Selection (HCELFS), combines the advantages of filters and wrappers to select the optimal feature subset from the original feature set. In HCELFS, Correlation Attribute Evaluator acts as filters to remove redundant features and Linear Forward Selection with SVM acts as the wrappers to select the ideal feature subset from the remaining features. This study conducted experiments in WEKA with ten fold cross validation. The experimental results with oral cancer data sets demonstrate that our proposed model has a better performance than well-known feature selection algorithms.

F. Mohd (✉) · N.M.M. Noor
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
21030 K. Terengganu, Terengganu, Malaysia
e-mail: mpfatihah@yahoo.com

N.M.M. Noor
e-mail: maizura@umt.edu.my

Z.A. Bakar
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM),
Selangor 40450 Shah Alam, Malaysia
e-mail: zainab@tmsk.uitm.edu.my

Z.A. Rajion · N. Saddki
School of Dental Sciences, Universiti Sains Malaysia (USM), Kubang Kerian 16150,
Kelantan, Malaysia
e-mail: zainul@kck.edu.my

N. Saddki
e-mail: fizah@kb.usm.my

77.1 Introduction

Feature selection (FS) as preprocessing steps to machine learning in real world data, is very useful in reducing dimensionality, removing irrelevant data and noise to improve result. It could directly reduce and remove irrelevant number of the original features by selecting a subset that contributes to the optimum information for classification. The FS algorithms are divided into two categories: the filter methods and wrappers methods [1]. Both methods have their own abilities and advantages. The filter method contributes high computational efficiency compared to the wrapper method. The wrapper could achieve better results than the filter approach. In this study, we combined both methods to propose a hybrid algorithm to gain the optimum selected features.

Head and neck (HNC) cancer is one of the major cancers worldwide. One part of HNC is oral cancer with the incidence rising in every country. Early clinical cancer diagnosis is seen as an important element in reducing the mortality rate of this deadly disease. The process of clinical diagnosis begins with information gathering or eliciting data from a patient's history. It includes data collection from patient's primary report of symptoms, past medical history, family history, and social history. In this process, sometimes decision making can be done, where the clinician can start the procedure of formulating a list of possible diagnoses [2]. Then, by doing a physical examination, the physician detects abnormalities by looking at, feeling, and listening to all parts of body. However, the patient's record is a collection of features and data that leads to problems for the diagnosis process. The challenge of applying computational solution to the data collected is in the conversion of data into an appropriate form, suitable for the diagnosis process [3]. Because of this, FS method is applied to reduce the irrelevant data and finally select the optimum features to diagnose the stage of oral cancer. This paper explains the development of a diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method in diagnosing the stage of oral cancer.

The remaining of the paper is organized as follows: related work is given in Sect. 77.2, while Sect. 77.3 gives a brief description about the FSA algorithms—Correlation Attribute Evaluator and LFS, SVM algorithm. Section 77.4 discusses the diagnostic model for oral cancer and Sect. 77.5 reports the results and discussion. The concluding remarks are given in Sect. 77.6 to address further research issues.

77.2 Related Work

Support Vector Machine (SVM) is an affective algorithm used in medical diagnosis for pattern recognition, machine learning and data mining. In the literature, there are some works related to medical diagnosis. Aruna et al. compared the

performance criterion of supervised learning classifiers such as Naïve Bayes, SVM RBF kernel, RBF neural networks, Decision trees J48 and Simple CART. The experiments conducted were found that SVM RBF Kernel produced highest result than other classifiers with respect to accuracy, sensitivity, specificity and precision [4]. Jaganathan et al. have proposed a feature selection method with improved F-score and SVM for breast cancer diagnosis and produced a classification accuracy of 95.565 %. This result is better than RBF Network (95.278 %) [5]. In other field, SVM is also applied in cyber-security. Maldonado and L'Huillier proposed an embedded approach for feature selection using SVM in phishing and spam classification. It outperforms other techniques in terms of classification accuracy by removing the features that affect on the generalization of the classifier by optimizing the Kernel function [6].

This related work is also focusing on the diagnosis of head and neck cancer using machine learning and data mining algorithm. For instance, Kawazu et al. [7] used neural network (NN) to predict lymph node metastasis of patients with oral cancer. They utilized histopathological data set of lymph nodes which saw an accuracy of 93.6 % in diagnosing patients. Boronti et al. produced four different results with three different methods such as SVM, Decision Trees (DTs), XCS and NN with accuracies of 75.5, 76.5, 79.2 and 71.3 % respectively [8]. In another study, they continued with other methods. They produced a classification result with DTs (70 %), XCS (79 %) and NN (78 %) [9]. Besides this, Exarchos et al. [10] employed a feature selection algorithm, Correlation-based Feature Subset selection (CFS) and the wrapper algorithm in order to omit redundant or possible irrelevant features and maintain the most informative and discriminatory ones. With the applications of Bayesian Networks, Artificial Neural Networks, SVM, DTs and Random Forests, the study produced an accuracy of (69.6 %), (66.1 %), (69.6 %), (66.1 %) and (58.9 %) respectively. However, with a hybrid model of Relief F-GA-ANFIS, Chang et al. produced a better classification accuracy with 93.81 % [11].

77.3 Materials and Methods

77.3.1 Features Selection Algorithm

In this study, feature selection for high-dimensional data are conducted in WEKA with tenfold cross validation. The main idea of feature selection functions are used to find the most significant attributes by removing features with little or no predictive information. The functions used for attribute evaluation (feature selection) within this study are as follows:

Correlation Attribute Evaluator. This algorithm evaluates the worth of an attribute by measuring the correlation between it and the class. Nominal attributes are considered on a value by value basis by treating each value as an indicator. An overall correlation for a nominal attribute is arrived at via a weighted average.

CFS Subset Evaluator. This algorithm evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

The Correlation Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. It measures subsets of features on the basis of the hypothesis, “A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other”. The following equation gives the merit of a feature subset S consisting of k features:

$$Merit_{sk} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

where $Merit_{sk}$ is the heuristic “merit” of a feature subset S containing k features, r_{cf} is the average value of all feature-classification correlations ($f \in S$), and r_{ff} is the average value of all feature-feature correlations. The numerator of (1) can be thought of as providing an indication of how predictive of the class a set of features are; the denominator of how much redundancy there is among the features [12].

All the attributes were searched using these algorithms:

Ranker. Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc.).

Linear Forward Selection (forward). Linear Forward Selection, a technique to reduce the number of attributes expansions in each forward selection step. This function is extension of Best First. It takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected. The search uses either the initial ordering to select the top k attributes, or performs a ranking (with the same evaluator the search uses later on). This algorithm starting from the empty set, sequentially add the feature x^+ that results in the highest objective function $J(Y_k + x^+)$ when combined with the features Y_k that have already been selected.

1. Start with the empty set $Y_0 = (\emptyset)$
2. Select the next best feature $X^+ = \arg \max [J(Y_k + x)]_{x \notin Y_k}$
3. Update $Y_{k+1} = Y_k + x^+$; $k = k + 1$
4. Go to 2

77.3.2 Oral Cancer Dataset

The study obtained a record review of oral cancer patients from the Otorhinolaryngology Clinic at Hospital Universiti Sains Malaysia (HUSM) in Kelantan respectively. The dataset is made up of 27 parameters and a primary tumor stage as attributes for the diagnosis of the patients. The study was conducted after the

Table 77.1 Description of the datasets

Attributes no.	Attributes name
1.	Age
2.	Gender
3.	Ethnicity
4.	Smoking
5.	Chewing betel quid
6.	Alcohol
7.	S1
8.	S2
9.	S3
10.	S4
11.	S5
12.	S6
13.	S7
14.	S8
15.	S9
16.	S10
17.	S11
18.	Site
19.	Size
20.	Lymph node
21.	Histological
22.	SCC
23.	T
24.	N
25.	M
26.	Stage (class label): Stage I, Stage II, Stage III, Stage IV

obtainment of the required approvals from the Research and Ethics Committee (Human), Universiti Sains Malaysia, No.236.4.(4.4) [13]. Number of instances was 210, and 27 features with patient_id was named as label and stage was named as class label. The numerical variables were analysed through the corresponding ranges of their values. Age was divided into five groups (group 1: below 30 years old; group 2: 30–39 years old; group 3: 40–49 years old; group 4: 50–59 years old and group 5: 60 years old and above) [14, 15]. The oral cancer regions included in this study were the tongue, buccal mucosa, palate, floor of mouth, maxilla, lip, cheek, mandible, tonsil, parotid gland, oropharynx and other unspecified parts. The details of the attributes found in this dataset for features selection listed in Table 77.1.

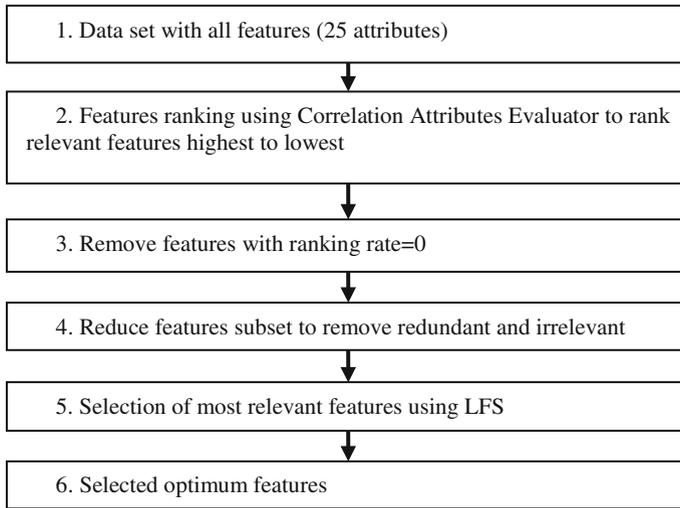


Fig. 77.1 Stage in hybrid correlation evaluator and linear forward selection (HCELFS) algorithm

77.4 Diagnostic Model: Hybrid Correlation Evaluator and Linear Forward Selection

A hybrid feature selection method, named HCELFS is proposed in this study. This FS method, hybrid Correlation Attribute Evaluator with ranker and CFS Subset Evaluator with Linear Forward Selection was applied for oral cancer diagnosis. It combines the advantages of both methods to select the optimal features subset from the original feature set. In the diagnostic model, the first step included the Correlation Attribute Evaluator method filtering the relevant features which resulted in reduced features subset. From this subset CFS Subset Evaluator with Linear Forward Selection (LFS) searched for the most relevant features resulting in optimum feature set used for diagnosing the cancer stage. Figure 77.1 shows the stage in the algorithm.

77.5 Results and Discussion

The experiments of features selection against oral cancer data set are conducted in WEKA with tenfold cross validation. Algorithm started with 25 features and 210 instances. With Correlation Ranking Filter, the algorithm ranked 25 features namely 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 7, 17, 3, 18, 5, 1, 13, 9, 11, 6, 25, 10, 14, 4 and 12 (see Table 77.2). We removed 1 feature namely 12 with ranking rate 0. With the resultant 24 features, the subset method then remove the redundant and

Table 77.2 Correlation ranking filter for oral cancer data set with 25 attributes

Correlation ranking filter	Ranked attributes (%)
20: Lymph node	0.4602
23: T	0.4319
21: Histological	0.3715
22: SCC	0.3611
16: S10	0.3561
19: Size	0.3558
24: N	0.3349
8: S2	0.3203
2: Gender	0.2751
15: S9	0.2660
7: S1	0.2644
17: S11	0.2345
3: Ethnicity	0.2323
18: Site	0.2297
5: Betel quid	0.1966
1: Age	0.1884
13: S7	0.1810
9: S3	0.1496
11: S5	0.1465
6: Alcohol	0.1042
25: M	0.1042
10: S4	0.0905
14: S8	0.0599
4: Smoking	0.0455
12: S6	0
Selected 25 attributes:	
20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 7, 17, 3, 18, 5, 1, 13, 9, 11, 6, 25, 10, 14, 4 and 12	

irrelevant 10 features namely, 7, 5, 1, 13, 11, 6, 25, 10, 14 and 4. This algorithm ended with 14 features namely 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 17, 3, 18 and 9 as optimum features set.

Table 77.3 summarize hybrid the features selection methods experimented in this study. It started with no features selection (FS0), Correlation Attribute Evaluator with Ranker (FS1), and then combined FS1 with CfsSubset Evaluator with Linear Forward Selection (FS2).

Table 77.3 Selected attributes with hybrid feature selection methods

FS	Method	Selected attributes
FS0	No features selection	All attributes
FS1	CorrelationAttributeEval	Ranked attributes: 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 7, 17, 3, 18, 5, 1, 13, 9, 11, 6, 25, 10, 14, 4, 12 (25 attributes)
	Ranker	Remove ranting value = 0 (attribute 12)
FS2	CfsSubsetEval	Remove irrelevant attributes = 10 attributes 7, 5, 1, 13, 11, 6, 25, 10, 14, 4
	LinearForwardSelection (forward)	Optimum features: 14 attributes 20, 23, 21, 22, 16, 19, 24, 8, 2, 15, 17, 3, 18, 9

Table 77.4 Accuracy performance for classification algorithms on oral cancer data set

Algorithm	FS0	FS1	FS2
Updateable Naïve Bayes	91.9048	91.9048	94.7619
	8.0952	8.0952	5.2381
MLP	94.2857	93.8095	95.2381
	5.7143	6.1905	4.7619
Lazy-IBK	86.1905	86.1905	91.4286
	13.8095	13.8095	8.5714
SMO- poly kernel (E-1.0)	93.3333	93.3333	96.1905
	6.6667	6.6667	3.8095

In order to evaluate the efficiency of the FS methods, performance measure of accuracy were considered. The measures are compiled by the Classification Accuracy (%) = $(TP + TN)/(TP + FP + FN + TN)$. In this study, four different machine learning algorithms were used to classify the oral cancer data set with three features selection methods and optimum features selected by the proposed hybrid algorithm, Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest neighbors (KNN), and SVM. NB classifier using estimator classes. MLP classifier uses backpropagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. K-Nearest neighbor classifier (lazy.IBk) can select appropriate value of K based on cross-validation. It can also do distance weighting. SVM or SMO-Poly Kernel (E-1.0) implemented globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Table 77.4 shows the results for the classifier. The empirical comparison shows that the features selected by the hybrid algorithm also improved the accuracy of the entire classifier algorithm used for the oral cancer data set. Table 77.5 shows the classification accuracies of our method and other classifiers from literature for the head and neck data set.

Table 77.5 Classification accuracies of this study method and other classifiers from literature

Author, year	Method	Accuracy (%)
Kawazu et al. 2003 [7]	Neural networks (NN)	93.6
Baronti et al. 2005 [8]	Support vector machines (SVM)	75.5
	Decision trees (C4.5)	76.5
	XCS	79.2
	NN	71.3
Tung and Quek 2005 [16]	FS based on wrapper	Above 90 %
	Monte Carlo evaluative selection (MCES)	
	Classification using	
	SVM with polynomial kernel, K-Nearest Neighbor (K-NN) classifier, artificial neural network (ANN) and the GenSoFNN-TVR(S) network	
Baronti and Starita, 2007 [9]	Naive Bayes (NB)	69.4
	C4.5	70
	NN	78
	XCS	79
	Hypothesis classifier systems (HCS)	83.8
Xie et al. 2010 [17]	Improved F-score and sequential forward floating search (IFSFFS)	100 (best)
	SVM	97.58 (avg)
Exarchos et al. 2011 [10]	Bayesian networks (BNs)	69.6
	ANN	66.1
	SVM	69.6
	Decision trees (DTs)	66.1
	Random forests (RFs)	58.9
Chang et al. 2013 [11]	1. Pearson’s correlation coefficient (CC) and Relief-F as the filter approach	93.81 %
	2. Genetic algorithm (GA) as the wrapper approach	
	3. CC-GA and ReliefF-GA as the hybrid approach.	
	Hybrid model of ReliefF-GA-ANFIS	
Calle-Alonso et al. 2013 [18]	Combines pairwise comparison, Bayesian regression and K-NN	97.74
This study	FS on erythemato squamous disease data set, combine	98.64
	1. CorrelationAttributeEval and Ranker	
	2. CfsSubsetEval and LinearForwardSelection (forward)	
This study	3. SMO- Poly Kernel (E-1.0)	96.19
	FS on oral cancer data set, combine	
	1. CorrelationAttributeEval and Ranker	
	2. CfsSubsetEval and LinearForwardSelection (forward)	
	3. SMO- Poly Kernel (E-1.0)	

77.6 Conclusion

In this study, it is noted that a diagnostic model based on Support Vector Machines (SVM) with a proposed hybrid feature selection method to diagnose the stage of oral cancer showed an increased of classification accuracy. The hybrid feature selection method, named HCELFS, combines Correlation Attribute Evaluator which acts as a filter and SBFS which acts as the wrapper to select the ideal feature subset from the remaining features.

The experimental results with oral cancer data sets demonstrate that the new hybrid feature selection method has a better performance than well-known feature selection algorithms. It obtained optimal classification accuracy with 14 features from a set of 25 features. The optimal feature subset obtained were then trained with various data mining algorithms such as Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest neighbors (KNN), and SVM to diagnose the stage of oral cancer. One direction for future studies is to consider proposing a hybrid algorithm with various dataset and other data mining classifier.

Acknowledgments This study has been supported in part the Exploratory Research Grant Scheme (ERGS) 600_RMI/ERGS 5/3 (3/2011) under the Malaysia Ministry of Higher Education (MOHE) and Universiti Teknologi MARA (UiTM) Malaysia. The authors would like to acknowledge all contributors, technical members at Hospital Universiti Sains Malaysia (HUSM) who have helped and greatly assisted in the completion of the study. Dr. Zailani Abdullah, data mining expertise from SIAM, UMT and anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

References

1. Alpaydin, E.: Introduction to machine learning. MIT Press, Cambridge (2004)
2. Neville, B., Damm, D., Allen, C., Bouguet, J.: Oral and maxillofacial pathology. Saunders Elsevier, St. Louis (2009)
3. Poolsawad, N., Kambhampati, C., Cleland, J.G.F.: Feature selection approaches with missing values handling for data mining—a case study of heart failure dataset. *World Acad. Sci. Eng. Technol.* **60**, 828–837 (2011)
4. Aruna, S., Rajagopalan, S.P., Nandakishore, L.V.: Knowledge based analysis of various statistical tools in detecting breast cancer. In: First International Conference on Computer Science, Engineering and Applications (CCSEA), pp. 37–45. Chennai, India (2011)
5. Jaganathan, P., Rajkumar, N., Kuppuchamy, R.: A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification. *Int. J. Mach. Learn. Comput.* **2**, 741–745 (2012)
6. Maldonado, S., L’Huillier, G.: SVM-based feature selection and classification for email filtering. In: Latorre Carmona, P., Sánchez, J.S., Fred A.L.N. (eds.) *Pattern Recognition—Applications and Methods*, pp. 135–148. Springer, Heidelberg (2013)
7. Kawazu, T., Araki, K., Yoshiura, K., Nakayama, E., Kanda, S.: Application of neural networks to the prediction of lymph node metastasis in oral cancer. *Oral Radiol.* **19**, 35–40 (2003)
8. Baronti, F., Colla, F., Maggini, V., Micheli, A., Passaro, A., Rossi, A.M.: Experimental comparison of machine learning approaches to medical domains: a case study of genotype

- influence on oral cancer development. In: European Conference on Emergent Aspects in Clinical Data Analysis (EACDA), pp. 81–86. Pisa, Italy (2005)
9. Baronti, F., Starita, A.: Hypothesis testing with classifier systems for rule-based risk prediction evolutionary computation. *Mach. Learn. Data Min. Bioinf.* **4447**, 24–34 (2007)
 10. Exarchos, K., Goletsis, Y., Fotiadis, D.: Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Trans. Inf. Technol. Biomed.* **16**(6), 1127–1134 (2011)
 11. Chang, S.W., Abdul Kareem, S., Merican, A., Zain, R.: Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinf.* **14**, 170 (2013)
 12. Hall, M.A.: Correlation-Based Feature Selection for Machine Learning. The University of Waikato, Hamilton, New Zealand (1999)
 13. Bakar, Z.A., Mohd, F., Noor, N.M.M., Rajion, Z.A.: Demographic profile of oral cancer patients in east coast of peninsular Malaysia. *Int. Med. J.* **20**, 362–364 (2013)
 14. Mohd, F., Bakar, Z.A., Noor, N.M.M., Rajion, Z.A.: Data preparation for pre-processing on oral cancer dataset. In: 13th International Conference on Control, Automation and Systems (ICCAS), p. 324. Gwangju, Korea (2013)
 15. Razak, A.A., Saddki, N., Naing, N.N., Abdullah, N.: Oral cancer presentation among Malay patients in hospital Universiti Sains Malaysia, Kelantan. *Asian Pac. J. Cancer Prev* **10**, 1131–1136 (2009)
 16. Tung, W.L., Quek, C.: GenSo-FDSS: a neural-fuzzy decision support system for pediatric all cancer subtype identification using gene expression data. *Artif. Intell. Med.* **33**, 61–88 (2005)
 17. Xie, J., Xie, W., Wang, C., Gao, X.: A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythematous-squamous diseases. *J. Mach. Learn. Res. Workshop Conf. Proc.* **11**, 142–151 (2010)
 18. Calle-Alonso, F., Pérez, C.J., Arias-Nicolás, J.P., Martín, J.: Computer-aided diagnosis system: a Bayesian hybrid classification method. *Comput. Methods Programs Biomed.* **112**, 104–113 (2013)

Chapter 78

A Linear Assignment Method of Simple Additive Weighting System in Linear Programming Approach Under Interval Type-2 Fuzzy Set Concepts for MCDM Problem

Nurnadiah Zamri and Lazim Abdullah

Abstract The ranking phase is valuable to examines the final alternative rankings of decision making problems. Based on simple additive weighting (SAW) and linear programming (LP) within the context of interval type-2 fuzzy sets (IT2 FSs), we develop a linear assignment method to produce the final ranking order of all alternatives for interval type-2 fuzzy TOPSIS (IT2 FTOPSIS) method. A numerical example is used to check the efficiency and applicability of the proposed method. The results shows consistent outcomes of the decision making process. Thus, the proposed method offers an alternative, user-friendly method that is robust in the decision making framework.

78.1 Introduction

Ranking phase is the step to examines the results of decision making problems. The interpretation of multiple attribute decision making (MADM) results can show the differences in the rankings of the alternatives. It was extensively applied and strengthened the theoretical part of aggregating phase by many authors. A few of them were; Gao et al. [1] developed a fuzzy approach based on the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), where in the ranking phase, the distances of each alternative from the fuzzy positive ideal solutions (PIS) and the fuzzy negative ideal solutions (NIS) are computed respectively with

N. Zamri (✉) · L. Abdullah

School of Informatics and Applied Mathematics, University Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia
e-mail: nadzlina@yahoo.co.uk

L. Abdullah

e-mail: lazim_m@umt.edu.my

a vertex method. Then, a closeness coefficient is obtained to rank order of all alternatives. Li [2] constructed nonlinear-programming models on the basis of the concepts of the relative-closeness coefficient and the weighted-Euclidean distance. Simpler auxiliary nonlinear-programming models were further deduced to calculate relative-closeness of intuitionistic fuzzy (IF) sets of alternatives to the interval-valued intuitionistic fuzzy-positive ideal solutions (IVIF-PIS), which can be used to generate the ranking order of alternatives. Jolai et al. [3], proposed the goal programming (GP) technique, and constructed a multi-objective mixed integer linear programming (MOMILP) model to determine the order quantities of each selected supplier for each product in each period.

Furthermore, for interval type-2 (IT2) fuzzy approach based on the TOPSIS, Chen and Lee [4] proposed a ranking value method to cumulative all the collective decisions and obtained the relative closeness through the traditional TOPSIS method computing process. However, the used of standard deviation in the ranking value method is believed influenced by extreme scores and the method is depended only on the dispersion's data. It is proved by Chen et al. [5], that the ranking value proposed by Chen and Lee [4] was difficult and higher in computational volume. Since that, various authors discussed on the ranking phase of IT2 FTOPSIS method. For example, Chen et al. [5] proposed a new method MADM based on the proposed ranking method of IT2FSs. Wang et al. [6] developed IT2 fuzzy weighted arithmetic averaging operator to aggregate all individual IT2 fuzzy decision matrices provided by the decision-makers (DMs) into the collective IT2 fuzzy decision matrix, then utilized the ranking-value measure to calculate the ranking value of each attribute value and constructed the ranking-value matrix of the collective IT2 fuzzy decision matrix. Chen and Wang [7] presented a new fuzzy ranking method based on the α -cuts of interval type-2 fuzzy sets (IT2 FSs). Chen [8] developed a new linear assignment method to produce an optimal preference ranking of the alternatives in accordance with a set of criterion-wise rankings and a set of criterion importance within the context of interval type-2 trapezoidal fuzzy numbers (IT2TrFNs) for MADM problems. However, little research has been conducted on the simple additive weighting (SAW) and linear programming (LP) for coping with IT2FSs. This linear assignment method (SAW and LP with IT2FSs method) is developed to handle the ranking phase of IT2 FTOPSIS method.

Thus, the purpose of this paper is to extend the SAW and LP methods in IT2FSs approach for ranking phase of IT2FTOPSIS. This paper proposes the linear assignment method with the identification of the SAW and LP methods to determine the final ranking orders respectively, for each pair of alternatives. The feasibility and the applicability of the proposed methods are illustrated using the MADM examples of Chen [9].

This paper is illustrated as follows. Section 78.2 discusses the concept of weighted average with linear programming. Section 78.3 proposes a linear assignment method based on SAW and LP methods in IT2 FSs concept. Section 78.4 illustrates a numerical example in order the check the efficiency of the proposed method. Finally, Sect. 78.5 presents the conclusions.

78.2 Weighted Average With Linear Programming

In the following, we recall basic notations and definitions of weighted average with linear programming.

Definition 78.1 [10, 11] The minimum and maximum for the fuzzy weighted average for each given α_j can be obtained by solving the following two fractional programming problems:

$$\begin{aligned} \min f_L &= \frac{w_1 a_1 + w_2 a_2 + \cdots + w_n a_n}{w_1 + w_2 + \cdots + w_n} \\ \text{s.t. } c_i &\leq w_i \leq d_i, \quad i = 1, 2, \dots, n, \end{aligned} \quad (78.1)$$

$$\begin{aligned} \max f_U &= \frac{w_1 b_1 + w_2 b_2 + \cdots + w_n b_n}{w_1 + w_2 + \cdots + w_n} \\ \text{s.t. } c_i &\leq w_i \leq d_i, \quad i = 1, 2, \dots, n, \end{aligned} \quad (78.2)$$

where c_i and d_i are the two end points of the interval w_i for a given α_j level cut.

The Charnes and Cooper's linear transformation is summarized in the following. Consider the following simple fractional programming problem:

$$\begin{aligned} \min \frac{px}{qx} \\ \text{s.t. } Ax \leq b, \quad x \geq 0, \end{aligned} \quad (78.3)$$

where p and q are two n -dimensional constant vectors, x is the n -dimensional variable, A is an $m \times n$ matrix, and b is an m -dimensional constant vector.

To transform the above fractional programming problem into a linear problem, let

$$z = \frac{1}{qx} \quad \text{and} \quad zx = y, \quad (78.4)$$

where we assume that $qx \neq 0$. Multiplying both the objective function and the constraints by z and using the definitions given in Eq. 78.5, we obtain

$$\begin{aligned} \min py \\ \text{s.t. } Ay \leq bz, \quad qy = 1, \quad y \geq 0, \quad z \geq 0, \end{aligned} \quad (78.5)$$

which is a linear programming problem.

This weighted average with linear programming is being used in defining a linear assignment method. Thus, the development of the proposed model is described in Sect. 78.3.

78.3 The Proposed Method

Original SAW method and LP are modified into an IT2 FS manner. Modifications are made to accommodate the objective of the research and also to simplify the computational procedure without losing the novelty of SAW and LP. The proposed method is then applied into a linear assignment method for IT2 FTOPSIS (MCDM method) to get the optimal preference ranking. This proposed method is believed to be more flexible rather than the existed model due to the fact that it is used the IT2 FS. On the other hand, this model is more suitable to represent uncertainties because it is involve end-users into the whole weighting process. Thus, suppose an IT2 FTOPSIS has n alternatives (A_1, \dots, A_n) and m decision criteria/attributes (C_1, \dots, C_m) . Each alternative is evaluated with respect to the m criteria/attributes. All the values/ratings assigned to the alternatives with respect to each criterion from a decision matrix, denoted by $S = (y_{ij})_{n \times m}$, and the relative weight vector about the criteria, denoted by $W = (w_1, \dots, w_m)$, that satisfying $\sum_{j=1}^m w_j = 1$. Therefore, the rest of the general process of this proposed method is listed as follows:

Rating State: *In this state, all the matrices are transformed into the IT2 FS concept.*

Step 1: Establish a decision matrix and weight matrix

Establish an IT2 decision matrix and IT2 fuzzy weight matrix.

Step 2: Comparable Scale

Construct a comparable scale for all elements in the decision matrix. The comparable scale is used to divide the outcome of a certain criterion by its maximum value, provided that the criteria are defined as benefit criteria. Therefore, the comparable scale is represented as follows:

$$\tilde{r}_{ij} = 1 / FOU(\tilde{r}_{ij}) = [\tilde{r}_{ij}, \tilde{r}_{ij}] \tag{78.6}$$

$$\text{For positive criteria } \tilde{r}_{ij}^* = \left(\left[\frac{\tilde{f}_{ij}}{\tilde{z}_{ij}^*} \right], \left[\frac{\tilde{f}_{ij}}{\tilde{z}_{ij}^*} \right] \right) \tag{78.7}$$

$$\text{For negative criteria } \tilde{r}_{ij} = \left(\left[\frac{\tilde{z}_{ij}^{\min}}{\tilde{f}_{ij}} \right], \left[\frac{\tilde{z}_{ij}^{\min}}{\tilde{f}_{ij}} \right] \right) \tag{78.8}$$

Then the decision matrix can be expressed as follows:

$$D = \begin{matrix} & C_1 & C_2 & \dots & C_n \\ x_1 & \begin{bmatrix} \tilde{r}_{11} \\ \tilde{r}_{21} \\ \vdots \\ \tilde{r}_{m1} \end{bmatrix} & \begin{bmatrix} \tilde{r}_{12} \\ \tilde{r}_{22} \\ \vdots \\ \tilde{r}_{m2} \end{bmatrix} & \begin{bmatrix} \dots \\ \dots \\ \dots \\ \dots \end{bmatrix} & \begin{bmatrix} \tilde{r}_{1n} \\ \tilde{r}_{2n} \\ \vdots \\ \tilde{r}_{mn} \end{bmatrix} \end{matrix} \tag{78.9}$$

where \tilde{r}_{mn} is the comparable scale value in the decision matrix.

Weighting State: Modification of the existed SAW method with “modified SAW in IT2 FS concepts”.

Step 3: Weight of attributes of SAW

Construct the weighting matrix W_p using the SAW formulae of the attributes of the decision-maker and construct the p th average weighting matrix \bar{W} .

Step 4: Weighted decision matrix

Construct the weighted decision matrix.

Aggregation State: Upgrading the calculation of separations of each alternative with linear programming concepts.

Step 5: Positive ideal solution and negative ideal solution

Determine the matrices that include positive and negative ideal solutions.

Step 6: Construct the separation of each alternative of SAW by linear programming approach

Calculate the separation of each alternative from the positive ideal solution I^* and negative ideal solution I^- using the formulae as follows:

$$C_i(b_{ij}, d_{ij}; s) = \frac{\sum_{j=1}^n [\tilde{w}_j \tilde{b}_{ij} + \tilde{\rho}_j (1 - \tilde{d}_{ij})]}{\sum_{j=1}^n [\tilde{w}_j \tilde{b}_{ij} + \tilde{\rho}_j (1 - \tilde{d}_{ij})] + \sum_{j=1}^n [\tilde{w}_j (1 - \tilde{b}_{ij}) + \tilde{\rho}_j (\tilde{d}_{ij})]} + s^l + s^u$$

$$= \frac{\sum_{j=1}^n [\tilde{w}_j \tilde{b}_{ij} + \tilde{\rho}_j (1 - \tilde{d}_{ij})]}{\sum_{j=1}^n [\tilde{w}_j + \tilde{\rho}_j]} + s^l + s^u \tag{78.10}$$

and let

$$z = \frac{1}{\sum_{j=1}^n [\tilde{w}_j + \tilde{\rho}_j]} \tag{78.11}$$

Assigned the value for

$$\tilde{t}_j = z\tilde{w}_j \tag{78.12}$$

and

$$\tilde{y}_j = z\tilde{\rho}_j \quad (j = 1, 2, \dots, n) \tag{78.13}$$

Since

$$z = \frac{1}{\sum_{j=1}^n [\tilde{w}_j + \tilde{\rho}_j]} \therefore \frac{1}{z} = \sum_{j=1}^n [\tilde{w}_j + \tilde{\rho}_j] \tag{78.14}$$

and

$$\tilde{t}_j = z\tilde{w}_j \quad \therefore \quad \tilde{w}_j = \frac{\tilde{t}_j}{z} \tag{78.15}$$

$$\tilde{y}_j = z\tilde{\rho}_j \quad \therefore \quad \tilde{\rho}_j = \frac{\tilde{y}_j}{z} \tag{78.16}$$

Thus, based on the above Charnes and Cooper’s transformations [12], Eq. 78.10 can be transformed into the equivalent linear programming models as follows:

$$C_i^u(\tilde{b}_{ij}, \tilde{d}_{ij}; s) = \max \left\{ \sum_{j=1}^n \tilde{t}_j b_{ij}^u + \tilde{y}_j (1 - d_{ij}^l) + s^l + s^u \right\}$$

$$\text{s.t.} \begin{cases} z\tilde{w}_j^l \leq \tilde{t}_j \leq z\tilde{w}_j^u & (j = 1, 2, \dots, n) \\ z\tilde{\rho}_j^l \leq \tilde{y}_j \leq z\tilde{\rho}_j^u & (j = 1, 2, \dots, n) \\ \sum_{j=1}^n (\tilde{t}_j + \tilde{y}_j) = 1 \\ z \geq 0 \\ s^l = n \quad (n = 0, \dots, 1) \\ s^u = n \quad (n = 0, \dots, 1) \end{cases} \tag{78.17}$$

and

$$\begin{aligned}
 C_i^l(\tilde{b}_{ij}, \tilde{d}_{ij}; s) &= \max \left\{ \sum_{j=1}^n \tilde{t}_i b_{ij}^l + \tilde{y}_j (1 - d_{ij}^u) + s^l + s^u \right\} \\
 \text{s.t. } \begin{cases} z\tilde{w}_j^l \leq \tilde{t}_j \leq z\tilde{w}_j^u & (j = 1, 2, \dots, n) \\ z\tilde{\rho}_j^l \leq \tilde{y}_j \leq z\tilde{\rho}_j^u & (j = 1, 2, \dots, n) \\ \sum_{j=1}^n (\tilde{t}_j + \tilde{y}_j) = 1 \\ z \geq 0 \\ s^l = n \quad (n = 0, \dots, 1) \\ s^u = n \quad (n = 0, \dots, 1) \end{cases} \tag{78.18}
 \end{aligned}$$

where $C_i(b_{ij}, d_{ij}; s)$ is an IT2 FS, denoted by $[C_i^l, C_i^u]$.

Step 7: Define the closeness coefficient

Calculate the relative degree of closeness to the ideal solution for each alternative.

Ranking State:

Step 8: Rank all alternatives

Sort the values of $(CC)_i$ in a descending sequence, where $1 \leq j \leq n$. The larger the value of $(CC)_i$, the higher the preference of the alternative for $(CC)_i$.

In this section, we have successfully introduced a new concept of linear assignment method. In order to check the efficiency of the proposed method, a numerical example is provided in Sect. 78.4 to illustrate the proposed method.

78.4 Illustrative Example

In this section, we used an example from Chen [9] to illustrate the proposed method. This numerical example is used to test the ability of the proposed method to handle the IT2 MCDM problems in many areas. All the relative importance weights in this numerical example are described using the linguistic variables which are defined in Table 78.1.

Table 78.1 Linguistic variables for the relative importance weights of criteria

Linguistic variable	Interval type-2 fuzzy number (IT2FN)
Very low (VL)	$((0,0.1;1), (0,0.5;1))$
Low (L)	$((0,0.3;1), (0.05,0.2;1))$
Medium low (ML)	$((0.1,0.5;1), (0.2,0.4;1))$
Medium (M)	$((0.3,0.7;1), (0.4,0.6;1))$
Medium high (MH)	$((0.5,0.9;1), (0.6,0.8;1))$
High (H)	$((0.7,1.0;1), (0.8, 0.95;1))$
Very high (VH)	$((0.9,1.0;1), (0.95,1.0,1))$

Table 78.2 Linguistic variables for the ratings of criteria

Linguistic variable	Interval type-2 fuzzy number (IT2FN)
Very poor (VP)	((0,1;1), (0,0.5;1))
Poor (P)	((0,3;1), (0.5,2;1))
Medium poor (MP)	((1,5;1), (2,4;1))
Medium (M)/fair (F)	((3,7;1), (4,6;1))
Medium good (MG)	((5,9;1), (6,8;1))
Good (G)	((7,10;1), (8,9.5;1))
Very good (VG)	((9,10;1), (9.5,10,1))

Table 78.3 Weights of the attributes evaluated by decision-makers

Attributes	Decision-makers		
	D ₁	D ₂	D ₃
C ₁	H	VH	MH
C ₂	VH	VH	VH
C ₃	VH	H	H
C ₄	VH	VH	VH
C ₅	M	MH	MH

Moreover, all the relative importance ratings (i.e. the criteria values) in this numerical example are described using the linguistic variables which are defined in Table 78.2.

Assume that there are three decision-makers, D_1 , D_2 , and D_3 of a software company to hire a system analysis engineer and assume that there are three alternatives x_1, x_2, x_3 and five attributes “Emotional Steadiness” (denoted by C_1), “Oral Communication Skill” (denoted by C_2), “Personality” (denoted by C_3), “Past Experience” (denoted by C_4), “Self-Confidence” (denoted by C_5). Let X be the set of alternatives, where $X = \{x_1, x_2, x_3\}$, and let F be the set of attributes, where $F = \{\text{Emotional Steadiness, Oral Communication Skill, Personality, Past Experience, Self-Confidence}\}$. Assume that there are three decision-makers D_1, D_2 , and D_3 used the linguistic terms shown in Table 78.1 to represent the weights of the four attributes, respectively, as shown in Table 78.3.

Then these three decision-makers D_1, D_2 , and D_3 used the linguistic terms shown in Table 78.2 to represent the evaluating values of the alternatives with respect to different attributes, respectively, as shown in Table 78.4.

Using the linguistic scales from Tables 78.3 and 78.4, and the eight steps of the proposed method (in Sect. 78.3), results for Chen [9] example is shown in Table 78.5. Table 78.5 shows the min value and max value from Step 6 and calculates the closeness coefficient $(CC)_i$ for each of alternatives.

Table 78.4 Linguistic of decision matrix

Attributes	Alternatives	Decision-makers		
		D ₁	D ₂	D ₃
C ₁	x ₁	MG	G	MG
	x ₂	G	G	MG
	x ₃	VG	G	F
C ₂	x ₁	G	MG	F
	x ₂	VG	VG	VG
	x ₃	MG	G	VG
C ₃	x ₁	F	G	G
	x ₂	VG	VG	G
	x ₃	G	MG	VG
C ₄	x ₁	VG	G	VG
	x ₂	VG	VG	VG
	x ₃	G	VG	MG
C ₅	x ₁	F	F	F
	x ₂	VG	MG	G
	x ₃	G	G	MG

Table 78.5 Final ranking order

	Min	Max	Closeness coefficient, (CC) _i
x ₁	2.3575	2.6364	0.4721
x ₂	2.3835	2.5944	0.4788
x ₃	2.3668	2.6216	0.4745

As shown in Table 78.5, results for the relative closeness of Chen’s method [9] for three alternatives are 0.4721 for x₁, 0.4788 for x₂ and 0.4745 for x₃; which lead to the ranking of x₂ > x₃ > x₁. Chen [9]’s result coincides with the proposed results.

78.5 Conclusion

This paper distributed a linear assignment method which consisted with the SAW and LP method for IT2 FTOPSIS. This proposed method is able to produce an optimal ranking order of the alternatives. Besides, we provided a numerical example to analyze the applicability of the proposed method. The proposed method can capture the imprecise and uncertain decision information instead of the optimal ranking orders. Furthermore, the proposed method offers an alternative ways of ranking phase for IT2 FTOPSIS method.

Acknowledgments This research is supported by MyBrain15 scholarship and Fundamental Research Grant, no. 59243. This support is gratefully acknowledged.

References

1. Gao, P., Feng, J., Yang, L.: Fuzzy TOPSIS algorithm for multiple criteria decision making with an application in information systems project selection. In: Proceeding of IEEE WiCOM '08. 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2008)
2. Li, D.: TOPSIS-based nonlinear-programming methodology for multiattribute decision making with interval-valued intuitionistic fuzzy sets. *IEEE Trans. Fuzzy Syst.* **18**, 299–311 (2010)
3. Jolai, F., Yazdian, S.A., Shahanaghi, K., Khojasteh, M.A.: Integrating fuzzy TOPSIS and multi-period goal programming for purchasing multiple products from multiple suppliers. *J. Purchas. Suppl. Manage.* **17**, 42–53 (2012)
4. Chen, S.-M., Lee, L.-W.: Fuzzy multiple attributes group decision-making based on the interval type-2 TOPSIS method. *Expert Syst. Appl.* **37**, 2790–2798 (2010)
5. Chen, S.-M., Yang, M.-W., Lee, L.-W., Yang, S.-W.: Fuzzy multiple attributes group decision-making based on ranking interval type-2 fuzzy sets. *Expert Syst. Appl.* **39**, 5295–5308 (2012)
6. Wang, W., Liu, X., Qin, Y.: Multi-attribute group decision making models under interval type-2 fuzzy environment. *Knowl.-Based Syst.* **30**, 121–128 (2012)
7. Chen, S.-M., Wang, C.-Y.: Fuzzy decision making systems based on interval type-2 fuzzy sets. *Inf. Sci.* **242**, 1–21 (2013)
8. Chen, T.-Y.: A linear assignment method for multiple-criteria decision analysis with interval type-2 fuzzy sets. *Appl. Soft Comput.* **13**, 2735–2748 (2013)
9. Chen, C.T.: Extension of the TOPSIS for group decision making under fuzzy environment. *Fuzzy Sets Syst.* **114**, 1–9 (2000)
10. Charness, A.A., Cooper, W.W.: Programming with linear fractional functionals. *J. Naval Res. Logist. Quart.* **9**, 181–185 (1962)
11. Charness, A.A., Cooper, W.W.: An explicit general solution in linear fractional programming. *J. Naval Res. Logist. Quart.* **20**, 181–185 (1973)
12. Guh, Y.-Y., Hon, C.-C., Lee, E.S.: Fuzzy weighted average: the linear programming approach via charnes and cooper's rule. *J. Fuzzy Sets Syst.* **117**, 157–160 (2001)

Chapter 79

Hybridization Denoising Method for Digital Image in Low-Light Condition

Suhaila Sari, Sharifah Zahidah Hasan Al Fakkri, Hazli Roslan
and Zarina Tukiran

Abstract A good noise reduction is a method that can reduce the noise level and preserve the details of the image. This paper proposes a denoising method through hybridization of bilateral filters and wavelet thresholding for digital image in low-light condition. The proposed method is experimented on selected night vision images and the performances are evaluated in terms of Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR) and visual effects. Results demonstrate that the proposed denoising method has improved the PSNR and MSE of average performance of bilateral filter by 0.97 dB and 1.33 respectively and the average performance of wavelet thresholding has improved by 0.98 dB and 1.19 respectively.

Keywords Image denoising · Edge preservation · Low-light condition · Poisson noise · Bilateral filter · Wavelet thresholding

79.1 Introduction

Low-light noise is a significant problem in photography. Most consumers' cameras have poor low-light characteristics, which typically result in images with noticeable noise artifacts. For example, digital images taken by consumer's digital camera

S. Sari (✉) · S.Z.H. Al Fakkri · H. Roslan · Z. Tukiran
Artificial Intelligence and Computer Vision Focus Group (AICOV), Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia
e-mail: suhailas@uthm.edu.my

S.Z.H. Al Fakkri
e-mail: eyzayda@gmail.com

H. Roslan
e-mail: hazli@uthm.edu.my

Z. Tukiran
e-mail: zarin@uthm.edu.my

may suffer from thermal sensor noise demosaicing noise and quantization noise [1]. Although camera technology has been improved extensively, noise still could not be eliminated completely [2].

In image denoising process, noise is undesired information that contaminates the image [3]. The information about the type of noise corrupting in the original image plays a significant role to determine the denoising technique. Typical images are corrupted with noise modeled as either or combination of the Gaussian noise, Salt and Pepper noise, Speckle noise, or Poisson noise.

Many denoising methods have been proposed over the years, such as the Wiener filter, wavelet thresholding [4, 5], anisotropic filtering [6], bilateral filter [7], total variation method [8], and non-local means methods [9]. Among these, wavelet thresholding has been reported to be a highly successful method. The wavelet thresholding crucial task is selection of threshold value. The wavelet decomposed a signal into low-frequency and high-frequency subbands, and the coefficients in the detail subbands are processed via hard or soft thresholding [10]. Despite its success, this method, however, experienced a problem on preserving the edge.

The bilateral filter was proposed by Tomasi and Manduchi [7] as an alternative to wavelet thresholding. It applies spatially weighted averaging. This is achieved by combining two Gaussian filters; one filter works in spatial domain, the other in the intensity domain. Therefore, not only the spatial distance, but the intensity distance is also important for the determination of weights [11]. Hence, these types of filters can remove the noise in an image effectively [12].

Combination of bilateral filter and wavelet thresholding was applied by Zhang and Guntruk to eliminate noise in real noisy images [12]. The similar method is also used by Roy, Sinha and Sen in order to eliminate noise in medical image [13]. The noise modeled in [12, 13] is Gaussian noise. Through these researches, the method has proven significant improvement in noise reduction for images corrupted by Gaussian noise.

Since the Gaussian noise distribution has some similar characteristic with Poisson noise [14], the method is expected to provide effective noise reduction for Poisson noise as well. Therefore, in this work, the method is proposed to reduce the Poisson noise on digital image captured in low-light condition.

The proposed method is based on bilateral filter and wavelet thresholding which exploits features of both bilateral filter and wavelet thresholding at the same time the limitations of both are overcome. To our knowledge, such studies in hybridization of bilateral filtering and wavelet thresholding for Poisson noise reduction in digital image have not yet been done.

The paper is organized as follows. The proposed hybrid denoising method is described in Sect. 79.2. Results are discussed in Sect. 79.3. Finally, concluding remarks are drawn in Sect. 79.4.

79.2 Proposed Method

In this work, hybridization of bilateral filter and wavelet thresholding is proposed to find the best solution in image denoising especially for digital image in low light condition. This proposed method utilizes the advantages of both bilateral filter and wavelet thresholding. Therefore the proposed method is expected to contribute a better preservation of fine details and edges in the image. While applying the threshold rule, the important features like edges, curves and textures can be identified.

79.2.1 Wavelet Thresholding

The wavelets play a major role in image compression and image denoising. Wavelet coefficients calculated by a wavelet transform represent change in the time series at a particular resolution. It is then possible to filter out noise by considering the time series at various resolutions. The term wavelet thresholding is explained as decomposition of the data or the image into wavelet coefficients, comparing the detail coefficients with a given threshold value, and shrinking these coefficients close to zero to take away the effect of noise in the data.

During thresholding, a wavelet coefficient is compared with a given threshold and is set to zero if its magnitude is less than the threshold; otherwise, it is retained or modified depending on the threshold rule. Thresholding distinguishes between the coefficients due to noise and the ones consisting of important signal information.

There are two general categories of thresholding which are hard thresholding and soft thresholding. Hard thresholding is a “keep or kill” procedure and is more intuitively appealing while soft thresholding shrinks coefficients above the threshold in absolute value. Wavelet threshold is effective in reducing certain amount of noise while preserving fine details and edges.

79.2.2 Bilateral Filter

The bilateral filter is a nonlinear weighted averaging filter. The weights depend on both the spatial distance and the intensity distance with respect to the center pixel. It smoothes images while preserving edges with the nonlinear combination of nearby pixel values. This can be achieved by combining two Gaussian filters; one filter works in spatial domain, the other filter works in intensity domain. Therefore, not only the spatial distance but the intensity distance is also important for the determination of weights.

The weakness of the bilateral filter is its inability to remove salt and pepper type of noise. It could not access to the different frequency components of a signal which fails to remove low frequency noise. Besides that, there is no study on the

optimal values of σ_d and σ_r , which are the parameters that control the behavior of the bilateral filter. Bilateral filter is known to have good performance in noise reduction, but tend to over smooth the fine details and edges.

79.2.3 Hybridization of Bilateral and Wavelet Thresholding

This work proposed hybridization of bilateral and wavelet thresholding (db4, Bayes soft thresholding, level of decomposition = 1) for digital image in low-light condition. The process flow for the proposed method is shown in Fig. 79.1.

79.3 Results and Discussion

We have conducted some tests to the selected night vision images of well-known landmark in Malaysia named as *Kuala Lumpur Conventional Center in Kuala Lumpur (klcc)*, *Dataran Bandaraya Johor Clock Tower in Johor (jbTower)* and *Dewan Undangan Negeri Kuching in Sarawak (kuchingHall)*. The image size of these images is 250×250 . In order to conduct the testing, these images are corrupted with Poisson noise. All of these images are tested with conventional methods which are bilateral filtering and wavelet threshold, as well as proposed hybrid denoising method.

To do a quantitative comparison, MATLAB software is used to simulate the test images under different Poisson distribution, $\lambda = 0.00001$, $\lambda = 0.001$, and $\lambda = 0.1$ which represents low, medium and high noise, respectively. From the test, there are three evaluation methods that we utilized, which are (a) Mean Squared Error (MSE), (b) Peak Signal to Noise Ratio (PSNR) and (c) visual comparison.

Equation (79.1) is used to calculate the MSE where A represents the original image while B represents the reconstructed image. Number of rows and columns in the input images are represented by M and N respectively.

$$\text{MSE} = \frac{1}{M \times N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [A(i,j) - B(i,j)]^2 \quad (79.1)$$

Meanwhile Eq. (79.2) is used to calculate the PSNR where MAX is the maximum possible pixel value of the image.

$$\text{PSNR} = 10 \times \log_{10}(MAX^2 / \text{MSE}) \quad (79.2)$$

The MSE and PSNR values obtained for all methods using all images stated in this section are shown in Tables 79.1 and 79.2 respectively.

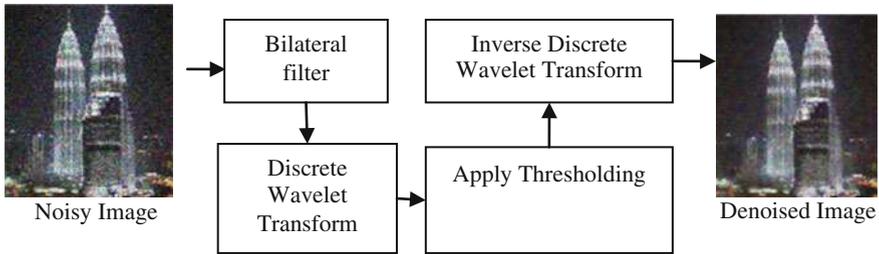


Fig. 79.1 Block diagram of image denoising using hybridization of bilateral filter and wavelet thresholding

Table 79.1 MSE evaluation for high Poisson noise ($\lambda = 0.1$)

Test images	Bilateral filter	Wavelet filter	Proposed method
<i>klcc</i>	0.11	0.10	0.08
<i>jbTower</i>	0.88	0.07	0.60
<i>kuchingHall</i>	0.10	0.09	0.08

Table 79.2 PSNR evaluation for high Poisson noise ($\lambda = 0.1$)

Test images	Bilateral filter	Wavelet filter	Proposed method
<i>klcc</i>	67.14	68.32	70.04
<i>jbTower</i>	70.36	71.07	72.26
<i>kuchingHall</i>	68.62	69.42	70.66

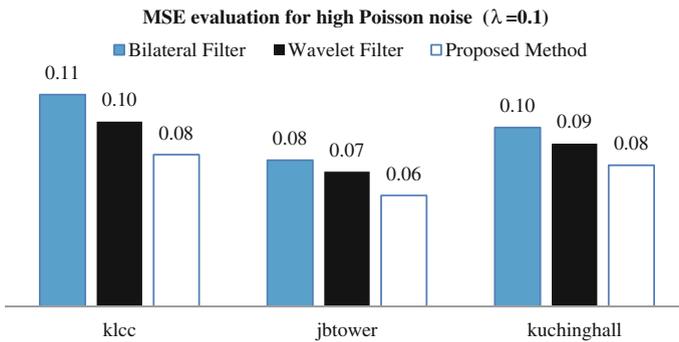


Fig. 79.2 Performance comparisons in terms of MSE

Figures 79.2 and 79.3 depict the graphical comparative performance in terms of MSE and PSNR. The results show that the proposed method has significantly better denoising performance in comparison to bilateral filter alone and wavelet thresholding alone.

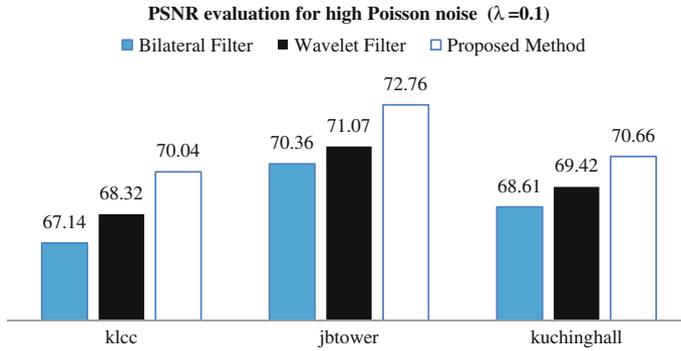


Fig. 79.3 Performance comparisons in terms of PSNR



Fig. 79.4 The visual effects comparison of *Kuching Hall* image for (top row) low Poisson noise level with $\lambda = 0.00001$, (middle row) medium Poisson noise level with $\lambda = 0.001$ and (bottom row) high Poisson noise level with $\lambda = 0.1$

Figure 79.4 shows the visual effects comparison of denoising result for different night vision images corrupted by low, medium and high Poisson noise level. Figure 79.5 shows the visual effects in comparison of all methods using all images with high Poisson noise.

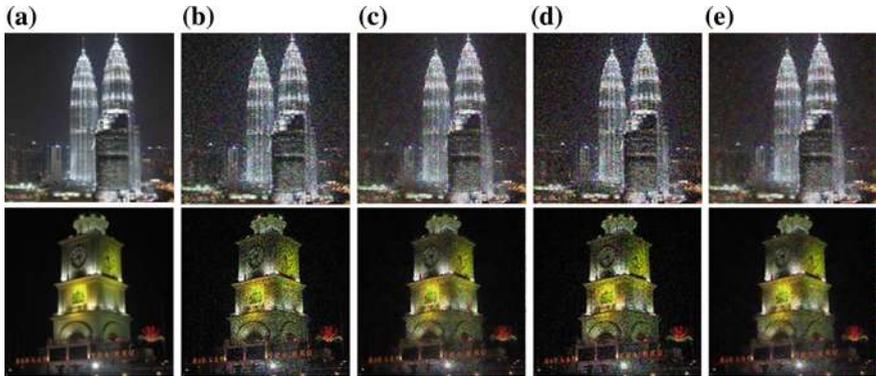


Fig. 79.5 The visual effects comparison of *klcc* image (above) and *jbTower* image (bottom) for, **a** Original image, **b** Noisy image with high Poisson noise level with $\lambda = 0.1$, **c** Denoised image using bilateral filter, **d** Denoised image using wavelet thresholding and **e** Denoised image using hybridization of bilateral filter and wavelet thresholding

The visual effects results show that although the output images from bilateral filter looks smooth and clear, but most of the fine details and edges are oversmoothed and the image blurred. Therefore, by applying wavelet thresholding, it multiplies the adjacent wavelet sub bands and strengthens the significant features in the image which indicates that our proposed method is better in image denoising.

79.4 Conclusion

In this paper, image denoising using hybridization, which integrates bilateral filter and wavelet thresholding is performed. From the result, we can conclude that the proposed method has provided better denoising for images in low-light condition corrupted by high level Poisson noise. The performance of the proposed method has improved the average performances of bilateral filter and wavelet thresholding in terms of the PSNR by 0.98 dB and 0.97 dB, respectively. On the other hand, the performance of the proposed method has improved the average performances of wavelet thresholding and bilateral filter in terms of the MSE by 1.19 and 1.33, respectively. As for the visual effects, the proposed method provides better denoising while preserving the fine details and edges.

Acknowledgements The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) (Grant vote: 1088) and Malaysia Government for the support and sponsor of this study.

References

1. Roy, S., Sinha, N., Sen, A.K.: A new hybrid image denoising method. *Int. J. Inf. Technol. Knowl. Manag.* **2**(2), 491–497 (2010)
2. Chatterjee, P., Joshi, N., Kang, S.B., Matsushita, Y.: Noise suppression in low-light images joint denoising and demosaicing. In: *Proceedings of IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pp. 321–328 (2011)
3. Teimouri, M., Vahedi, E., Avanaki, A.N., Shahi, Z.H.: An efficient denoising method for color images. In: *Proceedings of IEEE Conference on Signal Processing and Its Applications*, pp. 1–4 (2007)
4. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3), 425–455 (1994)
5. Ogawa, K., Sakata, M., Li, Y.: Adaptive noise reduction of scintigrams with a wavelet transform. *Int. J. Biomed. Imaging.* **2012** (2012). <http://dx.doi.org/10.1155/2012/130482>
6. Perona, P., Malik, J.: Scale-space and Edge detection using anisotropic diffusion. *Proc. IEEE Trans. Pattern Anal. Mach. Intell.* **12**(7), 629–639 (1990)
7. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: *Proceedings of International Conference on Computer Vision*, pp. 839–846 (1998)
8. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* **60**, 259–268 (1992)
9. Buades, A., Coll, B., Morel, J.: Neighborhood filters and PDE's. *Numer. Math.* **105**(1), 1–34 (2006)
10. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90**(432), 1200–1224 (1995)
11. Chang, S.G., Yu, B., Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression. *Proc. IEEE Trans. Image Process.* **9**(9), 1532–1546 (2000)
12. Zhang, M., Gunturk, B.: A new image denoising method based on the bilateral filter. In: *ICASSP, IEEE*, pp. 929–932 (2008)
13. Roy, S., Sinha, N., Sen, A.K.: An efficient denoising model based on wavelet and bilateral filters. *Int. J. Comput. Appl.* **53**(10), 0975–8887 (2012)
14. Portilla, J., Strela, V., Wainwright, M., Simoncelli, E.P.: Image denoising using scale mixtures of gaussians in the wavelet domain. *Proc. IEEE Trans. Image Process.* **12**(2), 1338–1351 (2003)

Chapter 80

The Improved Models of Internet Pricing Scheme of Multi Service Multi Link Networks with Various Capacity Links

Fitri Maya Puspita, Kamaruzzaman Seman and Bachok M. Taib

Abstract Internet Service Providers (ISPs) nowadays deal with high demand to promote good quality information. However, the knowledge to develop new pricing scheme that serve both customers and supplier is known, but only a few pricing plans involve QoS networks. This study will seek new proposed pricing plans are offered with multi service multi link networks involved. The multi service multi link Networks scheme is solved as an optimization model by comparing two models in multi QoS networks. The results showed that by fixing the base price and varying the quality premium or varying the base price and quality and setting up the equal capacity link values, ISP achieved the goal to maximize the profit.

80.1 Introduction

Recent works on multiple QoS networks are due to [1–4]. They described the pricing scheme based auction to allocate QoS and maximize ISP's revenue. The auction pricing scheme is actually scalability, efficiency and fairness in sharing resources. The solution of the optimization problem goes from single bottleneck link in the network and then she generalized into multiple bottleneck links using heuristic method. In this paper, she used only single QoS parameter-bandwidth, while in networks, there are many parameters affect QoS that can be considered.

F.M. Puspita (✉) · K. Seman · B.M. Taib
Faculty of Science and Technology, Universiti Sains Islam Malaysia,
71800 Nilai, Negeri Sembilan Darul Khusus, Malaysia
e-mail: pipitmac140201@gmail.com

K. Seman
e-mail: drkzaman@usim.edu.my

B.M. Taib
e-mail: bachok@usim.edu.my

Although QoS mechanisms are available in some researches, there are few practical QoS network. Even recently a work in this QoS network [5], it only applies simple network involving one single route from source to destination.

In previous discussion works on multiservice network proposed by [6–8], we work on single and multiple link networks to solve the internet pricing scheme. This work on multiple link multiservice networks can also be improved by considering all cases of capacity needed in the networks. The results show our improved method results in better optimal solution than previously conducted by other research.

So, we intend to improve the mathematical formulation of [5, 9] to be simpler formulation by taking into consideration the utility function, base price as fixed price or varies, quality premium, index performance, capacity and also bandwidth required by looking at all possibility of capacity needed in the network. Next we consider the problem of internet charging scheme as Mixed Integer Nonlinear Programming (MINLP). The solver LINGO 13.0 [10] were applied to assist the nonlinear programming solution to obtain the optimal solution.

80.2 Related Works

A number of related works has been listed for differentiated pricing scheme that works on multi QoS networks. In paper proposed by [11], they discussed about pricing scheme that is based on QoS level in different allocation to control congestion and load balance. Multiple class networks require differentiated pricing scheme to have allocation of different level of service traffic. The investigation of the connection between QoS characteristics at network with requirement of quality for users applying the network is proposed by [12]. The results of their findings mainly are that predictability, consistency of QoS is crucial, the pricing scheme is crucial to have QoS to be predicted, having reliable service protocols and new integrated service mechanism is to present alternative in solving the problem. There exist direct connection between QoS profile application identification in packet and users requesting the QoS. Models proposed by [13] viewed the relationship between congestion control, routing and scheduling of wired network as fair resource allocation. In the research explained in [14], they discussed the flat fee pricing scheme, and as the simplicity price is maximum revenue. The drawback of the rule is due to nonlinearity and does not reflect the price observed in reality.

Models proposed by [15] stated that in network, it is assumed that n users can be split into k categories. Each category can apply the same service offered by application server in a shared link with total bandwidth C_{tot} but has different demand framework and also difference price sensitivity. Alderson et al. [16] discussed issues related to ISP problems dealing with topology of the networks such as link cost, router technology which impact on availability of topology to

network creator then dealing with equipment of routing adopted to tackle network traffic flow.

Other research proposed by [17] stated about ways to solve internet optimization that includes system definition as an interest function and view it as different via points and system mathematical definition. Problem proposed by [18] focused on problem of arrangement of web services and explained about model of multi-dimension QoS. Framework presented by [19] analyzed the pricing problem in integrated service network having guaranteed QoS. The Method proposed by [20] define terms for performance prediction of service-base system that consist of performance showing the how fast completion time to finish a service request, time interval showing the time period to complete service request, dependability showing the capability of web service to conduct conditional required function, price setting up by ISP and reputation showing that user perception to the service. The method proposed by [21] explained about monopoly in pricing model strategy based on pay-per-volume and pay-per-time of network. They conclude that ISP will gain more benefit by providing pricing scheme based volume since this scheme is an alternative to numerous users and scheme of pay-per volume will benefit the network provider and can prevent from bursting the networks.

Study on multiservice was investigated by [9] which discussed problem of pricing of internet by considering network share, availability capacity in each service, the number of users available for the service and the QoS level. They solved the internet pricing by transforming the model into optimization model and solved using Cplex software. Recent works conducted by [22–27] also discussed internet charging scheme under multiple class QoS networks by comparing two models that involve base price as a fixed and variable set up by ISP. The model created by setting up base price as fixed price will yield higher optimal solution if ISP intended to recover the cost. But if ISP would like to compete in market, then the choice of model involving base price as variable price would be the best option to choose.

The work discussed on the botnet attacks detection by using nepenthes honeypots [28] is also crucial to be issued for problems increasing if we are dealing with security problem in multiservice networks.

80.3 Research Method

We attempt to apply optimization techniques in solving the problem in this paper. We also consider the optimization problem as MBINLP that can be solved by using optimization tools, LINGO 13.0. We transform the problem of pricing the internet in multi service networks into optimization model and attempt to solve it to get optimal solution. This solution will help us interpreting the current issues involving pricing, network share, base price, quality premium and also QoS level.

80.3.1 Model Formulation

The idea basically generates from [5, 8, 9] and is improved in multilink multi service networks by considering various cases where we can set up requirements for the capacity link.

80.3.2 Assumptions

Assume that there is only one single network from source to destination since concentrate on service pricing scheme. Assume that the routing schemes are already set up by the ISP. As [1–4] pointed out, we have 2 parts of utility function namely, base cost which does not depend on resource consumption and cost which depends on resource consumption. The parameters, decision variables and the models are adopted in [7, 8] and are described as follows.

The parameters are as follows.

- α_j base price for class j , can be fixed or variables
- β_j quality premium of class j that has I_j service performance
- C_l total capacity available in link l
- p_{il} price a user willing to pay for full QoS level service of i in link l

The decision variables are as follows.

- x_{il} number of users of service i in link l
- a_{il} reserved share of total capacity available for service i in link l
- I_i quality index of class i

Formulation when we assign α and β fixed is as follows.

$$\text{Max} \sum_{l=1}^L \sum_{i=1}^S (\alpha + \beta I_i) p_{il} x_{il} \tag{80.1}$$

Such that

$$I_i d_{il} x_{il} \leq a_{il} C_l, \quad i = 1, \dots, S, \quad l = 1, \dots, L \tag{80.2}$$

$$\sum_{l=1}^L \sum_{i=1}^S I_i d_{il} x_{il} \leq C_l, \quad i = 1, \dots, S; \quad l = 1, \dots, L \tag{80.3}$$

$$\sum_{l=1}^L a_{il} = 1, \quad i = 1, \dots, S \tag{80.4}$$

$$0 \leq a_{il} \leq 1, i = 1, \dots, S; l = 1, \dots, L \quad (80.5)$$

$$m_i \leq I_i \leq 1, i = 1, \dots, S \quad (80.6)$$

$$0 \leq x_{il} \leq n_i, i = 1, \dots, S; l = 1, \dots, L \quad (80.7)$$

With m_i and n_i are prescribed positive integer numbers.

$$\{x_{il}\} \text{integer} \quad (80.8)$$

Formulation when we assign α fixed and β vary is as follows.

$$\text{Max} \sum_{l=1}^L \sum_{i=1}^S (\alpha + \beta_i I_i) p_{il} x_{il} \quad (80.9)$$

Subject to Eqs. (80.2–80.8) with additional constraints as follows.

$$\beta_i I_i \geq \beta_{i-1} I_{i-1}, i > 1, i = 1, \dots, S \quad (80.10)$$

$$k \leq \beta_i \leq q, [k, q] \in [0, 1] \quad (80.11)$$

Formulation we have when α and β vary

$$\text{Max} \sum_{l=1}^L \sum_{i=1}^S (\alpha_i + \beta_i I_i) p_{il} x_{il} \quad (80.12)$$

Subject to Eqs. (80.2–80.8) and (80.10) with additional constraints

$$\alpha_i + \beta_i I_i \geq \alpha_{i-1} + \beta_{i-1} I_{i-1}, i > 1, i = 1, \dots, S \quad (80.13)$$

$$y \leq \alpha_i \leq z, [y, z] \in [0, 1] \quad (80.14)$$

Formulation when we have α vary and β fixed

$$\text{Max} \sum_{l=1}^L \sum_{i=1}^S (\alpha_i + \beta I_i) p_{il} x_{il} \quad (80.15)$$

Subject to Eqs. (80.2–80.8) and (80.13–80.14).

Since ISP wants to get revenue maximization by setting up the prices chargeable for α , β and QoS level to recover cost and to enable the users to choose services based on their preferences like stated in Eq. (80.1). Equation (80.2) shows that the required capacity of service does not exceed the network capacity reserved. Equation (80.3) explains that required capacity cannot be greater than the network capacity C in link l . Equation (80.4) guarantee that network capacity

has different location for each service that lies between 0 and 1 (80.5). Equation (80.6) explains that QoS level for each service is between the prescribed ranges set up by ISP. Equation (80.7) shows that users applying the service are nonnegative and cannot be greater than the highest possible users determined by ISP. Equation (80.8) states that the number of users should be positive integers. Equation (80.9) explains that ISP wants to get revenue maximization by setting up the prices chargeable for α , β and QoS level to recover cost and to enable the users to choose services based on their preferences. Equation (80.10) explains that β has different level for each service which is at least the same level or lower level. Equation (80.11) states that value of β lies between two prescribed values. ISP wants to get revenue maximization by setting up the prices chargeable for α , β and QoS level to recover cost and to enable the users to choose services based on their preferences like stated in Eq. (80.12). Equation (80.13) explains that the summation of α and β has different level for each service which is at least the same level or lower level. Equation (80.14) shows that the base price should lie between predetermined α set up by ISP. ISP wants to get revenue maximization by setting up the prices chargeable for α , β and QoS level to recover cost and to enable the users to choose services based on their preferences as stated in objective function (80.15).

80.4 Optimal Solution

Tables 80.1 and 80.3 below describes the solver status of model formulation in LINGO when considering base price to be fixed and model formulation when considering base price to be varied. Generated Memory Used (GMU) shows that how much the amount of memory use for generating a model. The total time used so far to generate and solve the model. In Table 80.1, the highest optimal solution of 811.2 is achieved when ISP sets up α to be fixed and vary β and equals the capacity link values with GMU = 32 K and ER = 1 s. In Table 80.3, the highest optimal solution of 912.6 is achieved when ISP varies α , β and equals the capacity link values with GMU = 33 K and ER = 1 s.

Tables 80.2 and 80.4 depict the optimal solutions of the formulation by setting up the base price value to be fixed or vary the base price. In Table 80.2, when the formulation of α to be fixed and vary β , the QoS level of 0.8 is achieved by service 1 and service 2 with $C_1 = C_2$. But only 10 users apply the service in service 2 on link 1 with 100 % network is reserved and 9 users on link 2 with 90 % network is reserved.

Table 80.1 Solver status and extended solver state of the models by considering α to be fixed for three capacity link cases

Solver status	α and β fixed			α fixed and β vary		
	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$
Model class	INLP	INLP	INLP	INLP	INLP	INLP
State	Local optimal	Local optimal	Local optimal	Local optimal	Local optimal	Local optimal
Objective	492.151	477.4	569.831	686.4	667.2	811.2
Infeasibility	2.7×10^{-13}	0.3×10^{-3}	1.8×10^{-7}	1.1×10^{-16}	0.1×10^{-15}	0.9×10^{-12}
Iterations	2,186	259	818	614	375	181
Solver type	B & B	B & B	B & B	B & B	B & B	B & B
Best Objective	492.151	477.4	569.831	686.4	667.2	811.2
Objective Bound	492.151	477.4	569.831	686.4	667.2	811.2
Steps	51	9	20	26	8	6
Active	0	0	0	0	0	0
Update interval	2	2	2	2	2	2
GMU (K)	30	30	30	32	31	32
ER (sec)	2	0	1	1	0	1

Table 80.2 Solutions of Models by considering α to be fixed for three capacity link cases

Var	$\alpha = 0.5$ and $\beta = 0.01$ (fixed)			$\alpha = 0.5$ (fixed) and β vary		
	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$
α_1	–	–	–	–	–	–
α_2	–	–	–	–	–	–
α_3	–	–	–	–	–	–
β_1	–	–	–	0.375	0.375	0.375
β_2	–	–	–	0.375	0.375	0.375
β_3	–	–	–	0.3	0.3	0.3
I_1	0.8	0.83	0.83	0.8	0.8	0.8
I_2	0.8	0.8	0.8	0.8	0.8	0.8
I_3	0.5	0.5	0.5	1	1	1
a_{11}	0.024	0.025	0.025	0.096	0	0
a_{21}	0.6	0.7	0.7	0.904	1	1
a_{31}	0.375	0.275	0.275	0	0	0
x_{11}	2	3	3	8	0	0
x_{21}	4	7	7	6	10	10
x_{31}	9	10	10	0	0	0
a_{12}	0.025	0.12	0.025	0.08	0.1	0.08
a_{22}	0.7	0.46	0.7	0.92	0.9	0.9
a_{32}	0.275	0.4	0.275	0	0	0.02
x_{12}	3	10	3	10	8	10
x_{22}	7	3	7	9	6	9
x_{32}	10	10	10	0	0	0

The Highest QoS level of 1 is achieved by service 3 but no users apply the service. When the formulation of varying α and β , as stated in Table 80.4, QoS level of 0.8 is achieved by service 1 and 2 also of 1 in service 3 but only in service 2, 10 users apply the service in link 1 and 9 users in link 2 with 100 % network reserved for service 2 in link 1 and 90 % network reserved in link 2.

To sum up, the objective of ISP to achieve the maximum profit will be reached if ISP set up the base price to be fixed and vary the quality premium or to vary the base price and quality premium with setting up the equal capacity link values ($C_1 = C_2$).

Table 80.3 Solver status and extended solver state of the models by considering α being varied for three capacity link cases

Solver status	α vary and β fixed			α and β vary		
	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$
Model class	INLP	INLP	INLP	INLP	INLP	INLP
State	Local optimal	Local optimal	Local optimal	Local optimal	Local optimal	Local optimal
Objective	585.144	563.2	677.421	772.2	750.6	912.6
Infeasibility	0.0015	0.3×10^{-3}	3.7×10^{-5}	8.7×10^{-13}	0	1.7×10^{-17}
Iterations	1,517	348	736	1,622	462	188
Solver type	B & B	B & B	B & B	B & B	B & B	B & B
Best Objective	585.144	563.2	677.421	772.2	750.6	912.6
Objective Bound	585.144	563.2	677.421	772.2	750.6	912.6
Steps	26	16	12	25	11	6
Active	0	0	0	0	2	0
Update interval	2	2	2	2	0	2
GMU (K)	32	32	32	33	33	33
ER (sec)	1	0	2	1	0	1

Table 80.4 Solutions of models by considering α being varied for three capacity link cases

Var	α vary and $\beta = 0.01$ (fixed)			α and β vary		
	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$	$C_1 < C_2$	$C_1 > C_2$	$C_1 = C_2$
α_1	0.5	0.5	0.5	0.25	0.26	0.26
α_2	0.597	0.59	0.59	0.53	0.53	0.63
α_3	0.6	0.6	0.6	0.6	0.6	0.6
β_1	–	–	–	0.8	0.8	0.8
β_2	–	–	–	0.45	0.45	0.34
β_3	–	–	–	0.3	0.3	0.3
I_1	0.8	0.83	0.8	0.8	0.8	0.8
I_2	0.8	0.8	0.8	0.8	0.8	0.8
I_3	0.5	0.5	0.5	1	1	1
a_{11}	0.024	0.025	0.024	0.096	0	0
a_{21}	0.602	0.7	0.7	0.9	1	1
a_{31}	0.373	0.275	0.276	0.003	0	0
x_{11}	2	3	3	8	0	0
x_{21}	4	7	7	6	10	10
x_{31}	9	10	10	0	0	0
a_{12}	0.024	0.12	0.024	0.08	0.1	0.08
a_{22}	0.7	0.46	0.7	0.91	0.9	0.9
a_{32}	0.276	0.4	0.276	0.012	0	0.02
x_{12}	3	10	3	10	8	10
x_{22}	7	3	7	9	6	9
x_{32}	10	10	10	0	0	0

80.5 Conclusion

The paper [9] be more upgraded by using our new approach using other tools. We obtain slightly increasing profit in several solutions we proposed. We also save human resources by only applying few users to apply the service and also we can save energy by only promote one service rather than two services. Our solutions show better profit with less idle time and number of users applied the services.

We have shown that by considering new parameters, more decision variables and constraints, we obtain better revenue maximization. The cases shown above basically are ISP strategy to vary its preference to achieve their goals. ISP is able to adopt the cases to suit their goals. But again, like stated in since it is more theoretical point of view and assumptions, we limit our result only static result in data changes, and cost preference is just based on our discrete data. Further research should address more generalization of the model to also consider numerous services offered or generalization of more services.

Acknowledgments The research leading to this study was financially supported by Ministry of Higher Education Malaysia for support through Fundamental Research Grant Scheme (FRGS) 2011, Research Code: USIM/FRGS-FST-5-50811.

References

1. Yang, W., Owen, H.L., Blough, D.M., Guan, Y.: An auction pricing strategy for differentiated service network. In: Proceedings of the IEEE Global Telecommunications Conference. IEEE (2003)
2. Yang, W.: Pricing Network resources in differentiated service networks. In: School of electrical and computer engineering. Phd Thesis. Georgia Institute of Technology. pp. 1–111 (2004)
3. Yang, W., Owen H., Blough, D.M.: A Comparison of auction and flat pricing for differentiated service networks. In: Proceedings of the IEEE International Conference on Communications (2004)
4. Yang, W., Owen, H.L., Blough, D.M.: Determining Differentiated services network pricing through auctions. In: Networking-ICN 2005, 4th International Conference on Networking April 2005 Proceedings, Part I. Reunion Island, France, Springer-Verlag, Berlin Heidelberg (2005)
5. Byun, J., Chatterjee, S.: A strategic pricing for quality of service (QoS) network business. In Proceedings of the Tenth Americas Conference on Information Systems, New York (2004)
6. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: A new approach of optimization model on internet charging scheme in multi service networks. *Int. J. Sci. Tech.* **2**(6), 391–394 (2012)
7. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: An improved optimization model of internet charging scheme in multi service networks. *TELKOMNIKA* **10**(3), 592–598 (2012)
8. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: An Improved model of internet pricing scheme of multi service network in multiple link QoS networks. In: The 2013 International Conference on Computer Science and Information Technology (CSIT-2013). Universitas Teknologi Yogyakarta (2013)
9. Sain, S., Herpers, S.: Profit Maximisation in multi service networks- an optimisation model. In: Proceedings of the 11th European Conference on Information Systems ECIS 2003. Naples, Italy (2003)
10. LINGO: LINGO 13.0.2.14. LINDO Systems, Inc: Chicago (2011)
11. Gu, C., Zhuang, S., Sun, Y.: Pricing incentive mechanism based on multistages traffic classification methodology for QoS-enabled networks. *J. Netw.* **6**(1), 163–171 (2011)
12. Bouch, A., Sasse, M.A.: Network quality of service—an integrated perspective. In: Proceedings of RTA's'99. Vancouver, 1–3 June 1999
13. Shakkottai, S., Srikant, R.: Network optimization and control. *Found. Trends Networking* **2**(3), 271–379 (2007)
14. Shakkottai, S., Srikant, R., Ozdaglar, A., Acemoglu, D.: The price of simplicity. *IEEE J. Sel. Areas Commun.* **26**(7), 1269–1276 (2008)
15. Eltarjaman, W., Ashibani, M., El-Jabu, B.: Towards optimized QoS based—charging model. In Southern African Telecommunication Networks and Applications Conference(SATNAC 2007). Sugar Beach Resort, Mauritius (2007)
16. Alderson, D., Willinger, W., Li, L., Doyle, J.: An Optimization-Based Approach to Modelling Internet Topology. *Telecommun. Plann. Innovations Pricing Netw. Des. Manage. Oper. Res./Comput. Sci. Interfaces Ser.* **33**, 101–136 (2006)
17. Garcell, M.A.G., Delgado, L.Y.M., Torres, L.A., Isaac, A.C.: Identifying and solving optimization problems on internet. *Вестник ТГТУ. Том 14. № 2. Transactions TSTU*, pp. 392–404 (2008)

18. Wu, M., Xiong, X., Ying, J., Jin, C., Yu, C.: A web services composition model for QoS global optimization. In Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCCT '09). P. R. China, Huangshan, 26–28 Dec 2009
19. Wang, Q., Peha, J.M., Sirbu, M.A.: Optimal Pricing for integrated-services networks with guaranteed quality of service. In: Bailey, J., McKnight, L. (eds.) *Internet Economics*. MIT Press, Cambridge (1996)
20. Marzolla, M., Mirandola, R.: *QoS Analysis for Web Service Applications: A Survey of Performance-Oriented Approaches from an Architectural Viewpoint*. Department of Computer Science University of Bologna, Bologna, Italy (2010)
21. Tektas, B., Kasap, N.: Time and volume based optimal pricing strategies for telecommunication networks. In: 17th International Conference on Management of Technology (IAMOT 2008). UAE, Dubai (2008)
22. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: The Improved formulation models of internet pricing scheme of multiple bottleneck link QoS Networks with various link capacity cases. In: Seminar Hasil Penyelidikan Sektor Pengajian Tinggi Kementerian Pendidikan Malaysia ke-3. Universiti Utara Malaysia (2013)
23. Puspita, F.M., Seman, K., Sanugi, B.: Internet Charging scheme under multiple QoS networks. In: The International Conference on Numerical Analysis & Optimization (ICeMATH 2011). Yogyakarta, Indonesia, Universita Ahmad dahlan, Yogyakarta, 6–8 June 2011
24. Puspita, F.M., Seman, K., Taib B.M.: A Comparison of optimization of charging scheme in multiple QoS networks. In: 1st AKEPT 1st Annual Young Reseachers International Conference and Exhibition (AYRC X3 2011) Beyond 2020: Today's Young Reseachers Tomorrow's Leader 19–20 DECEMBER 2011. 2011. PWTC, Kuala Lumpur (2011)
25. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Models of Internet charging scheme under multiple QoS networks. In: International Conferences on Mathematical Sciences and Computer Engineering 29–30 November 2012. Kuala Lumpur, Malaysia (2012)
26. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Improved Models of internet charging scheme of single bottleneck link in multi QoS networks. *J. Appl. Sci.* **13**(4), 572–579 (2013)
27. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Improved models of internet charging scheme of multi bottleneck links in multi QoS networks. *Aust. J. Basic Appl. Sci.* **7**(7), 928–937 (2013)
28. Kumar, S., Sehgal, R., Singh, P., Chaudhhary, A.: Nepenthes Honeypots based Botnet detection. *J. Adv. Inf. Technol.* **3**(4), 215–221 (2012)

Chapter 81

Improving the Models of Internet Charging in Single Link Multiple Class QoS Networks

Irmeilyana Saidi Ahmad, Indrawati, Fitri Maya Puspita
and Lisma Herdayana

Abstract In this paper, an improved internet charging scheme in multiple QoS networks will be discussed. The objective is to obtain better solution than previous results conducted by previous research. ISPs need a new charging scheme to maximize the revenue and provide better services to customers. The model is set up by fixing the fixed base price, varying the quality premium and fixing the sensitivity price for user in each class. The model is considered as Mixed Integer Nonlinear Programming (MINLP) and that can be solved by LINGO 11.0 to obtain the optimal solutions. We compare three cases of original, modified one and modified two models depending with the fixing or varying parameters or variables. The results show that by improving the pricing scheme model, the user' sensitivity price in modified two cases will yield maximum profit for ISPs.

81.1 Introduction

Previous works on pricing scheme of QoS networks is due to [1–4]. They described the pricing scheme based auction to allocate QoS and maximize ISP's revenue. The solution of the optimization problem goes from single bottleneck link in the network and then they generalized into multiple bottleneck links using

I.S. Ahmad · Indrawati · F.M. Puspita (✉) · L. Herdayana
Faculty of Mathematics and Natural Sciences, Universitas Sriwijaya,
Jln. Raya Palembang-Prabumulih, Indralaya, Ogan Ilir Sumatera Selatan 30662, Indonesia
e-mail: pipitmac140201@gmail.com

I.S. Ahmad
e-mail: imel_unsri@yahoo.co.id

Indrawati
e-mail: indrawati1006@gmail.com

L. Herdayana
e-mail: lisma_herdayana@ymail.com

heuristic method. In their study, they used single QoS parameter-bandwidth. In their discussion, they focus on auction algorithm to find the optimal solution. Based on their idea, it is attempted to improve their mathematical formulation and combine it with mathematical formulation discussed by Byun and Chatterjee [5] (see in [6–11]) to show that by improving the models, ISP improve the profit, with the advantages of availability of base price, quality premium and quality premium to be measured.

Recent studies have also been conducted to address problem of multiple service network, other kind of pricing scheme in network. Sain and Herpers [12] discussed problem of pricing in multiple service networks. They solve the internet pricing by transforming the model into optimization model and solved using Cplex software. Also, [13–15] discussed the new approach and new improved model of and got better results in getting profit maximization of ISP.

Although QoS mechanisms are available in some researches, there are few practical QoS network. Even recently a work in this QoS network proposed by [1–4], it only applies simple network involving one single route from source to destination.

So, the contribution is created by improving the mathematical formulation of to be simpler formulation in single link by taking into consideration the utility function, base price as fixed price or variable, quality premium as fixed prices and variable, index performance, capacity in one link, bandwidth required and also the user price sensitivity. The problem of internet charging scheme is considered as Mixed Integer Nonlinear Programming (MINLP) to obtain optimal solution by using LINGO 11.0 software. In this part, the comparison of two models is conducted in which whether decision variable is to be fixed of user admission to the class or not. This study focuses to fix the user's price sensitivity in each class. We consider cases of base price to be fixed and the quality premium to be fixed or vary depends on what target ISP would achieve. The Objective of ISP is to obtain maximum profit.

81.2 Research Method

The idea basically generates from [1–5] and are improved in single link multi class QoS networks. We attempt to improve the models when we consider the cases to fix the user price sensitivity in each class.

The steps are taken as follows.

1. Determine the parameters and decision variables for original and modified models.
2. Determine the constraints for the models.
3. Determine the model formulation of Steps 1 and 2.
4. Form the model formulation of base price and quality premium as the constant value and base price as the constant and quality premium as the variable.
5. Analyze the results and conclude the results.

81.3 Results and Discussion

81.3.1 Assumptions

Assume that there is only one single network from source to destination since concentrate on service pricing scheme. Assume that the routing schemes are already set up by the ISP. As [2] pointed out, we have 2 parts of utility function namely, base cost which does not depend on resource consumption and cost which depends on resource consumption. The parameters and decision variables we set up are presented in Tables 81.1 and 81.2.

81.3.2 Model Formulation

The model formulation follows from [10] except for \tilde{W}_{ij} and W_j we modify by varying or fixing the prices, for each case of original, modified and modified 1 with additional constraints if we set up \tilde{W}_{ij} and W_j as the parameters as follows.

$$\tilde{W}_{ij} = k, k \in R \quad (81.1)$$

$$W_j = 1 \quad (81.2)$$

81.3.3 The Solution for Original Model

In Table 81.3, the values of decision variables are given for original model. Final bandwidth (\hat{X}_{ij}) of user i is 1.234568. Minimum bandwidth for L_{M_1} and L_{M_2} is 1.234568 kbps. Sensitivity prices for class 1 and class 2, respectively (W_1 dan W_2) are 1.234568. We varies the base price 0.2/kbps and 0.3/kbps for all cases to promote the ISP goal to compete in market. Table 81.4 presents the solver status of the solver. Best objective is reached on value of 1.

Generated Memory Used (GMU) in Table 81.4 shows that the number of allocated memory used to run the solver. For original model, we have GMU of 28 K. Elapsed Runtime (ER) explains that the total time needed to solve the models and is affected by ether application running in the system. ER is 1 s in 4 iterations. We got total profit of only 1 unit price.

81.3.4 The Solutions for Modified Model with α and β Fixed

We modified the models into 3 groups when W_j and \tilde{W}_{ij} as parameters, W_j as and \tilde{W}_{ij} as variable and lastly when W_j as variable with \tilde{W}_{ij} as parameter. Table 81.5 explains the variable values for modified models. We obtain final bandwidth (\hat{X}_{ij})

Table 81.1 Parameters for each case of internet charging scheme

Parameter for original model	
Q	Total bandwidth
V_i	Minimum bandwidth needed by user i
α_j	Base price for class j
<i>Parameter for model modified 1 (α β constants)</i>	
α_j	Base price for class j
β_j	Premium quality having service performance I_j
Q	Total bandwidth
V_i	Minimum bandwidth needed by user i
c_j	Upper bound value for user i sensitivity price in class j
d_j	Upper bound for quality index in class j
<i>Parameter for model modified 2 (α constant, β variable)</i>	
Q	Total bandwidth
V_i	Minimum bandwidth needed by user i
α_j	Base price for class j
c_j	Upper bound value for user i sensitivity price in class j
d_j	Upper bound value for quality index in class j
f_i	Lower bound for premium quality in class j
g_i	Upper bound for premium quality in class j

for each user is 400.346 kbps. Premium quality for user 1 is 0.01 and for user 2 of 0.02. Minimum bandwidth L_{M_1} and L_{M_2} is 0.01 kbps. The price sensitivity for class 1 and 2 respectively (W_1 and W_2) are 13 when W_j as variable and \tilde{W}_{ij} as parameter. We obtain $W_1 = 10$ and $W_2 = 12$ when W_j and W_{ij} as parameters also W_j as variable and W_{ij} as parameter. It means that ISP can vary the base price of 0.2 unit price/kbps and 0.3 unit price/kbps, for all cases to compete in the market.

The highest GMU in this model is 29 K for each case as stated in Table 81.6. ER is 1 s when W_j and W_{ij} as parameter. Also, ER = 0 s for W_j as variable and \tilde{W}_{ij} as parameter, and last case when W_j as variable and W_{ij} as parameter. ESS is 0 since the solver applies the branch and bound solver. We can see that ISP can gain the maximum profit of 551.62 unit price if ISP sets W_j as variable and \tilde{W}_{ij} as parameter to enable ISP to recover cost.

81.3.5 The Solutions for Modified Model with α Fixed and β Vary

When we set up the modified model with α fixed and β vary, we group it into three categories namely when W_j and \tilde{W}_{ij} as parameter, W_j as parameter and \tilde{W}_{ij} as variable, dan lastly when W_j as variable and \tilde{W}_{ij} as variable. Table 81.7 shows the result of decision variables when we set up the modified model with α fixed and β

Table 81.2 Decision variables for each case of internet charging scheme

Variable for original model	$Z_{ij} : \begin{cases} 1, & \text{user } i \text{ is admitted to class } j \\ 0, & \text{otherwise} \end{cases}$
	$\tilde{X}_{ij} : \text{Final bandwidth obtained by user } i \text{ in class } j$
	$L_{Mj} : \text{Minimum bandwidth for class } j$
	$W_j : \text{Sensitivity price for class } j$
	$X_j : \text{Final bandwidth for class } j$
	$\tilde{W}_{ij} : \text{Sensitivity price for user } i \text{ in class } j$
Variable for modified model with α and β parameters	$Z_{ij} : \begin{cases} 1, & \text{user } i \text{ is admitted to class } j \\ 0, & \text{otherwise} \end{cases}$
	$\tilde{X}_{ij} : \text{Final bandwidth obtained by user } i \text{ in class } j$
	$L_{Mj} : \text{Minimum bandwidth for class } j$
	$W_j : \text{Sensitivity price for class } j$
	$X_j : \text{Final bandwidth for class } j$
	$\tilde{W}_{ij} : \text{Sensitivity price for user } i \text{ in class } j$
	$I_j : \text{Quality index of class } j$
Variable for modified model with α parameter and β variable	$Z_{ij} : \begin{cases} 1, & \text{user } i \text{ is admitted to class } j \\ 0, & \text{otherwise} \end{cases}$
	$\tilde{X}_{ij} : \text{Final bandwidth obtained by user } i \text{ in class } j$
	$L_{Mj} : \text{Minimum bandwidth for class } j$
	$W_j : \text{Sensitivity price for class } j$
	$X_j : \text{Final bandwidth for class } j$
	$\tilde{W}_{ij} : \text{Sensitivity price for user } i \text{ in class } j$
	$I_j : \text{Quality index of class } j$
	$\beta_j : \text{Premium quality of class } j \text{ having service performance of } I_j$

Table 81.3 Decision variable values for original model proposed by [2]

Decision variables values			
α_1	0.2	\hat{X}_{11}	1.234568
α_2	0.3	\hat{X}_{12}	1.234568
β_1	–	\hat{X}_{21}	1.234568
β_2	–	\hat{X}_{22}	1.234568
Z_{11}	1	L_{M1}	1.234568
Z_{12}	1	L_{M1}	1.234568
Z_{21}	1	X_1	1.234568
Z_{22}	1	X_2	1.234568
\tilde{W}_{11}	1.234568	I_1	–
\tilde{W}_{12}	1.234568	I_2	–
\tilde{W}_{21}	1.234568	W_1	1.234568
\tilde{W}_{22}	1.234568	W_2	1.234568

Table 81.4 Solver status of original model proposed by [2]

Solver status	Model class	INLP
	State	Local optimal
	Infeasibility	0
	Iterations	4
Extended solver state	Solver type	Branch and bound
	Active	0
	Update interval	2
	GMU (K)	28
	ER (sec)	1
	Best objective	1
	Objective bound	1
	ESS	0
	TSI	4

Table 81.5 Decision variable values for modified model with α and β fixed

Variable	Modified model with α and β fixed		
	W_j Par \tilde{W}_{ij} Par	W_j Par \tilde{W}_{ij} Var	W_j Var \tilde{W}_{ij} Par
α_1	0.2	0.2	0.2
α_2	0.3	0.3	0.3
β_1	0.01	0.01	0.01
β_2	0.02	0.02	0.02
Z_{11}	1	1	0
Z_{12}	1	1	0
Z_{21}	1	1	1
Z_{22}	1	1	1
W_1	10	10	13
W_2	12	12	13
\tilde{W}_{11}	12	10	12
\tilde{W}_{12}	12	12	12
\tilde{W}_{21}	15	10	15
\tilde{W}_{22}	15	12	15
\hat{X}_{11}	400.346	400.346	400.346
\hat{X}_{12}	400.346	400.346	400.346
\hat{X}_{21}	400.346	400.346	400.346
\hat{X}_{22}	400.346	400.346	400.346
L_{M_1}	0.01	0.01	0.01
L_{M_2}	0.01	0.01	0.01
X_1	400.346	400.346	400.346
X_2	400.346	400.346	400.346
I_1	0.9	0.9	0.9
I_2	0.8	0.8	0.8

Table 81.6 Solver status for modified model with α and β fixed

Solver status	W_j Par and \tilde{W}_{ij} Par	W_j Par \tilde{W}_{ij} Var	W_j Var \tilde{W}_{ij} Par
Model class	INLP		
State	Local optimal		
Infeasibility	7.38964e-012	7.38964e-012	0
Iterations	5	5	7
Solver type	Branch and bound		
Active	0	0	0
Update interval	2	2	2
GMU (K)	29	29	29
ER (sec)	1	0	0
Best objective	467.34	467.34	551.62
Objective bound	467.34	467.34	551.62
ESS	0	0	0
TSI	5	5	7

Table 81.7 Decision variables for modified model with α fixed and β vary

Var	Modified Model with α Fixed and β Vary		
	W_j Par \tilde{W}_{ij} Par	W_j Par \tilde{W}_{ij} Var	W_j Var \tilde{W}_{ij} Par
α_1	0.2	0.2	0.2
α_2	0.3	0.3	0.3
β_1	0.04	0.04	0.04
β_2	0.03	0.03	0.03
Z_{11}	1	1	0
Z_{12}	1	1	0
Z_{21}	1	1	1
Z_{22}	1	1	1
W_1	10	10	13
W_2	12	12	13
\tilde{W}_{11}	12	10	12
\tilde{W}_{12}	12	12	12
\tilde{W}_{21}	15	10	15
\tilde{W}_{22}	15	12	15
\hat{X}_{11}	400.346	400.346	400.346
\hat{X}_{12}	400.346	400.346	400.346
\hat{X}_{21}	400.346	400.346	400.346
\hat{X}_{22}	400.346	400.346	400.346
L_{M_1}	0.01	0,01	0.01
L_{M_2}	0.01	0,01	0.01
X_1	400.346	400,346	400.346
X_2	400.346	400,346	400.346
I_1	0.9	0.9	0.9
I_2	0.8	0.8	0.8

Table 81.8 Solver status of modified model with α fixed and β vary

Solver status	W_j Par \tilde{W}_{ij} Par	W_j Par \tilde{W}_{ij} Var	W_j Var \tilde{W}_{ij} Par
Model Class	INLP		
State	Local optimal		
Infeasibility	7.38964e-012	7.38964e-012	0
Iterations	5	5	7
<i>Extended solver state</i>			
Solver type	Branch and bound		
Active	0	0	0
Update interval	2	2	2
GMU (K)	29	30	29
ER (sec)	0	1	1
Best objective	467.41	467.41	551.69
Objective bound	467.41	467.41	551.69
ESS	0	0	0
TSI	5	5	7

vary. Final bandwidth (\hat{X}_{ij}) obtained by the users is 400.346 kbps. Premium quality for user 1 is 0.04 and for user 2 is 0.03. Minimum bandwidth for L_{M_1} and L_{M_2} is 0.01 kbps. Price sensitivity for class 1 and 2 (W_1 dan W_2) respectively is 13 when W_j as variable and \tilde{W}_{ij} as parameter while $W_1 = 10$ dan $W_2 = 12$ when W_j and W_{ij} as parameter and also when W_j as variable and W_{ij} as parameter. ISP enables to vary the base into 0.2/kbps dan 0.3/kbps to promote the available classes.

The highest GMU presented in Table 81.8 is 30 K when W_j as parameter and W_{ij} as variable, meanwhile GMU = 29 K when W_j and \tilde{W}_{ij} as parameter, also when W_j as variable and W_{ij} as parameter. ER is 0 s for the case when W_j and W_{ij} as parameter, while ER = 1 s for W_j as variable and \tilde{W}_{ij} as parameter, also W_j as variable and W_{ij} as parameter. ESS is 0 for all cases. ISP can obtain maximum profit of 551.69 unit price when ISP sets up the case when W_j as variable and \tilde{W}_{ij} as parameter to enable ISP to recover cost.

From Tables 81.4, 81.6 and 81.8 we can check that the best objective is achieved when ISP sets up either to fix the base price and quality premium to recover cost and to let user to choose the class; or to fix the base price and vary quality premium to recover cost and ISP can promote certain services by adding the condition to the models by setting up the sensitivity price for user j to be fixed and the sensitivity price for user i in class j to be varied.

81.4 Conclusion

From the above discussion, we can see that by considering the new parameters, decision variables and the constraints, we can obtain the better maximum profit. ISP can adopt either the model of modified by fixing α and β ; or fixing α and varying β for W_j as variable and fixing \tilde{W}_{ij} as parameter to attain maximum value of 323.78 bps for each file and web traffic data.

Acknowledgments The research leading to this study was financially supported by Directorate of Higher Education Indonesia (DIKTI) for support through Hibah Bersaing Tahun II, 2014.

References

1. Yang, W., Owen, H.L., Blough, D.M., Guan, Y.: An auction pricing strategy for differentiated service network. In: Proceedings of the IEEE Global Telecommunications Conference, IEEE (2003)
2. Yang, W.: Pricing network resources in differentiated service networks, in school of electrical and computer engineering. Ph.d. Thesis, Georgia Institute of Technology, pp. 1–111 (2004)
3. Yang, W., Owen, H., Blough, D.M.: A comparison of auction and flat pricing for differentiated service networks. In: Proceedings of the IEEE International Conference on Communications (2004)
4. Yang, W., Owen, H.L., Blough, D.M.: Determining differentiated services network pricing through auctions. In: Networking-ICN 2005, 4th International Conference on Networking April 2005 Proceedings, Part I. Reunion Island, France, Springer-Verlag Berlin Heidelberg (2005)
5. Byun, J., Chatterjee, S.: A strategic pricing for quality of service (QoS) network business. In: Proceedings of the Tenth Americas Conference on Information Systems, New York (2004)
6. Puspita, F.M., Seman, K., Sanugi, B.: Internet charging scheme under multiple QoS networks. In: The International Conference on Numerical Analysis & Optimization (ICeMATH 2011) 6–8 June 2011. Yogyakarta, Indonesia: Universita Ahmad dahlan, Yogyakarta (2011)
7. Puspita, F.M., Seman, K., Taib, B.M.: A comparison of optimization of charging scheme in multiple QoS networks. In: 1st AKEPT 1st Annual Young Researchers International Conference and Exhibition (AYRC X3 2011) Beyond 2020: Today's Young Reseacher Tomorrow's Leader 19–20 December 2011. PWTC, Kuala Lumpur (2011)
8. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Models of internet charging scheme under multiple QoS networks. In: International Conferences on Mathematical Sciences and Computer Engineering 29-30 November 2012. Kuala Lumpur, Malaysia (2012)
9. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: The improved formulation models of internet pricing scheme of multiple bottleneck link QoS networks with various link capacity cases. In: Seminar Hasil Penyelidikan Sektor Pengajian Tinggi Kementerian Pendidikan Malaysia ke-3 2013: Universiti Utara Malaysia
10. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Improved models of internet charging scheme of single bottleneck link in multi QoS networks. *J. Appl. Sci.* **13**(4), 572–579 (2013)
11. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: Improved models of internet charging scheme of multi bottleneck links in multi QoS networks. *Aust. J. Basic Appl. Sci.* **7**(7), 928–937 (2013)

12. Sain, S., Herpers, S.: Profit maximisation in multi service networks—An optimisation model. In: Proceedings of the 11th European Conference on Information Systems ECIS 2003. Naples, Italy (2003)
13. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: A new approach of optimization model on internet charging scheme in multi service networks. *Int. J. Sci. Technol.* **2**(6), 391–394 (2012)
14. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: An improved optimization model of internet charging scheme in multi service networks. *TELKOMNIKA* **10**(3), 592–598 (2012)
15. Puspita, F.M., Seman, K., Taib, B.M., Shafii, Z.: An improved model of internet pricing scheme of multi service network in multiple link QoS networks. In: The 2013 International Conference on Computer Science and Information Technology (CSIT-2013). Universitas Teknologi Yogyakarta (2013)

Chapter 82

A New Aggregating Phase for Interval Type-2 Fuzzy TOPSIS Using the ELECTRE I Method

Nurnadiah Zamri and Lazim Abdullah

Abstract Aggregating phase is considered as one of the important steps in interval type-2 fuzzy TOPSIS (IT2 FT) instead of ratings of alternatives under criteria and the importance weights of criteria and ranking of alternatives. However, some problems occur in aggregating phase of IT2 FT when it is have a large computational procedure due to the hardly defined in the second membership function. Therefore, we offer a more easier and practical in defining the new aggregating phase. Our proposed method is to establish a new aggregating phase for IT2 FT using the ELECTRE I method in the interval type-2 fuzzy set (IT2FS) concept. A numerical example is constructed to show the practicality and effectiveness of the proposed method.

82.1 Introduction

Interval type-2 fuzzy TOPSIS (IT2 FTOPSIS) is one of the multi-criteria decision making (MCDM) method and was first established by Chen and Lee [1] in the year of 2010. IT2 FTOPSIS is an extension from the original TOPSIS [2] and fuzzy TOPSIS [3]. IT2 FTOPSIS is believed can give more flexibility room due to the fact that it is uses interval type-2 fuzzy sets (IT2FSs) rather than type-1 fuzzy sets (T1FSs) to represent the uncertainties.

IT2 FTOPSIS can be divided into four main phases; rating phase, weighting phase, aggregating phase and ranking phase. Chen and Lee [1] introduced the concept on interval type-2 fuzzy sets [4] in the rating phase of IT2 FTOPSIS method.

N. Zamri (✉) · L. Abdullah

School of Informatics and Applied Mathematics, University Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia
e-mail: nadzlina@yahoo.co.uk

L. Abdullah

e-mail: lazim_m@umt.edu.my

The concept of IT2FSs [4] with subjective weight [2] were proposed in the weighting phase. The concept of ranking values was developed as a reduction method before the aggregating phase. Using the concept of positive-ideal solutions and negative-ideal solutions, optimal values of ranking was produced in the ranking phase. Unlike the IT2 FTOPSIS [1], our paper focuses on aggregating phase. We propose a new ELimination Et Choix Traduisant la REalite (ELimination and Choice Expressing the REality) (ELECTRE) with IT2FS concept in the aggregating phase.

Several papers have discussed on IT2 FTOPSIS' aggregating phase. For example, Chen et al. [5] proposed a new aggregating judgment method for ranking interval type-2 fuzzy sets. Besides, Chen [6] developed a new linear assignment method to produce an optimal preference ranking of the alternatives in accordance with a set of criterion-wise rankings and a set of criterion importance within the context of interval type-2 trapezoidal fuzzy numbers. Chen and Wang [7] established a new interval type-2 fuzzy TOPSIS (IT2 FT) with a new fuzzy ranking method based on the α -cuts of interval type-2 fuzzy sets. However, some problems occur in aggregating phase of IT2 FT when it is have a large computational procedure due to the hardly defined in the second membership function. Therefore, we propose a new ELECTRE I with IT2FS concept in the aggregating phase.

ELECTRE I was developed by Roy [8] as the first outranking method. Since that, it has been applied in various types of decision-making situations, including energy [9], environment management [10], project selection [11], and decision analysis [12]. Type-1 fuzzy ELECTRE was the extended method of ELECTRE. It can easily convert the linguistic preferences into the type-1 fuzzy numbers [13].

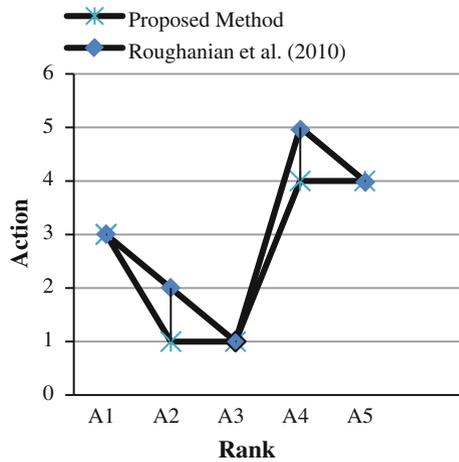
Various authors have applied type-1 fuzzy ELECTRE to the decision-making method. For example, Wu and Chen [14] used the ELECTRE method to solve the decision-making problems with intuitionistic fuzzy information. Botti and Peypoch [15] implemented an application of ELECTRE I in decision-making concept to tourism destinations. Hatami-Marbini and Tavana [16] proposed an alternative fuzzy outranking method by extending the ELECTRE I method to the Fuzzy TOPSIS (FT) method.

Our proposed work is to develop a new aggregating phase method for IT2 FTOPSIS [1] using the IT2 ELECTRE I. The rest of the paper is organized as follows. Section 82.2 reviews basic concepts related to ELECTRE I. In Sect. 82.3 we propose a new integrated ELECTRE I with IT2FS in IT2 FTOPSIS approaches. Section 82.4 illustrates the proposed method by applying it to a numerical example. Section 82.5 summarizes this study and presents future challenges.

82.2 Background of ELECTRE I

In this section, the basic decision of ELECTRE I [16], is briefly introduced. Then the proposed new ELECTRE I with IT2FS for IT2 FTOPSIS is presented in Sect. 82.3.

Fig. 82.1 The pictorial representation of the proposed method and the Roughanian et al. [17] rankings



Definition 1 Preference in ELECTRE I method is modelled by using binary outranking relations, S , whose meaning is “at least as good as”. Considering two actions x and y , four situations may arise:

- i. $x S y$ and not $y S x$, i.e., $x P y$ (x is strictly preferred to y),
- ii. $y S x$ and not $x S y$, i.e., $y P x$ (y is strictly preferred to x),
- iii. $x S y$ and $y S x$, i.e., $x I y$ (x is indifferent to y), and
- iv. not $x S y$ and not $y S x$, i.e., $x I y$ (x is incomparable to y).

Note that the incomparability preference is a useful relation to account for situations in which decision-makers (DMs) are not able to compare two actions.

Definition 2 According to ELECTRE I method, given two actions x and y , an outranking relation is based on two major concepts; the concordance and the discordance. The following statements provide insights into these concepts:

- The concordance concept: For an outranking $x S y$ to be validated, a sufficient majority of the criteria should be in favor of this assertion.
- The discordance concept: When the concordance condition holds, none of the criteria in the majority should oppose too strongly to the assertion $x S y$.

These two circumstances must be implemented for validating the assertion $x S y$.

We offer a new idea of aggregation phase for IT2 FTOPSIS by using the basic definition of ELECTRE I in Sect. 82.2. Therefore, the development of an IT2 ELECTRE I integrated IT2 FTOPSIS is shown in Sect. 82.3.

82.3 The Proposed Method

The proposed IT2 ELECTRE I integrated IT2 FTOPSIS starts with the constructed a decision matrix and weighting process using the IT2 linguistic scale as in Step 1 and Step 2. Then, the weighted decision matrix is constructed using the Step 1 and Step 2. Then the IT2 ELECTRE I is integrated with IT2 FTOPSIS as shown in Step 3 until Step 8. Full steps of the proposed method is shown as follows,

Assume that there is a set X of alternatives, where $X = \{x_1, x_2, \dots, x_n\}$, and assume that there is a set F attributes, where $F = \{f_1, f_2, \dots, f_m\}$. Assume that there are k decision-makers D_1, D_2, \dots , and D_k . The set F of attributes can be divided into two sets F_1 and F_2 , F_1 where denotes the set of benefit attributes, F_2 denotes the set of cost attributes. Below are the steps of IT2 ELECTRE I integrated IT2 FTOPSIS.

Step 1: Establish a decision matrix

Construct the design matrix Y_p of the p th decision-maker and construct the average decision matrix respectively.

Step 2: Calculate the weighting process

Construct the weighting matrix \tilde{W}_p of the attributes of the decision-maker and construct the p th average weighting matrix \tilde{W} .

Step 3: Construct the weighted DMs' matrix.

Construct the weighted decision matrix \tilde{V} .

Step 4 until Step 8 are the development of an ELECTRE I with IT2FS.

Step 4: Calculate the concordance matrix

Next, the concordance and discordance matrices are calculated using the weighted normalized IT2 fuzzy decision matrix (\tilde{V}) and the average decision matrix by DMs. Considering two actions A_g and A_j , the concordance set can be defined as $\tilde{J}_c = \{j | \tilde{v}_{gj} \geq \tilde{v}_{jj}\}$ where \tilde{J}_c is the index of all criteria belonging to the concordance coalition with the outranking relation $A_g S A_j$.

The discordance set can be defined as $\tilde{J}_D = \{j | \tilde{v}_{gj} \geq \tilde{v}_{jj}\}$ where \tilde{J}_D is the index of all criteria belonging to the discordance coalition and it is against the assertion “ A_g is at least as good as A_f .”

The concordance matrix is defined as follows,

$$B = \begin{bmatrix} - & \cdots & b_{1f} & \cdots & b_{1(m-1)} & b_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ b_{g1} & \cdots & b_{gf} & \cdots & b_{g(m-1)} & b_{gm} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ b_{m1} & \cdots & b_{mf} & \cdots & b_{m(m-1)} & - \end{bmatrix} \tag{82.1}$$

where

$$\tilde{c}_{gf} = \left(\tilde{c}_{gf}^l, \tilde{c}_{gf}^p, \tilde{c}_{gf}^q, \tilde{c}_{gf}^u \right) = \sum_{j \in J_c} \tilde{W}_j = \left(\sum_{j \in J_c} \tilde{w}_j^l, \sum_{j \in J_c} \tilde{w}_j^p, \sum_{j \in J_c} \tilde{w}_j^q, \sum_{j \in J_c} \tilde{w}_j^u \right)$$

In other words, the elements of concordance matrix are determined as the IT2 fuzzy summation of the IT2 fuzzy weights of all criteria in the concordance set.

Step 5: Calculate the discordance matrix

The discordance matrix is defined as

$$\tilde{D} = \begin{bmatrix} - & \cdots & \tilde{d}_{1f} & \cdots & \tilde{d}_{1(m-1)} & \tilde{d}_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{d}_{g1} & \cdots & \tilde{d}_{gf} & \cdots & \tilde{d}_{g(m-1)} & \tilde{d}_{gm} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{d}_{m1} & \cdots & \tilde{d}_{mf} & \cdots & \tilde{d}_{m(m-1)} & - \end{bmatrix} \tag{82.2}$$

where

$$\tilde{d}_{gf} = \frac{\max_{j \in J_D} |\tilde{v}_{gj} - \tilde{v}_{fj}|}{\max_j |\tilde{v}_{gj} - \tilde{v}_{fj}|} = \frac{\max_{j \in J_D} \tilde{d}(\max(\tilde{v}_{gj}, \tilde{v}_{fj}), \tilde{v}_{fj})}{\max_j \tilde{d}(\max(\tilde{v}_{gj}, \tilde{v}_{fj}), \tilde{v}_{fj})}$$

Note that there are prominent differences between the elements of \tilde{C} and \tilde{D} . The concordance matrix \tilde{C} reflects weights of the concordance criteria and the asymmetric discordance matrix \tilde{D} reflects most relative differences according to the discordance criteria. Both concordance and discordance indices have to be calculated for every pair of actions (\tilde{g}, \tilde{f}) , where $\tilde{g} \neq \tilde{f}$.

Step 6: Calculate the Boolean matrix for concordance

Now, we evaluate the value of the concordance matrix elements according to the concordance level. The concordance level, $\tilde{C} = (\tilde{c}_{gf}^l, \tilde{c}_{gf}^p, \tilde{c}_{gf}^q, \tilde{c}_{gf}^u)$, can be defined as the average of the elements in the concordance matrix, represented by $\tilde{c}^l = \sum_{f=1}^m \sum_{g=1}^m \tilde{c}_{gf}^l / m(m-1)$, $\tilde{c}^p = \sum_{f=1}^m \sum_{g=1}^m \tilde{c}_{gf}^p / m(m-1)$, $\tilde{c}^q = \sum_{f=1}^m \sum_{g=1}^m \tilde{c}_{gf}^q / m(m-1)$, and $\tilde{c}^u = \sum_{f=1}^m \sum_{g=1}^m \tilde{c}_{gf}^u / m(m-1)$. It is most desirable that the DMs achieve a consensus on the definition of the concordance level. If there is a disagreement among the DMs, then, the average value should be used for the definition.

Next, the Boolean matrix \tilde{B} is formed according to the minimum concordance level, \tilde{C} , as

$$\tilde{B} = \begin{bmatrix} - & \cdots & \tilde{b}_{1f} & \cdots & \tilde{b}_{1(m-1)} & \tilde{b}_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{b}_{g1} & \cdots & \tilde{b}_{gf} & \cdots & \tilde{b}_{g(m-1)} & \tilde{b}_{gm} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{b}_{m1} & \cdots & \tilde{b}_{mf} & \cdots & \tilde{b}_{m(m-1)} & - \end{bmatrix} \tag{82.3}$$

where

$$\begin{cases} \tilde{c}_{gf} \geq \tilde{C} \Leftrightarrow \tilde{b}_{gf} = 1 \\ \tilde{c}_{gf} < \tilde{C} \Leftrightarrow \tilde{b}_{gf} = 0 \end{cases}$$

Step 7: Calculate the Boolean matrix for discordance

Similarly, the elements of the discordance matrix are measured by a discordance level. The discordance level, $\tilde{D} = \sum_{f=1}^m \sum_{g=1}^m \tilde{d}_{gf} / m(m-1)$, can be defined as the average of the elements in discordance matrix. The Boolean matrix \tilde{H} is measured by a minimum discordance level as

$$\tilde{H} = \begin{bmatrix} - & \cdots & \tilde{h}_{1f} & \cdots & \tilde{h}_{1(m-1)} & \tilde{h}_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{h}_{g1} & \cdots & \tilde{h}_{gf} & \cdots & \tilde{h}_{g(m-1)} & \tilde{h}_{gm} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ \tilde{h}_{m1} & \cdots & \tilde{h}_{mf} & \cdots & \tilde{h}_{m(m-1)} & - \end{bmatrix} \tag{82.4}$$

where

$$\begin{cases} \tilde{d}_{gf} \geq \tilde{D} \Leftrightarrow \tilde{h}_{gf} = 1 \\ \tilde{d}_{gf} < \tilde{D} \Leftrightarrow \tilde{h}_{gf} = 0 \end{cases}$$

The elements of this matrix measures the power of the discordant coalition, meaning that if its element value surpasses a given level, \tilde{D} , the assertion is no longer valid. Discordant coalition exerts no power whenever $\tilde{d}_{gf} < \tilde{D}$. In other words, the elements of matrix \tilde{H} with the value of 1 show the dominance relations among the actions.

Step 8: Construct the global matrix

Next, the global matrix \tilde{Z} is calculated by peer to peer multiplication of the elements of the matrices \tilde{B} and \tilde{H} as follows:

$$\tilde{Z} = \tilde{B} \otimes \tilde{H} \tag{82.5}$$

where each element (\tilde{z}_{gf}) of matrix \tilde{Z} is obtained as

$$\tilde{z}_{gf} = \tilde{b}_{gf} \tilde{f}_{gf} \tag{82.6}$$

The final step of this procedure consists of exploitation of the above outranking relation (matrix \tilde{Z}) in order to identify as small as possible a subset of actions, from which the best compromise action could be selected.

This section describes the flow of a new IT2 ELECTRE I method (Step 4 till Step 8) is implemented to the existed IT2 FTOPSIS (starts from Step 1). With an attempt to consider the IT2FS framework, it would anticipate that the IT2 ELECTRE I integrated IT2 FTOPSIS makes a more comprehensive look.

82.4 Numerical Illustration

This numerical illustration is adapted from Roghanian et al. [17] is used to illustrate the procedures and feasibility of the proposed IT2 ELECTRE I integrated IT2 FTOPSIS framework. A company desired to select a suitable material supplier to purchase the key components of new products. After preliminary screening, five candidates (A1; A2; A3; A4, and A5) remain for further evaluation. A committee of three decision makers: D1, D2 and D3, has been formed to select the most suitable supplier.

Table 82.1 shows the linguistic terms for the rating of alternatives with seven scales “Very Poor (VP)”, “Poor (P)”, “Medium Poor (MP)”, “Fair (F)”, “Medium Good (MG)”, “Good (G)” and “Very Good (VG)”.

Moreover, Table 82.2 shows the linguistic terms for the weights of attributes also with seven scales “Very Low (VL)”, “Low (L)”, “Medium Low (ML)”, “Medium (M)”, “Medium High (MH)”, “High (H)” and “Very High (VH)”.

Table 82.1 Linguistic terms for the ratings and their corresponding interval type-2 fuzzy sets [1]

Linguistic terms	Interval type-1 fuzzy sets
Very poor (VP)	((0, 0, 0, 1; 1,1), (0, 0, 0, 1; 1, 1))
Poor (P)	((0, 1, 1, 3; 1,1), (0, 1, 1, 3; 1,1))
Medium poor (MP)	((1, 3, 3, 5; 1,1), (1, 3, 3, 5; 1, 1))
Fair (F)	((3, 5, 5, 7; 1,1), (3, 5, 5, 7; 1, 1))
Medium good (MG)	((5, 7, 7, 9; 1,1), (5, 7, 7, 9; 1, 1))
Good (G)	((7, 9, 0.9, 1; 1,1), (7, 9, 9, 1; 1, 1))
Very good (VG)	((9, 1, 1, 1; 1,1), (9, 1, 1, 1; 1, 1))

Table 82.2 Linguistic terms of weights of the attributes and their corresponding interval type-2 fuzzy sets [1]

Linguistic terms	Interval type-1 fuzzy sets
Very low (VL)	((0, 0, 0, 0.1; 1, 1), (0, 0, 0, 0.1; 1, 1))
Low (L)	((0, 0.1, 0.1, 0.3; 1, 1), (0, 0.1, 0.1, 0.3; 1,1))
Medium low (ML)	((0.1, 0.3, 0.3, 0.5; 1, 1), (0.1, 0.3, 0.3, 0.5; 1, 1))
Medium (M)	((0.3, 0.5, 0.5, 0.7; 1, 1), (0.3, 0.5, 0.5, 0.7; 1, 1))
Medium high (MH)	((0.5, 0.7, 0.7, 0.9; 1, 1), (0.5, 0.7, 0.7, 0.9; 1, 1))
High (H)	((0.7, 0.9, 0.9, 1; 1, 1), (0.7, 0.9, 0.9, 1; 1, 1))
Very high (VH)	((0.9, 1, 1, 1; 1, 1), (0.9, 1, 1, 1; 1, 1))

Table 82.3 The results comparison

Rank	Proposed method	Roghanian et al. [17]
A1	3	3
A2	1	2
A3	1	1
A4	4	5
A5	4	4

Using the linguistic scales from Tables 82.1 and 82.2, and the eight steps of the proposed method (in Sect. 82.3), result for Roghanian et al. example is shown in Table 82.3, as follows:

Result for Roghanian et al. example is graphed in pictorial representation a shown in Fig. 82.1, as follows:

Based on Roughanian et al. example, the derived Table 82.3 and graphed pictorial representation (Fig. 82.1) shows that there is a similarity between our proposed method and the results provided by Rouhghanian et al. [17]. Here, A2 and A3 are categorized in the first rank. Next, A1 is categorized in the second rank. The last prioritization belongs to A4 and A5.

82.5 Conclusions and Future Research

In this study, a new aggregating phase for IT2 FTOPSIS was developed by using the ELECTRE 1 method in terms of IT2FS. This proposed method is seen to ease the computational volume due to its ability to take ordinal scales into account without converting the original scales into the abstract ones by using the concordance and discordance steps. Besides, it is at the same time maintains the original verbal meaning of the decision problems [18]. A numerical example was used in order to test the efficiency of the proposed approach. In summary, the proposed approach produced some similar results with Roughanian et al. [17]. The efficiency of using new method was proven with a straight forward computation in

the case study. This approach is seen to provide a new perspective in aggregating phase of IT2 FTOPSIS. Moreover, they offer a practical, effective and low risk computation to produce a comprehensive judgment. A stream of future research can extend our algorithms to other variations of the ELECTRE methods such as ELECTRE II, III, IV, ELECTRE IS and ELECTRE TRI.

Acknowledgments This research is supported by MyBrain15 scholarship and Fundamental Research Grant, no. 59243. This support is gratefully acknowledged.

References

1. Chen, S.-M., Lee, L.-W.: Fuzzy multiple attributes group decision-making based on the interval type-2 TOPSIS method. *J. Expert Syst. Appl.* **37**, 2790–2798 (2010)
2. Chen, C.T.: Extension of the TOPSIS for group decision making under fuzzy environment. *Fuzzy Sets Syst.* **114**, 1–9 (2000)
3. Hwang, C.L., Yoon, K.: Manufacturing plant location analysis by multiple attribute decision making: single-plant strategy. *Int. J. Prod. Res.* **23**, 345–359 (1985)
4. Mendel, J.M., John, R.I., Liu, F.L.: Interval type-2 fuzzy logical systems made simple. *IEEE Trans. Fuzzy Syst.* **14**, 808–821 (2006)
5. Chen, S.-M., Yang, M.-W., Lee, L.-W., Yang, S.-W.: Fuzzy multiple attributes group decision-making based on ranking interval type-2 fuzzy sets. *Expert Syst. Appl.* **39**, 5295–5308 (2012)
6. Chen, T.-Y.: A linear assignment method for multiple-criteria decision analysis with interval type-2 fuzzy sets. *Appl. Soft Comput.* **13**, 2735–2748 (2013)
7. Chen, S.-M., Wang, C.-Y.: Fuzzy decision making systems based on interval type-2 fuzzy sets. *Inf. Sci.* **242**, 1–21 (2013)
8. Roy, B.: Classement et Choix en Présence de Points de vue Multiples (Laméthode ELECTRE). *RIRO*. **8**, 57–75 (1968)
9. Cavallaro, F.: A comparative assessment of thin-film photovoltaic production processes using the ELECTRE III method. *Energy Policy* **38**, 463–474 (2010)
10. Hokkanen, J., Salminen, P., Ettala, M.: The choice of a solid waste management system using the electre II decision-aid method. *Waste Manage. Res.* **13**, 175–193 (1995)
11. Figueira, J., Mousseau, V., Roy, B.: Electre methods. In: Figueira, J., Greco, S., Ehrgott, M. (eds.) *Multiple Criteria Decision Analysis. State of the Art Surveys*, pp. 133–162. Springer, New York (2005)
12. Karagiannidis, A., Perkoulidis, G.: A multi-criteria ranking of different technologies for the anaerobic digestion for energy recovery of the organic fraction of municipal solid wastes. *Bioresour. Technol.* **100**, 2355–2360 (2009)
13. Kaya, T., Kahraman, C.: An integrated fuzzy AHP–ELECTRE methodology for environmental impact assessment. *Expert Syst. Appl.* **38**, 8553–8562 (2011)
14. Wu, M.-C., Chen, T.-Y.: The ELECTRE multicriteria analysis approach based on Atanassov's intuitionistic fuzzy sets. *Expert Syst. Appl.* **38**, 12318–12327 (2011)
15. Botti, L., Peypoch, N.: Multi-criteria ELECTRE method and destination competitiveness. *Tourism Manage. Perspect.* **6**, 108–113 (2013)
16. Hatami-Marbini, A., Tavana, M.: An extension of the electre I method for group decision-making under a fuzzy environment. *Omega* **39**, 373–386 (2011)
17. Roghanian, E., Rahimi, J., Ansari, A.: Comparison of first aggregation and last aggregation in fuzzy group TOPSIS. *Appl. Math. Model.* **34**, 754–3766 (2010)
18. Greco, S., Matarazzo, B., Sowiński, R.: Rough sets theory for multi criteria decision analysis. *Eur. J. Oper. Res.* **129**, 1–47 (2001)

Chapter 83

The Role of Green IT and IT for Green Within Green Supply Chain Management: A Preliminary Finding from ISO14001 Companies in Malaysia

K.S. Savita, P.D.D. Dominic and Kalai Anand Ratnam

Abstract The purpose of this study is to investigate the role of Information Technology and Information System for Environmental Sustainability in Green Supply Chain Management within the context of manufacturing firms in Malaysia. More specifically, the pilot findings reveal that Green IS related activities among manufacturing firms in Malaysia are present, but measuring the extent of overall effectiveness of Green IS practices is not being carried out adequately. The data collection instrument used for the pilot study was distributed to the management team of 50 organizations and only 32 responses were received. With a response rate of 64 % for the pilot study, the results of the preliminary analysis indicates that manufacturing organizations in Malaysia are practicing green in Malaysia, but have yet to track and audit the exact energy efficiency level.

83.1 Introduction

Organizations today are obliged to respond to an increasing rate of change since product and technology life cycles are getting shorter with product differentiations [1]. Countries that are vigorously developing such as Asia is the main contributor to environmental issues. These developing countries are negatively impacting the

K.S. Savita (✉) · P.D.D. Dominic
Computer and Information Sciences Department, Universiti Teknologi Petronas,
31750, Ipoh, Perak, Malaysia
e-mail: savitasugathan@petronas.com.my

P.D.D. Dominic
e-mail: dhanapal_d@petronas.com.my

K.A. Ratnam
Asia Pacific University of Technology and Innovation (APU), Technology Park Malaysia,
Bukit Jalil, Kuala Lumpur 57000, Malaysia
e-mail: anand@apu.edu.my

ecological sustainability as compared to the developed countries [2]. The increased growth of CO₂ and Greenhouse Gas (GHG) emissions are significantly impacting the global climate change. This phenomena is in response to higher rates of economic growth and growing urban populations [3]. Manufacturing sectors is the key for the modernization of a country that differentiate the developed countries from the developing one [4]. The phenomena of globalization, industrialization and urbanization have associated impacts to worldwide ecological crisis. Nevertheless, companies are recognizing that environmental management is a key strategic movement that has a strong influence in enhancing organizational performance and gaining competitive edge in the industry [5].

Being a country that has emerged with multi-sector economy, Malaysia offers a cost-competitive location to manufacture advanced technological products for both regional and international markets. In 2011, The World Competitiveness Yearbook 2011 Report released by the Institute for Management Development (IMD) continued to rank Malaysia as among the top five most competitive nations in the Asia-Pacific region [6]. Manufacturing sector still remains a significant contributor to the growth of the Malaysia's economy [7]. Malaysia is making considerable progression in environmental management compared to other countries [6]. Despite that, Malaysia is facing growing challenge on energy consumption, air pollution from industrial emissions and poor waste management that leads to Greenhouse Gas (GHG) emissions that significantly impacting Malaysia's ecology [6, 8].

The impacts to the environment is largely through a product life cycle which is the primary source of today's environmental issues [2]. Manufacturing companies need an effective shift in its environmental management from end-of-the-pipe control and treatment of waste [9] to clean up and control at every stage of a product life cycle [10]. The supply chains of many modern-day firms cut across country boundaries, whereby the production and upstream of many supply chains exists in emerging and developing countries but the downstream may be located in developed countries [3]. Both the environmental management and supply chain management has its own root that complement each and must not be disregarded [11]. Therefore, adding the 'green' component to supply chain management focuses on activities that aims at minimizing ecological impacts of a product throughout its entire life cycle, namely environmental pollutions (air, water and land), waste of resources (energy, materials and products) as well as final use and disposal of products [5].

The implementation of Information System (IS) and Information Technology (IT) in Supply Chain Management (SCM) is becoming apparently important in an increasingly globalized and competitive economy. The advancement in the technology with high degree of automation in business processes is offering more opportunities than previously. IT as an infrastructure and a solution plays a critical factor in the improvement of SCM in within, downstream and upstream of the organization [12]. IT has leads to better performance in the supply chain with its power to provide timely, accurate, and reliable information, in which enabled real-time integration of SCM activities [12]. IT facilitates SCM by improving

integration and coordination of physical flow as well as the various information flow in the supply chain [13]. These highlight the importance of IT in functioning supply chain. The use of IT is considered as a prerequisite for an effective control of today's complex supply chain [14]. It is concluded that "IT is not an actual source of competitiveness but a source of competitive necessity", in which explained as IT implementation has become a necessity, not a choice [12]. The terms IS and IT are often used interchangeably, but both IT and IS contribute to the ecological differently and contrarily to an extent [15]. Therefore, in line with prior literature, this research differentiates "IT" and "IS" as two key elements of 'environmental sustainability in driving green practices in SCM.

The existing studies are lacking in assessing the role IT and IS in supporting organization's actions towards environmental and carbon compliance [16]. This is due to the general lack of urgency of the impacts of IT and IS have on environmental footprints [16]. Therefore, IT and IS are often not given the right ownership and ignored by organizations in their assessment on organization's environmental footprints [17]. Although, environmental conscious practices play a bigger role towards worldwide ecological sustainability, yet the role of IT and IS are disregarded in such undertakings, and much still remains unexplored [15, 18, 19]. Many of the research conducted in Malaysia are focusing either on green practices in SCM or green practices in IT and IS, and very few looks at integrating and incorporating both fields towards environmental and organizational sustainability. To the best of our knowledge, the research on IT and IS for environmental sustainability in supply chain in Malaysia is scant. Therefore, this study explores the role of IT and IS for environmental sustainability in driving the green practices in supply chain management within ISO 14001 manufacturing firms in Malaysia, and a preliminary finding from pilot study is presented.

83.2 Literature Review and Hypothesis Development

83.2.1 Green Supply Chain Management (GSCM)

In the early days, companies use typical supply chain to achieve economy of scale, efficiency, lower operating costs and mass produced products with very little concern on environmental consequence and ecological impacts [20, 21]. GSCM evolved significantly in the recent years in responding to the growing concern of environmental sustainability and compliance. GSCM covers activities such as 'green design', 'green sourcing/procurement', 'green operations' or 'green manufacturing', 'green distribution', 'logistics/marketing' and 'reverse logistics' [5]. The GSCM is categorized into inbound logistics, production or the internal supply chain, outbound logistics and reverse logistics that cover all supply chain activities, from green purchasing to the integration of life cycle management, through the manufacturer and customer, to closing the loop with reverse logistics [22]. Similarly, GSCM practices reflect the whole system of upstream, within, downstream

and transformation that focus on environmental operations management, suppliers collaboration (assessments and education), green procurement and logistics [23]. Other than that, GSCM practices covers internal environmental management, green purchasing, cooperation with customers, eco design and investment recovery [11]. The previous research that investigated on GSCM implementation in South East Region (Philippines, Indonesia, Malaysia, Thailand and Singapore), found that green practices exist but still at infancy stage [22]. Looking at Malaysia context, the fully owned Malaysian companies have a lower level of adoption of GSCM as compared to foreign based companies and Multinational Company (MNC) [24]. Also, [25] discovered that manufacturing industry in Malaysia has shown a strong awareness of environmental issues. However, very few companies were actually involved in the implementation plans, whereby, the larger firms, mostly multinational companies, are implementing environmental initiatives, although still limited and at early stage. This indicate the gaps to investigate in-depth the adoption of GSCM in developing countries, like Malaysia since the current findings is inconclusive as different sectors in different countries may be facing different pressure at various level and mode of implementation of GSCM practices [22, 26]. This further support [11] belief that, “study on Green SCM is timely and necessary to better aid the organization and have yet to be fully investigated”.

83.2.2 Green Information Technology

The impacts IT have on the environment are explained into direct or first-order effect known as Green IT and indirect and second-order effect refers to Green IS (IT for Green). Green IT is mainly focused on energy efficiency, carbon footprint and equipment utilization in contrast Green IS refers to the design and implementation of information systems that contribute to sustainable business processes [16, 27, 28]. Although, Green IT and Green IS are inter-related, they each have a different focus and purpose [29]. Green IT is conceptualized in various ways depending on its context and scope. Elliot and Binney [30] define Green IT in terms of design, production, operation and disposal of IT and non-IT-enabled products/services, which is not detrimental and beneficial to the environment during its entire life cycle. Murugesan [31] refers Green IT as environmentally sound IT, whereby it is a study and practice of designing, manufacturing, using and disposing of computers, servers and its associated subsystems with minimal or no impact on the environment (environmental sustainability), energy efficiency and total cost of ownership. Therefore, IT is considered as one of the key roles that transform supply chain management function in manufacturing organization, in which it has shorten the product life cycle [22]. Moreover, replacement of traditional technologies with IT will minimize the amount of resources used in manufacturing [28]. We therefore propose to contribute to the literature by exploring whether there are positive relationships Green IT in driving the Green Supply

Chain Management (GSCM) within ISO 14001 manufacturing firms in Malaysia. The following hypotheses are therefore proposed:

Hypotheses 1 (H 1) Green IT positively influences the implementation of green practices in SCM among ISO 14001 manufacturing firms in Malaysia

83.2.3 Green Information System (IS)

Recently, IS literature also began to realize the importance of sustainability, and proposed the concept of “Green IS” to better understand the role of IS in dealing with sustainability [18, 32]. Green IS focused not only on reducing the impact of IT but on the ways information systems can be used to help firms reduce their carbon footprint through automating and transforming products, business processes, business relationships and practices [19, 32]. Nowadays, companies increasingly integrate their business processes without realizing that Green IS initiatives are not be limited within companies, but extended to the externalities that need collaboration from the companies, partners, and even customers [33]. Green IS plays a significant role in making both business processes and the products they create environmentally sustainable [34]. Since, environmental sustainability is a supply chain commitment, therefore, Green IS are essential for creation, maintenance, and survival of environmental conscious practices in supply chain [35]. The information systems have the ability to enable interconnectedness, realign and reinvent business processes in support of productivity and efficiency towards ecological improvements [36]. Green IS encourages the application of information systems thinking and skills to initiatives across all functions of the organization, from logistics, to waste management, to communicating consumption information to customers [15]. Green IS is view as an enabler in inducing changes within and among business processes that decreases the environmental impacts through integration and coordination of IS throughout the SCM [34, 35]. These design and development of information systems represent the backbone of environmental management efforts that supports the firm’s environmental management systems [26]. As emphasized by [32], a much in-depth research is required to determine to what extent IS might improve sustainability in the realm of supply chains and logistics. Hence, the proposed hypothesis:

Hypothesis 2 (H 2) Green IS (IT for Green) positively influences the implementation of green practices in SCM among ISO 14001 manufacturing firms in Malaysia

83.3 Research Methodology

The chosen samples are from ISO 14001 manufacturing companies in Malaysia as they have been expected to have embarked on green practices in their operation. The companies are selected based on purposive sampling method. According to this method, the elements in the sample are selected for a specific purpose and in the best position to provide the information required by the researcher since the subjects has the expertise/knowledge on the topic being investigated [37]. The information on the companies is obtained from Federation of Malaysia Manufacturers Perak Directory of year 2012. The ISO 14001 certification detail is cross checked from SIRIM QAS website and company's website to ensure that the chosen companies are practicing green activities. For this preliminary research, the researchers' have chosen Ipoh and Kulim for the pilot testing in obtaining the initial findings. In both locations, the ISO 14001 manufacturing firms include of large enterprises and SMEs. The invitation to complete the questionnaire survey was send out using official email to the Human and Resources Department of the participating companies. The target respondents were company's senior management whom their scope of work includes managing the company's operation and production. After two weeks, 50 companies responded and agreed to participate in the survey. Subsequent emails were then sent to the identified personnel's which includes the link of the survey. The reason for adopting online survey is due to its potentially quicker response time with wider magnitude of coverage and higher cost savings compared to mail or direct survey [37]. A period of 2 months was given for the companies to respond to the survey. During that period, follow-up reminders via email and telephone calls are utilized in order to increase the participation rate of the survey. After 2 months, a total of 37 responses were received, whereby five of them were incomplete. For the pilot study, the final total of 32 questionnaires was used as the dataset to test the associated hypotheses.

83.4 Analysis

83.4.1 Profile of Respondents

The demographic profile of the participants is shown in Table 83.1, which indicates that the majority of the respondents were large firms, while the remainder were Small and Medium Enterprises (SMEs). The majority of the companies that participated in the survey were from Electrical and Electronics as well as Rubber and Plastics sectors. Most of the respondents are attached to operation department, where most of them hold top management posts.

Table 83.1 Assessments of organizations

	Frequency	Percentage
<i>Position</i>		
Director	3	9.4
Manager	16	50.0
Executive	13	40.6
<i>Sector</i>		
Electrical and electronics	8	25.0
Chemicals and chemical Products	3	9.4
Rubber and plastic products	7	21.9
Basic metal and fabricated metal products	5	15.6
Machinery and equipment	4	12.5
Others	5	15.6
<i>Size</i>		
Small and medium Enterprise (SME)	9	28.1
Large enterprise	23	71.9

83.4.2 Data Analysis and Discussion

This paper provides a preliminary insight on the role of Green IT and Green IS (IT for Green) in driving environmental sustainability in Green Supply Chain Management (GSCM). According to [18], Green IS (IT for Green) has a bigger influence than Green IT since it tackles a larger scope in making the entire systems more sustainable. Therefore, this paper choose to categorize the effects of IT into direct (Green IT) and indirect (Green IS/IT for Green) in order to obtain conceptual and practical understanding in influencing the Green SCM implementation.

Figure 83.1, illustrates the extent to which factors of Green IT that influences the Green SCM implementation. The findings obtained are largely consistent with both hypotheses. The environmentally friendly disposal of electronic wastes carried the highest significant percentage (84.4 %), and followed by procurement of IT hardware and equipment (56.3 %), print optimization (46.9 %), PC power management (34.4 %) and rightsizing IT equipment (28.1 %). The results obtained is consistent with study conducted among 143 organizations from Australia, New Zealand and the USA that indicated that disposal of IT in an environmentally friendly manner is the most adopted Green IT practice [38]. In addition, manufacturing companies in Malaysia is starting to recognize the importance on improving energy efficiency since use of virtualization technology is gaining its popularity. However, in data center, the design and use of energy efficient IT infrastructures are still lacking because most the large data center are based in the

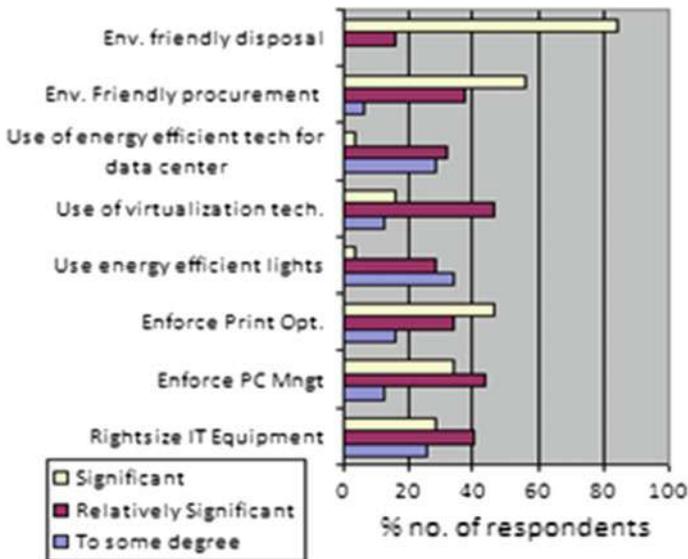


Fig. 83.1 To what extent do the following factors influence your firm in implementing environmentally conscious practices in supply chain management?

headquarters which are located in other countries. The use of energy efficient lights and other technologies in production floor is only considers among large enterprises, due to its rising cost and intangible returns. This is because, most these companies fails to measure the amount of the IT and its energy usage contribute in the electrical bills as well as actual cost savings from implementing energy efficient practices and technologies. The similar situation is being faced by 54 % of respondents (60 % Australian, 50 % New Zealand and 38 % US) that were unaware on the actual cost and savings from implementing Green IT [38]. As mentioned by [39], exact energy consumption and measurement of desktops, monitors, notebooks, networks, printers and communication equipment can only be estimated, and still remain a huge challenge for many organizations. Therefore, these preliminary results concluded that technologies that improve the energy efficiency of IT and ecological footprint from IT are not as widely adopted as expected among these ISO 14001 manufacturing companies.

As presented in Fig. 83.2, the highest scored Green IS activities is 81.3 % which is the application of IS in enhancing green practices within downstream, production and upstream of supply chain in integrating and coordinating physical and information flows by means of automation of business workflows and applications development. These enable the internal customers to get access to information anytime anywhere in a paperless environment followed by the use of video conferencing/telecommuting (68.8 %) and use of online groupware/collaboration tools (59.4 %) that facilitate more effective internal communication, apart from

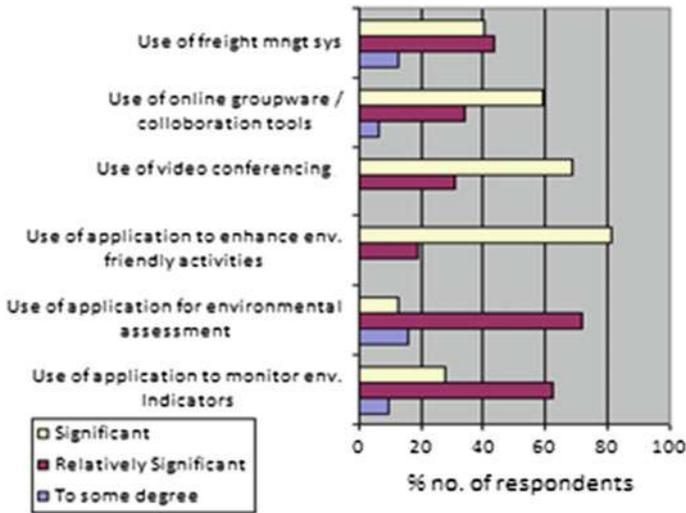


Fig. 83.2 To what extent do the following factors influence your firm in implementing environmentally conscious practices in supply chain management?

promote communication and collaboration seamlessly without having the need to travel, thus reducing carbon emissions enterprise-wide.

Green IS provide the information needed to coordinate within, and extended relationships with customers and suppliers that are necessary to make decisions about eco-design, production, packaging, transportation, reuse, recycling and recovery of materials [35]. This findings are consistent with [15] that categorized green IS practice into (1) use of systems for pollution reduction in business process; (2) use of system to enhance environmental friendliness at each stage supply chain management and (3) use of system for learning and innovation, such as knowledge management systems for sustainable business operations transformation. This trends suggested that the companies have been, using the applications from category 1 and 2 since few years ago without realizing is part of environmentally-friendly practice that are ecologically driven. However, the use of software to monitor, assess and report environmental indicators such as GHG emissions, waste, toxic and hazardous materials use is less significant to most of the companies. The lacks of its importance explained that companies are still lacking of proper governance and evaluation on overall environmental situation contributed by information technologies and systems. Butler [40] explained that between 65–80 % of companies lack an integrated IS infrastructure to track, audit and manage issues around product and process compliance to environmental criteria, which also mentioned in [39].

For technological context, the respondents were enquired on Green IT which covers the aspect of core technology followed by IT For Green (Green IS) which covers the aspect of software and applications. Table 83.1 presents the mean

Table 83.2 Technological context

Dimension	Mean	Std. Dev.	Cronbach's Alpha
Green IT	3.81	0.785	0.812
IT for Green	4.39	0.452	0.887

values and standard deviation for both Green IT and IT for Green. Based on the 5-point Likert-type scale, the mean and standard deviation were 3.81 and 0.785 respective for Green IT, followed by IT for Green with 4.39 and 0.452 respectively. Both Mean and standard deviation were higher for IT for Green as compared to Green IT, this inevitably means that manufacturing firms in Malaysia are inclined towards paying attention towards the green aspect of software systems as compared to hardware aspect (Table 83.2).

83.5 Conclusion

Most of the previous studies on GSCM, the key drivers that influence its implementation are organizational and environmental factors, which covers top management supports, regulations, customer pressures, expected business benefits and social responsibility. Despite from these drivers, Green Information Technologies and Systems have huge potential for addressing broader environmental issues in organizations, including supply chains and logistics. Information Technologies and systems are embedded in supply chain business processes, in which the effects of greening will impact many areas and overall sustainability of a business is remain unclear. The existing Green IT and Green IS research are fragmented and very little evaluation on the potential of Green IT and Green IS as the enablers for successful implementation of green practices in supply chains. In this study, the researchers combine the both Green IT and IT for Green (Green IS) in investigating their influence in driving the implementation of Green SCM within ISO14001 manufacturing firms in Malaysia. The results presented are at the preliminary stage, whereby it indicates that Green IS has greater influences over Green IT. This could be understood from the context of investments needed for such initiatives versus the tangible returns. Nevertheless, the companies showed receptive response towards the implementation of Green IT and IT for Green in enabling the green practices in supply chain activities. As such, a full scale study is been carried to determine to what extent Green IT and Green IS might improve environmental and business sustainability in the realm of supply chain management in Malaysia context.

References

1. Azevedo, S.G., Carvalho, H., Cruz Machado, V.: The influence of green practices on supply chain performance: a case study approach. *Transp. Res. Part E Logist. Transp. Rev.* **47**(6), 850–871 (2011)
2. Markandya, A., Halsnaes, K.: Developing countries and climate change. In: Owen, A.D., Hanley, N. (eds.) *The Economics of Climate Change*, pp. 239–258. Routledge, Taylor & Francis Group, London and New York (2004)
3. Gholami, R., Sulaiman, A.B., Ramayah, T., Molla, A.: Senior managers' perception on green information systems (IS) adoption and environmental performance: results from a field survey. *Inf. Manag.* **50**(7), 431–438 (2013)
4. El-Khasawneh, B.S.: Challenges and remedies of manufacturing enterprises in developing countries: Jordan as a case study. *J. Manuf. Technol. Manag.* **23**(3), 328–350 (2012)
5. Srivastava, S.K.: Green supply-chain management: a state-of-the-art literature review. *Int. J. Manag. Rev.* **9**(1), 53–80 (2007)
6. MGCC: 'Market watch 2012' the environmental sector in Malaysia (2012)
7. MIDA: Malaysia investment performance 2012: investment for transformation (2012)
8. Abidin, Z.Z., Jelani, N.: The impact of climate change and the Need for official statistics in Malaysia, pp. 1–19. 2011
9. Handfield, R., Sroufe, R., Walton, S.: Integrating environmental management and supply chain strategies. *Bus. Strateg. Environ.* **14**(1), 1–19 (2005)
10. Matos, S., Hall, J.: Integrating sustainable development in the supply chain: the case of life cycle assessment in oil and gas and agricultural biotechnology. *J. Oper. Manag.* **25**(6), 1083–1102 (2007)
11. Zhu, Q., Sarkis, J., Lai, K.: Confirmation of a measurement model for green supply chain management practices implementation. *Int. J. Prod. Econ.* **111**(2), 261–273 (2008)
12. Li, G., Yang, H., Sun, L., Sohal, A.S.: The impact of IT implementation on supply chain integration and performance. *Int. J. Prod. Econ.* **120**(1), 125–138 (2009)
13. Omar, R., Ramayah, T., May-Chuin, L., Sang, T.Y., Siron, R.: Information sharing, information quality and usage of information technology (IT) tools in Malaysian organizations. *Afr. J. Bus. Manag.* **4**(12), 2486–2499 (2010)
14. Fasanghari, M., Roudsari, F.H., Chaharsooghi, S.K.: Assessing the impact of information technology on supply chain management. *World Appl. Sci. J.* **4**(1), 87–93 (2008)
15. Chen, A.J., Watson, R.T., Boudreau, M.-C., Karahanna, E.: An institutional perspective on the adoption of green IS. *Aust. J. Inf. Syst.* **17**(1), 23–45 (2010)
16. Jenkin, T.A., Webster, J., McShane, L.: An agenda for 'Green' information technology and systems research. *Inf. Organ.* **21**(1), 17–40 (2011)
17. Huang, A.H.: A model for environmentally sustainable information systems development. In: *Pacific Asia Conference on Information Systems (PACIS)*, 2008
18. Watson, R., Boudreau, M., Chen, A., Huber, M.: Green IS: building sustainable business practices. In: Watson, R.T. (ed.) *Information Systems*, pp. 1–15. Global Text Project, Athens, GA (2008)
19. Chen, A.J.W., Boudreau, M.-C., Watson, R.T.: Information systems and ecological sustainability. *J. Syst. Inf. Technol.* **10**(3), 186–201 (2008)
20. Sarkis, J., Zhu, Q., Lai, K.: An organizational theoretic review of green supply chain management literature. *Int. J. Prod. Econ.* **130**(1), 1–15 (2011)
21. Nelson, D.M., Marsillac, E., Rao, S.S.: Antecedents and Evolution of the Green Supply Chain. *J. Oper. Supply Chain Manag.* **5**(Special Issue on Sustainability), 29–43 (2012)
22. Rao, P., Holt, D.: Do green supply chains lead to competitiveness and economic performance? *Int. J. Oper. Prod. Manag.* **25**(9), 898–916 (2005)
23. Holt, D., Ghobadian, A.: An empirical study of green supply chain management practices amongst UK manufacturers. *J. Manuf. Technol. Manag.* **20**(7), 933–956 (2009)

24. Eltayeb, T.K., Zailani, S.: Going green through green supply chain initiatives towards environmental sustainability. *Oper. Supply Manag.* **2**(2), 93–110 (2009)
25. Taha, Z., Sakundarini, N., Ariffin, R., Ghazila, R., Gonzales, J.: Eco design in Malaysian industries: challenges and opportunities. *J. Appl. Sci.* **6**(12):2143–2150 (2010)
26. Cai, S., Chen, X., Bose, I.: Exploring the role of IT for environmental sustainability in China: an empirical analysis. *Int. J. Prod. Econ.* **146**(2), 491–500 (2013)
27. Dedrick, J.: Green IS: concepts and issues for information systems research. *Commun. Assoc. Inf. Syst.* **27**(1), 173–184 (2010)
28. Faucheux, S., Nicolai, I.: IT for green and green IT: a proposed typology of eco-innovation. *Ecol. Econ.* **70**(11), 2020–2027 (2011)
29. Molla, A.: Identifying IT sustainability performance drivers: instrument development and validation. *Inf. Syst. Front.* **15**(5):705–723 (2013)
30. Elliot, S., Binney, D.: Environmentally sustainable ICT: developing corporate capabilities and an industry-relevant research agenda. In: *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*, Suzhou, China (2008)
31. Murugesan, San: Harnessing green IT: principles and practices. *IT Prof.* **10**(1), 24–33 (2008)
32. Melville, N.P.: Information systems innovation for environmental sustainability. *MIS Quarterly* **34**(1), 1–21 (2010)
33. Brooks, S., Wang, X., Sarker, S.: Unpacking green IS: a review of the existing literature and directions for the future. In: vom Brocke, J., Seidel, S., Recker, J. (eds.) *Green Business Process Management Towards the Sustainable Enterprise*, pp. 15–37. Springer, Berlin (2012)
34. Boudreau, M.C., Watson, R.T., Chen, A.: From green IT to green IS. In: Biros, B. et al. (ed.) *The Organizational Benefits of Green IT*, pp. 79–91. Cutter Information LLC, Arlington, MA (2008)
35. Green, K.W., Zelbst, P.J., Meacham, J., Bhaduria, V.S.: Green supply chain management practices: impact on performance. *Supply Chain Manag. An Int. J.* **17**(3):290–305 (2012)
36. Dwyer, C., Hasan, H.: Emergent solutions for global climate change: lessons from green IS research. *Int. J. Soc. Organ. Dyn. IT* **2**(2), 18–33 (2012)
37. Kumar, M., Salim, A.T., Ramayah, T.: *Business Research Methods*, pp. 1–417. Oxford Fajar, Oxford University Press (2013)
38. Molla, A., Pittayachawan, S., Corbitt, B.: Green IT diffusion: an international comparison. *Green IT Working Paper Series*, no. 1. pp. 1–15, 2009
39. Ereik, K.: From green IT to sustainable information systems management: managing and measuring sustainability in IT organisations. In *European, Mediterranean & Middle Eastern Conference on Information Systems*, vol. 2011, pp. 766–781 (2011)
40. Butler, T.: Compliance with institutional imperatives on environmental sustainability : building theory on the role of Green IS. *J. Strateg. Inf. Syst.* **20**(1), 6–26 (2011)
41. Zhang, X., Van Donk, D.P., Van Der Vaart, T.: Does ICT influence supply chain management and performance?: a review of survey-based research. *Int. J. Oper. Prod. Manag.* **31**(11), 1215–1247 (2011)

Chapter 84

Integrating e-Learning with Radio Frequency Identification (RFID) for Learning Disabilities: A Preliminary Study

**Wan Fatin Fatihah Yahya, Noor Maizura Mohamad Noor,
Mohd Pouzi Hamzah, Mohamad Nor Hassan,
Nur Fadila Akma Mamat and Mohd Arizal Shamsil Mat Rifin**

Abstract Integrating learning styles in adaptive educational systems are a growing trend in technology enhanced learning. Children have different learning styles, abilities, preferences that focus on different types of information and process new information in different ways. Providing adaptively based on learning styles can promote interest for learners and make learning easier for them. The purpose of our research is to adopt an e-Learning approach Radio Frequency Identification (RFID) technology in order to model the Visual, Auditory and Kinesthetic (VAK) learning style focused on Learning Disabilities (LD) children. Today's technology offers great chances to assist students with disabilities to live freely and learn more easily. Developing the learning environments assisted by technology is a new way in making their learning processes successful.

W.F.F. Yahya (✉) · N.M.M. Noor · M.P. Hamzah · M.N. Hassan · N.F.A. Mamat · M.A.S.M. Rifin
School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu,
21030 Kuala Terengganu, Terengganu, Malaysia
e-mail: wanfatinyahya@gmail.com

N.M.M. Noor
e-mail: maizura@umt.edu.my

M.P. Hamzah
e-mail: mph@umt.edu.my

M.N. Hassan
e-mail: mohamadnor@umt.edu.my

N.F.A. Mamat
e-mail: nurfadilaakma@yahoo.com

M.A.S.M. Rifin
e-mail: arizalshamsil@gmail.com

Keywords e-Learning · Learning styles · Learning disabilities · Radio frequency identification

84.1 Introduction

Computers have been applied to develop electronic learning (e-Learning) systems since 1980. The e-Learning market has developed and grown worldwide as the new paradigm of modern education. e-Learning can be defined as an activity to sustainance a learning practice by either developing or applying information and communication technology (ICT) and also involves the use of a number of technological tools that can be applied in various contexts [1, 2]. e-Learning is one important opportunity for supporting greater access for all children. Children have different learning styles, abilities, preferences that focus on different types of information and process new information in different ways [3, 4]. The concept of e-Learning is to provide children with learning disabilities access to great opportunity to improve the quality of their learning process. These children are a special minority group in the society. Therefore, setting [3] down the needs of the students with a learning disability is an important issue to support and it gives great chances for them to live freely and learn more easily.

Using learning styles in e-Learning environment is quite a new trend in technology enhanced learning. Providing adaptively based on learning styles can sustainance learners and make learning easier for children with Learning Disabilities (LD). e-Learning environments taking into learning style are more efficient than traditional e-Learning environment [5]. There are some problems with e-Learning materials, they do not strictly fulfill the conventional multi-sensory approaches to help Learning Disability children [6]. The element of touch, for example, is usually not available. A technique that has shown potential in creating the element of touch is by leveraging the state of the art Radio Frequency Identification (RFID) technology.

84.2 Related Works

84.2.1 *Review Stage Conventional Learning Versus e-Learning*

Conventional learning to conventional teaching and learning contrasts with e-Learning [5]. There is a disagreement that conventional learning is the best way of sustaining a learning process. However, there is no finding to support this disagreement, and research shows that technology supported models are at least as good as conventional learning [1]. Conventional learning assumes students to learn

Table 84.1 Comparison between conventional learning and e-Learning

Conventional Learning	e-Learning
Assumes students to learn skills at the same rate	Provide personalized student need
Book based	Use technological tools in learning
Teacher centered	Student centered
Passive learning	Active learning
Single media	Multi-sensory and media/interactive
Time and location constraints	Time and location flexibility

skills at the same rate. However, one student from another may learn in different ways from listening, watching, questioning and doing [1]. While both students have the ability to learn, conventional learning does not recognize this. Students who are quick learners often becoming uninterested or troublesome. Based on the result, it shows that both learners are impotent to meet their full potential. Table 84.1 show the comparison between conventional learning and e-Learning.

With the progress of ICT development, e-Learning is emerging as the paradigm of modern education. e-Learning means the delivery of information for education, learning or training program assisted by ICT [7, 8]. e-Learning involves the use of a number of technological tools that can be applied in various contexts [8]. Content is delivered via the Internet, intranet/extranet, audio or video tape, satellite TV, and CD-ROM. It can be self-paced or instructor led and includes media in the form of text, image, animation, streaming video and audio [1]. A common theme explored in e-Learning literature is on how technology plays the role in changing the learning paradigm, which resulted the process of acquiring knowledge becomes faster and more efficient [9].

Many studies have exposed that technology can play an important role in any work with specific disadvantaged groups [10–12]. Studies have also explored how information and communications technologies (ICT) can influence the education of students with LD and have shown that this technology can play an important and useful role in the learning process [10, 12, 13]. One particular problem with e-Learning materials is that it may contain multimedia elements, they do not strictly comply with conventional multi-sensory approaches to help learning disabilities. The element of touch, for example, is usually not available [12]. As an alternative, one of the technologies that have been successful bridging connections between the physical and virtual environment is RFID [7]. e-Learning can be viewed as a new style for delivering well designed, learner-centered, interactive, attractive, flexible, meaningful and facilitated learning environment [14]. One of the most best and complete comprehensive theoretical e-Learning models is Badrul e-Learning framework [15]. Figure 84.1 shows the Badrul e-Learning framework

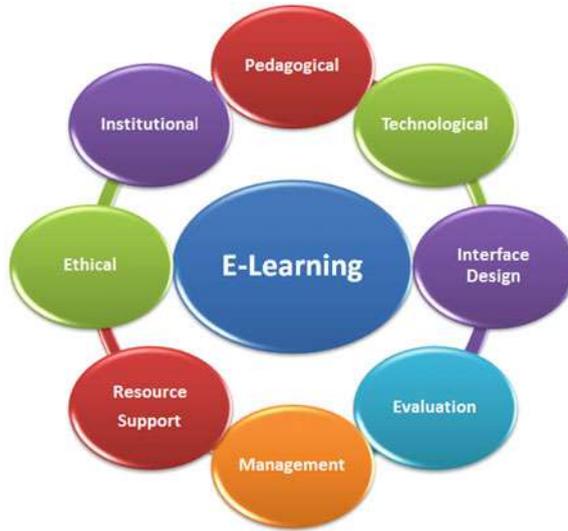


Fig. 84.1 Badrul e-Learning framework

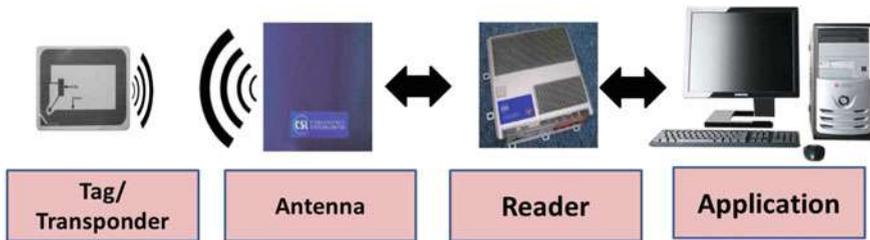
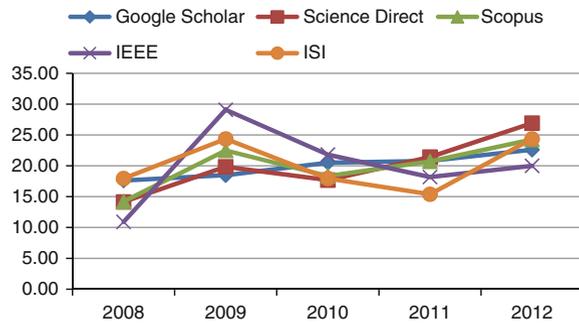


Fig. 84.2 RFID system component

84.2.2 Radio Frequency Identification

RFID describes systems that use radio waves to transmit an object’s identity. There are several methods of identifying objects by using RFID. The most common is store an ID or a serial number that identifies a specific product along with other information, on a tag, which is a small microchip that is attached to an antenna. The antenna enables the chip transmit whatever identification information it contains to a reader. The reader converts the radio waves from the RFID tag into digital information that software systems can use for processing [7]. Figure 84.2 shows the RFID system component. There are tag/transponder, antenna, reader and application.

Fig. 84.3 Growing number of radio frequency identification in learning between years 2008 and 2012



The emerging of the technology will revolutionize the development of learning process among the LD children. Information and Technology (ICT) and RFID, are among the examples that will instrument the learning revolution. We emphasize towards the active learning, where it gives the new experience to the LD children, to “actively” participate, simulate and interact with the tool and via-versa. The LD children will interact directly with the tool using more intuitive techniques of RFID that are attached to the learning items.

The term Learning Disability (LD) is used to refer to neurological disorders that affect one or more of the basic psychological development involved in the processes of speech, language, reading, writing, arithmetic, or other school subjects [8, 9]. These children may not be the same and shows a different combination and degree of difficulties. In difference to other disabilities, such as blindness, vision and motor impairments, a learning disability does not become immediately apparent in daily life [8]. If provided assistance related to children with learning disabilities, they can sparkle and they can be successful.

According to The National Joint Committee on Learning Disabilities (NJCLD) [10], the term ‘learning disability’ is defined as “A heterogeneous group of disorders manifested by significant difficulties in the acquisition and use of listening, speaking, reading, writing, reasoning or mathematical abilities. These disorders are intrinsic to the individual and presumed to be due to Central System Dysfunction.”

RFID has been used in variety of learning research. Result from the previous similar research show that the RFID technology can help children learners to have better learning experiences in terms of experiential learning, constructivist learning and more effective learning [16–18].

Figure 84.3 shows the number of publication related to RFID in learning, starting from 2008 to 2012. The analysis is done based on the number publication restricted in the Scopus, IEEE, Science Direct, ISI Web of Science, and Google Scholar. Figure 84.3 shows a significant increase number of publications, due to the features of RFID, it’s ideally suited for learning. Despite a decrease number of publication in 2009, however it is increased again starting from 2010.

84.2.3 Learning Disability

In Malaysia, children with learning disabilities are categorised as special needs children. The Malaysian Ministry of Education categorised special needs into three categories, those who are visually handicapped, or partially or fully deaf, or suffering from the disability to learn [11].

Traditional computer or web based learning environments offer the same content and they do not consider the individual differences, preferences and interests [12]. Modern technology offers great opportunities and continues to offer novelties that assist students with disabilities to take part in education and life-long learning. Emerging learning environments blended with technology offers a great potential to promote and enhance students learning processes to live freely and learn more easily. Many researches over the last 30 years have shown that technology can play an important role in specific disadvantaged groups such as the blind, those with movement disabilities and LD [6, 9, 12].

84.2.4 Learning Style

Learning is the process of gaining knowledge and active process, so it may be useful in life situations [2]. All people learn in different ways. Different learners have different cognitive processes, learning style preferences and past experiences that they apply when they are learning. An individual's learning style will affect the way in which information is processed and will effect on the learning effectiveness and efficiency [19].

Many researches showed that knowing learning styles of individuals will assist their learning process [13]. Learning style can be defined as individual preferences of learning and differences in students' learning and considered as one of the factors influencing a learner's achievement [13, 5]. Almost 71 different learning styles models have been stated [5]. Learning styles that are most used in literature are Felder–Silverman, Kolb, Dunn and Dunn, Honey and Mumford and Visual, Auditory and Kinesthetic (VAK) learning style [5, 14]. The main purpose for selecting VAK learning style for our system is the reason that this learning style is suitable for structural characteristics of topics in creating the content of the system. It can also apply e-Learning approaches that represent using media for all learning modalities. For example, visually, there are present words and images. For auditory, there are spoken words and sound explanation. Similarly, in kinesthetic, interactive animation is presented. For many years, multi-sensory approaches have been supported in order to help LD children [6].

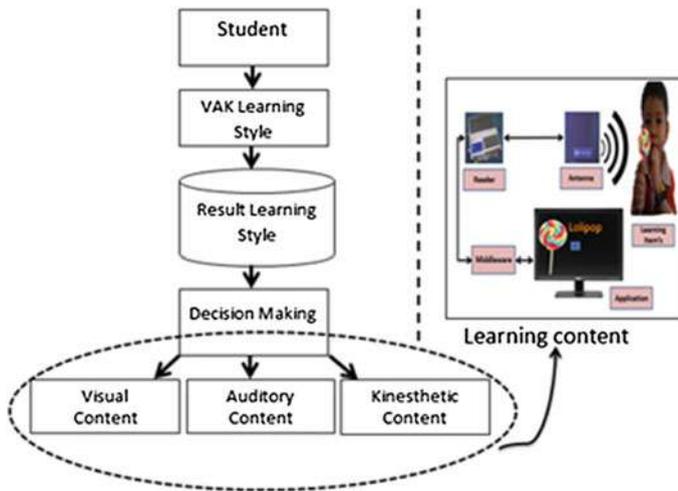


Fig. 84.4 Architecture of system

84.3 Proposed Approach

The purpose of our research is to combine an e-Learning with RFID technology in order to model the learning style focused for LD children. Today's technology offers great chances to assist students with disabilities to live freely and learn more easily. Developing the learning environments assisted by technology for these students is of great importance in improving their learning processes.

In this study, an adaptive e-Learning was proposed based on learning style of a learner using VAK learning style and RFID technology focused for LD children to improve their learning performance. Based on the VAK learning style, different contents will be prepared according to these three styles that are offered to the student.

This system also connects the digital and physical world by providing a platform for LD children to select tangible object and receive computer-based multimedia instructions. By using RFID technology, the children only needs to explore, select items, and simply move them over the RFID reader. A different content which will be automatically launched is prepared according to these three styles that are offered to the student to achieve greater impact on the learning process.

General architecture of this system is described. This system will be designed as an adaptive e-Learning environment where the content will be based on VAK learning style and is supported with RFID technology. The architecture of this system is shown in Fig. 84.4.

As a part of the research, scenarios appropriate for constructivist approach were prepared for the teaching of these subjects. These scenarios will be designed in

three different ways according to the sub-areas of VAK learning style. While transferring the activities into the digital environment according to sub-learning styles, characteristics of each area were made prominent. For example, by using RFID technology, the children only needs to explore, select items, and simply move them over the RFID reader. A different content which will be automatically launched is prepared according to this three styles that are offered which creates an activity for auditory learning style, sound lectures and instructions. While making an activity for visual learning style, more images, flow charts and animations were used. Similarly, while making an activity for kinesthetic learning style, learners were mainly presented with interactive animations and embedded with the storytelling concept.

84.4 Conclusions

This paper proposed a model for an adaptive e-Learning which is able to model the learning style of learners using the VAK learning style with RFID technology. Children learn the content that is appropriate to their own learning styles via this system. It is possible to say that this system provides a fully individualized environment to learners. In summary, the system offers what learners need. It creates a promise to significantly improve the learning result. Indeed, it provides a great opportunity improving the quality for learners to acquire knowledge faster and flexible, by assist learners to study in “the best way”.

Acknowledgments We would like to thank Senstech Sdn. Bhd. as contributing funds to conduct this research. The authors also would like to express a deep gratitude to the anonymous reviewers of this paper. Their useful comments have played a significant role in improving the quality of this work.

References

1. Ghadirli, H.M., Rastgarpour, M.: A model for an intelligent and adaptive tutor based on web by Jackson’s learning styles profiler and expert systems. In: Presented at the International MultiConference of Engineers and Computer Scientists 2012, Hong Kong (2012)
2. Sun, P.-C., et al.: What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction. *Comput. Educ.* 50, 1183–1202 (2008)
3. Akplotsyi, R., Mahdjoubi, L.: Effects of learning styles on engaging children in school projects. In: presented at the ARCOM, Bristol, UK, 5–7 Sept 2011
4. García, P., et al.: Evaluating Bayesian networks’ precision for detecting students’ learning styles. *Comput. Educ.* 49, 794–808 (2007)
5. Özyurt, Ö., et al.: Design and development of an innovative individualized adaptive and intelligent e-learning system for teaching–learning of probability unit: Details of UZWEBMAT. *Expert Syst. Appl.* 40, 2914–2940 (2013)
6. Beacham, N.A., Alty, J.L.: An investigation into the effects that digital media can have on the learning outcomes of individuals who have dyslexia. *Comput. Educ.* 47, 74–93 (2006)

7. Hanson, J: An introduction to RFID development Available at: <http://www.devx.com/enterprise/Article/31108> (2006)
8. Savidis, A., et al.: Developing inclusive e-learning and e-entertainment to effectively accommodate learning difficulties. *Univ. Access Inf. Soc.* 5, 401–419 (2007)
9. Adam, T., Tatnall, A.: Using ICT to improve the education of students with learning disabilities. 281, 63–70 (2008)
10. Sulaiman, T., et al.: The level of cognitive ability among learning disabilities children in Malacca Malaysia. *Int. J. Psychol. Stud.* 3(1), 69–77 (2011)
11. HJ, T., C, SK., P, Woo.: Student Learning Disability Experiences, Training and Services Needs of Secondary School Teachers. *Malays J. Psychiatry.* 17 (2010)
12. Polat, E., et al.: Adaptive web-assisted learning system for students with specific learning disabilities: a needs analysis study. In: *Educational Sciences: Theory & Practice*, pp. 3243–3258 (2012)
13. Ozyurt, O., et al.: Uzwebmat: a framework for expert system based on personalized adaptive and intelligent tutoring system for mathematics. In: *IADIS International Conference e-Learning*, pp. 173–180 (2011)
14. Mahdjoubi, L., Akplotsyi, R.: The impact of sensory learning modalities on children's sensitivity to sensory cues in the perception of their school environment. *J. Environ. Psychol.* 32, 208–215 (2012)
15. Khan, B.H.: ELearning-Chapter 5: Layout 1. In: *The Global e-Learning Framework ed, STRIDE* pp. 42–51 (2010)
16. Huang, K., et al: Breaking the sound barrier: designing an interactive tool for language acquisition in preschool deaf children. In: *Proceedings of the 7th International Conference on Interaction Design and Children, IDC '08* (2008)
17. Parton, B.S., Hancock, R.: Vision 3D: digital discovery for the deaf. In: *International Symposium on Instructional Technology and Education of the Deaf*, Rochester, New York, June 2008
18. Wei, C.W., et al. A joyful classroom learning system with robot learning companion for children to learn mathematics multiplication. *Turkish Online Journal of Education.* 11–23 (2011)
19. Bencheva, N.: Learning styles and e-learning face-to-face to the traditional learning. Vol.49 (2010)

Chapter 85

Palmprint Identification Using Invariant Moments Algorithm Based on Wavelet Transform

Inass Shahadha Hussein and M.J. Nordin

Abstract Because of the uniqueness of palmprints found on the palms of humans, palmprint identification has been used in several applications. It is usually associated with criminal identification, and has now become more popular in civilian applications. Therefore, the aim of the proposed model is to improve personal identification based on extracting shape feature using moments algorithm based on wavelet transform and matching algorithm, which is proposed in this model. This model has shown promising results without affecting rotation, translation and scaling of objects, because it is associated with the use of a good description of shape features. This system has been tested using databases from the Chinese Academy of Sciences (CASIA), in Beijing. By using false rejection rate (FRR) and false acceptance rate (FAR), we calculated the accuracy of identification. The experiment shows 98 % identification rate in the CASIA database.

85.1 Introduction

A wide variety of systems require reliable personal authentication schemes to either confirm or determine the identity of individuals requesting their services. The purpose of such schemes is to ensure that the rendered services are accessed by a legitimate user, and not anyone else. Traditionally, passwords (knowledge-based security) and identification cards (token-based security) have been used to restrict access to systems. However, security can be easily breached in these systems when a password is divulged to an unauthorized user or a card is stolen by

I.S. Hussein (✉) · M.J. Nordin
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia,
Bangi, Selangor, Malaysia
e-mail: inasshussin@yahoo.com

M.J. Nordin
e-mail: jan@ftsm.ukm.my

an impostor. Furthermore, simple passwords are easy to guess (by an impostor) and difficult passwords may be hard to recall (by a legitimate user). The emergence of biometrics has addressed the problems that plague traditional verification [1]. Among all biometric traits, palmprint has one of the highest levels of reliability and has been extensively used by forensic experts in criminal investigations. A palmprint is obtained from the central surface of the hand, between the wrist and the fingers. Figure 85.1 shows the shape and details of a palmprint.

A simple biometric system has four important modules [2], which are the following:

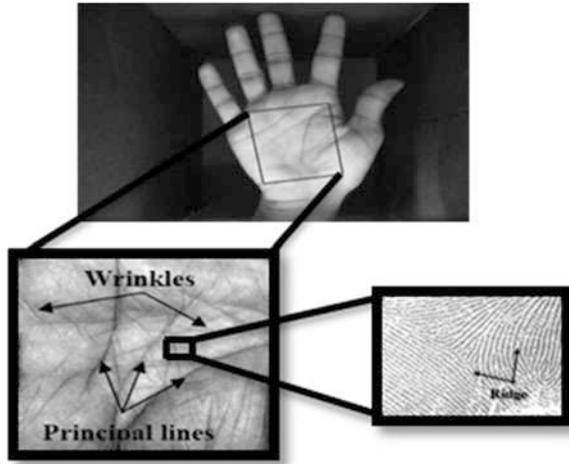
- The Sensor Module, which captures the biometric data of an individual. An example is a palmprint sensor that captures palmprint impressions of a user.
- The Feature Extraction Module, in which the acquired data is processed to extract feature values. For example, extract shape feature by using moments algorithm.
- The Matching Module, in which the feature values are compared against those in the template by generating a matching score. For example, in this module, the number of matching minutiae between the query and the template can be computed and treated as a matching score.
- The Decision-making Module, in which the user's claimed identity is either accepted or rejected based on the matching score generated in the matching module.

The rest of paper organized as follows: Sect. 85.2 explains in details the methodology that used in this paper. Section 85.3 introduces the results and discussion. Finally, Sect. 85.4 concludes the paper.

85.2 Methodology

This study explains and discusses the proposed method in performing palmprint recognition. The first stage is the acquisition of hand images from the CASIA. The second stage is pre-processing, which affects two main steps: including the binarizing hand image and extracting the palmprint region from the whole hand. This stage helps strip images from noise. The final stage includes feature extraction methods and matching techniques, such as invariant moments, used in computer vision applications. In this stage, the shape descriptor approach is employed in extracting global features of the palmprint, based on the wavelet transform, by decomposing each palmprint image into four bands and deriving seven invariant moments for each band. Finally, matching is carried out to perform identification. This method can generate more accurate results in less time. Points of interest in the identification of the palmprint are the features of an image, the change in the scale and rotation of the image's features.

Fig. 85.1 Shape and details of a palmprint



85.2.1 Hand Image Acquisition (Data Set)

The databases were obtained from the CASIA. Images of the hand were obtained using a flatbed scanner set at 256 grayscale. Grayscale images are mere representations of the image using only one color, which is gray, to cover the entire image. In a grayscale image, the density of the color component is used to represent the image. This component carries more information about the image. It also reduces the size of the image dimensions, for example, instead of three bytes, the image will be reduced gradually to a single byte (which is 0–255 in the decimal system). The image, when converted to this type of color range, it becomes easier to explain and promote. This property is also another feature of the grayscale mode [3]. In this study, we used the CASIA database. The CASIA database contains 800 images of the hands of 50 subjects. All images measure 640×480 pixels. Each user submitted eight pictures of their right hand and eight pictures of their left hand. The images show 256 gray levels on the flatbed scanner. The scanner is link-free, and therefore users are free to put their hands anywhere on the scanner as long as their hands are placed face down. Figure 85.2 shows samples of the left hand images. To ensure that the reproduction of your illustrations is of a reasonable quality, we advise against the use of shading. The contrast should be as pronounced as possible.

85.2.2 Image Pre-processing

Image processing is an area described by the need for intensified efforts to create an experimental feasibility of the proposed solutions to a specific problem. Image processing is a discipline in computer science. Each image has a value of more

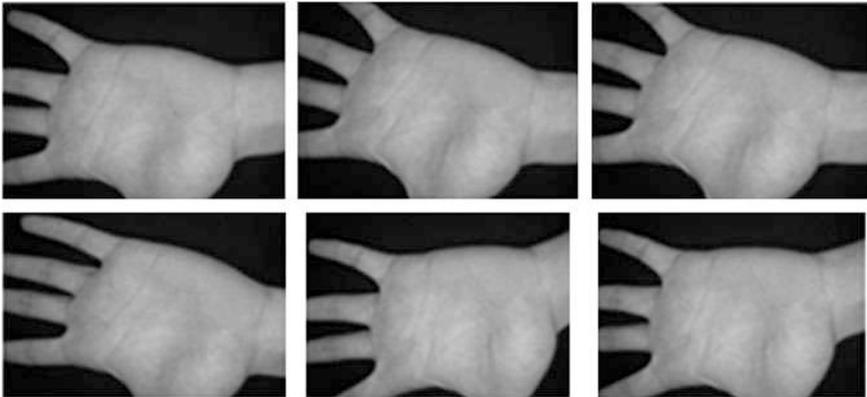


Fig. 85.2 Samples of the left hand images

than tens of thousands of words and provides a number of meanings. The digital images can be defined through the various definitions. For example, a digital image can be formulated as a function of two dimensions, and (X, Y) , where x and y are spatial coordinates [4]. Values $F(X, Y)$ represent the clarity or jamming of the image at a certain point. All the digital images of different elements, each element have a location and a value. These elements are referred to as image elements and pixels. The image processing is the study of the use of a variety of algorithms that were conducted on the pictures, where all of the input and output of the algorithms is the picture. First, we binarized hand image from the original image, at this stage a global threshold used to extract the image from rare hand. Figure 85.3 shows the original image and binarized image.

After that, extraction of the region from the whole hand will be done [5]. Figure 85.4 shows a sample of the extracted images.

85.2.3 2D Wavelet Transform

After extracting the region from the whole hand, 2D filter banks will apply to these regions, to get decomposed images into four bands [6], low–low pass filter, low–high pass filter, high–low pass filter and high-high pass filter (LL, LH, HL, HH) (Fig. 85.5).

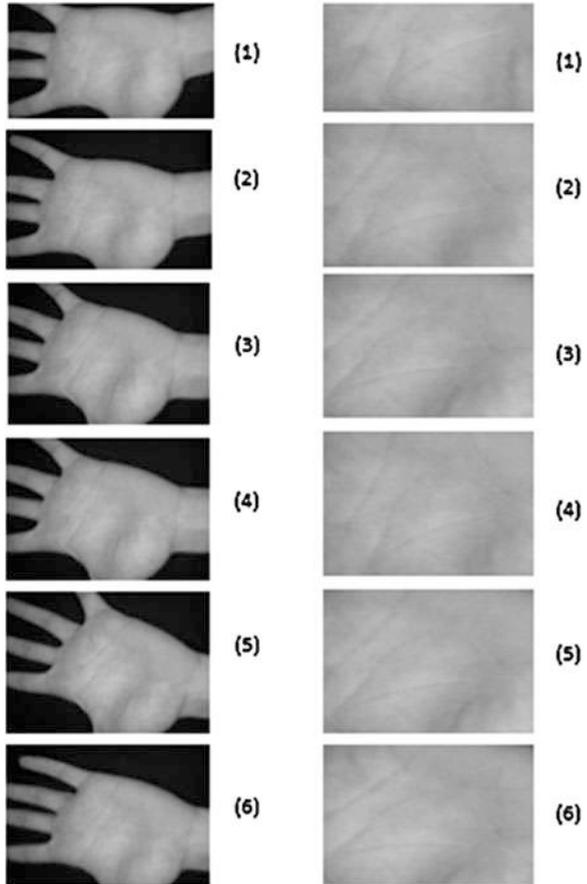
85.2.4 Invariant Moments Algorithm

Recognition of objects efficiently can extract feature points of objects (palmprint) to create a powerful feature descriptor or the representation of objects. Hu [7] has introduced a technique to extract the shape features form of images, which are



Fig. 85.3 Binarized hand images of the left and hands

Fig. 85.4 Samples of extracted left palmprint of one user from the CASIA database



called invariant moments. These features are invariance to translation, rotation, and change in scaling.

The advantage of the moments over other techniques is the implementation of the previous descriptors is straight forward, and they also carry a “physical” interpretation of boundary shape. Hu was first to set out the mathematical foundation for two-dimensional moment invariant and demonstrated their applications

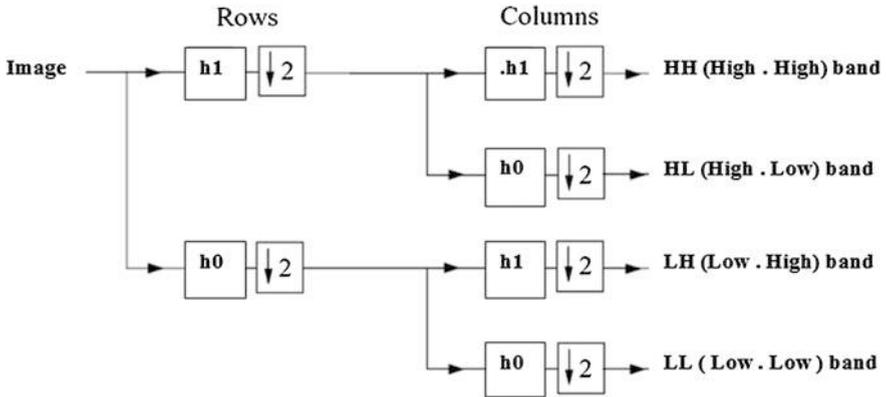


Fig. 85.5 2D filter bank analysis

to shape recognition. Invariant moments were first applied to aircraft shapes and have been proven to be fast and reliable. These values are invariant with respect to translation, rotation and shape scaling.

Hu defines seven of these shape descriptor values computed from central moments through order three that are independent to object translation, scale and rotation.

$$\Phi_1 = \eta_{20} + \eta_{02} \tag{85.1}$$

$$\Phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{211} \tag{85.2}$$

$$\Phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{85.3}$$

$$\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{85.4}$$

$$\begin{aligned} \Phi_5 = & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12}) [(\eta_{30} + \eta_{12})^2 \\ & - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\ & [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \tag{85.5}$$

$$\begin{aligned} \Phi_6 = & (\eta_{20} - \eta_{02}) [(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \end{aligned} \tag{85.6}$$

$$\begin{aligned} \Phi_7 = & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) [(\eta_{30} + \eta_{12})^2 \\ & - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}) \\ & [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \tag{85.7}$$

85.2.5 Identification

To obtain the identification accuracy of our palmprint system, each of the palmprint images was matched with all of the palmprint images in the database. A matching is noted as a correct matching if two palmprint images are from the same palm. If we suppose the query image is an X , and images in the database are N , we must compare X with N .

85.3 Results and Discussion

The main objective of this study is to obtain higher authentication accuracy by using a proposed algorithm after performing all previous stages. In this stage, 50 users from the CASIA database had features of their palmprints extracted using the invariant moment algorithm. This stage is considered fundamental in building the proposed system of recognition. To find out how accurate this algorithm is, we need to determine the accuracy rate of the identification phase. For determining accuracy rate, we have used two factors, which were FRR and FAR.

85.3.1 Results Based on Wavelet Transform

After applying wavelet transform (2D filter bank). The output is images that had been decomposed into four bands (LL, LH, HL, HH). Figure 85.6 shows an example of decomposed palmprint images.

85.3.2 Results Based on Moments Algorithm

After applying wavelet transform on palmprint images and decomposing these images into four bands (the output from the previous stage), The invariant moments algorithm was applied for deriving seven moments from each band. Table 85.1 shows values of the seven moments of four bands for one palmprint user (Table 85.2).

85.3.3 Accuracy Rate of Identification Phase

For the purpose of calculating the accuracy rate of the proposed method to perform the identification phase by matching query palmprint image with each one stored in the database based on FRR and FAR. To calculate the FRR factor, samples of

Fig. 85.6 Samples of decomposed palmprint for five users

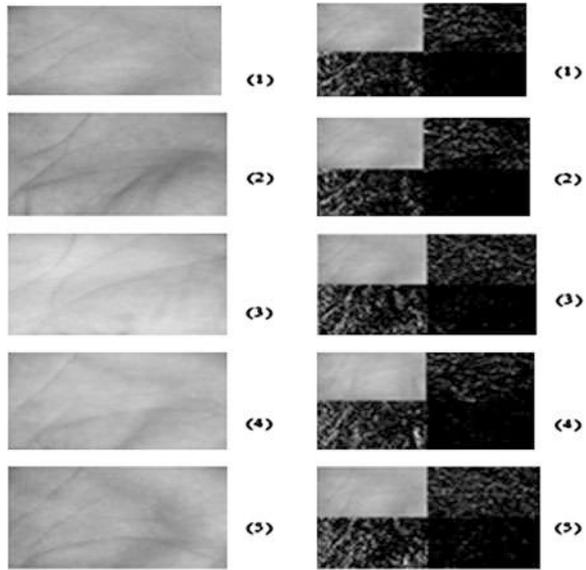


Table 85.1 Values of the seven moments of four bands for one palmprint user

Moments	LL	LH	HL	HH
First moment	9.66809E-04	0.72658299392	0.49137043951	34.0659780176
Second moment	6.78459E-12	3.3876777E-03	4.8144644E-04	457.609538309
Third moment	2.08395E-13	8.92198E-03	6.57074E-03	14562.2738551
Fourth moment	1.14094E-13	9.5015245E-03	3.8902797E-03	1665.46619179
Fifth moment	-3.1786E-20	7.7039764E-04	2.3161658E-04	-4608230.8966
Sixth moment	2.19293E-19	5.1328263E-04	-4.361702E-05	27447.6711358
Seventh moment	-1.1909E-26	-6.772396E-05	-1.531511E-05	6526394.74011

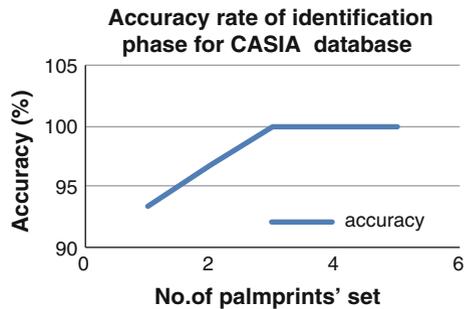
Table 85.2 Final seven moments which were derived from Table 85.1 for one palmprint

Moment	Value
First moment	8.82122456503965
Second moment	114.403351858058
Third moment	3640.57233696204
Forth moment	416.369895895987
Fifth moment	-1152057.72390458
Sixth moment	6861.91790138718
Seventh moment	1631598.68500723

Table 85.3 Accuracy type

No. set	FRR (%)	FAR (%)	Accuracy (%)
1	6.66	6.66	93.33
2	6.66	0	96.66
3	0	0	100
4	0	0	100
5	0	0	100
Rate	1.33	2.66	98

Fig. 85.7 Accuracy rate for CASIA database of identification phase



five sets were taken, the size of each set was 15 palmprint templates. These samples belong to the same samples stored in the database. To calculate the FAR, samples of five sets were taken, the size of each set was 15 palmprint templates. These samples were of out the stored database. Finally, accuracy rate was calculated based on FRR and FAR as: $Accuracy = (100 - (FRR + FAR))/2$.

According to the results in Table 85.3, it was observable that the highest accuracy was at set no. 3, 4 and 5, with a value of 100 %. The results had the highest FRR, which was 6.66 %, which means, just one palmprint of the fifteen valid users was rejected, and with the same high rate (6.66 %) of the FAR, means that one of fifteen invalid palmprints was accepted. The accuracy rate is 98 % (Fig. 85.7).

85.4 Conclusion

All methods used in this study are based on important steps in the processing and recognizing of palmprints. The major important steps include preprocessing, feature extraction, and matching. All of the aforementioned processes were carried out in a variety of ways and algorithms, which are suitable for the structural condition of the image itself. The results of the experiment show that the overall goal of using invariant moments on the basis of the wavelet transform is

successfully implemented on palmprints. It was found that moments of extracting palmprint to be invariant to rotation, translation and scaling. Transformation was applied to decompose palmprint images, that made moments more stable implemented to achieve its goal.

Acknowledgment The authors wish to thank Universiti Kebangsaan Malaysia (UKM) and Ministry of Higher Education Malaysia for supporting this work by research grant ERGS/1/2012/STG07/UKM/02/9.

References

1. Ross, A.: Information Fusion in Fingerprints Authentication. Michigan State University, Michigan (2003)
2. Jain, A., Hong, L., Pankanti, S., Bolle, R.: An Identity Authentication System Using Fingerprints Pattern Recognition and Image Processing Laboratory. Michigan State University, Michigan (1997)
3. Abdullah, A., Veltkamp, R.C., Wiering, M.A.: Fixed partitioning and salient points with Mpeg-7 cluster correlograms for image categorization. *Pattern Recogn.* **43**(3), 650–662 (2010)
4. González, R.C., Woods, R.E.: *Digital Image Processing*, 1st edn., pp. 703. Addison- Wesley, Reading, Massachusetts (1993)
5. Pavlidis, T.: *Algorithms for Graphics and Image Processing*, p. 416. Computer Science Press, Rockville, Maryland (1982)
6. Zainal, Z.A., Manaf, M., Shibghatullah, A.S., Jusoff, K., Ahmad, R., Ayop, Z., Anawar, S., Shaaban, A., Yusoff, M.: A New Hybrid Embedding Method in Iris Biometric System *Australian. J. Basic Appl. Sci.* **7**(3), 46–50 (2013)
7. Hu, D., Feng, G., Zhou, Z.: Two-dimensional locality preserving projection (2dlpp) with its application to palmprint recognition. *Pattern Recogn.* **40**(1), 339–342 (2007)

Chapter 86

Auto Mobile Ad Hoc Mechanism in Delay Tolerant Network

Muhammad Affandy Azman, Sharifah Hafizah Syed Ariffin,
Norsheila Fisal, Mazlan Abbas, Mohd Husaini Mohd Fauzi
and Sharifah K. Syed-Yusof

Abstract Delay Tolerant Network (DTN) is known as the solution to an Internet network where connectivity is an issue. There are existing project which uses smartphone as a physical transport of data between these intermittent networks. However, each smart phone needs to connect to each other via an infrastructure which will result in lower successful transfer rate. An automated mechanism is proposed for Ad Hoc connection between the smart-phones is to ensure connectivity which leads to higher successful transfer rate. This report presents the automated mechanism called Auto Mobile Ad Hoc Network in Delay Tolerant Network that is able to provide better reliability for data transmitted through DTN. This mechanism also allows application of the current Delay Tolerant Network to be connected to other networks and other nodes automatically using Ad Hoc mode.

M.A. Azman · S.H.S. Ariffin (✉) · N. Fisal · M.H.M. Fauzi · S.K. Syed-Yusof
UTM-MIMOS Center of Excellence, Universiti Teknologi Malaysia, UTM Skudai,
81310 Johor Bahru, Johor, Malaysia
e-mail: sharifah@fke.utm.my

M.A. Azman
e-mail: maffandyaz@gmail.com

N. Fisal
e-mail: sheila@fke.utm.my

M.H.M. Fauzi
e-mail: hus6012@gmail.com

S.K. Syed-Yusof
e-mail: kamilah@fke.utm.my

M. Abbas
MIMOS (M) Bhd, Technology Park, 57000 Kuala Lumpur, Malaysia
e-mail: mazlanabbas@mimos.com.my

86.1 Introduction

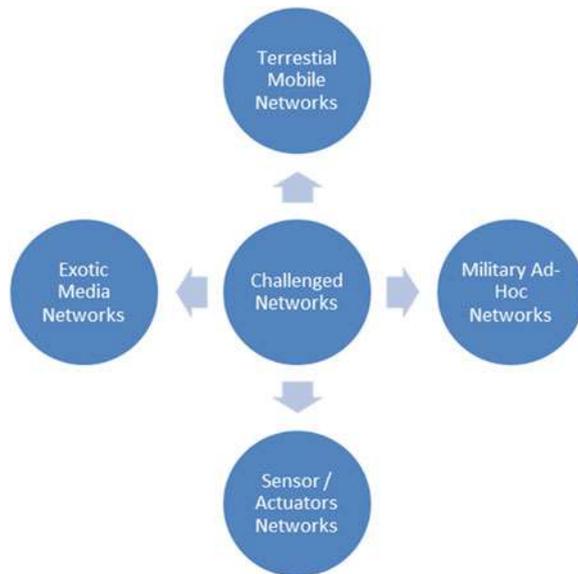
Internet plays an important role to connect people regardless of distance. However, not everyone has the opportunity to use the Internet, because of the challenging issues due to delay and disruptions. One of the most challenging issues for a certain geographical area to have communication infrastructure is the lack of infrastructural and facilities. Delay Tolerant Network (DTN) architecture for challenged network are categorized in [1] where DTN can be implemented in four different challenged networks shown in Fig. 86.1. There are terrestrial mobile networks where users are mobile and commuting from one network to another, there is also exotic media networks where near-earth satellite communication that is prone to latencies. There are two other challenged networks in DTN which are the military Ad Hoc network where intentional jamming might be the cause of the disconnection and network need to compete for bandwidth, and sensor/actuators network where this networks are consist of millions of low powered nodes and communication within this network often scheduled to conserve power.

There are projects which uses the DTN approaches such as in [2–4] to apprehend these challenges. One of the projects is Bytewalla: DTN on Android [5] that uses mobile devices to transfer data between nodes. This project aims at connecting African rural villages using Android phones with delay-tolerant networking. The idea is that people carrying their mobile phones and travel from villages to cities will carry data with them. Once they reach the city, they will connect to any WiFi access point and upload the data. For the past years, Bytewalla had evolved from version 1–5, each with new features such as security and routing protocols. However, there are several issues in Bytewalla that can be improved. In the current version of Bytewalla both nodes (i.e. transmitter and receiver) need to be connected manually to an infrastructure or the same network to allow data to be transferred. If both nodes are not connected in a same network, the data will not be forwarded, and will be stored in the transmitter node until the data expires, which affects the reliability of data transfer.

In the present report, we will discuss the implementation of an automatic mechanism called Auto Mobile Ad Hoc function using Bytewalla. This mechanism is to improve the probability of successful bundle (a protocol data unit of the DTN bundle protocol) send and receive by enabling all nodes to connect to each other automatically rather than to wait for an available Access Point (AP). Each node will act dynamically according to the certain situation. This implementation is still using the conventional way of transferring data by adding another layer in the Internet protocol suite which is the Bundle Protocol [6] and will be implemented on a smart-phone. This method is to enhance the existing Bytewalla software and to increase reliability data transfer as well as reduce transmission delay. Figure 86.1 shows the challenged network in DTN.

The report is organized as follows: Sect. 86.2 describes the protocols used in DTN and the transmission routings. Section 86.3 presents the proposed concept of

Fig. 86.1 Challenged network in delay tolerant network



Auto Mobile Ad Hoc Mechanism. In Sect. 86.4 development and implementation in experimental DTN test bed is presented in detail. In Sect. 86.5, results and discussion are elaborated. Last but not least, Sect. 86.6 concludes the report.

86.2 Data Protocols and Routings in DTN

The DTN is introduced to solve technical issues in networks that may lack of continuous network connectivity. The DTN is designed to operate effectively over extreme distances. An example of networks that are facing issue is the space-network, such as Inter Planetary (IPN) Internet Project [7, 8]. The DTN is used to tackle problems such as the intermittent connectivity, a long and variable delay, asymmetric data rates, and high error rates by using the store and forward switching [9, 10]. In order to implements this method, a device with a persistent storage is needed to hold the message indefinitely. Example of the device is a smart-phone. Figure 86.2 shows the store and forward method used.

It is possible for DTN to implement the store and forward method by overlaying a new protocol layer which is the Bundle Protocol (BP), RFC5050 on top of the Transport Layer. Basically, a bundle is a packet which contains a source-application user's data, control information and a bundle header. Each node is identified by Endpoint Identifiers (EID), which can be treated as address. There are several practical applications in DTN that have been done by other researchers. One of them is DakNet, which was developed by the MIT media Lab researchers [11]. Their goal was to provide remote villagers with low-cost digital communication by



Fig. 86.2 Store and forward method

equipping busses with a mobile access point that travels between village kiosk and city collecting data.

Routing is the main aspect of DTN. It determines the delivery success rate and the delay of the bundles. Currently there are two routing strategies in DTN. The first strategy is flooding, where the message is replicated to several nodes in order for destination EID to be received. Each node acts as a relay to store the bundle until they are able to contact with another node. One of the routing protocols that use flooding concept is Epidemic routing [12], where nodes replicate and transmit messages to newly discovered contacts continuously. The problem with Epidemic routing is that it may congest the network in clustered areas while wasting network resource (bandwidth, storage and energy).

The second strategy is forwarding strategy which uses the best path to the destination by making use of network topology and local or global knowledge to find the best route path to deliver the message to the destination without replication which can lead to less bandwidth and consumption, and faster. One of the routing protocols that use forwarding strategy is the Probabilistic Routing Protocol using History of Encounters and Transitivity (PRoPHET) [13]. This routing protocol uses algorithm which maintain the set of probabilities for successful delivery for known the destination. The downside for this protocol is that the actual value for the probability is crucial for successful bundle transfer. The routing protocol used in this project is the Epidemic routing. This is to ensure the successful transfer of bundle because of the high transfer probability rate in Epidemic routing.

86.3 Auto Mobile Ad Hoc Network

With the current DTN mechanism, all transmission from one mobile device to another need to be done manually even though there are potential node that can take the data to be forwarded later. A mechanism is added to solve this connectivity issue where the application will not connected to any network/node, if it is not initiated manually. This proposed mechanism will scan and connect automatically to another network/node.

The architecture in the experiment consists of three nodes (i.e. Android phones) which had been rooted (i.e. ability to access to system files and modify). Figure 86.3 illustrates the overall experiment scenario architecture. Each node will store and forward received bundle. The bundle will keep on forwarded until it has reached the destination EID. Each of the nodes is connected with each other through Ad Hoc network.

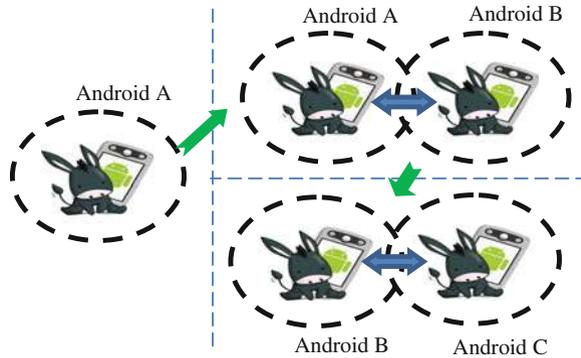


Fig. 86.3 Experimental scenario of the auto mobile Ad Hoc network

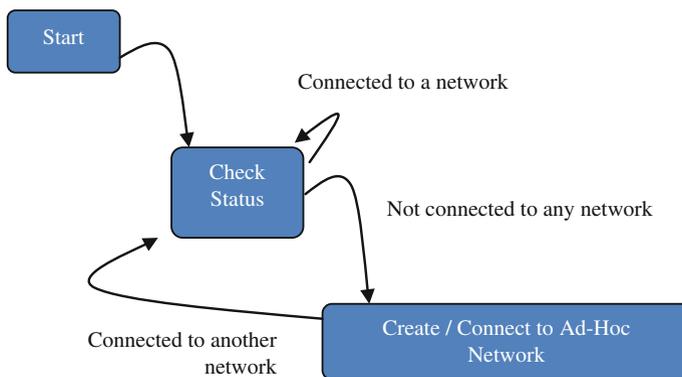


Fig. 86.4 State diagram of the auto mobile Ad Hoc network

Figure 86.4 shows the state diagram for Auto Mobile Ad Hoc mechanism in the DTN, which shows the flow for each node’s network connection. The node will continuously check for its own network status, if the node is connected to any network, the node will stay connected and resume the checking phase. If the node is not connected to any network, the node will then scan and attempt to connect to any nearby Ad Hoc node. If there are no Ad Hoc nodes nearby, the node will then create its own Ad Hoc network.

86.4 Experimental DTN Test Bed

The experimental DTN test bed consist of DTN Servers and mobile node that have been configured and embedded with auto mobile Ad Hoc function.

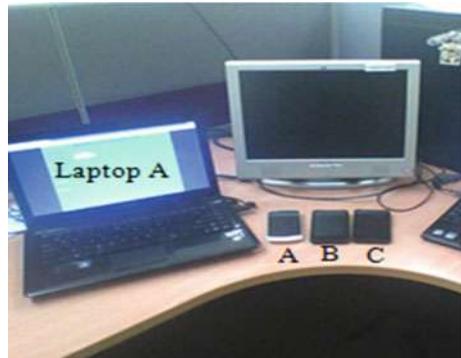
86.4.1 DTN Server

The DTN 2 application uses DHCP server, DTN2 software on Ubuntu Platform. In order to send a bundle, an EID and a payload (message) is needed. Each node has their own unique EID. For the smartphone, the EID is determined by the MAC address. An EID can be treated as an address for each node. Referring to Fig. 86.3, The bundle will be sent to the mobile node that is in the same network with the server (i.e. network A), which is Android A.

86.4.2 Mobile Node

For the mobile side, an application is called Bytewalla (Bytewalla is a DTN application for Android¹) is used. Each mobile node is installed with a modified version of Bytewalla which has the Auto Mobile Ad Hoc functionality. Each mobile node is also rooted, to enable the use of Ad Hoc in mobile node. Each node has its own unique EID. For the Android phones, the EID is determined by the MAC address. An EID can be treated as an address for each node. Referring to Fig. 86.3, a bundle will be sent to Android C which is the EID, originated from Android A. All of the nodes are not originally connected to each other. When Android A is in range with a nearby node (in this case Android B), Android A and Android B will initiate a connection automatically and the bundle will be sent to the Android B from Android A. The bundle will then store and forwarded to the next node until it has reached the EID (Fig. 86.5).

Fig. 86.5 Test bed set up



¹ Android-A linux based operating system for touch screen phones.

86.5 Results and Discussions

We investigate the effects of message size and number of nodes to the overall delay. The overall delay is calculated by calculating the time taken for a bundle to arrive to an EID from source. The measurement of the overall delay for each message size and number of nodes is carried out three times.

In the first experiment, we determined the delay based on the message size and different approaches, Android A and B were used as the nodes. Android A sent the bundle to Android B and both nodes are originally not in the same network. In this experiment we compared the delay of data transmission between infrastructure network and Auto Mobile Ad Hoc network.

Figure 86.6 shows the results of the first experimental setup for the delay of transmitted data with variable data size. As the data size increased, the delay also increased. Using the infrastructure setup, the minimum delay for 200 kB of message size is about 14 s. The delay increased gradually based on the data size to about 23 s for 1000 kB of message size. For the Ad Hoc setup, the minimum delay is about 34 s for 200 kB of message size and increased gradually to about 42 s for 1000 kB of message size. The difference in delay between Ad Hoc and Infrastructure is because of the time needed for an Android phone to create/establish an Ad Hoc network which takes around 30–32 s.

In the second experiment, we determined the delay based on the number of nodes and here, Android A, Android B and Android C were used. In this setup, all of the nodes are originally not in the same network. Android A will be the source sending the bundle to Android B or Android C which is the EID. In the second experimental setup, we compared between 2 and 3 nodes of data transmission from source to destination. In the 2-node setup, Android A and B are used. In the 3-node setup, Android A, Android B, and Android C are used.

Figure 86.7 shows the results from the second experiment of the delay with variable number of nodes. For 3 nodes, the minimum delay is about 65 s for 200 kB of message size which increases to about 83 s of delay for 1,000 kB of message size. As we increased the number of nodes, the overall delay also increased almost double. The delay increases from 2 to 3 nodes because all of the nodes are not originally in the same network, thus each nodes creates their own Ad Hoc which consume time.

Figure 86.8 shows the delay compared between the original Bytewalla and the Auto Mobile Ad Hoc Network. As shown in Fig. 86.8, the Auto Mobile Ad Hoc, the delays increased as we increase the message size to a maximum of about 42 s. However, for the original Bytewalla, because of the nodes originally not in a same network, the nodes will not be connected to any other node and the bundle will not be send to the next node, hence, the delay is infinity.

Fig. 86.6 Delay versus message size for infrastructure and Ad Hoc mode DTN

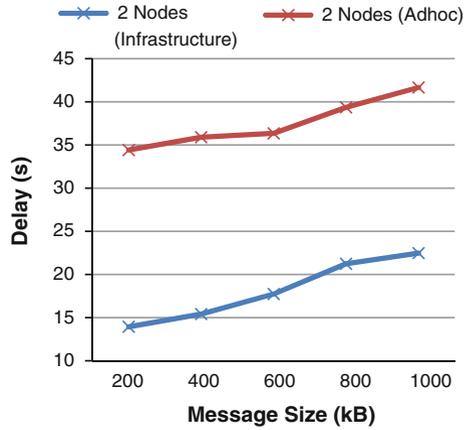


Fig. 86.7 Delay versus the number of mobile nodes the message are transmitted in DTN

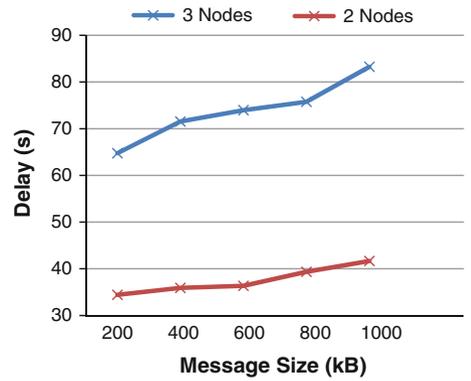
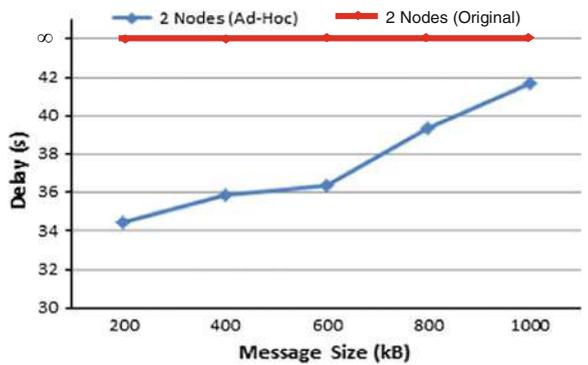


Fig. 86.8 Reliability of the Ad Hoc function in DTN compared with the original version of DTN



86.6 Conclusion

This report has presented the implementation of Auto Mobile Ad Hoc function in the existing DTN application using Bytewalla software embedded in mobile nodes. The Auto mobile Ad Hoc function is to improve the probability of successful bundle transfer to the destination by adding the ability to connect to each node automatically. This network concept can be very fruitful for the current situation since the bundle will keep on transferring to the next node instead of relayed only on one node. From the results shown, the Auto Mobile Ad Hoc Network function is able to connect automatically to other nodes without user prompt which improved the probability of successful bundle transfer. However, there are delays involved, which is the time needed for a node to create its own Ad Hoc network and connects to other Ad Hoc network. The Auto Mobile Ad Hoc Network certainly has its drawback by having about 32 s of delay for the Ad Hoc network creation. However, it will ensure higher reliability of data transfer through the automated connection of nodes through Ad Hoc network.

Acknowledgments The authors would like to thank all those who contributed toward making this research successful. The authors wish to express their gratitude to Ministry of Higher Education (MOHE), Malaysia Research Management Center (RMC) for the sponsorships, Telematic Research Group (TRG), Universiti Teknologi Malaysia and MIMOS (M) Bhd for the financial support and advice of this project. (Vot number Q.J130000.7823.4L550).

References

1. Fall, K.: A Delay-Tolerant Network Architecture for Challenged Internets. ACM SIGCOMM 2003, Germany
2. Fall, K., Farrell, S.: DTN: An architectural retrospective. *IEEE J. Sel. Areas Commun* **26**(5), 826–828 (2008)
3. MacMahon, S.F.: Delay-and disruption-tolerant networking. *IEEE Internet Comput.* **13**(6), 82–87 (2009)
4. Cerf, V., Hooke, A., Torgerson, L., Dust, R., Scott, K., Fall, K., Weiss, H.: Delay-tolerant networking architecture. IETF RFC 4838 (2007)
5. Ntareme, H., Zennaro, M., Pehrson, B.: Delay tolerant network on smartphones: applications for communication challenged areas. *ExtremeComm11'*, ACM International Conferences Proceedings Series, doi[10.1145/2414393.2414407](https://doi.org/10.1145/2414393.2414407)
6. Scott, K., Burleigh, S.: Bundle Protocol Specification IETF RFC 5050 (2007)
7. Akyildiz, I.F., Akan, O.B., Chen, C., Fang, J., Su, W.: InterPlaNetary internet: state-of-the-art and research challenges. *Comput. Networks J. (Elsevier Science)* **43**(2), 75–112 (2003)
8. Burleigh, S., Hooke, A., Torgerson, L., Fall, K., Cerf, V., Durst, B., Scott, K.: Delay-tolerant networking: an approach to Interplanetary Internet. *IEEE Commun. Mag.* **41**(6), 128–136 (2003)
9. Chuah, M., Yang, P., Davidson, B.D.: Store-and Forward performance in DTN, *IEEE VTC* (2006)
10. Macmahon, A., Farrell, S.: Delay and disruption-tolerant network. *Internet Computing IEEE* (2009)

11. Pentland, A., Fletcher, R., Hasson, A.: Daknet: rethinking connectivity in developing Nations. *Computer* **37**(1), (2004) doi:[10.1109/MC.2004.1260729](https://doi.org/10.1109/MC.2004.1260729)
12. Jain, S., Fall, K., Patra, R.: Routing in a delay tolerant network. In: SIGCOMM '04 Proceedings of the 2004 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, vol. 34 (4), pp. 145–158
13. Vasilakos, A.V., Zhang, Y., Spyropoulos, T.: Delay Tolerant Network: Protocols and Applications. CRC Press, ISBN 1439811083 (2012)

Chapter 87

An Exploratory Study on Blind Users' Mental Model in Computer Accessibility

Manoranjitham Muniandy and Suziah Sulaiman

Abstract The growing needs in the area of HCI has led to many new researches focusing on user-centric design, investigating the problems faced by a computer user and ways to produce an effective yet efficient design. Often, designers need to acquire the knowledge and the experience of the users to produce a good interface design. This can be referred as a wise exploitation of user's mental model. The paper serves as a preliminary study which explores the mental model of blind users as a contribution in improving the accessibility of computer applications. Based on the blind user's perception, imagination and interpretation, the study indicates that touch sensation plays an important role in improving the representation of a computer application to them. Associating touch sensation with information familiar to the blind users enhances the learning process.

87.1 Introduction

The term "computer accessibility" in human-computer interaction spells out clearly the accessibility of computer system to all people, regardless of disability or severity of impairment [1]. The user interface of a computer should be accessible by anyone, independent of their physical, perceptual-motor, social and cultural capabilities [2]. In [3], it is stated that blind users usually use software known as "screen readers" which works with all the programs installed in a computer to read out the interface and details of an application to them. However, not all can be

M. Muniandy (✉) · S. Sulaiman
Department of Computer and Information Sciences, Faculty of Science and Information
Technology, Universiti Teknologi PETRONAS, Bandar Seri Iskandar,
31750 Tronoh, Perak, Malaysia
e-mail: bhannu18@yahoo.com; manoranm@utar.edu.my

S. Sulaiman
e-mail: suziah@petronas.com.my

read by the screen readers. Braille devices able to produce the output on the computer display however, some important information such as application names and pop-ups are not displayed by the Braille device.

World Health Organization (WHO) recorded the estimated visually impaired population as 285 million; with the blind holding the record of 39 million and the rest occupied by the low vision category [4]. Eweek.com did address an important question on why the blind community are being ignored when they are able to create a big market? [5]. The challenges in life and acquiring a new skill especially on IT field seem to be an issue to the blind. It is important to allow the blind to further venture into the IT equipped world by providing an easy learning platform for them. As a start, this paper presents a study that identifies the usability problems based on users' experience when interacting with a computer application. The intention is to understand the blind users' mental model. The study further investigates the extent in which touch sensation could be used in accessing the application.

87.2 Literature

Designers tend to address a user's mental model as a valuable piece of information due to the ability of these users to formulate the mental model using their experience and expectations. In order to design a user's preferred system, one need to utilize the user's mental model as this information is found to be greatly contributing during the development stage. Every designer considers a mental model to be an absolutely necessary element as it helps them to understand a user's view on system content [6, 7]. Usually, when a user interacts with an existing or familiar environment, they tend to use the previously developed mental model and if they come across an unfamiliar or a new environment then they will create a new mental model altogether [6].

Familiarity can be achieved when a user's mental model is analyzed and some of the important questions are answered such as (1) how a system should look alike? Or (2) how a system supposedly should behave? Or (3) how a user expects a system to respond? Mental model of a blind can be in any form, such as a diagram, an image, a theory, a set of concepts, some guidelines or etc. [8]. This representation is said to be a great help in enhancing the understanding of usability of any system [9, 10]. Usually, a system is build according to a designers' view. In another word, the developed system is based on a designer's mental model. In this study, mental model of blind users refers to the representation formed by the blind when they use computer application.

According to [2], usability issues highlight the easy use of an interface which is measured through the interaction between a user and the system. In [11], it was addressed the need of assistive technology when a blind computer user accesses a computer application. And that it is important for the designer to design a system

or an application which is “software” friendly for the screen readers to be able to read the presented content.

Often the blind users' mental models are not taken into account and the developed applications are tested without considering the fact that blind users have the tendency to navigate through an application instead of listening carefully the information that is being read [12]. A developed computer application must be oriented to usability and accessibility for the users from any background. Referring to Table 87.1, usability or also referred to as accessibility of an application is defined in terms of five quality components [13]. Usually, an application is designed by a designer having these five components in their mind as a benchmark to ensure good applications with great accessibility level is produced for the users. Current computer applications do not address the issues face by the blind. It can be due to the fact that, the availability of screen reader is believed to perform its' duty. And another reason can be due to the lack of awareness among the developers and also visually impaired are not even considered as a heavy computer user but research has proven that a large number of blind users are now very independent and learning new skills for survival.

87.3 Related Work

87.3.1 *Mental Model and Blind Internet Users*

In [14], the two-dimensional mental model created by the blind people when they are using touch screen with audio feedback was discovered. These authors have agreed that little work has been done on mental model for blind people. Mental model of blind people created from surfing a two-dimensional web page is investigated using two ways: firstly, is by using a touch screen display with audio feedback and the other using a screen-reader only. Users were given 10 to 15 min training on the each of this web page before the actual test was conducted. Five users are required to surf the pages twice using screen reader for the first time and using touch screen with audio feedback for the second time.

And another five did the same but using the touch screen with audio feedback for the first time and screen reader for the second time. After that, all users are given a set of foam blocks containing rough and smooth surface to construct the diagrammatic representation of the pages. Rough surfaced blocks to represent the headings while smooth surfaced blocks to represent the data. After the analysis of these diagrammatic representation, it shows that previously (using screen reader) the mental model of a blind user are in single column structure as quite a number of the representation stacked up the headings foam blocks in one column or both headings foam and data foam blocks in one column.

The same users had two dimensional mental model after using the touch screen with audio feedback which shows that these assistive technology able to deliver

Table 87.1 The five quality components [13]

Components	Description
Learnability	How easy is it for users to accomplish basic tasks the first time they encounter the design?
Efficiency	Once users have learned the design, how quickly can they perform tasks?
Memorability	When users return to the design after a period of not using it, how easily can they reestablish proficiency?
Errors	How many errors do users make, how severe are these errors and how easily can they recover from the errors?
Satisfaction	How pleasant is it to use the design?

the content of a layout better than a screen reader. As these users, tend to stack up the headings foam and data foam blocks in two-dimensional (table) format. However, their research also shows that the theory is applicable better if the layout is simple and not much different can be seen if the layout is complex. Therefore, it can be concluded that the two-dimensional mental model of the blind people is highly influenced by the complexity of a web page.

In [15], it was highlighted the issues faced by the blind society when navigating web pages. They assess the web pages based on one important accessibility point which is “Ease of navigation”. They have also highlighted that “information scent” plays an important role whereby related information are grouped or arranged together and if a blind user skips their targeted topic, they will know it by realizing the scent of the current topic which is different from the previous topic. Though the research used a systematic evaluation process, the final outcome of this research was merely an identification of the accessibility level among the e-commerce sites in three different countries.

87.3.2 Difficulties Faced by the Blind Users Involving the Sighted Individuals

A study by Fernando Alonso et al. focuses on the development of dual graphical user interfaces for both visually impaired and sighted communities. They concentrated on how to come up with a design that will be able to satisfy both types of users [16]. The main concern of this project was to deliver a set of guidelines that can be used in the dual interface development in accordance with design for all principle. Even though these authors did take into consideration for the blind users, but the crossed checked guidelines had to drop out some issues that exist only in the visually impaired environment since those issues does not co-exist in the sighted user environment.

A research on the application of various stimuli to produce an effective pattern which is believed to increase the effectiveness of represented information to the user was conducted in [17]. The study suggested a design principle defining the

stimulus pattern. It was a good study; however, the authors have failed to provide the same experiment for all the users which are essential in deriving a conclusion. To eliminate biasness during the evaluation, a change in the sequence of the tactile pattern should be done instead of not giving certain patterns at all for some users. Therefore, this shows that it is not right to conclude if the actual test subject is differs for all the users.

87.3.3 Audio Metaphors for the Blind Users

A conversational metaphor was proposed in [18], in which the metaphor focuses on the environment of 3D aural to allow the visually impaired for easy access of information. The proposed hypermedial model contains direct graph nodes and links which is used to map to certain speakers. Documents are represented by the nodes and the semantic relationships between the documents are represented using the links. A speaker is used to explain the detail of the nodes and links to the user. Many samples of real people were used to generate the 3D voice of speakers using an off-line way [18]. The authors have developed a metaphor for the visually impaired, the proposed idea focused only in audio environment and did not address the problems relating to computer accessibility or other usability issues related to blind users.

87.3.4 General Improvement of the Surrounding Objects for the Visually Impaired

A study was focused on improving the interactions between a blind and everyday technological manmade object such as hand phones, software applications and etc. In this article, interviews and observations were used to address the obstacles faced by the blind in using the technological gadgets [19]. In order to improve the gadgets, it is important to know the obstacles faced by the blind community while using the surrounding objects. In another review, Gregory Petit et al. has generated tactile graphics by translating the illustrations in the schoolbook to be accessed by the visually impaired students. Two types of tactile hardware were used to generate the three different tactile rendering of an illustration [20]. A significant result was shown in displaying the tactile illustration by applying a workable methodology; however, they did not propose any framework that can be used by others at the end of the project.

87.4 Motivation of the Study

Previous researches focus on the problems faced by the blind in using a screen reader thus not highlighting the inaccessibility of computer application to them [14, 18]. Some other findings relating to touch sensation were lacking of empirical study, hence not supporting the contribution of touch sensation in learning computer application [14, 17, 20]. Our initial investigation reveals that in order for a blind user to select and to open an application such as Microsoft Word, the user need to depend on the screen reader to read the file name or the name of the application. This entire process takes plenty of time as user need to wait and listen to the entire icon description to be read. Based on our findings, blind users possess rich tactile sensation in dealing with their daily life. There is a need to conduct an exploratory study on blind users to address their problems in accessing computer applications and to evident the presence of the touch sensation as an important factor that contributes effectively in improving the accessibility of computer applications.

87.5 The Study

The objectives of this study are twofold:

- (i) **To examine blind user's experience on computer accessibility.**

A thorough study will be conducted to identify the issues faced by the blind in accessing computer applications.

- (ii) **To examine how touch feedback may assist in accessing computer applications.**

Touch sensation is known to work well with blind, however, in this study, it will highlight to what extent touch can contribute in accessing computer applications.

87.6 Study Methods

A group of blind users with the need to learn IT skills was selected as the study object. As a good site to study, Malaysian Association of the Blinds (MAB) a voluntary organization serving the blind was chosen. In here, some basic skills such as mobility, Braille, English language and computer literacy are taught. The researchers conducted their field work to study extensively the blind users' experience and imagination during this entire rehabilitation course for one semester (duration of five months). The following section discusses on the three data collection methods used in this study.

87.6.1 Personal Interviews and in-Depth Interviews

Several informal interviews and four in depth interviews were conducted with seven blind users. One of them is the chairman, two of them are the instructors and four of them are the students. The age group of the interviewees is above 18 and their occupation varies tremendously. Interviews were chosen due to the nature of this method which has a very high response rate and also the method encourages the collection of true and correct responses addressing the nature of a problem [21]. The section below describes some of the questions asked during the interview session:

- *Can you easily open a Microsoft word application?*
- *How do you know that you are opening a word document instead of an excel sheet?*
- *Have you ever made a mistake of opening other documents instead of the intended document?*

The questions somehow addressing the same issue which is on the users' accessibility level when accessing an application during the learning process. The main objective of the researcher for asking the similar questions is to ensure the respondent's response is consistent and correct directing to the main problem of this research. Four in-depth interviews were also conducted to get the respondents to further open up about their concern and dissatisfactions when dealing with computer applications. In-depth interviews are useful when detailed information about a person's thoughts and imagination is required [21]. The section below highlights two of the questions asked during the in-depth interview:

- *In your imagination, how flexible do you want/prefer the existing computer applications to be presented?*
- *In your opinion, what kind of senses seems appeals the most for a blind user?*

These questions are designed to trigger the user to communicate their thoughts and expectations. When a user describes their thoughts and expectations in detail, it allows the researcher to identify the issues with the current application which does not satisfy the user's need.

87.6.2 Observation

Apart from interviews, observation was also done for the entire period of Braille class and computer literacy class with the same set of participants from the interview session except for the chairman. The observation was also video recorded for multiple reviews. Observation was conducted due to the ability of this technique recording the actual users' behavior. Observation eliminates the tendency of a respondent being bias or exaggerates their problems, captures the latest happenings and only actual behavior is being captured [21]. The facial reactions

and eagerness or lack of interest and irritations when a user encounters an issue while working with computer applications is studied to identify the negative emotions of a user and the factors which cause this negative emotion during their learning process. By identifying a users' emotion and factors which cause this emotion, more accurate information is obtained by the researcher to support the research questions of this research.

87.6.3 Hands-on Activities

Hands-on activities in the form of two simple games were conducted with four participants to learn more about a blind person's habit in applying their existing experience on discovering knowledge. The experiment was performed with the intention to explore and to identify the effect of touch sensation to the blind users. Polystyrene blocks resembling the Braille cells are created and referring to Table 87.2, blocks were used to represent some of the frequently used applications. The representations are new patterns and are not represented in Braille code. The pattern of pop-up blocker and pattern of applications (apps) is a new creation due to no suitable Braille code is available for the representation while the combination of application (apps) code and alphabets referring to the first word of an application is used for the complete representation. These inventions are used to test the touch sensitivity level among the participants to explore the mental model of the blind users to show that touch sensation able to contribute in improving the accessibility of computer applications.

87.7 Results and Discussion

87.7.1 Results of Interviews

The information presented in Table 87.3, highlights the views of each key person in terms of issues faced by the blinds in accessing computer applications. Referring to the feedback, they have admitted that almost all of their students are unable to open a simple application such as ms.word. This simple application needed repetitive actions even with the aid of a screen reader such as listening and clicking repetitively.

87.7.2 Results of Observation

The below analysis was done based on the observation of these selected group of participants and instructors. The video recorded observations were reviewed few times in order to ensure that only real issues are captured for analysis purpose.

Table 87.2 Patterns of the blocks to represent applications

Applications/description	Block pattern	Applications/description	Block pattern
(1) Pop-up blocker (new)		(4) Internet explorer (apps + i)	
(2) Ms. Word (apps + w)		(5) File (f + e)	
(3) Ms. Excel (apps + e)		(6) Folder (f + r)	

Table 87.3 Issues and views of key person on the blinds

Role	Issues and views
Chairman	<ul style="list-style-type: none"> • The difficulty of the participants in following the entire program and some abandoned half way due to the complex learning process
Instructors	<p><i>General issues faced by the students:</i></p> <ul style="list-style-type: none"> • Totally helpless when a pop up blocker pops out and inability to understand the displayed graphics as the screen reader reads it out as “graphics”. Expressed the need for additional assistive technology <p><i>Issues faced by the beginner students</i></p> <ul style="list-style-type: none"> • Inability to open a simple application such as Ms. Word and inability to understand the words read by the screen reader due to the foreign accent
Participants/ students	<ul style="list-style-type: none"> • Inability to understand the word read by the screen reader, especially when a word spelled using local language (Malay language) is read • Unable to figure out a pop-up blocker and inability to open some simple applications. Express the need of some other additional technology

The second presented information in Table 87.4, points out the observed nature or characteristics of the key persons during both interview session and observation session.

The facial reaction and the body language of the participants were analyzed carefully to derive the below emotions as stated in the Table 87.4. The observation reveals that the instructor’s emotion is divided into both positive and negative while the emotions of the students is mainly negative. This can be due to the reason that the instructor is an advanced user while the students are just the beginners. Based on Tables 87.3 and 87.4, the main factors contributing to the “inaccessible” elements for the blinds are identified.

The main factors contributing more to the “inaccessible” elements for the blinds are arranged to the most frequently encountered to the less frequently encountered by the users and also according to the most frustrating to the less frustrating emotion felt by the users:

- *Inability to open the right application.*
- *Blocks by the pop-up blocker.*
- *Having to listen repetitively the same information when missed out for the first time.*
- *Inability to know the actual content of a graphic.*
- *Inability to understand the foreign accent.*

Table 87.4 Observed analysis of key person

Role	Observed nature
Instructors	<ul style="list-style-type: none"> • Appears calm and confident in moving around the applications and able to confidently open some computer application and demo the class activities • Seems a little impatient while waiting for the screen reader to read each app and seems unhappy when a screen reader reads out a displayed graphic as “graphic”. Appears not happy when a pop up blocker blocks their input
Participants/ students	<ul style="list-style-type: none"> • Appears nervous and diffident when trying to open an app and appears frustrated when unable to open the right apps or when there is a pop blocker block the screen. Appears blank when a screen reader reads out a displayed graphic as “graphic” • Looks annoyed when unable to understand the word read by the screen reader and having to listen again the same word in the attempt to understand

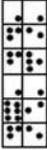
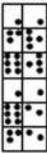
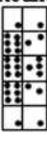
87.7.3 Results of Hands-on Activities

Two types of hands on activities were conducted with four players. Each player was given about 50 blocks of polystyrene. For the first hands-on activity (arrange the blocks activity), a list of applications names was read out and the players find the blocks that represent the particular application and arranged it in front of them. The second hands-on activity (Guess the blocks activity), blocks representing an application are presented in front of the players and they feel the blocks and guessed the application names. All the participants were given only 5–10 min to learn the representation of blocks pattern for each application.

Referring to Table 87.5, three rounds of arranging the blocks activity were conducted. At each round, five to six applications’ names were read out and the players searched through the blocks given to them and arranged the correct representation of the apps. Comparison between the arranged blocks by the players and the actual blocks (as shown in Table 87.5) shows that the players are able to arrange most of the blocks correctly. Referring to Table 87.6, player 1 and player 3 are presented with blocks of description (a) and player 2 and player 4 is presented with blocks of description (b). Based on the observed result, all players are able to correctly guess the presented blocks except for player 2 as player 2 is someone who has become blind just recently and trying to adapt to blindness. Throughout this activity, the players expressed their opinion by pointing out that having Braille blocks to represent applications and pop-up blockers truly facilitates them in learning computer skills.

Most of the previous research on assistive technologies is focused on the problems faced by the blind in using a screen reader thus not highlighting the inaccessibility of computer application to them [14, 18]. Some other findings relating to touch sensation were lacking of empirical study, hence not supporting the contribution of touch sensation in learning computer application [14, 17, 20].

Table 87.5 Arrange the blocks activity

Desc/ Apps	Player 1 & Player 2	Player 3 & Player 4
<p>1st Round</p> 		
<p>2nd Round</p> 		
<p>3rd Round</p> 		

Throughout this activity, the players or better known as the blind participants seems more confident and excited in using the blocks to represent computer applications. They have also expressed their opinion by pointing out that having braille blocks to represent applications and pop-up blockers truly facilitates them in learning computer skills. Having only screen readers as their assistive technology in learning computer seems complex for the blinds. Combination of screen reader and braille blocks will definitely help the blinds in learning the applications faster and in an easier manner.

Table 87.6 Guess the blocks activity

Desc/apps and icon names	Player 1 (a)	Player 2 (b)	Player 3 (c)	Player 4 (d)
(a) Internet explorer	√	√	√	√
(b) Pop-up blocker				
(a) File	√	√	√	√
(b) Ms. Word				
(a) Pop-up blocker	√	√	√	√
(b) File				
(a) Ms. Word	√	√	√	√
(b) Internet explorer				
(a) Folder	√	X	√	√
(b) Ms. Excel				
(a) Ms. Excel	√	√	√	√
(b) Pop-up blocker				
(a) Pop-up blocker	√	X	√	√
(b) Folder				
(a) Internet explorer	√	√	√	√
(b) Ms. Word				
(a) Ms. Word	√	X	√	√
(b) File				
(a) File	√	X	√	√
(b) Internet explorer				

√ Able to guess correctly X Unable to guess correctly

87.8 Conclusion and Future Work

This exploratory research investigated the issues and problems faced by the blinds society by conducting a series of interviews, recorded observations and hands on activities. Challenges faced by the blinds in performing a simple task such as opening an application in the computer has been highlighted. Views from various key personnel such as chairman, instructors and blind participants from the blind association has taken into consideration in concluding the problems of the blinds which shows that the usual representation of computer accessibility is not suitable for the blind IT users. And the analysis of the collected result has shown that the touch sensation of a blind can be used in identification of an object. In this case, braille blocks were used as the object. Thus, representation of a computer application using a braille blocks improves the learning process of a blind person. The hands on activity has also evident that touch sensation is an effective alert mechanism for the blinds in navigating through computer applications.

Hence, this study discovers the mental model of a blind user as an important contribution in improving the accessibility of computer applications. As a future work, the formulation of a usable model representing the mental model of the blind

users will be focused onto. This model will be established to overcome the identified factors which contribute to the “inaccessible” element of an application for the blind. It is foreseen that a developer may use the model as a system development guide and as well as a system assessment guide in future.

References

1. http://en.wikipedia.org/wiki/Computer_accessibility—Accessed 30/09/2013/
2. Ferreira, S.B.L., Nunes, R.R., da Silva, D.S.: Aligning usability requirements with the accessibility guideline focusing on the Visually-Impaired. In: 4th International Conference on Software Development for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2012), Elsevier (2012)
3. Queiroz, M.A.:—Bengala legal—<http://www.bengalalegal.com/>—Access 2/4/2008
4. World Health Organization, Global data on visual impairments 2010, WHO/NMH/PBD/12.01
5. Vaas, L.: Web blind spots. eWeek.com. Retrieved October 24, 2012, (2000) from <http://www.eweek.com/c/a/Web-Services-Web-20-and-SOA/Web-Blind-Spots/1/>
6. Sulaiman, S., Hasbullah, H.: Siti Nur Syazana Mat Saei, Mental model of blind users to assist designers in system development. In: ITSIM (2010)
7. Potosnak, K.: Human factor: mental model: helping users understand software (1989)
8. Mental Models and Usability, Retrieved on December 20, 2011 from: <http://www.lauradove.info/reports/mental%20models.htm>
9. Kurniawan, S.H., Sutcliffe, A.G.: Mental models of blind users in the windows environment. In: Proceedings ICCHP, pp. 568–574 (2002)
10. Roth, S.P., Schmutz, P., Pauwels, S.L., Bargas-Avila, J.A., Opwis, K.: Mental models for web objects: where do users expect to find the most frequent objects in online shops, news portals, and company web pages? In: Elsevier (2009)
11. Moloney, K.P., Jacko, J.A., Vidakovic, B., Sainfort, F., Leonard, V.K.E. Shi, B.: Leveraging data complexity: pupillary behavior of older adults with visual impairment during HCI. In: ACM transaction on computer-human interaction, pp. 376–402 (2006)
12. Takagi, H., Asakawa, C., Fukuda, K., Maeda, J.: Accessibility designer: visualizing usability for the blinds. In: Proceedings of the ACM SIGACCESS Conference on Computers and Accessibility, pp. 22–26, Atlanta (2004)
13. Nielsen, J.: Usability 101: Introduction to Usability by Nielsen Norman Group, <http://www.nngroup.com>, 4th Jan 2012—Accessed 02/10/2013
14. Abidin, A.H.Z., Xie, H., Wai, Wong, K.W.: Blind users' mental model of web page using touch screen augmented with audio feedback. In: International Conference on Computer and Information Science (ICCIS) (2012)
15. Takagi, H., Saito, S., Fukuda, K., Asakawa, C.: Analysis of Navigability of web applications for improving blind usability. In: ACM Transaction on Computer-Human Interaction, vol. 14 (2007)
16. Alonso, F., Fuertes, J.L., González, Á.L., Martínez, L.: A framework for blind user interfacing. In: ICCHP pp. 1031–1038 (2006)
17. Nishino, H., Goto, R., Kagawa, T., Yoshida, K., Utsumiya, K., Hirooka, J., Osada, T., Nagatomo, N., Aoki, E.: A touch screen interface design with tactile feedback. In: International Conference on Complex, Intelligent, and Software Intensive Systems, IEEE (2011)
18. Lumberas, M., Rossi, G.: A metaphor for the visually impaired: browsing information in a 3D auditory environment. In: CHI'95 (1995)
19. Shinohara, K.: Designing assistive technology for blind users. In: ASSETS (2006)

20. Petit, G., Dufresne, A., Levesque, V., Hayward, V., Trudeau, N.: Refreshable tactile graphics applied to schoolbook illustration for students with visual impairment. In: ASSETS'08 (2008)
21. Kumar, M., Talib, S.A., Ramayah, T.: Business Research Methods. Oxford University Press, Oxford (2012)

Chapter 88

Case-Based Reasoning and Profiling System for Learning Mathematics (CBR-PROMATH)

Nur Azlina Mohamed Mokmin and Mona Masood

Abstract This paper discusses the architecture of a case-based reasoning profiling system for learning mathematics (CBR-PROMATH). The adaptive system has the ability to suggest suitable learning materials based on previous cases of learner profiles and individual learning styles. The developed learning materials use a learning tool which consists of so-called mastery, understanding, interpersonal and self-expressive styles. Two sets of experiments were carried out to test the system's functionality. The first consisted of 10 sets of learners' profile cases, stored previously in the database. The second presented the system with 10 real new cases. The system compared and calculated similarity values between the new and stored cases. The learning material that was most similar was presented as a solution for the new case. The experiment showed that the CBR algorithm was successfully applied in the development of the CBR-PROMATH.

Keywords Case-based reasoning · Mathematics · Mathematics learning styles

88.1 Introduction

Mathematics is one of the basic pillars of scientific progress. Therefore, mathematical literacy is crucial in building the foundation of a more complex skill [1]. Hodgen and Marks suggest that the ability to benefit from higher education and then to be able to play a productive role in the workplace depends on the mathematical competence of the individual [2]. The lack of mathematical literacy will

N.A.M. Mokmin (✉) · M. Masood
Center for Instructional Technology and Multimedia, Universiti Sains Malaysia,
11800 Pulau Pinang, Malaysia
e-mail: namm12_tpm010@student.usm.my

M. Masood
e-mail: msmona@usm.my

subsequently cause incorrect application or interpretation of mathematics especially in the science and engineering fields [3, 4].

One reason why mathematic records poor achievement is because of the failure to differentiate between the learning styles of students [5–8]. A student's learning style is the method s/he uses to understand a lesson [9]. Some students prefer visual diagrams and pictures, whereas others prefer the hands-on approach. Educators must be able to identify the student's preferred learning style to ensure the effectiveness of the teaching and learning process [10].

Hundreds of applications have been created in the field of computer science to assist students with their lessons. The advancement of technology in education has led to the development of educational software that applies various learning style models [11–13]. Alves believes that using case-based reasoning (CBR) can support educators to design appropriate learning activities to enhance students' learning experience [14]. Therefore, this paper will first discuss the application of CBR and related past research as well as the algorithm for developing an intelligent profiling system that can suggest a learning material based on a specific student profile. The architecture of the CBR profiling system for learning mathematics (CBR-PROMATH), students' profile, and the learning engine are elaborated. In addition, the results of an experiment on the system's functionality are also highlighted.

88.1.1 Case-Based Reasoning (CBR)

Artificial intelligence (AI) is the computer's attempt at imitating human thoughts and responses. The adoption of AI in education started at the end of the 1980s, in the form of autonomous agents, intelligent tutoring systems, and educational theories [14]. Case-based reasoning (CBR) is a problem-solving AI paradigm that enables utilization of specific knowledge and experience [15, 16]. The field of CBR is based on cognitive science. It focuses on how humans generate hypotheses on new situations based on past experiences [17].

According to the CBR theory, an encountered problem (the new case) prompts the individual to retrieve cases from memory, and reuse the old case, which then suggests a solution. If the new solution is found to be effective, the knowledge is then stored in memory for later use. Embedded within each case is a series of indices which aids memory retrieval. If, however, the attempt to solve the problem fails, the algorithm will find reasons for the failure and store the result to avoid repeating similar mistakes [16, 18].

A case indicates a problem situation. Figure 88.1 show the CBR cycle. When a new case is presented to the system, it will go through a CBR cycle starting from establishment as a new case and ending when it has been solved by the system. The problem-solving lifecycle in a CBR system consists of the following four stages [19];

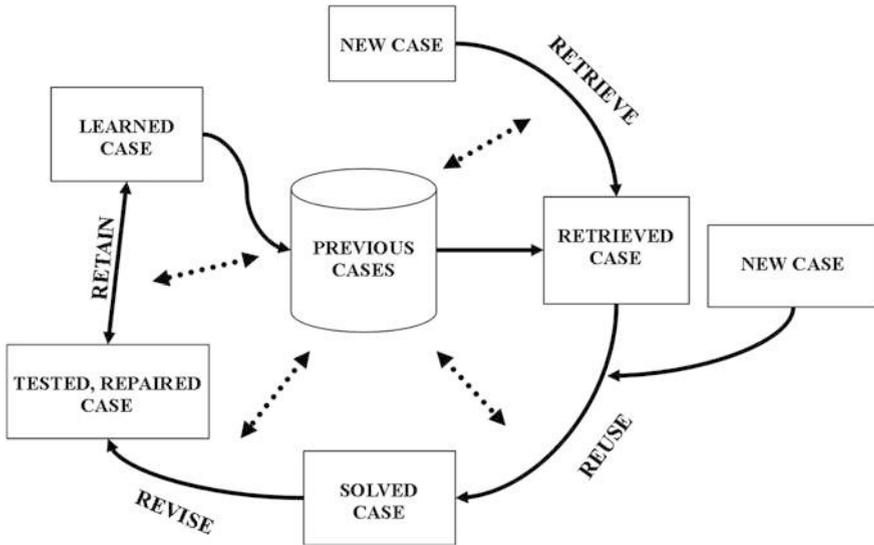


Fig. 88.1 The CBR cycle [18]

- Retrieve: retrieving similar cases from the existing knowledge base whose problem is identified as similar to a given new case
- Reuse: reusing the solutions from the retrieved cases for the new case
- Revise: revising the solutions retrieved in an attempt to solve the new case
- Retain: retaining the new case once it has been solved.

88.1.2 Previous Studies

Many systems are developed for educational purposes; nonetheless, they commonly present the same materials without specifically considering individual student differences. Conversely, many researchers in the education field have developed systems that cater for differences in the learning process. These systems have adopted the different learning styles identified.

Yang and Yen propose a framework for multi-agent CBR (MACBR) for personalized recommendation of e-learning resources [20]. This system gives personalized recommendation to the learner by taking account of the learner's characteristics. It works by noting students' learning requirements, retrieving related cases and submitting the case to the similarity calculator agents. The system then calculates the degree of similarity between the target case and the related cases and seeks a case similar to the student's requirements by identifying similar values.

An adaptive web-based educational system named Domus was developed in [21] that applied CBR and fuzzy logic to adapt e-learning contents and contexts according to student learning styles and individual needs. The Domus system works by searching in the database for similar cases based on student learning style, assuming that students with similar profiles have similar difficulties in the learning process. Domus functions by providing adapted learning, recommending the accomplishment of certain learning activities, collecting opinions about a subject, suggesting Web resources, alerting the student to events, and supporting agenda management.

In [22] a personalized learning system was constructed by means of a genetic algorithm and a CBR approach. In this personalized e-learning system, CBR worked by providing a summative assessment for each individual learner after several units of instruction. The research considered the curriculum's difficulty level and the continuity of successive curricula in mastery learning.

Most adaptive applications apply a CBR algorithm developed for e-learning content but very few applications are developed exclusively for mathematics learning. An adaptive mathematics learning application must include mathematics learning styles and mathematics achievements in the learning engine. Therefore, the current study uses the mathematics learning style recommended by [23] in the design and development process.

88.1.3 Similarity and Weighting in CBR

The learning engine uses outcomes to analyze similarities of a given problem in order to re-weight case features. A similarity measure is a critical component in any CBR system [24, 25]. Similarity measures were described in [26] as a means to compare two cases. Local similarity measures are similar on the feature level and global similarity measures are similar on the case or object level.

88.2 Architecture of Case-Based Reasoning Profiling System for Learning Mathematics (CBR-PROMATH)

CBR-PROMATH is a web-based system programmed in Hypertext Preprocessor (PHP), a server-side scripting language designed not only for web development but also as a general-purpose programming language. The related database management system applied in this system is MySQL, which is based on Structured Query Language (SQL). MySQL and PHP are both free for download and use because they are open-sourced. The PHP and MySQL combination has become a popular

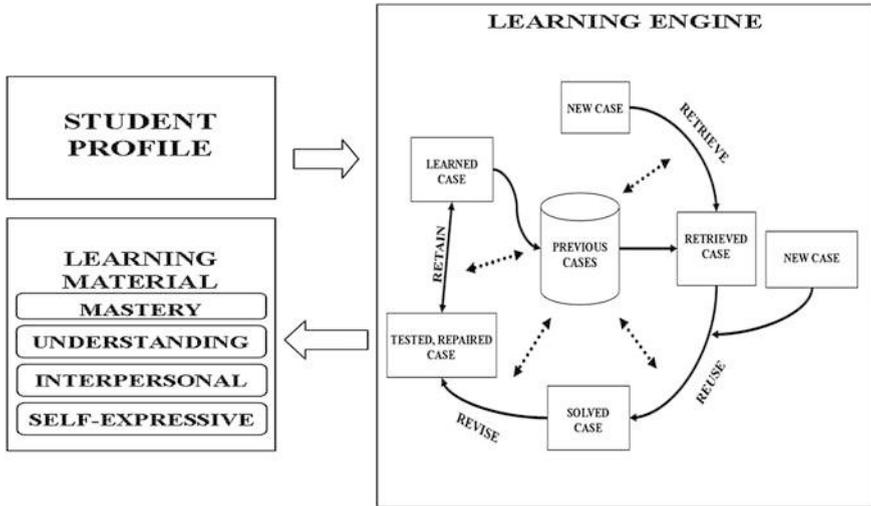


Fig. 88.2 The CBR-PROMATH architecture

choice for database-driven websites. In contrast, the learning materials are developed from multimedia software.

The main characteristic of CBR-PROMATH is that it can give suggestions on the appropriate learning materials based on learning styles and student profiles. The system has three basic components; the student profile database, the learning engine and the learning material database (Fig. 88.2). These three components interact to adapt the different cases (problems) submitted by the user to the system. These submitted data are transformed by the learning engine into variables that are processed as information before the best learning material is suggested.

88.2.1 Student Profile

The student’s profile consists of information such as their identity card, the Malaysian Certificate of Education or “Sijil Pelajaran Malaysia” (SPM) mathematics result, learning styles and test results. The student’s profile is where information on the learner is obtained and processed into variables by the system. This information is then used by the learning engine to compute similarity values. Figure 88.3 shows the graphical user interface (GUI) for the student’s profile. Figure 88.4 shows the set of data that is stored in the system in the form of cases for consultation by the learning engine.

The screenshot shows a web form for a student's profile. It consists of four horizontal input sections stacked vertically, each with a label and a corresponding input field. The first section is labeled 'ID:' and has a text input box. The second section is labeled 'MATHEMATICS SPM RESULT' and has a dropdown menu showing 'A'. The third section is labeled 'LEARNING STYLE' and has a dropdown menu showing 'MASTERY'. The fourth section is labeled 'TEST RESULT' and has a dropdown menu showing 'A'. Below these sections is a 'Submit' button.

Fig. 88.3 GUI of the student's profile

The screenshot shows a table with 11 rows of data. The first row is the header, and the following 10 rows represent individual student cases. Each row contains five columns of data: ID, Learning Style, SPM, Result, and Material. The data is as follows:

[id]	[LEARNING STYLE]	[SPM]	[RESULT]	[MATERIAL]
[AHMAD]	[INTERPERSONAL]	[D]	[D]	[MATERIALINTERPERSONAL]
[AIDA332]	[UNDERSTANDING]	[B]	[C]	[MATERIALUNDERSTANDING]
[AILING77]	[MASTERY]	[B]	[A]	[LEARNINGMASTERY]
[ASIYAH22]	[SELF-EXPRESSIVE]	[C]	[C]	[MATERIALSELFEXPRESSIVE]
[gopal]	[MASTERY]	[A]	[A]	[MATERIALMASTERY]
[HAIDA55]	[UNDERSTANDING]	[A]	[B]	[MATERIALUNDERSTANDING]
[JUNE88]	[INTERPERSONAL]	[C]	[D]	[MATERIALINTERPERSONAL]
[LAN33]	[SEFL-EXPRESSIVE]	[C]	[D]	[MATERIALSELFEXPRESSIVE]
[SITI2]	[UNDERSTANDING]	[B]	[C]	[MATERIALUNDERSTANDING]
[SITIAHMAD89]	[INTERPERSONAL]	[D]	[D]	[MATERIALINTERPERSONAL]

Fig. 88.4 Set of stored cases

88.2.2 Learning Engine

The learning engine processes the information keyed in by the user, and converts the information into variables for calculation by the system. When a new case is presented to the system, the CBR algorithm calculates the local similarity value and the global similarity value between each case for comparison with the new case. The case with the highest similarity value is selected. The learning materials suggested are used as a reference for the new case. Personalized learning materials can then be recommended to a student with a similar learning profile.

88.2.3 *Learning Materials*

In [27] it was found that students who learn with adaptive learning materials have significantly better learning achievement than those who learn with non-adaptive materials. Providing students with learning materials and personalized recommendations in terms of their preferred learning methods can make learning easier for them [11]. Researchers involved in the development of personalized and adaptive learning materials include learning styles as important factors in their work [25, 28, 29]. It is important to diagnose the student's learning style because some students learn more effectively when taught according to their preferred method. Information about different learning styles can help the system become more sensitive towards students who use the system. As stated by [30], a student's learning style represents the type of learner they become.

In the development of the CBR-PROMATH, the mathematics learning style that was suggested by [23] was used in the design of the learning materials. The learning materials have four distinct styles of instruction:

- Mastery learning style (MLS) that emphasizes skill acquisition and retention of critical mathematical terms.
- Understanding learning style (ULS) that builds student's capacities to find patterns and explain mathematical concepts.
- Self-expressive learning style (SLS) that capitalizes on students' powers of imagination and creativity
- Interpersonal learning Style (ILS) that invites students to find personal meaning in mathematics by working together as part of a community of problem-solvers.

The materials suggested by the system are from the learning engine. At the end of every learning material presented, marks or grades for every assessment are stored in the database. A low mark indicates the unsuitability of the learning materials suggested, in which case the student is presented with the second learning material which has the second-highest similarity value. If the result is a pass, the learning material suggested will be used as a reference for the next case. Figures 88.5 and 88.6 show an example of the learning material that has been developed.

88.2.4 *Experiments*

In order to test the system's functionality, a series of experiments was carried out. In particular, it was set up to see whether the system is capable of carrying out effective calculations when given a new case. Table 88.1 shows the set of new cases submitted, learning materials suggested and the case referred to previously. The experiments and calculation values given by the system suggest that CBR-PROMATH successfully referred the cases with the highest similarity value to the new case.

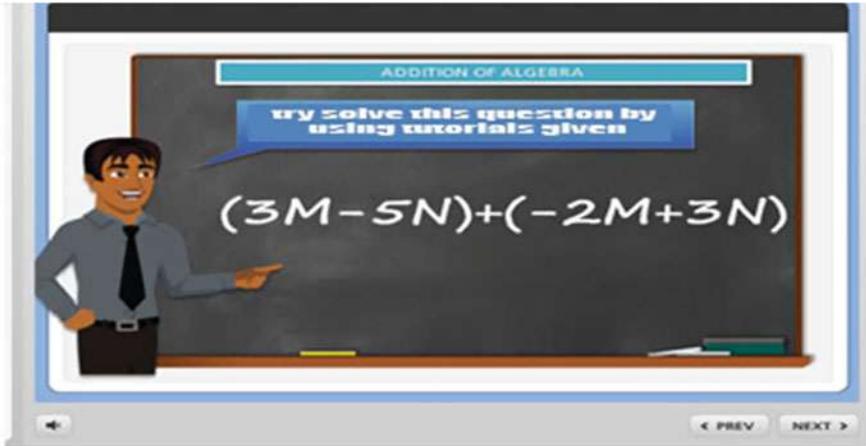


Fig. 88.5 Example of learning material 1

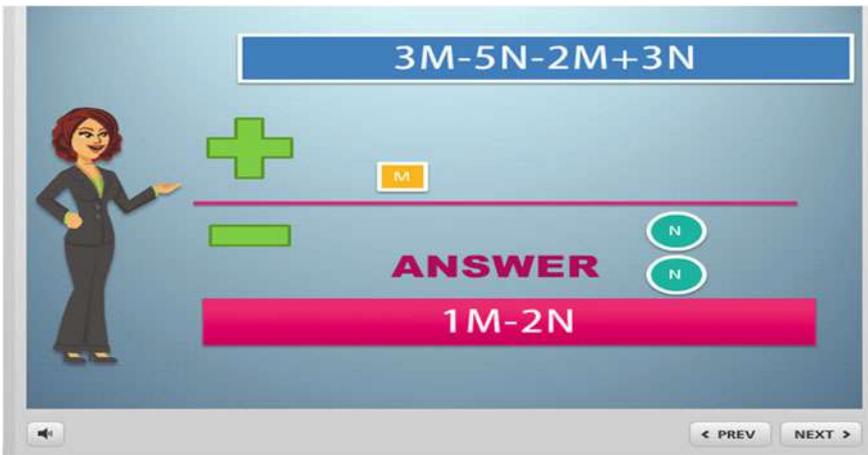


Fig. 88.6 Example of learning material 2

The suggested learning materials in Table 88.1 are the result of the similarity value of each stored case in the database. In the table, Case Id Aida332 has the highest similarity in terms of value to the New Case Id Nora12; therefore the learning material for Aida332 is used as a learning material suggestion for New Id Nora12.

Table 88.1 New cases, learning material suggestions and reference case

New case Id	Learning style	Mathematics result	Test result	Reference to case Id in database	Learning material suggestion
Nora12	ULS	A	C	Aida332	Understanding
Akila225	ILS	D	D	Ahmad	Interpersonal
Danny67	SLS	C	B	Asiyah22	Self-Expressive
Johnny89	MLS	B	A	Ailing77	Mastery
Sitihajar34	ILS	C	D	June88	Interpersonal
Kamal44	ILS	B	C	Aida332	Understanding
Badri66	ILS	D	B	Ahmad	Interpersonal
Haikal33	SLS	C	B	Asiyah22	Self-Expressive
Fiza78	ULS	C	A	Ailing77	Mastery
Alia99	ILS	C	B	Asiyah22	Self-Expressive

88.2.5 Conclusion and Future Work

This work shows that the design and development of CBR-PROMATH can suggest suitable learning materials for the user based on the cases submitted to the system. In case-based reasoning, a reasoning engine remembers a previous situation similar to the current one and uses that to solve the new problem [15]. The learning materials that are developed are of value to math educators and researchers. A useful suggestion for future research is the assessment of all learning materials suggested by the system in terms of their effectiveness in the teaching and learning process.

References

1. Kashefi, H. et al.: Creative problem solving in engineering mathematics through computer-based tools. In: 2nd International Seminar on Quality and Affordable Education, pp. 207–211 (2013)
2. Hodgen, J., Marks, R.: The employment equation: why our young people need more maths for today's jobs. The Sutton Trust, London (2013) [Online]. Available: <http://www.suttontrust.com/news/publications/the-employment-equation-why-our-young-people-need-more-maths/>
3. Haider, H., et al.: How we use what we learn in math: an integrative account of the development of commutatively frontline learning research I. *Frontline Learn. Res.* **2**(1), 1–21 (2014)
4. Hong, K.S., et al.: Status of mathematics teaching and learning in Malaysia. *Int. J. Math. Educ. Sci. Technol.* **40**(1), 59–72 (2009)
5. Aral, A., Cataltepe, Z.: Learning styles for K-12 mathematics e-Learning. In: CSEDU 2012 - 4th International Conference on Computer Supported Education, pp. 317–322 (2012)
6. Middleton, K. et al.: Examining the Relationship between Learning Style Preferences and Attitudes Toward Mathematics Among Students in Higher Education. *Institute for Learning Styles Journal*, Howard University, vol. 1 (2013)

7. Havola, L.: Assessment and learning styles in engineering mathematics education. Licentiate thesis, Aalto University, Espoo, Finland (2012)
8. Murat, G.: The effect of students' learning styles on their academic success. *Creative Educ.* **4**(10), 627–632 (2013)
9. Mohamad, B., Hashim, I.: *Gaya Pengajaran Dan Pembelajaran*. PTS Professional Publishing Sdn. Bhd, Kuala Lumpur (2010)
10. Adnan, M., et al.: Learning style and mathematics achievement among high performance school students. *World Appl. Sci. J.* **28**(3), 392–399 (2013)
11. Kinshuk, S.G.: Providing adaptive courses in learning management systems with respect to learning styles. In: *Proceedings of the World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (e-Learn)*, pp. 2576–2583 (2007)
12. Mahnane, L., et al.: A model of adaptive e-learning hypermedia system based on thinking and learning styles. *Int. J. Multimedia Ubiquitous Eng.* **8**(3), 339–350 (2013)
13. Sarrafzadeh, A., et al.: How do you know that i don't understand? a look at the future of intelligent tutoring systems. *Comput. Hum. Behav.* **24**(4), 1342–1363 (2008)
14. Alves, P.: Advances in Artificial Intelligence to Model Student-Centred VLEs, *Advances in Learning Processes, INTECH*, pp. 113–124 (2010)
15. Kolodner, J.L.: An introduction to case-based reasoning. *Artif. Intell. Rev.* **6**, 3–34 (1992)
16. Andrew, A.T., Keene, C.W.: Applying case-based reasoning theory to support problem-based learning. *J. Appl. Instr. Des.* **3**(2), 31–40 (2013)
17. Shiu, S.C., Pal, S.K.: Case-based reasoning: concepts, features and soft computing. *Appl. Intell.* **21**(3), 233–238 (2004)
18. Aamodt, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun.* **7**(1), 39–59 (1994)
19. Swanson, R., Gordon, A.S.: Say anything: using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2**(3), 16 (2012)
20. Yang, L., Yan, Z.: Personalized recommendation for learning resources based on case reasoning agents. *Electr. Control Eng. (ICECE)* **3**, 6689–6692 (2011)
21. Alves P. et al.: Case-based reasoning approach to adaptive web-based educational systems. In: *ICALT '08: Proceedings of the 2008 Eighth IEEE Int. Conference on Advanced Learning Technologies*, pp. 260–261 (2008)
22. Huang, M., et al.: Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Syst. Appl.* **33**(3), 551–564 (2007)
23. Silver, H.F., et al.: *Math Tools, Grade 3-12; 60 + Ways to Build Mathematical Practices, Differentiate Instruction, and Increase Student Engagement*, 2nd edn. Corwin, Thousand Oaks (2012)
24. Long et al. J.: Adaptive similarity metrics in case-based reasoning. In: *Proceedings of Intelligent Systems and Control, Honolulu, Hawaii, USA* (2004)
25. Soh et al. W.M.: Implicit detection of learning styles—the smart way. In: *Proceedings of the 9th International CDIO Conference, Massachusetts Institute of Technology and Harvard University School of Engineering and Applied Sciences, Cambridge, USA* (2013)
26. Bergmann, R., et al.: Representation in case-based reasoning. *Knowl. Eng. Rev.* **20**(3), 209–213 (2005)
27. Hwang, G.J., et al.: A learning style perspective to investigate the necessity of developing adaptive learning systems. *Educ. Technol. Soc.* **16**(2), 188–197 (2013)
28. Anthony, P., et al.: Learning how to program in c using adaptive hypermedia system. *Int. J. Inf. Educ. Technol.* **3**(2), 151–155 (2013)
29. Sudhana, K.M., et al.: An architectural model for context aware adaptive delivery of learning material. *Int. J. Ad. Comput. Sci. Appl.* **4**(10), 80–87 (2013)
30. Mohamad, M.M., et al.: Disparity of learning styles and cognitive abilities in vocational education. *Int. J. Soc. Hum. Sci. Eng.* **8**(1), 6–9 (2014)

Chapter 89

What Is the Influence of Users' Characteristics on Their Ability to Detect Phishing Emails?

Ibrahim Alseadoon, M.F.I. Othman and Taizan Chan

Abstract Phishing emails cause significant losses for organisations and victims. To fight back against phishing emails, victims' detection behaviours must be identified and improved. Then, the impact of victims' characteristics on their detection behaviours must be measured. Three methods, namely experiments, surveys and semi-structured interviews, were applied in our research to gain a richer understanding of victims' behaviours concerning phishing emails. Several user characteristics were measured using the deception detection model. The results suggest that users' characteristics either increase or decrease their suspicions. Characteristics such as user extraversion, trust and submissiveness represent variables that prevent victims from suspecting phishing emails. In contrast, email experience increases victims' suspicion of phishing emails. Furthermore, victims' personality characteristics and a variable called the susceptibility variable play important roles in increasing the tendency of victims to execute the actions requested in phishing emails.

89.1 Introduction

Phishing emails are designed to imitate legitimate entities, causing users to trust in their validity. Then, users are asked to perform an action included in the phishing email, which is most likely connected to a fake website, in order to obtain victims' sensitive information. A phishing attack has two aspects: the first is the way in which the attacker reaches his or her victims; the second depends on the victims themselves and whether they choose to respond to these emails [1–3]. Securing users against phishing emails begins with the use of technology to detect such emails before they reach users [4–6]. However, the perpetrators of phishing emails

I. Alseadoon (✉) · M.F.I. Othman · T. Chan
Queensland University of Technology, Brisbane, QLD 4000, Australia
e-mail: ialseadoon@gmail.com

are always finding new ways to enable these emails to reach users [7, 8]. Undetected phishing emails make users the last defence line of defence against these emails.

The number of phishing websites detected by the Anti-Phishing Working Group in April of 2012 was 63,253, which is believed to be the highest number ever recorded [9]. The financial losses incurred due to phishing attacks are considerable. These losses can vary from \$61 million to approximately \$3 billion per year [10, 11]. Increased investigation and prevention of phishing attacks must occur to minimise these losses.

The purpose of our research is to identify users' detection behaviours,¹ as well as user characteristics that impact their detection behaviours. Several user characteristics have been identified and measured using the deception detection model proposed by Grazioli in 2004. If victims are found to have individual characteristics that predisposed them to falling victim to deception, then knowing this will aid in identifying other individuals who are at risk.

The paper is organised as follows: the next section describes relevant research on phishing emails. Then, the theoretical framework section introduces the deception detection model used and user characteristics measured using the model. The methodology section details the steps taken to conduct our research. The findings are described in the results section. Finally, the discussion and conclusion sections conclude this paper.

89.2 Related Work

Education programs have been provided to increase users' knowledge about phishing emails. Such programs have increased the rate at which users become detectors by 40 % [12]. Educating users on the design features of phishing emails has helped reduce the rate at which users fall victim to phishing emails [13–16]. The main drawback of phishing email education is that some users who received education programs still became victims [7, 8]. This suggested that we should attempt to look for weaknesses that predisposed users to becoming victims.

Security tools have been provided to warn users about phishing attempts. In such studies, the final decision concerning the validity of an email is made by the user, but this is done with the help of security tools [17, 18]. The drawback of these approaches is that they are limited to users' discretion; one study revealed that users often ignore these kinds of warnings and thus still fall victim to the deceit [19, 20]. To help prevent users from becoming victims, we must understand users' detection behaviours and which user characteristics influence their detection behaviours [21].

In other words, the main focus in phishing email studies has been improving users' knowledge about phishing emails and developing new security tools. Unfortunately, user characteristics and detection behaviours have not been well-investigated [21].

¹ Users' detection behaviours are the behaviours that users perform to detect phishing emails, which include susceptibility, confirmation and response.

89.3 Theoretical Framework

The theory of deception explains the process by which users detect deception [22, 23]. According to this theory, detection is a cognitive process that involves examining various cues (e.g., words, tone and body language). Detectors look for cues that lead them to conclude that the received message is a deceptive message. Grazioli [24] applied this theory in a computer-based environment and found that the detection process goes through four phases (see Fig. 89.1). Interestingly, Grazioli found that detectors and victims both go through the four phases of the deception detection model. However, only detectors were able to detect deception, whereas victims failed to detect it [24]. In this research project, we aim to discover the reasons victims failed to successfully detect deceptive messages.

We applied Grazioli's model in order to understand users' detection behaviours when faced with phishing emails. From examining the Grazioli model, we were able to determine users' detection behaviours and divide them into three phases: susceptibility, confirmation and response (see Fig. 89.1). Additionally, our research is interested in finding the impact of users' characteristics on their detection behaviours (see Fig. 89.2), as well as the impact of the confirmation channel.

89.3.1 Variables Related to Users' Detection Behaviours

These variables can be divided into three main variables, as explained in the previous section: susceptibility, confirmation and response.

89.3.1.1 Response

Response is the last behavior, and it indicates that users have responded to the phishing email. Ultimately, users were classified as either detectors or victims. Victims were those who chose to perform the action in the phishing emails. Detectors were those users who saw the phishing email and chose to ignore it.

89.3.1.2 Confirmation

Users who become doubtful about a phishing email generate hypotheses and evaluate them. Hypothesis generation is a mental process that is not measured in this research. On the other hand, hypothesis evaluation is a process that can be captured by examining the ways in which users choose to act. For example, when questioning the legitimacy of an email, some users contact organisations by phone, while others email their friends. Confirmation was measured by asking users to self-report the type of confirmation channels they used. Confirmation channels are

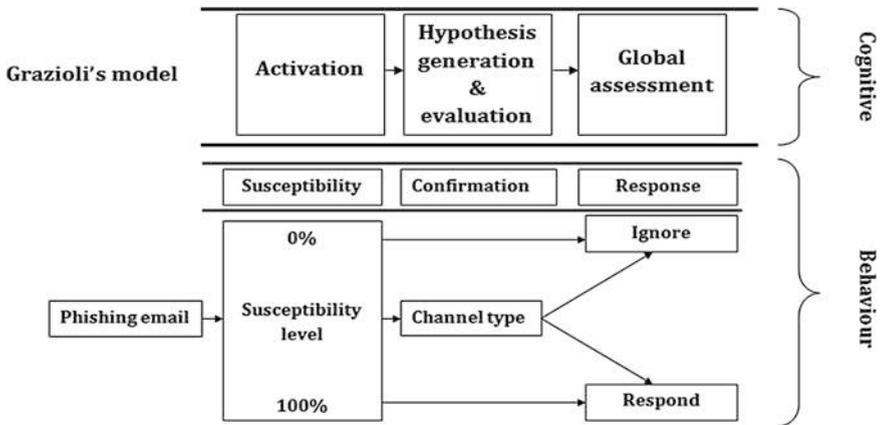


Fig. 89.1 Users' detection, cognitive and behaviour processes

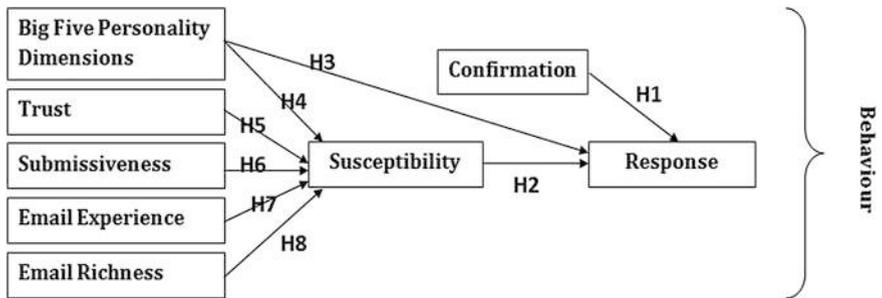


Fig. 89.2 Research model

classified based on their richness (face-to-face, telephone, email or self-investigation). Face-to-face is considered to be a highly rich medium [25].

H1. Use of a highly rich confirmation channel decreases user response rates to phishing emails.

89.3.1.3 Susceptibility

Susceptibility is the first phase in user detection behaviour. Users begin their detection behaviour by suspecting the phishing email. Based on the level of susceptibility, users will decide upon their intended behaviour (see Fig. 89.1). To measure users' susceptibility, five phishing emails were presented to participants in the form of a picture. Participants were asked to play the role of someone called John and were told that John had received these five emails. Participants were asked to rate the likelihood of responding to each of the five emails, with 1 being 'definitely

would ignore the email' and 7 being 'definitely would respond to the email'. This variable was measured in the first survey, meaning that the participants had not been informed about the phishing email experiment or that these five emails were phishing emails. The aim was to capture users' normal responses to emails and to determine whether they were able to determine these emails to be phishing emails.

H2. A high level of susceptibility to phishing emails increases user response rates to phishing emails.

89.3.2 Related Users' Characteristics

Our research examines the impact of users' characteristics on their detection behaviours. Figure 89.2 illustrates the proposed model used in this research. The diagram below presents related user characteristics.

89.3.2.1 Big Five Personality Dimensions

The Big Five personality dimensions divide a user's personality into five main dimensions: (1) extraversion, which describes a person who is more interactive with others (2) agreeableness, which describes a person who is more kind and warm to others (3) conscientiousness, which describes a person who is more determined to finish tasks (4) emotional stability, which describes a person who is more likely to be calm and (5) openness, which describes a person who is more open to new experiences [26]. These dimensions have been suggested to influence users' susceptibility to phishing emails [27]. Our research investigates the impact of extraversion (H3a, H4a), agreeableness (H3b, H4b), conscientiousness (H3c, H4c), emotional stability (H3d, H4d) and openness (H3e, H4e) on users' detection behaviours. The measures used were developed by Gosling et al. [28], whose work contained ten items.

H3. Certain big five personality dimensions impact user response to phishing emails.

H4. Certain big five personality dimensions impact user susceptibility to phishing emails.

89.3.2.2 Trust

Trust measures users' tendency to trust others. One drawback of relying on victims to make decisions regarding phishing emails is their trust that these emails are legitimate. Our research aims to discover the impact of trust on users' susceptibility to phishing emails. Trust was measured using three items developed by McKnight et al. [29].

H5. Trust increases users' susceptibility to phishing emails.

89.3.2.3 Submissiveness

Wright et al. [30] performed a study on phishing emails, and based on interviews they conducted, they suggested that submissiveness has an effect on users becoming victims of phishing emails. The nature of phishing emails is such that they cause users to believe that they are in a situation that demands their compliance with the requests of the phishing emails. Wright et al. [30] indicate that users with a high level of submissiveness are different than others. Submissiveness is measured using 16 items developed by Allan and Gillbert [31].

H6. Submissiveness increases user susceptibility to phishing emails.

89.3.2.4 Email Experience and Email Richness

The theory of media richness supports the idea that users are more able to detect deception carried out via media with a high degree of richness as compared to media with a low degree of richness [25]. Richness is measured by the ability to reproduce the information carried by the medium. Because email is considered to have a low degree of richness, deception detection in emails tends to be difficult for users. However, the theory of channel expansion suggests that the more experience a user has with a medium, the more he or she will perceive the medium to be rich [32]. Therefore, our research will measure the impact of email experience and richness on user ability to detect deception via email. The measures include six items for email experience and four items for email richness developed by Carlson and Zmud [32].

H7. Email experience decreases user susceptibility to phishing emails.

H8. Email richness decreases user susceptibility to phishing emails.

89.4 Methodology

A mixed methods approach was applied by using a combination of qualitative and quantitative methods. The following subsections provide more details.

89.4.1 Subjects

The targeted participants were university undergraduate students. Aiming for this age group was crucial because it has been reported in the literature that students in this age group are more susceptible to phishing emails than those of older age groups [33, 34]. A total of 187 students out of 430 contacted provided complete answers to all the phases required by our research.

89.4.2 The Survey

The survey was designed to collect information regarding participants' characteristics and detection behaviors, which required us to divide the survey into two parts. The first survey was delivered before sending the phishing email to collect information regarding the participants' characteristics. The second survey was delivered after sending the phishing email to collect information regarding participants' detection behaviours.

89.4.2.1 The First Survey

The first survey, which did not inform participants about the experiment, was designed to collect information regarding participants' characteristics, such as the Big Five personality dimensions, trust, submissiveness, email experience and email richness. In this survey, participants were informed that the survey was designed to capture the impact of users' characteristics on their normal day-to-day email behaviours.

89.4.2.2 The Second Survey

This survey was designed to collect information regarding users' detection behaviors and recorded the type of confirmation channel used. The survey was sent to participants after a sufficient amount of time had passed (i.e., around two weeks) since their receipt of the phishing email. In this survey, the real goal of the experiment was revealed to the participants (i.e., a phishing email study).

89.4.3 The Phishing Email

The aim of sending the phishing email was to classify participants into two groups: detectors and victims. The phishing email was sent on behalf of the blog maintenance team. The body of the phishing email was designed as follows: the phishing email informed participants about a problem that occurred in a blog, which resulted in damaging the blog database that held students' account information. The phishing email emphasised that the students who had received this email were those who had been affected by this recent problem and that they needed to act quickly to resolve it. After students were given an explanation as to why they had received the email, they were prompted with a solution that for this purpose of this research, imitating a real phishing email design. Two solutions were provided; students were either asked to reply to an email address with the requested information or to click on a link embedded in the phishing email.

Both the reply email address and the link were hosted outside the university domain (on a malicious domain).

89.4.4 Interviews

After collecting information about participants' responses to the phishing email and the two surveys, participants were sent an invitation to be interviewed about their experiences with the phishing email. In these semi-structured interviews, participants were asked questions regarding their responses when they received the phishing email. The interviews helped to shed light on certain aspect of users' detection behaviours that could not be identified in the previous surveys.

89.5 Results

Participants were randomly divided into two main groups: those who received reply phishing emails and those receiving click phishing emails. The percentage of victims was 26 % of the whole population. The results suggest that phishing emails that ask about private information directly are less harmful than those that ask about private information indirectly. Two-thirds (69.5 %) of the 187 participants were under 26 years of age, and 77.5 % (N = 145) were male. Some 57.2 % (N = 107) spoke English as their first language, and 58.3 % (N = 109) identified their nationality as Australian.

We applied regression analysis to measure the relationships between variables. There were two dependent variables: susceptibility and response (see Fig. 89.2). Linear regression was applied to measure the dependent variable, susceptibility, and logistic regression was applied to measure the dependent variable, response. The results show that R-square is equal to 0.643 for the linear regression (susceptibility) and that R-square is equal to 0.492 for the logistic regression (response).

89.5.1 Susceptibility

The results of multiple regressions show that there are two positive, significant relationships: one between submissiveness and susceptibility and another between trust and susceptibility. There is also a negative, significant relationship between perceived email richness and susceptibility. The results show that submissiveness, trust and email richness explain 64 % of the variance in susceptibility (see Table 89.1).

Table 89.1 Summary of the hypotheses tested regarding susceptibility

Hypotheses	Variables	Significance	Conclusion
Hypothesis 5	+ Trust	$\beta = 0.475$ $P = 0.000$	Supported
Hypothesis 6	+ Submissiveness	$\beta = 0.190$ $P = 0.012$	Supported
Hypothesis 8	– Email Richness	$\beta = -0.199$ $P = 0.007$	Supported

Extraversion refers to a person who is more interactive. Due to the nature of extraverted users, they are more susceptible to phishing emails. Those users with high levels of trust are more likely to trust others. Phishing emails ask their victims to trust them (i.e., phishing emails impersonate trustworthy entities). Submissiveness measures users' obedience to others. Phishing emails directly ask their victims to perform an action. High levels of submissiveness will increase the chances of complying with the direct order received via phishing emails. Email experience's impact is opposite that of the previous three variables (extraversion, trust and submissiveness). High email experience decreases user's susceptibility to phishing emails. Email experience may improve the baseline for distinguishing between legitimate emails and illegitimate ones. Observing inconsistent cues is an essential aspect of making users suspect phishing emails, which can be a result of high levels of email experience. Finally, email richness has not been indicated to have any significant impact on users' susceptibility.

89.5.2 Response

The results of the linear regressions show that there are four positive, significant relationships between the Big Five personality dimensions, susceptibility and confirmation channel. There is also a negative, significant relationship between confirmation channel and response. The results show that the identified variables explain 49 % of the variance in responses (see Table 89.2).

As indicated in Table 89.2, susceptibility, openness, extraversion, agreeableness and confirmation channel predicted the response as an outcome at the $p < 0.05$ level. Those with higher susceptibility and openness scores were two and half times more likely than others (Exp (B)) to respond to the phishing emails. Those with higher extraversion scores were two times more likely than others (Exp (B)) to respond to the phishing emails. Those with higher agreeableness scores were four times more likely than others (Exp (B)) to respond to the phishing emails. Those with higher confirmation channel scores were one time more likely than others (Exp (B)) to detect phishing emails.

Phishing emails ask their victims to interact with them and comply with their requests, which may be the reason that extraversion and agreeableness significantly impacting users' responses to phishing emails. Openness describes users who are more open to new experience and take more risks. Phishing emails are a risk for users who comply with their requests. It can be suggested that openness

Table 89.2 Summary of hypotheses tested with responses

Hypothesis	Variables	Significance	Conclusion
Hypothesis 3a	+ Extraversion	B = 0.540 P = 0.002	Supported
Hypothesis 3b	+ Agreeableness	B = 1.289 P = 0.000	Supported
Hypothesis 3e	+ Openness	B = 0.859 P = 0.001	Supported
Hypothesis 2	+ Susceptibility	B = 0.914 P = 0.000	Supported
Hypothesis 1	Confirmation	B = -0.432 P = 0.031	Supported

increases users' tendency to take the risks included in phishing emails. High levels of susceptibility will decrease users' suspicion levels, which will increase users' vulnerability. Finally, the type of confirmation channel did not have any significant effect on users being victims.

89.6 Discussion

From the results above, it can be seen that users' characteristics play an important role in influencing users' detection behaviours. The process of detection includes various phases, and each phase is influenced by certain user characteristics. The following paragraphs describe the impact of user characteristics on their detection behaviours.

The first phase in the deception detection model is activation, which is measured by susceptibility. There are two variables responsible for increasing users' susceptibility to phishing emails. These variables prevent users from suspecting phishing emails. A lack of suspicion means that these victims are not able to tell the difference between phishing emails and legitimate ones. In contrast, email richness reduced users' susceptibility to phishing emails and caused them to engage in detection behaviour. Engaging in detection behaviour does not guarantee that users will be detectors; users must develop an explanation as to why there are differences between what they observe and what they expected to observe. This is called inconsistency, which is essential for detection [35].

Hypothesis generation and hypothesis evaluation are the next phases in the deception detection model. Generating a weak hypothesis and performing weak hypothesis evaluation inhibits users from detecting phishing emails. In this research project, one victim who responded to the phishing email reported that he suspected the phishing email to be illegitimate. However, he became a victim because he generated the following hypothesis: "If other students received this email, then it is genuine." After contacting one of his friends, who confirmed having received the phishing email, he decided to respond to the phishing email simply because other students had received the email. It can be seen that this victim had a combination of weak hypothesis generation and weak hypothesis evaluation. Phishing emails can easily fulfill this assumption because they target an enormous number of users. In contrast, detectors have shown a high level of strong hypothesis

generation. One detector forwarded the phishing email to the owner of the blog, asking him whether or not the email was genuine. Confirming the legitimacy of a suspected email with the legitimate entity that the phishing emails impersonate will help in detection.

The final phase in the deception detection model is global assessment. In this phase, users come to a decision regarding whether to respond to the phishing email. In this research, four variables were found to increase victims' tendency to perform the action requested in the phishing emails. Most of these variables are related to users' personalities, which leads to the assumption that these victims are risk takers. One participant acknowledged his knowledge that sending passwords over emails was a low-security behaviour, but he still chose to send them. In the interviews, our research found that some victims suspected the phishing email and commenced the process of detection until they reached the final phase. However, these victims still chose to respond to the phishing email. It can be seen that personality characteristics influenced users' decisions to perform risk-taking behaviours. This means that these victims were aware of the risk involved in their low-security behaviour but, due to their personality, chose to take the risk.

89.7 Limitations

Our research examines users who have a relationship with the impersonated entity. It is assumed that these kinds of phishing email attacks are difficult to detect. Emails received from impersonated entities that users have no relationship with are easily rejected.

Our research tested two types of phishing emails. However, in the real world, phishing emails come in various forms. Therefore, more experiments and time are needed.

Because the number of participants in our research was limited, the results cannot be generalized to the wider population of Internet users in Australia. To generate broader results, similar experiments involving an enormous number of participants must be conducted.

89.8 Conclusion

Our research aimed to identify user characteristics that cause users to become victims of phishing emails, as well as their detection behaviour. Several variables have been measured using the deception detection model to discover the impact of user characteristics on the process of detecting phishing emails. The findings of the research revealed that trust and submissiveness increase users' susceptibility to phishing emails (activation), which is the first phase in the deception detection model. The personality dimensions of extraversion, agreeableness and openness, in

addition to susceptibility, are responsible for increasing users' tendency to proceed with the requested actions included in the phishing email, which is the last phase of the deception detection model. Our research suggests that users' ability to detect phishing emails is complicated and should be investigated. Detectors are able to conclude that they should ignore phishing emails on their own. In contrast, victims need to be assisted with their decision making to increase their resistance to phishing emails.

The overall suggestion produced by our research is that user characteristics play an important role in improving users' ability to detect phishing emails. Various user characteristics impact different phases of the process of detecting phishing emails. To increase user security against phishing emails, the user characteristics that are important for each phase must be improved. Assisting users in detecting phishing emails can be done by (1) educating users about phishing emails and their techniques to decrease their susceptibility and (2) performing re-assistance to reduce their risk-taking behaviours and thus decrease their tendency to respond to phishing emails.

Acknowledgments Ibrahim Alseadoon would like to acknowledge the Ministry of Higher Education, Saudi Arabia and the University of Ha'il for sponsoring his PhD studies at Queensland University of Technology, Brisbane, Australia. Mohd Fairuz Iskandar Othman would like to acknowledge the Ministry of Higher Education, Malaysia and Universiti Teknikal Malaysia Melaka for sponsoring his PhD studies at Queensland University of Technology, Brisbane, Australia.

References

1. Zhang, W., Luo, X., Burd, S.D., Seazzu, A.F.: How Could I Fall for That? Exploring Phishing Victimization with the Heuristic-Systematic Model. In *System Science (HICSS)*, 45th Hawaii International Conference on, 2012, pp. 2374–2380 (2012)
2. Xun, D., Clark, J.A., Jacob, J.: Modelling User-phishing Interaction. In: *Proceedings of Human System Interactions May 25-27, 2008, Kraków, Poland* (2008)
3. Wang, J., et al.: Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *Profess. Commun. IEEE Trans.* **99**, 1 (2012)
4. Fette, I., Sadeh, N., Tomasic, A.: Learning to detect phishing emails. In: *Proceedings of the 16th International Conference on World Wide Web*. pp. 649–656. ACM: Banff, Alberta, Canada (2007)
5. Cook, D., Gurbani, V., Daniluk, M.: Phishwish A Stateless Phishing Filter Using Minimal Rules. In: *Financial Cryptography and Data Security*. pp. 182–186. Springer, Heidelberg (2009)
6. Bergholz, A. et al.: Detecting known and new salting tricks in unwanted emails. In: *CEAS 2008: Proceedings of the Fifth Conference on Email and Anti-Spam*. Citeseer (2008)
7. Wang, J., et al.: An exploration of the design features of phishing attacks. *Inf. Assur. Secur. Priv. Ser.* **4**, 29 (2009)
8. Sharma, K.: An anatomy of phishing messages as deceiving persuasion: a categorical content and semantic network study. *EDPACS* **42**(6), 1–19 (2010)
9. Anti Phishing Working Group Phishing Activity Trends Report, 2nd Quarter (2012)

10. Pettey, C.: Gartner Says Number of Phishing Attacks on U.S. Consumers Increased 40 Percent in 2008 (2006) [cited 2009 2/9]; Available from: <http://www.gartner.com/it/page.jsp?id=936913>
11. Herley, C., Florencio, D.: A profitless endeavor: phishing as tragedy of the commons. In: Proceedings of the 2008 Workshop on New Security Paradigms. ACM: Lake Tahoe, California, USA. pp. 59–70 (2008)
12. Sheng, S. et al.: Anti-Phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In: Proceedings of the 3rd Symposium on Usable Privacy and Security. ACM: Pittsburgh, Pennsylvania (2007)
13. Arachchilage, N.A.G., Love S., Scott M.: Designing a Mobile Game to Teach Conceptual Knowledge of Avoiding “Phishing Attacks”. *Int. J. e-Learn. Secur.* **2**, 127–132 (2012)
14. Kumaraguru, P. et al.: School of phish: a real-world evaluation of anti-phishing training. In: Proceedings of the 5th Symposium on Usable Privacy and Security. ACM: Mountain View, California, pp. 1–12 (2009)
15. Bekkering, E., Hutchison, D., Werner, L.: A follow-up study of detecting phishing emails. In: Proceedings of the Conference on Information Systems Applied Research. Washington DC (2009)
16. Sven, D. et al.: Phishing IQ tests measure fear, not ability. In: *Financial Cryptography and Data Security*. pp. 362–366. Springer, Heidelberg (2007)
17. Wu, M., Miller, R.C., Little, G.: Web wallet: preventing phishing attacks by revealing user intentions. In: SOUPS '06: Proceedings of the Second Symposium on Usable Privacy and Security. ACM, New York (2006)
18. Kirda, E., Kruegel, C.: Protecting users against phishing attacks. *Comput. J.* **49**(5), 554 (2006)
19. Kim, Y.-G., et al.: Method for evaluating the security risk of a website against phishing attacks. *Intell. Secur. Inf.* **5075**, 21–31 (2008)
20. Wu, M., Miller, R.C., Garfinkel, S.L.: Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 601–610. ACM, Montreal, Quebec, Canada (2006)
21. Purkait, S.: Phishing counter measures and their effectiveness—literature review. *Inf. Manag. Comput. Secur.* **20**(5), 382–420 (2012)
22. Johnson, P.E., Grazioli, K.S.: Fraud detection: intentionality and deception in cognition. *Acc. Organ. Soc.* **18**(5), 467–488 (1993)
23. Johnson, P.E., et al.: Success and failure in expert reasoning. *Organ. Behav. Hum. Decis. Process.* **53**(2), 173–203 (1992)
24. Grazioli, S.: Where did they go wrong? an analysis of the failure of knowledgeable internet consumers to detect deception over the internet. *Group Decis. Negot.* **13**(2), 149–172 (2004)
25. Daft, R.L., Lengel, R.H.: Organizational information requirements, media richness and structural design. *Manage. Sci.* **32**(5), 554–571 (1986)
26. Costa, P.T., McCrae, R.R.: Four ways five factors are basic. *Pers. Individ. Differ.* **13**(6), 653–665 (1992)
27. Parrish Jr, J.L., Bailey, J.L., Courtney, J.F.: *A Personality Based Model for Determining Susceptibility to Phishing Attacks*. Little Rock: University of Arkansas (2009)
28. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. *J. Res. Pers.* **37**(6), 504–528 (2003)
29. McKnight, H., Kacmar, C., Choudhury, V.: Whoops. Did i use the wrong concept to predict e-commerce trust? modeling the risk-related effects of trust versus distrust concepts. In: System Sciences, Proceedings of the 36th Annual Hawaii International Conference. IEEE, Waikoloa, Hawaii, USA (2003)
30. Wright, R., Chakraborty, S., Basoglu, A., Marett, K.: Where Did They Go Right? Understanding the Deception in Phishing Communications. *Group Decision and Negotiation.* **19**(4), 391–416 (2009)
31. Allan, S., Gilbert, P.: Submissive behaviour and psychopathology. *Br. J. Clin. Psychol.* **36**(4), 467–488 (1997)

32. Carlson, J.R., Zmud, R.W.: Channel expansion theory and the experiential nature of media richness perceptions. *Acad. Manag. J.* **42**(2), 153–170 (1999)
33. Sheng, S. et al.: Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In: *Proceedings of the 28th International Conference on Human Factors in Computing Systems*. ACM, Atlanta, Georgia, USA (2010)
34. Kumaraguru, P. et al.: Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In: *Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit*, pp. 70-81. ACM, Pittsburgh, Pennsylvania (2007)
35. Buller, D.B., Burgoon, J.K.: Interpersonal deception theory. *Commun. Theory* **6**(3), 203–242 (1996)

Chapter 90

Adaptive and Dynamic Service Composition for Cloud-Based Mobile Application

R. Kanesaraj Ramasamy, Fang-Fang Chua and Su-Cheng Haw

Abstract Adaptive and dynamic service composition are the main research challenges in Service-oriented Architecture (SOA). We are looking into improving the efficiency of dynamic composition and the methods on how to improve the discovery of web services. Currently the approaches which are available do not compose while checking for errors. In this paper, we will be composing web services using Multi Agent Methods (MAS) and Petri Net for error checking to improve the efficiency of the composition. We are implementing the composition engine using Business Process Model (BPM) to compose web services while MAS is used to locate web services. Proposed architecture is expected to produce a dynamic web service composition engine and improve reusability for Cloud-Based Mobile Application.

90.1 Introduction

Web services are becoming more important and have received great considerations from the research community. Web services are reachable network interface to application programs. They are generally created by using the standard Internet technologies. Web services are also self-describing software applications that can be advertised, located, and used across the Internet using a set of standards such as SOAP, WSDL, and UDDI [1]. Web services composition allow application to communicate on various platform.

R.K. Ramasamy (✉) · F.-F. Chua · S.-C. Haw
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
e-mail: kanes87@gmail.com

F.-F. Chua
e-mail: ffchua@mmu.edu.my

S.-C. Haw
e-mail: schaw@mmu.edu.my

Service-oriented Computing paradigm (SOC), emphasis the web service composition, which exploits services as important basics for developing applications. Web service helps to achieve interoperable Business to Business (B2B) interaction offered by multiple business partners based on business processes [2]. This interconnection of web services is called as web service composition. There are two types of composition which are static and dynamic or automatic composition. In this paper, we have elaborated on two different web service compositions and its methods. The web service composition languages such as Business Process Execution Language for Web Services (BPEL4WS), Business Process Modeling Language (BPML), Web Service Choreography Interface (WSCI), XML language (XLANG), and Web Services Flow Language (WSFI) are used to specify process models. The remainder of this paper is organized as follows. Section 90.2 introduce service composition methods. Section 90.3 discusses the analysis of web service composition methods. Section 90.4 describes about cloud service composition and Sect. 90.5 presents the proposed solution.

90.2 Service Composition Methods

90.2.1 Multi Agent Method

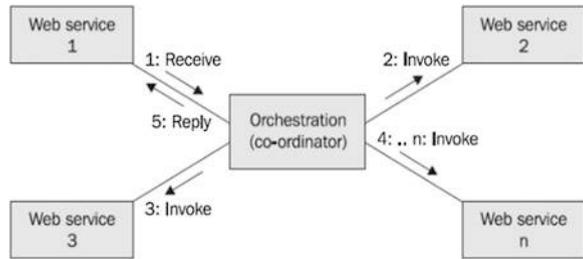
Intelligent behavior is intended as the ability of processing the input requirements as introduced by the user, whereas the system is able to analyze the input and identify the web services. It also carry out an automatic composition of web services for corresponding business processes [3]. Agents are built with intelligent behavior based on respective task to be completed. Agents are important to web service composition because it being context-aware, autonomous and able to interpret semantic with the help of ontological knowledge representation. Agents interact with web services by requirement layer which is one of the three layers in multi agent system [4]. Requirements are set to describe the interaction and roles whereas contracts are used to describe the responsibilities of the roles in the requirement layer which have three types of contract as follow:

- Permission shows the permit behavior of roles
- Obligation goals that the roles must accomplished
- Prohibitions describe the forbidden behavior of the roles

Agent layer is the second layer in MAS, which consists of role agent and manage agent. Role agent will act according to the web service and manage agent will monitor the agent and system behavior. Besides that, there are two approaches that can be used to compose web services known as orchestration and choreography.

Figure 90.1 shows orchestration approach. Orchestration can interact with both internal and external web services and it occurs at the message level. Orchestration

Fig. 90.1 Orchestration approach [4]



also represents control from one party’s perspective which is responsible for coordinating the whole process of the system and does the composition while the agent doesn’t need to know whom to contact. In the orchestration approach, agent coordinator handles the composition. Orchestration is dynamic as the agent coordinator take full responsibility of the new services instead of each service or agent aware of it. Multi agent can self-update and self-compose the web services and each agent knows its role hence it will be beneficial for us to use multi agent method as it is autonomous and highly reliable. Figure 90.2 shows Choreography approach. Choreography is collaborative and allows each invoked party to describe its part in the interaction and track message sequence among multiple parties and sources. Moreover, in this approach, agent is more independent compared to orchestration approach where the entire agent is controlled by coordinator. All service or agents in choreography approach have to be aware of business process, exchange message, operation and timing of exchange message [4].

90.2.2 Complex Event Processing Method

Complex Event Processing (CEP) helps to control the events while multi events can compose the complex event. According to David B. Robins, CEP is a part of Event-Driven Architecture (EDA), which is an architecture dealing generally with the production, detection, consumption of, and reaction to events [5]. CEP is using detection of multiple event stream process and heterogeneous event. The events occur when messages are sent and received Simple Object Access Protocol (SOAP) messages and CEP adaptation needs the platform to consume incoming messages, process them and send the results to its destination. Besides that, CEP has the ability to recognize the complex events. This technology discovers a relationship between events through the analysis and correlation of the multiple events. CEP contains simple techniques and use new techniques such as patterns of events including event relationships. A large part of CEP is pattern matching and it consists filtering, which is to determine events of interest and extracting properties from the event [5]. Table 90.1 describes the basic transformation patterns.

The CEP has higher level performances and capacity compared to other approaches. It is an excellent platform which can carry many events and processes

Fig. 90.2 Choreography approach [4]

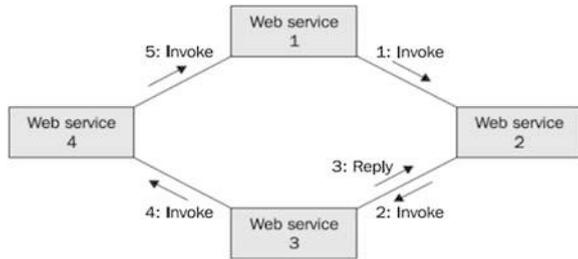


Table 90.1 Five basic transformation patterns that can reconcile mis-matches

Patterns	Description
Match-make	It translates one message type to another
	Solve the one-to-one transformation
Split	The approach for the one-to-many patterns
	It separates one message sent by the source into more than one to be received separately
Merge operator	The solution for the many-to-many patterns
	It combines more than two messages into one message
Aggregate	Solution for the one to one transformation
Disaggregate	Solutions for the many to many transformations

(1000 to 100 k messages per second). The CEP plays its role as a great platform in real time business implication however it has low efficiency in composing large amount of services.

90.2.3 Petri Net Method

Petri net-based method is used for the data validation of web services composition. It is developed with BPEL4WS. Petri net is used as validation tool in web service composition to detect errors, misconceptions and failures. This method is focused on selecting the best possible path for web services composition based on the user request. In this method, the movement of token (black dots) from one place to another caused by firing of transition. The firing represents an occurrence of the event or an action taken. Petri net is a process modeling technique that has a formal semantic that has been used to model and analyze several types of processes including protocols and business process. Every node is either a place or a transition and it is connected and directed. When a token occupied in place, it is able to place at least one token that connected to the transition. The enabled transition fire one token from every input place and depositing one token from every output place [6].

The early forms of high level petri net are Coloured Petri Nets (CPNs). CPNs are used as an underlying formalism and BPEL as inputs and define service workflow net as a kind of CPNs-as a united formalism to describe services, composition and mediator [7]. In [7], they separated the control flow with the message exchange and provide a united formalism to describe services, composition, and mediator. Then High-level Petri net (HPN) was developed to overcome the problems in web service, such as the use of complex structured data as tokens and using algebraic expressions to annotate net elements. Figure 90.3 shows the process flow of Petri net.

Basically petri net has three steps:

- Input places allowing messages to get into the process
- Output places, to go out from the process
- Internal places allowing for modeling of the process behavior.

Transitions represent the dynamic element (activity of the business process). Based on the customer requests from the Petri net, once the services are identified through Petri Net, the corresponding aspects developed are bonded to the join points in the main code. It also can be used to evaluate the effectiveness of services composition. The tokens describe the state of the net. Hereby an arc connects a condition (place) with an event (transition) and vice versa [3].

If an arc goes from a condition to an event, it means that the condition is necessary for the occurrence of the event. In this case, we have a pre-condition. Analogously an arc from an event to a condition means, that the occurrence of the event makes this condition true, and it is called a post-condition. In large number of places, there are several places with same semantic meaning. Figure 90.4 shows the normal process in web service that has two web services “BookInfo” Query and Book Deal. Both the web services require the same input “BookName” from user. Figure 90.4 is using different places for both the services even it requires the same semantic. Petri net method will group in one place in which the semantic have the same meaning with the tokens as shown in Fig. 90.5. If the transition is not effective then there is no matching between the transitions and inputs or desire outputs, the service can be deleted because it will not be used.

Petri net graph gives all possible combination of composition available with the selected web services. By implementing this method in web service composition, it improves the reliability of web service selection. Petri net also reduce service selection scope and improve the composition efficiency.

90.2.4 SWORD Method

According to [8], SWORD is a set of tools for the composition of Web services including “information-providing” services. SWORD used rules to represent a service that able to produce specific output based on certain input given to the service. To determine automatically whether a desired composite can be

Fig. 90.3 Petri net process flow [3]

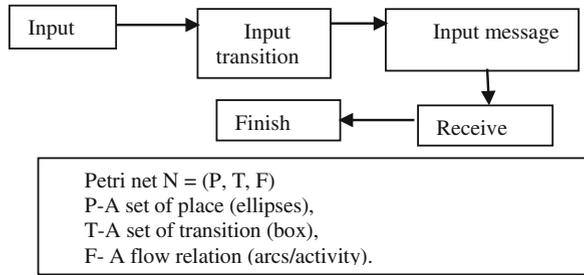


Fig. 90.4 Normal web service process [11]

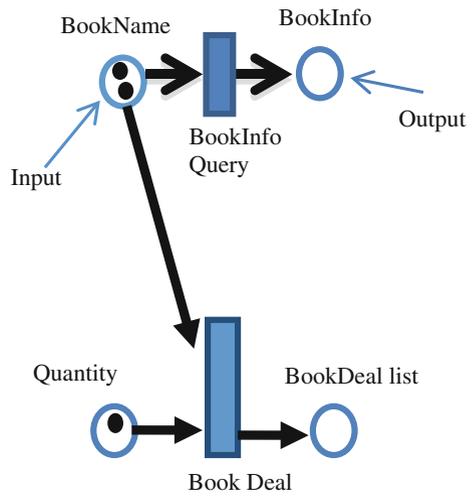
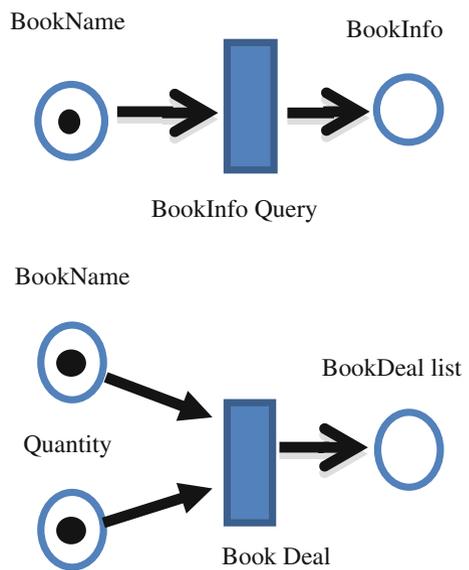


Fig. 90.5 Using petri net (merge same semantic) [11]



appreciated with the existing services, a rule—based expert system is established. Additionally, SWORD can compose information by providing services such as information, email and other services. It also generates a functional composition plan given the functional requirements. SWORD uses Entity-Relation (ER) model to specify the Web services. From the inputs and outputs of service, the rule based expert systems defined whether which outputs can be obtained by the service which indicates the inputs. Developer needs to specify the inputs and outputs of the composite service in the world model and submit it to SWORD at the phase of deployment. It is to create and deploy a new composite service. SWORD generates a composition plan for the composite service. Moreover, the developer can view the generated plan and the developer can also request for a persistent representation of the plan generated if the plan is appropriate [9].

The sequence of services is included in the representation of the plan that need to be invoked in order to obtain the composite service outputs from its inputs. According to [9], when actual requests for the composite service is received, the services specified in the plan are executed, starting with the known inputs, in order to compute the desired outputs. SWORD has the ability to self-compose when given certain inputs and it is efficient in responding to its environment. It allows developers to compose the existing web services efficiently to produce new web services composition.

90.3 Analysis of Web Service Composition Method

MAS is autonomous because it has two approaches that can control its own agent. For the first approach, the coordinator will take full responsibility of the services, whereas the second approach, each agent acts independently. MAS can stay operational and carry out the task immediately when the user gives an input. Moreover, the agent will give the output based on what user request and robustness in response to the environment. MAS will be highly complex if combined with HPN. Table 90.2 shows the characteristic comparison between methods.

CEP which uses pattern such as aggregate. CEP is autonomous because it is independent but low in efficiency when composing an input. It keep composing till receive checkout messages that indicate the input has been sent. Sword and MAS uses mediator however, Sword is rule based method. It composes based on the rules that has been set. Petri net is neither autonomous nor independent because there are no simultaneous images of the flow of parts, resources, information and control through the system. This method offers the ability to visualize the system flow, which makes it easy to understand and learn. Petri Net method efficiently reduces service selection scope by finding the shortest path or critical path to compose the service but it is highly complex because of its formula and distributed choice of transition has to be executed [7]. Petri net unable to self-compose because the flow of web service have different interface or mismatches in the interface. Based on Table 90.2, the recommended method to be used to compose

Table 90.2 Comparison between methods

Characteristics/methods	Multi agent (MAS)	CEP/patterns	Sword	Petri net
Autonomous	Yes	Yes	Yes	No
Reliability	Yes	Yes	Yes	Yes
Efficiency	Yes	Low	Yes	Yes
Complexity	Low	High	High	Very High
Self-compose	Yes	No	Yes	No

web service is MAS. MAS is autonomy where it can self-update and highly reliable. It is more efficient compared to other methods because of its robustness in response to the dynamic environment. Moreover, MAS has lower complexity than CEP, Sword and Petri net.

90.4 Cloud Service Composition

There are a lot of different definitions available for cloud. We accept the definition of cloud, in which provided by the National Institute of Standards and Technology (NIST) [10]. The NIST cloud computing definition: *Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models* [10].

Cloud services work with all resources called as participants such as consumers, brokers and service providers but participants may have incomplete knowledge of all possible Cloud resources and even to other participants, due to constantly changing Cloud environments and their distributed nature [4]. An efficient Cloud service composition method should dynamically select the most appropriate services to create the cheapest composition solution. As service providers charge a fee for providing a service and this service fees may change over the time, based on supply and demand. Referring to [4], to satisfy incoming client requirements, there is a collaboration needed between brokers and service providers. Those requirements can be received via Cloud resources which can be accessed by using web service in an automated manner. However, constantly changing cloud computing environments, which are deployed as self-contained components, are normally partial solutions that must be composed to provide a single virtualized service to Cloud consumers. This composition of services should be carried out in a dynamic and automated manner to satisfy consumer's requirements. The idea of adopting software agents for managing Cloud services was first introduced in [10]. Mell and Grance [10] are the earliest efforts in designing and constructing

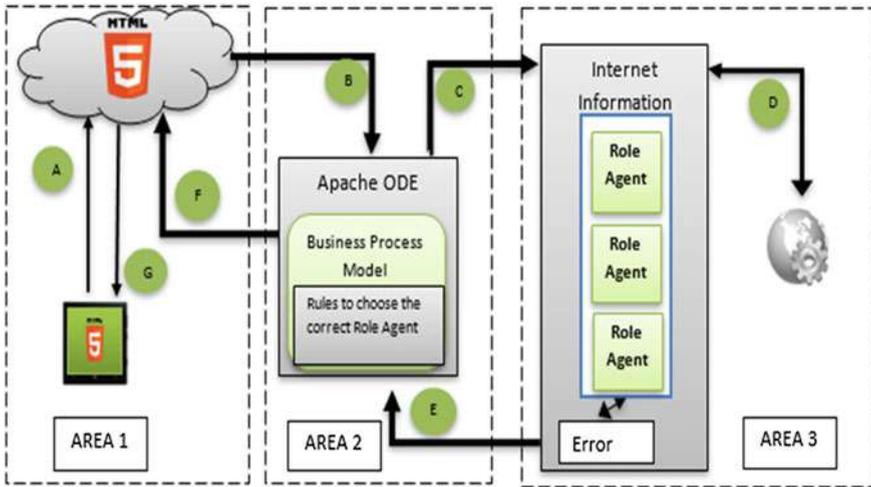


Fig. 90.6 Architecture of web service composition method

Table 90.3 Explanation for stages

Data flow	Description
A	User input from mobile phone to cloud (HTML 5 Application)
B	User inputs transferred to business process model
C	Business process model with embedded rules will decide which role agent to trigger
D	Role agent will search for the respective web service to be used
E	Selected web service will be sent to business process model to perform operation
F	Upon completion of all the processes in business process model, the output will be replied back to HTML 5 application which is placed in CLOUD
G	The output received in HTML 5 application will be reflected on user’s mobile phone

negotiation agents for supporting the establishment of service-level agreements among Cloud participants, this present research introduces a self-organizing agent-based approach for dealing with Cloud service composition.

90.5 Proposed Solution

Based on the analysis from various sources as discussed above, we have proposed a new architecture for a cloud-based mobile application which have the capability to compose web services dynamically. Besides that, our focus is also on designing the architecture that is adaptive for the provided environment. Figure 90.6 shows the architecture of the proposed idea. Table 90.3 provides the explanation of data flow based on Fig. 90.6.

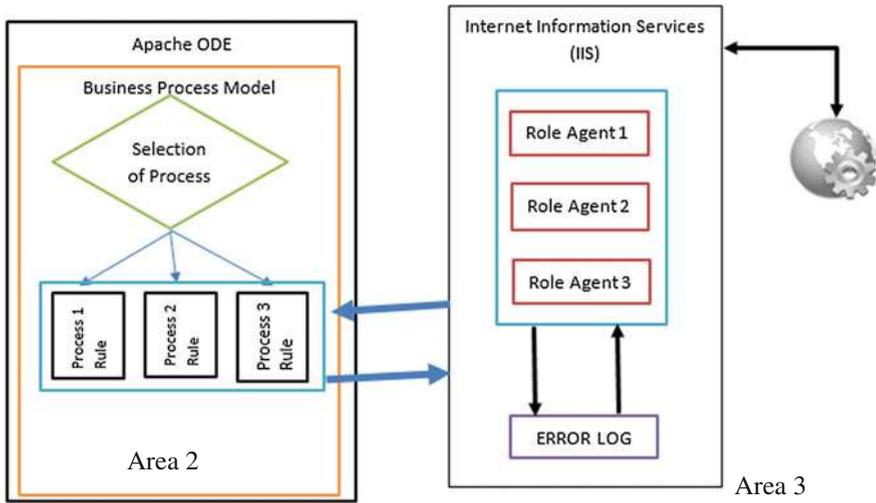


Fig. 90.7 Area 2 and area 3 (adaptive and dynamic process)

Referring to Fig. 90.6, the proposed architecture is divided to 3 main areas. The first area is the user input or end user input for the model. Second area is the adaptable area finally the dynamic searching area.

Adaptability happens in area 2. We have adapted the orchestration approach as used in [11] to improve the availability. HPN as proposed by [3], will be used to reduce errors. We aim to reduce service selection scope and improve on the composition efficiency. Business Process Execution Language (BPEL) will be used to execute the designed Business Process Model. Our proposed solution will cover exceptional behavior handling such as faults, events and compensation.

As shown in area 3 in Fig. 90.7, dynamic composition can be achieved by using automated tools to analyze a user query, and select and assemble Web service interfaces to fulfill the user demand. Choreography approach will be used to create the dynamic environment. Agents will receive requests from BPM and it will search the correct service and return it back to BPM. Agents will also update the error log when the requested semantic was not found during the searching process. The agent will refer to the error log before the searching is started to check on the availability and the location of the services.

90.6 Conclusion

The proposed system will be tested using a real time mobile application such as early warning system for natural disasters. Besides that, it will also support various mobile platform such as Windows Phone, IOS and Android. As for future work,

we are planning for performance analysis using QoS attributes to test the efficiency of adaptive service composition strategy. It is expected to provide an improved search result and reduce composition time.

References

1. Liu, Y., Ngu A.H., Zeng L.Z.: QoS computation and policing in dynamic web service selection. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, pp. 66-73. ACM (2004)
2. Papazoglou, M.P., Van Den Heuvel, W.: Service-oriented computing: state-of-the-art and open research issues. *IEEE Comput.* **40**(11), 1086 (2003)
3. García Coria, J.A., Castellanos-Garzón, J.A., Corchado, J.M.: Intelligent business processes composition based on multi-agent systems. *Expert Syst. Appl.* **41**, 1189–1205 (2014)
4. Maalal, S., Addou, M.: A new approach of designing multi-agent systems. *Int. J. Adv. Comput. Sci. Appl.* **2**, 148–149, 149–150 (2011)
5. Taher, Y., Parkin, M., Papazoglou, M.P., Van Den Heuvel, W.: *Adaptation of Web Service Interactions Using Complex Event Processing Patterns*, pp. 601–609. Springer, Heidelberg (2011)
6. Zhao, L., Li, B., Wu, J.: A multi-agent collaborative model for web services composition. In: ISECS International Colloquium on Computing, Communication, Control, and Management. CCCM 2009, vol. 1, pp. 145–148. IEEE (2009)
7. Bencomo, N., Whittle J., Sawyer, P., Finkelstein, A., Letier, E.: Requirements reflection: requirements as runtime entities. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, Vol. 2, pp. 199–202. ACM (2010)
8. Chan, P.P., Lyu, M.R.: Dynamic web service composition: a new approach in building reliable web service. In: 22nd International Conference on Advanced Information Networking and Applications. AINA 2008, pp. 20–25. IEEE (2008)
9. Ponnekanti, S.R., Fox, A.: Sword: a developer toolkit for web service composition: In Proceedings of the Eleventh International World Wide Web Conference, Honolulu, HI, vol. 45. (2002)
10. Mell, P., Grance, T.: *The NIST Definition of Cloud Computing*. Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, p. 50 (2009)
11. Yamaguchi, H., El-Fakih, K., Bochmann, G.V., Higashino, T.: A Petri net based method for deriving distributed specification with optimal allocation of resources. In: Proceedings of ASIC International Conference on Software Engineering Applied to Networking and Parallel/ Distributed Computing (SNPD'00), pp. 19–26 (2000)

Chapter 91

Web Service Composition Using Windows Workflow for Cloud-Based Mobile Application

R. Kanesaraj Ramasamy, Fang-Fang Chua and Su-Cheng Haw

Abstract There is a lot of support provided by the mobile OS giants in cloud services. Besides providing their own services, they encourage developers to upload services into their cloud environment. However, when more services are being consumed in their cloud, more charges imposed to the users. Meanwhile, there exists a large amount of services which are available in public web service repositories. The cost of mobile development can be reduced by invoking and composing the services in sequence and upload or use them as a single service. We have used Windows Workflow for web service composition and the WSDL file produced can be hosted in our cloud environment and to be used by all. In this paper, we will be discussing on how to use Windows Workflow for web service composition for cloud-based mobile application. With the proposed system we will be able to save battery life time and increase the storage capacity.

91.1 Introduction

Web services are unique as it can be supported by various platforms although it was developed in one particular platform. Web services which are available in service repositories can be used in one single system, which is called as web service composition. In other word, various web services can be composed in one application [1]. In fact, web service interfaces are like remote procedure call and the interaction protocols are manually written [2]. Application can be developed

R.K. Ramasamy (✉) · F.-F. Chua · S.-C. Haw
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Malaysia
e-mail: kanes87@gmail.com

F.-F. Chua
e-mail: ffchua@mmu.edu.my

S.-C. Haw
e-mail: schaw@mmu.edu.my

by invoking a couple of services using the composition method to reduce development time. There are plenty of support being provided by various service providers such as Google and Microsoft since web services can be called remotely. They have provided a platform with all the prebuilt services to be used for our daily development whereby they also encourage the developer to add on more services into their cloud environment. This indicates the widespread of cloud environment in current mobile development.

Besides that, service composition can be executed by composing elementary or composite services. When composing web services, the sequence of events that occur in any business process will be implemented by few services. This will lead to workflow management, where the logic is comprehended by composing autonomous applications. These composite services also can be in turn recursively composed with other services into higher level solutions, and so on [3]. Furthermore, this recursive composition feature allows developing new solution rapidly based on the existing business services and becoming one of the most significant features of SOA.

Deadlock, mismatching and availability of web service are the major issues to be faced by developers in the development process of web service composition for complex and distributed environment [4]. In order to solve the problems, they divided the composition process into three parts: constraints before composition, constraints in composition and constraints after composition. They used finite state machine to model web services to discuss the constraints in web service composition. Before the composition occurs, the status of web service should be confirmed as active even though it is difficult to be determined.

The paper discusses about the development of web service composition for Cloud-Based mobile application. We had tested the efficiency of stand-alone mobile application with cloud-based mobile application. With this implementation, it is expected to have complex mobile application running on mobile devices without compromising the resources of the mobile device.

The remainder of this paper is generally organized as follow: Sect. 91.2 will introduce the related works which have been done by other researchers and ideas that provide motivation for the proposed solution. Section 91.3 shows the proposed solution on how it's being implemented. Section 91.4, shows the results of the implementation and discussion. In Sect. 91.5, analysis of the result will be discussed to identify the problems which have been fixed with the proposed method. Finally, a conclusion is drawn and future work is described.

91.2 Related Works

Web service composition provides an opportunity for enterprises to increase the ability to adapt themselves to frequent changes in users' requirements by integrating existing services [5]. According to [6], they present a novel algorithm called HRLPLA (Hierarchical Reinforcement Learning) for composing Web services.

This algorithm deliberates functional and QoS properties simultaneously. However, the plans generated by AI planning are not adaptive to running environment [6]. This leads to failure of composition plan when a component service stops functioning. Additionally, it is not suitable for dynamic environment. To solve this problem, they introduced MAXQ, a HRLPLA method where it provides efficient service composition and can work with a large scale of services. Since it is using hierarchical structure, information obtained in designed subtasks can be reused by super tasks; therefore it accelerates the process of learning. Their approach is proven through experimental result to achieve higher efficiency and effectiveness than normal reinforced learning and AI planning.

Though every method provides different level of automation in service composition, they are not to prove that the higher automation is the best. This is because the Web service environment is highly complex and it is not feasible to generate everything in an automatic way [7]. Jinghai and Xiaomeng [7] presents a technique to generate composite services from high level declarative description. This method used composition rules to define whether two services can be composed. This method contains 4 phases; first specification phase where Composite Service Specification Language (CSSL) language used. Secondly, in matchmaking phase, it accepts service requester's specifications by generating compositions plan using composition rules. Thirdly is the selection phase where service requester select particular plan generated in selection phase based on quality of composition. The final phase is the generation phase. In this phase, a detailed description of the composite service is automatically generated and presented to the service requester [7]. The composition rules deliberate the syntactic and semantic properties of Web services and at the same time contribute the Web service's attributes that could be used in service composition. These rules can be used as a guideline for other Web service methods also [7].

Another service composition approach is non-QoS based approaches. Jinghai et al. [8] presents a mixed initiative framework for semantic web service discovery and composition which allow user involvement over many key decisions by suggesting and identifying the inconsistencies. In this approach, users decide the number of functionalities to be supported. The quality and accuracy of the support provided by their framework is intended to improve over time, as users learn to develop richer and more accurate annotations. For this composition engine, it combines Web Ontology Language (OWL) ontologies where it plans functionality with Jess based on the GraphPlan algorithm. GraphPlan provides reach-ability analysis to determine whether a given state can be reached from another state and disjunctive refinement [8].

In [9], the authors utilize the semantic web service languages with the model driven methodology to build composite web services. There are four phases and at the end of first phase, abstract composite model is attained. It contains all the vital information used for service discovery and selection. On the second phase, discovery process depends on semantic descriptions where appropriate web services are handled. While on third phase, a concrete composite model is attained to handle mismatch between output of one service and input of other service. In the last phase,

different descriptions of the concrete composition model are used for the composed service [9]. The methodology also deliberates a syntactic and semantic description about the interfaces of service candidates at the same time processes QoS requirements from the developer and offerings from the service providers [9]. The benefit is this methodology provide good documentation of composition in the form of graphical models. Executable compositions and the ability generate semantically from a graphical model offers valuable advantage to the service developers, where they don't have to write a lot of low-level XML code.

91.3 Proposed Solution

Significantly, stand-alone mobile application responds faster compared to cloud based application. This is because, a stand-alone mobile application only loads internally meanwhile a cloud-based mobile application will load from cloud server. We would like to test the performance of the applications towards Business Process which are created by composing a few services. Figure 91.1 shows the proposed architecture for our Cloud-Based mobile application. Based on Fig. 91.1, the engine for the mobile application will be placed in cloud environment. We consider Windows Workflow Foundation (WF) and HTML 5 pages as the engine. WF is used to compose web services together to provide a complex mobile application. Table 91.1 shows the description for proposed architecture for cloud-based mobile application.

WF is a Microsoft technology that provides an API, an in-process workflow engine, and a rehostable designer to implement long-running processes as workflows within .NET applications. The current version of WF was released as part of the .NET Framework version 4 and is referred to as (WF4) [10]. The term "Workflow" here refers to the root activity that is executed by the host. In addition, Workflows can use both out-of-box activities and custom activities [10].

For the deployment and testing, we are using Windows Phone (Nokia Lumia 920) and Android Emulator to test the efficiency in completing the composition and getting the results. The test environment is being set up in a personal computer which runs on Intel i7 2.66 GHz with 8 GB RAM.

We used BMI calculator as the domain for testing because it's easy to understand but it has more than one service to be composed. Proposed system uses three different web services. Table 91.2 shows the description of the services used in this project.

Proposed workflow will receive four values from the mobile application which are two values with double data type and two values with the metric used in the system for the measurements. In the workflow, the system will convert the width to kilogram and height to centimetres using the service provided in workflow. Then the converted value will be send to BMI calculator for BMI calculation and return the BMI value to the user. (91.1) explains the formula being used for this system or the calculation of BMI in this system. To ensure the values meeting the required

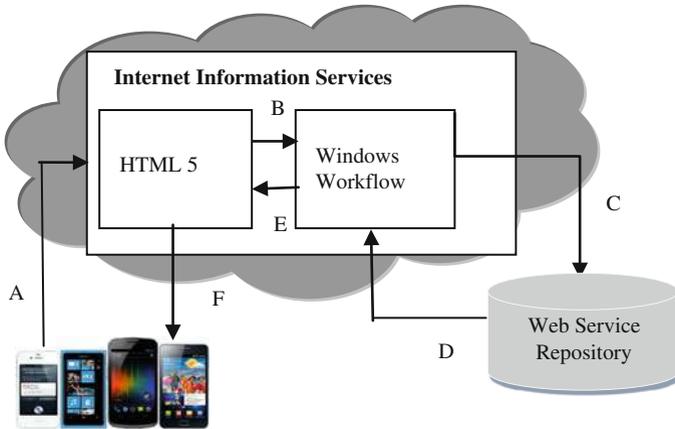


Fig. 91.1 Proposed architecture for cloud-based mobile application

Table 91.1 Description for proposed architecture for cloud-based mobile application

Stage	Description
A	User input from mobile device will be sent to HTML 5 page
B	HTML 5 will trigger the WSDL produced from WF
C	WF will invoke web service related to functions in workflow
D	The processed data in the related web service will be returned to WF
E	The WSDL will be return the processed data to HTML 5 page
F	Based on the processed data HTML 5 will produce the output to mobile device

Table 91.2 Description of services used

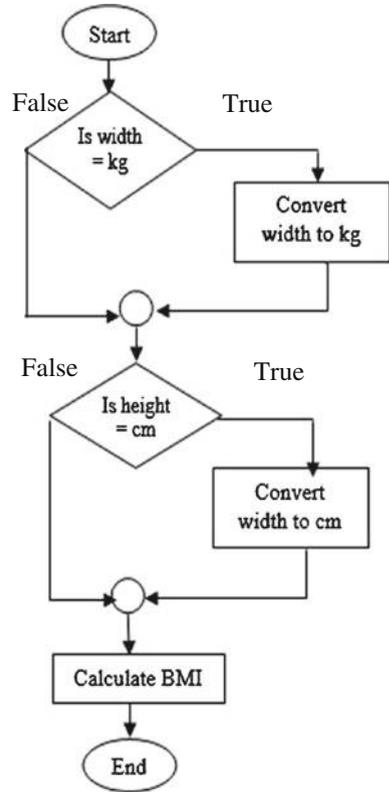
Service	Description
BMI	Calculates BMI but only receives weight in kilogram and height in centimetres
Width converter	Converts all the values received to desired width metric as required by user
Length converter	Converts all the values received to desired length metric as required by user

metrics for the calculation, we have included two separate services from a web service repository to perform the conversion and provide the value to BMI calculator service.

$$BMI = Weight[kg] / ((Height [cm] / 100)^2). \tag{91.1}$$

Figure 91.2 shows the flowchart of the Business Process (BP) where the composition takes place. Every process in flowchart will be using a web service from various web service repositories. In Fig. 91.3, we have shown the BP, which

Fig. 91.2 BMI BP flowchart



was developed on Windows Workflow based on Fig. 91.2. This BP will be deployed in a form WSDL. We will have 3 different WSDL which have been combined as one single WSDL which will call and perform the operation. This will improve the reusability of the code as developers do not have to call all the web services manually.

The composition is done using Windows Workflow 4.0 running on .NET Framework 4.5. We have used Visual Studio 2013 RC to develop the Windows Workflow project. The design that we have developed is able to handle the error or exception catching. We have used the try catch blocks to develop the workflow so that any error will be handled in workflow level and will not interrupt users operation or generate any system error for the user. Furthermore, to improve the performance in workflow, we have also included IF ELSE blocks, if user have selected the required metrics, it doesn't go through the conversion again. Figure 91.3 shows the implemented workflow or the composition diagram for this system.

Fig. 91.3 BMI BP in windows workflow



91.4 Implementation and Testing Results

As shown in Fig. 91.4, a stand-alone Windows Phone Application was developed and tested. Figures 91.5 and 91.6 shows the Cloud-Based Mobile Application which was developed for this testing. Using this emulator and Nokia phone we have collected the data of total response time to receive user input, invoke composed WDSL and finally display the result to user's screen in milliseconds as shown in Table 91.3. As for the test data, we have fixed the same 10 sets as we would like to test in all areas in this system such as converting the value, based on the metric required by the service. Since we have IF ELSE blocks in our workflow it is expected that it will provide a shorter response time compared to the flow that requires metric conversion.

Results in Table 91.3, indicates that Cloud Based application respond faster compared to a stand-alone application when there is WSDL to be executed from another server. The set of data used are the same for the entire test. We have fixed

Fig. 91.4 WP stand-alone application

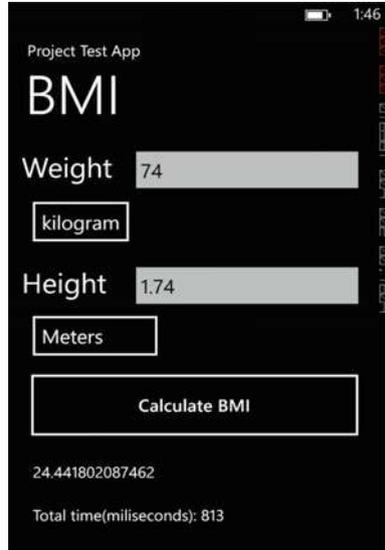


Fig. 91.5 Cloud based application in WP

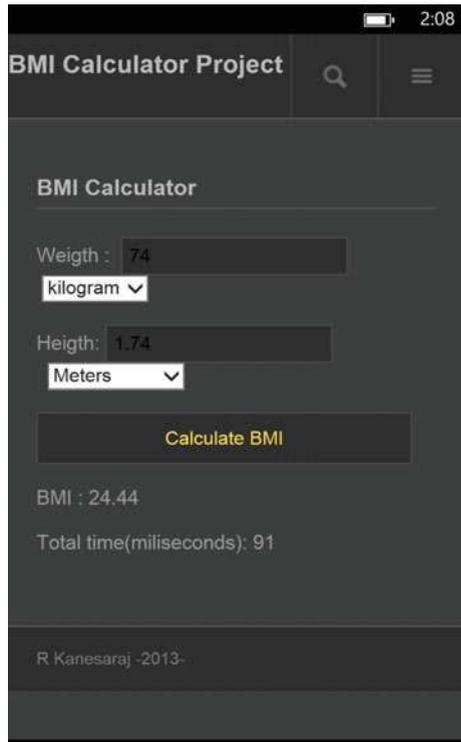


Fig. 91.6 Cloud based application in android



Table 91.3 Test result

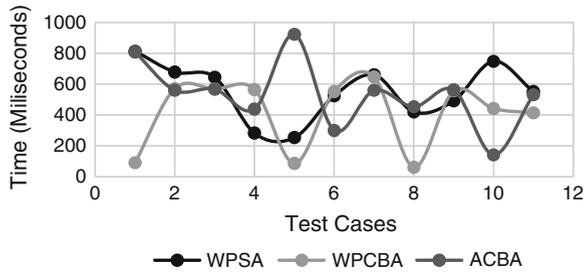
Weight	Height	Time (ms)		
		WPSA	WPCBA	ACBA
74 (kg)	1.74 (m)	813	91	808
57 (kg)	157 (cm)	678	571	560
59 (kg)	172 (cm)	645	568	568
71 (kg)	175 (cm)	283	565	438
65 (kg)	1.61 (m)	254	87	923
52 (kg)	166.5 (cm)	525	552	300
71 (kg)	156 (cm)	660	647	560
71000 (g)	172 (cm)	420	62	454
69 (kg)	173 (cm)	492	557	563
68 (kg)	1.65 (m)	748	442	142
Average		551.8	414.2	531.6

WPSA windows phone stand-alone application
 WPCBA windows phone cloud-based application
 ACBA android cloud based application

the measurement metrics for test data as the Business Process will convert the metric if it is not the same as required by the BMI service.

Figure 91.7 was generated based on Table 91.3. We are able to identify the pattern of total response time needed for every type of application. By average it shows that Cloud Based Mobile Application takes shorter time compared to stand-alone mobile application when there is a WSDL to be triggered or invoked.

Fig. 91.7 Test result



The results are also affected by the network load such as number of PC and the operations in the network. The experiment was not conducted in an empty network, as we were considering the network where usually mobile devices are used. Most mobile devices are used in 3G network rather than office or home private WIFI which is much slower compared High Speed Broadband.

91.5 Analysis and Discussion

Based on the result obtained, we are able to have a single process for all the application in various platforms. This was a problem raised by [11] where programmers have to develop some process again for multiple platform. Besides that, using service composition for mobile application will also help the developer to build any web applications in future using the same Business Process which is being created for mobile application.

Furthermore, as stated in [4], getting the status of the web service before a composition happen is very important but it's impossible to check the status of the web service every time we would like to use it. It will increase the burden to the server. For this experiment, we have implemented TRY CATCH blocks in our Business Process. We will catch the error in workflow level. If the developer will like to log the errors, they could request it from the return values in the workflow.

We are using the HRLPLA method in our service composition. It's a hierarchical based composition where the data used by parent process will be assigned to the child process. By using Cloud-Based Mobile application for service composition, it will increase the battery life time for the mobile device compared to stand-alone mobile application. This is because all the processing power will be handled by the cloud servers as compared to the stand-alone application which will use up the mobile device resources. For storage purpose, stand-alone application are not able to store a large amount of data compared to Cloud-Based Mobile Application as it will be using the cloud server storage to place all the data. With the proposed solution, mobile user can have more applications installed in the mobile device and have longer battery life time even with complex applications are loaded in the phone.

91.6 Conclusion

Based on the experiment that we have completed, we have proved that Cloud-Based Mobile application responds faster compared to a stand-alone application. This has been considered as a very important test as mobile application that has shorter loading time are the stand-alone applications, however when it comes to processing from cloud servers, Cloud-Based mobile application have shorter loading time. This application might not load as fast as stand-alone application for the UI, but it responds faster when it comes to invoking web services from different servers.

This project will be further improved to have more complex Business Process where it is able to dynamically select web services based on the mobile application requirement. This selection will be done in workflow level using software agents in multiple levels. There are a couple of issues in dynamic service composition for mobile application such as semantic discovery for web services and availability issues. We will be looking into all these issues in our future work and will be tested using more complex case studies.

References

1. Samwell, J.: What is web service composition. <http://stackoverflow.com/questions/1244425/what-is-web-service-composition>. Accessed 16 Oct 2013
2. Srivastava, B., Koehler, J.: Web service composition. Current solutions and open problems. In: ICAPS 2003 vol. 1, p. 23 (2003)
3. Lublinsky, B.: Service composition. <http://www.infoq.com/articles/lublinsky-soa-composition>. Accessed 25 Oct 2013]
4. Yan, H., Hui, W.: Constraints in web services composition, networking and mobile computing. In: 4th International Conference Wireless Communications, pp. 1–4 (2008)
5. Roy, G., Michael, J.C.: Model-driven semantic web service composition. In: Asia-Pacific Software Engineering Conference (APSEC'05), vol. 12, pp. 79–86 (2005)
6. Hongbing, W., Xiaohui, G.: An adaptive solution for web service composition. In: Services (SERVICES-1), pp. 503–510 (2010)
7. Jinghai, R., Xiaomeng, S.: A survey of automated web service composition methods. In: Proceedings of First International Workshop on Semantic Web Services and Web Process Composition. vol. 1(2), p. 67 (2004)
8. Jinghai, R., Dimitar, D., Paul, H., Norman, S.: A mixed initiative approach to semantic web service discovery and composition. In: SAPs Guided Procedures Framework. IEEE International Conference on Web Services, pp. 401–410 (2006)
9. Groenm, R., Jaeger, M.C.: Model-driven semantic Web service composition. In: 12th Asia-Pacific Software Engineering Conference (APSEC 2005), p. 8 (2005)
10. Pluralsight, M.: A developer's introduction to windows workflow foundation (WF) in .NET 4. <http://msdn.microsoft.com/en-us/library/ee342461.aspx>. Accessed 17 Oct 2013
11. Anthony I.W.: Software Engineering Issues for Mobile Application Development. Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn, pp. 397–400. ACM, Berlin (1996, 2010)

Chapter 92

An Effective Image Retrieval Method Based on Fractal Dimension Using Kernel Density Estimation

Zhang Qin, Huang Xiaoqing and Liu Wenbo

Abstract Fractal coding has been proved useful for image compression, and it is also proved effective for image retrieval. In the paper, we present a statistical method called variable bandwidth kernel density estimation to analyze fractal coding parameters. Then retrieve images using the retrieval index constructed with this method. Experimental results show that the proposed method with a variable optimized bandwidth performs better than those with a fixed bandwidth and the histogram method both in retrieval rate and retrieval speed. In this paper, the Average Retrieval Rate (ARR) can reach 72.40 %, which is more than that obtained by the existing methods.

Keywords Fractal coding · Kernel density estimation · Variable optimized bandwidth · Image retrieval

92.1 Introduction

Image retrieval has been an active research area for years, there are various kinds of image retrieval technique mainly based on text, content and semantic. The Content-Based Image Retrieval (CBIR) technique [1] is used to retrieve images with image features directly, so we can find the most similar images from the database through the comparison between the image features. Fractal coding parameters can effectively represent essential features of images. Fractal coding, as a new image compression technique, has been applied into image retrieval. Fractal features provide geometric information of an image that is irrelevant to the shape and size of an object in the image, therefore, fractal features are more robust than

Z. Qin (✉) · H. Xiaoqing · L. Wenbo
College of Automation Engineering, Nanjing University of Aeronautics and Astronautics
(NUAA), Nanjing, China
e-mail: Daisyzhangq@126.com

color and texture features. Meanwhile, retrieving images in fractal domain can be faster and more effective, especially for the compressed images.

Fractal image coding is a block-based scheme that exploits the self-similarity hiding within an image. Fractal features generated by the block-based scheme are quantitative measurements of self-similarity, therefore they can be used to construct image features. Fractal image compression was originally developed by Barnsley and Sloan [2]. Jacquin [3] implemented a block-based fractal compression, which is popularly known as fractal block coding. And fractal block coding has been applied into image retrieval. Pi MH proposed to employ the histogram of range block means and the 2D joint histogram of range block means and contrast scaling parameters as an image index [4, 5], and this technique greatly improved the retrieval rate. Some scholars proposed the histogram of collage error as an image signature and combined fractal parameters with collage error to improve the retrieval rate [6].

The features of an image can be acquired effectively with the statistical characteristics of fractal coding parameters, and the performance in image retrieval has already been confirmed. A statistical method called kernel density estimation is proposed, which can estimate the density of samples more accurately. Compared with the commonly used histogram method, the kernel density estimation can be more accurate and smooth. Therefore, we apply this method into image retrieval. Since the bandwidth of kernel function plays an important role in kernel density estimation, we propose the method with a variable optimized bandwidth in conformity with data [7]. Experimental results show that this method has not only higher retrieval rate but also less retrieval time than the existing methods.

The rest of the paper is organized as follows. The Sect. 92.2 introduces fractal coding and collage error. Section 92.3 presents the proposed method. Section 92.4 presents the performance evaluation. The Sect. 92.5 presents conclusions and future work.

92.2 Fractal Coding and Collage Error

92.2.1 Fractal Coding

In this paper, the orthogonalization fractal coding method is adopted [8]. An image ($M \times M$) is first segmented into non-overlapping blocks of size $B \times B$ called range blocks, recorded as R_1, R_2, \dots, R_m . A domain block pool Ω is a set of domain blocks of size $2B \times 2B$, generated by dividing the ($M \times M$) image into overlapped blocks. And the domain blocks are recorded as D_1, D_2, \dots, D_n . In general, $B = 2^t$, t is an integer. After the 4-neighborhood pixel average and compression transform, the domain blocks are mapped into the images with size $B \times B$. To improve the quality of the images, eight kinds of isometric transform are applied into the domain blocks. (In Jacquin's scheme, rotation transformation of $0^\circ, 90^\circ, 180^\circ,$

270°, vertical midline, horizontal midline and diagonal reflection transformation of 45°, 135° are proposed)

According to the Partitioned Iterated Function System (PIFS), we can find out the domain blocks matched with the range blocks using affine transformation iterations.

$$R' = \bar{r}U + s\rho(D - \bar{d}U) \tag{92.1}$$

For each range block R , orthogonalization fractal block coding is obtained by minimizing the following equation

$$E(R, D) = \|R - \bar{r}_i U - s_j \rho(D - \bar{d}U)\|^2 \tag{92.2}$$

The above minimization is performed over $D \in \Omega$ by working with a set of pre-quantized fractal parameters $\{\bar{r}_i\}_{i=1}^I$ and $\{s_j\}_{j=1}^J$ (I and J are the quantization levels for \bar{r}_i and s_j , respectively). Note that U is a matrix whose elements are all ones, s is a contrast scaling parameter, ρ is the isometric transform, $\|\cdot\|$ is the 2-norm and \bar{r} and \bar{d} are the average of range block and domain block respectively. \bar{r}_i is the average of the i th range block. Since $\langle U, D - \bar{d}U \rangle = 0$, we define Eq. (92.2) as orthogonalization fractal block coding. Then range block R can be written as $(\bar{r}, s, x_D, y_D) = \arg \min_{D \in \Omega} E(R, D)$. Where (x_D, y_D) is the top-left corner coordinate of the ‘best-matching’ domain block.

When all the domain blocks matched with the range blocks are found, the fractal coding of the whole image is completed.

92.2.2 Collage Error

We define the collage error as follows:

$$e = \hat{E}(R) = \min_{D \in \Omega} \sqrt{\frac{E(R, D)}{B \times B}} \tag{92.3}$$

Collage error is a quantitative measure of the similarity between range block and “best-matching” domain block. It is relatively robust compared with other fractal parameters which can be quite sensitive to changes in domain block pool. Pi MH [9] has proved that the proposed indices not only reduce computational complexities, but also enhance the retrieval rate, compared with the existing fractal-based retrieval methods.

92.2.3 Program Code

```

[imagem imagen]=size(Image1);
Sr=4;Sd=8;
Rnum=(imagem/Sr)*(imagen/Sr);
Dnum=(imagem/Sd)*(imagen/Sd);
Image2=zeros(Dnum,Sr,Sr);
Image2=blkproc(Image1,[Sd/Sr,Sd/Sr],'mean(mean(x)');
RBlocks=zeros(Rnum,Sr,Sr);
DBlocks=zeros(Dnum,Sd,Sd);
DBlocksReduce=zeros(Dnum*8,Sr,Sr);
for i=1:image/Sr
    for j=1:image/Sr
        k=(i-1)*imagen/Sr+j;
        RBlocks(k,:,:) = Image1((i-1)*Sr+1:i*Sr,(j-1)*Sr+1:j*Sr);
    end
end
for i=1:image/Sd
    for j=1:image/Sd
        k=(i-1)*imagen/Sd+j;
        m=Sr; n=Sr;
        DBlocksReduce(k,:,:) = Image2((i-1)*Sr+1:i*Sr,(j-1)*Sr+1:j*Sr);
        DBlocksReduce(k+Dnum,:,:) = DBlocksReduce(k,m:-1:1,:);
        DBlocksReduce(k+2*Dnum,:,:) = DBlocksReduce(k,: ,n:-1:1);
        DBlocksReduce(k+3*Dnum,:,:) = DBlocksReduce(k,m:-1:1,n:-1:1);
    DBlocksReduce(k+4*Dnum,:,:) = reshape(DBlocksReduce(k,:,:),Sr,Sr)';
        A=reshape(DBlocksReduce(k+3*Dnum,:,:),Sr,Sr)';
        DBlocksReduce(k+5*Dnum,:,:) = A(: ,n:-1:1);
    DBlocksReduce(k+6*Dnum,:,:) =
    imrotate(reshape(DBlocksReduce(k,:,:),Sr,Sr),90);
    DBlocksReduce(k+7*Dnum,:,:) =
    imrotate(reshape(DBlocksReduce(k,:,:),Sr,Sr),270);
        DBlocks(k,:,:) = Image1((i-1)*Sd+1:i*Sd,(j-1)*Sd+1:j*Sd);
    end
end
RandDbest=zeros(Rnum,1)+256^3;
RandDbests=zeros(Rnum,1);
RandDbesto=zeros(Rnum,1);
RandDbestj=zeros(Rnum,1);
for i=1:Rnum
    x=reshape(RBlocks(i,:,:),Sr*Sr,1);
    meanx=mean(x);
    for j=1:Dnum*8
        y=reshape(DBlocksReduce(j,:,:),Sr*Sr,1);
        meany=mean(y);
        s=(x-meanx)'*(y-mey)/( (y-mey)'*(y-mey) );
        o=(meanx-s*meany);
        e=(x-s*y-o)'*(x-s*y-o);
        if (RandDbest(i)>e)
            RandDbest(i)=e;
            RandDbests(i)=s;
            RandDbesto(i)=o;
            RandDbestj(i)=j;
        end
    end
end
end
end

```

92.3 Proposed Method

Kernel Density Estimation (KDE), as a popular nonparametric density estimation, is widely used in the field of pattern recognition, classification and image processing. The histogram is the simplest non-parametric density estimation method which is frequently used. It has been demonstrated that histograms of fractal parameters capture statistical characteristic of texture images effectively [4]. Since the histograms heavily depends on width of bins and end points of bins, different bins and end points may result in different histogram distribution, meanwhile, different distributions lead to different results of image retrieval, and thereby affect the retrieval rate.

On the condition that the densities of data are unknown, the kernel density estimation is used, and it has been applied into statistical characteristics of images in massive literatures [10, 11] and has obtained considerable results. The properties of kernel density estimation are, as compared to histograms, smooth, no end points and they depend on bandwidths heavily. This method is more simple and effective than the histogram method, which reduces the computational complexity of data in image retrieval.

92.3.1 Kernel Density Estimation of Fractal Parameters

Let x_1, x_2, \dots, x_n be an i.i.d (independent and identically distributed) sample drawn from some distribution with an unknown density $p(x)$. We are interested in estimating the shape of this function $p(x)$. Its kernel density estimation is

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (92.4)$$

where $K(\bullet)$ is the kernel—a symmetric but not necessarily positive function that integrates to one—and h is a smoothing parameter called the bandwidth. Intuitively we want to choose h as small as possible, which will lead to instability. However, we should balance the smoothness and stability of the estimation. However, it's the bandwidth not the function of the kernel that exhibits a strong influence on the estimation results. A bandwidth h of the kernel may alter the density estimation, and it can accordingly affect the goodness-of-fit of the density function $p_n(x)$ to the unknown underlying density $p_n(x)$. Generally the most common optimal criterion used to select bandwidth is the Mean Integrated Squared Error (MISE) [12]. This principle is applied to select a fixed optimal bandwidth. The optimal bandwidth is the argument that minimizes the MISE.

$$MISE(\hat{p}_n(x)) = E \int [\hat{p}_n(x) - p_n(x)]^2 dx \tag{92.5}$$

The integrand of the MISE can be decomposed into three parts: $E\hat{p}_n^2 - 2p_nE\hat{p}_n + p_n^2$. Then we will subtract p_n^2 from the MISE since it is the underlying density and does not depend on the choice of a kernel, thus the cost function as a function of the bandwidth is defined as

$$MISE' = MISE - \int_a^b p_n^2(x)dx = \int_a^b E\hat{p}_n^2 dx - 2 \int_a^b p_n E\hat{p}_n dx \tag{92.6}$$

Here (a, b) is an interval of interest, and the interval length is H . The minimum of the cost function Eq. (92.6) is an estimate of the fixed optimal bandwidth, which is denoted by h^* .

After obtaining the fixed bandwidth, we will introduce the proposed method to obtain a variable bandwidth. First we define a formula:

$$y = \frac{1}{n} \sum_{i=1}^N \delta(x - x_i) \tag{92.7}$$

where n is the number of estimated point. $\delta(t)$ is the Dirac delta function. The kernel density estimation is obtained by convoluting a Gaussian kernel $k(s)$ to y .

$$\hat{p}_n = \int y_{x-s}k(s)ds \tag{92.8}$$

The most commonly used kernel is Gaussian kernel function:

$$k(s) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{s^2}{2h^2}\right) \tag{92.9}$$

As we know, the fixed bandwidth is selected from an entire observation interval (a, b), however, the estimation may be improved by using a kernel bandwidth which is adaptively selected in conformity with data. Thus, the kernel density estimation with the variable bandwidth h_x is expressed as

$$\hat{p}_n = \int y_{x-s}k_{h_x}(s)ds \tag{92.10}$$

Here, we provide a method for obtaining the variable bandwidth h_x that minimizes the MISE by optimizing a local interval length among which the variable bandwidth can be regarded as a fixed one. To conduct the local optimization, we introduce the local MISE criterion as

$$localMISE = E \int [\hat{p}_n(x) - p_n(x)]^2 \rho_H dx \tag{92.11}$$

The weight function ρ_H localizes the integration in the interval H . According to Eq. (92.6), we introduce the local cost function by subtracting the term irrelevant for the choice of h as

$$C_n(h, H) = localMISE - \int p_n^2 \rho_H dx = \frac{1}{n^2} \sum_{i,j} \psi_{h,H}(x_i, x_j) - \frac{2}{n^2} \sum_{i \neq j} k_h(x_i - x_j) \rho_H^{x_i - x} \tag{92.12}$$

where

$$\psi_{h,H}(x_i, x_j) = \int k_h(u - x_i) k_h(u - x_j) \rho_H^{u-x} du \tag{92.13}$$

The optimal bandwidth h^* varies according to different interval length H . We suggest selecting an interval length that scales with the optimal bandwidth as $\gamma^{-1}h^*$, The parameter γ is a smoothing parameter for the variable bandwidth, it regulates the interval length for local optimization. With small γ , the variable bandwidth fluctuates slightly, while with large γ , the variable bandwidth fluctuates significantly.

In order to select a variable kernel bandwidth, firstly, compute the local cost function $\psi_{h,H}(x_i, x_j)$ in Eq. (92.13) and find h^* that minimize the Eq. (92.13). Then repeat the procedure above while changing H . Change H to find H^* that satisfies $H = \gamma^{-1}h^*$. We could obtain the variable bandwidth by computing the cost function

$$\hat{C}_n(\gamma) = \int_a^b \hat{p}_n^2 dx - \frac{2}{n^2} \sum_{i \neq j} k_{h_\gamma}(x_i - x_j) \tag{92.14}$$

where $\hat{p}_n = \sum_i k_{h_\gamma}(x - x_i)$. At last, we should repeat the procedure above to find γ^* that minimizes $\hat{C}_n(\gamma)$, then apply it to obtain the variable bandwidth h_γ^* . The bandwidth is what we want to calculate kernel density estimation more precisely.

It has been proved that range block mean \bar{R} , contrast scaling parameter s and collage error e [4, 6, 9] are effectively used to retrieve images. In this paper, we apply these parameters into kernel density estimation directly. The optimal uniform quantization for \bar{R} is $\{0, 1, 2, \dots, 63\}$ [13] and to ensure the convergence of the decoding, the scaling factor s is restricted to the interval $(-S_{max}, S_{max})$, where $0 < S_{max} < 1$, as for collage error e , it is real-valued, hence, before we calculate kernel density estimation of collage errors, they are rounded into the closest integer if collage errors are smaller than $T - 1$, or are set as $T - 1$ if they exceed $T - 1$ (T is a user-specified threshold). In this paper, we set $T = 20$. Then we

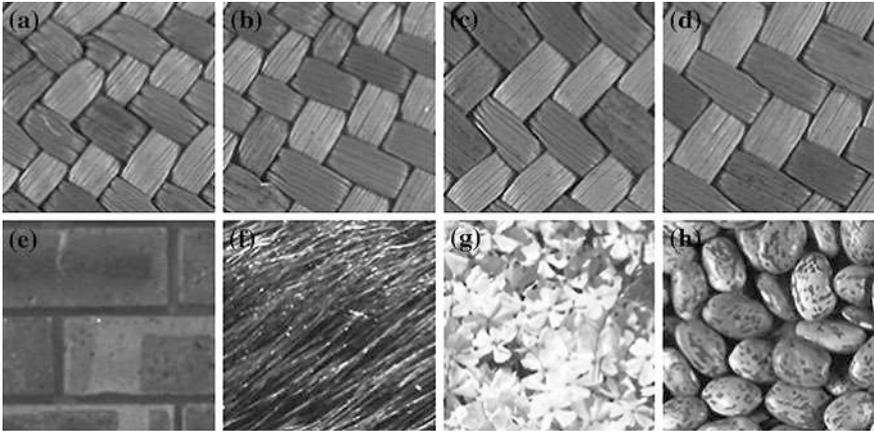


Fig. 92.1 Examples of 128×128 images. **a–d** four similar images; **e–h** four different images

calculate kernel density estimation with the processed collage errors and fractal coding parameters.

Figure 92.1 shows four similar and four different texture images from VisTex texture database [14]. Kernel density estimation of range block mean \bar{R} , contrast scaling parameter s and collage error e corresponding to these images are plotted in Fig. 92.2. In most cases, the curves are close for similar texture images, and different for the dissimilar texture images.

The left column of Fig. 92.2 shows the kernel density estimation of range block mean \bar{R} , contrast scaling parameter s and collage error e respectively according to the first four similar images (a–d). Obviously the curves are close for the similar texture images.

The right column of Fig. 92.2 shows the kernel density estimation of range block mean \bar{R} , contrast scaling parameter s and collage error e respectively according to the other four different images(e–h). We can see that the curves are different for the dissimilar texture images.

92.3.2 Similarity Measurement

We define that $V_Q(\bullet) = \{u_1, u_2, \dots, u_H\}$ and $V_C(\bullet) = \{v_1, v_2, \dots, v_H\}$ are the features of the query and candidate images respectively. The vectors of range block mean \bar{R} , contrast scaling parameter s and collage error e of a query image are expressed as $\{u_{R1}, u_{R2}, \dots, u_{RH}\}$, $\{u_{S1}, u_{S2}, \dots, u_{SH}\}$ and $\{u_{e1}, u_{e2}, \dots, u_{eH}\}$ respectively. The same with vector V_C for candidate images.

In addition, the subscript variable H represents the amount of range block mean \bar{R} , contrast scaling parameter s or collage error e of an image. To measure the

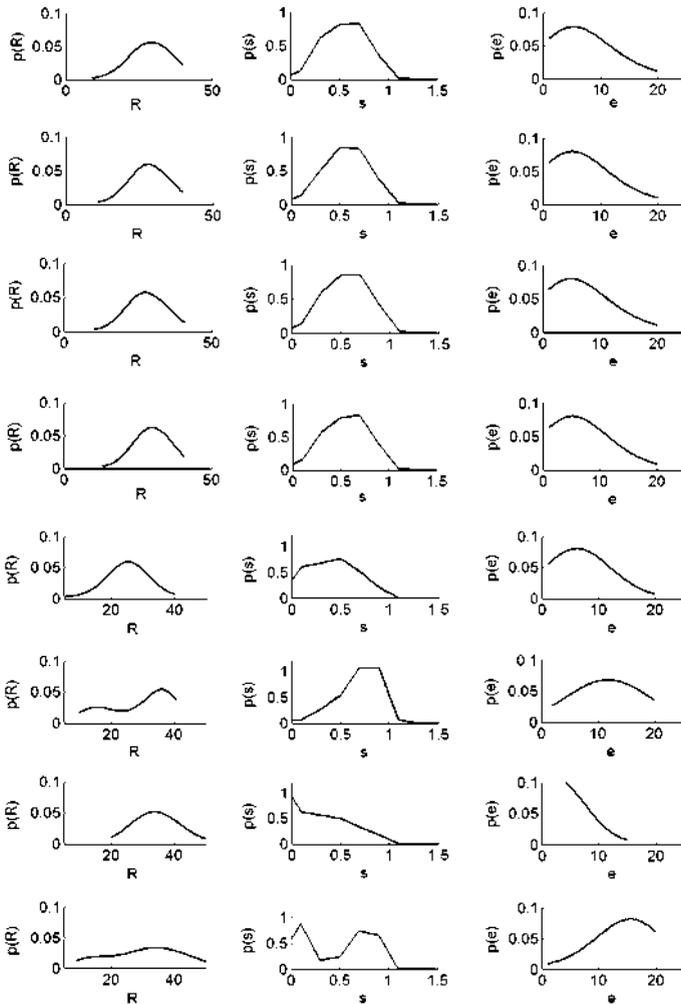


Fig. 92.2 Comparison of the KDE of fractal coding parameters

similarity between two images, we can calculate the deviation between their fractal coding parameters instead since these parameters can express images compactly.

In this paper, we adopt the most commonly used L_2 and KLD (Kullback-Leibler divergence) [15] as the distance criterion to measure the similarity between the query and the candidate images, the distance between the two images is calculated as follows:

$$d_{L_2}(Q, C) = \sqrt{\sum_{b=1}^H (u_b - v_b)^2} \quad (92.15)$$

$$KLD(Q, C) = \sum_{b=1}^H u_b \log\left(\frac{u_b}{v_b}\right) \quad (92.16)$$

The distances between similar images are much smaller than those between dissimilar images. Experiments show that the similarity measurement Eq. (92.7) performs much better than Eq. (92.8). Thus we only discuss the retrieval rate using L_2 distance. The obtained distances are sorted in an ascending order.

92.4 Performance Evaluation

We have performed experiments on VisTex texture database. The set of VisTex is the classical selection of 40 classes of texture images that are used by many literatures for image retrieval [16]. The images are listed as follows: Bark0, Bark6, Bark8, Bark9, Brick1, Brick4, Brick5, Buildings9, Fabric0, Fabric4, Fabric7, Fabric9, Fabric11, Fabric14, Fabric15, Fabric17, Fabric18, Flowers5, Food0, Food5, Food8, Grass1, Leaves8, Leaves10, Leaves11, Leaves12, Leaves16, Metal0, Metal2, Misc2, Sand0, Stone1, Stone4, Terrain10, Tile1, Tile4, Tile7, Water5, Wood1, and Wood2. These are real world 512×512 images from different natural scenes. Only gray-scale levels of the images are used. As for our experiments each image is divided into sixteen non-overlapping 128×128 sub-images, thus creating a test database of 640 texture images.

In the retrieval experiments, each sub-image in the database is used once as a query image. For comparison purpose, retrieved images are the first 16 most similar images for each query. The relevant images for each query consist of all the sub-images from the same original texture image. All experiments are conducted on a 2 GHz PC using Matlab7.8.0 as a programming tool.

The retrieval method using feature vectors of range block mean, contrast scaling and collage error are respectively named as KM (kernel density estimations of range block mean), KS (kernel density estimations of contrast scaling) and KE (kernel density estimations of collage error). A large number of texture images in database are used to do testing, from which we could conclude that the KE method works much better than the KM and KS. Therefore, we apply the KE method into images retrieval in this paper.

Table 92.1 Retrieval rate of three methods (%)

Image	HE	FKE	VKE
Bark0	16.80	21.09	52.34
Bark6	14.84	11.72	37.50
Bark8	22.27	20.31	47.65
Bark9	21.48	22.66	46.87
Brick1	22.27	39.06	64.84
Brick4	34.77	30.86	83.98
Brick5	21.48	24.22	76.56
Buildings9	21.88	26.17	75.39
Fabric0	46.88	41.80	76.56
Fabric4	36.80	41.48	75.00
Fabric7	33.69	38.91	75.78
Fabric9	58.20	59.53	74.21
Fabric11	38.28	42.58	76.56
Fabric14	35.94	40.63	74.60
Fabric15	52.34	46.88	76.56
Fabric17	30.47	38.67	74.60
Fabric18	25.00	46.88	75.39
Flowers5	57.03	58.59	73.82
Food0	67.19	63.44	81.17
Food5	65.40	28.52	73.43
Food8	46.48	51.17	76.17
Grass1	32.03	42.50	73.82
Leaves8	36.72	64.06	76.17
Leaves10	23.83	27.34	75.78
Leaves11	48.44	57.81	75.00
Leaves12	42.58	57.81	76.56
Leaves16	50.47	32.03	71.87
Metal0	38.67	54.53	81.17
Metal2	34.77	48.67	76.09
Misc2	39.45	46.25	77.73
Sand0	37.11	35.16	73.04
Stone1	26.56	33.44	78.12
Stone4	31.25	33.98	75.00
Terrain10	37.97	22.66	77.34

(continued)

Table 92.1 (continued)

Image	HE	FKE	VKE
Tile1	72.27	63.28	85.82
Tile4	69.92	74.22	88.04
Tile7	63.28	71.41	86.60
Water5	25.00	22.27	70.31
Wood1	22.66	34.27	75.39
Wood2	38.75	38.13	87.10

Table 92.2 Average Retrieval Rate (ARR) compared with other literatures (%)

Method	HE	FKE	literature [3] GGD + MM	literature [4] rmm + CT	Proposed method (VKE)
ARR (%)	35.42	39.88	67.27	69.52	72.40

92.4.1 Average Retrieval Rate and Retrieval Speed

We use 40 512 × 512 VisTex texture images. Each image is divided into 16 128 × 128 non-overlapping sub-images. Finally a test database of Z = 640 texture images is created. Each sub-image is encoded using full search. Let the number of ideally retrieved images of one class be denoted by F (in this case F = 16) and m_z be the number of correctly retrieved images of one class from the top 16 images at the z-th test. The performance is measured in Average Retrieval Rate (ARR) that is defined the same with literature [6], which is then calculated as

$$ARR = \frac{\sum_{z=1}^Z m_z}{F \times Z} \tag{92.17}$$

Experiments show that KE (kernel density estimation of collage error) method works better than the others. We can see from Table 92.1 and Table 92.2 that our method performs better than HE (histogram estimation) method, FKE (fixed bandwidth kernel density estimation of collage error) method and other methods.

From Table 92.2 we can see that the ARR of the proposed method is 72.40 %, which is more that the other listed methods. The runtime of the retrieval, which is completely determined by the performance of the similarity measurement process, is also a key index to indicate the performance. Compared with literature [16], our runtime is largely reduced since the basic arithmetic operations are adopted in our method, while the computationally expensive log, e^x and x^r operations with more iterations are applied in literature [16], which leads to an increase in computation time.

92.5 Conclusion

In this paper, we apply orthogonalization fractal coding algorithm into image retrieval, which has been verified that the decoding speed is higher than that of the basic fractal coding. Meanwhile, we propose an image retrieval method based on fractal coding parameter with a variable optimized bandwidth kernel density estimation method. The kernel bandwidth can adjust according to the data distribution. Thus, the statistical characteristics of fractal coding parameters are employed as retrieval indices. Experiments show the superiority in both retrieval rate and retrieval speed when compared with the existing methods. In the future, we will combine some other features of images with fractal parameters to improve performance of image retrieval.

References

1. Chun, Y.D., Kim, N.C., Jang, I.H.: Content-based image retrieval using multi-resolution color and texture features. *IEEE Trans. Multimedia* **10**, 1073–1084 (2008)
2. Barnsley, M., Sloan, A.D.: A better way to compress images. *Byte* **13**, 215–233 (1988)
3. Jacquin A.E.: Fractal image coding: a review. *Proc. IEEE* **81**(10), 1451–1465 (1993)
4. Pi M, Mandal M, Basu A.: Image retrieval based on histogram of new fractal parameters. In: *Proceedings of the ICASSP*, vol. 3, pp. 585–588 (2003)
5. Pi M., Mandal M., Basu A.: Image retrieval based on histogram of fractal parameters. *IEEE Trans. Multimed.* **7**(4), 597–605 (2005)
6. Pi, M., Li, H.: Fractal indexing with the joint statistical properties and its application in texture image retrieval [J]. *IET Image Proc.* **2**(4), 218–230 (2008)
7. Shimazaki, H., Shinomoto, S.: Kernel bandwidth optimization in spike rate estimation. *J. Comput. Neurosci.* **29**, 171–182 (2010)
8. Øien, G.E., LepsØy, S.: A class of fractal image coders with fast decoder convergence. *Sign. Process.* **40**(1), 105–117 (1994)
9. Pi, M., Tong, C.S., Basu, A.: Improving fractal codes based image retrieval using histogram of collage errors. In: *2nd International Conference on Image and Video Retrieval*, vol. 2728, pp. 121–130 (2003)
10. Xiaoqing, H., Qin, Z., Wenbo, L.: A new method for image retrieval based on analyzing fractal coding characters. *J. Vis. Commun. Image Represent.* **24**(1), 42–47 (2013)
11. Zhi, L., Ran, S., Liquan, S.: Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. *IEEE Trans. Multimedia* **14**(4), 1275–1289 (2012)
12. Deng Biao, Yu Chuanqiang, Li Tianshi, Su Wenbin, Pan Yangke.: Dual-bandwidth kernel density estimation algorithm based on estimate points. *Chinese Journal of Scientific Instrument*, vol.32(3), pp.615-620 (2011)
13. Tong, C.S., Pi, M.: Fast fractal image compression using adaptive search. *IEEE Trans. Image Process.* **10**, 1269–1277 (2001)
14. VisTex texture database. <http://vismod.media.mit.edu/vismod/imagery/VisionTexture/>
15. Do, M.N., Vetterli, M.: Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process.* **11**(2), 146–158 (2002)
16. Kwitt, R., Uhl, A.: Lightweight probabilistic texture retrieval. *IEEE Trans. Image Process.* **19**(1), 241–253 (2010)

Chapter 93

Bio Terapi Solat: 3D Integration in Solat Technique for Therapeutic Means

Arifah Fasha Rosmani, Noor Azura Zainuddin, Siti Zulaiha Ahmad and Siti Zubaida Ramli

Abstract Solat has been known for its health benefits, however not many people are aware of it. Solat involves stretching that is in harmony with the body needs for health care. Therefore, this research focuses on solat guidance in three dimensional (3D) ways. The purpose of this research is to show the complete solat techniques as a therapy to the human body and health using 3D model. This application is developed using 3D Autodesk Maya software that includes the human model, movement (rigging) and other multimedia elements that can create interactivity in this application. The application shows a guideline of learning a perfect Solat Technique and the benefit of solat in the therapeutic means. Usability test is performed to observe the usability and learnability aspects of the application. Respondents were selected among adults to experience this application. It eventually guides the users to focus on solat therapy techniques and its benefits.

93.1 Introduction

There are two types of solat: solat fardhu and solat sunat, solat fardhu is the second Rukun Islam that is compulsory to all Muslims. Solat is mandatory and it must be performed five times every day. Solat (dhuhr, asr, maghrib, isha', and fajr) is important to each Muslim because through solat, Muslims can get closer to Allah S.W.T and avoid doing bad things. In the Al-Quran it is also stated that solat is mandatory to be performed as follows:

A.F. Rosmani (✉) · S.Z. Ahmad · S.Z. Ramli
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 02600 Arau,
Perlis, Malaysia
e-mail: arifah840@perlis.uitm.edu.my

N.A. Zainuddin
Academy of Contemporary Islamic Studies, Universiti Teknologi MARA, 02600 Arau,
Perlis, Malaysia



Which means: Maintain with care the [obligatory] prayer and [in particular] the middle solat and stand before Allah, devoutly obedient. (Al-Baqarah, 238)

The benefits from solat can make it an option for medical therapies for those who perform it. Unfortunately, many people do not realize that solat can be a good health treatment and they do not know that the solat technique can be a therapy to relax their mind [1].

Solat has many techniques and since Islam is simple, it does not burden Muslims in performing the solat. There are many techniques for solat in this world according to different *mazhabs* and manners of people. However, people do not know which one is a better technique for them to follow. As a Muslim, solat has to be carried out in the manner of solat advocated by the prophet (Nabi Muhammad S.A.W) [2].

Most of the guides for learning the solat technique for solat therapy are available in written form. The materials, such as books, make the readers bored and they are not interested to learn solat. An effective way to make it more interesting and engaging is by illustrating Solat Technique using 3D animation. 3D model of animation is intended to connect users with the application. There are many 3D softwares available in the market. However, 3D Autodesk Maya has been chosen based on the extensive creative features such as 3D computer animation, character modeling, visual effects, rendering, and compositing process.

Therefore, this study intends to develop an application to show the solat technique in 3D animation because the existing application uses limited materials for learning solat such as books, video-based learning, and 2D models. As the 3D is very popular and interesting in the real world environment, this project utilizes the technology to develop 3D animation as an alternative to learn solat and becomes a technique for therapy.

93.2 Research Background

This project provides an alternative way to learn the solat technique for therapeutic means for all Muslims. It focuses on the 3D animation as a way of learning Solat Technique. This project is suitable for all users but the main targets of users are children from 7 years old to teenagers around 18 years old. The Prophet (Nabi Muhammad S.A.W) said: *Order your children to pray at the age of seven. And beat them (lightly) if they do not do so by the age of ten.* From the hadith it tells that children should start practicing solat from the age of 7. Hence, the development of this project enables children to interact and understand how to learn solat.

Children can learn the correct way or the technique of solat as the correct way of Solat Technique is good for health.

Besides that, this project can also be used by Muslims to introduce the benefit from solat technique to the other Muslims who are not aware of it. The development of this project is also to illustrate the right movement and the technique of breathing during solat.

93.2.1 Solat

Solat in the Arabic language is a worshiping act with physical and body movements as well as a silent Quranic recitation though mind and soul. This Muslim way of worshiping also involves some *quranic* recitations and as well as supplications (*du'a*). These specific recitations and supplications must be verbalized when the worshipper assumes certain positions and performs movements between positions [3].

The Messenger of Allah S.W.T, Rasulullah S.A.W, says: *There are five solat that Allah obligated the slaves to perform. Whoever performs them properly without belittling their obligation, Allah promised to admit him into Paradise. Whoever leaves them out does not have a promise from Allah to have Paradise without torture before. If He willed, He tortures him, and if He willed, He forgives him.* (Narrated by Ahmad in his Musnad.)

It is obligatory to perform each of these five solat in its due time. It is better to perform each solat earlier in its due time. Solat is also physical as well as a spiritual act which involves total obedience and submission to Allah S.W.T.

93.2.2 Solat Techniques

Each solat has continuous sequences of body movements consisting of bowing, kneeling, prostrating, and sitting. When performing solat, most of the muscles and joints are in action. All the movements when performing the solat fardhu 17 *raka'at* per day, is equal to 119 physical postures per day, 3570 per month, and 43,435 per year. The movements increase went performing the other optional solat. Through the solat, unique postures are possible which can give significant reactions to the human body.

93.2.3 Solat as a Therapy

Solat helps to remove all sources of tension due to the constant change of movements. It is known that such changes lead to an important physiological

relaxation. Therefore, the Prophet commands that the Muslims, when afflicted with a state of anger, should resort to prayers. It is proved that prayer has an immediate effect on the nervous system as it calms agitation and maintains balance. They are as well a successful treatment for insomnia resulting from nervous dysfunction [4].

In order to visualize the implementation of therapeutic benefit in Solat, 3D model with animation should be integrated to increase the learnability and awareness among Muslim.

93.2.4 Dimensional (3D) Model

Three Dimensional is one of a techniques to create or build interactive multimedia and animation film or cartoon [5]. Multimedia is more interesting when 3D animation is used rather than 2D animation. According to Yongguang et al. [6], 3D describes an image that provides the view in depth. When 3D images are made interactive, the users will feel involved with the scene. Nowadays, 3D has already developed in the film industry and cartoon production and now is growing in the field of education, game, and also in advertisement.

93.3 Related Works

A few observations have been done on the previous available solat applications to find the advantages and disadvantages of each application before developing our own application. Figure 93.1 shows solatSim application which is a review of each position in *Zuhur* obligatory Solat. It also contains male and female respondent's image illustration solat position. When the respondents do the technique of solat, SolatSim also view the blood circulation flow and the muscle condition for each position of solat [4]. In SolatSim it include five (5) main topics which are "Male", "Female", "Research", "Credit" and "Solat Info & Demonstration".

"The Right Way to Pray" is another application that is related with this research as shown in Fig. 93.2. It includes all the techniques of solat that can guide users to learn solat and includes information about Islam. This application is developed by Islamic finder at <http://www.islamicfinder.org/prayer/index.html>. This application has provided a user friendly interface to users. It is also interactive and users can control the page. It includes buttons such as "back", "repeat" and "next". It can be repeated by user if they did not understand on certain topics. However, the disadvantage of this application is the model is not clear. The developer used a blur model to describe each of the movements of solat. Users might not see clearly the right position of solat. Moreover, it also does not include the Arabic writing of the positions of solat.



Fig. 93.1 SolatSim interface



Fig. 93.2 The right way to pray interface

93.4 Methodology

In this research, three phases were involved which are Designing, Developing, and Testing. The particular activities of all phases are briefly explained as below.

93.4.1 Design Site Map

A site map provides the flow of content of all processes that are included in the application. Figure 93.3 shows the site map of this project. The site map shows the contents of the application. This application has 3 menus which are “Sunnah Rasulullah S.A.W”, “Teknik Solat” and “Glosari”. “Teknik Solat” is the main menu of this application. In the site map, each page contains a “KELUAR” button to ease users to exit this application. When a user clicks the “KELUAR” button, a message box will be appeared to ask users for confirmation to exit the application. Besides that, this application includes maximizing and minimize button to ease its usage.

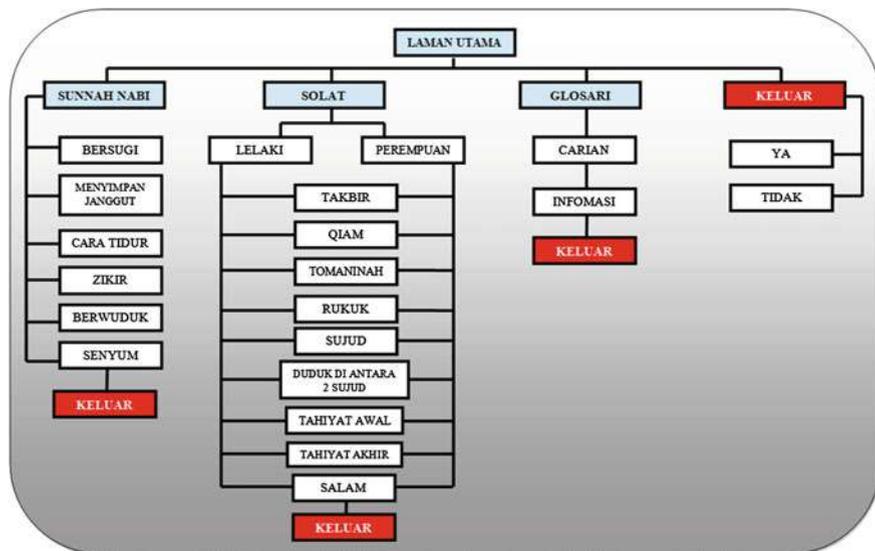


Fig. 93.3 Site map of “Bio Terapi Solat” application

93.4.2 Design Storyboard

The storyboard is designed to show the flow of the application. Usually it is sketched on a piece of paper before designing. It is easier to design the interface and get an idea when it is sketched on paper. When the storyboard has been completed, the sketch of the interface will be scanned using a scanner to make it in a digital form.

93.4.3 Development

The storyboard made will be used as a guide to draw the button and the interface. The button plays an important role to enable users to navigate in an application. Each element in the application must match the interface. Color and appropriate font is made to attract users to use this application. The home icon is used to identify the “Laman utama” application “Bio Terapi Solat.” The book icon is used to identify a button of “Sunnah Rasulullah S.A.W”. “Solat Teknik” icon that shows a picture of the bow position is used to show the button for “Solat Teknik”. While the icon “Glosari” is used to visualize the search to find a piece of information. Figure 93.4 shows the button in the application using icon.



Fig. 93.4 Button for “Bio Terapi Solat” application

Fig. 93.5 Interface of “Laman Utama”



93.4.3.1 Main Menu

The home page is the first interface for this application. On the homepage, there is information about this application. It also has 4 buttons including “Laman Utama” (Main Menu), “Sunnah Rasulullah S.A.W”, “Teknik Solat” (Solat Technique) and “Glosari” (Glossary). In addition, there are buttons to exit, minimize and maximize the application. Figure 93.5 shows the interface of “Laman Utama”.

93.4.3.2 Sunnah Rasulullah Page

On this page, there are six buttons of “Sunnah Rasulullah S.A.W”, users can click on each button to access all the available “Sunnah”. This application comes with “Sunnah Rasulullah S.A.W” to provide additional information to the users. When a user clicks the button, the information will be displayed on the “Sunnah”. Figure 93.6 shows an example interface for “Sunnah” information.

93.4.3.3 Terapi Solat Page

Figure 93.7 is the interface to see prayer techniques in 3D animation. Users can select the desired prayer technique. There are nine techniques provided as “Takbir”, “Qiam”, “Tomaninah”, “rukuk”, “sujud”, “duduk di antara dua



Fig. 93.6 “Sunnah Rasulullah” page



Fig. 93.7 Terapi solat interface

sujud”, “Tahiyad Awal”, “Tahiyad Akhir” and “Salam”. For every single movement in “Solat”, there are many beneficial elements that can provide the therapeutic means for Muslims that have been emphasized in the application. For example as shown in Fig. 93.7, the benefits for solat movements have been included in “Terapi” and “Teknik” buttons where the correct way of performing solat and its benefits are highlighted.

93.4.3.4 Glossary Page

The glossary is a page that allows the user to access information contained in this multimedia application. It contains information such as a dictionary and only contains jargons such as medical terms that may not be understood by the users. Users can enter a word in the textbox and if the requested information is in the database then it will be displayed as shown in Fig. 93.8.

93.5 Usability Testing

After the completion of Development Phase, usability test is carried out to get the feedback on flaws of the application. Generally, usability testing is the method used to test prototypes and find out the ease of use level. The participants in

Fig. 93.8 Glossary page



Usability Test are real users. From the testing, the developer reviews the attitude of the users towards the project. Participants are required to complete the designed tasks under simulation scenarios—usually in the laboratory. It can help to find out the problems and modify the design in the early stages of product design [7].

The usability testing was done by the lecturers of UiTM, Perlis. The application has been tested on the lecturers in Center of Islamic Teachings and Understanding (CITU) and Computer Science Department. Besides that, participants of usability testing are the students of UiTM, Perlis. This usability testing had been done by determining the navigation, design, sound, graphic, animation, and interactivity between the user and the application. Participants were asked to perform a questionnaire after exploring the application; they were also required to complete the testing. After exploring applications, developer’s has distributed a set of questionnaires to the participants. They were given 5–10 min to complete the relevant questionnaire.

93.5.1 Usability Test Result

From the observation, it is found that the participants were confused with the Glossary interface. They did not know what terms that are provided in the Glossary and it does not provide any direction or instruction to the participants. The “Terapi” and “Teknik” buttons also made the participant uncomfortable with them. The buttons required the user to hold the mouse click to enable them to read the information. However, these drawbacks have been refined and suitable enhancement has been implemented.

Based on the usability test result as shown in Table 93.1, the user is satisfied with the application. From the observation, the participants obtained information regarding Solat as a therapy. They are able to use the application as their guidance to perform solat. Additionally, the therapeutic benefits and the integration of 3D model have enhanced their learnability.

Table 93.1 Usability result

	Description	Feedback
Task 1	Easy to click on the button	<i>Button</i> Most evaluator satisfied with the use of a button. The icon of a button makes it interesting to the participant
Task 2	Understand the usage of each button provided	
Task 3	Easily navigates from one screen to another screen	<i>Navigation</i> The evaluator finds the menu navigation easily
Task 4	Interface suitable to the application	<i>Interface</i> The interface includes interactivity with the user
Task 4	Text clear and readable	<i>Text</i> Most participants agree the font suitable to the application. It's easy to read and suitable with the background
Task 5	Suitable text color	
Task 6	Easy to understand	
Task 7	Sound could assist in the learning process	<i>Sound</i> Sound is redundant when user clicks the button (sound button in "solat teknik")
Task 8	Sound is clear	
Task 9	Suitable with application	<i>Color</i> Satisfied with the color
Task 10	3D model useful in learning process	<i>3D</i> The participants agree 3D model could easily to be understood

93.6 Conclusion and Future Works

Bio Terapi Solat can be used as a guideline for Muslim or others for learning and gaining knowledge about the benefits of solat. Many people did not realize solat can be utilized as a therapy and is beneficial for us. Therefore, this application will guide and show users how to perform the correct way of solat. The application also provides information on techniques and highlights the therapeutic benefits that can be achieved from the prayer technique that has been done.

In addition, the 3D model performs solat in 360° view. Each technique of solat, including the therapy is displayed in this application. This application differs with other applications because it includes the additional information such as therapy of solat, Sunnah Rasulullah S.A.W and glossary.

As a conclusion, the objective of this project has been achieved. The suitable multimedia element has been developed as interactive multimedia for guideline and learning of solat as a therapy. The model integrates with a 3D animation human model for making it interesting and more understandable. This application can be used by Muslims as a guide for them on the right technique to perform solat. The therapy is a good treatment to us if performed correctly during solat. This application which provides information about the prayer therapy could guide and educate users while using this application.

Furthermore, the evaluators recommend a 3D model over virtual reality modes. Users can control the 3D model where they can select the different part or angle of the model.

References

1. Doufesh, H., Faisal, T., et al.: EEG spectral analysis on muslim prayers. *Appl. Psychophysiol. Biofeedback* **37**(1), 11–18 (2012)
2. Rahman, U.Z.A.: *Formula Solat Sempurna*. Sri Damansara, Kuala Lumpur, Telaga Biru Sdn. Bhd (2008)
3. Salleh, N.A., Lim, K.S., et al.: AR modeling as EEG spectral analysis on prostration. In: 2009 International Conference for Technical Postgraduates (TECHPOS) (2009)
4. Aziz, N.A.A., Samsudin, S.: Computerized simulation development for blood circulation and bodily movement during obligatory prayers (SolatSim). In: Second World Congress on Software Engineering (WCSE) (2010)
5. Baran, I., Popovi, J., et al.: Automatic rigging and animation of 3D characters. *ACM Trans. Graph.* **26**(3), 72 (2007)
6. Yongguang, L., Mingquan, Z., et al.: Using depth image in 3D model retrieval system. *Adv. Mater. Res. (Comput. Mater. Sci.)* **268–270**, 981–987 (2011)
7. Wang, H., Yan, B.: A data-processing mechanism for scenario-based usability testing. In: 2011 IEEE 2nd International Conference on Computing, Control and Industrial Engineering (CCIE) (2011)

Chapter 94

Enhanced Interactive Mathematical Learning Courseware Using Mental Arithmetic for Preschool Children

Siti Zulaiha Ahmad, Noor Asmaliyana Ahmad,
Arifah Fasha Rosmani, Umi Hanim Mazlan
and Mohammad Hafiz Ismail

Abstract Interactive Mathematical Learning Courseware 2.0 (iMLc2.0) is an enhanced version of multimedia application which aims to expose the pre-school children in mental arithmetic technique with; (i) larger range of numbers, (ii) implementation of standard written method and, (iii) video support to visualize both techniques. The application integrates all multimedia elements and mental arithmetic techniques in interactive and supportive ways which are suitable for pre-school children. It is divided into three modules, addition, subtraction, and quizzes. The addition operation has been designed to apply the finger-brain approach which is a part of mental arithmetic technique while, mental-imagery approach is adapted for subtraction operation. Experimental test and user acceptance test have been conducted to evaluate the application. The implementation of iMLC2.0 in enriched interactive multimedia environment can be used as additional teaching and learning tool since it is supportive and attractive. In addition, it can motivate the preschooler to be prepared for future mathematical learning.

94.1 Introduction

The process of learning mathematics, such as numbers and basic operations like addition and subtraction, takes a long time to build for new learners, especially for preschoolers. Some of them found it hard to memorize numbers and mathematical concepts even though they already knew and understood them. Technology and computer are main aspirations and they are now integrated in education curriculums [1].

S.Z. Ahmad (✉) · N.A. Ahmad · A.F. Rosmani · U.H. Mazlan · M.H. Ismail
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Perlis, Perlis, Malaysia
e-mail: sitizulaiha@perlis.uitm.edu.my

With the evolution of computer technology, learning of mathematics for pre-school children can be improved. Generally, childrens' attention and interest can be developed if the learning process is assisted by multimedia tools. Furthermore, children nowadays are already exposed to computer at an early age.

Multimedia is an interactive computer-based environment with the combination of texts, voices, pictures and animations [2]. Thus the teachers can teach more consistently and effectively by using it as an alternative teaching tool. It is different from the traditional method, where only the exercise book is used as their single source in learning mathematics due to lack of tools. This results a boring environment and causing children to lose their focus.

Mental Arithmetic Technique can help children to build cognitive thinking when they need to use mental representations and fingers in order to do basic operation exercises. With the support of multimedia, their imagination of the mathematical concepts can be at enhanced. This improves their performance and allows them to be more active and excited due to the utilization of multimedia combination elements during learning sessions.

Most children in traditional method environment use only exercise books as their source of information. It is hard to develop their understanding and imagination using mental arithmetic if this method is used as a single approach in the classroom. Without an alternative learning tool, children will easily get bored and lose their interest. As a result, their focus on the mental representations of mathematical concepts decreases. This may affect their mathematical skills performance in the future, especially when they are in primary school. Therefore, the purpose of this study is to improve the design of previous iMLC by introducing mental arithmetic with standard written method using additional interactive elements for pre-school children.

94.2 Literature Review

94.2.1 *Mental Arithmetic Technique*

Mental arithmetic can be defined as the action of adding numbers together, multiplying them and other mathematical operations by using the brain, without the involvement of writing or the use of calculators [3]. This technique is usually used in mathematical operations such as simple addition, subtraction, multiplication, and division.

Mental arithmetic is also one of the methods that use the movement of fingers for counting numerical values [4, 5] as well as the working brain. A previous study proved that children move their real fingers to start the counting process after they build their imagination in the brain [6]. In addition, a study on introductory of mental arithmetic using simple application has shown a significant improvement in basic mathematic addition operation [7]. Therefore in this study, we would like to apply mental arithmetic with a large range of numbers using standard written method.

94.2.2 Children's Cognitive Thinking

Every child develops their learning preference along mental process simultaneously in day-to-day activities. Cognitive or mental processes are the processes of recognizing, understanding, and learning something [3]. The evolution of technology become as a catalyst towards development of children's cognitive thinking and skill to enable them progresses well in learning process.

Furthermore, cognitive skills in children can be sharpened through interaction and communication with computers [8]. Based on Piaget's Theory, children around five to eight years old are able to obtain the skills regarding objects, events, people, and use the symbols to imagine and represent real life and some examples are the symbols of words, numbers, and images. Therefore, children can establish their cognitive thinking and build mental representations when learning sessions are supported with multimedia elements such as sound, audio, video, images, graphics, and animations.

94.2.3 Multimedia

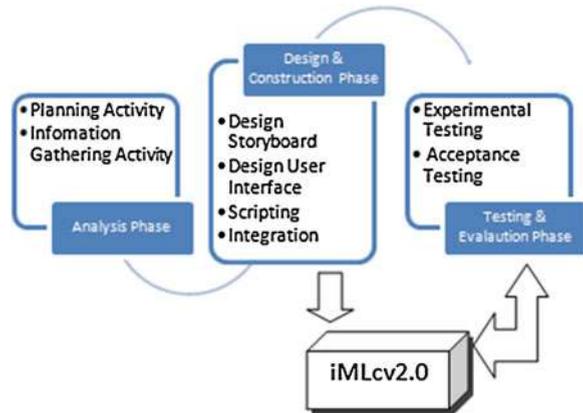
Multimedia is defined as the interactive computer-base, which is included with the texts, images, audio, video, and animations [2, 7, 9, 10] The use of multimedia is able to attract the children's interest and attention towards learning mathematics.

94.2.3.1 Elements in Multimedia

An interactive multimedia is composed of many elements that are important to deliver attractive prototype to children's learning. Animation is primarily used to illustrate the ideas and concepts. At an early age, children usually like to learn from the moving images because of their characters, which are attractive and motivating [11]. Animation is essential to improve understanding and brings out the interests of young learners [11, 12]. Next, video is an alternative element that could provide better visualization on specific activity. In any instructional courseware we need to match learning preferences with different types of multimedia instruction in order to provide easy-manage learning environment [13].

Images or graphics are considered to be part of multimedia where creativity and imagination are needed in a learning session. Basically, the use of graphics can help to sharpen memory [11] and stimulate mental model representations [11, 14].

The audio or sound also assists the children to give optimum attention and provides a way to learn mathematical concepts or terms from the oral speech. The implementation of sound can improve learning and pronunciation skills after listening to the spoken explanation [14]. Besides that, text is one of the important element in providing interaction and information as well as an effective way to communicate [10].

Fig. 94.1 Research model

94.3 Methodology

The study was conducted by applying research model as depicted in Fig. 94.1. The activities have been subdivided into three main phases:- Analysis, Design & Construction and Testing & Evaluation.

94.3.1 Analysis Phase

The first phase was to collect information on three areas of study:- preschoolers' preferences, multimedia elements principles and mental arithmetic. Interview and observation technique were applied in order to gain insight of the target users and the classroom environment. Teachers from two pre-schools in Kedah were selected in the interview session and preschoolers have been observed during mathematics classes. The observation was focused on the implementation of mathematic class, children behavior and their reaction towards learning process.

94.3.2 Design and Construction Phase

The second phase consists of two main activities: design and construction. Design activity involved storyboard design process by sketching interface design concept on a paper. It is based on the information gathered from the previous stage. The following process was designing user interface that incorporated multimedia elements such as graphics, sounds, texts, animations and videos using Adobe Flash, Adobe Photoshop and Adobe Illustrator software. Those elements were specifically chosen to stimulate the children's learning experience and to entertain them. Next, the design was translated into multimedia application in a construction activity.

Scripting and integrating the designed user interface were the main tasks of construction activity using Adobe Flash, Audacity and AVS Video Editor. iMLc2.0 was delivered as a multimedia application prototype.

94.3.3 Testing and Evaluation Phase

The final phase involved a series of testing and evaluation processes. The prototype was tested using usability testing and some refinements on the prototype were done based on the users' comments. Next, experimental testing was conducted to measure the children's performance which consist of Pre-Test and Post-Test evaluation. In Pre-Test session, the children were tested with mathematic question using normal learning process while during the Post-Test session, they had to answer a different set of question by applying mental arithmetic learning process. In order to assess courseware's acceptability among the targeted audiences, an acceptance test was conducted with kindergarten teachers. They were given a set of acceptance test questionnaires consist of 13 Likert Scale Based questions. The result was then analyzed to determine the acceptance rate of the application.

94.4 Courseware Architecture

The courseware consists of three main menus: (1) Number, (2) Operation, and (3) Activity modules. The content has been enhanced in all aspects including the design, the interaction approach and the range of number. Figure 94.2 shows the architecture of Interactive Mathematical Courseware application.

Fig. 94.2 The architecture of interactive mathematical courseware application

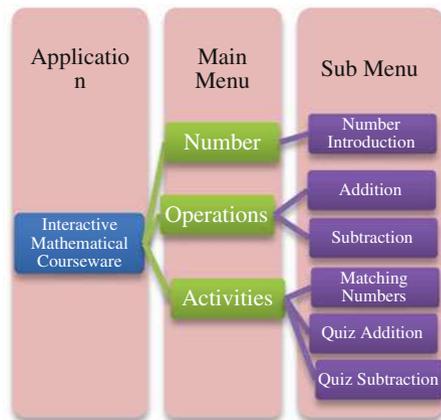


Fig. 94.3 Number introduction interface



94.4.1 Number

Number menu consists of Number Introduction sub menu as shown in Fig. 94.3. The children can learn starting from number one until number nine. It is also provided with the animation object and spelling that represents the numbers. The interface design includes graphic that familiar to children environment in order to attract their attention.

94.4.2 Operation

There are two sub menu involved in the operation menu of the application. They are addition and subtraction operation. Figures 94.4a, b show the addition interface which applied mental arithmetic technique, provided with the video for addition process. In order to introduce larger range of number, we include an example as a tutorial shown in Fig. 94.4c which is blended with standard written method. In Fig. 94.5, the mental-imagery of subtraction shows the subtraction operation interface. It implemented object animation with narration to visualize the subtraction process in three different range of number examples.

94.4.3 Activity

Activity menu consists of three different sub menus which are matching numbers, addition quizzes and subtraction quizzes. The concepts that were applied in the activity menu are drag and drop, perfect choice and writing concept. One of the

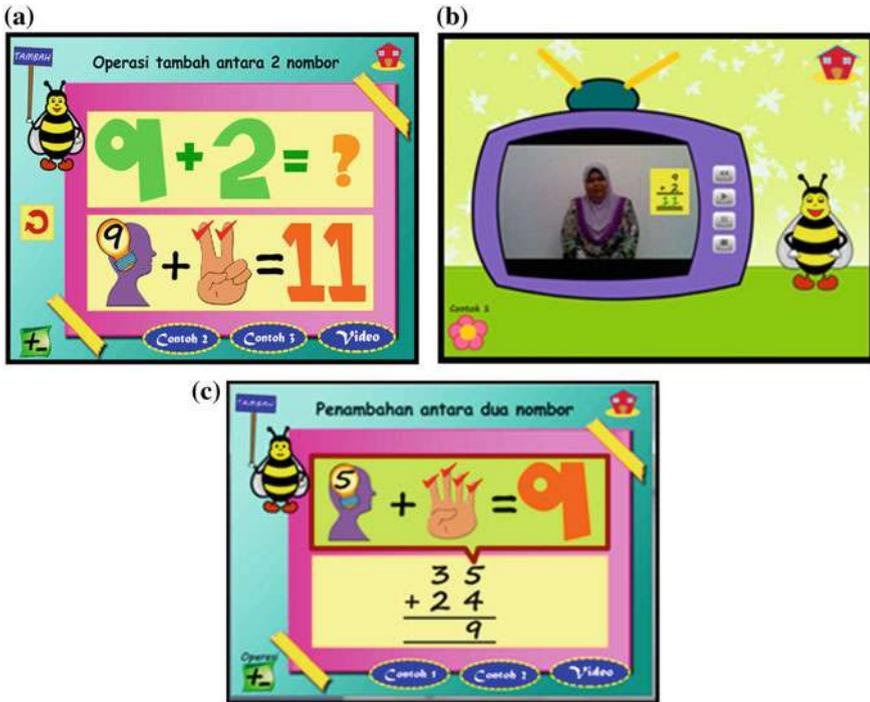


Fig. 94.4 a Addition operation interface with video assistance. b Addition operation interface. c Addition operation interface in larger range of number to adapt mental arithmetic technique using standard written method

Fig. 94.5 Subtraction operation interface

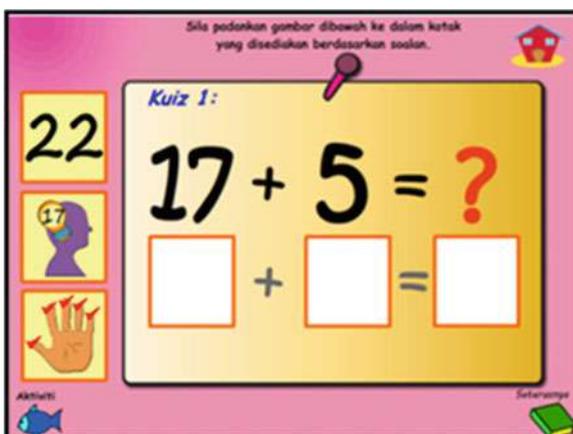


activities is shown in Fig. 94.6, number matching activity, which uses drag and drop concept. Figure 94.7 shows other activity to strengthen mental arithmetic skill in order to solve addition operation.

Fig. 94.6 Drag and drop interface: number matching



Fig. 94.7 Drag and drop interface: reinforcement of mental arithmetic skill



94.5 Finding and Discussion

Experimental test and User Acceptance test were conducted during testing phase of the application. Experimental test was done with preschoolers at two schools in Kedah to measure their understanding level towards mental arithmetic technique using iMLc2.0 by carrying out pre-test (conventional method) and post-test (mental arithmetic technique). The level of understanding for both schools were represented by the mean score of pre-test and post-test as shown in Fig. 94.8. It shows that children are capable to achieve higher scores after being exposed with the mental arithmetic technique. Besides that, as depicted in Fig. 94.9, the children took more than 15 min to solve basic mathematic questions during pre-test session by using conventional method such as stick bar counting or finger counting.

Fig. 94.8 Comparison of understanding level for pre-test and post-test

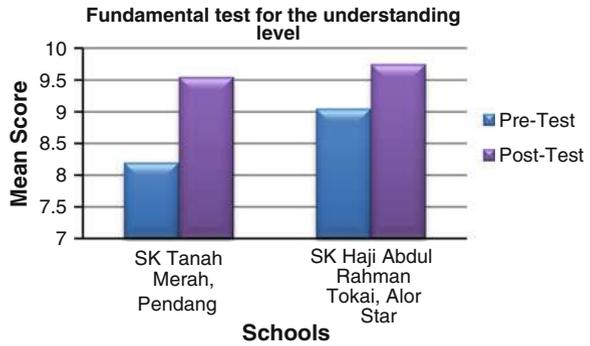
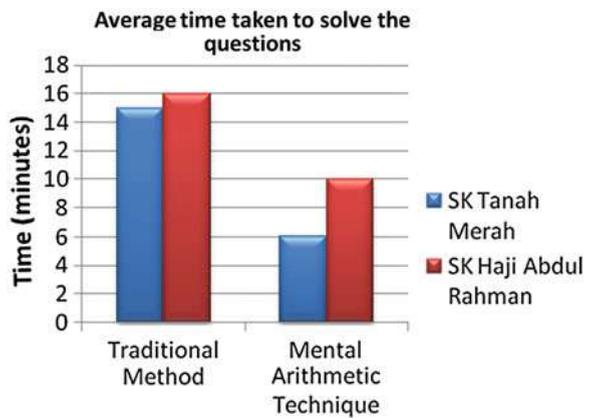


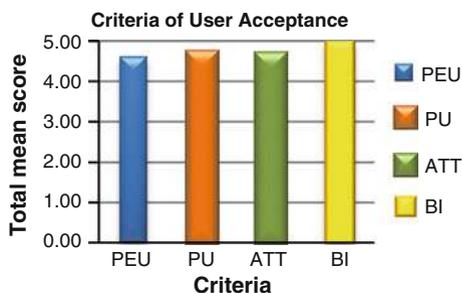
Fig. 94.9 Comparison of time taken using two different methods



However, in the post-test session, average time taken for the children to solve the mathematical problem using mental arithmetic obviously improved as they only took less than 10 min to complete all questions. Time improvement and higher score in post-test session proved that participants can think faster and accurate as they can simplify the calculation for the given questions.

User Acceptance Test was conducted with preschool teachers in order to determine the confidence level of end users of the application and to ascertain whether it can be accepted as a teaching tool in preschool. They were given time to explore the application before answering a set of questionnaires that covers four criteria. They are Perceived Ease Use (PEU), Perceived Usefulness (PU), Attitude (ATT), and Intention to Use (BI). Mean score for all criteria were calculated and plotted into a bar chart as shown in Fig. 94.10. The highest total mean score is BI, which is 5.00 where all participants strongly agreed to adopt the application in future as a teaching aid. The total mean score for PEU is 4.63 where most of the participants agreed that the application was easy to use (user friendly) and learnable. The total mean for PU is 4.71 which reveals that the application is useful

Fig. 94.10 Bar chart of criteria for user acceptance test



to attract children's interest and attention. The total mean score for ATT is 4.67, which means that users are more likely to use the application because it uses multimedia elements.

94.6 Conclusion

The main contribution in this research is more towards enhancing the interaction design and problem solving technique of the previous application [7] which are the utilization of multimedia elements (video and text animation) and content enrichment (larger range of number and standard written method). Based on this study, it can be summarized that the iMLC2.0 application was successfully accepted to be used in teaching and learning mathematics in preschool.

The inclusion of video element enable the children to increase understanding in mental arithmetic technique as they can visualize it better compared to the utilization of animations, text and sound alone. The result proved that learning mathematics through multimedia application can increase the fun and interest during the learning session. Yet, it could improve their mathematical skill in addition and subtraction operations.

The application achieves its target to be a helpful tool in improving users' performance and sharpening their cognitive thinking. In addition, this technique could be implemented in different school levels: pre-school, primary school (lower and higher level) and secondary school. However, the approach might be different for each level.

In conclusion, this application achieves the objectives of this research, where an Interactive Mathematical Learning Courseware was designed and developed by adapting suitable multimedia elements such as text, graphics, animation, sound, and video. The implementation of mental arithmetic technique and standard written method were successfully introduced in the application. Our next research paper will be focusing specifically on instructional design diversity and its effectiveness in implementing mobile iMLC.

References

1. Ktoridou, D., Etekleous, N., Gregoriou, G.: Preschoolers Developing Mathematical Understanding through Computer-Based Activities. In: The International Conference on Computer as a Tool, pp. 787–790 (2005)
2. Weiss, I., Kramarski, B., Talis, S.: Effects of multimedia environments on kindergarten children's mathematical achievements and style of learning. *Educ. Media Int.* **43**(1), 3–17 (2006)
3. Longman: Longman dictionary of contemporary english online (Pearson Education Limited). <http://www.ldoceonline.com/>. Accessed October 2011
4. Wu, S.S., Meyer, M.L., Maeda, U., Salimpoor, V., Tomiyama, S., Geary, D.C.: Standardized assessment of strategy use and working memory in early mental arithmetic performance. *Dev. Neuropsychol.* **33**(3), 365–393 (2008)
5. Klein, E., Moeller, K., Willmes, K., Nuerk, H., Domahs, F.: The influence of implicit hand-based representations on mental arithmetic. *Frontiers Psychol. Cogn.* **2**, 1–7 (2011)
6. Cheah, B.L., Ong, S.L.: Perbandingan Kebolehan Menyelesaikan Masalah Matematik Antara Murid Yang Belajar Abakus-Aritmetik Mental Dengan Murid Yang Tidak Belajar Abakus-Aritmetik Mental. *Jurnal Pendidik dan Pendidikan. Universiti Sains Malaysia*, vol. 21, pp. 85–100 (2006)
7. Ahmad, S.Z., Rosmani, A.F., Ismail, M.H., Suraya, M.S.: An introductory of mental arithmetic using interactive multimedia. *Comput. Inform. Sci.* **3**(4), 72–79 (2010)
8. Gelderblom, H., Kotze, P.: Designing technology for young children: what we can learn from theories of cognitive development. In: Proceeding of SAICSIT '08 (South African Institute of Computer Scientists and Information Technologists), pp. 66–75 (2008)
9. Shujuan, Z., Wei, S., Zheng, L., Qibo, H.: the principles of multimedia teaching design based on cognitive load theory. In: 2010 2nd International Conference on Education Technology and Computer (ICETC), pp. 110–112 (2010)
10. Segers, E., Verhoeven, L., Hulstijn, N.-H.: Cognitive processes in children's multimedia text learning. *Appl. Cogn. Psychol. J.* **22**, 375–387 (2008)
11. Betrancourt, M., Chassot, A.: Making sense of animation: how do children explore multimedia instruction? cognitive processes in children's multimedia text learning. *Appl. Cogn. Psychol.* **22**, 375–387 (2008)
12. Fang, H., Li, X., Helge, H.: Research on Improving Learning Interests for Elementary Students Based-on Mental Mode by Using Mathematics Animations. In: International Conference on Application of Information and Communication Technologies, pp. 1–5 (2009)
13. Zaidel, M., Luo, X.: Effectiveness of multimedia elements in computer supported instruction: analysis of personalization effects, student's performance and cost. *J. Coll. Teach. Learn.* **7**(2), 11–16 (2010)
14. Ali, B.B., Badioze H.Z.: Framework of adaptive multimedia mathematics courseware. Proceeding of the 2nd IMT-GT Regional Conference on Mathematics, Statistics and Applications (2006)

Chapter 95

Comparative Evaluation of Ensemble Learning and Supervised Learning in Android Malwares Using Network-Based Analysis

Ali Feizollah, Nor Badrul Anuar, Rosli Salleh and Fairuz Amalina

Abstract With the prevalence of mobile devices, the security threats are growing in number and seriousness. Among the mobile operating systems, Google's Android has been attacked more than others have. From April 2013 until June 2013, the number of malwares were doubled for the Android. In this paper, we evaluate the mobile malwares detection using the ensemble learning and supervised learning. Furthermore, we compare the two learning approaches based on the experimental results. We compared our experimental results with a similar work. The network traffic generated by mobile malwares are analyzed. We use 600 malware samples from the MalGenome data sample to build the dataset. We use two versions of random forest algorithm as our evaluating algorithm, ensemble learning and supervised learning. The empirical results show that the ensemble learning improves the detection of the Android malwares. The ensemble learning achieved 99.6 % of true positive rate while the supervised learning attained 99.4 %.

Keywords Mobile malware · Ensemble learning · Supervised learning · Android · Network-based analysis · Malgenome

A. Feizollah (✉) · N.B. Anuar · R. Salleh · F. Amalina
Security Research Group (SECREg), University of Malaya, Kuala Lumpur, Malaysia
e-mail: ali.feizollah@siswa.um.edu.my

N.B. Anuar
e-mail: badrul@um.edu.my

R. Salleh
e-mail: rosli_salleh@um.edu.my

F. Amalina
e-mail: fairuzamalina@siswa.um.edu.my

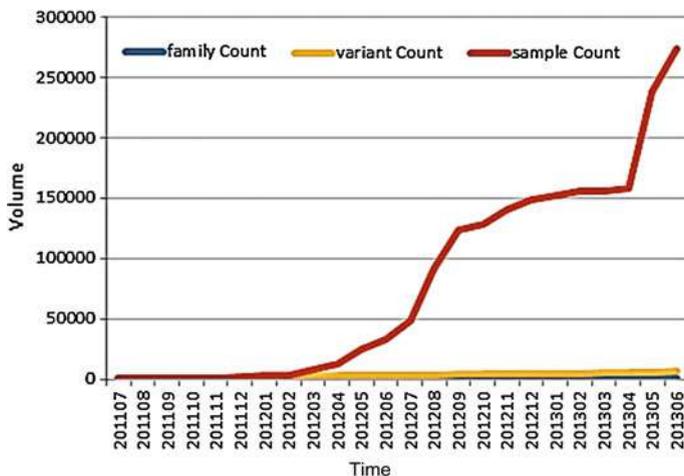


Fig. 95.1 Android Malware Growth [1]

95.1 Introduction

The heavy use of mobile devices is undeniable in today's world. They have facilitated life in many aspects. People are able to email, chat, facebook and do mobile banking on the go from their mobile devices. However, there is a drawback inherent in such conveniences. Since the mobile devices contain sensitive and personal information, they are prone to attacks by hackers. Based on Symantec report that was published in October 2013, in just 2 months, from April 2013 until June 2013, the volume of Android malwares has doubled [1]. The Fig. 95.1 shows the growth in the Android malwares between July 2011 and June 2013.

Among the mobile operating systems, the Google's Android operating system has been attacked more than other mobile operating systems since it is open source [2]. F-secure corporation reported that Android operating system comprises of 79 % of malwares in 2012 [3]. It is estimated that the sales of Android based mobile devices will grow by 12 % in 2014 compare to other operating systems [4]. Such growth in malwares and the sale of Android based mobile devices raise concerns over the security of mobile devices.

In addition, the official Android market, known as Google Play, contains malwares. The report from Symantec [1] shows the top ten application categories with highest percentage of malwares. The Google Play is included in this report as well. The Fig. 95.2 represents the application categories with percentage of malwares.

We used MalGenome data sample as our dataset for this study. It is one of the well-accepted data samples of Android malwares among the research community. It is a collection of Android malwares collected by the University of North Carolina in the period of August 2010 and October 2011 [5]. They published the paper

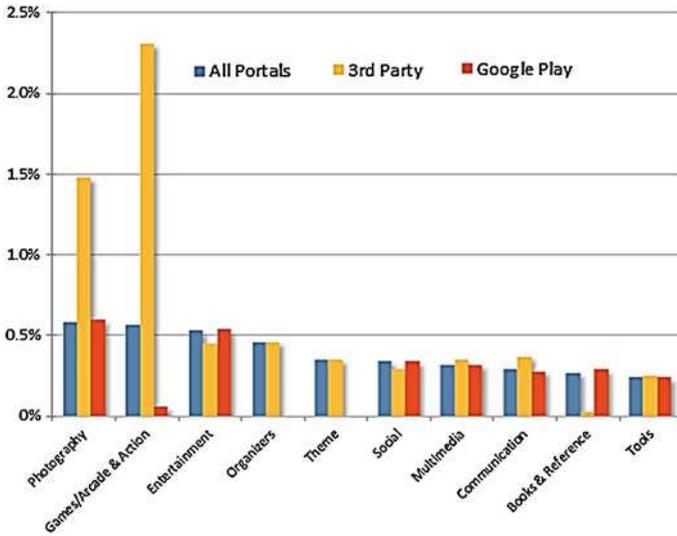


Fig. 95.2 Application Category with the percentage of malwares [1]

along with the data sample in 2012. It is a collection of 1260 malwares in 49 different malware families. In this paper, 600 malware samples were selected from the MalGenome data sample.

The supervised learning is defined as methods applied on a labelled data set. It entails two stages: training and testing. In the training stage, the algorithm tries to find a pattern in data with the same label in order to group them. In the test stage, the learned algorithm is used with the new set of data to examine how successful the algorithm is.

The ensemble learning comprises of the individually trained algorithms. In the testing stage, the prediction is the combination of trained algorithms. In other words, the selected algorithms are trained separately but the decision on new data is done together. There are two methods in ensemble learning: bagging and boosting. In the bagging method, the individual algorithms are trained with the random distribution of the training sets. The distribution of the training sets is independent of the previous algorithm’s performance. The Fig. 95.3 shows a sample of assigning training sets in the bagging algorithm [6].

The boosting method is similar to the bagging method, but the difference is that the assignment of the training sets for each algorithm is based on the performance of the previous algorithm. For instance, the Fig. 95.4 depicts a sample of boosting method in which the set C is repeated in the second algorithm onward. The reason is that the boosting method keeps track of the incorrectly classified data. When the first algorithm incorrectly classifies the data in C, the boosting method repeats the C set in the training set for second algorithm. The boosting method is proved to

Fig. 95.3 Training sets for bagging method

Original Training Set	A,B,C,D,E
Training Set for 1st Algorithm	B,C,A,E,D
Training Set for 2nd Algorithm	D,C,E,B,A
Training Set for 3rd Algorithm	E,D,A,D,B

Fig. 95.4 Training sets for boosting method

Original Training Set	A,B,C,D,E
Training Set for 1st Algorithm	B,C,A,E,D
Training Set for 2nd Algorithm	D,C,E,C,A
Training Set for 3rd Algorithm	C,D,C,D,C

produce better results than the bagging method [6]. In this paper, we have selected the Adaboost algorithm, which uses the boosting method.

In analysis of the Android malwares, there are two methods: behavior-based and network-based. The behavior-based analysis refers to the internal behavior of the malwares in the mobile operating system such as system calls. Malwares represent some specific behavior by which analyzing them makes it easier to detect the malwares. The network-based analysis, on the other hand, focuses on the malwares' behavior in the network traffic. Every mobile application, including malicious one, has to communicate through the network. The communication of the malicious applications have some features that differentiate them from normal network traffic [7].

Thus, there is a vital need in analyzing the malwares and developing effective methods to confront the mobile malwares. This paper aims at evaluating the ensemble learning and supervised learning in mobile malware detection using the network-based analysis. It is imperative to study various types of algorithms in order to determine the best one to confront the massive growth of the mobile malwares.

This paper is organized as follows. Section 95.2 discusses some of the related works done by other researchers. Section 95.3 presents backgrounds of the algorithms used in this paper. The methodology and the flow of the proposed method is described in the Sect. 95.4. The results of this work is presented in the Sect. 95.5 along with discussions. Section 95.6 discusses the future works that can be done.

95.2 Related Works

95.2.1 Supervised Learning Versus Ensemble Learning

In a study [8], researchers applied ensemble learning to the generic algorithm and feature clustering. They experimented with the ensemble learning. The results showed an improvement in the detection.

Yang Li et al. [9] published a study in which they employed the ensemble learning on the KDD CUP 99 dataset. The KDD dataset is the collection of the DDoS attacks. They reported a significant increase in the detection rate via the ensemble learning. Thus, the ensemble learning is proven to increase the effectiveness of other algorithms. In this paper, we prove that the same effect of the ensemble learning is applied to the Android malwares.

A study was published in 2008 [10] in which authors used the ensemble learning and supervised learning in order to predict breast cancer survivability. The results show that the ensemble learning was more successful than the supervised learning.

95.2.2 Behavior-Based Versus Network-Based Analysis

As discussed in the introduction, there are two methods in analyzing the Android malwares: behavior-based and network-based. The behavior-based analysis involves analyzing the system calls inside the mobile operating system such as read, write and open.

The Crowdroid [7] analyzed the system calls by collecting them as a log file and processing them to discover anomalies. They tested the Crowdroid using self-written malwares as well as real malwares. The results were as high as 100 %, but the downside is that they used self-written malwares that are not as realistic as the real world malwares.

However, the Android operating system is based on Linux kernel [11] and collecting the system calls is a toil [7]. Whereas, in the network-based analysis, collecting the network traffic is as simple as installing the tPacketCapture application [12].

Sue et al. [13] extracted nine features from the network traffic and used two algorithms, decision tree and random forest, for the experiments. The features are the average and standard deviation of the number of sent/received packets, the average and standard deviation of the number of bytes sent/received and the average TCP/IP session duration. They achieved 96.70 % of detection rate with random forest.

Similarly, a study was done by Feizollah et al. [14] in which they analyzed the network traffic of 100 malwares for malware detection. They chose three network features namely, packet size, connection duration and number of GET/POST parameters. The results were as high as 99.94 % of true positive rate. Thus, the network-based approach was chosen for this paper.

95.3 Background of Algorithms

We have used adaboost algorithm along with the random forest for the ensemble learning approach.

95.3.1 Adaboost

It is one the most successful and popular ensemble algorithms. It is a self-rated algorithm meaning that it scores the reliability of its own prediction. It has a high flexibility for combining with other methods such as supervised learning. It also requires less knowledge about the data set in order to improve the accuracy of the algorithm [15]. In this work, we employed the Adaboost.M1 [16] in our experiment. As discussed in the introduction section, the boosting algorithm works by assigning the weight on the training sets. Suppose training set $(x_1, y_1), \dots, (x_n, y_n)$ where each x_i belongs to instance space x and each label y_i is in the label set y . The Adaboost.M1 assigns the weight on the training sets. Initially, the weights are equal. Based on the performance of each algorithm, the weight of misclassified training sets are increased in order to focus on the misclassified sets in the next rounds of training [10].

95.3.2 Random Forest

Random forest algorithm is one of the most popular among the researchers. It has been proven that it is a popular and powerful algorithm in the pattern matching and machine learning [17]. The random forest algorithm constructs a collection of decision trees, which use the classification and regression methods. They are sets of rule-based methods to generate the decision trees. The trees are developed independent of each other and they vote for the best class to form the random forest. The error in the random forest algorithm depends on the robustness of individual trees and the correlation between them [18].

95.3.3 Adaboost and Random Forest

The combination of Adaboost and random forest was utilized in this paper. Adaboost is the boosting algorithm that is used to improve the performance of the random forest algorithm. For instance, researchers used the Adaboost with the random forest and the final results show improved performance [10].

95.4 Methodology

The methodology of this study comprises of three phases: building the data set, feature selection and extraction, and the supervised learning and ensemble learning. Figure 95.5 shows the architecture of the methodology. The details of each phase are explained in the following sections.

95.4.1 Building the Data Set

The data set consists of the selected features in the network traffic. The first step is to collect the network traffic generated by the malwares as well as normal applications in order to construct the data set. The capturing process of the network traffic is done by installing the tPacketCapture Pro [12] on a real device and running the applications for 30 min. The generated network traffic is captured via the application and the result is prepared in the pcap file format. We then transferred the file to the computer to extract selected features. We used 600 malware samples from the MalGenome data sample in this paper ranging from the simplest malware to the most sophisticated ones like AnserverBot.

95.4.2 Feature Selection and Extraction

Among the massive number of network features, we have selected four of the most important ones. They are packet size, connection duration, frame length and the source port.

The packet size is the size of each packet; the packet that leaks user’s data from a device to the hacker has larger size than normal packets. Most of the time,

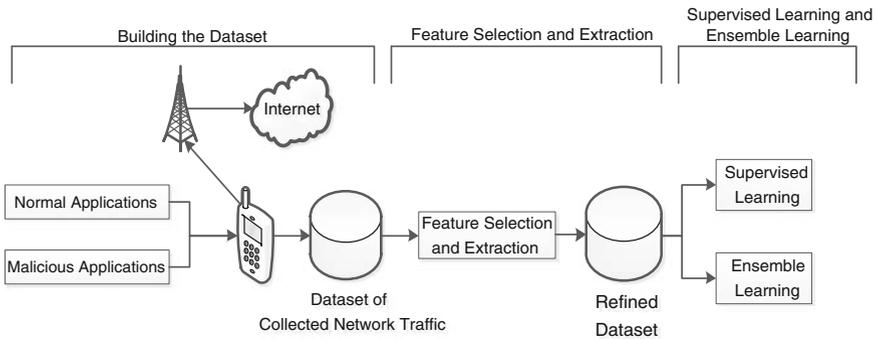


Fig. 95.5 Methodology architecture

Table 95.1 Results of the experiment

Algorithm	TPR (%)	FPR (%)
Adaboost and random forest	99.6	0.4
Random forest	99.4	0.6

Android malwares connect to the hacker to check for new commands. Thus, most connections are in the form of simple handshakes, which is a plausible reason to choose the connection duration for this study. The frame length is the length of each frame in bytes. The source port is important in the detection of the malwares.

The selected features were extracted using the tshark [19] program, which is a command line version of the Wireshark program.

95.4.3 Supervised Learning and Ensemble Learning

The prepared data set is fed to the supervised algorithm as well as ensemble algorithm. We chose the random forest algorithm for the supervised phase and the adaboost and random forest for the ensemble phase. The purpose is to compare the results of two learning methods and to prove that the ensemble learning provides better results and performance than the supervised learning in the Android malwares detection.

95.5 Results and Discussions

The results of the experiment are expressed in terms of true positive rate (TPR) and false positive rate (FPR). The random forest algorithm achieved the TPR of 99.4 % while the Adaboost and random forest algorithm resulted in 99.6 % of the TPR. Table 95.1 shows the results of the study.

Receiver operating characteristic (ROC) curve is one of the common methods of analyzing the results of an experiment. It indicates how the true positive rate changes as the internal thresholds of the algorithm are varied to generate more or fewer false alarm. It plots true positive rate against false positive probability. ROC curves show the tradeoff between false positive rate and true positive rate that means any increase in the true positive rate is accompanied with a decrease in the false positive rate. The top left corner of the diagram is considered the optimum solution. The Figs. 95.5 and 95.6 illustrate the ROC curve for our experiment (Fig. 95.7).

As presented, the ROC curves are very similar. In order to analyze ROC curves, the area under the curve (AUC) is examined. In this case, the AUC for Adaboost and random forest is 0.999 and the AUC for the random forest is 0.998. As the

Fig. 95.6 Adaboost and random forest ROC

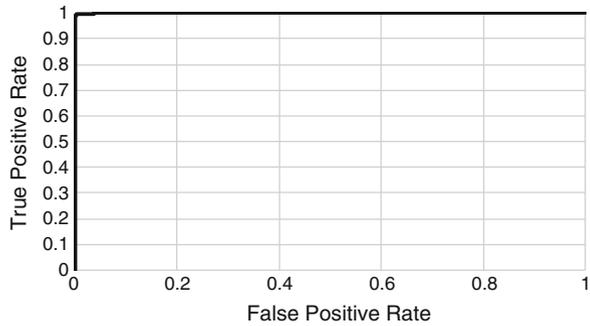
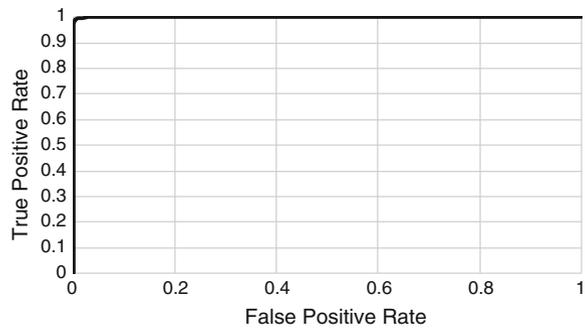


Fig. 95.7 Random forest ROC



AUC approaches to 1.00, the algorithm has better performance. Thus, the Adaboost and random forest has the better performance than the random forest.

In order to authenticate this work, a comparison is done with a similar work. In 2013, a study was conducted [20] on Android malwares for classification purpose. The authors used ensemble learning approach with the random forest as the chosen algorithm. They achieved 99 % of detection rate in their experiment. However, in this work, we attained 99.6 % of detection rate. In addition, we conducted another experiment with supervised learning to prove that the ensemble learning is more effective than the supervised learning.

Overall, it is evident that the ensemble learning, adaboost and random forest, performs better than the supervised learning, random forest, in the Android malware detection and the ensemble learning has improved the Android malware detection over the supervised learning.

95.6 Conclusion and Future Works

In this paper, we analyzed the network traffic generated by the Android malwares using two versions of random forest algorithm, which are supervised learning approach and ensemble learning approach. The experimental results proved that

the ensemble learning improved the Android malware detection. In addition, we compared our work with a similar work done by other researchers and we concluded that our work achieved higher detection rate.

As a future work, the real time experiment on the mobile malwares can be conducted. It is worth mentioning that experimenting with more data samples is suggested for future works.

Acknowledgments This work was supported in part by the Ministry of Higher Education, Malaysia, under Grant FRGS FP034-2012A and the Ministry of Science, Technology and Innovation, under Grant eScienceFund 01-01-03-SF0914.

References

1. Symantec: <http://www.symantec.com/connect/blogs/android-madware-and-malware-trends>
2. Teufl, P., Ferk, M., Fitzek, A., Hein, D., Kraxberger, S., Orthacker, C.: Malware detection by applying knowledge discovery processes to application metadata on the Android Market (Google Play). *Security and Communication Networks*, (2013)
3. Techcrunch: <http://techcrunch.com/2013/03/07/f-secure-android-accounted-for-79-of-all-mobile-malware-in-2012-96-in-q4-alone/>
4. Cnet-News: http://news.cnet.com/8301-1035_3-57614451-94/heads-up-apple-android-to-make-big-gains-in-2014-idc-predicts/?part=rss&subj=news&tag=title
5. Yajin, Z., Xuxian, J.: Dissecting android malware: characterization and evolution. In: 2012 IEEE Symposium on Security and Privacy (SP), pp. 95–109. (2012)
6. Maclin, R., Opitz, D.: Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999)
7. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: Crowdroid: behavior-based malware detection system for Android. In: 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 15–26. ACM, 2046619 (2011)
8. Tao, H., Ma, X.-p., Qiao, M.-y.: Support vector machine selective ensemble learning on feature clustering and genetic algorithm. In: Wang, X., Wang, F., Zhong, S. (eds.) *Electrical, Information Engineering and Mechatronics 2011*, vol. 138, pp. 1617–1625. Springer, London (2012)
9. Li, Yang, Jian Lin, Li, Song Jie, Yue, Wang, Z.: Research of intrusion detection based on ensemble learning model. *Appl. Mech. Mater.* **336**, 2376–2380 (2013)
10. Thongkam, J., Guandong, X., Yanchun, Z.: Ada boost algorithm with random forests for predicting breast cancer survivability. In: *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 3062–3069. (2008)
11. Shabtai, A., Fledel, Y., Kanonov, U., Elovici, Y., Dolev, S., Glezer, C.: Google android: a comprehensive security assessment. *Secur. Priv. IEEE* **8**, 35–44 (2010)
12. tPacketCapturePro: <https://play.google.com/store/apps/details?id=jp.co.taosoftware.android.packetcapturepro>
13. Su, X., Chuah, M., Tan, G.: Smartphone dual defense protection framework: detecting malicious applications in android markets. In: 2012 Eighth International Conference on Mobile Ad hoc and Sensor Networks (MSN), pp. 153–160. (2012)
14. Feizollah, A., Anuar, N.B., Salleh, R., Amalina, F., Ma'arof, RuR, Shamshirband, S.: A study of machine learning classifiers for anomaly-based mobile botnet detection. *Malays. J. Comput. Sci.* **26**, 251–265 (2013)
15. Schapire, R.E.: A brief introduction to boosting. In: *the International Joint Conference on Artificial Intelligence*, pp. 1401–1405. (1999)

16. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, pp. 148–156. (1996)
17. Meinshausen, N.: Quantile regression forests. *Mach. Lear. Res.* **7**, 983–999 (2006)
18. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001)
19. tshark: <http://www.wireshark.org/docs/man-pages/tshark.html>
20. Alam, M.S., Vuong, S.T.: Random forest classification for detecting android malware. In: 2013 IEEE and Internet of Things Green Computing and Communications (GreenCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing (iThings/CPSCoM), pp. 663–669. (2013)

Chapter 96

Tailored MFCCs for Sound Environment Classification in Hearing Aids

Roberto Gil-Pita, Beatriz López-Garrido and Manuel Rosa-Zurera

Abstract Hearing aids have to work at low clock rates in order to minimize the power consumption and maximize battery life. The implementation of signal processing techniques on hearing aids is strongly constrained by the small number of instructions per second to implement the algorithms in the digital signal processor the hearing aid is based on. In this respect, the objective of this paper is the proposal of a set of approximations in order to optimize the implementation of standard Mel Frequency Cepstral Coefficient based sound environment classifiers in real hearing aids. After a theoretical analysis of these coefficients and a set of experiments under different classification schemes, we demonstrate that the suppression of the Discrete Cosine Transform from the feature extraction process is suitable, since its use does not suppose an improvement in terms of error rate, and it supposes a high computational load. Furthermore, the use of the most significative bit instead of the logarithm also supposes a considerable reduction in the computational load while obtaining comparable results in terms of error rate.

96.1 Introduction

A hearing aid capable of automatically classifying the acoustic environment that surrounds his/her user, and selecting the amplification “program” that is best adapted to such environment (“self-adaptation”) would improve the user’s

R. Gil-Pita (✉) · M. Rosa-Zurera
Signal Theory and Communications Department, University of Alcalá, Madrid, Spain
e-mail: roberto.gil@uah.es

M. Rosa-Zurera
e-mail: manuel.rosa@uah.es

B. López-Garrido
Servicio de Salud de Castilla La Mancha (SESCAM), Castilla La Mancha, Spain
e-mail: lopesbea@hotmail.com

comfort [1]. The “manual” approach, in which the user has to identify the acoustic surroundings, and to choose the adequate program, is very uncomfortable and frequently exceeds the abilities of many hearing aid users [2]. Furthermore, sound classification is also used in modern hearing aids as a support for the noise reduction and source separation stages, like, for example, in voice activity detection [3], in which the objective is to extract information from the sound in order to improve the performance of these systems.

There is a number of interesting features that could potentially exhibit different behavior for speech, music and noise and thus may help the system classify the sound signal. One of the most typical features used for information extraction in audio analysis are the Mel Frequency Cepstral Coefficients (MFCCs), that have already been used for sound environment classification in hearing aids [4]. The problem of implementing an MFCC based sound classifier in a hearing aid is that DSP-based hearing aids have constraints in terms of computational capability and memory. The hearing aid has to work at very low clock rates in order to minimize the power consumption and thus maximize the battery life. Additionally, the restrictions become stronger because a considerable part of the DSP computational capabilities are already being used for running the algorithms aiming to compensate the hearing losses.

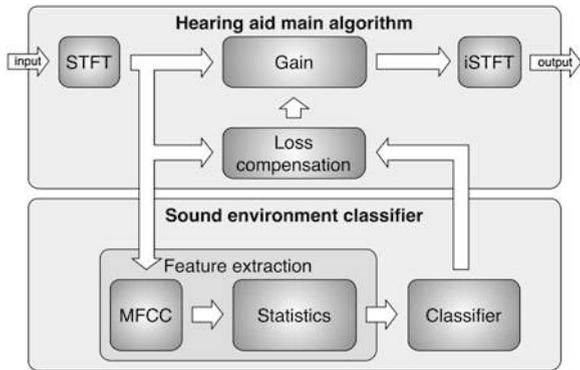
Therefore, the design of any automatic sound classifier is strongly constrained to the use of the remaining resources of the DSP: roughly speaking, the computational power does not use to exceed 5 MIPS. This restriction in number of operations per second enforces us to put special emphasis on signal processing techniques and algorithms tailored for properly classifying while using a reduced number of operations.

This paper presents a tailored efficient implementation of MFCC based sound environment classifier for digital hearing aids. The idea is to propose solutions for the implementation of these systems, bearing with the finite-precision limitations and the computational cost, with the aim of establishing a set of tradeoffs between performance in terms of error rate and number of required assembler instructions per second.

96.2 Description of the Sound Environment Classifier

Figure 96.1 shows the block diagram of the sound environment classifier, in relationship with the main loss compensation algorithm of a hearing aid. As we can appreciate, the sound environment classifier basically consists of a feature extraction block, and the aforementioned classifier. Each of this block will be studied in detail in this section.

Fig. 96.1 Block diagram of the proposed system showing the different stages of the hearing aid



96.2.1 Feature Extraction

There is a number of interesting features that could potentially exhibit different behavior for speech, music, and noise and thus may help the system classify the sound signal. Classical approaches use tailored specific features such as the spectral centroid, spectral flux, voice to white or short time energy, which are computationally efficient and they obtain considerably good results in the problem at hand [5]. In a different approach, MFCCs render very good results in terms of error rate, but with a considerably increase of the required computational resources [6]. In this paper we will focus on the last ones, with the aim of analyzing their drawbacks in terms of computation complexity, allowing a posterior definition of a computationally simplified version of the MFCCs.

Obtaining MFCC coefficients [7] has been regarded as one of the techniques of parameterization most important used in speech processing. They provide a compact representation of the spectral envelope, so that most of the energy is concentrated in the first coefficients. Perceptual analysis emulates human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands. Mel cepstral analysis uses the Mel scale and a cepstral smoothing in order to get the final smoothed spectrum.

The main stages of MFCC analysis are:

- Short Time Fourier Transform (STFT): MFCCs are evaluated using information derived from the STFT. In order to overcome the non-stationary of speech, it is necessary to analyze the signal in short time periods, in which it can be considered almost stationary. So, time frames or segments are obtained dividing the signal in frames, which usually have an overlap factor of 50 %. Then, DFT is calculated to each windowed time frame, generating $X_i[k]$, a time-frequency decomposition.
- Mel scale non uniform filter bank: In order to determine the MFCCs, phase information is discarded, since we work with the energy of STFT, $|X_i[k]|^2$. This signal $|X_i[k]|^2$ is then multiplied by a triangular filter bank, using Eq. (96.1).

$$E_{mt} = \sum_{k=0}^{N/2} |X_t[k]|^2 H_m[k], 1 \leq m \leq F, \quad (96.1)$$

where $H_m[k]$ are the triangular filter responses, whose area is unity. These triangles are spaced according to the Mel frequency scale, matching F the number of final MFCCs [8].

- Discrete Cosine Transform (DCT): Through the DCT, expressed in Eq. (96.2), the logarithm of the spectral coefficients are re-transformed, so the mel frequency spectral coefficients are converted to a cepstral domain.

$$MFCC_{mt} = \sum_{k=1}^F \log(E_{mt}) \cos(m(k - 1/2)\pi/N), m = 1, \dots, F \quad (96.2)$$

Once MFCCs are evaluated, features are determined from statistics of each MFCC. Some of the most common used statistics are the mean and the standard deviation, which have been successfully used in sound environment problems for hearing aids [5]. So, in this paper we study the implementation of mean and standard deviation of the MFCCs as features, since we have found that these values obtain very good results with a considerably low computational complexity.

96.2.2 Classification Algorithms

In order to analyze the performance of the proposals under different classification scenarios, four different classifiers are studied: the Least Squares Linear Classifier (LSLC), the Least Squares Diagonal Quadratic Classifier (LSDQC), the Multilayer Perceptron (MLP) and the Quadratic Multilayer Perceptron (QMLP).

Least Square Classifiers are a classifiers that render very good results with a very fast learning process and low computational complexity after training and therefore they have been selected for the experiments carried out in this paper. Let us consider a set of training patterns $\mathbf{x} = [x_1, x_2, \dots, x_L]^T$ (for instance, and as we described above, the mean and/or standard deviation of the used features), where each of these patterns is assigned to one of the C possible classes, $c = 1, \dots, C$. In a linear or quadratic classifier, the decision rule can be obtaining looking for the class that maximizes a combination from a set of C combinations, as shows Eq. (96.3).

$$y_c = v_{c0} + \sum_{n=1}^L v_{cn} x_n + \sum_{n=1}^L \sum_{p=1}^L w_{cnp} x_n x_p, \quad (96.3)$$

where v_{cn} are the linear weighting values, and w_{cnp} are the quadratic weighting values. So, in the LSLC the terms w_{cnp} is zero, giving a linear combination of the input features.

Another particular case is the LSDQC which is referred to the use of only the diagonal coefficients of the matrix of quadratic terms ($w_{cnp} = 0, \forall n \neq p$). In the least squares approach, the weights are adjusted in order to minimize the mean square error, which leads to the solution of the equations of Wiener-Hopf [9]. So, the use of the LSDQC allows obtaining more complex solutions with the drawback of duplicating the effective features of the classifier.

Concerning MLPs, they are feed forward artificial neural network models that have successfully been implemented in hearing aids as sound environment classifiers [5]. They consist of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Each node is denominated neuron, and it processes data applying a nonlinear function called activation function to a linear combination of the inputs of the node [10]. MLPs are typically trained to minimize the mean squared error of the outputs using back propagation algorithms. In this paper two-layer MLPs have been trained using the Levenberg-Marquardt optimization algorithm [11]. In the MLP, each neuron of the first layer divides the input space in two by means of a hyperplane, and the second layer combines these hyperplanes to generate more complex boundaries. So, the complexity of the solutions implemented in an MLP can be controlled by the number of available neurons.

At last, QMLPs, also known as second order neural networks [12], are a variant of the MLPs inspired in the differences between the LSLC and the LSDQC. It consists in an MLP whose inputs are doubled adding quadratic terms of the input pattern. So, in this case each neuron in the first layer can divide the input space in two by a quadratic function, highly increasing the capability of the classification system.

It is important to highlight all the four classifiers described in this paper are not influenced by shifts or scaling of the input patterns, since these kind of changes will only suppose a corresponding change in the input weights and they will not alter the performance of the classifier. This property will be explored in the next section in order to reduce the computational cost of the implemented solution.

96.3 Description of the Proposals

In this section we will focus on two specific points that highly increase the computational complexity of the sound environment classifier, and we will propose possible solutions to overcome each problem.

In order to determine the MFCCs in a hearing aid, we start from the squared values of the STFT. These values are usually calculated in the main algorithms of the DSP, since they are used for the multi-band compression-expansion algorithm

used to adapt sounds to the hearing losses of the patient. In some cases and in order to optimize the power consumption, specific coprocessors are implemented in order to determine the values of $|X_r[k]|^2$ [13]. So, the computational complexity of the evaluation of the terms $|X_r[k]|^2$ will not be considered in the present paper.

The computational cost associated to the evaluation of the Mel scale triangular filters supposes a number of instructions that must be taken into account. Ideally, the number of Multiplication/AC cumulation (MAC) instructions required can be proportional to the product $F(N/2 + 1)$, being F the number of filter-banks and N the frame length (see Eq. 96.1). Fortunately, since most of the terms of $H_m[k]$ are zero, the number of operations is drastically reduced. For instance, in the particular case at hand the proportion of non-zero terms of $H_m[k]$ is around 7.3 % of the total values.

The next group of operations required to determine the MFCCs consists in the evaluation of the DCT and the logarithm at the end of the MFCC extraction. These two parts of the feature extraction process usually consume a high proportion of the available DSP power, and therefore they will be deeply studied in this section.

96.3.1 Proposal 1: Approximated Evaluation of the Logarithm

The logarithms are implemented in a typical DSP architecture using 13 assembler instructions. Taking into account that every time frame we must determine the logarithm of F terms (see Eq. 96.2), this operation requires a considerable amount of instructions in the problem at hand. In order to reduce the number of instructions, in this paper we propose the use of the Most Significant Bit (MSB) (the position of the first non zero bit in the binary representation of the register) in order to approximate the logarithm function. This operation only requires one assembler instruction, and it can be seen as an approach of the logarithm function. So, we can obtain a very fast way of approximating the \log_2 by means of the position of the most significant bit in the accumulator using Eq. (96.4).

$$\log(a) = \log(2) \cdot \log_2(a) \simeq \log(2)(msb(a) + 0.5), \quad (96.4)$$

where $msb(\oplus)$ represents the MSB operation.

The multiplication by the constant $\log(2)$ and the addition of 0.5 can be suppressed, since it only supposes a scale and shift over the features of the classifier, and its use will only change the values of the trained weights of the classifiers, but it will not alter the decision rule. So, the proposed replacement of the logarithm function in Eqs. (96.2) and (96.5) by the MSB function significantly reduces the number of required assembler instructions.

96.3.2 Proposal 2: Removal of the DCT Block

The DCT is evaluated by means of a set of linear combinations of the F filter banks using Eq. (96.2). This supposes F^2 MAC operations for each time frame.

In the particular case of determining the mean value of the MFCCs as features with a linear classifier, the use of the DCT can be suppressed, since the performance of the classifier without the DCT is completely equivalent. The evaluation of the mean value supposes a linear combination of the DCT, and the classifier implements linear combinations of these linear combinations. Since the weights of the linear classifier are determined in order to minimize the mean square error, the use of the DCT supposes a change in the values of the weights v_{cn} but it will not alter the values of the linear combination of the classifiers (given by Eq. (96.3) with $w_{cnp} = 0, \forall n \neq p$), since a linear combination of a linear combination is another linear combination. Therefore, the final error of this particular classification scheme (mean of the MFCCs and linear classifier) is completely independent of the use of the DCT in the feature extraction process.

So, the proposed tailored MFCCs removes the DCT block extraction, so that Eq. (96.2) is replaced by Eq. (96.5).

$$MFCC'_{mt} = \log(E_{mt}) \quad (96.5)$$

In the case of a more complex classifier, or in case of the use of the standard deviation of the features, the error rate might vary with the suppression of the DCT block, but the changes in terms of error rate could not compensate the drawback of the added computational complexity.

96.4 Experimental Work

Prior to the description of the different experiments we have carried out, it is worth having a look at the sound database we have used. It consists of a total of 2343 s of audio, including both speech in quiet, speech in noise, speech in music, vocal music, instrumental music and noise. The database was manually labelled, obtaining a total of 781 s of speech in quiet, speech in music and speech in noise, 781 s of music, and 781 s of noise, so we will work in a three class problem ($C = 3$, speech, music and noise). All audio files are monophonic, and were sampled with a sampling frequency of $F_s = 16$ kHz and 16 bits per sample. Speech and music files were provided by D. Ellis, and recorded by E. Scheirer and M. Slaney [14]. This database has already been used in a number of different works [15]. Speech was recorded by digitally sampling FM radio stations, using a variety of stations, content styles and levels, and contains samples from both male and female speakers. The sound files present different input levels, with a range of 30 dB between the lowest and the highest, which allows us to test the robustness of

the classification system against different sound input levels. Music includes samples of jazz, pop, country, salsa, reggae, classical, various non-Western styles, various sorts of rock, and new age music, both with and without vocals. Finally, noise files include sounds from the following environments: aircraft, bus, cafe, car, kindergarten, living room, nature, school, shop, sports, traffic, train, and train station. These noise sources have been artificially mixed with those of speech files (with varying degrees of reverberation) at different Signal to Noise Ratios (SNRs) ranging from 0 to 10 dB. In a number of experiments, these values have been found to be representative enough regarding the following perceptual criteria: lower SNRs could be treated by the hearing aid as noise, and higher SNRs could be considered as clean speech.

For training and testing, it is necessary for the database to be divided into two different sets. 1414 s (H 60 %) for training, and 930 s (H 40 %) for testing. This division has been randomly carried out, ensuring that the relative proportion of files of each category is preserved for each set. The training set is used to determine the weights of the classifiers in the training process, and the test set is used to assess the classifier's quality after training. The test set has remained unaltered for all the experiments described in this paper.

Each file was processed using the hearing aid simulator described in [16] without feedback. The features were computed from the output of the Weighted Overlap-Add (WOLA) filter-bank with $N = 128$ DFT points and analysis and synthesis window lengths of 256 samples. So, the time/frequency decomposition is performed with 65 frequency bands. Concerning the architecture, the simulator has been configured for a 16-bit word-length Harvard Architecture with a MAC unit that multiplies 16-bit registers and stores the result in a 40-bit accumulator.

In order to study the effects of the limited precision, the classifiers were configured for taking a decision with time slots of 20 ms. The objective is to study the effects of the limited computational capability in a classification scenario in which a small time scale is required like, for example, in noise reduction or sound source separation applications. The results we have illustrated below show the average probability of classification error for the test set and the computational complexity of the considered in percentage of computational load for a 5 MIPS standard hearing aid DSP. The probability of classification error represents the average number of time slots that are misclassified in the test set.

It is important to highlight that in a real classification system the classification evidence can be accumulated across the time for achieving lower error rates. This fact makes necessary a study of the tradeoff between the selected time scale, the integration of decision for consecutive time slots, the performance of the final system and the required computational complexity. This analysis is out of the scope of this paper, since our aim is not to propose a particular classification system, that must be tuned for the considered hearing aid application, but to illustrate a set of tools and strategies that can be used for determining the way an MFCC based classifiers can efficiently be implemented in real time for sound environment classification tasks with limited computational capabilities.

With the aim of carrying out an experimental validation of the two proposals, several experiments have been carried out under different scenarios. In these experiments, the objective is to determine the effectiveness of the proposed approximations over both the error rate and the computational cost. The parameters of this second group of experiments are:

- *Features*: We have carried out experiments with 25 MFCCs using the Proposal 1 (MSB approximation of the logarithmic function) and the Proposal 2 (removal of the DCT block in the MFCC estimation), described in the above section. Furthermore, we have also carried out experiments combining both proposals. For comparison purposes, we also run experiments using (1) a set of four classic features (spectral centroid, spectral flux, voice to white and short time energy) described in [5], (2) the 25 standard MFCCs, and the first 12 MFCCs, as it was described in [6].
- *Statistics*: In order to evaluate the importance of the selected statistics of the features, two different choices have been considered and compared: first, the use of the mean value of the different features along the 20 ms time slot has been studied; second, both the mean and the standard deviation of the features have been used as input vector for the corresponding classifiers.
- *Classifiers*: Four different classifiers have been evaluated for each feature and statistic combination: the LSLC, LSDQC, a two layer MLP, and a two layer QMLP. Both the MLP and the QMLP have been configured with 10 tan-sigmoidal neurons in the first layer and three linear neurons in the output layer (corresponding to the three classes considered in the experiments).

Table 96.1 shows the results obtained in the experiments. As we can see, the use of the 4 classic features supposes the worst results in terms of error rate, but the associated computational cost is quite low when compared to the use of 25 standard MFCCs. On the other hand, the use of the first 12 MFCCs supposes a reduction in the DSP load, but in most of the cases with a consequent increment in the average error rates.

Concerning the approximation of the logarithm by the MSB operator (Proposal 1), we can see that it supposes a slight increment in the error rate in all the cases when compared to the standard use of MFCCs (an average relative increase of 1.6 % in error rate), but a considerable reduction in the computational cost (an average relative reduction of 24 % in the DSP load). This fact makes the proposed MSB approximation suitable for most of the real time hearing aid scenarios.

As it was expected, we can check that the results achieved with the mean value of 25 standard MFCCs and with the Proposal 2 for the LSLC are equal. As we demonstrated in the last section, in that case the DCT block does not alter the classifier performance and therefore it can be removed. But, furthermore, in the case of removing the DCT (Proposal 2) with other classifier or in the case of also using the standard deviation, the error rate is not only different but even lower than in the case of the standard use of MFCCs (average relative reduction of 2.66 %). Moreover, the removal of the DCT block supposes an average reduction of 50 %

Table 96.1 Test error rate and DSP load for the experiments described in the paper

Features	Classifier (%)	Mean of the features					Mean and standard deviation of the features				
		LSLC (%)	LSDQC (%)	MLP (%)	QMLP (%)	LMLP (%)	LSLC (%)	LSDQC (%)	MLP (%)	QMLP (%)	LMLP (%)
Referenced	4 classic features [5]	37.23	35.83	29.01	28.43	36.40	34.52	28.34	27.01		
	DSP load	1.43	1.45	1.60	1.65	2.02	2.06	2.22	2.31		
	25 standard MFCCs [8]	22.64	19.23	15.32	16.62	20.60	17.20	14.24	12.72		
	DSP load	5.55	5.68	5.87	6.17	6.20	6.45	6.70	7.30		
	First 12 MFCCs [6]	24.49	21.58	16.35	15.85	21.95	18.81	15.15	13.92		
Proposed	DSP load	3.89	3.95	4.12	4.26	4.50	4.62	4.81	5.10		
	Error rate	22.63	19.26	15.57	14.36	20.67	17.36	14.26	13.30		
	DSP load	4.05	4.18	4.37	4.67	4.70	4.95	5.20	5.80		
	Proposal 1: MSB-MFCCs	22.64	19.38	15.03	13.67	19.80	16.15	13.83	11.75		
	Proposal 2: NODCT	2.43	2.55	2.75	3.05	3.08	3.33	3.57	4.17		
Proposed 1 + 2	Error rate	22.63	19.40	15.34	13.99	19.96	16.27	13.22	12.35		
	DSP load	0.93	1.05	1.25	1.55	1.58	1.83	2.07	2.67		

of the DSP load, which is quite important in order to minimize power consumption.

When both proposals are implemented (MSB approach and DCT removal) we get the lowest computational costs (74 % lower than the use of standard MFCCs, and 14 % lower than the use of the classic features), and the error rates are even lower than those obtained by the standard MFCCs. This choice might be suitable for those applications in which the computational cost is more constrained.

Concerning the comparison between classifiers, we can see that their use allow to establish tradeoffs between computational cost and error rate. So, the best result in terms of computational cost is obtained by the LSLC classifier with the mean value of the MFCCs under Proposals 1 and 2 (22.63 % of error rate with a computational load of 0.93 %), and in those cases in which we desire to get the best performance in terms of error rate, the QMLP achieves the lowest error rates (for instance, 12.35 % with the mean and standard deviation of the same features with only a 2.67 % of the DSP load).

96.5 Conclusions

This paper has been motivated by the fact that the implementation of signal processing techniques on hearing aids is strongly constrained by the small number of instructions per second to implement the algorithms on the digital signal processor the hearing aid is based on. In this respect, the objective of this paper has been the proposal of a set of approximations in order to optimize the implementation of standard MFCC based sound environment classifiers in real hearing aids. The performance of the proposed solutions show a balance between keeping error classification probability within low values (in order to not disturb the user's comfort) and achieving this by using a small number of instructions per second. The reason underlying these restrictions is that hearing aids have to work at low clock rates in order to minimize the power consumption and maximize battery life.

The final, global conclusion is that the suppression of the DCT block from the MFCC extraction process is suitable, since its use does not suppose an improvement in terms of error rate, and it supposes a high computational load. Furthermore, the use of the MSB operator instead of the logarithm in the MFCC supposes a considerable reduction in the computational load while obtaining comparable results in terms of error rate.

Acknowledgments This work has been partially funded by the University of Alcalá (CCG2013/EXP-074), the Spanish Ministry of Economy and Competitiveness (TEC2012-38142-C04-02) and the Spanish Ministry of Defense (DN8644-ATREC).

References

1. Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., Rass, U.: Signal processing in high-end hearing aids: state of the art, challenges, and future trends. *EURASIP J. Appl. Sig. Process.* **18**, 2915–2929 (2005)
2. Büchler, M., Allegro, S., Launer, S., Dillier, N.: Sound classification in hearing aids inspired by auditory scene analysis. *EURASIP J. Appl. Sig. Process.* **18**, 2991–3002 (2005)
3. Marzinzik, M.: Noise reduction schemes for digital hearing aids and their use for hearing impaired. PhD thesis, Carl von Ossietzky University Oldenburg (2000)
4. Dong, R., Hermann, D., Cornu, E., Chau, E.: Low-power implementation of an hmm-based sound environment classification algorithm for hearing aid application. In: *Proceedings of the 15th European Signal Processing Conference (EUSIPCO 2007)*, vol. 4 (2007)
5. Gil-Pita, R., Alexandre, E., Cuadra, L., Vicen, R., Rosa-Zurera, M.: Analysis of the effects of finite precision in neural network-based sound classifiers for digital hearing aids. *EURASIP Journal on Advances in Signal Processing*, vol. 2009. (2009)
6. Xiang, J.J., McKinney, M.F., Fitz, K., Zhang, T.: Evaluation of sound classification algorithms for hearing aid applications. In: *Acoustics Speech and Signal Processing (ICASSP)*, 2010 IEEE International Conference on, IEEE, pp. 185–188. (2010)
7. Hunt, M., Lennig, M., Mermelstein, P.: Experiments in syllable-based recognition of continuous speech. In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80, IEEE* vol. 5, pp. 880–883. (1980)
8. Mohino-Herranz, I., Gil-Pita, R., Alonso-Diaz, S., Rosa-Zurera, M.: Synthetical enlargement of mfcc based training sets for emotion recognition. *Int. J. Comput. Sci. Inf. Technol.* **6**(1), 249–259 (2014)
9. Ye, J.: Least squares linear discriminant analysis. In: *Proceedings of the 24th international conference on Machine learning, ACM* pp. 1087–1093. (2007)
10. Bishop, C.M., Nasrabadi, N.M.: *Pattern recognition and machine learning*, vol. 1. Springer, New York (2006)
11. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**(2), 431–441 (1963)
12. Kosmatopoulos, E.B., Polycarpou, M.M., Christodoulou, M.A., Ioannou, P.A.: High-order neural network structures for identification of dynamical systems. *Neural Networks, IEEE Trans.* **6**(2), 422–431 (1995)
13. Cuadra, L., Alexandre, E., Gil-Pita, R., Vicen-Bueno, R., Álvarez-Pérez, L.: Influence of acoustic feedback on the learning strategies of neural network-based sound classifiers in digital hearing aids. *EURASIP Journal on Advances in Signal Processing*, pp. 14. (2009)
14. Scheirer, E., Slaney, M.: Construction and evaluation of a robust multifeature speech/music discriminator. In: *ICASSP*, pp. 1331–1334. (1997)
15. Thoshkahna, B., Sudha, V., Reemigration, K.: A speech-music discriminator using hilm model based features. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 425–428 (2006)
16. Vicen-Bueno, R., Gil-Pita, R., Utrilla-Manso, M., Álvarez-Pérez, L.: A hearing aid simulator to test adaptive signal processing algorithms. In: *IEEE Int. Symposium on Intelligent Signal Processing (WISP)*, pp. 619–624. (2007)

Chapter 97

Metamodelling Architecture for Modelling Domains with Different Mathematical Structure

Vitaliy Mezhujev

Abstract The new metamodelling approach for domain specific modelling is proposed in the paper. The additional level of the metamodelling architecture is introduced, which gives the possibility of metamodels development in the different mathematical semantics. This allows to take into account the mathematical structure of modelled domains, and to use the mathematical operations for development of new effective methods for solving domain specific tasks. The applicability of the approach for development of metamodels for modelling different domains is shown.

Keywords Domain specific modelling · Metamodel · Metamodelling architecture · Mathematical structure · Formal system

97.1 Introduction

The methodology of Domain Specific Modelling (DSM) becomes more and more popular today, allowing to overcome the known issues of the “universal” modelling approach [1]. The sense of DSM is development of Domain Specific Languages (DSLs), applicable for modelling properties of particular domains. A DSL is built inside a so called metamodel, defining the concrete syntax of the language. The abstract syntax of a DSL is defined in the frame of the meta-metamodel as e.g. MOF [2], GOPRR [3], MGA [4] etc.

V. Mezhujev (✉)

Faculty of Computer Systems and Software Engineering, University Malaysia Pahang,
Gambang, Malaysia
e-mail: mejuev@ukr.net

Emphasizing the power of the existing DSM approaches, they have a number of issues, caused by the lack of generalisation and formalisation:

- the metamodel based DSLs are mostly descriptive, i.e. not expressive for the definition of methods for solving domain specific tasks;
- the applicability of a DSL by the generation of software data and code is limited;
- while the DSM approach is intended for using by domain experts, the obligatory involvement of IT specialists for development of code generators is needed;
- for code generation an additional external language should be used, which is not linked with specifics of a modelled domain;
- the meta-metamodel, used for metamodels development, does not reflect the mathematical structure of a considered domain and is hardcoded inside a DSM tool.

Let's consider the principles of the proposed approach to the metamodels development, allowing to overcome the specified above issues:

- the formal definition of the object of modelling—the domain, as the set of entities, linked by the forming mathematical structure and the domain specific relationships;
- the definition of the meta-metamodel and the metamodel as the formal systems, allowing to fix correspondingly the structural and domain specific properties;
- the mathematical structure of a domain is defined at the meta-metamodel level and next is used as the carrier of domain specific properties;
- the additional level of the metamodelling architecture is introduced, which allows to develop the meta-metamodels, having different mathematical semantics.

While the existing metamodelling approaches use the predefined mathematical formalisms (mostly, graphs) for structuring domain properties, here the development of meta-metamodels in the different mathematical semantics is possible. Additional level of the metamodelling architecture allows to express properties of domains in terms of set theory and to reflect different mathematical structures (algebraic, topological, differential, geometrical etc.). Corresponding mathematical operations are integrated in the metamodel and used for solving domain specific tasks. Generation of software data and code becomes the partial case of the proposed metamodelling approach.

The paper is organized as follows. First the new metamodelling architecture is discussed in comparison with existing approaches. Section 97.3 of the paper shows applicability of the proposed approach for producing the graph based metamodels for modelling software systems. Section 97.4 expands the practical applications for requirements engineering, business process modelling and solving tasks of multidimensional physical domains. The conclusion, plan of future research and references list finalize the paper.

97.2 Metamodelling Architectures

The methodology MOF (Meta Object Facility) [2] was used by the OMG (Object Management Group) consortium for development of the Unified Modelling Language (UML). MOF has the four levels of the metamodelling architecture. The top level is the meta-metamodel (M3), defining the language for development of the metamodels (having the level M2). The level M2 (here, UML) used for development of the domain models of the level M1 (the UML-models). The last is the level of data (M0), describing the concrete instances of M1. The MOF architecture is based on the object-oriented methodology of software systems design.

The meta-metamodel GOPPRR (Graph-Object-Property-Port-Role-Relationship) allows to produce metamodels inside the graph based notations, by means of connection of objects by relationships, definition of domain properties (attributes) and roles [3]. Each of the GOPPRR concepts a metatype is called. As MOF, the metamodelling architecture of the GOPPRR in four levels can be shown (see the Fig. 97.1).

The proposed approach also has the multiple-level metamodelling architecture, but its semantics differs from the existing methodologies. All of the metamodels are considered to be formal systems; they contain an alphabet of types, a grammar and operations. We introduce the additional level of the metamodelling architecture—the meta-meta-metamodel (M4), as a formal system, that is built on the basis of set theory. M4 includes the meta-metatype “element of a set”, set operations and grammar rules, which (taken together) allow us to specify a set structure. This approach allows us to consider a domain as a set of heterogeneous entities, having domain specific properties and linked by different kinds of mathematical structures.

Formally, we define a domain as a set of entities D , linked by structural S and domain specific P relationships:

$$D = \{d_1, d_2, \dots, d_N\}, S, P \subseteq D \times D \quad (97.1)$$

where N is a power of D . Each element of D can have attributes, which we consider as unary relationships on D . 0-ary relationships are used to identify elements of D . Binary and other relationships are used to fix mathematical structure of D .

All of the levels of the proposed metamodelling architecture contain not only descriptive elements, such as in MOF or GOPPRR, but also procedural part, implemented with software functions.

Following our proposal, the architecture for development of the graph based metamodel on the Fig. 97.2 is shown. Here a node and an edge of a graph serve as the mathematical metatypes for development of domain specific metamodels types (an attribute is the inherent part of a node and of an edge). The node and the edge are produced from the meta-meta-metamodel as the having algebraic structure subsets of the composing domain entities. Note, while GOPPRR [3] and MGA [4] also use the graphs for structuring domain specific properties, this is a partial case

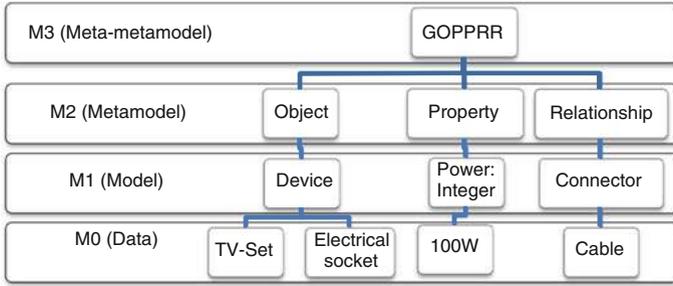


Fig. 97.1 The GOPPRR metamodeling architecture

of the proposed approach, where development and using the different mathematical structures is possible.

The implementation of mathematical operations of the metamodels at all levels of the proposed architecture, forms the Application Program Interface (API) of the corresponding software tool. The API of M4 contains the methods for manipulation with the elements of a set of composing domain entities. The API of M3 is the operations with subsets (e.g., with a node and an edge of a graph, and in the general case with any model objects of the considered domain). For M2, the API contains the metamodel processing routines (here, the metatypes of the level M3 become domain-specific types, i.e., to the mathematical subsets the semantics of the domain is assigned). M1 contains instances of the types and definitions of domain-specific methods, implemented with the APIs of all the previous levels. M0 is data values and processes in the computer memory (instances of the methods, defined at the level M1).

97.3 Development of Graph-Based Metamodels

Let us consider the mathematical method for producing the graph-based meta-metamodel in the context of proposed approach. Its alphabet includes the meta-types node N and edge E of the graph $Gr = (N, E)$; the grammar G_{Gr} is the set of rules, defining the possibility $\{true, false\}$ of the connection of nodes n_i, n_j by the edge $e_k = (n_i, n_j)$, $n \in N, e \in E$

$$G_{Gr} = \{ (n_i, n_j) | g_k \in \{true, false\}, n_i, n_j \in N, i, j = 1..M, k = 1..K \} \quad (97.2)$$

where M is a power of N . The number of rules K depends on the properties of the graph Gr (is it directed, are loops possible, etc.).

At the level of metamodel development, to the nodes and the edges of the meta-metamodel the semantics of domain is assigned. For example, the node N can be the metatype for definition of the types of software tasks and synchronization

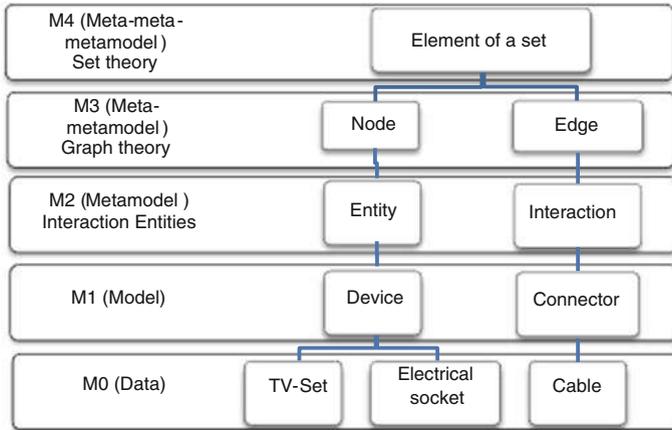


Fig. 97.2 The levels of the proposed metamodeling architecture

objects, and the edge E can be the metatype for definition of the types of channels (communication protocols) between tasks and synchronization objects. This metamodel will include the alphabet, containing typical for parallel programming synchronisation objects (critical section, mutex, semaphore, resource, FIFO etc.) and software tasks (driver, application etc.); the grammar rules, specifying the valid interactions of software tasks via synchronisation objects, and operations, used for definition of code generation functions.

Table 97.1 shows an example of the definition of the metamodel for modelling the parallel concurrent software system inside the graph based meta-metamodel.

In this example, a Node and an Edge are the mathematical metatypes of graph based meta-metamodel M3. Domain specific types are the nodes Task, Sync and the edges PutData, GetData, which compose the alphabet of M2 metamodel and are used to create instances at the M1 level. M2 also defines the grammar rules for combining instances of the types by using predicates PutData(Task, Sync) and GetData(Sync, Task). These grammar rules correspond to the edges of the graph-based meta-metamodel and are used for development of code generation methods (implemented by walking the graph based model M1). The M1 model of software system includes instances of $Task_1, Task_2 \dots Task_T$ and synchronization objects $Sync_1, Sync_2 \dots Sync_S$, linked by the channels of interaction PutData, GetData (where T, S —are the number of tasks and the number of synchronization objects in the model respectively).

For the interesting reader, to show the applicability of described graph based metamodel, we can refer to the metamodel of interacting entities [5], which was used for development of a real-time operation system [6] and for modelling distributed parallel real-time software [7].

Table 97.1 Levels of metamodelling architecture for a software system modelling

Level		Alphabet	Grammar	Operations/ Methods
M4	Math structures	Elements d of the set D	The rules of grammar, based on the relations $d \in D, \{d\} \subseteq D$	Create /delete element d , subset $\{d\}$
M3		Node $n \in Node$ and edge $e \in Edge$ of graph $G = (Node, Edge)$, $Node, Edge \subset D$	Connection of nodes by edges $e_k (n_i, n_j), n_i, n_j \in Node, e_k \in Edge, i, j = 1.. Node , i \neq j, k = 1.. Edge $	Add edge $G' = G + e$ Delete edge $G' = G - e$ Add node $G' = G + n$ Delete node $G' = G - n$
M2	Domain specific properties	Edge PutData, GetData; Node Task, Sync;	PutData (Task, Sync); GetData (Sync, Task)	Add /delete a type of task Task /sync object Sync , create communication channel PutData , GetData
M1		Task Task1, Task2; Sync Sync1; PutData (Task1, Sync1); GetData (Sync1, Task1);		

The definition of the metamodel alphabet as the set of attributed types and the domain model as the instances of the types, having the concrete values of attributes, make possible the formal checking a model in its state space. Due to including mathematical methods in the metamodel the checking properties of behaviour of a real-time system (e.g. absence of deadlocks) was applied. The graph based methods (e.g. Dijkstra's algorithm) for development of the code generation functions (e.g. routing table of a real time operation system) were used.

97.4 Other Applications of the Metamodelling Approach

Except development of graph-based metamodel for software systems design, the applicability of the proposed approach was proven for the next domains:

- requirements engineering (RE), where conceptual metamodelling for systems specification was used. The set of the typical for the RE concepts formed the alphabet of the metamodel, which symbols were the types for instantiation—definition of the concrete statements describing a system properties and behaviour. The methods of the graph based meta-metamodel were used to

check correspondence of the graph of architectural decomposition to the graph of initial requirements, generate the document of systems specifications, made the control of versions etc. The conceptual metamodel was further expanded by the Finite State Machine formalism [5]. This allows us to build the domain specific models of processes on the base of the ontology of a domain. To each concept of ontology the state transition attribute was added. The process grammar was the set of rules, defining the state transitions of conceptual model of a system description. e.g. only after capturing requirements user can move to the specification stage, next to the phase of architectural modelling etc. Such the approach allows us to manage users activity to achieve the goal of a process in a given time (up to deadline);

- development of the metamodel, based on the vector algebra and the logic of syllogisms. Here vectors were used as the metatypes for producing the logical types of the metamodel alphabet. In the practical implementation [8], the alphabet of the metamodel on the base of the types of categorical syllogisms was developed. Due to using vector algebra for the definition of the metamodel, the operations on syllogisms as operations on vectors in linear vector space were implemented. This allows us to develop the algorithm for automatic geometrical theorem proving. The approach was used for development of the logic for optical computers, where at physical level vectors were implemented as laser beams;
- development of the metamodel for multidimensional physical domains [9]. The alphabet of the meta-metamodel was defined as the set of the basic (corresponding to the dimensions of the physical space) geometrical objects, i.e. point, line, surface and 3D region. For metamodels development we set distributions of physical properties among the defined with the meta-metamodel geometrical structures. Due to considering objects as the sets of geometrical points in the physical space, the grammar of the metamodel in the terms of Boolean operations on geometrical subsets was defined. This grammar limits the possible compositions of the geometrical objects in the 3D space. The mathematical methods of the metamodel correspond to the solutions of multidimensional tasks of the integral and differential calculus. As the interesting application of the metamodelling approach for physical domains the design of metamaterials (artificial composites with specific optical properties) can be mentioned [10].

97.5 Plan of Future Research

The plan of research is further exploring the properties of the metamodels, allowing to fix different mathematical structures:

- the formal definition of the metamodels, the mathematical structure of its types, grammars and operations at all levels of the metamodelling architecture;

- learning the linguistic properties of the metamodels, incl. the possibility of reduction of the grammars into the normal Chomsky form;
- definition of the method for metamodels composition, allowing to combine the declarative and imperative constructs (alphabet, grammar and operations);
- exploring the textual and the visual forms of expression of metamodels and development of the method for its combination;
- expansion of the approach on the other types of mathematical structures (metric, geometrical, differential, topological, etc.).

97.6 Conclusion

The new approach for metamodels development is proposed. The metamodelling architecture is decomposed into the layers, allowing to fix the structural and the domain specific properties. This allows to take into account the mathematical structure of considered domains. The additional set-based level of the metamodelling architecture is introduced, which allows to define the meta-metamodels in the different mathematical semantics.

References

1. France, R.B., Ghosh, S., Dinh-Trong, T., Solberg, A.: Model-Driven development using UML 2.0: promises and pitfalls. *IEEE Comp.* **39**(2):59–66 (2006)
2. ISO/IEC 19502:2005, Information technology. Meta object facility.—ANSI, p. 292 (2007)
3. Kelly, S., Juha-Pekka, T.: Domain-Specific Modeling: Enabling Full Code Generation, p. 427. Wiley-IEEE Computer Society Pr. (2008)
4. Nordstrom, DD.: Metamodeling—Rapid Design and Evolution of Domain-Specific Modeling Environments/Dissertation for the Degree of Doctor of Philosophy in Electrical Engineering, p. 170. Nashville, Tennessee (1999)
5. Mezhuyev, V., Sputh, B., Verhulst, E.: Interacting entities modelling methodology for robust systems design. In: Second International Conference on Advances in System Testing and Validation Lifecycle, pp. 75–80. CPS Publishing (2010)
6. Boute, R.T., Miguel, J., Faria, S., Sputh, B.H.C.: Vitaliy Mezhuyev Formal Development of a Network-Centric RTOS/Eric Verhulst, pp. 227. Springer, Berlin (2011)
7. Mezhuyev, V.: domain specific modelling distributed parallel real time applications. *Syst. Inf. Process.* **5**(86), 98–103 (2010)
8. Mezhuyev, V.: Vector logic: Theoretical Principles and Practical Implementations, pp. 91–97. The papers of Zaporizzia National University, Zaporizzia, ZNU (2006)
9. Mezhuyev, V., Lytvyn, O.: Metamodel for visual modelling multidimensional domains and its practical applications. *Control Syst. Mach.* **4**, 31–43 (2010)
10. Mezhuyev, V., Pérez-Rodríguez, F.: Visual Environment for Metamaterials Modelling. Some Current Topics in Condensed Matter Physics, pp. 1–13. Universidad Autónoma del Estado de Morelos (2010)

Chapter 98

Use Case Based Approach to Analyze Software Change Impact and Its Regression Test Effort Estimation

Avinash Gupta, Aprna Tripathi and Dharmendra Singh Kuswaha

Abstract Software needs to be maintained and changed to satisfy the new requirement and existing faults. Without analyzed changes, change implemented to software often cause unexpected ripple effects. To avoid this and diminish the risk of performing undesirable changes, an impact analysis of the change is done. Software Change Impact Analysis (SCIA) needs to be computed at every change request for software systems, to access the impact information for several critical software engineering tasks such as risk analysis, effort estimation, and regression testing. The use of UML analysis/design models on large projects lead to a large number of interdependent UML diagrams. Paper proposes a UML model based approach strictly to use use-case diagram for impact analysis that is applicable in early decision making and change planning. Later, by using the impact set we estimate the regression test effort required for the effected change in the software. The reduction in test effort observed ranges from 47 to 95 %, saving significant software testing cost.

Keywords Impact analysis · Use-case · Test case · Regression testing · SCIA

98.1 Introduction

Requirement of users are increasing day by day. To fulfil user's requirements and to compete in market, software companies need to enhance their product's performance and functionality before let it being obsolete in market. So software change

A. Gupta · A. Tripathi (✉) · D.S. Kuswaha
MNNIT Allahabad, Allahabad, India
e-mail: aprnatrpathi@gmail.com

A. Gupta
e-mail: rcs1051@mnnit.ac.in

D.S. Kuswaha
e-mail: dsk@mnnit.ac.in

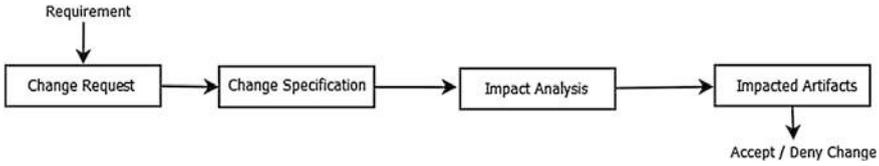


Fig. 98.1 Software change impact analysis process model

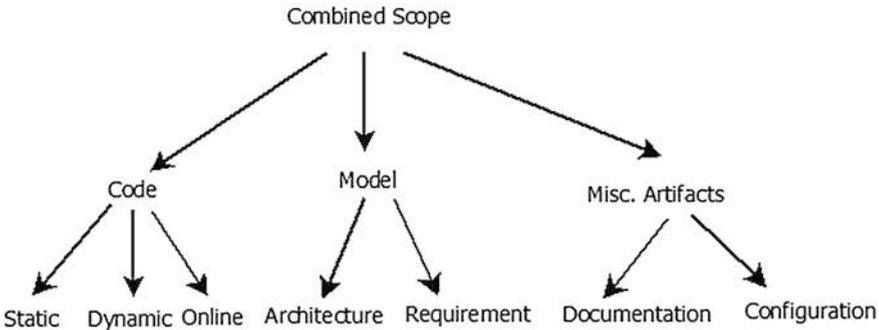


Fig. 98.2 Impact analysis classification by Lehnart

management is essential activity to enrich software age. SCIA is an essential activity before implementing change in software. It allows to measure the effort required implementing a change and its ripple effects, impact analysis suggests those software artefacts which may be changed, and helps to identify test cases which should be re-executed to ensure that the change was implemented correctly [1].

SCIA has following steps as shown in Fig. 98.1.

Common techniques used to implement change impact analysis are based on either traceability or dependency relationships between the software artefacts. Traceability-based impact analysis techniques work on analyzing the relationships between requirements and other development artefacts (such as design, implementation and test cases) to determine the scope of the anticipated changes, dependency-based impact analysis techniques work on a more detailed level by analyzing the relationships between the artefacts of the same development phase. Lehnart et al. [2] proposes impact analysis technique’s classification based on evaluation artifacts as shown in Fig. 98.2.

98.1.1 Model Based Software Change Impact Analysis

Model based analysis provides facility of SCIA for software models to keep their quality and correctness at early stage. Models, such as UML diagrams, enable the assessment of architectural changes on a more abstract level than source code.

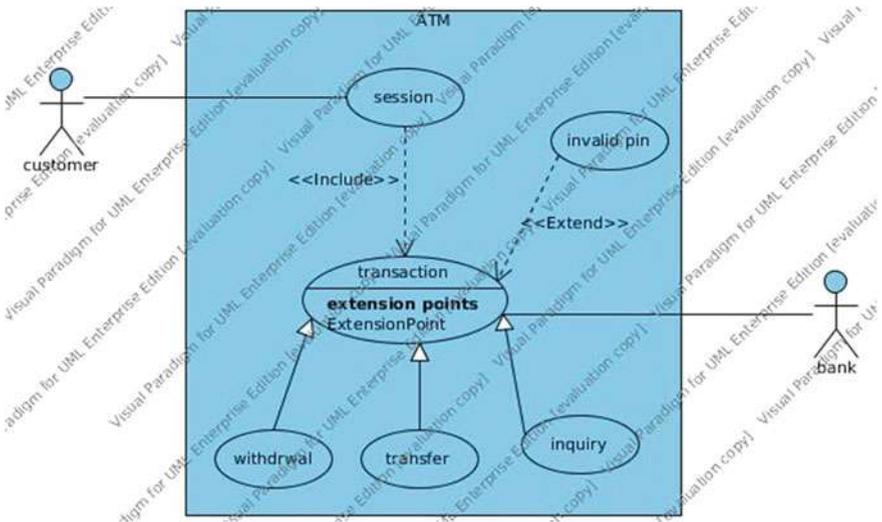


Fig. 98.3 Use-case diagram designed by visual paradigm tool

This enables SCIA in earlier stages of development and in MDD, which has become more important in recent years. But dependent on the underlying modelling language, even model based analysis provides effective impact analysis results, for example when analyzing detailed UML class diagrams. Thus, SCIA is an essential activity of software Maintenance. Proposed approach is concerned towards the automation of SCIA in MDD environment that provides an automated SCIA approach for use-case model scheme by using flow of events written in Visual Paradigm [3] Modelling Tool.

This paper has been organized into the following sections. Section 98.2 describes the related work done in this field. Section 98.3 provides the proposed approaches and implementation. Section 98.4 presents simulation and comparison results. Finally, Sect. 98.5 includes the conclusion and future directions (Fig. 98.3).

98.2 State of the Art

There are various strategies for performing SCIA, These strategies are based on some parameters which is considered during their requirements engineering process the execution phase. These strategies can be broadly classified as following: Automatable and Manual. These strategies require fewer infrastructures, but may be harder in their impact estimation than the automatable ones.

98.2.1 Traceability and Dependency Analysis

Antoniol et al. [4] and Corley et al. [5] proposes that traceability and dependency analysis both involve examining relationships among entities in the software. They differ in scope and detail level; traceability analysis is the analysis of relationships among all types of models, while dependency analysis is the analysis of low-level dependencies extracted from source code. By extracting dependencies from source code, it is possible to obtain call graphs, control structures, data graphs and so on [6]. Since source code is the most exact representation of the system [5], any analysis based on it can very precisely predict the impact of a change. Dependency analysis is also the most effective strategy for impact analysis available. The drawback of using source code is that it is not available until late in the project [7], which makes dependency analysis narrow in its field of application. The identification of the primary fundamental point of traceability is based on a predefined search strategy and a multi-step selection process [8] When requirements traceability exists down to the source, it can, however, be very efficient to use source code dependencies in order to determine the impact of requirements changes. A drawback is that very large systems have huge amounts of source code dependencies, which make the dependency relationship difficult to use.

98.2.2 Program Slicing Technique

Slicing [9] attempts to understand dependencies using different independent slices of the program. The whole program is sliced into a decomposition slice, which contains the place of the change, and the rest of the program, a complement slice. Slicing is based on data and control dependencies in the program. Changes made to the decomposition slice around the variable that the slice is based on are guaranteed not to affect the complement slice. Slicing technique limits the scope for propagation of change and makes that scope explicit. Slicing technique is also used for slicing of documents in order to account for ripple effects as a part of impact analysis. Slicing techniques can be useful in requirements engineering to isolate the impact of a requirements change to a specific part of the system. In order to provide a starting point for the slicing technique, the direct impact of the change must first be assessed.

98.2.3 Model Driven Development

Model-Driven Development (MDD) is a model-based software development approach which aims at improving build time and quality of software artifacts by focusing on UML models as abstract level artifacts rather than code [10].

UML models can be defined at different levels of abstraction to represent various aspects of the system. Dependencies between software life cycle objects are becoming more complex as many software systems grow beyond a million of lines of code. So, MDD is helpful in such type of systems. The potential benefits of using models are significantly greater in software than in other engineering disciplines because of the potential for a seamless link between models and the systems they represent.

In the context of MDD, models also used for many reasons such as, to handle immediate errors, to add new functional requirements, to enhance some quality aspects, or to adapt to a new technological or architectural environment. UML is a standard for modelling software systems and it is extensively used in the area of model-driven development [11]. For this reason, our work is restricted to UML models. There are many UML Models e.g. class diagram, use case diagram, sequence diagram, activity diagram etc. some author considered the hierarchies between the models. Use case model is considered at a very abstract level UML model, and this is very close to main functionality of the software system. Class diagram and sequence diagram is considered the next level of UML models. So our research is focused on use case diagram. A measure of distance between a changed element and potentially impacted elements is proposed by Briand et al. [12] to prioritize the results of impact analysis according to their likelihood of occurrence.

98.2.4 Informational Retrieval

SCIA approach in this thesis, is based on Information Retrieval discussed in [14], used to derive the information contained in use-case flow events, with respect to requested change. Information retrieval processes contains following steps: (1) Building a corpus, (2) Natural-language processing (NLP), (3) Querying, and (4) Estimating an impact set.

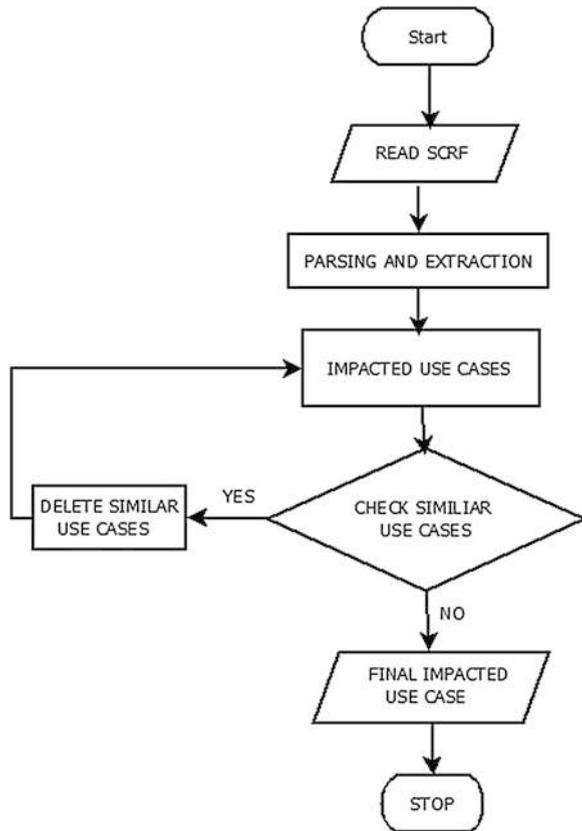
98.3 Proposed Approach

In the proposed approach, the use-case diagram with descriptions is used in SCIA. A software system contains one or more use cases. Each use case has its main flow of event and alternate flow of event, Details of core functionality of use cases in defined in terms of flow events.

Dependency of use cases are utilized in the proposed approach. In the proposed approach, flow of events of use cases is written by using the tool Visual Paradigm Trial version software [3]. The proposed approach includes the steps shown in Fig. 98.4.

Read SCRF: Users or Developer, who want some change in the developed system, fill SCRF as per requirement. After getting the filled SCRF, Controller

Fig. 98.4 Flow chart of proposed approach



Configuration tool (CCT) will analyze the SCRF. The important field of the SCRF i.e. the change requested field is stored in a new directory.

Parsing and Extraction: In this phase, change requested file is parsed. Parser parses the stop words like is, are, am, this, that etc., from the stop word file. The parsed keywords are stored in an output file. **Impacted Use Cases:** This phase is the information retrieval phase. After parsing the SCRF, for all extracted keywords, we search the flow of events directory and gives the impacted use case with respected to each keyword, this process follows recursively run for all keywords. For each flow of event file, respected use case name is stored.

Check Similar Use Cases: In this step similar use cases names are checked. **Delete Similar Use Cases:** In previous step, it might be possible that, there may be some redundant use cases. In this phase these redundant use cases are removed to avoid the ambiguity. **Final Impacted Use Cases:** The final outcome of this step is the name of impacted use cases.

Table 98.1 Mapping of classes with use cases

Use-cases	Mapped classes
Transaction	Log, Logpanel, Receipt, Simreceiptprinter, Transaction, Inquiry, Balances,
Inquiry	Message
Session	Simcardreader, Simdisplay, Simcarderader, Card, Cardreader, Carddispencer,
Withdrawal	Session, Simkeyboard, Accountinformation, Withdrawal, Money, Balance,
Transfer	Deposit, Transfer
Invalidpin	Invalid, Authentication, Status, Checker

98.3.1 Implementation and Results

Proposed scheme is illustrated by the ten open source java based and some self made projects.

The proposed approach has been implemented in three phases. To gather the change request from the user or developer, SCRF is needed, thus in first phase a SCRF is designed in java language. This SCRF form contains many fields like priority, Type of change, Document affected etc.

In second phase, a NLP parser is written in java programming language. This NLP parses the stop words (i.e. in computing, stop words are words which are filtered out prior to, or after, processing of natural language data (text) [2]. There is not one definite list of stop words which all tools used. Stop word list are updates as per CCT decisions. More than 600 stop words are managed to parse. These stop words will growing, based on decisions taken by the CCT. After parsing change request we got some keywords as per desired change requested. To validate our approach we have taken ten examples, these examples flow of events are written in visual paradigm trial version software [3].

In third phase i.e. IR part, with the help of parsed keywords, dependent use case flow of events are filtered with respect to change request. Extracted keywords of change request are used and retrieve impacted use case names from the flow of events of the system. For this, a C program is developed. Output of this program is names of the impacted use cases with respected to the change request.

98.4 Comparison

Proposed approach is compared with Yin Li et al. [13] approach, which is based on dependences between the system classes. For comparison, Li approach is applied on the project that is used to evaluate proposed approach. Since, results achieved through proposed approach are based on the use-case diagram, thus use-cases are mapped with the classes. Table 98.2 maps the use-cases with classes of project ATM. Table 98.1 filters the impacted classes with respect to the impacted use-cases of project ATM for the change request “Ask for PAN validation if transfer or

Table 98.2 Impacted use-cases identified through proposed approach and respected classes for Project “ATM”

Impacted use-cases	Impacted classes
Transaction	Log, Logpanel, Receipt, Simreceiptprinter, Transaction,
Withdrawal	Accountinformation, withdrawal
Transfer	Money, Balance, Deposit, transfer

Table 98.3 Comparison of proposed approach and Yin Li Existing approach impacted classes for project “ATM”

Proposed approach impacted classes	Yin Li impacted classes
Log, Logpanel, Receipt Simreceiptprinter, Transaction Accountinformation, Withdrawal, Money, Balance, Deposit, Transfer	Log, Logpanel, Receipt Simreceiptprinter, Transaction Accountinformation, Withdrawal, Money, Balance, Deposit, Transfer

withdrawal is more than 3,000”. Table 98.2 illustrates the impacted classes found by proposed approach and Yin Li approach. After analysing the Table 98.3, it is found that the results from proposed approach are equivalent to Li’s approach. From this comparison, it is concluded that proposed approach is more efficient in terms of abstraction.

The next section elaborates a case study to illustrate the reduction in test cases.

98.4.1 Estimation of Regression Testing Effort from the Impact Set

When a change is made to software, the total effort made is the sum of modification effort and testing effort. The testing effort can be reduced if we can identify those test cases from the test suite that may be impacted due to the modifications made. Other test cases need not be run again and again. Once we have the impact set of use cases and hence the classes, when a change is made in any use case, effort required for regression testing and total effort reduction in testing is computed. Effort Required in Regression Testing (E') = (Total number of test cases of impacted classes / Total number of test cases in Test suite) * E

$$\% \text{ Reduction in Test Effort} = \frac{E - E'}{E} * 100$$

where E is the effort required to run the existing test suite. In the next section we will analyze the area of the class diagram impacted because of the change class and its impact on regression test effort (Table 98.4).

Table 98.4 Classes and their impact area along with percentage reduction in effort required during regression testing

Classes	Number of use cases/ classes impacted	Number of test cases after reduction	% reduction of test cases
Arithmetic	6	99	47
Equality	5	86	54
Bitwise	4	86	54
Additive	4	60	68
Multiplicative	3	47	75
Prefix	3	34	82
Shift	3	34	82
Unary	3	34	82
Compute	2	8	95

98.5 Simulation and Results

We demonstrate the proposed methodology work using a small application “Computer Operations Tutorial (COT)” developed in JAVA that takes a mathematical expression as input, parses it and displays the solution as the output. The exit condition for the application is to give the input “exit”.

For COT, we have following classes whose names are stored in an array A:

$A[] = \{ \text{“Additive”, “Arithmetic”, “Bitwise”, “Compute”, “Main”, “Multiplicative”, “Prefix”, “Shift”, “Unary”, “Equality”} \}$.

Now suppose we make a change to the *Equality* class for implementing the requested change, then the impact set (S) thus produced would be: Equality, Bitwise, Compute, Multiplicative and Main. This can be verified by applying Breadth First Search algorithm to find all connected components with the id of *Equality* class as source, on matrix DM shown in Fig. 98.5. The zero value of $DM[i][j]$ represents the independency between class i and class j while the one shows that the class j depends on the class i.

The requested change can impact the system in two ways: directly or indirectly, as has been depicted in the dependency matrix shown above. When the equality class is considered as the element of change set, the directly impacted class will be bitwise, while compute, multiplicative and main will be indirectly impacted classes. This can be easily observed from Fig. 98.5. The brown colored class (row labelled E in matrix) is the initial class to which the change is requested. The green colored class (row labelled B in matrix) represent the directly impacted classes. Blue colored classes (row labelled M and N in matrix) are the indirectly impacted classes, and the black colored classes are those that are unaffected by the change that has been requested.

For a change requested in Equality classes, in the existing test suite there are 190 test cases, out of which we require to re-run only 86 test cases, thus required

Fig. 98.5 Dependency matrix for equality class (E) of COT

	A	A	B	C	M	M	P	S	U	E
A	0	0	0	1	0	0	1	0	0	0
A	1	0	0	0	0	1	0	0	0	0
B	0	0	0	1	0	1	0	0	0	0
C	0	0	0	0	1	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0
M	0	0	0	1	0	0	0	0	0	0
P	0	0	0	1	0	0	0	0	0	0
S	0	0	0	1	0	0	0	0	0	0
U	0	0	0	1	0	0	0	0	0	0
E	0	0	1	0	0	0	0	0	0	0

regression testing effort for the change will be 0.46E where E is the effort required to run the existing test suite and the percent of average reduction in testing effort will be 54 % of the existing system testing effort.

As the number of impacted classes rises for a class, more the effort is required for regression testing.

98.6 Conclusion and Future Work

Proposed approach reduces the effort that is unnecessary required in source code based SCIA. With the help of proposed approach, impacted use cases are found in an early stage. SCIA is important activity of SCM process. Software system documentation and coding is increasing in massive rate, it is very difficult to impact analysis of whole system manually, for small systems manual SCIA might be useful, but for Large System, SCIA is a tedious job. The approach used in this paper finds the regression test effort as well as test suite required for regression testing based on the impact set that is the sub set of the existing test suite of the system. Obtained results have coarse granularity over the existing approaches. Benefits of Proposed approach, at the use case level, save time and effort required in later stages of change management of the system. Thus the proposed approach allows efficient SCIA of the systems and reduces effort with a significant improvement over the traditional methods. In the proposed approach we are using use case diagram for impact analysis. To refine the SCIA, this approach may be applied in other UML models like in class diagram, sequence diagram, activity diagram. To find the

impact use case, search based information retrieval approach is used. By using advance approaches like artificial intelligence etc., in information retrieval techniques, SCIA will be more efficient and improved. The reduction in test effort observed ranges from 47 to 95 %, saving significant software testing cost.

References

1. Arnold, R., Bohner, S.: Impact analysis-towards a framework for comparison. In *Software Maintenance, 1993. CSM-93, Proceedings., Conference on*, pp. 292–301, (1993)
2. Lehnert, S.: A taxonomy for software change impact analysis. In *Proceedings of the 12th International Workshop on Principles of Software Evolution and the 7th annual ERCIM Workshop on Software Evolution, IWPSE-EVOL'11, ACM*, pp. 41–50, New York, NY, USA, (2011)
3. Visual paradigm: <http://www.visual-paradigm.com/>. [Online; accessed 03-June- 2013]
4. Antonioli, G., Canfora, G., Casazza, G., De Lucia, A., Merlo, E.: Recovering traceability links between code and documentation. *Soft. Eng. IEEE Trans.* **28**(10), 970–983 (2002)
5. Corley, C.S., Kraft, N.A., Etkorn, L.H., Lukins, S.K.: Recovering traceability links between source code and fixed bugs via patch analysis. In *Proceedings of the 6th International Workshop on Traceability in Emerging Forms of Software Engineering*, pp. 31–37. ACM, (2011)
6. Bohner, S.A.: A graph traceability approach for software change impact analysis. Ph.D. Thesis, Fairfax, VA, USA, UMI Order No. GAX95-42995 (1995)
7. Khalil, A., Dingel, J.: Supporting the evolution of UML models in model driven software developmeny: A Survey, Technical Report, School of Computing. Queens University, Canada (2013)
8. Breivold, H.P., Crnkovic, I., Larsson, M.: A systematic review of software architecture evolution research. *Inf. Softw. Technol.* **54**(1), 16–40 (2012)
9. Tip, F.: A Survey of Program Slicing Techniques. Technical Report. CWI, Amsterdam, The Netherlands (1994)
10. Göknil, A., Kurtev, I., van den Berg, K.G.: Change impact analysis based on formalization of trace relations for requirements. In: *ECMDA Traceability Workshop (ECMDA-TW)*, pp. 59–75. Berlin, Germany (2008)
11. Fowler, M.: *UML Distilled: A Brief Guide to the Standard Object Modeling Language*, 3rd edn. Addison-Wesley Longman Publishing Co., Inc, Boston (2003)
12. Briand, L.C., Labiche, Y., O’Sullivan, L.: Impact analysis and change management of uml models. In *Proceedings of the International Conference on Software Maintenance, ICSM'03, IEEE Computer Society*, pp. 256–, Washington, DC, USA, (2003)
13. Li, Y., Li, J., Yang, Y., Li, M.: Requirement-centric traceability for change impact analysis: a case study. In: Wang, Q., Pfahl, D., Raffo, D. (eds.) *Making Globally Distributed Software Development a Success Story. Lecture Notes in Computer Science*, vol. 5007, pp. 100–111. Springer, Berlin Heidelberg (2008)
14. Marcus, Sergeyev, A., Rajlich, V., Maletic, J.I.: An information retrieval approach to concept location in source code. In: *Reverse Engineering, 2004. Proceedings. 11th Working Conference on*, pp. 214–223. IEEE, (2004)

Chapter 99

A Review of Image Segmentation Methodologies in Medical Image

Lay Khoon Lee, Siau Chuin Liew and Weng Jie Thong

Abstract A precise segmentation of medical image is an important stage in contouring throughout radiotherapy preparation. Medical images are mostly used as radiographic techniques in diagnosis, clinical studies and treatment planning. This review paper defines the limitation and strength of each methods currently existing for the segmentation of medical images.

Keywords Advantages · Image segmentation · Medical image

99.1 Introduction

In medical fields nowadays, medical imaging is a crucial component in a many applications. Such applications take place throughout the clinical track of events; not only within diagnostic settings, but prominently in the area of preparation, carrying out and evaluation before surgical operations.

Generally, image segmentation is the procedure of separating an image into several parts. Instead of considering the whole data presented in an image all at once, it is better to focus on a certain region-based semantic object in image segmentation. Image segmentation has been widely implemented in medical imaging to separate homogeneous area. Studied proposed by Liew et al. in 2012 shows that region of interest (ROI) segmentation plays a crucial role in multilevel authentication [1]. Thus the goal of image segmentation is to find the regions that

L.K. Lee (✉) · S.C. Liew · W.J. Thong
Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang,
Lebuhraya Tun Razak, 26300 Gambang Kuantan, Pahang Darul Makmur, Malaysia
e-mail: jesscefyn@hotmail.com

S.C. Liew
e-mail: liewsc@ump.edu.my

W.J. Thong
e-mail: briantwj@gmail.com

represent meaningful parts of objects for easier analyzation purpose .In this review paper, author aims to gather and analyze methods used in image segmentation. So, in general this paper will summarize suitable image segmentation methods to be used for each types of medical images scan.

99.2 Method

In this section, several techniques that are being widely used on medical image segmentation had been briefly described by the author. Segmentations are divided mainly in four different techniques, which are thresholding-based, region-based, edge-based, and clustering-based. Additionally there are also other methods for image segmentations. The Fig. 99.1 below illustrates the types of image segmentation available.

99.2.1 Thresholding-Based

99.2.1.1 Gray-Level Thresholding

The work of Beveridge [2] and his friends offered a decent example of a procedure that integrates gray level thresholding which is a technique under the thresholding-based image segmentation. In their paper, an input image can either be a grayscale or colored image. The image is then divided into sectors of fixed size and fixed location. An intensity histogram is calculated for each sector (and on colour images, for each colour channel), and used to produce a local segmentation. For every sector, information from its neighbors is used to detect clusters for which there may not be enough local support due to the artificially induced partition of the image.

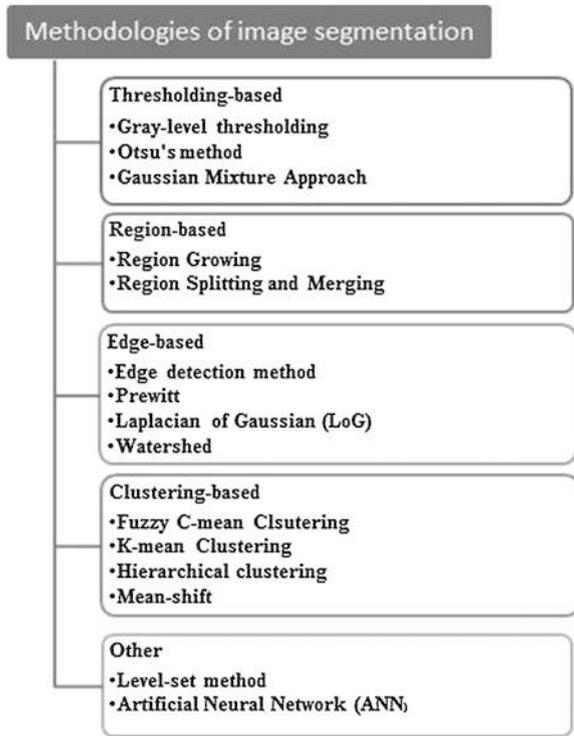
99.2.1.2 Ostu's Method

In 1979, Ostu's [3] research shows that thresholding method is based on a very simple idea, which is to figure out the threshold that minimizes the weighted within-class variance. In the experiment, Ostu's method resulted in an advance in threshold-based technique by converting a gray scale image into a binary image by automatically.

99.2.1.3 Gaussian Mixture Method

Gaussian mixture approach for image segmentation is a method to estimate the number of components with their means and covariance sequentially without requiring any initialization as proposed by Nicole and Alexander in 2012 [4]. The experiment

Fig. 99.1 Methodologies of image segmentation



procedure starts from a single mixture component covering the whole data set and sequentially splits incrementally during the expectation maximization steps. This Gaussian mixture method successfully shows its effectiveness after many experiments.

99.2.2 Region-Based

99.2.2.1 Region Growing

In 2001, Cootes and Taylor proved that region growing method can be accomplished by adding an offset and by scaling to transform all gray values to zero-mean and unit-variance. As an alternative to work with intensities, Cootes and Taylor suggested a method which is storing the gradients' direction and strength, the latter is mapped non-linearly, this method proved to be better than the normalized intensity. In 2002, Boasch et.al described a non-linear procedure in his paper. The latter method proved especially useful in images with a strongly non-Gaussian distribution as encountered, e.g. in ultrasound images. Studies continued by Scott et al. in 2003, used gradient orientation, corner and edge strength for the detection of vertebrae in dual energy X-ray images [5–7].

99.2.2.2 Region Splitting and Merging

Region splitting and merging technique is well known in region-based approach. This technique is a combination of region splitting and merging. Manousakes et al. in 1998 had [8] applied splitting and merging technique in trying to overcome the difficulties occurred when using homogeneity measures. This experiment successfully verified that the split-and-merge methods can be implemented in 2D and 3D MRI with a directionally uniform performance. Improvements involving simulated annealing and boundary elimination can also be applied effectively in 3D or 2D MRI. The execution process is less time consuming.

99.2.3 Edge-Based

99.2.3.1 Edge Detection

In general edge detection methods are the process of identifying and locating sharp discontinuities in an image. Edge detection is important for the object recognition of human organs in medical images. In the years of 2006, Y.Q Zhang et al. [9] introduced basic mathematical morphological theory and operations, the novel mathematical morphological edge detection is proposed to detect the edge of lungs CT image with salt -and -pepper noise. The experimental result shows that the method proposed is more efficient for both medical image de-noising and edge detection than the existing edge detection method.

99.2.3.2 Prewitt Edge Detection

According to Prewitt and J.M, 1970 [10], the Prewitt edge detector is an appropriate way to estimate the magnitude and orientation of an edge. Although differential gradient edge detection needs a rather time consuming calculation to estimate the orientation from the magnitudes in the x and y-directions, the compass edge detection obtains the orientation directly from the kernel with the maximum response. The Prewitt operator is limited to eight possible orientations; however experiments show that most direct orientation estimations are not that accurate. This gradient based edge detector is estimated in the 3×3 neighbourhood for eight directions. All the eight convolution masks are calculated. One convolution mask is then selected, namely that with the largest module. In this experiment, Prewitt detector is able to clearly define the edges.

99.2.3.3 Laplacian of Gaussian

Laplacian of Gaussian was firstly introduced by Marr and Hildreth in 1980 who combined Gaussian filtering with the Laplacian technique. This algorithm is not often used in machine vision. Researchers that continued this method were Berzins in 1984, Shah, Sood and Jain in 1986, Huertas and Medioni in 1986 [11, 12].

99.2.3.4 Watershed

Watershed is an image segmentation method introduced by S. Beucher and F. Meyer in 1990 which separates overlapping objects [13]. They proposed using mathematical morphology in image segmentation. The objective of the paper is to avoid the problem of oversegmentation. This method involved two tools which are watershed transform and the homotopy modification. The number of experiments and the positive results were attached in the author's published paper.

99.2.4 Clustering-Based

99.2.4.1 Fuzzy C-Mean Clustering Method

The fuzzy set theory was introduced by Zadeh, and successfully applied in image segmentation. The fuzzy c-means algorithm was proposed by Bezdek in 1981 based on the fuzzy theory, it is the most widely studied and used algorithm in image segmentation for its simplicity and the ability to obtain more information from images. The studies was extended by many researchers, one of the latest researchers' studies was done by Krinidis and Chatzis in 2010. They proposed a fuzzy local information c-means (FLICM) to overcome the problem of setting parameters in the FCM-based methods. This algorithm uses both spatial and gray-level local information, and is fully free of parameter adjustment (except for the number of clusters) [14–16].

99.2.4.2 K-Means Clustering

According to Kaus et al. in 2004, [17] K-means Clustering Method is a process to classify a given data set through a certain numbers of K-clusters. After clustering all features into n classes with the k-means algorithm, each landmark is assigned to the cluster containing the largest number of training features from that point. In the research, the features sampled at a certain landmark only need to be compared with the assigned cluster. Different appearance models for different boundary segments are thus supported automatically.

99.2.4.3 Hierarchical Clustering

Hierarchical clustering is one of the edge based method developed by D. Cordes et al. in 2002. The author had done a research by using hierarchical clustering to measure connectivity in functional MRI. This method is able to detect similarities of low-frequency fluctuations, and the results indicated that the patterns of functional connectivity can be obtained with hierarchical clustering that resembles known neuronal connections [18].

99.2.4.4 Mean-Shift

In the year of 2002, Comaniciu and Meer described a segmentation method based on the mean-shift algorithm. The mean-shift algorithm by Cheng in 1995 is designed to locate points of locally-maximal density in feature space. Feature vectors containing gray-scale or colour information as well as pixel coordinates are computed for each pixel [19, 20].

99.2.5 Other Method

99.2.5.1 Level-Set Method

Level-set method were presented by Osher and Sethian in 1988 then extended by Malladi et al. in 1995. This paper features an implied shape representation and can be employed with regional or edge-based features. It is slightly different with the research by Leventon et al. whereby they extended the original energy formulation by an additional term which deforms the contour towards a previously learned shape model. A frequent criticism is that the signed distance maps which the shape model is based on, do not form a linear space, which can lead to invalid shapes if training samples vary too much. In 2006, Pohl et al. presented a method of embedding the signed distance maps into the linear Log Odds space, which could solve the modeling problems. To keep this review at a reasonable length, they had to ignore level-set theory and techniques: The conceptual differences between the implicit representation and the discrete models they intend to focus on would have required a special treatment for all following sections [21–23].

99.2.5.2 Artificial Neural Network

In 2003, Pal and Pal [24] had done a research on ANNs method used for segmentation. As what Pal and Pal have predicted, ANN becomes widely applied in image processing. Lately the research of this method has been continued by Indira, S.U., and Ramesh in the years of 2011 [25]. In this paper, an unsupervised method

Kohonen's Self-Organizing Maps (SOM) form ANN and (Genetic Algorithm) GA were combined to identify the main featured present in the image. SOM combined with GA and some of the variants of SOM like the Variable Structure SOM (VSSOM), Parameterless SOM (PLSOM) are compared and their performance is evaluated. A new unsupervised, nonparametric method is developed by combining the advantages of VSSOM and PLSOM. The experiments performed on the satellite image shows that the modified PLSOM is efficient and the time taken for the segmentation is less when compared to the other methods.

99.3 Modality

The imaging modalities employed can be divided into two global categories: anatomical and functional. This paper only covered anatomical modalities. Images are presented in 2D as well as in 3D domain. In the 2D domain each element is called pixel, while in 3D domain it is called voxel [26]. Medical imaging is performed in various modalities and in this paper the author will discuss images such as CT, MRI, X-Ray and ultrasound.

CT (Computed tomography) is the image of sectional planes (tomography) which are harder to interpret. CT can visualize small density difference, e.g. grey matter, white matter and CSF. CT can detect and diagnose disease that cannot be seen with X-ray. However it is more expensive than X-ray, lower resolution and lower ionizing radiation [27].

Magnetic resonance imaging (MRI) is a sophisticated medical imaging technique that uses magnetic fields and radiofrequency to visualize the body's internal structures. Magnetic field gradients cause signals from different parts of the body to have different frequencies. Signals collected with multiple gradients are processed by computer to produce an image [27]. MR imaging of the body is performed to get the structural details of brain, liver, chest, abdomen and pelvis which helps in diagnosis or treatment.

In ultrasound imaging, computerized images are produced by sound waves reflected by organs and other interior body parts in real time. It can be used for interventional procedures. Ultrasound has no known harmful effects (at levels used in clinical imaging). Ultrasound equipment is inexpensive, but many anatomical regions cannot be visualized with ultrasound, for instance, head [28].

X-ray is the oldest non-invasive imaging of internal structures. It is rapid, short exposure time and cheaper than other modalities. However X-ray is unable to distinguish between soft tissues in head and abdomen. Real time X-ray imaging is possible and used during interventional procedures. X-Ray contains ionizing radiation which is a cause of cancer [29].

99.4 Discussion

Table 99.1 below summarizes the advantages and limitation of each methodology.

99.5 Analysis

In this section, Table 99.2 is presented as recommendations of methodologies for CT scan, ultrasound, MRI and X-ray. Analysis is done according to the advantages and disadvantages of each methods has mentioned above.

Segmentation of CT scan usually involves three main image related problems; noise that can alter the intensity of a pixel such that its classification becomes uncertain, intensity inhomogeneity where the intensity level of single tissue class varies gradually over the extent of the image, and image that have finite pixel size and are subject to partial volume averaging where individual pixel volumes contain a mixture of tissue classes so that the intensity pixel in the image may not be consistent with any one class. For the segmentation of bone via CT scan, it is recommended to use thresholding-method and region-based, or a combination of region growing and watershed methodologies. By using these two methods, image renderings are often used to provide detailed visualization of skeletal structure.

The main goal in brain MR segmentation is to segment gray matter, white matter and cerebrospinal fluid. Segmentation is also used to find out the regions corresponding to lesions tumors, cyst, edema, and other pathologies. Thus the recommended method is clustering-based method. This is because this method is based on the intensity of information, and a major concern is the presence of intensity inhomogeneity.

For 3D MRI, it is slightly different with the 2D MRI, it is recommended to use ANN or level-set method, due to complexity of dimensionality.

Thresholding-based method has been performed mostly in ultrasound for extracting a variety of structures. The thresholding of intensity and texture statistics is used to segment ovarian cysts which will result in a good output, since is simple to use.

Edge-based and watershed method is recommended for X-ray segmentation. The edge-based methods works well for noise-free image, but the performance degrades with noisy images or when fake and weak edges are present in the X-ray imaging. Better result could be obtained when the edge-based method is combined with watershed method.

Table 99.1 Advantages and limitation of methodologies

Methodologies		Advantages	Disadvantages
Thresholding-based	Gray level thresholding-based	Easy to implement and efficient. No need prior information or calculation	Noisy and blurred edges. Lack of sensitivity and sharpness, hard to define, complex for multidimensional
	Otsu's method	Functions well and stably. General: no specific histogram shape assumed. Able to extend to multilevel thresholding	False maxima may occur during optimization function. The method tends to artificially enlarge small classes to obtain 'better separation'; small classes might be merged and missed
	Gaussian mixture approach	Relatively general histogram model. When model is invalid, minimizes classification error probability. Applicable to small-size classes	Many histograms are not Gaussian. Intensities are finite and non-negative. Difficult to detect close and flat modes. Requires significant simplification of model form extension to multithresholding
Region-based	Region growing	Perform better in noise image. Easy to compute	Seed point must be specified. Costly. Images may be undersegmentated or oversegmentated
	Region merging and splitting	The image can be split progressively according to demanded resolution. Can split the image by mean or variance of segment pixel value. The merging criteria can be different to the splitting criteria	May produce blocky segments
Edge-based	Edge detection	Enclose large areas. Applicable to images with uneven illumination	Only applicable to simple background. closed contours are not guaranteed
	Prewitt edge detection	Simplicity and able to detect edges and their orientations	Sensitivity to noise and inaccurate
	Laplacian of Gaussian	Able to find the correct places of edges. Can be used to test wider area around the pixel	Malfunctioning at corners, curves and where the gray level intensity function varies. Not finding the orientation of edge because of using the Laplcian filter
	Watershed	Perform better in noise image. Fast speed and output reliable	Seed point must be specified and easily over-segmentated. Only applied on gradient and time consuming

(continued)

Table 99.1 (continued)

Methodologies		Advantages	Disadvantages
Clustering-based	Fuzzy C-mean clustering	Simple and easy to understand	Undefined optimal solution. Sensitive initialization process. Not compatible with noisy data
	Hierarchical clustering	Simple, output reliable. The process and relationships can check the dendrogram. Only need to compute the distances between each pattern, instead of calculating the centroid of clusters	Involves in detailed level, the fatal problem is computation time. For the reason that hierarchical clustering involves in detailed level, the fatal problem is the computation time
	K-mean clustering	Fast number of clustering is fixed. K-means algorithm is easy to implement. It is faster than the hierarchical clustering	User has to select the number of desired output clusters before starting to classify data. The result is sensitive to the selection of the initial random centroids. Cannot show the clustering details
	Mean shift	An extremely versatile tool for feature space analysis. Suitable for arbitrary feature spaces	The kernel bandwidth is the only factor that can control the output. The computation time is quite long
Other method	Level set method	Versatile, robust, accurate, and efficient	Require considerable thought in order to construct appropriate velocities for advancing the level set function
	Neural network based	Applicable to a variety of problems. Easy to implement	Lack of profound theoretical basis for ANNs. Problem in choosing the best architecture. Black-box problem

Table 99.2 Recommended methodologies

Modalities	Recommended method
CT Scan	Thresholding and region-based
	Region-growing and watershed
3D MRI	ANN
	Level set method
MRI	Clustering-based
Ultrasound	Thresholding-based
X-ray	Edge-based and watershed

99.6 Conclusion

In conclusion, this paper surveys on the existing methods of image segmentation and recommends image segmentation methods to be applied on CT scan, 3D MRI, MRI, ultrasound and X-Ray. The scope only includes digital watermarking in medical image. Each image segmentation method has its own limitations. Hence this paper can be used as a reference. However, there are certain limitations for image segmentation in digital watermarking for medical images. The accuracy of the segmentation remains the most concerned issue in determining critical cases such as detection of tumor via medical imaging. Recommendation for future works includes improving the accuracy and the speed of image segmentation for the digital watermarking in medical imaging.

References

1. Liew, S.C., Liew, S.W., Zain, J.M.: Tamper localization and lossless recovery watermarking scheme with ROI segmentation and multilevel authentication. *J. Digit. Imaging* **26**(2), 316–325 (Springer) (2012). DOI: [10.1007/s10278-012-9484-4](https://doi.org/10.1007/s10278-012-9484-4)
2. Beveridge, J.R., Griffith, J., Kohler, R.R., Hanson, A.R., Rise-man, E.M.: Segmenting images using localized histograms and region merging. *Int. J. Comput. Vision* **2**(3), 311–347 (1989)
3. Ostu, N.: A threshold selection method from gray-level histogram. *IEEE Trans. Syst. Man Cybern.* **SMC-8**, 62–66 (1978)
4. Nicola, Greggio, et al.: Fast estimation of Gaussian mixture models for image segmentation. *Mach. Vis. Appl.* **23**(4), 773–789 (2012)
5. Cootes, T.F., Taylor, C.J.: On representing edge structure for model matching In: Proceedings of the IEEE CVPR, vol. 1, (2001)
6. Bosch, J., Mitchell, S., Lelieveldt, B., Nijland, F., Kamp, O., Sonka, M., Reiber, J: Automatic segmentation of echocardiographic sequences by active appearance motion models. *IEEE Trans. Med. Imaging (Cootes, T.F, Taylor, ging)* **21**(11), 1374–1383 (2002)
7. Scott, I.M., Cootes, T.F., Taylor, C.J.: Improving appearance model matching using local image structure. In: Proceedings of the IPMI.LNCS, vol. 2732, Springer, Heidelberg (2003)
8. Manousakkas, I.N., Undrill, P.E., Cameron, G.G., Redpath, T.W.: Department of Biomedical Physics and Bioengineering, University of Aberdeen, Foresterhill, Aberdeen, AB25 2ZD, Scotland, United Kingdom Received December 30, (1996)
9. Yu-qian, Z., Wei-hua, G., Zhen-cheng, C., Jing-tian, T., Ling-Yun, L.: Medical images edge detection based on mathematical morphology. In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE, pp. 6492–6495. (2006)
10. Prewitt, J.M.: Object Enhancement and Extraction, vol. 75. Academic Press, New York (1970)
11. Albvik : Handbook of Image and Video Processing. Academic Press, New York (2000)
12. Haralick, R.M., Shapiro, L.G.: Computer and Robot Vision. vol. 1, Addition-Wesley Publishing Company Inc., Boston (1992)
13. Beucher, S., Meyer, F.: The morphological approach to segmentation: The watershed transform. In: Dougherty, E.R. (ed.) *Mathematical Morphology in Image Processing*, vol. 12, pp. 433–481. Marcel Dekker, New York (1993)
14. Zadeh, L.: Fuzzy sets. *Inform. Control* **8**, 338–353 (1965)

15. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell (1981)
16. Krinidis, S., Chatzis, V.: A robust fuzzy local information c-means clustering algorithm. *IEEE Trans. Image Process.* **19**, 1328–1337 (2010)
17. Kaus, M.R., von Berg, J., Weese, J., Niessen, W., Pekar, V.: Automated segmentation of the left ventricle in cardiac MRI. *Med. Image Anal.* **8**(3), 245–254 (2004)
18. Cordes, D., Haughton, V., Carew, J.D., Arfanakis, K., Maravilla, K.: Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magn. Reson. Imaging* **20**(4), 305–317 (2002)
19. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(5), 603–619 (2002)
20. Cheng, Y.: Mean shift, mode seeking and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(8), 790–799 (1995)
21. Malladi, R., Sethian, J., Vemuri, B.: Shape modeling with front propagation: a level set approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(2), 158–174 (1995)
22. Leventon, M.E., Grimson, W.E.L., Faugeras, O.: Statistical shape influence in geodesic active contours. In: *Proceedings of the IEEE CVPR*, vol. 1, (2000)
23. Pohl, K.M., Fisher, J., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M.: Logarithm odds maps for shape representation. In: *Proceedings of the MICCAI LNCS*, vol. 4191, Springer, Heidelberg (2006)
24. Pal, N.R., Pal, S.K.: A review on image segmentation techniques. *Pattern Recogn.* **26**(9), 1277–1294 (1993)
25. Indira, S.U., Ramesh, A.C.: Image segmentation using artificial neural network and genetic algorithm: a comparative analysis. In: *Process Automation, Control and Computing (PACC)*, 2011 International Conference on IEEE, pp. 1–6. (2011)
26. Subach, B.R., et al.: Reliability and accuracy of fine-cut computed tomography scans to determine the status of anterior interbody fusions with metallic cages
27. McRobbie, D.W.: *MRI from Picture to Proton*. Cambridge University Press, Cambridge (2007)
28. Semelka, R.C., Armao, D.M., Elias, J., Huda, W.: Imaging strategies to reduce the risk of radiation in CT studies, including selective substitution with MRI. *J. Magn. Reson. Imaging* **25**(5), 900–909 (2007)
29. Spiegel, P.K.: The first clinical X-ray made in America—100 years. *Am. J. Roentgenol.* (Leesburg, VA: American Roentgen Ray Society) **164**(1), 241–242 (1995)

Chapter 100

The Utilization of Template Matching Method for License Plate Recognition: A Case Study in Malaysia

Norazira A. Jalil, A.S.H. Basari, Sazilah Salam, Nuzulha Khilwani
Ibrahim and Mohd Adili Norasikin

Abstract Automatic License plate detection and recognition system is special form of optical character recognition and has been an active research domain in image processing field. However, the accuracy is varied due to different styles of number plates endorsed. Besides, the characters on Malaysia license plate are in one or two lines. Thus, the proposed license plate recognition (LPR) technique of this research is able to achieve the best recognition performance based on Malaysia license plate vehicle registration number characters. This paper presents a study of applying the template matching method for character image recognition. The database of characters and license plate image has been created by collecting images from various type of car. The initial pre-processing involves image enhancement, binarization, filtering and segmenting of license plate. There are 100 license plates that contain 693 characters have been tested, and the result shown that 92.78 % of all characters is correctly recognized. Thus, template matching can be classified as one of the promising algorithm for recognizing Malaysia license plate.

Keywords Malaysia license plate · Image processing · Template matching

N.A. Jalil (✉) · A.S.H. Basari · S. Salam · N.K. Ibrahim · M.A. Norasikin
Center for Advanced Computing Technology (C-ACT), Faculty of Information and
Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal,
Melaka, Malaysia
e-mail: norazira90@gmail.com

A.S.H. Basari
e-mail: abdsamad@utem.edu.my

S. Salam
e-mail: saizilah@utem.edu.my

N.K. Ibrahim
e-mail: nuzulha@utem.edu.my

M.A. Norasikin
e-mail: adili@utem.edu.my

100.1 Introduction

The advanced of computer application processed more than textual data solving everyday problems. Image processing techniques has been applied in many real-life applications with great societal value. It has been applied in numerous domain applications such as medical, security, engineering, forensic and biometric purposes. The ability to process image or picture and translate it into something meaningful has made the image processing fields become an active research area until today. A picture is worth a thousand words. Thus, with the great demands in intelligent transportation system, the technology of image processing has adopted for vehicle and traffic surveillance [1], managing vehicle parking system and highway electronic toll collection. The fundamental issues in real-time license plate recognition are the accuracy and the recognition speed in different environmental conditions.

A license plate is the unique identification of a vehicle. License plate recognition system is a complex image processing application which recognizes the characters on a car plate based on the given conditions and situation. The license plate recognition system is installed in many places with multiple purposes and even polices are using this application to detect speeding vehicles and monitoring surveillance from distance away. These systems are beneficial because it can automate car park management [2], improve the security of car park operator and the users as well. Moreover, this system can eliminate the usage of swipe cards and parking tickets that lead to green technologies [3], improve traffic flow during peak hours, detect speeding cars on highways, and detect cars which run over red traffic lights. License plate recognition system has been intensively studied worldwide; however, in Malaysia cases, the implementation of these systems are less accurate due to the different styles of car plates applied. The main reason is the developments of image processing applications are still inadequate where it is unable to reach the 100 % accuracy in recognition. Due to that, it is recommended that research is still conducted for this application because of the importance of car plate recognition.

The Road Transport Department of Malaysia has endorsed a specification for car plates that are allowed for use includes the font and size of characters that must be followed by car owners. In Malaysia, the standard number plate has black color for the background and white color for the font. All number plate is in the form of vertical and horizontal with some specified font and size. Figure 100.1 below shows samples of common and special Malaysian car plates. These various fonts and sizes of characters will lead to problems during recognition phase.

One of the factors that contribute to the failure in achieving 100 % accuracy in recognition was unable to recognize similar pattern characters. The common confused characters such as in the case of recognizing character 'B' and '8' or '3', '0' and 'D', 'E' and 'F', 'G' and '6' or 'Q', 'A' as '4', '6' with 'S' and '7' with 'T' or '1'. However, to avoid confusion between the letters and numbers, certain letters are not used in the Malaysian registration system. The alphabets 'I' and 'O' are omitted from the alphabetical sequences due to their similarities with the numbers



Fig. 100.1 Vehicle number plate specification approved by Jabatan Pengangkutan Jalan Malaysia (JPJ)

‘1’ and ‘0’. The alphabet ‘Z’ is omitted and reserved for use on Malaysian military vehicles. Besides, there can be no leading zeroes in the number sequence.

Therefore, the experiments are conducted by using template matching to recognize the characters of license plate. Both vertical and horizontal Malaysia license plate is applied as testing images.

The rest of the paper is organized as follows: the entire recognition processes are briefly described in Sect. 100.2. Testing and results were discussed in Sect. 100.3 and finally concluding remarks were given in Sect. 100.4.

100.2 Materials and Methods

Based on the previous studies [4, 5], the overview of image processing technique in license plate recognition system includes 5 stages: (a) pre-processing, (b) filtering, (c) feature extraction, (d) segmentation and (e) character recognition. The final output of the sample experiment is to recognize the alphanumeric characters on the license plate.

100.2.1 Preprocessing

Digital image preprocessing is an initial step to image processing improving the data image quality for more suitable for visual perception or computational processing. Preprocessing is an important step of applying a number of procedures for smoothing and enhancing the image [6], removing background noise, normalizing the intensity of individual image particles, image blurring and remove image

reflections. This phase is used for making a digital image usable by subsequent algorithm in order to improve their readability for the next phases. Preprocessing for car license plate number uses three common sub processes [1], which are geometric operation, gray scaling process and binarization process.

100.2.1.1 Geometric Operation

Geometric operation is a process to locate the car license plate. The purpose of this operation is to localize the car plate for faster character identification over a small region. The input image is first cropped manually to minimize the processing time. Only the lower part of the input image that contains the required information will be processed [1, 5, 7].

100.2.1.2 Grayscale Process

Grayscale is a process to produce shades of gray image from a multi-colour image. Grayscale is usually the preferred format for image processing. Often, the grayscale intensity is stored as an 8-bit integer giving 256 possible different shades of gray from black to white. Basically, grayscale images are entirely sufficient for many tasks and so there is no need to use more complicated and harder-to-process colour images.

The method for the conversion of RGB to grayscale, as follows [8, 9]:

$$\text{Intensity value} = 0.299 * R + 0.587 * G + 0.114 * B \quad (100.1)$$

100.2.1.3 Binarization Process

Binarization is a process of converting grayscale image into black and white image or “0” and “1”. In digital image processing, binarization process enable of removing unwanted information, detect the information of target area and increase the processing speed [10]. Previously, the grayscale image consists of different level of gray values; from 0 to 255. To improve the quality and extract some information from the image, the image needs to be process a few times and thus make the binary image more useful. Gray threshold value of an image is required in the binarization process as it is important to determine whether the pixels that having gray values will be converted to black or white.

The purpose of binarization is to identify the extent of objects and to concentrate on the shape analysis, where the shape of a region is more significant than the intensity of pixels [6]. To avoid from data loss during the binarization process, as both desired (characters) and undesired image (noise) will appeared; a variable thresholding technique [11] is applied. This technique helps in determines the local optimal threshold value for each image pixel in order to avoid problem originating from non-uniform illumination [12].

100.2.2 Filtering

In this stage, filtering process is used for blurring and for noise reduction while preserving the sharpness of the image. Blurring is used in filtering process to remove and clean-up the small noise and boundaries. In order to reduce the background noise, softening is applied in low pass filter and thus making the image blur.

A low-pass filter [13, 14] is also known as blurring or smoothing filter. The method is used by calculating the average of a pixel and all of its eight immediate neighbours. The result will replace the original value of the pixel. The process is repeated for every pixel in the image.

100.2.3 Feature Extraction

In order to obtain an accurate location of the license plate, localization and extraction of needed area from the vehicle image is necessary. This phase will give a greatly affects for recognition rate and overall speed of the whole process. After removing most of the unwanted object in an image, the next step is to search for regions with high values which are most likely contain a license plate [15]. This is done by constructing a horizontal and vertical projection profile of the edge image. Due to the presence of some characters and text, the region containing the number plate has a higher concentration of edges than the surrounding regions [16]. By combining both horizontal and vertical projection [17], the location of the number plate within the region of interest can be detected. The number of white pixels presence in each row and column is counted. The vertical boundaries of the number plate manifest as peaks in the graph of the vertical projection. The horizontal boundaries of the number plate manifest as troughs in the graph of the horizontal projection indicate the regions has strong edged and can be used for horizontal position of license plate.

100.2.4 Image Segmentation

Segmentation is an important and crucial [15, 18] stage in License Plate Recognition system that influences the accuracy of the whole system. The goal of character segmentation is to find and segment the isolated characters on the plates, without losing features of the characters. The correct segmentation process will ease the process of character recognition. There are many factors that make the character segmentation task difficult, such as image noise, plate frame, rivet, and rotation and illumination variance. Thus, preprocessing is very important for ensuring the good performance of character segmentation.

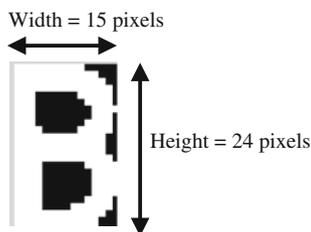


Fig. 100.2 Size of segmented character

For Connected Component Analysis [19], the threshold of the plate image is calculated and search for the connected components in the image, each connected component will be assigned a special label in order to distinguish between different connected components. Then, the character is resized to the standard height and width in order to be used for recognition system as illustrated in Fig. 100.2.

100.2.5 Character Recognition

Character recognition is the most important task in recognizing the plate number [6]. The recognition of characters has been a problem that has received much attention in the fields of image processing, pattern recognition and artificial intelligence. It is because there is a lot of possibility that the character produced from the normalization step differ from the database. The same characters may differ in sizes, shape and style [20] that could result in recognition of false character, and affect the accuracy and complexity of the whole system. In Malaysian car plate, there are two groups of character, which is alphabet and numeric. It is important for the system to differentiate the character correctly as sometimes the system may confuse due to the similarities in the form of shape.

100.2.5.1 Template Matching

Template matching [1, 21, 22] is one of the most and common and easy classification methods for recognizing the segmented characters [23]. This technique is applied by finding the small part of image that match with the template. The size of each character image is normalized according to the template stored in the database. Each character is compared to its corresponding pixel in the template and the highest coefficient result is identified as the character of the input [24]. This method has shown high accuracy but requires efficient searching method and needs a large storage to save all the numbers and character templates [25].

However, due to processing time, the small size of template is commonly used and thus may lead to inaccurate detection. Besides, slight deviations in shape, size,

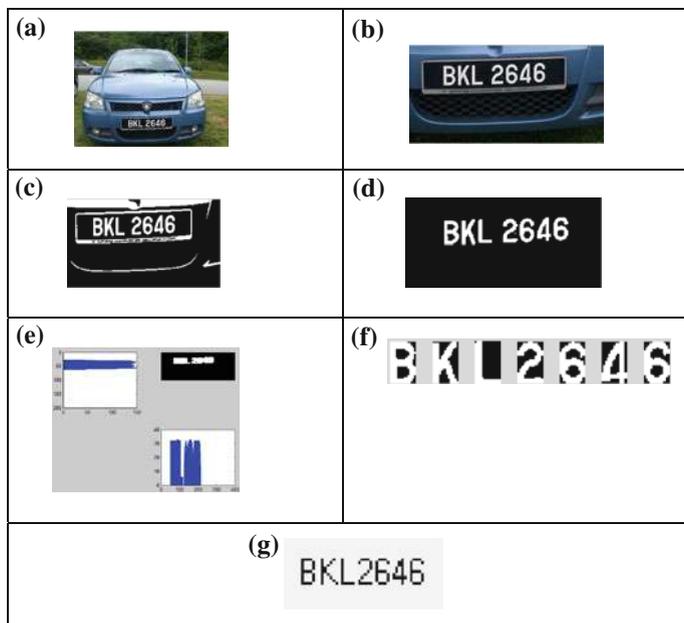


Fig. 100.3 Proposed technique, **a** Input image, **b** Cropped input image, **c** Binary image, **d** Filtered number plate, **e** Extracted number plate, **f** Segmented character, **g** Output from template matching

and orientation, would prevent template matches from reading even the limited 0–9 numbers and A–Z English alphabets.

The recognition process using template matching is grouped into two main sub-modules which are template creation and template matching. Template database contain the characters that may be present in a license plate are from A to Y and the digits that may also be present in a license plate are from 0 to 9. There are four set of 33 alphanumeric characters (23 alphabets and 10 digits) is used for template creation. The alphabets ‘I’, ‘O’ and ‘Z’ are omitted and reserved due to their similarities with few digits and for Malaysian military vehicles purposes only. The number of templates can be increased in order to increase the accuracy of the license plate recognition system. Fitting approach is necessary for template matching. For matching the characters with the database, input images must be equal-sized with the database characters. The matching is done on a pixel by pixel basis. In order to achieve optimum recognition rate, the template characters fit to 15 × 24 pixels is prepared. The segmented number plate region must be at the same size of the template, although the real number plate region is bigger or smaller than the template size [1]. The extracted characters segmented from plate and the characters on database are now equal-sized [26].

The next step is template matching. Template matching is an effective algorithm for recognition of characters. The character image is compared with the ones

in the database and the highest correlation value is selected. Correlation is an effective technique for image recognition [27]. This method measure the degree of similarity corresponding pixel values of two images, template and input image. An image with the highest correlation value means strong relation between input image and templates that will produces the best match.

This type of comparison does not involve much computation. It is restricted to the number of samples to be compared and the number of templates available in the database, proportionate to which the duration of time taken to compare and throw up a result would vary. Another limitation of the method is that when exact matches are not found the probability of misinterpreted increases. In this method the computational time increases as the number of sample increases [28]. Figure 100.3 shows the detail process flow.

100.3 Results and Discussion

In this experiment, 693 characters from 100 license plate images were employed and the size of each segmented images has been resized to 15×24 pixels. For these images, all of them were taken by digital camera from various scenes and under different lighting conditions of the real world, varied distances from the vehicle and weather. The distance between the camera and the vehicle varied from 3 to 7 m. The success rate has been achieved where the accuracy of license plate recognition using template matching yield up to 92.78 %.

From 693 characters tested, 50 have been wrongly recognized. Out of 33 characters, there are 11 characters that are wrongly recognized which are B, G , M, Q, U, 0, 3, 5, 6, 8 and 9. Out of 50 errors, number 8 and 9 has the highest misrecognized characters where system tends to recognize as 6 and 0 respectively. Other common misrecognized characters are M and 0. The failure of this section is caused by some defectives on input image that lead to failure. Based on the result, the recognition accuracy and efficiency of the system can be increases with more number of training samples.

The highest degree of similarity between each character is 1. Thus, based on the experiment, even though the degree of similarity closes to 1, yet there are errors detected. The highest degree of misrecognizes character is 0.8157 while the lowest is 0.4005.

100.4 Conclusion

In this paper, a simple yet efficient method of detecting and recognizing characters image in Malaysia license plate has been presented. The proposed method utilizes template matching to perform the recognition. Firstly, pre-processing and filtering is conducted to remove the unwanted images. Next, extraction is used to locate the

position of license plate characters, and then segmentation is done to separate the plate characters individually. Finally, template matching with the use of correlation is applied for recognition of plate characters. The performance of template matching with Malaysia license plate has been tested and evaluated. The result has shown that the developed algorithm works well and has gained the high accuracy value up to 92.78 %.

As future work, it will be interesting to test the method with more number of English alphabet and numeric samples. Another suggestion is to improve the performance of template matching by hybridizing with back-propagation neural network.

Acknowledgments This work is part of Master by research in Information and Communication Technology supported by MyBrain Scheme of Universiti Teknikal Malaysia Melaka and the Malaysian Technical University Network—Centre of Excellence (MTUN CoE) funding grant, MTUN/2012/UTeM-FTMK/11 M00019.

References

1. Zakaria, M.F., Suandi, S.A.: Malaysian car number plate detection system based on template matching and colour information. *Int. J. Comput. Sci. Eng.* **02**(04), 1159–1164 (2010)
2. Ibrahim, N.K., Kasmuri, E., Norazira, A.: License plate recognition (LPR): a review with experiments for Malaysia case study. *Int. J. Soft Comput. Softw. Eng.* **3**(3), (2013)
3. Tare, S.R.: Review paper on carpooling using android operating system—a step towards green environment. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**(4), 54–57 (2013)
4. Dandu, B.R., Chopra, A.: Vehicular number plate recognition using edge detection and characteristic analysis of national number plates. *Int. J. Comput. Eng. Res.* **2**(3), 795–799 (2012)
5. Teo, L.A.K.T.K., Wong, F.: Smearing algorithm for vehicle parking management system. In: *Proceedings of the 2nd Seminar on Engineering and Information Technology*, pp. 331–337. (2009)
6. Indira, B., Shalini, M., Murthy, M.V.R., Shaik, M.S.: “Classification and recognition of printed hindi characters using artificial neural networks. *Int. J. Image Graph. Signal Process.* **4**(6), 15–21 (2012)
7. Mousa, A.: Canny edge-detection based vehicle plate recognition. *Int. J. Signal Process. Image Process. Pattern Recognit.* **5**(3), 1–8 (2012)
8. Sedighi, A., Vafadust, M.: A new and robust method for character segmentation and recognition in license plate images. *Expert Syst. Appl.* **38**, 13497–13504 (2011)
9. Dashtban, M.H.: A novel approach for vehicle license plate localization and recognition. *Int. J. Comput. Appl.* **26**(11), 22–30 (2011)
10. Jin, L., Xian, H., Bie, J., Sun, Y., Hou, H., Niu, Q.: License plate recognition algorithm for passenger cars in Chinese residential areas. *Sensors (Basel)* **12**(6), 8355–8370 (2012)
11. Mukherjee, A.: Enhancement of image resolution by binarization. *Int. J. Comput. Appl.* **10**(10), 15–19 (2010)
12. Sivanandan, S., Saiyyad, Y.: Automatic vehicle identification using license plate recognition for Indian vehicles. *Int. J. Comput. Appl.* **2012**, 23–28 (2012)
13. Goyal, R., Kaur, A.: A Review of optimal binarization techniques on documents with damaged background. *Int. J. Comput. Sci. Technol.* **4333**, 237–239 (2011)

14. Sun, G., Sun, X., Han, X.: A new method for edge detection based on the criterion of separability. *J. Multimed.* **6**(1), 66–73 (2011)
15. Gaur, S.B.C.: Comparison of edge detection techniques for segmenting car license plates. *Int. J. Comput. Appl. Electron. Inf. Commun. Eng.* **5**, 8–12 (2011)
16. Li, P., Connan, J.: Numberplate detection using double segmentation. In: *Proceedings of the 2010 Annual Research Conference of the South African Institute for Computer Scientists and Information Technologists—SAICSIT'10*, pp. 386–389. (2010)
17. Zheng, L., He, X., Samali, B., Yang, L.T.: An algorithm for accuracy enhancement of license plate recognition. *J. Comput. Syst. Sci.* **79**(2), 245–255 (2013)
18. Akter, S.: Automatic license plate recognition (ALPR) for Bangladeshi vehicles. *Glob. J. Comput. Sci. Technol.* **11**(21), (2011)
19. Yilmaz, K.: A smart hybrid license plate recognition system based on image processing using neural network and image correlation. In: *2011 International Symposium on Innovations in Intelligent System and Applications (INISTA)*, pp. 148–153. (2011)
20. Prasad, K., Nigam, D.C., Lakhotiya, A., Umre, D., Durg, B.I.T.: Character recognition using matlab's neural network toolbox. *Int. J. u- e- Serv. Sci. Technol.* **6**(1), 13–20 (2013)
21. Adebayo Daramola, T.F.S., Adetiba, E., Adoghe, A.U., Badejo, J.A., Samuel, I.A.: Automatic vehicle identification system using license plate. *Int. J. Eng. Sci. Technol.* **3**(2), 1712–1719 (2011)
22. Kranthi, S., Pranathi, K., Srisaila, A.: Automatic number plate recognition. *Int. J. Adv. Technol.* **2**(3), 408–422 (2011)
23. Reshma, P.: Noise removal and blob identification approach for number plate recognition. *Int. J. Comput. Appl.* **47**(8), 13–17 (2012)
24. Mahalakshmi, T., Muthaiah, R., Swaminathan, P., Nadu, T.: Review article : an overview of template matching technique in image processing. *Res. J. Appl. Sci. Eng. Technol.* **4**(24), 5469–5473 (2012)
25. Khalil, M.I.: Car plate recognition using the template matching method. *Int. J. Comput. Theory Eng.* **2**(5), 3–7 (2010)
26. Ibrahim, N.K., Kasmuri, E., Jalil, N.A.: A review on license plate recognition with experiments for Malaysia case study. *Middle-East J. Sci. Res.* **14**(3), 409–422 (2013)
27. Gilly, D., Raimond, D.K.: License plate recognition- a template matching method. *Int. J. Eng. Res. Appl.* **3**(2), 1240–1245 (2013)
28. Suvarna, M.: Diagnosis of burn images using template matching, k-nearest neighbor and artificial neural network. *Int. J. Image Process.* **7**, 191–202 (2013)