Lecture Notes in Networks and Systems 24

J. K. Mandal
Goutam Saha
Debdatta Kandar
Arnab Kumar Maji *Editors*

Proceedings of the International Conference on Computing and Communication Systems

I3CS 2016, NEHU, Shillong, India

Springer

Lecture Notes in Networks and Systems

Volume 24

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Advisory Board

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil
e-mail: gomide@dca.fee.unicamp.br

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey
e-mail: okyay.kaynak@boun.edu.tr

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA and Institute of Automation, Chinese Academy of Sciences, Beijing, China
e-mail: derong@uic.edu

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada and Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
e-mail: wpedrycz@ualberta.ca

Marios M. Polycarpou, KIOS Research Center for Intelligent Systems and Networks, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus
e-mail: mpolycar@ucy.ac.cy

Imre J. Rudas, Óbuda University, Budapest Hungary
e-mail: rudas@uni-obuda.hu

Jun Wang, Department of Computer Science, City University of Hong Kong
Kowloon, Hong Kong
e-mail: jwang.cs@cityu.edu.hk

More information about this series at http://www.springer.com/series/15179

J. K. Mandal · Goutam Saha
Debdatta Kandar · Arnab Kumar Maji
Editors

Proceedings of the International Conference on Computing and Communication Systems

I3CS 2016, NEHU, Shillong, India

Springer

Editors
J. K. Mandal
Department of Computer Science
and Engineering
University of Kalyani
Kalyani, West Bengal
India

Goutam Saha
Department of Information Technology
North-Eastern Hill University
Shillong, Meghalaya
India

Debdatta Kandar
Department of Information Technology
North-Eastern Hill University
Shillong, Meghalaya
India

Arnab Kumar Maji
Department of Information Technology
North-Eastern Hill University
Shillong, Meghalaya
India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-10-6889-8 ISBN 978-981-10-6890-4 (eBook)
https://doi.org/10.1007/978-981-10-6890-4

Library of Congress Control Number: 2017958721

© Springer Nature Singapore Pte Ltd. 2018
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.
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.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The Second International Conference on Computing and Communication Systems (I3CS 2016) is an annual event organized by the Department of Information Technology, North-Eastern Hill University (NEHU), Shillong, Meghalaya, India. I3CS 2016 aims to provide an interdisciplinary forum for engineers and scientists to discuss and promote research and technology in the emerging areas of the core topics of computing, information, communication, and their applications. The conference brings researchers, educators, professional engineers and technologists under a single forum to discuss developments in research.

After a successful conduction of its first edition of I3CS 2015, the Department of Information Technology, NEHU organized second I3CS 2016 during November 11–13, 2016. More than 150 researchers across the country and from abroad have participated in the event to exchange and share their research findings in the areas of computation and communication systems.

This is one of the rare international events took place in the North-Eastern part of the country. Three days of technical sessions included scientific deliberations and intellectual discussions. This event occupies a unique place in India to empower, advance, guide students and young faculties in this remote region. The I3CS 2016 organizers received more than 330 papers. All papers were reviewed thoroughly by experts of review committee, and finally 81 papers were considered for the publication based on the quality of research and outcomes.

Kalyani, India — J. K. Mandal
Shillong, India — Goutam Saha
Shillong, India — Debdatta Kandar
Shillong, India — Arnab Kumar Maji

Organizing Committee

The Second International Conference on Computing and Communication Systems (I3CS'16) was held during 11–13th November 2016 at the Department of Information Technology North-Eastern Hill University, Shillong, Meghalaya, India.

Chief Patron

Prof. Sri Krishna Srivastava, V.C., NEHU

Patron

Prof. P. V. Koparkar, Dean SOT, NEHU

General Chair

Dr. H. K. Kalita, I.T., NEHU

Organizing Chair & Co-chair

Dr. D. Kandar, I.T., NEHU
Dr. A. K. Maji, I.T., NEHU

Editorial Board

Prof. J. K. Mandal, CSE, KU. (Corr. Edtr.)
Prof. G. Saha, I.T., NEHU
Dr. D. Kandar, I.T., NEHU
Dr. A. K. Maji, I.T., NEHU

Program Chair

Prof. G. Saha, I.T., NEHU
Dr. B. Bhuyan, I.T., NEHU
Dr. S. Roy, I.T., NEHU

Finance Chair

Dr. I. Hussain, I.T., NEHU

Program Committee (Partial List)

Prof. M. G. Romay, UPV, Spain
Prof. M. K. Bhuyan, IUPUI, USA
Prof. M. S. Rahman, BUET, Bangladesh
Prof. S. Bhattacharya, UOL, UK
Prof. S. Basu, J.U., India
Prof. P. Pakray, NIT, Mizoram, India
Prof. D. Koley, JGEC, India

Organizing Committee (Partial List)

Mr. S. Neog, I.T., NEHU
Mr. S. Das, I.T., NEHU
Mr. S. C. Sahana, I.T., NEHU
Mr. K. Amitab, I.T., NEHU
Mr. A. Nath, I.T., NEHU
Mr. N. R. Pal, I.T., NEHU

Advisory Committee (Partial List)

Prof. Gai-Ge Wang, Jiangsu Normal University, Jiangsu, China
Prof. Simon Fong, University of Macau, Taipa, Macau
Prof. Nikolaos Matsatsinis, Technical University of Crete, Greece
Prof. George K. Karagiannidis Aristotle, University of Thessaloniki, Greece
Dr. Pietro H. Guzzi, University of Catanzaro, Italy
Prof. Jimson Mathew, University of Bristol, UK
Prof. Clive Rosen, University of Derby, UK
Prof. Goutam Chakraborty, Iwate Prefectural University, Japan
Prof. Jugal K. Kalita, University of CCS, Colorado, USA
Prof. Venkat Ramachandran, Concordia University, Canada

Acknowledgements

We would like to thank Prof. Sri Krishna Srivastava, Vice Chancellor, NEHU, for his continuous support. We heartily thank Prof. J. K. Mandal, Department of Computer Science and Engineering, University of Kalyani, who has always been the guiding force behind this event's success. It was his dream that we have strived to make a reality. Our special thanks to Springer and especially the editorial staff who were patient, meticulous, and friendly with their constructive criticism on the quality of papers and outright rejection at times without compromising on the quality of the papers as they are always known for publishing the best international papers.

Finally, we would like to express our heartfelt gratitude and warmest thanks to the I3CS 2016 Organizing Committee for all the hard work and outstanding contributions as a member of the organizing committee. We know how much time and energy this assignment demanded, and we deeply appreciate all of your efforts to make it a grand success.

Our special thanks to all the authors who have contributed to publish their research work here in this conference and participated to make this conference a success.

About the Conference

The Second International Conference on Computing and Communication Systems (I3CS 2016) is an annual event organized by the Department of Information Technology, North-Eastern Hill University (NEHU), Shillong, Meghalaya, India.

I3CS 2016 provided an interdisciplinary forum for researchers, engineers, and scientists to discuss and promote research and technology in the thrust areas of computing, information, communication and applications. The conference brought researchers, educators, professional engineers and technologists into a single forum in order to discuss and debate on the emerging research in the above emerging areas.

The original and unpublished research contributions submitted by authors were checked through peer review process and authors of the selected papers presented their papers during the conference.

The keynote talks, oral presentations, invited sessions on the applications and theory of computing and communication systems and related areas were presented in I3CS 2016. It provided excellent opportunities for the presentation of interesting new research results and discussion about them, leading to knowledge transfer and the generation of new ideas. The conference proceedings is published by renowned publishing house, **Springer** in its one of the popular series **LNNS**. Being one of the premier institutes in North–East, hosting this conference NEHU was able to make an imprint in the global map as well as in the country as one of the exchange venue for information and communication engineers.

Contents

About the Editors

J. K. Mandal received M.Tech. from Department of Computer Science, University of Calcutta, India and Ph.D. from Jadavpur University, India in the field of Data Compression and Error Correction Techniques. He is Professor in Computer Science and Engineering, Director, IQAC, Former Dean, FETM, University of Kalyani. Professor Mandal has 29 years of teaching and research experience. He is working in the field of Network Security, Steganography, Image Processing, Wireless and Sensor Networks. He has guided 20 Ph.D. scholars, published six books, and published more than 360 papers including 148 publications in various International journals, and edited 19 volumes as volume editor from Science Direct, Springer, CSI, etc.

Goutam Saha received B.E. (Electrical Engineering) from University of Calcutta, India, M.E. (Electronics and Telecommunication Engineering) from University of Calcutta, India, Ph.D. (Engineering) from IIT Kharagpur, India, and P.D.F. from Ben Gurion University, Israel in the field of Bioprocess Engineering and Bioinformatics. He is a Professor in the Department of Information Technology, Dean, School of Technology, North-Eastern Hill University, Shillong. Dr. Saha has 29 years of teaching and research experience. He works in the field of Systems Biology, Computational Biology and Bioinformatics, and Bioprocess Controller. He has guided 11 Ph.D. Scholars. He has published one book, more than 25 papers in conference proceedings and various national and international journals, and also has one patent.

Debdatta Kandar M.Sc. in Physics from Vidyasagar University in 2001, M.Tech. in Information Technology from A.A.I.D.U., Allahabad in 2006, he was awarded Ph.D. (Engg.) in Electronics and Communication Engineering by Jadavpur University in 2011. Presently he is working as Associate Professor and Head of Information Technology, North-Eastern Hill University, Shillong, Meghalaya. He is also working as Post Doctoral Research Fellow (PDRF) at Faculty of Engineering and the Built Environment, University of Johannesburg, Johannesburg.

He has 15 years of teaching and industrial experience. Wireless Mobile Communication, Soft Computing and Radar Operation and Imaging are the area of

specializations. Before joining NEHU, he worked at S.K.P. Engineering College, Tiruvannamalai, Tamil Nadu and at Sikkim Manipal Institute of Technology, Sikkim. He also worked in DRDO sponsored project at SMIT. He was involved in the development of Digital RADAR system on that project. He has supervised 05 M.Tech. students, one Ph.D. scholar submitted and 6 are perusing.

He has several years of industry experience too. He worked in the development of Multimedia Show Control System in Premier Irrigation Equipment Ltd., Kolkata. The system is a controlling system of a complete multimedia show which includes LASER show, Musical Fountain, Video Show, Light Show etc and also combination of all these shows. He also worked at M/S Netguru India Pvt. Ltd., Kolkata in various projects.

He has been awarded 'Young Scientist' award from Union Radio Science International (URSI GA-2005) at Vigyan Bhaban, Delhi, for his research work. President of India, Dr. A. P. J. Abdul Kalam invited him at his residence on that occasion.

Arnab Kumar Maji received Ph.D. in Engineering from Assam University and M.Tech. in Information Technology from IIEST. He has 13 years of professional experience. He is currently working as Assistant Professor, Department of Information Technology, North-Eastern Hill University, Shillong. He is working in the field of NP Complete Puzzles, Image Processing, and Image Security. One Ph.D. scholar is currently pursuing Doctoral Degree under his supervision. He has also guided six M.Tech. theses. He has published 10 book chapters and more than 30 papers in conference proceedings and various national and international journals.

Part I
Network Technologies

Chapter 1
Effective Verification Scheme for Filtering Injected False Data in Wireless Sensor Networks

Gayathri Santhosh and Yogesh Palanichamy

Abstract Wireless Sensor Networks (WSN) are used to sense the events that may occur in a given environment. The environment considered here is either unattended or hostile. The nodes that sense the events forward the data related to the sensed events to the sink through the en-routers. Attackers can compromise either the route nodes or the en-router nodes and send the false data to the sink. Sink thus receives the false report about the event. Many verification schemes are available for filtering such injected false data by checking the false reports in the en-routers before it reaches the sink. This paper provides an effective method for preventing the compromised en-routers form injecting false reports before forwarding. Dynamic keys are generated by the sink and source node use these keys to encrypt the reports about the events before forwarding the reports to the sink. These encrypted reports can be decrypted and verified only by the sink. Thus the proposed scheme is able to prevent the compromised en-router from injecting false data into the network. This scheme extends the lifetime of the network and also provides efficient forwarding of the report.

Keywords WSN · Sink · En-router · Gang injection

Introduction

Wireless sensor network is a collection of nodes organized together to form a cooperative network [1]. Each node has various sensors, actuator, a Radio Frequency (RF) receiver, a micro controller and power sources. The advancements in micro electronics and wireless communications have led to the creation of the

G. Santhosh (✉) · Y. Palanichamy
Department of Information Science and Technology, CEG Campus Anna University, Chennai 600025, Tamil Nadu, India
e-mail: m.gayath@yahoo.com

Y. Palanichamy
e-mail: yogesh@auist.net

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_1

Wireless Sensor Network (WSN) technology. Wireless Sensor Networks are widely used to monitor the unattended and hostile environments. Some popular applications that belong to this domain are habitat monitoring, forest fire detection and so on. Recently WSNs are gaining importance in urban and sensing urban areas also. Applications such as disaster management, military surveillance and emergency response come under this category.

The networks that are deployed in an unattended environment are subjected to different types of malicious attacks such as sybil attack, wormhole attack, DoS attack, black hole attack etc. One of the major attacks in sensor environment is false data injection attack. In this type of attack, the attacker will compromise the sensor node [2] which is detecting an event or it may compromise the en-routers which are used for forwarding the report. Event based report generation is compromised with the false data injected by the attacker. Hence the sink cannot receive the true data about an event. These type of attacks not only fake the report but also drain the energy by making them to forward unnecessary reports to the sink. Moreover these attacks make the sink to receive large and unwanted data besides the required data. In selective forwarding attacks, the forwarding nodes will drop the report about an event. In report disruption attack, node purposely contaminate the report and make the forwarding nodes to drop the report. Several scheme have been suggested for verifying the injected false data in the network. Many of these schemes suffer from two major limitations: a single compromised node reveals the details of all other nodes present in the network and the compromised en-routers go unnoticed/undetected.

In the proposed scheme two types of reports are generated and forwarded to the sink. The source node forwards two types of reports in the proposed scheme. One type of the reports are verified by the en-routers using shared keys and the other types of reports are verified by the sink using dynamic keys. The sink has the required information to identify the compromised en-routers. It is very difficult to attack the source node due to the dynamic keys generated by the sink. The rest of the paper is as follows. Section "Related Works" discusses about the related works. Section "System Models and Assumptions" explains about the system model and assumptions. Section "Proposed System" describes the proposed scheme and how its working in an unattended environment. Section "Conclusion and Future Work" discusses the results and conclusion.

Related Works

Several verification schemes have been suggested for filtering the false data in WSN. Statistical En-route Filtering (SEF) is one of the earliest filtering schemes for verifying the false data. SEF [3] has global key pools for authentication in which the keys are divided into non-overlapping partitions, Using these authentication

keys the reports are generated based on events. Nodes generate Message Authentication Code (MAC) using keys. Report with insufficient number of MAC are not forwarded. Sink verifies the MAC because it knows the key pool. Reports are checked by the en-router only if it shares a key with report generating nodes. Moreover this scheme is inefficient if the number of compromised nodes exceeds the threshold value of non-compromised nodes. In Interleaved Hop by Hop Authentication (IHA) scheme [4], base station enables the association process to generate the pair wise keys between the report generating nodes and en-routers. Reports that are forwarded by the nodes are verified only by the corresponding association nodes. Association repair and maintenance has been done periodically. The drawback of IHA scheme is that it requires fixed path between the base station and every cluster head for association maintenance.

Dynamic En-route Filtering (DEF) scheme [5] works only for clusters. Nodes generate authentication keys for report verification. Reports are generated in rounds. Many reports are generated in all the rounds. Keys are disseminated to all forwarding nodes for every rounds. Report forwarding and disclosing keys are used to validate the reports. The overhead here is due to the additional number of keys and control messages. Extra control messages triples the delay of the report. Secure Ticket based En-route Filtering scheme (STEF) [6] works based on query. Sink randomly selects the area and sends a query containing a ticket. It selects a random value 'C', hashed it and forward to the Cluster Head (CH). C is exclusively for CH because 'C' is encrypted using the personal key that is stored only between the node and the sink. CH receives the query and generate report and forward it to the neighbours. Now the cluster members generate their reports with their personal keys and forwards it to the CH. En-route nodes check the report by calculating the hash value for C and compare with the 'C' which was stored early. En-router does not check the content of the report. It checks only the validity of 'C' which is a major limitation.

Wang et al. [7] have proposed a scheme to focus how the neighboring nodes endorses the report whether it is logical or not. Sink knows all keys, location and relative position of nodes. Legitimacy of the nodes are calculated by the enrouter based on the location of nodes and the location of the event. However this scheme checks only whether the report is legitimate or not and does not analyze the details of the report. Effective prevention scheme for false data injection has been proposed in this paper. The sensor nodes of the sensing domain are organized into clusters based on their vicinity to the events that occur in the network. One among the nodes in the cluster is selected and designated as the source node to generate the report related to the event in coordination with the neighbours. The source node forwards the generated reports to the sink through the en-routers in shortest path for energy conservation. The verification of the report is done by both the sink and the en-routers.

System Models and Assumptions

System Model

Sensor nodes have been sowed in an unattended environment to sense the events taking place in the environment. The nodes of the WSN are static and dense. It has been assumed that the communication range of a sensor node is a circle of radius r and this communication range is considered as the sensing region of the sensor node. Another important assumption is that the communication links between the neighbours are bidirectional in nature i.e., if node u can hear node v then node v can also hear node u [4]. Multiple nodes in the vicinity of an event form a cluster and within this cluster a sensor node is nominated as the cluster head or the source node that is responsible for the generation of the report related to the detected event. All nodes within a cluster take turns to serve as the source node [8] and this implies that the source node is an ordinary sensor node which takes the responsibility of forwarding the report besides the usual function of detecting events. Sink has the information such as keys, location, neighbours and energy level of the nodes in the network.

Threat Model

It is assumed that the sink is able to broadcast the security parameters and other related information to the sensor nodes in a secure way. It is not possible for the attackers to launch the attack during the report generation since the source node knows the keys and location of the member nodes present in the cluster. However it is possible for the attacker to compromise the en-router and inject the false data through the en-router. Compromising the en-routers is possible since the en-routers do not share security parameters among themselves. We further assume that sink could not be compromised by the attacker since the sink node is well fortified.

Scenario

Figure 1.1 depicts the scenario where the sensor nodes are randomly deployed in the network. These nodes sense the events. Multiple nodes may detect the same event. N0 acts as the source node for forwarding the reports to reduce redundancy. N1, N2.… are the selected neighbours which are used to authenticate the report. R1, R2 …. are the en-routers which are used to forward the report.

After detecting an events such as forest fire, temperature change etc., sensor node generates a report about an event. Sensornode sends the report to the sink via

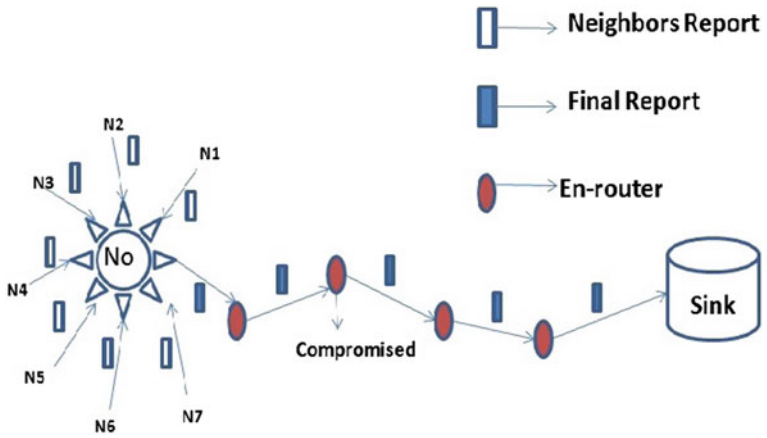


Fig. 1.1 Report generation and forwarding of reports

an established route through the en-routers. Attacker may compromise an en-router and send false data through this compromised en-router.

Proposed System

Sensor Node Deployment

Sink is responsible for assigning security parameters for all the nodes that are to be deployed in an environment. Elliptical Curve Cryptography (ECC) based public key cryptographic scheme [9] has been chosen for secure communication between the nodes. Let 'p' be a large prime and E(Fp) be a elliptical curve over prime field. Base point 'G' should be taken from the elliptical group. In ECC [9], private key 'x' is an integer randomly selected from the interval [1, p − 1]. Public key 'Y' is the product of xG, where x is a private key and G is the base point. All deployed nodes in the network will have its own private and public keypair.

Shortest Path Estimation

Source node estimates the shortest path to the sink. The source node calculates multiple path to the sink by marking the distance of the starting node as permanent and all the other distances as temporary. Setting the starting node as active the source node calculates the temporary distances by calculating the round trip time

(RTT). This is calculated using the beacon messages to the neighboring nodes by the source. This step is also called update and is central idea of Dijkstra's shortest path algorithm [8].

Report Generation

If an event occurs, nodes that are closer to the event formed as a cluster. One among the cluster node will be selected as Source node randomly [8]. Source node will selects the neighbor for report generation. Threshold level of number of neighbouring nodes is $K \leq 6$. Selection of neighbouring node is to provide verification of the reports. Source node sends the routing information to all the neighbouring nodes for the generation of shared keys. Shared keys are used by the en-router to verify the report by itself. Neighbouring nodes generate the Message Authentication Code (MAC) with the shared keys.

Shared Keys For Mac (Neighbors) Shared keys are generated by the neighbouring nodes for en-router verification. For a hostile environment, key independency is an important criterion for secure communication. Key independency is a property that ensures that a single compromised node can't reveal the identity of other uncompromised neighbours. We have applied the Elliptical Curve Diffie Hellman (ECDH) non-interactive key pair establishment to generate the shared key [9]. Let us take two nodes i and j. The shared key between these two nodes has been created as in Eq. 1.1.

$$K_{ij} = x_i Y_j \tag{1.1}$$

where x_i is the private key of node i and Y_j is the public key of node j. We know that Y_j is the public key of node j. We know that Y_j is the product xG and hence Eq. 1.1 can be written as in Eq. 1.2.

$$K_{ij} = x_i Y_j = x_i x_j G \tag{1.2}$$

Moreover we observe that k_{ij} and k_{ji} are one and the same since

$$K_{ij} = x_i Y_j = x_i x_j G = x_j x_i G = x_j Y_i = K_{ji} \tag{1.3}$$

Using these shared keys, neighbouring nodes generate Message Authentication Code (MAC) as in Eq. 1.4.

$$\mathrm{MAC}(m, k, n) = h(m \| k) \bmod 2^n \tag{1.4}$$

where m, k, n are message, key and an adjustable parameter respectively. The adjustable parameter can vary between 1 to α, where α is the number of en-routers. Each neighbouring node generates reports for all en-routers using shared keys [7].

The scenario depicted in Fig. 1.1 has 4 en-routers and 7 neighbouring nodes. Each neighbouring node generates 4 reports and forward all the reports to the source node. Node 1 generates the report as follows.

$$\text{mac}_{11} = (\text{m}\|\text{T}, \text{K}_{\text{iR1}}, \text{n}) \text{ for En} - \text{router R1} \tag{1.5}$$

$$\text{mac}_{12} = (\text{m}\|\text{T}, \text{K}_{\text{iR2}}, \text{n}) \text{ for En} - \text{router R2} \tag{1.6}$$

$$\text{mac}_{13} = (\text{m}\|\text{T}, \text{K}_{\text{iR3}}, \text{n}) \text{ for En} - \text{router R3} \tag{1.7}$$

$$\text{mac}_{14} = (\text{m}\|\text{T}, \text{K}_{\text{iR4}}, \text{n}) \text{ for En} - \text{router R4} \tag{1.8}$$

Besides the above MACs, node 1 generates mac1s for sink using its private key. All other neighbour nodes from 2 to 7 also repeat the above procedure. Like above reports, 7 neighbouring nodes will generate the report and forward it to the source node.

Secured Report from Source Node After receiving the reports from the neighbours, source node has to forward the report to the sink through the en-routers. To ensure the security the source node encrypts the report and forward the encrypted report. The sink generates the dynamic key(dk) and forwards the key to the source through a different path other than the shortest path [10]. The dynamic key(dk) is calculated based on the id of the source node, the group to which the source node belongs, the hop count from the sink, the capacity of the channel and the usage of the channel and the sequence number to ensure the freshness. The source node encrypts the generated report R using the dynamic key [dk] and the encrypted report is denoted as [*R*]*dk*. The format of the report sent from the source node to the sink is (m, T, MAC, [*R*]*dk*) where m and T are the message and the timestamp respectively. MAC holds the report of all the report generating nodes. Hence the source node formats the report from the neighboring nodes as a row vector as shown in Eq. (1.9).

$$\text{MAC} = \begin{bmatrix} \text{Row}_1 \\ \text{Row}_2 \\ \vdots \\ \text{Row}_7 \end{bmatrix} = \begin{bmatrix} \text{mac}_{11} & \text{mac}_{12} & \dots & \text{mac}_{1\text{s}} \\ \text{mac}_{21} & \text{mac}_{22} & \dots & \text{mac}_{2\text{s}} \\ \vdots & \vdots & \ddots & \vdots \\ \text{mac}_{71} & \text{mac}_{72} & \dots & \text{mac}_{7\text{s}} \end{bmatrix} \tag{1.9}$$

The above equation can be generalized for a scenario where there are ‘e’ en-routers and ’n neighbouring nodes as in Eq. (1.10).

$$\text{MAC} = \begin{bmatrix} \text{Row}_1 \\ \text{Row}_2 \\ \vdots \\ \text{Row}_\text{n} \end{bmatrix} = \begin{bmatrix} \text{mac}_{11} & \text{mac}_{12} & \dots & \text{mac}_{1\text{s}} \\ \vdots & \vdots & \ddots & \vdots \\ \text{mac}_{\text{k}1} & \text{mac}_{\text{k}2} & \dots & \text{mac}_{\text{ks}} \\ \vdots & \vdots & \ddots & \vdots \\ \text{mac}_{\text{n}1} & \text{mac}_{\text{n}2} & \dots & \text{mac}_{\text{ns}} \end{bmatrix} \tag{1.10}$$

En-Route Filtering The en-router receives the report from the upstream node and checks for the integrity of the message 'm' and the validity of the timestamp 'T'. If timestamp 'T' is outdated then the report is dropped by the en-router. If timestamp 'T' is valid the en-router checks the integrity of the message. En-router had already received the information about the neighbours of the source node and the information about the shortest path from the source node. Based o this information, the en-router calculates the shared keys of all the report generating nodes and calculate $\overline{\text{mac}}_{ij}$ as in Eq. (1.11).

$$\overline{\text{mac}}_{ij} = \text{MAC}(m\|T, k_{ij}, n) \tag{1.11}$$

The en-router checks whether the already computed mac_{ij} and this $\overline{\text{mac}}_{ij}$ are one and the same or not by XORing these two. If both *macij* and $\overline{\text{mac}}_{ij}$ are equal then $\text{mac}_{ij} \oplus \overline{\text{mac}}_{ij}$ results in zero, otherwise $\text{mac}_{ij} \oplus \overline{\text{mac}}_{ij}$ results in a non-zero value. The condition $\text{mac}_{ij} \oplus \overline{\text{mac}}_{ij} = 0$ should be true for all the entries in a row for the report to be a valid one. If the report is a valid one, then the en-router forwards the report to the downstream node, otherwise the report is dropped.

Sink Verification Sink receives the report from the en-router and checks for the integrity of the message 'm' and validity of the timestamp 'T'. Sink checks the final column of the report with the keys that is shared between itself and the neighbouring nodes. It should be noted that the sink possesses the private keys of all the nodes. With these private keys, the sink checks the forwarding reports of all the en-routers. Moreover sink has the dynamic key 'dk' which has been used by the source node for its report generation. Using this 'dk', sink decrypts the report of source node and checks whether the content of all the reports are same. Usage of dynamic key is to assure that the source node is not compromised. En-router verifies only the reports of the neighbouring nodes which are generated using shared keys. Reports that are generated using dynamic keys are not verified by the en-router. Hence there is a less chance for the compromised en-router to inject the false data into the network.

Performance Evaluation Table 1.1 shows the various network parameters used in simulations. The proposed scheme filters the false data that are possibly injected by compromised en-router and also by the neighbours of the source node. Moreover this scheme is able to conserve energy of the nodes since the report is forwarded to the sink using the shortest path to the sink. The scheme adapts itself to the dynamic topology and hence able to reduce the computational complexity associated with key generation and distribution. In many of the existing schemes, group of keys are assigned and shared among the neighbouring nodes whereas the keys are generated and shared only when the necessity arises.

The proposed scheme has used Adhoc On demand Distance Vector (AODV) routing protocol as the base routing protocol. Elliptical Curve Cryptography (ECC) has been followed in this scheme for the generation and distribution of the keys. The scheme has generated ECC as a package in ns-2 simulator and the simulations have been carried out. We have named the proposed scheme as ECC AODV.

Table 1.1 Simulation parameters

Deployment parameters	Values
Area	500 × 500
Number of nodes	53
No. of normal nodes	50
Sink	1
Antenna	Omni directional antenna
Radio propagation model	Two way ground
Network interface type	Wireless Phy
Transmission range	>30 m
MAC type	802.11 MAC
Number of malicious nodes	1 to many
Sensing time interval	<50 s
Node to sense	Any
Node to receive	Sink
Wireless bandwidth	2 Mbps
Traffic type	512 byte-CBR traffic
Average neighbours	>6

We have compared ECC AODV with the original AODV in terms of throughput and energy consumption in Figs. 1.2 and 1.3. The throughput is higher in ECC AODV since the reports are sent to the sink through the uncompromised nodes. The average energy consumption in ECC AODV is less than the original AODV since the reports are sent through the shortest path and unnecessary forwarding of falsely injected data is avoided.

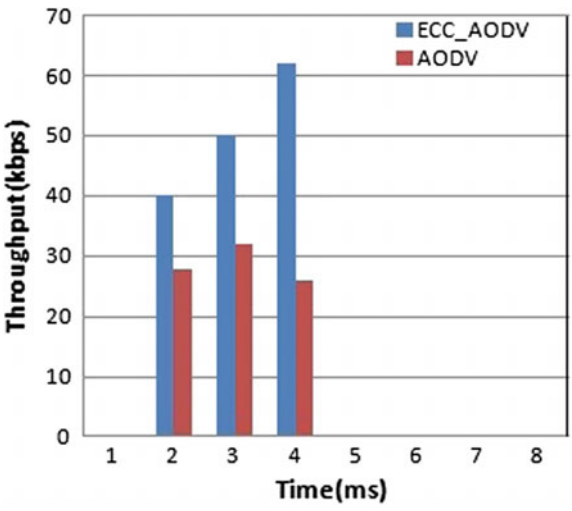


Fig. 1.2 Throughput (AODV versus ECC_AODV)

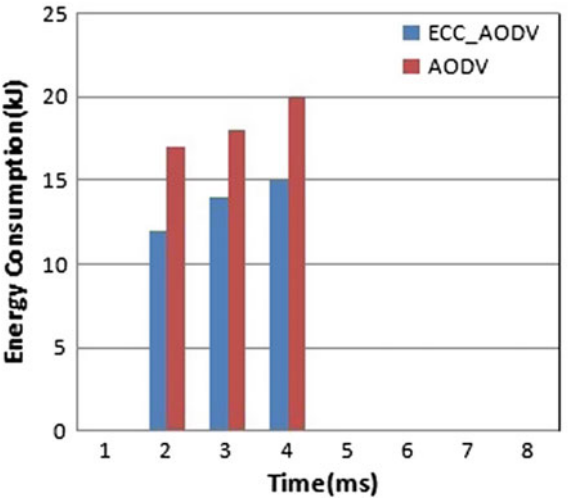


Fig. 1.3 Energy consumption (AODV versus ECC AODV)

Conclusion and Future Work

Developing verification scheme for filtering the false data injected either by the attackers or by the compromising nodes is difficult. The effective verification scheme proposed in this paper is able to overcome various difficulties of the existing schemes. In the proposed scheme the nodes are randomly deployed in the environment. Attacker nodes are simulated as mobile nodes that try to capture the data transmitted in the network and modify the data. The attacker node randomly selects an en-router to compromise the data. For every event two reports are forwarded to the sink through the shortest path. One report includes the MAC generated using the shared keys between the nodes and the other report includes the MAC generated using the dynamic keys. The en-router verifies the report that includes the MAC generated using the dynamic keys. The report is dropped either by an en-router or by a sink whenever a mismatch occurs. This paper assumes that the source node is a reliable node and the future work shall also consider the case of non-reliable source node. Moreover this work assumes that it is difficult or impossible for the attackers to compromise all the neighbours of the source node. The future work can relax these assumption and come up with more sophisticated techniques to deal with these situation.

References

1. Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine, 40, 8, 102–114 2002.
2. Zhou, L., Haas, Z. J.: Securing ad hoc networks. IEEE network, 13, 6, 24–30 1999.

3. Ye, F., Luo, H., Lu, S., Zhang, L.: Statistical en-route filtering of injected false data in sensor networks.: IEEE Journal on Selected Areas in Communications, 23, 839–850 2005.
4. Zhu, S., Setia, S., Jajodia, S., Ning, P.: An interleaved hop-by-hop authentication scheme for filtering of injected false data in sensor networks. In: IEEE Symposium on Security and privacy, pp. 259–271. IEEE Press, 2004.
5. Yu, Z., Guan, Y.:. A dynamic en-route filtering scheme for data reporting in wireless sensor networks. IEEE/ACM Transactions on Networking (ToN), 18, 1, 150–163 2010.
6. Kraub, C., Schneider, M., Bayarou, K., Eckert, C.: STEF: A secure ticket-based en-route filtering scheme for wireless sensor networks. In: 2nd IEEE International Conference on Availability, reliability and security, pp. 310–317. IEEE Press, 2007.
7. Wang, J., Liu, Z., Zhang, S., Zhang, X.: Defending collaborative false data injection attacks in wireless sensor networks. Information Sciences, 254, 39–53 2014.
8. Lu, R., Lin, X., Zhu, H., Liang, X., Shen, X.: BECAN: a bandwidth-efficient cooperative authentication scheme for filtering injected false data in wireless sensor networks. IEEE transactions on parallel and distributed systems, 23, 1, 32–43 2012.
9. Ahirwal, R. R., Ahke, M.: Elliptic curve diffie-hellman key exchange algorithm for securing hypertext information on wide area network. International Journal of Computer Science and Information Technologies, 4, 2, 363–368 2013.
10. Chan, H., Perrig, A., Song, D.: Random key predistribution schemes for sensor net- works. In IEEE Symposium on Security and Privacy, pp. 197–213. IEEE Press, 2003.

Chapter 2
ISCP: Inter State Changing Problem in Wireless Sensor Network with Dumb-Behaving Nodes

Pushpendu Kar, Subhabrata Barman and Subhransu Das

Abstract In a wireless sensor network (WSN), different types of nodes (i.e., active, misbehaving, dead) may coexist. To ensure seamless performance of the network, network administrators often use several mechanisms to recover misbehaving or faulty nodes. As it is impossible for a node to predict the state of another node, the recovery attempts may not be successful every time. Such unsuccessful attempts may bring detrimental effects to the network. In this paper, we introduce three of such problems which occur due to change in state (such as dumb to dead, etc.) with such attempts. We name the problems as Recovery Attempt for Dead (RAD) node, Insecure Recovery Attempt (IRA), and Data Burst (DB). The detrimental effects of these problems are evaluated using network simulator 3 (ns3) simulations. Additionally, this paper also suggests some solution guidelines for these problems. The analysis and simulation experiments led us to enlist some open research issues in the concluding section.

Keywords WSN · Misbehaving nodes · RAD · IRA · DB

P. Kar
Energy Research Institute, Nanyang Technological University, Singapore, Singapore
e-mail: pushpendu.kar@gmail.com

S. Barman
Department of Computer Science and Engineering, Haldia Institute of Technology, Haldia, West Bengal, India
e-mail: subha.barman@gmail.com

S. Das (✉)
Department of Computer Application, Contai College of Learning & Management Science, Purba Medinipur, Contai 721401, West Bengal, India
e-mail: subhransu.cs@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_2

Introduction

Ubiquitous deployment, ad hoc nature, and low-cost implementation have given wireless sensor network enormous popularity [1]. Several applications (such as active volcano detection [2], weather monitoring [3], underwater monitoring, etc.) often deploy sensor networks in hostile areas. Sensor nodes often experience adverse environmental conditions such as rainfall, fog, high temperature, etc., due to outdoor deployment. Such adverse environmental conditions often cause shrinkage in the communication range of a sensor node [4]. When a node becomes disconnected from other nodes due to environmental adverseness, it exhibits dumb behavior [5]. Dumb nodes have detrimental effects on the network's performance. In order to maintain smoothness in the networks performance, dumb nodes need to be detected and healed. On the other hand, sensor nodes are tiny, battery-operated device with limited life span. In order to save energy, they switch from one state to another. For example, a sensor node may switch to idle (or sleep) state from active state, when its transmission and reception (T_x/R_x) activities are stop [6]. At the end of life cycle of the battery, sensor nodes permanently move to dead state. Wireless sensor nodes are prone to security threats also [7]. Nodes may be physically hijacked or replicated by the adversaries [8]. An interesting issue of such attacks is that if a node is hijacked or replicated, it might exhibit dumb behavior to all other nodes until a false node is placed back. For a sensor node, it is impossible to predict the exact state of its neighboring node. Therefor, if it initiates to recover a misbehaving node (like dumb node), the recovery attempt may not succeed every time additionally such unsuccessful recovery attempt may bring detrimental effects to the network also. This paper introduces three of such problems namely—RAD, IRA, and DB. Section of this paper describes problems, and section "Performance Evaluation" evaluates the performance of different problems using simulation. Out of the remaining parts of the paper, section "Sensor Nodes States" describe the states of wireless sensor networks and Section refers a guideline to overcome these problems.

Contributions

- This paper introduces three problems (namely i. Recovery Attempt for Dead (RAD) node, ii. Insecure Recovery Attempt (IRA), and iii. Data Burst (DB). These problems may occur in wireless sensor network with change in state of the nodes.
- Analysis of the detrimental effects of such problems using simulation.
- A solution guideline is also provided for solving these problems.

Sensor Nodes States

This section describes the states of wireless sensor devices according to Fig. 2.1.
i. Idle: In this state, a sensor node remains active but does not perform any activity. Energy consumption during ideal state is very minimum and only to activate processing unit and low-power receiving unit.
ii. Transmit/Receive (Tx/Rx): In this state, a sensor node remains fully active to transmit its own sensed information, receive, or forward others' information. Energy consumption of sensor node in this state is maximum.
iii. Dumb: In this state, sensor networks exhibit misbehavior. Sensor nodes cannot transmit any packet to their neighbors due to temporal shrinkage in their transmission range under adverse environmental conditions [5]. Such behavior is denoted by ψ_d in Eq. 2.1, (notations related to this equation are also explained in [5]).

$$\Psi_d = \begin{cases} 1, & \{(0 < d_{min} \leq r_c(t_i) \leq R)\} \wedge \{0 \leq r_c(t_j) < d_{min} < R)\} \quad \forall t_i \forall t_j \quad t_i \neq t_j \\ 0, & otherwise \end{cases} \tag{2.1}$$

iv. Compromised: In this state, a sensor node pretends to be a node working for benefit of the network, but it works for an external entity to diminish the network performance and steal information. A node enters this state due to external attack.
v. Dead: In this state, a sensor node cannot perform any activity due to damage or insufficient energy to perform any activity. If a node is dead, it cannot become active again without external intervention.

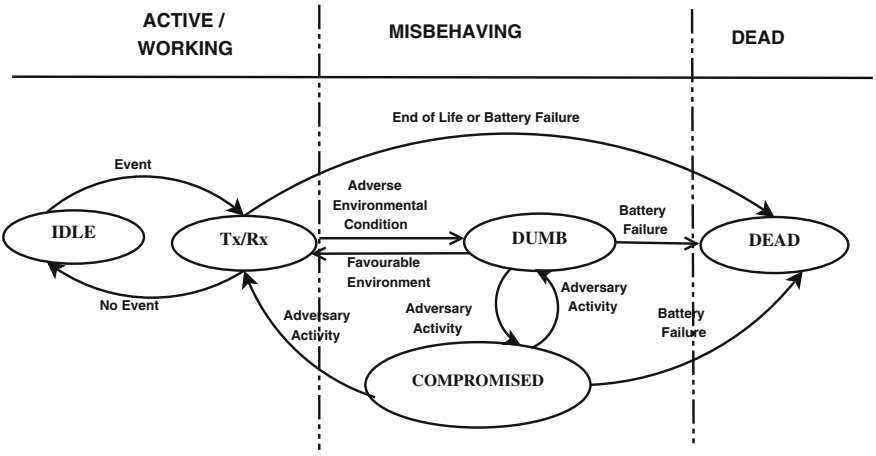


Fig. 2.1 State diagram of wireless sensor nodes

Description of the Problems

Inter State Changing Problem (ISCP) occurs when a node is assumed to be in dumb state, but practically, its state has changed from dumb to dead or compromised. This section describes these problems.

RAD: Recovery Attempt of Dead Nodes

Existence of dumb or disconnected node is harmful for network's performance [9]. Distributed approach may be applied to recover dumb nodes [10]. Under this scheme, a number of hello packets are exchanged to recover a dumb node. One major problem with this approach is state change. If the node moves from dumb state to dead state then the total effort becomes useless. The distributed approach to recover from dumb state becomes an useless Recovery Attempt of Dead (RAD) node.

Theorem 1 *Degradation in channel utilization increases with the number of RAD nodes.*

Proof We assume that in a network of N nodes, every node can transmit P packets in t time slot with P_s successful transmission probability. Now, two incidents can take place in the network.

• **Case 1**: When all nodes participate in the network data transfer, then network throughput (τ) becomes

$$\tau = \frac{\sum_{i=1}^{N} P_i P_s}{t} \tag{2.2}$$

• **Case 2**: Let n_1 nodes participate in network data transfer and n_2 nodes participate in RAD operation, where $N = n_1 + n_2$. The overall throughput (τ) divides in network throughput (τ_i) and RAD throughput (τ_{RAD}), such as $\tau = \tau_i + \tau_{RAD}$

$$\tau = \frac{\sum_{i=1}^{n_1} P_i P_s}{t} + \frac{\sum_{i=1}^{n_2} P_i P_s}{t} \tag{2.3}$$

Here, t and P_s can be considered as constant. So τ_{RAD} becomes

$$\tau_{RAD} = \frac{P_s}{t} \sum_{i=1}^{n_2} P_i \tag{2.4}$$

Hence, from 2.3 and 2.4, we conclude that increment of n_2 will significantly increase RAD efforts minimizing network throughput τ_i.

IRA: Insecure Recovery Attempt

Wireless sensor networks are prone to attack in form of node hijack or node replication. An adversary may physically hijack a node and read its security credentials. After understanding the security arrangements of the network, the adversary may put the node back into its position to insert mal data in the network. It is obvious that a compromised node should behave like a dumb node during its hijack. But if the detection process does not include security check, the adversary may easily intrude a mal node in the network. Another important aspect of this attack is related to the payload size. From [11] we know that, networks saturation throughput(S_{th}) of the network is a function of the number of payload information bits successfully transmitted within a single virtual time slot (E_{pl}), and the length of the virtual time slot L_{ts}. The relation between them is shown in Eq. 2.5

$$S_{th} = \frac{E_{PL}}{L_{ts}} \tag{2.5}$$

Therefore, it is evident that if an adversary can inject large data frames in the network, he can insert large number of mal data in the network. Figure 2.3b reflects this issue and validates our notion.

DB: Data Burst

Data burst is a phenomenon, which occurs on the return of favorable weather condition. When a dumb node suddenly recovers its transmission and reception capability, it broadcasts common packets to all its neighboring nodes. Hence, it increases routing overhead as well as traffic load in the network.

Solution Guideline

Table. 2.1 prescribes possible solution guidelines which can be adopted to overcome such problems.

Table 2.1 Solution guideline for Inter State Changing Problems

Problem	Solution guideline
RAD	One possible solution of such problem is to implement a self-detection mechanism Under a self-detection mechanism, a node detects its state and tries to recover itself
IRA	The only solution to this problem is to implement a security framework for the misbehavior detection algorithm. The security protocol should check the authenticity of a node whenever it is recovered from dead state
DB	Data Burst is a natural network phenomenon, and it cannot be avoided completely But any routing protocol, which periodically updates routing table can ameliorate the effect of Data Burst (DB) to some extent

Performance Evaluation

Simulation Setup

In order to verify our target issues, we performed simulation experiments in network simulator-3 (ns-3) [12]. The simulation was performed with several nodes randomly deployed over 500 m × 500 m surface area. Packet size is considered as 512 bytes, where hello packet size was approx 50 bytes. Some common simulation parameters are mentioned in Table 2.2.

Table 2.2 Simulation parameters

Parameter	Values
Number of nodes	10–30
Percentage of faulty node	10–30%
Percentage of node in detection (RAD nodes)	20–60%
Deployment area	500 m × 500 m
Data packet size (λ)	512 bytes
Hello packet size	100 byte (approx.)
Mal-data packet size (α)	1000–1500 bytes
Number of packets (per application)	10,000
Inter-packet interval	0.001 s
Placement of the nodes	Random
Total initial energy	10 J

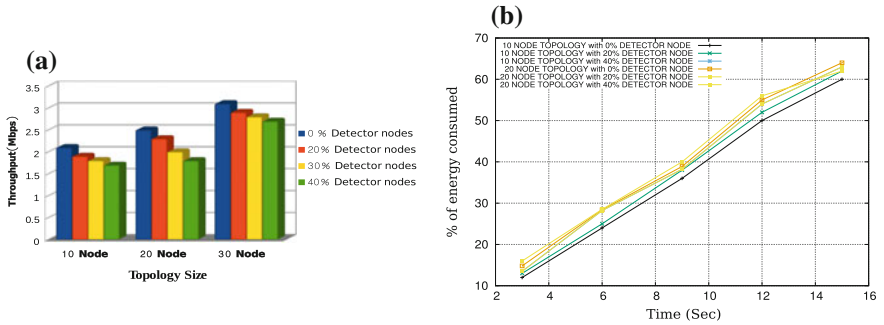


Fig. 2.2 Test of RAD **a**: degradation in network throughput, **b**: energy consumption

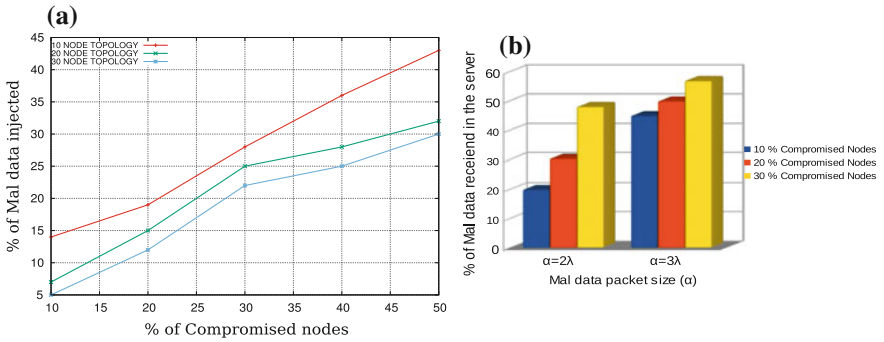


Fig. 2.3 Test of IRA **a**: insertion of mal data **b**: amount of mal data consumed with increasing mal-data size

In our first experiment, we test RAD. We deploy 10–30 nodes randomly over 500 m × 500 m area, where 0–40% nodes were engaged in dumb node detection. Packet queue for every node was full. Additionally, detector nodes have a responsibility to transmit hello packet to detect dumb nodes. As our target dumb node has actually become a dead node due to change of state (Fig. 2.1) Hence, all efforts provided for this detection consumes some bandwidth which is useless in practice. For example, Fig. 2.2a exhibits that, in a topology of 20 nodes, 34% drop in throughput having 40% detector nodes in the network. This validates the statement of Theorem 1. Figure 2.2b exhibits a slight increment in the energy consumption with the detection process. The nodes which are exchanging hello packet are consuming more energy from the source to accomplish T_x/R_x operation.

In our second experiment, we place compromised nodes randomly in the network. These nodes inject similar data packet to the original. We notice that a compromised node may declare itself as a recovered dumb node. If the detection process does not imply sufficient security check, compromised nodes may easily enter into the

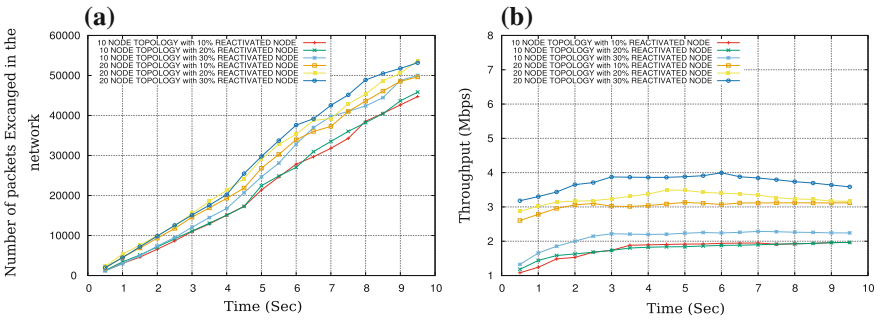


Fig. 2.4 Test of DB **a**: increment in packets in the network **b**: change of throughput with DB

network declaring them as recovered dumb nodes. This experiment exhibits IRA problem. We create similar data packet as the originals and inject them in the network. Results of this experiment are presented in Fig. 2.3a. The plot shows that if an adversary hijacks and replicated almost 40% of the nodes in a network, then he can easily inject more than 40% mal data in the network. This effect becomes even worse when an adversary inputs larger data frames than the normal one. We doubled and tripled the mal-data packet size(α) and observed the output in Fig. 2.3b. The result shows that if the adversary uses mal-data packet, which are three times in size of the original packet (i.e., $\alpha = 3 \times \lambda$), he can inject approx. 57% mal data in the network by replicating only 10% nodes of the network.

Our final experiment represents Data Burst (DB) problem in the wireless sensor network. In this experiment, a certain percentage of nodes become activate between 4.5–5 s. Figure 2.4a shows that the hike in received packet count in the network during this time period. This happens due to redundant data flow via the detector nodes. When a node gains back its transmission and reception capability under favorable weather condition, all of its neighbors receive same packet from the node due to the broadcast nature of wireless sensor network. This hike in received packet does not contribute any significant improvement in the network's throughput, as shown in Fig. 2.4b. For this reason, we call this phenomenon as Data Burst problem.

Conclusion

Since the last few years, sensor networks are being deployed in countless fields of application. When sensor nodes are deployed in adverse environmental conditions or hostile areas, they experience several problems. In this paper, we introduce three of such problems namely RAD, IRA and DB which occur due to change of state of sensor nodes. The solution of these problems may be subject to further research, but it is necessary to solve this problems for security and performance issues of the network. During real-life deployment, sensor networks often face new problems.

Advancement of research on this field solves such problems. Additionally, it introduces new features of sensor devices. This trend has made wireless sensor network a popular medium of communication and useful in our life.

References

1. S. Das, S. Barman, S. Bhunia, Performance analysis of ieee 802.11 rate adaptation algorithms categorized under rate controlling parameters, in: Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, ACM, 2014, p. 8.
2. G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, M. Welsh, Fidelity and yield in a volcano monitoring sensor network, in: Proceedings of the 7th symposium on Operating systems design and implementation, USENIX Association, 2006, pp. 381–396.
3. R.-H. Ma, Y.-H. Wang, C.-Y. Lee, Wireless remote weather monitoring system based on mems technologies, Sensors 11 (3) (2011) 2715–2727.
4. K. Bannister, G. Giorgetti, S. K. Gupta, Wireless sensor networking for hot applications: Effects of temperature on signal strength, data collection and localization, in: Proceedings of the 5th Workshop on Embedded Networked Sensors (HotEmNets 08), Citeseer, 2008.
5. S. Misra, P. Kar, A. Roy, M. S. Obaidat, Existence of dumb nodes in stationary wireless sensor networks, Journal of Systems and Software 91 (2014) 135–146.
6. H.-Y. Zhou, D.-Y. Luo, Y. Gao, D.-C. Zuo, Modeling of node energy consumption for wireless sensor networks, Wireless Sensor Network 3 (1) (2011) 18.
7. Q. Zhang, T. Yu, P. Ning, A framework for identifying compromised nodes in wireless sensor networks, ACM Transactions on Information and System Security (TISSEC) 11 (3) (2008) 12.
8. W. T. Zhu, J. Zhou, R. H. Deng, F. Bao, Detecting node replication attacks in wireless sensor networks: a survey, Journal of Network and Computer Applications 35 (3) (2012) 1022–1034.
9. C. Zhu, C. Zheng, L. Shu, G. Han, A survey on coverage and connectivity issues in wireless sensor networks, Journal of Network and Computer Applications 35 (2) (2012) 619–632.
10. A. Roy, P. Kar, S. Misra, M. S. Obaidat, D3: Distributed approach for the detection of dumb nodes in wireless sensor networks, International Journal of Communication Systems.
11. Y. Lin, V. W. Wong, Wsn01-1: frame aggregation and optimal frame size adaptation for ieee 802.11 n wlans, in: IEEE globecom 2006, IEEE, 2006, pp. 1–6.
12. T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, J. Kopena, Network simulations with the ns-3 simulator, SIGCOMM demonstration 15 (2008) 17.

Chapter 3
Effects of Persistent Misbehaving Nodes in Wireless Sensor Networks

Subhransu Das, Jayashri Deb Sinha and Subhabrata Barman

Abstract In a wireless sensor network (WSN), sensor nodes may temporarily exhibit mal-behavior (e.g., dumb behavior) under adverse environmental conditions. A misbehaving node may gain back its work ability with the return of favorable weather conditions. In this paper, we consider some scenarios where sensor nodes exhibit mal-behavior for a significant duration of time. We call such misbehaving nodes as "persistent misbehaving nodes". We present an analytical model which shows the severeness of the detrimental effects caused by persistent misbehaving nodes. The performance analysis corroborates our model and compares the effects of temporary and persistence misbehaving nodes. The results of the performance analysis also show that the existence of persistent misbehaving nodes causes 38% degradation in throughput and 18% extra energy depreciation than the temporary misbehaving nodes. The analysis and performance analysis led us to enlist some open research issues in the concluding section.

Keywords WSN · Persistent misbehaving nodes · ns3 · Dumb nodes

S. Das
Department of Computer Application, Contai College of Learning & Management Science, Purba Medinipur, Contai 721401, West Bengal, India
e-mail: subhransu.cs@gmail.com

J. D. Sinha
Department of Computer Science Engineering, Bengal Institute of Technology & Management, Shantiniketan, West Bengal, India
e-mail: jdebsinha@gmail.com

S. Barman (✉)
Department of Computer Science Engineering, Haldia Institute of Technology, Haldia, West Bengal, India
e-mail: subha.barman@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_3

Introduction

In countless fields of applications, sensor nodes are deployed in hostile areas, where they suffer environmental adverseness like high temperature, rainfall, etc. [1, 2]. Under such environmental hazards, sensor nodes may lose their sensing capabilities and exhibit trans-faulty behavior [3] or their communication range may shrink and they may exhibit dumb behavior [4]. The detrimental effect of environmental hazards on the performance and life time of sensor nodes is a proven phenomenon. For example, the attenuation of signal under rainfall and snowfall is formulated in [5]. High environmental temperature has a crucial impact on the performance of the sensor nodes. In [6], it is shown that the communication radius of a node may shrink upto 60% with increasing temperature. We adhere to the valuable results of this paper to model our experiments. Dumb [4] and trans-faulty [3] nature are the two temporal misbehaviors of sensor nodes observed under environmental adverseness. Dumb and trans-faulty nodes may gain back their work ability under favorable weather conditions. Therefore, they are considered as temporary misbehavior. In this paper, we move a step forward and analyze the effect of persistent misbehavior of sensor nodes. We call a node as persistence misbehaving node if it cannot recover itself within a given time period Δt. For example in a manufacturing industry, the room temperature may remain high through out the production hours. During these hours, sensor nodes will experience a shrinkage in their communication range and will exhibit dumb behavior for a significantly long time. In order to analyze the effects of persistent misbehaving nodes, we outline the remaining parts of the paper as follows. In section "Difference Between Temporary and Persistent Misbehavior", we distinguish the effects temporary and persistent misbehavior. Where, in section "Performance Evaluation", we analyze the effects of persistent misbehaving nodes in terms of network throughput, delay, average inter-packet arrival delay (*AIPAD*), and energy consumption using ns-3 simulations [7]. The outcome of this analysis and performance evaluation allowed us to enlist some open research issues in the concluding section.

Difference Between Temporary and Persistent Misbehavior

A temporary misbehaving node may work properly with the return of favorable weather conditions. For example, a dumb node may regain its transmission and reception capabilities with the return of favorable weather conditions. But when malbehavior of a node persists for a significantly high time duration (Δt), it becomes persistent misbehaving node. The selection of the Δt is a flexible choice for the network administrator. The throughput of a communication link depends on the misbehavior duration of a receiver node.

Theorem 1 *The throughput of a communication link depends on the misbehavior interval of the receiver.*

Proof Let us assume that a node receives p packets in t time interval with uniform inter-packet arrival interval. Therefore, average packet arrival rate of the link (λ) can be expressed as

$$\lambda = \frac{p}{t}$$

By definition, a misbehaving (dumb node) cannot receive any packet during misbehavior interval. Using this notion, we model the probability of misbehaving (dumb) node (P_{dumb}) using Poisson distribution (as shown in Eq. 3.1)

$$P_{dumb} = P\{x = 0\} = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{3.1}$$

Here, x is the expected number of packets in specified time interval which is zero ($x = 0$) for misbehaving (dumb) nodes and $\lambda = \frac{p}{t}$ therefore, we rewrite 3.1 as 3.2.

$$P_{dumb} = e^{-\frac{p}{t}} \tag{3.2}$$

Assumes that a sender transmits P_t packets in T time interval. Out of this T time interval, the receiver remains T_{active} time active and exhibits misbehavior for $T_{misbehavior}$ time periods, where

$$T = T_{active} + T_{misbehavior}$$

Hence, the throughput of the link (T_h) is expressed as shown in Eq. 3.3

$$T_h = \frac{P_t.(1 - e^{-\frac{p}{t}})}{T_{active} + T_{misbehavior}} \tag{3.3}$$

From Eq. 3.3, it is evident that the throughput of a link is inversely proportional to the misbehavior interval of the receiver. A corollary of this theorem is that "the throughput of a link depends on the probability of misbehavior of the sender also". We observe that persistent misbehaving nodes delivers less network throughput then temporary misbehaving nodes in the performance evaluation section.

The changes in Activity Node Ratio (*ANR*) is significantly high in the network with temporary misbehaving nodes compared to the persistent misbehaving nodes.

Definition 1 *Activity Node Ratio* (*ANR*): It is the ratio between properly behaving nodes ($N_{properly\_behaving}$) and total nodes (N) of the network.

$$ANR = \frac{N_{properly\_behaving}}{N}$$

The values *ANR* range from 0 to 1. Figure 3.1a depicts the changes in *ANR* in different networks. One obvious effect of this fluctuation in *ANR* can be observed

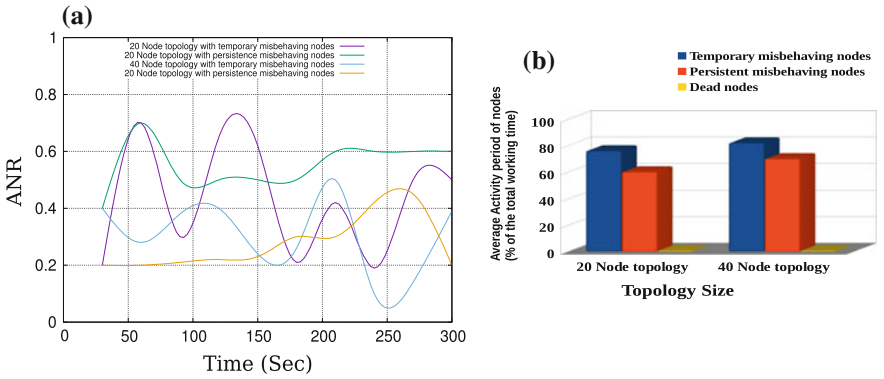


Fig. 3.1 Analysis of persistent misbehaving nodes **a**: changes in *ANR* **b**: average activity period of the nodes

the average activity period of the nodes. In Fig. 3.1b, we notice that a temporary misbehaving node remains active longer than a persistence misbehaving node.

Average inter-packet arrival duration (*AIPAD*) also depends on the persistence nature of misbehavior exhibited by a sensor node.

Lemma 1 *Average inter-packet arrival duration* (*AIPAD*) *depends on the misbehavior interval of the receiver.*

Proof Let us assume that in a network, a temporary misbehaving node receives P packets over a communication link in T time. T is represented as

$$T = \Delta t_{temp} + \Delta t_{active}$$

where Δt_{temp} and Δt_{active} are time slots for exhibiting disconnected misbehavior and active nature. Therefore, average inter-packet arrival duration for temporary misbehaving node ($AIPAD_{temp}$) can be calculated as shown in Eq. 3.4

$$AIPAD_{temp} = \frac{T}{P} = \frac{\Delta t_{temp} + \Delta t_{active}}{P} \tag{3.4}$$

Similarly, we derive average inter-packet arrival duration for persistence misbehaving node ($AIPAD_{persi}$) in Eq. 3.5

$$AIPAD_{persi} = \frac{T}{P} = \frac{\Delta t_{persi} + \Delta t_{active}}{P} \tag{3.5}$$

Now,

$$\Delta t_{persi} > \Delta t_{temp}$$

Hence, for a common Δt_{active} and P,

$$AIPAD_{persi} > AIPAD_{temp} \tag{3.6}$$

Thus, Eq. 3.6 proves the lemma.

We verify this notion in the following performance evaluation section.

Performance Evaluation

Simulation Setup

In this section, we analyze the effect of persistence misbehaving nodes over the networks performance. We perform several simulation experiments in network simulator 3 (ns-3) for this purpose. We deploy 20–100 nodes in a 300 m × 300 m area with some common simulation parameters as shown in Table 3.1. We simulated "dumb behavior" as a misbehavior to represent its temporary and persistent effects over the network.

Our first set of simulations analyzes network throughput with the existence of temporary and persistent misbehaving nodes. We deployed 20–50 nodes uniform randomly over the simulation area. Every node transmitted UDP packets of size 512 bytes. The inter-packet interval for this experiment was 1 s. Figure 3.2a, b shows the change in network throughput. As the fluctuation of *ANR* is remarkably high in networks with temporary misbehaving nodes, the throughput of the network (observed in uniform time interval) changes more frequently.

Temporary misbehaving nodes can recover themselves sooner than the persistent misbehaving nodes. This phenomenon affects the overall throughput of the network. Figure 3.3a depicts that a network with temporary misbehaving node can even achieve 28% more throughput than the network with permanent misbehaving nodes. This results corroborates our notion of Theorem 1.

Table 3.1 Simulation parameters

Parameter	Values
Number of nodes	20–100
Placement of the nodes	Uniform random
Deployment area	300 m × 300 m
Data packet size (λ)	512 bytes
Inter-packet interval	0.01–1 s
Number of packets (per application)	10,000
Total initial energy	50 J

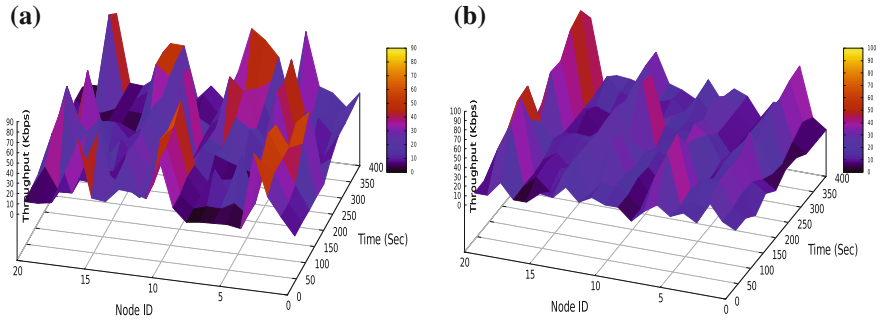


Fig. 3.2 Analysis of throughput on every node **a**: with temporary misbehaving nodes **b**: with persistent nodes

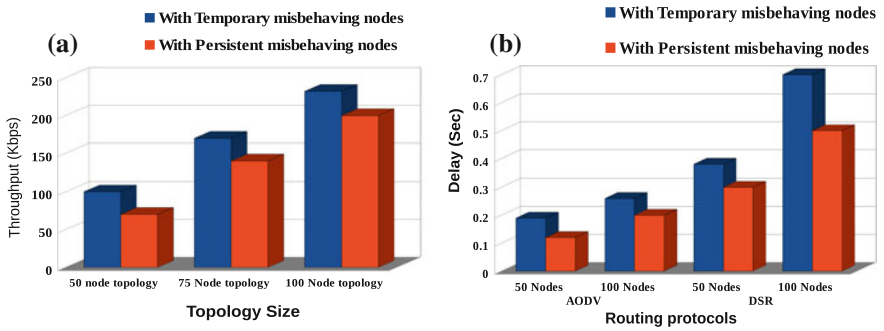


Fig. 3.3 Analysis of persistent misbehaving nodes **a**: network throughput **b**: delay against different routing protocols

Frequent changes in *ANR* may create delay in the routing decisions. Reconstruction of routing table and repeated route discovery increase delay in the network with temporary misbehaving nodes. Figure 3.3b shows the delay with temporary and persistent misbehaving nodes. We observe a 38% extra delay in the network of 100 nodes with temporary misbehaving nodes under DSR protocol. In almost every test case, AODV [8] outperforms DSR [9]. Out of these experiments, the minimum delay (0.102 s) occurred in 50 node topology with AODV protocol (with persistent misbehaving nodes).

To corroborate Lemma 1, we simulated three network scenarios where inter-packet sending interval ranges from 0.01 to 0.03 s. Lemma 1 shows that $AIPAD_{temp}$ will be higher than $AIPAD_{persi}$. In Fig. 3.4a, we observe in all test cases the value of $AIPAD_{persi}$ is greater than $AIPAD_{temp}$. We also notice a 28% increase in AIPAD over a link when the inter-packet sending interval was 0.01 s.

In our final set of experiments, we analyze the energy consumption of the network with temporary and persistent misbehaving nodes. Misbehaving nodes (dumb nodes) are clearly distinguishable from dead nodes. They are active in nature but their work ability is suspended due to adverse environmental conditions. We

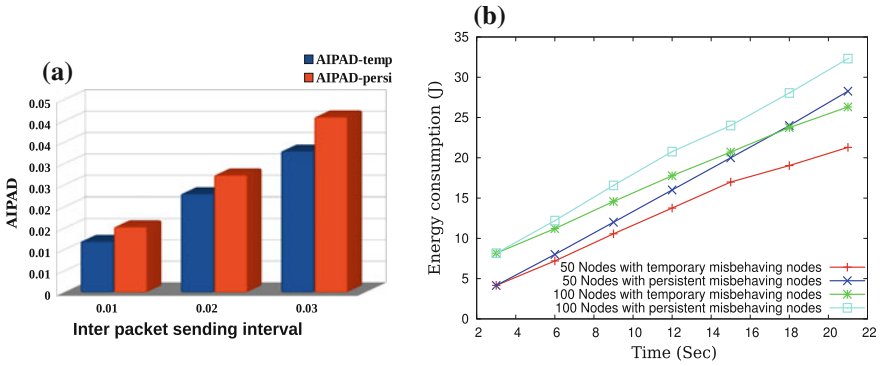


Fig. 3.4 Analysis of persistent misbehaving nodes **a**: AIPAD **b**: energy consumptions

simulated two network scenarios with 50 and 100 nodes. In both scenarios, the energy consumption of the network with persistence dumb node is higher than the other. For example, we notice 18% extra energy consumption with persistent misbehaving node with 100 nodes during 20–22 s time interval.

Conclusion

This paper analyzes the effects of persistence misbehaving nodes over a network. Additionally, it distinguishes the effects of temporary and persistent misbehaving nodes using simulation experiments. We summarize that the detrimental effects of persistent misbehaving nodes over network throughput and energy consumption are more severe then the effects of temporary misbehaving nodes existing in the network. However, a statistical measurement can corroborate our results more precisely. Additionally, a statistical measurement based system for detection and avoidance of persistent misbehaving nodes can be an interesting project goal for future work.

References

1. J. Yick, B. Mukherjee, D. Ghosal, Wireless sensor network survey, Computer networks 52 (12) (2008) 2292–2330.
2. T. Arampatzis, J. Lygeros, S. Manesis, A survey of applications of wireless sensors and wireless sensor networks, in: Intelligent Control, 2005. Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation, IEEE, 2005, pp. 719–724.
3. P. Kar, S. Misra, Reliable and efficient data acquisition in wireless sensor networks in the presence of transfaulty nodes, IEEE Transactions on Network and Service Management 13 (1) (2016) 99–112.

4. S. Misra, P. Kar, A. Roy, M. S. Obaidat, Existence of dumb nodes in stationary wireless sensor networks, Journal of Systems and Software 91 (2014) 135–146.
5. F. Nadeem, E. Leitgeb, M. Awan, S. Chessa, Comparing the life time of terrestrial wireless sensor networks by employing hybrid fso/rf and only rf access networks, in: Wireless and Mobile Communications, 2009. ICWMC'09. Fifth International Conference on, IEEE, 2009, pp. 134–139.
6. K. Bannister, G. Giorgetti, S. K. Gupta, Wireless sensor networking for hot applications: Effects of temperature on signal strength, data collection and localization, in: Proceedings of the 5th Workshop on Embedded Networked Sensors (HotEmNets 08), Citeseer, 2008.
7. T. R. Henderson, M. Lacage, G. F. Riley, C. Dowell, J. Kopena, Network simulations with the ns-3 simulator, SIGCOMM demonstration 15 (2008) 17.
8. C. Perkins, E. Belding-Royer, S. Das, et al., Rfc 3561-ad hoc on-demand distance vector (aodv) routing, Internet RFCs (2003) 1–38.
9. S. Goswami, S. Joardar, C. B. Das, Performance comparision of different routing protocols with fuzzy inference system in manet, Advances in Science and Technology Research Journal 9 (27).

Chapter 4
Performance Comparison of Routing Protocols in Mobile Ad Hoc Networks

Alak Roy and Titan Deb

Abstract Mobile Ad Hoc Networks (MANETs) are self-sorted intercommunicating systems shaped by portable hosts, without any settled framework. In MANETs, data transmission requires several hosts linking sender and receiver for effective routing of data packets. The nodes in MANETs have the capacity to move arbitrarily and compose themselves subjectively in the network. This makes the networks more complicated in nature (hence, require an efficient routing protocol to deliver data). There is plenty of routing techniques that are available for MANETs. It is hard to decide efficient protocol for a constrained situation. In this paper, a comprehensive survey of routing protocols that are available in recent literature with classification is given. This paper gives portrayals of the protocols with exchanges of preferences and drawbacks of the various routing protocols. Finally, a simulation-based performance evaluation of the routing is presented to compare their performance based on average throughput, end-to-end delay, and packet delivery fraction in various data rates.

Keywords Mobile ad hoc networks · Routing protocols

Introduction

MANETs [1] are the most encouraging fields for research. Since 1970, remote systems get to be more famous and a fast development of research hobbies in MANETs began in the 1990s. In the last couple of years, the utilization of remote systems has ended up more mainstream. MANETs are multi-jump impermanent foundation less self-arranging system which conveys through remote connection. In MANETs, mobility causes stochastic changes of topologies. Therefore, routing becomes dif-

A. Roy (✉) · T. Deb
Department of Information Technology, Tripura University, Agartala 799022, Tripura, India
e-mail: alakroy@tripurauniv.in
URL: http://alakroy.webs.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_4

ficult in MANETs. Every mobile node is able to travel in any direction randomly. This results in changes of routes between source and destination rapidly and unpredictably. In this manner, each portable node goes about as a switch and host in the meantime in MANETs. Routing is the center issue in MANETs for sending information from source to destination. Communication in MANETs requires efficient and prominent routing protocols between two nodes. In order to provide network connectivity in MANETs, it uses dynamic routing in case of a static infrastructure. The objectives of any routing protocols in MANETs are to route data packet among the mobile nodes effectively and efficiently. Two main activities of MANETs are as follows: First is determining the optimal route and second is the transferring of information to the destination node. One of the principle points of interest of MANETs is that it gives access to data paying little mind to geographic territory and it can be set up at wherever. Use of MANETs incorporates the areas where no communication infrastructure, Civilian environments, mine site operation, emergency operations, military environments, personal area networking like cell phone, laptop even in rescue operation. Plenteous routing protocol has been proposed for MANETs. The aim of this paper is to present a comparative performance analysis of few routing protocols available in recent literature for MANETs with their characteristics, advantages and disadvantages. The rest of the paper is organized as follows: section "Background Details" gives background details of MANETs; section "Related Work" includes the related works and existing research work. Section "Routing Protocol in MANETs" contains literature survey of different types of proposed routing protocols that are listed in two parts: proactive and reactive. Section "Performance Evaluation" includes performance evaluation metrics and discussion of simulation results. The paper concludes with future research direction in section "Conclusion".

Background Details

In some of the MANETs, routing protocols use beacon technique as shown in Fig. 4.1, where each node periodically generates beacon or hello message. Beacon is used to announce the existence of the nodes. In MANETs, routing information are maintained in route cache and route table. The routing table includes the details of routes to all of the networks where the router has learned about. It maintains this table whether there is traffic flowing to those networks or not. The routing table is only concerned with the route to a destination, so it does not keep track of source addresses. On the other hand, the routing cache stores recently used routing entries and a flow is defined by the source and destination IP addresses. In some cases, format of a beacon may contain the following: (i) Validity Time (8 bits): It is the time where a node is active for receiving messages, (ii) Address Block: It contains the used addresses over which the Hello message is transmitted, (iii) Interval Time (8 bits): It specifies the interval at which Hello messages are being generated by the originator node, and (iv) One or more Address Blocks: It consists of addresses of non-neighbor nodes (Table 4.1).

Related Work

In this section, we intend to detail various presented research works done on MANETs routing protocol. In [2], the performances of AODV, DSR, and ZRP are evaluated. They analyzed these routings with different pause times and TTL base hop counts. They concluded that AODV performs better in the case of EED. In the case of packet delivery fraction, AODV and DSR perform better than ZRP. DSR performs best with less a number of hops in comparing TTL-based hop count. In [3], performances are analyzed for AODV and DSR protocols using different network loads, network sizes, and mobilities. In this paper, it concluded that DSR performs better in constrained situation. In [4], performances are compared for three routing protocols: AODV, DSDV, and DSR. They analyzed these three routing protocols with node mobility, various interval times, and packet size. With expanding packet volume, PDR and throughput are decreased in AODV. DSR protocol performs better at less packet size but routing load is high. They concluded that the performance of DSDV is always less than DSR and AODV. In [5], performances are compared for AODV, AOMDV, LAR, and LAMR protocols with varying pause time and speed. Their simulation result shows that best performance is given by AOMDV in order to PDR and EED. In [6], in this paper, performances are analyzed for DSDV, DSR, and Eff-DSDV routing protocol under various numbers of source and node speed. They concluded that Eff-DSDV performs better with varying speed of mobile nodes. In [7], it is concluded that reactive protocol gives better result in case of PDR and AEED under both TCP and CBR traffic patterns. Experiment is carried out by Network Simulator 2. In [8], performance is evaluated for DSDV and AODV routing protocols with different numbers of source and various pause times. They concluded that AODV is suitable when communication is done under UDP protocol. But in certain case, AEED of DSDV is less than AODV. In [9], performance of AODV, DSR, DSDV, and OLSR is evaluated under node mobility and number of nodes in the network. They concluded that for much less volume of node in addition to low range of mobility DSDVs, effectiveness can be more. From best of our knowledge, no research work is done on performance evaluation under data rate. Data rate is the essential component of data communication. In [2–9], no comparison is done under static node characteristics. Drawbacks of routing protocols are not included in the existing work. A detailed comparison of various routing protocols is listed in Table 4.2. That is why, we have evaluated performance analysis under data rate and fixed node position using Network Simulator 2. In this paper, we have listed the various drawbacks and advantages of various routing protocols in MANETs.

Routing Protocol in MANETs

In recent literature, several routing protocols are available in MANETs, which are generally applied based on their network circumstances. The main theme behind

every protocol is to expand the throughput while minimizing the packet loss, energy consumption by sensor node, and decreasing network burden. In order to compare and analyze the available routing protocols in MANETs, a classification among them is required, which is shown in Fig. 4.2.

Reactive Routing Protocol

In case of dynamic changes of topologies, traditional table-driven routing protocols are not efficient for communication in MANETs. Thus, requirements come to develop on-demand routing protocols. Reactive protocol discovers opposite routes when a packet has to be forwarded. It maintains the currently involved routes. Therefore, routes are built on-demand when a node is flooded with route request and route reply is received. It needs to be determined by the route just before the packet sending begins and thus results in a delay of the first data packet to receive. These offices low overhead and diminish load on the network. Certain reactive routing protocols are Dynamic Source Routing (DSR), Ad Hoc On-Demand Distance Vector (AODV), Location-Aided Routing (LAR), and Associativity-Based Routing (ABR). Among a few receptive routing protocols, we have selected Ad Hoc On-Demand Distance Vector (AODV) and Dynamic Source Routing Protocol (DSR) for execution assessment.

Ad hoc On-Demand Distance Vector (AODV) [9] is a famous on-demand reactive routing protocol. AODV finds out route only before data transfer begins between the peers. The route finding process begins when there are no pre-existing routes between the communication peer in the routing table. Finding routes in the source node is flooded with a route request (RREQ) indicating the destination. This process proceeds until it reaches the receiver or lifespan has finished. Once RREQ is received, receiver sends back route reply (RREP) by specifying a reverse path to the source that initiates the RREQ [3, 5, 6, 8]. The RREQ packet incorporates source arrangement number and last destination succession number known to source. An entrance of forward way that is not utilized for a dynamic route time interim is cleansed from the routing table. For route support process, source node can reinitiate a route disclosure process when it changes its location. In the event that any middle node of this system moves inside a specific route, the neighbor of the floated node can distinguish the connection disappointment and sends a connection disappointment notice to its upstream neighbor. This methodology proceeds until the disappointment notice or data achieves the source node. In the view of achieved data, the source may choose to re-launch the route disclosure stage. Favorable circumstances include (i) the playing point of AODV is that it is suitable to exceptionally element systems. (ii) It underpins both multicast and unicast packet transmissions notwithstanding for nodes in steady development. (iii) It additionally reacts rapidly to the topological change that influences the dynamic routes, and (iv) AODV takes less time to set up association between conveying nodes. However, it has few drawbacks as follows: (i) Nodes might confront extensive deferrals amid route discovery, and connection

disappointment may launch another route disclosure, which leads to delay and expends more transfer speed as the span of the system increases and (ii) heavy control overhead.

Destination Source Routing (DSR) [10] protocol is an on-demand reactive unicast source routing. In DSR, source nodes have the capacity to locate the precise succession by which a packet engenders. Nodes are obliged to keep up route stores in DSR. The route store is overhauled when a route is known for a specific section in the route reserve. In this, routing is done utilizing two stages: route disclosure and route support. At the point, a source must send out new packet to a vacation spot, and first counsels its route store to figure out if it thinks about any route to the destination or not. If not, it launches a route demand telecast. This solicitation incorporates the source address, destination location, and an exceptional ID number. In the middle of the road, node checks its own route store to discover whether it thinks about the destination or not. In the event that it does not think about the destination, the middle of the road node again advances the packet and inevitably this acquires the communicating end. A route answer is sent by the destination or by any of the middle of the road nodes when it thinks about how to achieve the destination. Diverse packet may have distinctive routes, notwithstanding having the same source and destination. In this way, the name is source routing. Favorable circumstances of DSR protocols include (i) a playing point is that every node can keep different routes, which implies that the sender can check its route reserve for a legitimate route before starting route disclosure, and if a substantial route is observed there is no requirement for route revelation. (ii) Another playing point of DSR is that it does not require any occasional beaconing; consequently, node may go to sleep mode to save energy. This additionally spares a measure of data transmission in the system and (iii) One of the principle focal points is no compelling reason to continue routing table in order to route given information packet. Meanwhile, some drawbacks are as follows: (i) routing overheads are proportionally increased with increase in distance between source and destination nodes. (ii) It requires altogether more preparing assets than most different protocols. To get the routing data, every node must invest packet of energy to process any control information. (iii) Using stale cached route by intermediate nodes, other nodes caches may be polluted [4, 7] and (iv) packet header size grows with route length and RREQ flooding.

Temporarily Ordered Routing Algorithm (TORA) [11] is sorted as hybrid routing protocol. TORA is a blend of reactive and proactive upgrades. It is an exceedingly versatile, versatile, and circle-free routing protocol; subsequently, it can work in a multi-bounce system. TORA is in view of the idea of “Connection Reversal”. To create connection between nodes in TORA, Directional Acyclic Graph (DAG) is utilized. For this situation, a route disclosure inquiry is shown and engendered all through the system until it achieves the destination or a node that has data about how to achieve the destination. TORA characterizes a parameter, termed stature. The parameter tallness is a measure of the separation of the reacting node’s separation up to the obliged receiving node. In the route revelation stage, this parameter comes back to the questioning node. As the inquiry reaction or result engenders back, in the middle of the road, every node redesigns its TORA table with the route and tall-

ness to the end node. The source node then uses the tallness to choose the best route toward the destination. This protocol has an intriguing property that it habitually picks the most advantageous route, as opposed to the briefest route. For every one of these endeavors, TORA tries to minimize the routing administration movement overhead. In TORA, numerous routes exist in the middle of each source and destination. Therefore, connect disappointment or node disappointment can be recognized and understood rapidly. Favorable circumstances include (i) advantage of TORA is that it additionally backs multicasting and (ii) numerous routes among just about any source desired destination match are upheld by this protocol. Consequently, disappointment of any of the nodes is immediately determined. Meanwhile, few drawbacks include the following: (i) the calculation of route might likewise create interim invalid route and (ii) it depends on synchronization of clock between the nodes in the network. Therefore, Internet encapsulation protocol is required at immediate down layer of TORA.

Location-Aided Routing (LAR) [12] is an on-interest source routing. LAR protocols constrain the control overhead of route revelation by using the area data of MANETs. To acquire the area data, LAR utilizes Global Positioning System (GPS). With the accessibility of GPS, it is simple for every host to know its physical area. At whatever point, a source needs to correspond with destination, instead of flooding the routing information, source node configures an expected zone for destination node. Depending on the availability of location information, a request zone may be formed. In this request zone, all nodes are able to forward route discovery packets. Thus, expected zone is smaller in size than the request zone. Favorable circumstances are as follows: this approach (i) decreases routing overhead, (ii) reduces communication complexity, and (iii) decreases overhead of route discovery. Disadvantage is that route errors are generated due to route break.

Associativity-based routing (ABR) [13] is another source launched routing protocol, which additionally utilizes a question–answer procedure to focus routes to the obliged destinations. On the other hand, in ABR route determination is principally in light of steadiness. To pick stable course, every hub keeps up an associativity tick with their neighbors, and the associations with higher associativity tick are picked in slant to the once with lower associativity tick. Every node often produces reference node which indicates that it is present. After getting the sign information, a new neighbor node renovates associativity table. For any reference level, the associativity beats the accepting node using beaconing node. An increased opinion regarding associativity beat for virtually any certain beaconing node guarantees the node to be relatively static. Associativity beat is reset to zero any time; any nearby node moves outside the area of a few other nodes. This approach gives stable route including few drawbacks as follows: (i) it obliges intermittent beaconing to focus the quality of associativity of the connections which requires extra power utilization. (ii) Another detriment is that it does not keep up different routes or a route reserve, which implies that backup ways to go will not be quickly accessible, and a route disclosure will be obliged utilizing connection disappointment and (iii) Only suitable for small or medium network.

Dube et al. designed Signal Stability-Based Adaptive Routing (SSBR) [14] protocol. The main routing mechanism of SSBR is the location and signal stability. In case of route discovery, signal strength of nearby node plays a vital role. Static Routing Protocol (SRP) and Dynamic Routing Protocol (DRP) are two sub-protocols of SSBR. In this protocol, communication is done based on the best signal strength route. Hence, packet delivery and throughput increase.

Ad Hoc QoS On-Demand Routing (AQOR) [15] was proposed by Xue and Ganz. It is a reactive protocol enabling the QoS support in respect of end-to-end delay. Resource reservation is done according to estimated delay and bandwidth. AQOR provides QoS support in MANETs by enabling hop-by-hop routing. Advantages include low control overhead, limited flooding mechanism, and reduced route discovery overhead.

Source Routing for Roofnet (SrcRR) is proposed by Aguayo et al. The main idea behind the design techniques of SrcRR [16] is to utilize the high-throughput route for routing data packet. SrcRR was implemented by Roofnet test bed [17]. The design principle of SrcRR was inspired by DSR. However, finding best path for various routes between source and destination SrcRR uses the shortest path.

Interference-Aware Load-Balancing Routing (IALBR) [18] is based on AODV protocol. In this routing algorithm, routing metric consists of the sum of traffic load which helps to find the routes with fewer loads from sender to receiver in the network. Each intermediate node calculates the load on receiving RREQ and a new entry in the routing table if it does not exist. In IALBR, transmission is done through the lowest load route instead of shortest path like AODV.

Proactive Routing Protocol

A proactive routing protocol is likewise called “table-driven” routing protocol. Keeping in mind the end goal to hold the security, in this routing protocol, every node in a system consists of routing tables that are overhauled routinely. Whenever there is an adjustment in the system topology, a message is telecasted by each node to the whole system. In any case, it brings about extra overhead cost because of keeping up forward data, and thus throughput of the system may be influenced; however, it gives the genuine data to the accessibility of the system. In this way, a source node can perceive routing way in a flash in the event that they obliged at whatever time quickly in the event that they obliged whenever. This sort of protocol is not effective for bigger systems, as they need to keep up every node passage in the routing table. In this manner, it builds routing overhead and utilization of transmission capacity. Couples of proactive routing protocols are Global State Routing (GSR), Destination-Sequenced Distance Vector (DSDV), Wireless Routing Protocol (WRP), and Hierarchical State Routing (HSR). Among these, we have used Destination-Sequenced Distance Vector (DSDV) for execution assessment.

Destination-Sequenced Distance Vector (DSDV) [19] is a table-driven routing plan. DSDV is in light of the Bellman–Ford calculation. It was produced by

C. Perkins and P. Bhagwat in 1994. In DSDV, every node keeps up a route table comprising next jump address for every destination address, which redesigns continually and intermittently, not on interest. Each node intermittently telecasts their routing table data with neighboring nodes. Every node keeps up expense metric for the way to destination in routing table. Biggest arrangement number is utilized to mark the route. DSDV uses arrangement number began by destination. In this manner, circling issue brought about by stale routing is lessened which implies that it gives a solitary way to a destination and no dormancy because of route revelation. The routing redesigns in two ways: one is known as a "full dump" and incremental. If there arise an occurrence of full dump, the whole routing table of every node is transferred to the nearer nodes and the incremental packet conveys just the data changed following the last full dump. It keeps up two tables. The issue is that it obliges bidirectional connection for correspondence. It gives a solitary way to a destination, that is, it gives circle-free routes. Few disadvantages are as follows: (i) Due to the necessity of the occasional redesigns, DSDV acquaints a lot of overhead with the system and (ii) large segment of the system transfer speed is utilized as a part of the redesigning.

Wireless Routing Protocol (WRP) [20] is likewise a circle-free proactive protocol. WRP requires every node to keep up four routing tables: connection expense table, a separation table, a routing table, and a message retransmission list (MRL). It presents memory overhead. In this protocol, every node creates hello message or reference points and these welcome messages are traded between neighboring nodes. Like DSDV, it also provides loop-free routes. Few disadvantages are (i) every node keeps four types of routing table, and it introduces memory overhead. (ii) It uses hello message to ensure connectivity between the nodes and (iii) hello message consumes bandwidth and more energy.

In Optimized Link State Routing (OLSR) [21], every node keeps three tables: initial one for routing, second for neighbor nodes, and third for topology. It performs hop-by-hop routing that implies every node in the system that uses its latest data to send a packet. It is in light of the idea of multipoint relays (MPRs). It just proclaims the arrangement of connections with its neighbors that are its "multipoint transfers" as opposed to announcing its whole connections to all nodes in the system. Along these lines, OLSR [21] decreases the control overhead. MRPs nodes cover all the nodes. To choose the MPRs nodes, every node intermittently telecasts a rundown of its one bounce neighbor utilizing hello messages. Then, every node selects a subset of nodes that cover every single other node. Just MRPs of a node can retransmit its telecast messages and different nodes can read and, however, cannot retransmit it. Accordingly, this protocol diminishes the quantity of retransmissions in a telecast methodology. Every node decides an ideal route to each destination utilizing its topology, and stores this data in a routing table. Favorable circumstances include (i) OLSR lessens control overhead and associations. (ii) This protocol lessens the quantity of retransmissions in a broadcast procedure and (iii) in order to manage routing, it does not require central administrative system. While other disadvantages are (i) OLSR requires more time to re-finding a broken connection. (ii) For discovering an alternate route, OLSR requires more processing power and (iii) OLSR consumes more bandwidth.

In Global State Routing (GSR) [22], each node sustains global data of the entire topology, which depends on link state routing. Link state information is periodically exchanged with local neighbors, in order to detect topology changes. Therefore, the smaller sequence number of the table entries is replaced by larger sequence number of that table. The topology map maintained at each node is used to calculate the shortest path. This protocol reduces control overhead, while increasing network size is the main disadvantage of GSR.

Fisheye State Routing (FSR) [23] is a proactive routing protocol, in which the data of each node is gathered from the neighboring nodes and afterward the routing table is figured. It is in view of the connection state routing and is a change of Global State Routing (GSR). Favorable circumstances are (i) it trades halfway routing redesign data with neighbors just. Henceforth, FSR fundamentally decreases the devoured transmission capacity and (ii) it will not lead to any control information for connection disappointment. Accordingly, no change in the routing table will happen regardless of the possibility that there is any connection disappointment.

Source-Tree Adaptive Routing (STAR) [24] is developed by Garcia-Luna-Aceves and Spohn. In this routing, a source tree is established by each node depending on the links to all destination nodes. Two novel advantages provided by STAR are Optimum Routing (ORA) which detects shortest path for routing; meanwhile, packet overhead is maintained by Least Overhead Routing (LORA). Later on, STAR protocol is extended as SOAR [25].

OLSR with Quality of Service (QOLSR) [26] protocol was proposed by Fonseca and Munarctto. QoS parameters are included in terms of delay and bandwidth is added with the base routing protocol OLSR. Further, three more advanced protocols developed based on the QOLSR are QOLSR1, QOLSR2, and QOLSR3. The route consisting of lowest delay is preferred for routing packets.

Hybrid Routing Protocols

Hybrid routing protocols combine the features of reactive protocols and proactive.

Zone Routing Protocol (ZRP) [27] is a hybrid routing protocol. In order to achieve maximum efficiency and scalability, ZRP fuses the benefits of table-driven and on-demand routing protocol. This feature also helps to discover the local neighborhood of every node and communication among the neighborhoods, respectively. In ZRP, each node may be able to set up its own zone side and it is defined as number of hops to the zone perimeter. Each node may be in multiple overlapping zones of variable size. ZRP consists of the following components: (i) The reactive Inter-zone Routing Protocol (IERP), (ii) the proactive Inter-zone Routing Protocol (IARP), and (iii) Broadcast Resolution Protocol (BRP). These three components work independently of the others to provide efficient routing features to ZRP. Details of ZRP can be found in. Favorable circumstances are as follows: (i) it reduces the communication overhead, (ii) it is suitable for large network and diverse mobility pattern, and (iii) it

provides loop-free routes; while drawbacks are as follows: (i) the main disadvantage is increasing complexity and (ii) its performance depends on zone radius.

Pei et al. designed Landmark Ad Hoc Routing (LANMAR) [28] which creates a subset of nodes to move together. In each subset, a landmark node is selected, like FSR. In this routing protocol, routing table consists of nodes within the range and landmark node. Depending on landmark, a distance vector is calculated and attached to every updated packet.

Relative Distance Micro-Discovery Ad Hoc Routing (RDMAR) [29] is developed by Aggelre and Tafazolli. In this protocol based on the relative distance from source to destination, maximum hops are calculated. Using the maximum number of broadcast messages is limited. RDMAR routing table consists of time stamp, flag next hop, and relative distance field, respectively. Transmitted packets are kept in a buffer until a received reply comes from destination.

Wang et al. designed the novel hybrid protocol known as Hybrid Ant Colony Optimization (HOPNET) [30] on the basis of Ant Colony Optimization (ACO) and zone routing. This protocol combines the features of ZRP and DSR routing protocols with ACO-based schemes. The idea behind this distinct protocol is taken from the ant hopping from one place to another place. Two routing tables as inter-zone and intra-zone are maintained to handle link failure. This protocol increases delivery of packet.

Other Routing Protocols

Location-Aware Routing Protocol: In this category, routing is done based on the global positioning system. It is assumed that each node in the network is capable to acquire geographical location of other nodes in the network. Location-Aided Routing (LAR) [31, 32] is designed by Ko and Vaidya in which route discovery is done by using the location information. LAR uses GPS to decrease route discovery overhead. Basagni et al. designed Distance Routing Effect Algorithm for Mobility (DREAM) [33] protocol which also uses GPS. However, it combines the benefits of proactive and reactive routing protocols. DREAM optimized the flooding for RREQ by limiting the forwarding neighbor. DREAM is a robust and loop-free protocol. Both the bandwidth and energy are efficiently utilized by DREAM protocol.

Multipath Routing Protocol: Multipath routing uses multiple paths instead of single route from source to destination for packet delivery; as a result, bandwidth is used more efficiently. Hence, packet delivery ratio is increased.

Hierarchical Routing Protocol: In hierarchical routing techniques, nodes are grouped into cluster. Each cluster must have at least one Cluster Head (CH). CH is responsible for communication with other CH with the help of gateway node. Hierarchical State Routing (HSR) [34] protocol is introduced by Iwata et al. It is developed based on the multilayer clustering. The main theme of design of HSR routing is to reduce the flooding of various control information using the local information of CH. Hierarchical addressing techniques are used for routing in HSR.

Multicast Routing Protocol: Multicast routing is the process of concurrently transmitting the data from one source to multiple destinations. The novel technique is developed by Steve Deering. The extended multicast version of AODV routing protocol is Multicast Ad Hoc On-Demand Distance Vector Routing Protocol (MAODV) [35]. Route discovery is initiated only before beginning data transmission. The advanced version of multicast routing is geographical multicast routing protocol [36]. In geocast, node used geographical location information such as GPS. Earlier, survey work on geocast is explained in [36] by C. Maihoter. An and Papavassiliou developed Direction Guided Routing (DGR) [37]. For routing packets, clusters are formed and CH is selected dynamically. It reduces the control overhead.
Power-Aware Routing Protocol: In WSNs, battery power of node is a crucial problem. In power-aware routing nodes, mobility and routing depend on the energy remaining. Maximizing the lifetime of each node with the constrained energy sources the motive of this protocol in this class.

Performance Evaluation

To evaluate the comparative performance analysis of three diverse routing protocols AODV, DSR, and DSDV, we have used Network Simulator 2 [38] for MANETs under constrained condition where source and destination nodes' positions are fixed. We have used three performance matrixes for comparison evaluation which comprises average end-to-end delay, average throughput, and packet delivery fraction (pdr * 100).

Simulation Setup and Parameters

For simulation, we have assigned packet size of 512 bytes at varying rates of 4, 8, 12, 15, 20, 25, and 30 packet/s. 25 uniform nodes are taken for simulation with simulation time of 100 ms. 12 Constant Bit Rate (CBR) traffic sources are used. Parameters used for simulation are detailed in Fig. 4.1.

Simulation Results

Metrics used for performance comparison include the following:
Average End-to-End Delay: End-to-end delay is defined as the delay experienced by a packet generated by (CBR) source till it is received at the destination node. Average end-to-end delays are measured by the ratio of end-to-end delay and the received packets, including almost all possible delays during route discovery, propagation time, and queuing at interface queue. In our experiment, we have found that

Parameters	Values
Simulation Time (in seconds)	100
Simulation Area	500 X 500
CBR Source	12
Numbers of Nodes	25
Packet Size (in bytes)	512
Routing Protocols	AODV, DSR, DSDV
Data Rates (in packets/seconds)	4,8,12,15,20,25,30
Traffic Type	CBR

Fig. 4.1 Simulation parameters

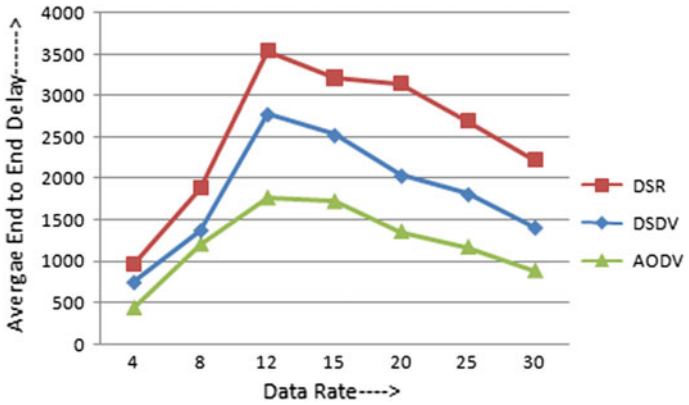


Fig. 4.2 Average end-to-end delay of routing protocols

with the increasing data rate AODV outperforms over DSR and DSDV as shown in Fig. 4.2. Average end-to-end delay is inversely proportional to the data rate. The main reason behind higher end-to-end delay in DSR is that it is on-demand source routing protocol, where route is discovered before communication begins between source and destination and there is a route finding mechanism that occurs every time. It also has to carry a large overhead each time, thus the higher delay. Another reason may be that the path between communicating peer may not be the shortest path. Thus, results in more delay occur in delivery. The whole routing table needs to be updated periodically in DSDV, which leads more delay in delivery compared to AODV. During transmission, packets must stay in buffer for more time in DSDV which causes more delay. Hop-by-hop mechanism in AODV helps to reduce the EED.

Packet Delivery Fraction (PDR): It is determined as total number of packet obtained on desired destination node dividing by total number of packet generated by CBR source. Packet delivery fraction is calculated as PDR * 100%. DSDV has slightly higher PDF compared to AODV and DSR as shown in Fig. 4.3, which may

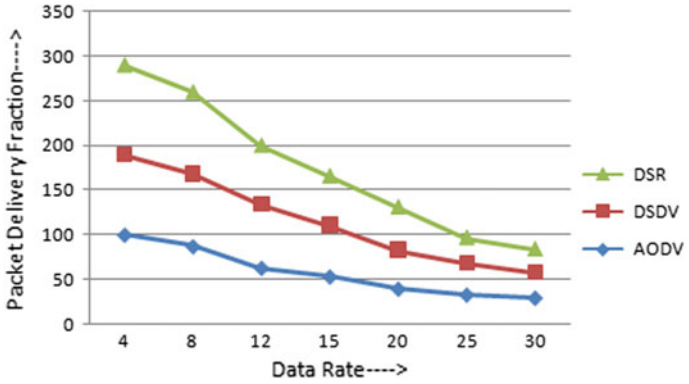


Fig. 4.3 Packet delivery fraction of routing protocols

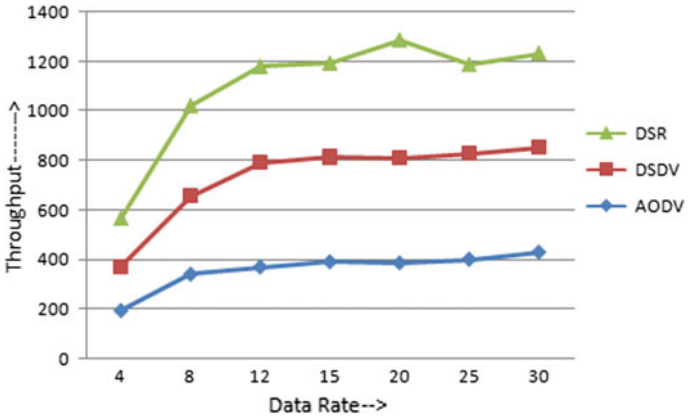


Fig. 4.4 Throughput of routing protocols

be due to being a table-driven protocol and slightly more reliable. With the increase in data rate, PDF is decreasing, and in case of DSR, it shows a drastic decrease when data rate is in between 12 and 20.

Average Throughput: Average throughput is average number of packet obtained at the receiver end in per unit time. In our experiment, we have found that the throughput of DSDV outperforms over DSR and AODV with increasing data rate as shown in Fig. 4.4. The throughput of DSR oscillates because it uses route cache; and stored route may be stale after some time. Stale cache routing in DSR leads to fluctuation in average throughput.

Table 4.1 Classification of routing protocols for MANETs

Proactive	DSDV, R-DSDV, OLSR, HOLSR, CGSR, WRP, GSR, STAR, QOLSR, LCA, MMRP, LQSR, WAR, DFR, DBF, AWDS, IARP,TBRPF,
Reactive	DSR,AODV,TORA,LMR,ABR,SSBR,AQOR,ARA,ROAM,FORP,DAR,FDG,LLR,QMRB,AODV-ABR,LBAQ,LSR, SWORP,RPR,GPR,SLR,LDR,RBR,DBR2P,LBCR,IALBR
Hybrid	ZRP, FSR, LANMAR, DST, ZHLS, RDMAR, SLURP, DDR, A4LP, HOPNET, LRHR, FZRP, ANSI, HSLS, HRPLS HARP
Location Aware	LAR, DREAM, GPSR, DRM, ALARM, REGR, LAKER, MORA, OGRP, SOLAR, LBLSP, GLR, MER,TBR
Multi-path	CHAMP, AOMDV, SMR, NTBR, TMRP, SMORT, SecMR, REEF, MuSeQoR, MDR, DYMO, OMR
Hierarchical	HSR, CEDAR, H-LANMAR, DART, ATR
Multicast	DCMP, ADMR, AMRoute, QMRPCAH, Fireworks, PPMA, ALMA, AQM, CBM, DDM, ROMANT, EraMobile, ABAM, ADMR, AMRIS, BEMRP, MOLSR, AMRoute, CBM, FGMP, DGR, GAMER, GeoGRID, GeoTORA, GPSAL, GRLI, GDSTR,BVGF
Power-aware	DEAR, Scott & Bombos, MEHDSR, PARO, DSRPA, EADSR, ISAIAH

Table 4.2 A comparison of routing protocols for UWSNs

Parameters	**DSDV**	**DSR**	**AODV**	**LAR**	**TORA**	**ZRP**	**ABR**	**OLSR**	**WRP**	**FSR**
Protocol type	Table driven	On-demand	On-demand	On-demand	Link-reversal	Hybrid, DV and LS	On-demand	Proactive	Table Driven	Proactive
Loop Free	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes, but not instantaneous	Yes
Multiple Route	No	Yes	No	Yes	No	No	No	No	Yes	Yes
Routing Over head	Medium	Low	High	Low	High	Medium	Low	Medium	High	High
Hop count	Medium	Very high	Normal	High	Medium	Medium	Less	Less	Medium	Less
Error Message	More	Less	More	Less		Medium			Less	
Security	No	No	No	No	No	No	No	No	No	No
Periodic broadcast	Yes	No	Possible	Possible	Possible	No	NA	Yes	NA	NA
Congestion	High	Low	Medium	Medium	Low	Medium	Medium	Medium	Medium	Medium
Multicast Capability	No	No	Yes	Yes	No	Partly	No	Yes	No	No
Routes maintains in	Route table	Route cache	Route table	Routing table	Route table	Intrazone and Interzone tables	Route table	Route table	Route cache	Route Table
QoS support	No	No	No	No	No	No	No	Yes	No	Yes
Use sequence number	Yes	No	Yes	No	No	No	No	Yes	Yes	Yes
Routing metric	Shortest distance	Shortest path	Shortest path	Shortest path	Shortest path	Local Shortest path	Associativity and Shortest Path	Shortest distance	Shortest path	Shortest path

Conclusion

Routing is a crucial problem for collecting and delivering data packet efficiently in the network. It is a significant element for data transmission in MANETs. This paper provides a comprehensive survey of the existing routing protocols for MANETs, which are broadly classified as proactive, reactive, and hybrid. Through simulation studies, we performed a comparative study of AODV, DSDV, and DSR protocols in respect of end-to-end delay, PDR, and throughput under various data rates. The simulation results reveal that DSR can be used for constrained network situation where better throughput and PDR are crucial requirements, whereas AODV can be used for low-delay network. In the future work, we will briefly explain advantages and disadvantages of other available routing protocols and also we will propose a

new routing protocol which will solve the problems of existing routing protocols for MANETs. Our proposed protocol should be able to utilize and increase throughput and PDF, and efficiently decrease end-to-end delay. This paper will help research community to design a new routing protocol.

References

1. Cordeiro, D. M and Agrawal, D. P.,: Mobile ad hoc networking, Center for Distributed and Mobile Computing, ECECS, University of Cincinnati, (2002).
2. Mittal, S., and Kaur, P.: Performance Comparison of AODV, DSR and ZRP Routing Protocols in MANET's, Advances in Computing, Control, & Telecommunication Technologies, ACT'09, International Conference, IEEE (2009).
3. R. Misra and C. R. Mandal, Performance Comparison of AODV/DSR On-demand Routing Protocols for Ad Hoc Networks in Constrained Situation, (IEEE 2005).
4. Tuteja, A., Gujral, R. and Thalia, S.: Comparative performance analysis of DSDV, AODV and DSR routing protocols in MANET using NS2, In Advances in Computer Engineering (ACE), International Conference, pp. 330–333. IEEE, (June 2010).
5. Trung, H. D., Watit, B., and Duc, P. M.: Performance evaluation and comparison of different ad hoc routing protocols, Computer Communications, vol. 30(11), pp. 2478–2496. (2007).
6. Khan, K., Zaman, R. U., Reddy, K. A., and Harsha, T. S.: An efficient DSDV routing protocol for wireless mobile ad hoc networks and its performance comparison, In Computer Modeling and Simulation, EMS'08. Second UKSIM European Symposium, pp. 506–511. IEEE (September 2008).
7. Ramesh, V., Subbaiah, D. P., Rao, N. K., and Raju, M. J.: Performance Comparison and Analysis of DSDV and AODV for MANET. International Journal on Computer Science and Engineering, vol. 2(02), pp. 183–188. (2010).
8. Ade, S. A., and Tijare, P. A.: Performance comparison of AODV, DSDV, OLSR and DSR routing protocols in mobile ad hoc networks, International Journal of Information Technology and Knowledge Management, vol. 2(2), pp. 545–548. (2010).
9. Perkins, E., and Royer, E. M.: Ad-hoc on-demand distance vector routing, In Mobile Computing Systems and Applications, Proceedings. WMCSA'99. Second IEEE Workshop, pp. 90–100. IEE (1999).
10. Johnson, D. B., and Maltz, D. A.: Dynamic source routing in ad hoc wireless networks, In Mobile computing, pp. 153–181. Springer US (1996).
11. Park, V. D., and Corson, M. S.: A highly adaptive distributed routing algorithm for mobile wireless networks, In INFOCOM'97, Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies, Driving the Information Revolution., Proceedings IEEE, Vol. 3, pp. 1405–1413. IEEE (April 1997).
12. YBae Ko, and Nitin, H. V.: Location Aided Routing (LAR) in mobile ad hoc networks, wireless networks vol. 6(4), pp. 307–321. (2000).
13. Toh, C. K.: Associativity-based routing for ad hoc mobile networks, Wireless Personal Communications 4, no. 2, pp. 103–139. (1997).
14. Dube, R., Rais, C.D., Wang, K.Y., Tripathi, S.K.: Signal stability-based adaptive routing (SSA) for ad hoc mobile networks, IEEE Personal Communications Magazine, pp. 3645. IEEE (1997).
15. Xue, Q., Ganz, A.: Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks, Journal of Parallel Distributed Computing, vol. 63(2), pp. 154165. (2003).
16. Aguayo, D., Bicket, R., Morris, SrcRR: A High-Throughput Routing Protocol for 802.11 Mesh Networks, (2003).
17. Chambers, B.A.: The Grid Roofnet: a Rooftop Ad Hoc Wireless Network, Massachusetts Institute of Technology (May 2002).

18. Feng, J., Xia, Zhou, R. H.: Interference-aware load balanced routing in wireless mesh networks, in: Proceedings of the International Conference Wireless Communications, Networking and Mobile Computing (WiCom07), pp. 17301734. (September 2007).
19. Perkins, C.E., and Bhagwat, P.: Highly Dynamic Destination Sequence-Vector Routing (DSDV) for Mobile Computers, Computer Communication Review, vol. 24(4), pp. 234–244. (1994).
20. Murthy, Shree, and Jose Joaquin Garcia-Luna-Aceves. "An efficient routing protocol for wireless networks", Mobile Networks and Applications, vol. 1, no. 2, pp. 183–197. (1996).
21. Clausen, Thomas, et al., "Optimized link state routing protocol (OLSR)", (2003).
22. T-Wei Chen, and Gerla, M.: "Global state routing: A new routing scheme for ad-hoc wireless networks." In Communications, ICC 98. Conference Record. International Conference, vol. 1, pp. 171–175. IEEE (1998).
23. Pei, Gerla, G., M., & Chen, T. W.: Fisheye state routing: A routing scheme for ad hoc wireless networks. In Communications, 2000. ICC 2000. IEEE International Conference, Vol. 1, pp. 70–74. IEEE (2000).
24. Garcia-Luna-Aceves, J.J., Spohn, M.: Source-tree routing in wireless networks, in: Proceedings of IEEE ICNP, pp. 273282. (OctoberNovember 1999).
25. Roy S., Garcia-Luna-Aceves, J. J.: Using minimal source trees for on demand routing in ad hoc networks, in: Proceedings of IEEE INFOCOM, pp. 11721181. IEEE (April 2001).
26. Munaretto, A., Fonseca, M.: Routing and quality of service support for mobile ad hoc networks, Computer Networks, vol. 51(11), pp. 31423156. (2007).
27. Haas, Z. J., Pearlman, M. R., and Samar, P.: The zone routing protocol (ZRP) for ad hoc networks, draft-ietf-manet-zone-zrp-04, Txt, (2002).
28. Pei G., Gerla, M., and Hong X.: LANMAR: Landmark routing for large scale wireless ad hoc networks with group mobility, in: Proceedings of ACM MobiHoc, pp. 1118. (August 2000).
29. Aggelou G. N, Tafazolli, R.,: RDMAR: A bandwidth-efficient routing protocol for mobile ad hoc networks, in: Proceedings of ACM WOWMOM, pp. 2633. (August 1999).
30. Wang J., Osagie, E., Thulasiraman, P., Thulasiram R.: Hopnet: a hybrid ant colony optimization routing algorithm for mobile ad hoc network, Ad Hoc Networks, vol. 7(4), pp. 690705. (2009).
31. Ko, Y., Vaidya, N.: Geocasting in mobile ad hoc networks: Locationbased multicast algorithms, Technical report, TR-98-018, Texas AM University, (September 1998).
32. Ko Y., Vaidya N.: Location-Aided Routing (LAR) in mobile ad hoc networks, in: Proceedings of ACM MobiCom, pp. 6675. (October 1998).
33. Basagni S., Chlamtac I., Syrotiuk V.R., Woodward B.A.: A distance routing effect algorithm for mobility (DREAM), in: Proceedings of the ACM MOBICOM, pp. 7684, (October 1998).
34. Iwata A., Chiang, C., Pei, G., Gerla, M., Chen, T.: Scalable routing strategies for ad hoc wireless networks, IEEE Journal on Selected Areas in Communications 17 (8), pp. 13691379. IEEE (1999).
35. Elizabeth Royer M. and Charles E. Perkins: Multicast Operation of the Ad-hoc On-Demand Distance Vector Routing Protocol, In Proc. of the 5th annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), pp. 207218. IEEE (August 1999).
36. Maihfer C.: A survey on geocast routing protocols, IEEE Communications Surveys and Tutorials 6 (2), pp. 3242. IEEE (2004). An B., Papavassiliou S., Geomulticast: architectures and protocolsfor mobile ad hoc wireless networks, Journal of Parallel Distributed Computing 63(2), pp.182195, 2003.
37. An B., Papavassiliou S., Geomulticast: architectures and protocolsfor mobile ad hoc wireless networks, Journal of Parallel Distributed Computing 63(2), pp. 182195, 2003.
38. Network Simulator 2, available via http://www.isi.edu/nsnam/ns.

Chapter 5
Secure Adhoc On-Demand Multipath Distance Vector Routing in MANET

V. Vinoth Kumar and S. Ramamoorthy

Abstract A Mobile Ad hoc Network (MANET) is a formation of the dynamic network with a group of mobile nodes and communicating through wireless links. The key issue of the mobile MANET is lack of infrastructure and resource constraint that causes fault node. Hence, MANET is exposed to dangerous attacks. In order to improve the secure routing in the network and reduce hazards by fault nodes, we enhance the existing protocol Ad hoc On-Demand Multipath Distance Vector (AOMDV) and designed routing protocol named Trusted Security Adhoc On-Demand Multipath Distance Vector (TS-AOMDV). The TS-AOMDV protocol finding and removing fault node attacks like flooding, blackhole, and grayhole attacks in the network. The detection and segregation of attacks done by using IDS and secure routing in route finding and data transmission phases. An IDS monitor data and control packets those participated in data transmission and route finding phases. In order to enhance the performance of routing in the MANET, An IDS combines and calculate data in the AOMDV for detecting attacks. This improves the TS-AOMDV to give a better performance of routing and protection in MANET. At last, the proposed protocol TS-AOMDV is evaluated with obtainable protocol AOMDV, the performance is analyzed by using the NS-2 simulator and improves routing performance like throughput by 59.1%. The result observed from a simulation that the proposed method enhances the performance of routing like as end-to-end delay, throughput, control overhead, and packet delivery ratio.

Keywords QoS · IDS (Intrusion detection System) · Fault node Routing attacks · TS-AOMDV · NS-2

V. Vinoth Kumar (✉)
Department of Information Technology, Dr. M.G.R. Educational and Research Institute University, Chennai, India
e-mail: vinoagra@gmail.com

S. Ramamoorthy
Department of Computer Science and Engineering, Dr. M.G.R. Educational and Research Institute University, Chennai, India
e-mail: srm24071959@yahoo.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_5

Introduction

Wireless networks are becoming famous nowadays in different areas like military applications, environment applications, etc. The rapid development of wireless communication enlarges the number of users, easy to deploy, and introduce new generations. The generation of new technologies like Wifi, WiGig, and Zigbee improves range of the wireless networks. Security is an important aspect in MANETand needs to improve to detect attacks and isolation of threats. Nowadays, the traditional approaches like encryption and key management techniques are used to provide security to users. Hence, MANET takes network communication with available nodes and forming dynamic topology in nature and the nodes are so much resource constraint that it is difficult to employ security solutions. In Infrastructure less network, nature of dynamic topology changing the mobile nodes, so that network connectivity changes in order to lack AP. Every mobile device in a network transferring data using single path or multipath mode within the communication range [1–3].

Hence, the routing protocol aim is to discover a best appropriate route from source to destination node. The routing process involves efficiently in the occurrence of changing topology, moving of nodes, power and energy, data rate and security, self-directed architecture and reserve controlled infrastructure. The major challenges in MANET are lack of security in routing, needs to improve QoS parameters, cost, and energy consumption. The limited range of communication and absence of communications are the main reason for combining transmission model. In a MANET, a group of nodes creating self-motivated topology and nodes present within the communication level called as neighbor nodes. Each and every node transmits information packets to all neighbors inside the transmission level. The node transfers information packets with continuous of single or multiple hops through in-between, then whenever source node needs to throw information packets to the destination device. The range of latent uses in network supports a wide level of routing protocols to complete the QoS necessities. There is an most important challenges of routing are efficiency and performance of the protocol in the occurrence of a lively infrastructure [4, 5].

Security Needs in Routing Protocols

In multihop routing, the essential requirement is node cooperation for secure data transfer. The noncooperative node leads to fault node and malicious behavior resulting unsecure data transfer and routing attacks. The fault node drops several data packets transmitting through it. The key issues of MANET are lack of infrastructure, access points, and reserve constriction causes a fault and misbehaving nodes. Hence, the MANET is serious in the direction of vulnerable harms [5, 6]. The various number of the possible steering attacks be a wormhole attack,

replay attack, blackhole attack, silent attack, and hurrying attack [7, 8]. The MANET wants to make available trustworthy and safe routing under operation-significant environments similar to disaster and armed applications. While ad hoc network becomes an insecure network the communication established by using infrastructure then transmission of data packets as well as control packets may cause by potential threats and it is exposed to security breaches [9].

The important parameters in MANET are network overload, processing time, and energy consumption. The other factors of safety provisioning are reliability, ease of use, and self-healing. Ease of use means the possible transmission right to use at any time and reliability ensures guarantees the data delivery. In sole pathway routing is defenseless to defense coercion compare to multipath routing because of requirements refuge parameters like as ease of use, trustworthiness and self-healing [10–12].

The attacks may avoid information passing through broken link among the nodes in a routing path, and destination node does not receive the information packets be passing through on its own path between source to the objective node. The modern routing protocols start a route finding mechanism using control packets and from source to destination node. This causes time and energy consumption problem in MANET. This is not acceptable in dangerous mission purposes since it necessary continuous monitoring infrastructure and take appropriate conclusion making. Due to ease of use in displace paths, multipath routing is always flexible to various attacks [13–16].

Identification of Problem

Several routing protocols have been proposed for preventing security attacks in MANET. Such protocols are not suitable for mission critical applications like health care and military applications. These protocols are work well in MANET if movable nodes are exceedingly trust. Henceforth, the modification is required toward these protocols to provide secure and reliable routing in MANET. To maximize the throughput in the network, we make all nodes are available one without any fault node. Cooperation among the nodes for data forwarding is a key issue in MANET. The major challenge in self-organizing network is detecting security attacks. Attacks established through noncooperative nodes and routing performance degrades in the network. To improve the performance of the network, design a routing protocol to detect attacking nodes and eliminate from the network. Another important challenge in infrastructure less network is mobility of the node. The node mobility causes frequent link break and needs to repair the broken link, the routing protocol again send and receive control packets for connection establishment. This causes network degrades performance and decreasing throughput. The Existing

multipath routing protocols are not designed security provisioning in mind. The conventional routing protocols are not considered in security attributes. In mission critical situations, the multipath routing perform the important task of load balancing and maintaining consistent throughput level.

Proposed Methodology

The anticipated protocol TS-AOMDV seek to detect and eliminate security threats such that flooding, blackhole, grayhole attacks. The IDS provides to secure routing, the attack detection, and elimination conceded in data forwarding and route finding phases. In route finding phase, the fault node floods the RREQ packets with absent of target IP. The IDS continuously monitoring source node's packet generation ratio and give an indirect relative rate of RREQ packet count as the "source-trust rate" of the equivalent source. Whenever the trust rate reaches verge level, then attacker drops route request packet. During data transmitting in the network, a fault node acts as a router and drops the data packets as an alternative of forwarding. An Intrusion Detection System monitoring router's data delivery process and assign "router-trust rate" as a relative amount and counts of forwarding packets and receiving packets. Whenever its trust value reaches a threshold level, it selects an alternative path which is moved in the table of the router and start again packet delivery using updated one.

Detection of Attacks by Using Intrusion Detecting System (IDS)

The proposed protocol TS-AOMDV, Intrusion Detection System monitors the routing activity of each and every node to detect routing attacks at network layer [15, 16]. It continuously monitoring data transfer and measuring the rate of packet generation and identify the behavior of the node and predict information transferring statistics. The IDS provide total safety in routing by monitoring both information packets and control packets those engage in data forwarding and route detection phases. The predicted data are combined together in existing protocol AOMDV intended for the finding of attacks. To enhance the performance of routing, the predicted statistics value is converted into trust rate by using IDS and it is used in the data transfer and route identification phases of TS-AOMDV. Due to the modification, the TS-AOMDV provide security and improved performance of routing in Mobile Ad hoc Networks (Fig. 5.1).

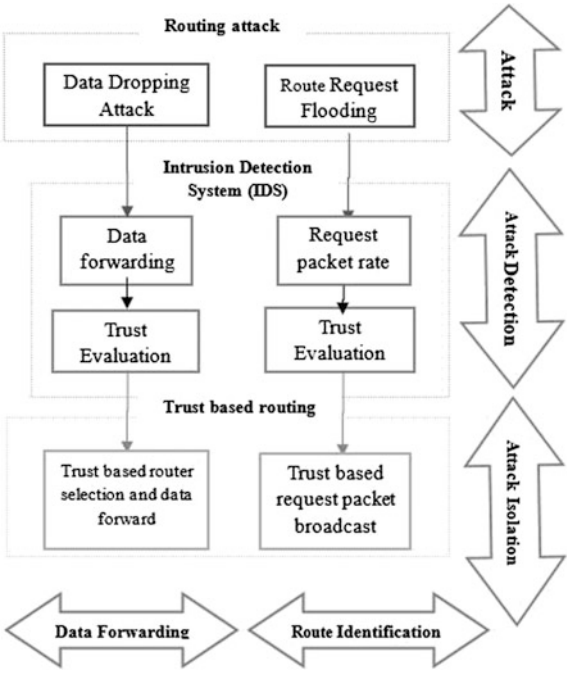


Fig. 5.1 TS-AOMDV system architecture

Determine Routing packet generation rate

The attackers flood the huge amount of control packet (RREQ) breaching the rule of routing. Moreover, the attacker uses deceptive Logical addresses for the target nodes for fault route discoveries. During route discovery phase to authenticate the performance of sending node, the proposed protocol TS-AOMDV count the amount of packets generated using source node and amount of RREQ packets expected at each node. While getting the control packets, the IDS getting the ID of the first generated node and source node increments the Route request value by 1. Thus, the value of data packet production does not exceed the threshold value. In network layer, the IDS keep up the count of number of Route request packets generated from the sending nodes.

Observation of Neighbor node's packet drop

During packet transferring process, the malicious node involves in the router and drops the packet instead of forwarding. The grayhole attack partially drops the packet and blackhole attack continuously dropping data packets. However, in data forwarding phase, thus IDS continuously monitors the data communication of neighbor nodes to identify both blackhole and grayhole attack. It counts the amount of forwarded packets and amount of received packets of a neighbor node to fix dependence trust rate. Owing to that IDS gives the router-trust rate based on the difference between the amount of packets sent and the amount of packets acknowledged and gives notification to the network layer.

Trust Rate Estimation

Based on the network traffic, the threshold value of data packet production and transmissions rate of data are changed. By using IDS, during the high data rate and traffic, the value of the threshold is controlled. To improve the routing concert in proposed protocol, the detecting system calculates the router-trust rate give an indirect relative rate count of RREQ, the difference between received and onward data packet rate. Then, trust rate of intruders are limited and immediately deleted from routing.

$$\text{Source} - \text{Trust} = (\text{RREQ Count})^{-1} \tag{5.1}$$

$$\text{Router} - \text{Trust} = \text{No of packets sent/No of packets Received} \tag{5.2}$$

Based on the Eq. (5.1) the Source-trust rates are gathered in trust table and it is applied to decide propagating RREQ packets. In the result of Eq. (5.2) the trust rate of the router is applied to find the optimal path for packet transmission. A node detects intruders whenever the router trust rate or sending node go beyond the maximum limits and intruders are isolated from the adjacent list and delete every entry of routers from routing table.

Routing Activity of Trusted Secure AOMDV

By using trust-based route finding and IDS, the detection of attacks, and removing attacker nodes in TS-AOMDV are conceded in route finding and data transmission phases. The IDS taking a decision based on the source node's control packet and select an optimal secure route for packet transmission.

Trusted path finding

At the initial stage, every node in the network initiated the value of trust rate by "1" to all neighbors and IDS counts the actual trust value of all nodes and notifies to the network layer. When the trust rate of node less than threshold rates after that the source node's Route request packet is declined due to overflow activity. In Fig. 5.2 the Source node keep trust rate using neighbor nodes.

Trusted Packet Transmission

Before data transmission, every node checks the router trust rate in order to avoid blackhole and grayhole attackers. Then, trust rate of the router specifies consistency of the data forwarding. When the router trust rate is equal to the source node, then source node forward data otherwise fault router is blocked. Consequently, trusted router resumes data transmission through routing table. In Fig. 5.3 trust rate of the neighbor node A value A is less than node B, it selects node B for data forwarding. By this way both IDS and trust-based routing improves the performance of routing and security with TS-AOMDV in MANET.

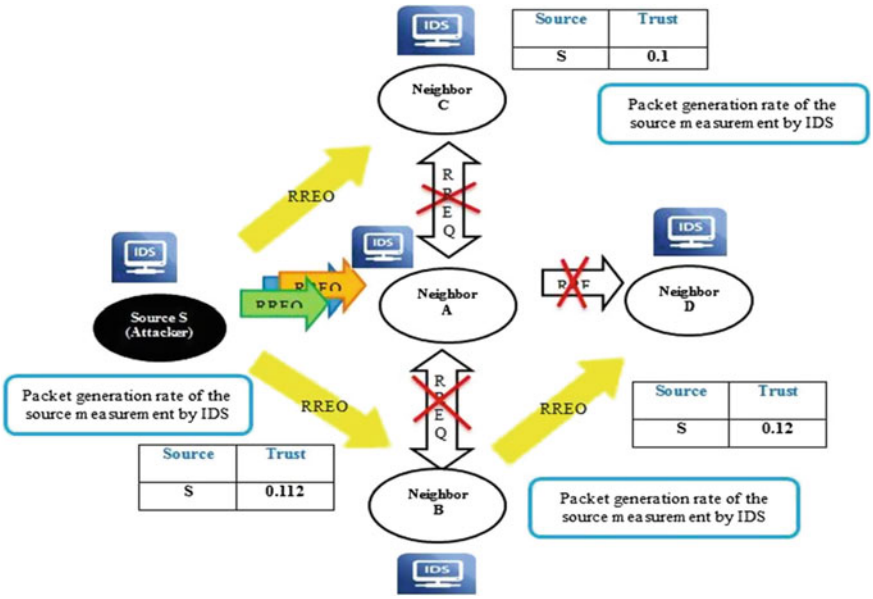


Fig. 5.2 Path finding process: RREQ flooding attack finding and removing

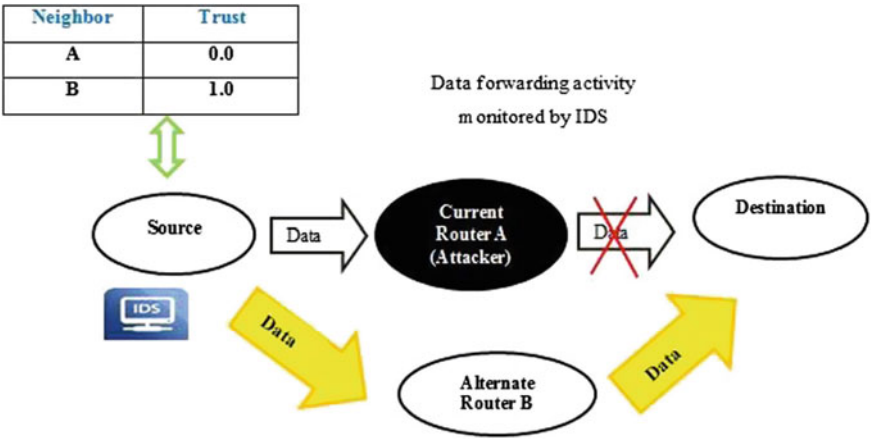


Fig. 5.3 Data forwarding phase: blackhole and grayhole attack finding and removing

Implementing TS-AOMDV

The proposed protocol identifying and isolating blackhole, grayhole attack, and flooding attack. The actual process of TS-AOMDV is experimented using NS2 with QoS parameters, such as throughput, route discovery time, control overhead, energy

consumption. The simulation outputs employed with a variety of simulation models.

Blackhole and Grayhole Attack Detection

Hence, data transmission is initiated from source node 0 and destination node 1 for specific time period. By using AOMDV routing protocol routing path is evaluated and one of the nodes designed as an attacker for drop the packets at a specific time. IDS enabled with every node in a network due to act all neighbor nodes monitors routing process of every router. All the nodes trust value initialized by 1 and router threshold value is 0.8 based on the traffic level.

1. *Route Discovery Time*

The nodes ranges from 50 to 90 based on the network size with transmission range is 250 m. Figures 5.4 and 5.5 show the simulation result of Trusted Secure AOMDV both Grayhole and Blackhole attacks (Tables 5.1 and 5.2).

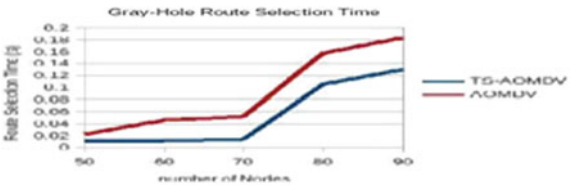


Fig. 5.4 Path finding time for grayhole attack

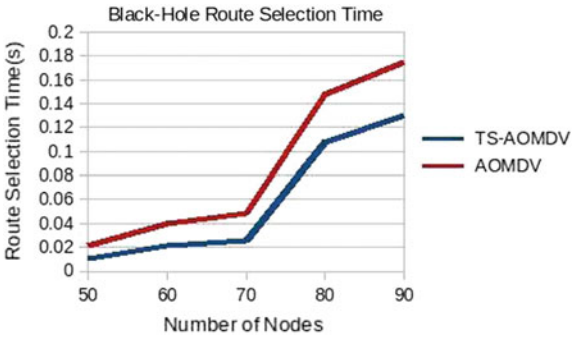


Fig. 5.5 Path finding time for blackhole attack

Table 5.1 Path finding time for grayhole attack

Number of nodes	Route selection time (s)	
	TS-AOMDV	AOMDV
50	0.021	0.01
60	0.04	0.0208
70	0.0490	0.026
80	0.148	0.107
90	0.175	0.130

Table 5.2 Path finding time for blackhole attack

Number of nodes	Route selection time (s)	
	TS-AOMDV	AOMDV
50	0.0218	0.01
60	0.0465	0.0118
70	0.0519	0.0140
80	0.158	0.106
90	0.184	0.1298

2. *Throughput*

Based on the Tables 5.3 and 5.4 the throughput analyzed in Figs. 5.6 and 5.7 with blackhole and grayhole attacks compared by TS-AOMDV and AOMDV protocols.

From observed above simulation results, throughput increases in TS-AOMDV compared to AOMDV by 0.0361 rate.

Table 5.3 Throughput level in grayhole attack

Number of nodes	Throughtput (Mbps)	
	TS-AOMDV	AOMDV
50	0.0361	0.0145
60	0.038	0.0152
70	0.0391	0.0161
8O	0.042	0.0172
90	0.0465	0.0198

Table 5.4 Throughput level in blackhole attack

Number of nodes	Throughput (Mbps)	
	TS-AOMDV	AOMDV
50	0.0285	0.0145
60	0.0298	0.0156
70	0.0312	0.0169
80	0.0348	0.0179
90	0.0396	0.0201

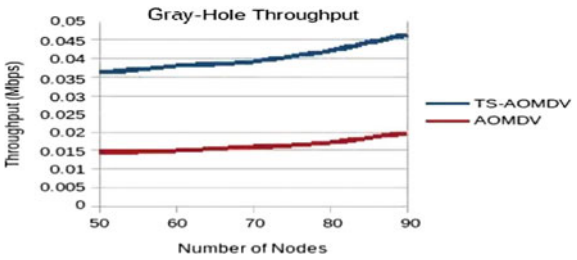


Fig. 5.6 Throughput level in grayhole attack

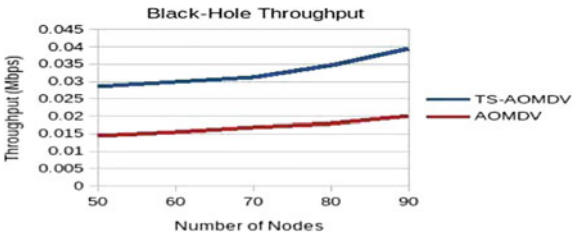


Fig. 5.7 Throughput level in blackhole attack

3. *Control Overhead*

Based on the Table 5.5 the throughput analyzed in Fig. 5.8 with blackhole and Grayhole attacks.

From observed above simulation results both TS-AOMDV and AOMDV have same overhead value at a number of nodes 50. However, the number of nodes at 90, it added the value up to 2700.

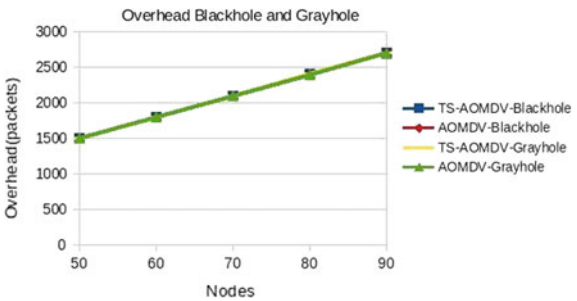


Fig. 5.8 Control overhead blackhole and grayhole

Table 5.5 Control overhead blackhole and grayhole attacks

Number of nodes	Trust non-utilization factor (Packets)			
	Blackhole attack		Grayhole attack	
	TS-AOMDV	AOMDV	TS-AOMDV	AOMDV
50	1500	1500	1501	1500
60	1800	1800	1801	1800
70	2100	2100	2101	2100
80	2400	2400	2401	2400
90	2700	2700	2701	2700

Detection of Flooding Attacks

In a network, one node act as a flooding attacker and IDS is enabled every node in the network. The route request packets floods at every 0.01 s. Each node observes its neighbor nodes RREQ packet generation and threshold limit has been set at all nodes. When the rate of RREQ exceeds threshold value the node detected as a flooding attacker.

1. *Trusted Non-utilization Factor*

The nodes range from 50 to 90 based on the network size with transmission range is 250 m. Figure 5.9 shows the simulation result of TS-AOMDV based on the Table 5.6.

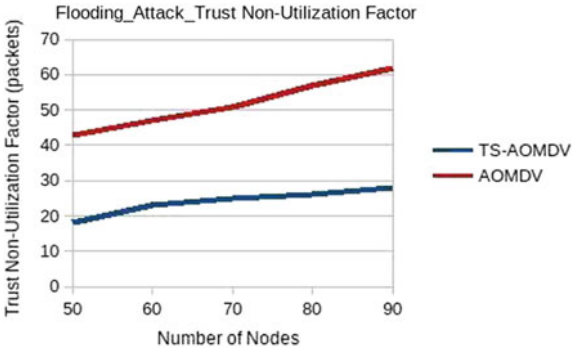


Fig. 5.9 Trusted non-utilization factor flooding attack

Based on the above observation Fig. 5.9 shows, first 18 packets are forwarded and gradually increased 28 packets at 90 nodes. As per the result, while the node number at 90, the trusted non-utilization factor of AOMDV number of packets increases to 62. But at same number of node TS-AOMDV protocol achieve only 28 packets.

Table 5.6 Trusted non-utilization factor flooding attack detection

Number of nodes	Trust non-utlllzation factor (Packets)	
	TS-AOMDV	AOMDV
50	18	43
60	23	47
70	25	51
80	26	57
90	28	62

2. *Energy Consumption*

A number of nodes ranges from 50 to 90 based on the network area considered for energy consumption. Figure 5.10 shows energy consumption of flooding attacks by using Table 5.7.

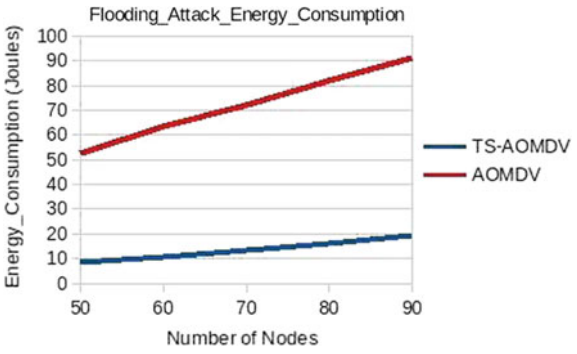


Fig. 5.10 Energy consumption for flooding attack

Table 5.7 Energy consumption for flooding attack

Number of nodes	Energy consumption (J)	
	TS-AOMDV	AOMDV
50	8.52	52.80
60	10.67	63.49
70	13.37	72.45
80	16.30	82.20
90	19.17	91.38

In table, the result shows the trusted secure AOMDV consume 8.5 J when the number of nodes at 50 and improved to 19.1 J when nodes at 90. Compared protocol Ad hoc on-Demand Multipath Distance Vector consumes 52.80 J at 50 nodes and gradually increases 91.38 J at 90 nodes.

3. *Control overhead*

Figure 5.11 shows the resulting graph based on the Table 5.8 value and results are compared between AOMDV and TS-AOMDV protocol.

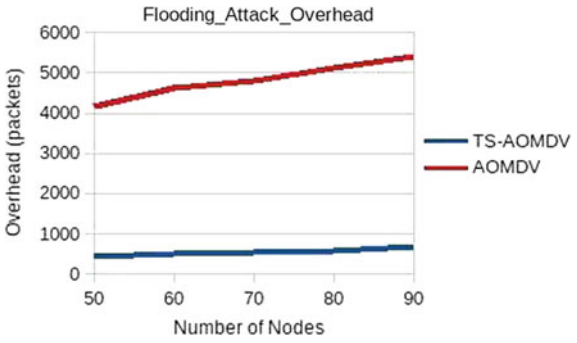


Fig. 5.11 Control overhead for flooding detection

Table 5.8 Control overhead for flooding attack

Number of nodes	Overhead	(Packets)
	TS-AOMDV	AOMDV
50	434	4156
60	494	4618
70	530	4808
80	568	5109
90	666	5395

From above simulation result observed using nodes ranges from 50 to 90 on 600 m × 600 m area considered for overhead. From Fig. 5.11, the AOMDV protocol overhead value is 4156 at 50 nodes and escalates the routing overhead to 5395 at 90 nodes. But the TS-AOMDV shows overhead value 434 at 50 nodes and gradually increased to 666 at 90 nodes. The result shows TS-AOMDV protocol have less overhead value compared to AOMDV. It is observed that the AOMDV protocol, not aware flooding attacks.

Conclusion

In this research, to improve the secure routing in MANET, the newly designed protocol a TS-AOMDV implemented in MANET. The newly designed protocol detects and isolates the security attacks named as grayhole and blackhole attacks on data packets and flooding attacks on route request packets. The IDS embedded with every node, perform operations such as measuring RREQ value of the source at

during route finding and predict packet forwarding rate of the neighbor nodes during data forwarding phase. The IDS compares these results with threshold value it detects attackers. The predicted data is integrated to trust rates and discovering secure routing for improves QoS of the proposed protocol. The newly designed protocol is evaluated using NS2. The QoS metrics in conditions of time taken for path discovery, level of throughput, trusted non-utilization factor, control overhead, and energy consumption are achieved the best result in TS-AOMDV protocol compared with AOMDV.

References

1. Erciyes, K. "Distributed Graph Algorithms for Computer Networks", Computer Communications and Networks, London: Springer, pp. 259–275, (2013)
2. S. Abdel Hamid, H. Hassanein and G. Takahara, "Routing for Wireless Multi-Hop Networks: Unifying Features", SpringerBriefs in Computer Science, pp. 11–23, (2013)
3. Hamid, S. A., Hassanein, H., & Takahara, G., "Routing for Wireless Multi Hop Networks–Unifying and Distinguishing Features", School of Comp.—Queen's University, Canada, report 583, (2011)
4. Habib, S., Saleem, S., & Saqib, K. M., "Review on MANET routing protocols and challenges", IEEE Student Conference on Research and Development SCOReD, pp. 529–533, (2013)
5. A. Ahmed, K. Abu Bakar, M. Channa, K. Haseeb and A. Khan, "A survey on trust based detection and isolation of malicious nodes in ad-hoc and sensor networks", Frontiers of Computer Science, vol. 9, no. 2, pp. 280–296, (2015)
6. I. Abdel-Halim, H. Fahmy and A. Bahaa-Eldin, "Agent-based trusted on-demand routing protocol for mobile ad-hoc networks", Wireless Netw, vol. 21, no. 2, pp. 467–483, (2015)
7. Mahmoud, Mohamed MEA, and Xuemin Sherman Shen. "Secure routing protocols." Security for Multi-hop Wireless Networks. Springer International Publishing, pp. 63–93, (2014)
8. Shanmuganathan, V., and T. Anand. "A Survey on Gray Hole Attack in MANET." IRACST–International Journal of Computer Networks and Wireless Communications (IJCNWC), pp. 2250–3501, (2012)
9. Tayal, S., & Gupta, V., "A Survey of Attacks on Manet Routing Protocols", International Journal of Innovative Research in Science, Engineering and Technology, vol. 2, no. 6, pp. 2280–2285, (2013)
10. Vaidya, Binod, et al. "Secure multipath routing scheme for mobile ad hoc network." Third IEEE International Symposium on Dependable, Autonomic and Secure Computing, pp. 163–171, (2007)
11. C. Tachtatzis and D. Harle, "Performance evaluation of multi-path and single-path routing protocols for mobile ad-hoc networks", Performance Evaluation of Computer and Telecommunication Systems,2008. SPECTS 2008. International Symposium on, pp. 173–180, (2008)
12. K. Yu, C. Yu and S. Yan, "An Ad Hoc Routing Protocol with Multiple Backup Routes", Wireless Personal Communication, vol. 57, no. 4, pp. 533–551, (2011)
13. Eliana Stavrou, Andreas Pitsillides,"survey on secure multipath routing protocols in WSNs," Computer Networks, vol. 54, no. 13, pp 2215–2238, (2010)
14. Mitchell, R., & Chen, R, "A survey of intrusion detection in wireless network applications", Computer Communications, vol. 42, pp. 1–23, (2014)

15. Mitchell, R.; Chen, I.-R., "Behavior Rule Specification-Based Intrusion Detection for Safety Critical Medical Cyber Physical Systems," in Dependable and Secure Computing, IEEE Transactions on, vol. 12, no. 1, pp. 16–30, (2015)
16. Mitchell, R., & Chen, I. R., "Specification based intrusion detection for unmanned aircraft systems", Proceedings of the first ACM MobiHoc workshop on Airborne Networks and Communications, pp. 31–36, 2012.

Chapter 6
Reputation-Based Trust for Selection of Trustworthy Cloud Service Providers

Monoj Kumar Muchahari and Smriti Kumar Sinha

Abstract In cloud computing—an emerging paradigm, the challenges of trust management are the prime issue. Inadequate security assertion, loss of asset control, and transparency compel enterprise to question cloud services. Cloud environment being dynamic and transient, establishment of trust becomes a necessity. In this paper, we present a reputation-based trust management architecture for cloud computing. In the proposed framework, the selection of trustworthy Cloud Service Provider is made on the basis of direct, indirect, and transitive trust values. History of reputation rating is used as weight to calculate the trust values. The effectiveness of our approach is demonstrated with experimental results.

Keywords Reputation · Trust · Cloud computing

Introduction

The agility of cloud has captivated many large and small enterprises but carries along, the risk of data privacy and security. To take advantage of the remunerative features of cloud, an effective trust management is necessary. The reason that makes the selection of a trustworthy provider a difficult task is the increasing number of Cloud Service Providers (CSPs) with varying offers [1]. Feedback and ratings from Cloud Service Consumers (CSCs) and CSPs can play a major role in building trust and reputation of prospective CSP. Trust and reputation is either considered as a single entity or are being closely linked. Reputation is defined to be an estimation of consequential interaction of an entity and trust is judged on prior insight of anticipated reputation of a source. Reputation-based trust is believed to shun malicious

M. K. Muchahari (✉) · S. K. Sinha
Department of Computer Science and Engineering, Tezpur University, Napaam, Tezpur, Sonitpur 784028, India
e-mail: memonoj01@yahoo.com

S. K. Sinha
e-mail: smritikumarsinha@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_6

players, minimize the menace of dishonest behavior, and protect a networked computing system from possible misuses and abuses [2]. Decision of CSC to consume a service will immensely depend on the reputation of the service providing CSP and therefore CSP will try to build and sustain higher reputation.

In this paper, we propose a reputation-based trust management framework where selection of trustworthy CSPs by prospective CSCs is made on the basis of direct, indirect, and transitive trust values. Using explicit experiences and feedback provided, we gauge trust values of CSPs by employing reputation as weight to the calculation. Cloud being a scalable network of nodes, requires more robust trust model against inaccurate or malicious feedback and its credibility must be judged extensively [3]. To cope with malicious feedback, credibility of feedback and ratings are checked. A recommender that consists group of experts, represents CSC preferences, for the purpose of suggesting suitable trustworthy CSPs.

The rest of the paper is organized as follows. In section "Literature Review", related works in literature are presented. Section "Proposed Framework" illustrates the proposed framework and the experimental results are presented in section "Experimental Results". The conclusions and future directions are given in section "Conclusion".

Literature Review

From age-old markets to today's e-commerce, trust and reputation are the two premier concepts of decision-making in many fields [4, 5]. Hwang et al. [6] suggest a reputation system for building trust between CSPs and data proprietors employing a trust overlay network over various data centers though reputation is being used for only data center protection. Hwang et al. [7] also recommend data access defense of cloud through trust management with reputation systems. Trust Manager in [8] collects and maintains reputation rating of CSPs based on direct and indirect observation and experience. Sanchez et al. [9] introduce trust management based on reputation within their identity management model. An architecture called Hatman (HAdoop Trust MANager) is proposed in [10] that is a data-centric, reputation-based trust management system for Hadoop clouds. Some similar works are discussed in [11]. Most of the works do not discuss how reputation can affect the trust value. Reputation rating history is also an important factor that needs to be addressed for a robust reputation value.

Proposed Framework

The proposed framework as depicted in Fig. 6.1 has Cloud Service Registry and Discovery (CSRD) that works as a repository of registered CSPs amalgamated with archive of direct, indirect, and transitive trust values. Direct and indirect trust comes

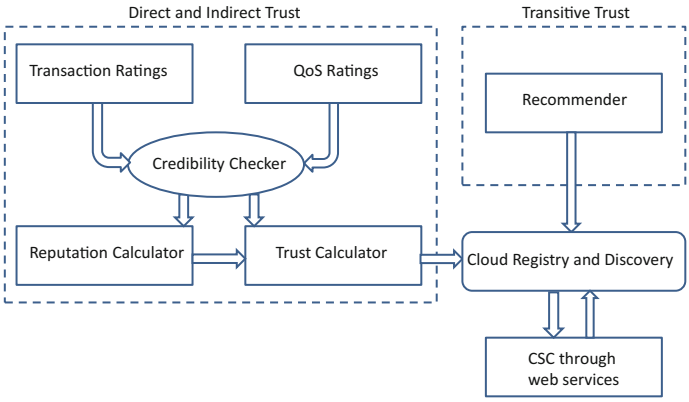


Fig. 6.1 Proposed framework

from Reputation Calculator (RC) and Trust Calculator (TC) whereas transitive trust comes from Recommendation Accumulator (RA). Credibility Checker (CC) is used to overcome malicious or inaccurate ratings and feedback from RC and TC. The process of query for trustworthy CSP is invoked through web service.

Cloud Registry and Discovery

Cloud computing implies services being delivered through Internet [12] and so, we believe CSRD is imperative. CSPs register themselves in CSRD based on their service delivery models namely Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) (Table 6.1).

Credibility Checker

For calculation of credibility of entity, the number of services provided and the duration since when the entity started providing services or the number of cloud services being consumed and the duration since when the entity started consuming services are taken into account as in [13]. Feedback of entity whose credibility value is higher than the mean of all calculated credibility is contemplated to calculate the trust value of a provider.

Table 6.1 Symbol table

Symbol	Meaning
D^c	Current reputation dataset
D^p	Previous reputation dataset
$\mathfrak{R}$	Reputation value of a CSP
R_p	Number of positive reputation ratings of a CSP
R_z	Number of neutral reputation ratings of a CSP
R_n	Number of negative reputation ratings of a CSP
R_t	Total number of reputation ratings for a CSP
R_{pmax}	Maximum number of positive rating of current and previous dataset
R_{zmax}	Maximum number of neutral rating of current and previous dataset
R_{nmax}	Maximum number of negative rating of current and previous dataset
F^c	Current feedback dataset
F^p	Previous feedback dataset
T_r	Calculated trust value
CSP^i	CSPs where $i = 1, 2, \ldots, n$
$P^i(j)$	jth feedback parameter of CSP i where $j = 1, 2, \ldots, m$
$M^c(j)$	Mean of parameter j of each CSP in $F_i^c, \forall i = 1, 2, \ldots, n$
$S^p(j)$	Standard deviation of parameter j of each CSP in $F_i^p, \forall i = 1, 2, \ldots, n$

Algorithm 1: Calculation of credibility of RE (E)

1: Credibility of E,

$$C_E = \frac{\sum_{s=1}^{n} E_s \times D_E}{\sum_{s=1}^{n} E_s + D_E}$$

where E_s is the number of cloud services consumed or provided by E, D_E is the duration since E started consuming or providing the service E_s.

2: Find the median of C_E.

3: Select all E whose C_E value is larger than the median.

Reputation Calculator

The reputation percentage of the rated CSP ($\mathfrak{R}$) is calculated using Algorithm 2. The rating values for RC are drawn from personal transactions. For reputation, CSPs and CSCs or Rating Entities (REs) rate CSPs after each transaction with three heads namely positive, neutral, and negative (+1, 0, −1).

To evade malicious ratings, we allow only registered credible REs to rate a particular CSP, with no repetitive but reversible rating. No repetitive rating means that the registered REs if have given their rating once are not allowed to rate again but are allowed to reverse or modify their ratings. Reputation changes with time and so

Algorithm 2: Reputation Calculator

Data: D^p, D^c
Result: $\mathfrak{R}$
begin
  Calculate R_p^p, R_z^p, R_n^p and R_t^p from D^p;
  Calculate R_p^c, R_z^c, R_n^c and R_t^c from D^c;
  Calculate $R_{pmax} = \max(R_p^p, R_p^c)$, $R_{zmax} = \max(R_z^p, R_z^c)$ and $R_{nmax} = \max(R_n^p, R_n^c)$;
  Calculation $Fr_p^p = \frac{R_p^p}{R_t^p} \times R_{pmax}$, $Fr_z^p = \frac{R_z^p}{R_t^p} \times R_{zmax}$ and
  $Fr_n^p = \frac{R_n^p}{R_t^p} \times R_{nmax}$;
  Calculation $Fr_p^c = \frac{R_p^c}{R_t^c} \times R_{pmax}$, $Fr_z^c = \frac{R_z^c}{R_t^c} \times R_{zmax}$ and
  $Fr_n^c = \frac{R_n^c}{R_t^c} \times R_{nmax}$;
  Calculate $R_{max} = \max(\max(Fr_p^c, Fr_p^p), \max(Fr_z^c, Fr_z^p), \max(Fr_n^c, Fr_n^p))$;
  if $R_{max} = R_p$ **then**
    $\mathfrak{R} = \frac{R_p}{R_t} \times 100$
  else if $R_{max} = R_z$ **then**
    $\mathfrak{R} = \frac{R_z}{R_t} \times 100$
  else
    $\mathfrak{R} = \frac{R_n}{R_t} \times (-100)$
  end
end

reputation history is quantitatively taken into account such that the effect can be seen in the newer reputation value.

Trust Calculator

TC gauge the trust value for each registered CSP. Reputation percentage R^p from RC is used as weight to evaluate the trust value in TC. Trust values of all registered CSP are calculated using Algorithm 3. Trust value changes with time and transaction. To know the deviation of feedback value of current dataset from previous dataset, first mean of each CSP $M^c(j)$ for each parameter j from current dataset is computed and then using $M^c(j)$, the standard deviation of the previous dataset is determined. Averaged value of sum or difference of deviation along with current feedback value and R^p provide the trust values.

Recommender

Recommender aids in the selection of cloud services. The recommender may consist of cloud experts, cloud market experts, cloud service experts, etc., that form a group.

Algorithm 3: Trust Calculator Algorithm

Data: $F^p, F^c, \mathfrak{R}$
Result: T_r
begin
  Compute $M^c(j)$
  Compute $S^p(j)$
  Compute $diff(j) = M^c(j) - S^p(j)$
  foreach *parameter j in* CSP^i *in* F^c **do**
    $P^i(j) = F_i^c(j) + diff(j)$
  end
  Compute $M = \sum_{j=1}^{m} \frac{P_i(j)}{m}$
  if $\mathfrak{R} = 0$ **then**
    $T_r = M$
  else
    $T_r = M \times \frac{\mathfrak{R}}{100}$
  end
end

The recommender would find out which of the CSP would be most suitable and trustworthy for consumers.

Experimental Results

Experiments were carried out on a machine of 3.40 GHz Intel i7 processor with 16 GB main memory on 64-bit Windows 10 Pro operating system. The proposed algorithms are implemented using MATLAB 2015a software.

Dataset Used

Two datasets, namely Trust Feedbacks Dataset[1] by "Cloud Armor Project" and "Rating" taken from the well-known product review website Ciao[2] for QoS feedback, are considered. For reputation rating, we use Jester Dataset.[3] We considered only those ratings whose values were −1, 0 and +1, discarding the rest of the ratings. For experimental purpose, two synthetic dataset namely "*Synthetic 1*" and "*Synthetic 2*" are considered. Both "*Synthetic 1*" and "*Synthetic 2*" were generated using random

[1] http://cs.adelaide.edu.au/cloudarmor/ds.html.
[2] http://www.ciao.co.uk/.
[3] http://www.ieor.berkeley.edu/goldberg/jester-data/.

Table 6.2 Statistics of dataset used

(a) Statistics of Armor and Ciao dataset

	Armor		Ciao
No. of CSCs	7312	No.of users	7375
No. of CSPs	117	No.of products	106797
No. of QoS attributes	12	No.of category	12
Rating range	0–5	Rating range	0–5

(b) Statistics of Jester dataset

	Jester-data-1		Jester-data-2
No. of users	24,983	No. of users	23,500
No. of jokes	36	No. of jokes	36
Rating range	−10 to +10	Rating range	−10 to +10

(c) Statistics of synthetic dataset

	Synthetic 1	Synthetic 2
No. of CSCs	10080	282649
No. of CSPs	235	235
No. of QoS attributes	25	25
Rating range	1-3	1-5

(d) Statistics of synthetic rating dataset

	SynthR 1		SynthR 2
No. of users	10080	No.of users	24990
No. of jokes	117	No.of jokes	235
Rating	−1, 0, +1	Rating	−1, 0, +1

integer number function in MATLAB with rating value range of 1–3 and 1–5 respectively. The statistics of dataset are shown in Table 6.2.

Results

To evaluate our model, we use average of feedback trust value as a metric. We considered the "Ciao" as earlier and the "Armor" as present dataset for trust percentage calculation whereas "jester-data-1" is used to act as the earlier rating and "jester-data-1" as the current rating for reputation percentage. In Fig. 6.2a, the reputation weighted trust percentage of CSPs from "Armor" and "Ciao" is shown. Figure 6.2b depicts the average feedback percentage and reputation from ratings of the similar CSPs. To show how reputation affect the traditionally calculated average trust value, a comparison is plotted in Fig. 6.2c.

In Fig. 6.2a, we see an improved feedback in "Synthetic 1" than the earlier "Synthetic 2" dataset and "SynthR 2" has ratings of three values (−1, 0, +1) compared to only one value (−1) ratings in "SynthR 1". The reputation weighted trust values of "Synthetic 1", "Synthetic 2", "SynthR 1", and "SynthR 2" are given in Fig. 6.3a and b, comparison of reputation weighted trust and reputation percentage is also given. The improved feedback and reputation ratings from the earlier dataset show consequential reputation weighted trust as shown in Fig. 6.3c.

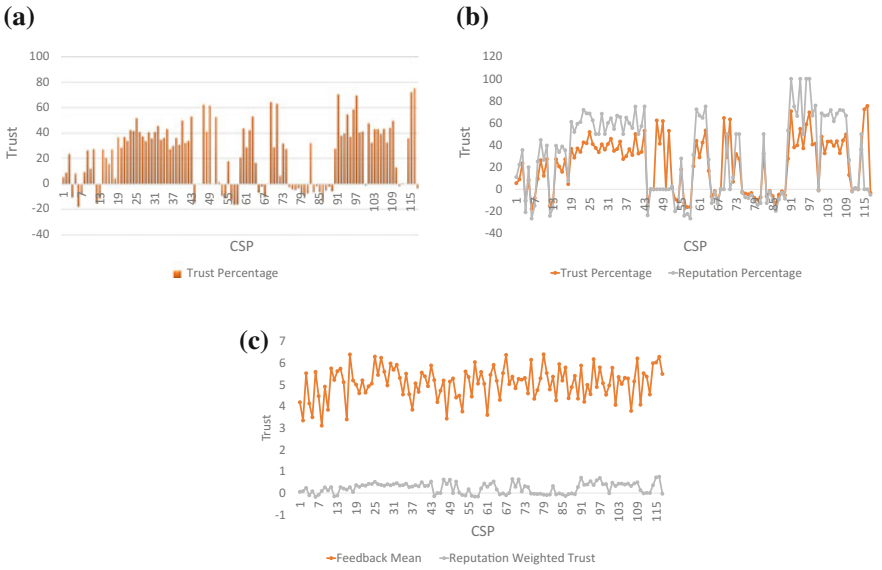


Fig. 6.2 Results using Armor and Ciao data

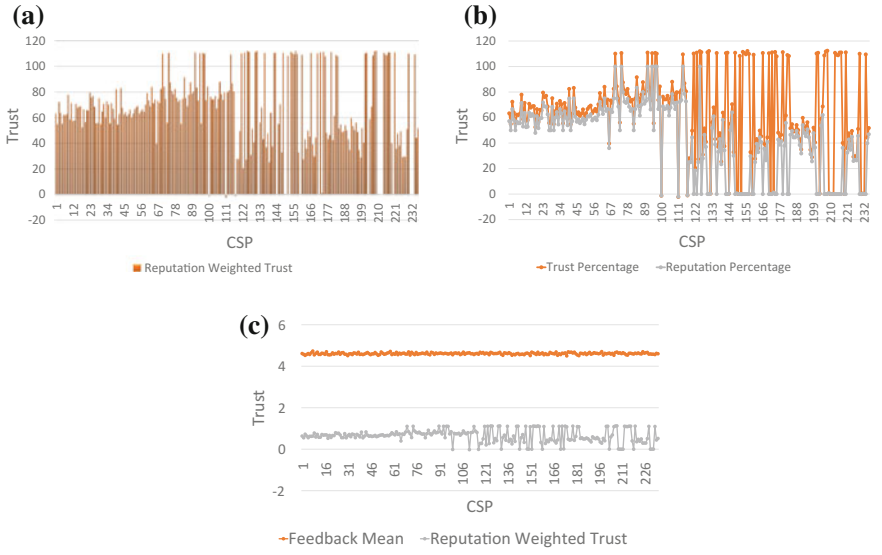


Fig. 6.3 Results from Synthetic 1 and Synthetic 2 data

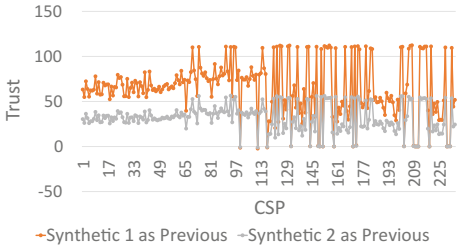


Fig. 6.4 Result with Synthetic 1, SynthR 1 as previous and Synthetic 2, SynthR 2 as current and vice versa

Figure 6.4 depicts the result when we consider Synthetic 1, SynthR 1 as previous dataset and Synthetic 2, SynthR 2 as current dataset and vice versa that shows how previous trust history data affects the current data.

Conclusion

In this paper, we have presented a trust management framework that uses reputation as weight to calculate the trust value of CSPs. The proposed framework consists of CSRD that acts as a repository of registered CSPs and information of direct, indirect, and transitive trust values. For credibility level, CC is equipped with credibility algorithm to contend against inaccurate or malicious feedback. The process of query for trustworthy CSP is invoked through web service. The work has been evaluated in different sets of experiments and experimental results demonstrate the applicability of our approach. In future, we intend to deal with new providers that are devoid of reputation rating history. Robust credibility and recommender system are another focus of our work.

References

1. Muchahari, M.K., Sinha, S.K.: Mmh: An effective clustering algorithm for trustworthy cloud service provider selection. International Journal of Computer Science and Information Security **14**(5) (2016) 281–291
2. Liu, L., Shi, W.: Trust and reputation management. Internet Computing, IEEE **14**(5) (2010) 10–13
3. Muchahari, M.K., Sinha, S.K.: A survey on web services and trust in cloud computing environment. In: National Workshop on Network Security 2013, Tezpur University. Volume 2., Narosa Publishing House (2013) 91–102
4. Habib, S.M., Ries, S., Mühlhäuser, M.: Cloud computing landscape and research challenges regarding trust and reputation. In: Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2010 7th International Conference on, IEEE (2010) 410–415
5. Habib, S.M., Hauke, S., Ries, S., Mühlhäuser, M.: Trust as a facilitator in cloud computing: a survey. Journal of Cloud Computing **1**(1) (2012) 1–18

6. Hwang, K., Li, D.: Trusted cloud computing with secure resources and data coloring. Internet Computing, IEEE **14**(5) (2010) 14–22
7. Hwang, K., Kulkareni, S., Hu, Y.: Cloud security with virtualized defense and reputation-based trust mangement. In: Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on, IEEE (2009) 717–722
8. Abawajy, J.: Establishing trust in hybrid cloud computing environments. In: Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on, IEEE (2011) 118–125
9. Sanchez, R., Almenares, F., Arias, P., Díaz-Sánchez, D., Marín, A.: Enhancing privacy and dynamic federation in idm for consumer cloud computing. Consumer Electronics, IEEE Transactions on **58**(1) (2012) 95–103
10. Khan, S.M., Hamlen, K.W.: Hatman: Intra-cloud trust management for hadoop. In: Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, IEEE (2012) 494–501
11. Noor, T.H., Sheng, Q.Z., Zeadally, S., Yu, J.: Trust management of services in cloud environments: Obstacles and solutions. ACM Computing Surveys (CSUR) **46**(1) (2013) 12
12. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., et al.: A view of cloud computing. Communications of the ACM **53**(4) (2010) 50–58
13. Muchahari, M.K., Sinha, S.K.: A new trust management architecture for cloud computing environment. In: Cloud and Services Computing (ISCOS), 2012 International Symposium on, IEEE (2012) 136–140

Chapter 7
Global Common Sequence Alignment Using Dynamic Window Algorithm

Lalit Kumar Behera

Abstract The DNA sequencing technology is evolving day by day. It results in the creation of large repository of sequence data. One important aspect of analyzing these sequence data is sequence alignment. In this research paper, a method of DNA sequence alignment with the help of a dynamic window algorithm has been discussed. The dynamic window will help in producing the comparative score of different alignment scheme resulting into the best acceptable alignments efficiently.

Keywords DNA · Sequence alignment · Dynamic algorithm

Introduction

In the past decade, a lot of research work has been done in the field of DNA sequence alignment. As a result, a large number of alignment algorithms and related software are available at present [1]. Basically, a sequence alignment problem deals with matching the nitrogen base of a sequence with a reference sequence. The processes help in finding out the similar functional regions that have great importance in the field of bioinformatics and its related field.

DNA Sequencing and Alignment

DNA is generally a long sequence of base pairs, polymer of repeating subunits called nucleotide. Each nucleotide contains five-carbon sugar, a phosphate group, and a base. Different bases are adenine (A), guanine (G), thymine (T), cytosine (C), and uracil (U). Nitrogen base combines with the five-carbon sugar to form nucleoside. Nucleoside in turn combines with phosphate groups to form nucleotide [2].

L. K. Behera (✉)
Department of Statistics, Center for IT, Utkal University, Odisha, India
e-mail: whomelalit@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_7

During the sequencing, DNA is represented as sequence of these nitrogen bases. Various methods are there for sequencing the DNA like Sanger's method, Maxam & Gilbert method, Pyro-sequencing, etc. [3]. Apart from these conventional methods, there are many other interdisciplinary approaches like graphical representation techniques [4], Voss representation, 2-bit binary, the 4-bit binary, the paired nucleotide, the 12-letter alphabet, the digital Z-signals, and the phase-specific Z-curve [5].

Sequence alignment is the simultaneous matching of two or more nucleotides. There are different algorithms available for this purpose. Broadly, those can be classified as algorithms based on hash tables, algorithms based on suffix trees and algorithms based on merge sorting [6]. Optimal alignment methods aim to maximize the score or minimize the cost of the process. But in case of dynamic programming approach, the time complexity of the algorithm is proportional to length of the sequence in exponential terms [7, 8]. In such approach, if the gap cost is nonlinear then complexity is extremely more [9].

Proposed Method

In this research paper, two DNA sequences have been aligned using a dynamic window algorithm. The major functions associated with this algorithm are LengthEqualiser (X, Y), dynamic window (p, q, X, Y), Modify-Y (i, j, X, Y), Modify-X (i, j, X, Y), and score (hit, c).

Algorithm of length equalizer function

```
Length_Equaliser (X,Y)
1.If |X| != |Y|
2.    do  NIG ← ||X| ~ |Y||
3.          if  |X| < |Y|
4.               append NIG to X
5.               else
6.               append NIG to Y
```

The function Length Equalizer takes the two sequences as arguments and tries to make them the same length by introducing the gaps at proper place. NIG represents the number of introductory gaps. X and Y are the two sequences,| | represents their cardinality.

Algorithm of Dynamic window function

```
1. P←1
2. q←1
3.       Dynamic_Window (p,q,X,Y)
4.         {seq← ø
5.           for  i← p to ( i- |X|)
6.           for  j← q to  (j- |Y|)
7.                 do  {seq ← ø + X(i)
8.                         i ←i+1
9.                         j←j+1
10.                    }
11.        hit ←i-p
12.           Return  seq to List , hit to Score( )
13.        }
14.            Modify-Y (i,j,X,Y)
15.            Modify-X (i,j,X,Y)
```

The dynamic window function tries to match the possible common sequences. The dynamic nature of the window is maintained with the help of sub-functions—Modify-Y (i, j, X, Y) and Modify-X (i, j, X, Y). Apart from this, the function sends the hits calculated to the score function to calculate the score of the process and keeps the different common sub-sequences obtained.

Algorithm of Modify-Y (i,j,X,Y) function

```
Modify-Y (i,j,X,Y)
1.{
2. for m ← j to m ≤ |Y|
3.       do {
4.                while ( X(i) != Y(m) )
5.                            c ← j +1
6.                            m ← m+1
7.            }
8.                      Insert gap (c-i) before X(I)
9.                       X(i) ← X(c)
10.                  P←i
11.                q←i
12.               X← X (i, |X|+c-i)
13.            Length_Equaliser (X,Y)
14.         Dynamic_Window (p,q,X,Y)
15.    Return  c to Score(hit, c )
16.  score (hit, c)
17. }
```

Algorithm of Modify-X (i,j,X,Y) function

```
Modify-X (i,j,X,Y)
1.{
2. for n ← i to m ≤ |X|
3.       do {
4.                while ( Y(j) != X(n) )
5.                            c ← i +1
6.                            n ← n+1
7.            }
8.                      Insert gap (c-i) before Y(j)
9.                       Y(j) ← Y(c)
10.                  P←j
11.                q←j
12.               Y← Y (i, |Y|+c-i)
13.            Length_Equaliser (X,Y)
14.         Dynamic_Window (p,q,X,Y)
15.    Return  c to Score(hit, c )
16.  Score (hit, c)
17. }
```

These two Modify functions are accountable for the dynamic nature of the window. First of all, these functions search the right position for the shifting of a base by calculating the number of gaps required which is represented by "c". Then after introducing the gaps in the either sequence X OR Y, they revoke the Length_Equaliser (X, Y) and Dynamic_Window (p, q, X, Y) which work recursively. They also return the c-value to Score function.

```
Algorithm of  Score (hit, c)

Score (hit, c)
   1. {
   2. Cost- hit ←hit x weight-of- base
   3. Penalty ← c  x  weight-of- gap
   4. Score % ←( Cost- hit ~ Penalty) / |X|
   5. }
```

The Score () function takes hit from Dynamic_Window (p, q, X, Y) and c from Modify-X (i, j, X, Y) or Modify-Y (i, j, X, Y) and calculates the score. Here, the cost of hit is calculated using the cost of each base in the sequence. Similarly, the penalty is calculated using the cost of a single gap and cumulative number of gaps introduced. Figure 7.1 represents the schematic view of the working of the algorithm.

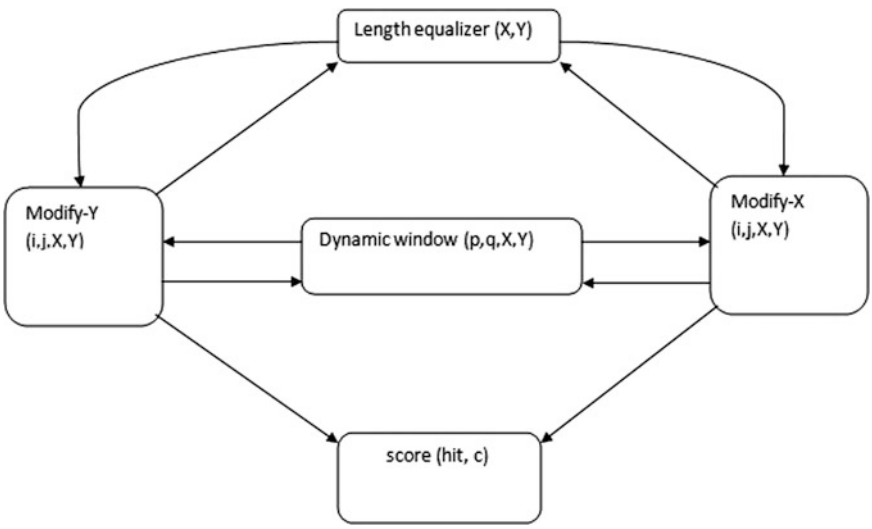


Fig. 7.1 Schematic view of the working of the dynamic window algorithm

Table 7.1 The coding sequence of the exon1 of beta-globin gene of eight different species

Species	Coding sequence
Human (92 bases)	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
	GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum (92 bases)	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCT
	GGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus (92 bases)	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCT
	GGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur (92 bases)	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGT
	GGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Rat (92 bases)	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGT
	GGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Rabbit (90 bases)	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTG
	GGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Goat (86 bases)	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCA
	AGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Chimpanzee (105 bases)	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA
	GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

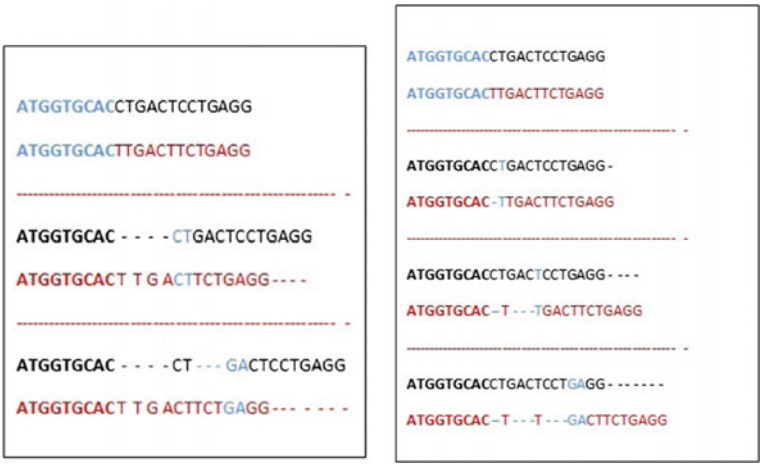


Fig. 7.2 Dynamic_Window algorithm functioning; left side represents the working of Modify-Y (i, j, X, Y) function; right side represents the working of Modify-X (i, j, X, Y) function

Result and Discussion

The performance of the dynamic window algorithm has been analyzed by different combinations of the following DNA sequences of β-globin genes of eight species given in Table 7.1 [10]. A part of the algorithm functioning is given in Fig. 7.2.

As a result of the working of the Dynamic_Window algorithm on a sub-sequences of Human and Opossum β-globin genes, it has been found that the hits are 11, gaps are 7 using Modify-Y (i, j, X, Y) function whereas the hits are 12, gaps are 7 using Modify-X (i, j, X, Y) function. By taking the cost of a base as +1 and that of a gap as −1, the score % in first case is 13.79 (left side in Fig. 7.2) whereas in second case, it is 17.29 (right side in Fig. 7.2). The common sequences in first case in the List are {ATGGTGCAC, CT, GA} and in the second case are {ATGGTGCAC, T, GA}.

Conclusion

In this study, the common sequence alignment using dynamic window algorithm has been proposed. The structure and the working of the algorithm have been discussed. The algorithm has been used for different sets of subsequences of β-globin genes and the score % has been obtained. The dynamic nature is clear from the comparative analysis of the score percentages.

References

1. Bioinformatics: sequence and genome analysis, David W. Mount, Cold Spring Harbor Laboratory Press.
2. Fundamental Molecular Biology, Lizabeth A. Allison, Blackwell Publishing Ltd.
3. A review of DNA sequencing techniques, Lilian T. C. Franca et al, Quarterly Reviews of Biophysics 35, 2 (2002), pp. 169–200. 2002 Cambridge University Press.
4. DB-Curve: a novel 2D method of DNA sequence visualization and representation, Chemical Physics Letters 367 (2003) 170–176.
5. Advanced Numerical Representation of DNA Sequences, 2012 International Conference on Bioscience, Biochemistry and Bioinformatics, IPCBEE vol. 3 1(2012) © (2012) IACSIT Press, Singapore.
6. A survey of sequence alignment algorithms for next-generation sequencing, Heng Li and Nils Homer, Briefings in bioinformatics.vol 11. no 5. 473–483, 2010.
7. Bioinformatics: sequence, structure & databanks, A practical approach, D. Higgins, W. Taylor, Oxford university press.
8. Sequence alignment and Markov's Model, K.R. Sharma, The Mc-Graw Hill.
9. Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments, Tanya Golubchik etal, Mol. Biol. Evol. 24(11):2433–2442. 2007.
10. Directed graphs of DNA sequences and their numerical characterization, Chun Li, Journal of Theoretical Biology 241 (2006) 173–177.

Part II
Artificial Intelligence and Soft Computing

Chapter 8
Generic Document Classification Using Clustering, Centrality, and Voting

Aakanksha Sharaff, Anshul Verma and Hari Shrawgi

Abstract Documents are an indispensable information source used widely to store and access crucial as well as trivial information. To enhance the ease of access of information, prim classification of these documents is a necessity. As documents are used for a variety of purposes and in varying contexts, classification is required to be generic. In this study, a novel approach is proposed which employs score-based voting to classify a document generically using a trained classifier. The classifier can be trained for any set of categories and is thus capable of classifying text documents into generic categories. These categories can be chosen to suit the purpose and requirements of the environment. Our classifier classifies a document based on the scores of its sentences; each sentence has a vote equivalent to its score. Scoring of sentences is performed using clustering and cosine centrality.

Keywords Document classification · Cosine centrality · Clustering
Cosine similarity · Score-based voting · Generic categories

Introduction

The digital world is suffused with documents. Documents are everywhere from opinions to editorials, from reports to research papers, and from applications to offer letters. There is an increasing demand for document classifiers which can categorize and organize documents. The nature and context of a document are not constant, but are contingent upon innumerable factors. Classification of documents thus

A. Sharaff · A. Verma (✉) · H. Shrawgi
Department of Computer Science and Engineering, National Institute of Technology Raipur, Raipur 492010, Chattisgarh, India
e-mail: anshulverma6978@gmail.com

A. Sharaff
e-mail: asharaff.cs@nitrr.ac.in

H. Shrawgi
e-mail: shrawgi.hari@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_8

becomes an intricate task owing to the shifting nature of documents based on their purpose.

Classification is an essential task in the field of machine learning and data mining. It entails labeling the given input into certain categories unambiguously. The caveat is that documents are not strictly categorical and the categories may vary with context and time. The proposed document classifier addresses this issue and categorizes the input documents into generic categories which can be altered to suit the user's needs.

In this work, feature words are extracted to train our classifier according to the categories into which documents are to be classified. Documents from each category are fed to the classifier from which it generates feature words based on TF-IDF model [1] and is thus trained to recognize documents from the respective categories. Further, cosine centrality is used along with clustering to assign score to each sentence in the input document. Then our classifier classifies this document based on its content and ranking of sentences. Higher ranked sentences are given more importance by the classifier; thus they highly influence the overall classification of the document.

Related Work

Automatic document classification [2] is a field of persistent interest since several past decades. This field has been researched continually since several decades with various techniques being implemented to improve classification. Artificial intelligence has been widely used [3]; from clustering to neural networks [4], different concepts have been applied.

Techniques related to automatic document classification or in general text classification are used widely as tools for wide-ranging applications. It is used in spam filtering to help filter out unwanted mails. It is used as an identification tool to discern the language or sometimes the genre of text. Search engines like Google and Yahoo use it to find relevant webpages related to the query provided by the users. Information retrieval is yet another field which has borrowed techniques from text classification and vice versa. Indexing is a concept which is quite similar to classification. The difference lies in the fact that indexing matches subjects with the text rather than the other way around. Indexing leverages some index terms to identify or summarize the documents. Indexing is mainly used in information retrieval [5].

This field has developed from its origins in Library Science. From the works in this field, several conclusions are drawn which continue to be relevant in designing modern classifiers. But the major chunk of development in this field has resulted from data mining [6] and machine learning [7], or in general artificial intelligence. Techniques such as Naïve Bayes Classifiers [8], support vector machines (SVM) [9], and clustering have been widely used to improve text document classification.

Centroid-based technique for document classification is a dynamic approach [10] which outperforms the above-mentioned techniques in general setting. It adapts to varying classes and is adroit for text classification. Our model builds upon this approach but introduces some new aspects to improve its flexibility and robustness.

Methodology

First and foremost, documents related to each category are obtained. Then the classifier is trained for each category based on these documents. A number of feature words are generated for each category; they act as a label for the category. Feature words are generated by a combination of term frequency (TF) and inverse document frequency (IDF). The algorithm for training our classifier and thus generating the feature words is as follows:

Step1: *Preprocessing of training documents*

1a. Perform Stop Words removal
1b. Perform Stemming

Step2: *Calculating TF and IDF*

2a. For each category, perform steps 2b and 2c
2b. Each training document is represented in vector space with each word as a vector
2c. The term frequency (TF) and inverse document frequency (IDF) of all the documents in that category are calculated

Step3: *Selecting Feature Words (label)*

3a. For each category, perform steps 3b, 3c and 3d
3b. P words having highest IDF is selected
3c. From each document, Q words are selected having highest TF
3d. From each document, R words are selected having highest TF * IDF

Where,

N is the total number of feature words
P are the feature words representing all the documents of a category
Q are the most frequent words from each document of a category
R are the feature words from each document with respect to other documents

Now that the classifier is trained, it can classify any document or a set of documents. The classification process begins with preprocessing and then passes through a twofold process of clustering and ranking based on cosine centrality. These ranked features or sentences are then fed to the classifier which labels the document based on these ranked features. The workflow of classification process is represented in the form of flowchart in Fig. 8.1.

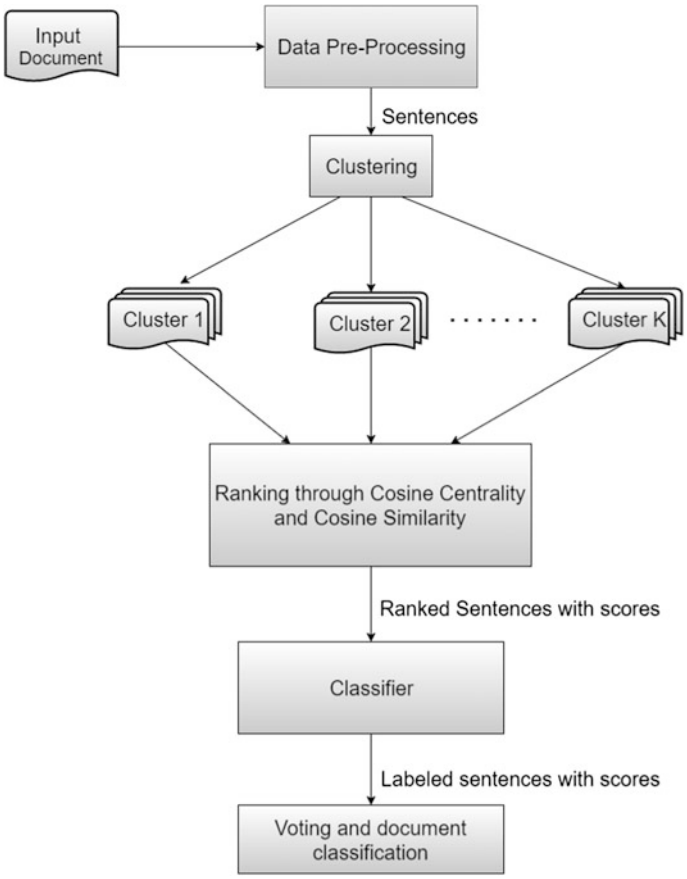


Fig. 8.1 Workflow of classification process of an input document

A. Data Preprocessing

Feature Extraction: Features used in the classifier are sentences and thus we parse/extract sentences from the input document. Sentences are treated as features of a document by the classifier.

Stop Words Removal: Words which are not relevant to the context or do not impart any meaning to the text are called stop words. These are filtered out from text before using the text in the classifier. Removing stop words increases both the effectiveness and efficiency of our process.

Stemming: The meaning of a word is contained in its root word. In a document, there are multiple words with the same root word and it is prudent to consider them all as the same. Thus, reducing all the words into their respective root words is consistent and increases the effectiveness. Reduction of a word to its root word is called stemming. Stemming was performed using the Porter's modified algorithm [11].

B. Cosine Similarity

Cosine similarity is used to measure the similarity between two vectors. It measures the similarity between two vectors on a scale of 0 to 1. 0 means both the vectors are completely different and 1 means they overlap each other. It uses the fact that the cosine of an angle varies from 0 to 1 as the angle varies from 0° to 90°.

The cosine angle between two vectors **X** and **Y** is determined by the dot product of the vectors. The dot product of two vectors is given by the formula

$$\mathbf{X}.\mathbf{Y} = ||\mathbf{X}||*||\mathbf{Y}||*\cos\theta$$

So, the similarity between two vectors is given by

$$\text{similarity} = \cos\theta = \frac{\mathbf{X}.\mathbf{Y}}{||\mathbf{X}||*||\mathbf{Y}||}$$

where, **X** and **Y** are two vectors and θ is the angle between them.

We use term frequency (TF) and inverse document frequency (IDF) for expressing the documents in the vector form and then their relative similarity is calculated by the above formula.

C. Clustering

A document is made of several sentences and together they convey the meaning of the document. Clustering of these sentences based on their similarity would result in different clusters having similar sentences; each cluster conveying a part of the overall gist of the document. Classification of the whole document would depend on these clusters and how the sentences of each cluster are classified.

Clustering is performed using the K-Means clustering algorithm [12] along with cosine similarity as the distance metric [13]. This results in K clusters containing some sentences each.

D. Cosine Centrality and Ranking

Sentences in each cluster are ranked so as to impart a score to each sentence which would govern its influence on the classification. Higher the sentence is ranked; higher would be its vote while classifying the document into categories. Cosine centrality and cosine similarity are used to rank these sentences.

For each cluster, the central sentence is discovered using cosine centrality [14] and it is given a score of 1. The central sentence is the most important sentence of the cluster and it represents the gist of the cluster. Further, all other sentences in the cluster are given a score which is the measure of their similarity to the central sentence. These scores would range from 0 to 1. Sentences are ranked on the basis of these scores within each cluster; central sentence of each cluster would always be the top-ranked in its cluster.

$$CS_K = \text{cosine\_centrality}(C_K)$$
$$\text{Score}(CS_K) = 1$$
$$\text{Score}(S_{iK}) = \text{similarity}(S_{iK}, CS_K),$$

where

C_K	is the Kth cluster
CS_K	is the central sentence of the Kth cluster
S_{iK}	is the ith sentence of the Kth cluster
Score (S_{iK})	represents the score of that sentence
Similarity(x,y)	represents the similarity measure between x and y vectors

E. Classification

Classification of the document would result from the integration of the classification of its constituent sentences. Classification of each sentence would be carried out by our trained classifier. Each word in our sentence would be mapped to the labels of each category generated in our classifier. The category which has the highest number of word mappings to the sentence would be chosen as the category of that sentence.

After ranking each sentence is given a vote based on their score. The trained classifier will classify all sentences into one of the categories. Each sentence then votes for its own category based on their score. Thus, a higher ranked sentence would have a bigger weightage in voting and thus would highly influence the classification. All the votes are added up and then the category with highest score is chosen as the category of the document.

F. Algorithm for Classification

Step1: *Preprocessing*

1a. Extract sentences from the input document
1b. Perform stop words removal
1c. Perform stemming

Step2: *Perform clustering*

2a. Select K random sentences as centroids for each of the K clusters
2b. Repeat steps 2c, 2d, 2e until the process converges that is same set of centroids is selected again
2c. Calculate distance (using cosine similarity) of all the sentences to every centroid
2d. Assign each sentence to the centroid (cluster) that is closest to it
2e. For each cluster, select a new centroid as the sentence which is most similar to others

Step3: *Select Central Sentences*

Step4: *Assign a score (rank) to all the sentences*

4a. For each cluster, do Steps 4b and 4c
4b. Calculate the similarity of each sentence to the central sentence of the cluster
4c. All the sentences have their score as the similarity calculated in the previous step

Step5: *Classification of sentences*

5a. For all sentences in all clusters, do the Steps 5b and 5c
5b. Map each word of the sentence to one of the category labels generated by the classifier
5c. Classify the sentence into the category which is mapped to maximum number of words in the sentence

Step6: *Voting*

6a. Assign a vote to each sentence equivalent to its score
6b. Each sentence votes for its own category selected in Step5
6b. Each sentence votes for its own category selected in Step5

Step7: *Classification of document*

7a. Count the votes for each category
7b. The input document is classified into the category with maximum number of votes.

G. Evaluation

Assigning the utmost importance to correctness of classification is essential for any utilitarian classifier. Thus, we have evaluated different parameters of our algorithm based on the accuracy of classification. Number of feature words selected for each category is one parameter while number of clusters of sentences is another.

Our aim was to achieve the highest accuracy possible by adjusting these two parameters in the proposed model.

Result

The proposed model of text document classification is contingent on two parts: training the classifier and classifying the input document. Both of these were implemented to test the coherence of these two parts. Three categories, namely technology, business, and sports were chosen by us for our experiment. Any set of categories can be chosen to train the classifier to categorize documents, we chose the aforementioned three. We present the result of our experiments according to the

parameter that was varied for a particular experiment. We evaluate accuracy of classification for varying parameters to assess the performance of our classifier.

The critical part of training the classifier is selecting the number of feature words (N) to be generated for each category. It directly impacts the classifiers' performance as the more feature words it can generate, better information it would have about that particular category. As can be seen from Fig. 8.2, accuracy increases as the classifier is allowed to generate more feature words.

During classification, a document is clustered into sentences, and this represents the pivotal part of the classification process. As number of clusters (K) is allowed to increase, the variability of topic within a document increases. K should be high enough to allow the relevant topics to come to the fore, but low enough to curb the importance of irrelevant topics. Thus, we see in Fig. 8.3 that accuracy increases with increasing K up to a saturation point, but then steadily decreases. Plummeting of accuracy can be understood by considering the many topics that these different clusters represent. If we increase clusters then unimportant topics would also have significant influence and thus would decrease the accuracy of classification.

The average accuracy achieved by this model is lower than some of the state-of-the-art classifiers, but our experiments are based on a relatively smaller dataset. As the size of dataset would increase, our classifier would have better training and thus its accuracy would increase. But our experiments do indicate that this approach can be leveraged to classify documents into varied categories which is an essential requirement of this age.

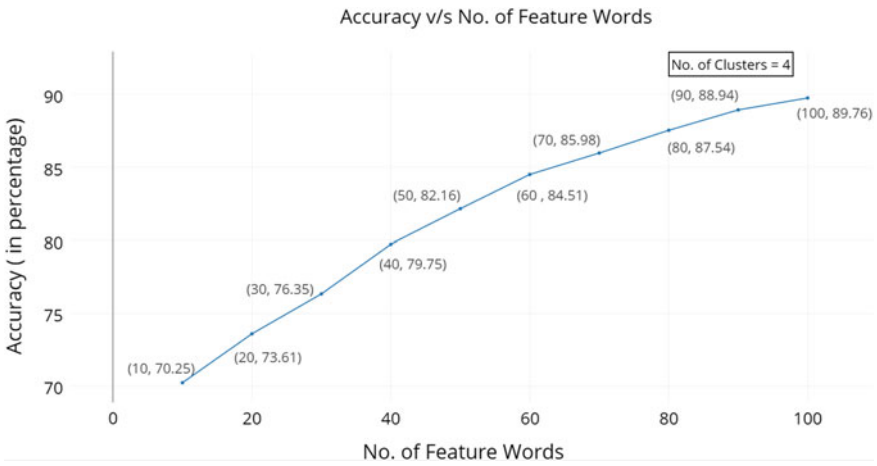


Fig. 8.2 Measure of accuracy with varying number of feature words generated

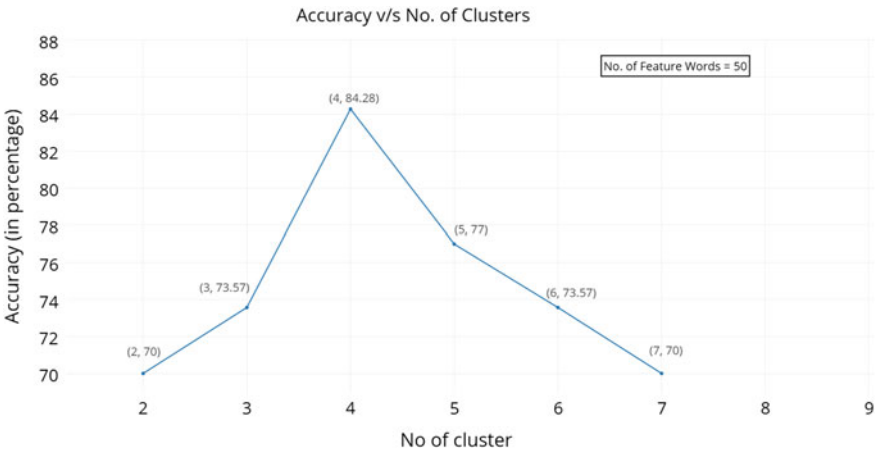


Fig. 8.3 Measure of accuracy with varying number of clusters in a document

Conclusion and Future Work

Above results show that the proposed technique is able to effectively classify the documents into the three selected categories viz. Technology, Business and Sports. The proposed model also incorporates a lot of parameters such as the no. of feature words, no. of clusters, similarity threshold etc. Presence of such parameters provides this model with customizability and adaptiveness. The model can be easily calibrated to be used in different contexts and setting by adjusting the said parameters.

This study evaluates the proposed model on the basis of accuracy. More evaluation measures like precision, recall and F-measure can be used to provide insights into the scalability and robustness of the model. Three categories have been used in this work, but more categories with more test cases can be used to further establish the performance of the proposed model. The study can also be extended by providing a comparison of performance with other suggested models with similar objectives.

References

1. Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), 13. (2008).
2. Borko, H., Bernick, M.: Automatic Document Classification. Journal of the ACM. 10, 151–162 (1963).
3. Bilski, A.: A Review of Artificial Intelligence Algorithms in Document Classification. International Journal of Electronics and Telecommunications. 57, (2011).

4. Manevitz, L., & Yousef, M.: One-class document classification via neural networks. Neurocomputing, 70(7), 1466–1481. (2007).
5. Maron, M. E., & Kuhns, J. L.: On relevance, probabilistic indexing and information retrieval. Journal of the ACM (JACM), 7(3), 216–244. (1960).
6. Tsoumakas, G., & Katakis, I.: Multi-label classification: An overview. Dept. of Informatics, Aristotle University of Thessaloniki, Greece. (2006).
7. Sebastiani, F.: Machine learning in automated text categorization. CSUR. 34, 1–47 (2002).
8. Kim, S., Han, K., Rim, H., Myaeng, S.H.: Some Effective Techniques for Naive Bayes Text Classification. IEEE Transactions on Knowledge and Data Engineering. 18, 1457–1466 (2006).
9. Manevitz, L. M., & Yousef, M.: One-class SVMs for document classification. Journal of Machine Learning Research, 2(Dec), 139–154. (2001).
10. Han, E., Karypis, G.: Centroid-based document classification: Analysis and experimental results. European conference on principles of data mining and knowledge discovery. Springer Berlin Heidelberg. (2000).
11. Willett, P.: The Porter stemming algorithm: then and now. Program: electronic library and information systems. 40, 219–223 (2006).
12. Huang, A.: Similarity measures for text document clustering. Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand. (2008).
13. Hartigan, J., Wong, M.: Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1, 100–108 (1979).
14. Erkan, Radev, D.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research. (2004).

Chapter 9
A Game for Shared Ontology Evolution

Dipika Boro and Zubin Bhuyan

Abstract With the advent of the *Semantic Web* and emergence of extremely large knowledge repositories, community-driven ontology development and support has become inevitable. However, such multi-agent paradigms can be prone to unwarranted and unjustified modifications by *individually rational* or *selfish* agents. Such changes can result in an inaccurate modeling of the domain being represented. We propose a framework, based on an adaption of the social *Ultimatum Game*, for restricting such illegitimate modifications to shared ontologies. Simulation of this model is done on two standard ontologies, and the results show that the proposed framework is able to segregate update requests from *legitimate* and *rogue* agents.

Keywords Ontology evolution · Shared ontology · Ontology maintenance

Introduction

Ontologies have emerged as the prime mover of the development and deployment of Semantic Web, enabling efficient machine–machine, human–machine, and even human–human communication [1, 2]. It has found usage in diverse fields such as AI, robotics, bioinformatics, and astronomy. With the sizes of existing ontologies growing at such a fast pace, it has become inevitable that such knowledge resources be developed and sustained in a decentralized, community-driven manner by humans and computer systems. Web Protégé [3] is such a tool designed for collective ontology development. For a survey of collaborative ontology building methodologies, readers are referred to [4].

D. Boro (✉)
Department of Computer Science and Engineering, NETES Institute of Technology and Science, Mirza, India
e-mail: dipika@nitsmirza.ac.in

Z. Bhuyan
Biomimetic and Cognitive Robotics Lab, Department of Computer Science and Engineering, Tezpur University, Tezpur, India
e-mail: zubin@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_9

Collaborative ontology maintenance leads to the issue of trust among the various agents involved in the enterprise. The framework presented in this paper deals with such multi-agent scenarios where there is a possibility of individuals transgressing, and leading to lapses in the ontology being shared.

In the following section, we shall discuss ontology as a domain representation tool and some basic game theoretic concepts. We shall focus on the *ultimatum game* as the work presented in this paper bears resemblance to a social form of the game. In section "Proposed Framework for Evolution of Shared Ontology", we explain our framework and how it is able to discriminate *fair* ontology updates from the *unfair* ones. Section "Simulation Setup and Results" describes the simulation setup and the results obtained. Thereafter, we conclude with a note on future direction for this research.

Background

Ontology for Domain Representation

A description of a domain via its existent concepts, attributes of these concepts, and constraints on attributes comprises an ontology. Ontologies often also have *individuals* belonging to the defined concepts. Gruber in [5] describes *ontology* as

> "a description of the concepts and relationships that can formally exist for an agent or a community of agents".

Ontology change refers to modifications to an ontology, due to the need for improved conceptualization of the domain of discourse, a change in user requirement, or a dynamic emendation of the domain's understanding [6]. A process-centric survey of ontology evolution is presented Zablith et al. in [7].

Game Theory

Game theory is the study of strategic decision-making of intelligent agents. It provides us with mathematical structure of investigation for arriving at rational decisions in competitive environments. Few of the areas where game theory has been applied include biology, battlefield analysis, economics, and computer science.

A *game*, within the bounds of game theory, is a mathematical object which is usually used to model scenarios to study their potential solutions or equilibrium states.

The Ultimatum Game The Ultimatum Game, first proposed in [8], is an experiment to analyze bargaining behavior of two participating agents. In this game, a sum of money has to be divided between two players, *the proposer* and the second player, *the responder*. *The proposer* proposes how the amount should be divided between

them; *the responder* then can either accept or reject this proposal. If the *responder* accepts, the money is split according to the proposal. But if he or she rejects the proposal, neither player receives any money. Several versions and extensions of this game have been proposed and studied extensively.

Our proposed framework is also based on an adaptation of the Ultimatum Game. We consider the agents involved in the ontology development and maintenance to be players. Whenever an agent proposes a change to be made in this shared ontology, the proposal is put forward for consideration by the other agents. Only when a predetermined fraction of the agents agree to the change, will it be executed.

Proposed Framework for Evolution of Shared Ontology

Motivation

It is often advantageous to build and maintain an ontology in a decentralized manner, with contributions from a distributed community of individuals or agents, as it ensures fast, parallel, and efficient operations. However, such multi-agent paradigms can be susceptible to schemes of *individually rational agents* or selfish agents, leading to an inaccurate or undesirable global ontology which fail to fully describe the domain being modeled. An agent might prefer a specific part of the domain to be represented in a particular way so that it might have certain minimal advantages, regardless of how it affects the community.

The framework presented in this paper addresses this problem of unfair ontology modification in a multi-agent paradigm. The stakeholder agents are considered together as a community which interacts among one another before deciding whether to execute an *update request*.

Modeling Ontology Evolution as a Game

We propose a novel framework for ontology evolution for scenarios where multiple agent attempt to modify or update parts of a shared ontology. Agents submit requests for modification to a central *Update Manager*, which then initiates the process of verifying whether the request should be carried out. This is illustrated in Fig. 9.1.

We consider an ontology $\mathbb{O}$, and a community of agents $\mathbb{P} = \{P_1, P_2, \ldots, P_N\}$. Agent P_z requests the *Update Manager* to incorporate a change, ρ, to the ontology. The *Update Manager* identifies O, the segment of the ontology on which ρ is to be applied. It then informs all stakeholder agents of the sub-ontology O about the proposed modification of O to O^{ρ}. This we represent by the notation $O \xrightarrow[P_z]{\rho} O^{\rho}$.

As shown in (9.1), an utility function, *Util*() is defined for each agent to compute a utility score of a modified ontology for that agent.

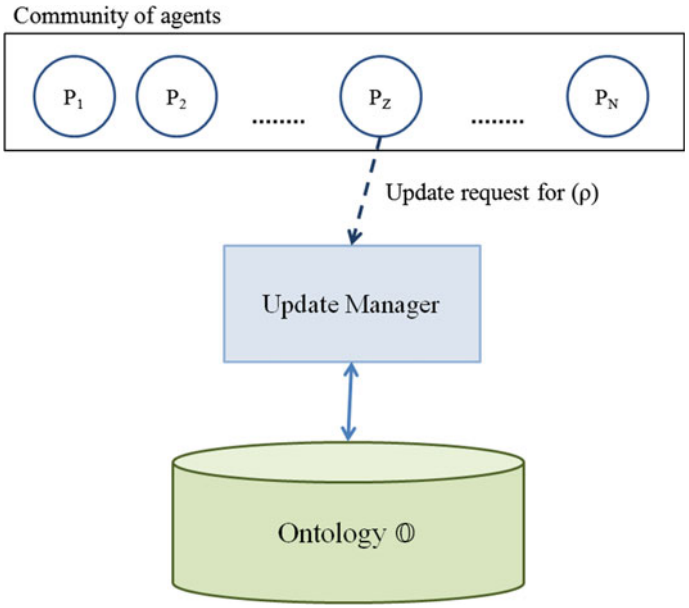


Fig. 9.1 Basic block diagram of the framework

$$\phi_i^{\rho} = Util_i(O^{\rho}) \tag{9.1}$$

ϕ_i^{max} is the maximum payoff or utility an agent P_i can receive for modified sub-ontology $O\rho$. This, however, may not be acceptable to others as it might be so that utilities are computed for other agents are not same, i.e., for any agent P_i and P_j, it is possible that $\phi_i^{\rho} \neq \phi_j^{\rho}$.

If, for an agent P_i, the difference is tolerable, i.e., $(\phi_i^{max} - \phi_i^{\rho}) < \tau$, it votes for the proposed change, else the agent rejects it. We denote this decision as a binary variable $d_i^{\rho} = \{0, 1\}$, where 1 is assigned for acceptance of a request, and 0 otherwise. If the number of agents accepting is greater than a predefined threshold, λ, the change is adopted and necessary modifications to the ontology is made. This process is presented in Algorithm 1.

Simulation Setup and Results

We have simulated our algorithm with two standard ontologies, viz., the wine ontology[1] and the countries ontology.[2] The simulation environment was programmed in

[1] https://www.w3.org/TR/owl-guide/wine.rdf.

[2] https://www.w3.org/Consortium/Offices/Presentations/RDFTutorial/rdfs/Countries.owl.

Algorithm 1 Ontology Update Game

procedure ONTOLOGY–UPDATE–DECISION
  Identify O from $\mathbb{O}$ for change ρ_z
  Compute set of stakeholder agents P_O from $\mathbb{P}$
  $O^{\rho} =$ Change in ontology for change ρ_z
  $D = 0$
  for each agent $P_i \in \{P_O - P_z\}$ **do**
    $\phi_i^{\rho} = P_i.Util_i(O^{\rho})$
    if $(\phi_i^{max} - \phi_i^{\rho}) < \tau$ **then**
      $D = D + 1$
    end if
  end for
  if $D > \lambda$ **then**
    Commit proposed change
  else
    Reject proposed change
  end if
end procedure

Python 3.5. For both the ontologies, we iteratively increased the number of agents and recorded the observations.

In a real-world scenario, agents would be exploring the domain or monitoring any observable changes. These observations, when presented as an update request to the *Update Manager*, would be supported by *legitimate* peer agents. However, an unwarranted change requests from a *rogue* agents would be supported by none or a very small number of agents.

Before commencing with the simulation, we removed small sections of the ontology, and each agent was assigned as a stakeholder for a part of this ontology. The sections removed comprised roughly of 25% of the ontology primitives. A 5–10% of these agents were randomly marked as *rogue*, and another 15–30% agents were designated as *legitimate* modifiers. The *rogue* agents would attempt to make unwarranted changes to the ontology. The *legitimate* agents were equipped with some partial information about the missing parts which enabled them to judge, with some accuracy, whether an update request was legitimate or not. Based on this information, they were also able to propose valid changes as well.

We compare such a multi-agent scenario with and without the proposed Ontology Update Game. We compute the accuracy score based on how similar the resulting ontology is compared to the respective unaltered ontology. For each setting, the agents would repeatedly send update requests to the *Update Manager*. The accuracy value for a setting was calculated by averaging the value obtained for three independent executions.

Figures 9.2 and 9.3 show how the accuracy varies with increasing number of agents. It is observed that initially when the number of *rogue* agents are less, they do not have a considerable impact on the ontology. However, as the number of agents increases, our proposed framework performs far better than the situation where there is no policy for restricting unwarranted updates. With 120 agents, the accuracy for

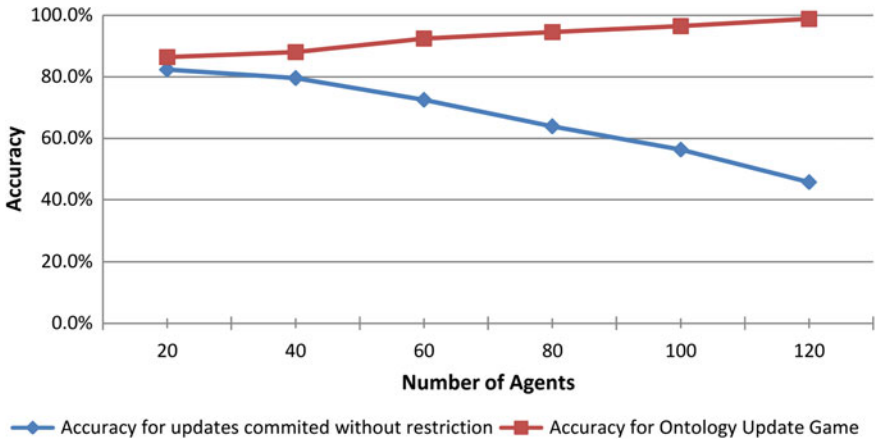


Fig. 9.2 Accuracy of updating the wine ontology with varying number of agents

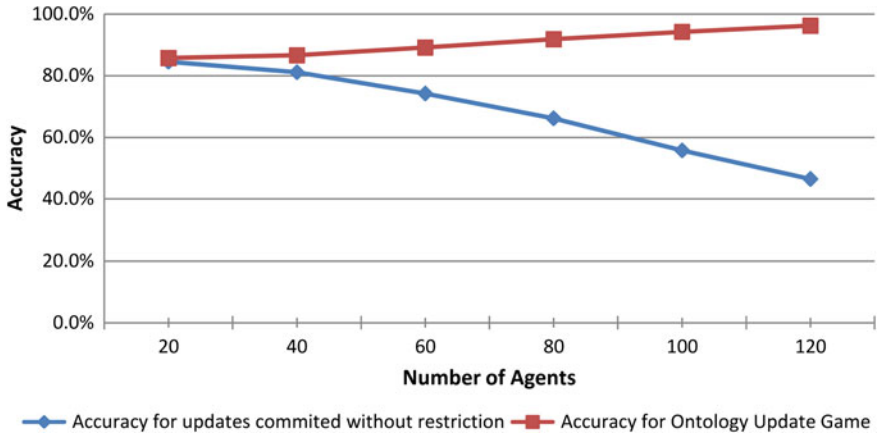


Fig. 9.3 Accuracy of updating the country ontology with varying number of agents

the Country Ontology stands respectively at 46.4 and 96.2% for updates without restriction and updates in our framework. These values for the Wine Ontology are 45.7 and 98.8%, respectively.

From these observation, we can infer that for a shared ontology, with a significant number of agents involved in its evolution, it is essential to have a moderation policy which can arrest unnecessary and unjustified updates.

Conclusion and Future Directions

We have presented a framework for updating ontology in a multi-agent paradigm which enables stakeholder agents to work together and distinguish changes which benefit the community from the malfeasance ones. It was also shown that it performed considerably better than the scenario without a mechanism for controlling undesirable modifications.

We plan to extend this work by incorporating the notion of *reputation* and *notoriety*. An agent consistently making fair offers of change has a higher reputation than a *selfish* agent. The *reputation* score can be used to dynamically decide the number of agents required to approve a proposed change, i.e., value of λ.

References

1. Horrocks, I., F. Patel-Schneider, P., L. McGuinness, D. and A. Welty, C. OWL: A Description Logic Based Ontology Language for the Semantic Web. In *The Description Logic Handbook (2nd Ed.).* Cambridge Univ. Press, 2007.
2. Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P. and Sattler, U. OWL 2: The next step for OWL. *J. of Web Semantics*, 6(4):309–322, November 2008.
3. Tudorache, T., Nyulas, C., Noy, N. F. and Musen, M. A. Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semant. web*, 4(1):89–99, 2013.
4. Simperl, E. and Luczak-Rösch, M. Collaborative ontology engineering: A survey. *The Knowledge Engineering Review*, 29(1):101–131, 005 2013.
5. Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
6. Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D. and Antoniou, G. Ontology change: Classification and survey. *Knowl. Eng. Rev.*, 23(2):117–152, 2008.
7. Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D. and Sabou, M. Ontology evolution: a process-centric survey. *The Knowledge Engineering Review*, 30(1), 45–75, 008 2013.
8. Guth, W., Schmittberger, R. and Schwarze, B. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3(4):367–388, 1982.

Chapter 10
An Adaptive Genetic Algorithm Based Neural Network for Traversability Analysis Under Uncertain Terrain Conditions

N. Samanvita and H. M. Ravikumar

Abstract The high pace emergence of technologies, hardware availability and primarily vision-based decision processes, has given rise to a new engineering dimension called autonomy that employs traversability assessment and path planning strategies for navigation. Autonomy serves in an array of applications ranging from home or industrial automation to space exploration missions. To enable effective navigation, terrain classification and its traversability analysis are of great significance. In this paper, a novel evolutionary computing (EC)-based traversability analysis scheme has been developed. The proposed scheme at first applies 2D discrete wavelet decomposition (2D-DWPD) schemes for feature extraction, which is followed by gray-level co-occurrence metrics (GLCM) estimation. Four terrain features, energy, contrast, correlation, and entropy or homogeneity, have been retrieved from estimate GLCM statistical information. Further, an adaptive genetic algorithm based neural network (AGA-ANN) classifier has been developed that classifies the sub-terrain regions as traversable or non-traversable. Performance evaluation with NASA's Mars rover datasets has affirmed that the proposed AGA-ANN scheme outperforms existing K-means clustering-based terrain classifier. The results affirm that the proposed scheme can be used for traversability analysis of terrains under uncertain conditions.

Keywords Autonomous navigation · Traversability · Terrain classification Adaptive genetic algorithm · Neural network · K-means clustering

N. Samanvita (✉) · H. M. Ravikumar
Department of Electrical and Electronics Engineering, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India
e-mail: Samanvita_n@yahoo.com

H. M. Ravikumar
e-mail: hmrgama@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_10

Introduction

The high pace emergences in technologies and associated applications have given rise to numerous advance techniques, such as autonomy or autonomous navigation that intends to serve civil automation utilities, major industries, defence as well as space missions. In the present scenarios, robotics and autonomous robots have become exponentially significant for major utilities serving automotive and manufacturing industries, defence, surveillance applications, battlefield reconnaissance, medical, space technologies, scientific research, and numerous household utilities. Considering upper end applications, such as space exploration mission, no doubt autonomy is one of the vital technologies required. Numerous space agencies such as NASA, ISRO, etc. are working toward developing efficient autonomy to enable optimal autonomous navigation, safe path planning, and route decision. Traversability analysis is one of the driving mechanisms that apply obstacle detection and avoidance, terrain classification and change detection, object recognition techniques, etc., for efficient autonomous navigation and path planning.

The most critical activity in autonomous robotic navigation system, particularly in spacecraft and rovers, is the real-time exploration and analysis of terrain under uncertain conditions. The real-time assessment and quantification of terrain environment enables safe navigation toward destination. In any space exploration mission, the autonomous rover intends to reach certain targeted planetary surface and requires real-time terrain assessment and safety analysis. On the other hand, in such missions, the terrain's environmental conditions do vary. In case of rough and deformable terrain, the insufficient navigational strategies might result in hazards and hence loses mobility. The incidents such as NASA's robotic mission (April 29, 2005 and May 30, 2006, NASA/JPL) are the key evidences where rover got trapped in loose drift material. Such jeopardize results in huge economical and strategical looses. To avoid such losses, it is significant to develop a novel terrain assessment approach for exploring the real-time terrain and identify the best traversable path for navigation. To deal with such issues, traversability analysis under uncertain terrain environment has emerged as one of the advanced areas for research community. Our work is motivated by such issues and intends to develop a novel terrain classification and traversability analysis model for optimal path planning and autonomous navigation. To enable safe landing and navigation, a number of embedded techniques incorporating sensors, transducers, battery-powered actuators, etc. are applied. Still, enabling efficient and safe autonomous navigation cannot be accomplished without certain effective artificial intelligence approaches to assess the real-time terrain's traversability conditions. In many situations, the autonomous navigation rovers are required to navigate through different complex and unpredictable environmental conditions. Rovers require the prior assessment of natural terrain and its quality, also called terrain features so as to avoid obstacles and make optimal path planning. To achieve these needs, the effective traversability analysis scheme can play a vital role. It can significantly classify the natural terrain

as navigable or non-navigable and the rough terrain could be avoided for speed optimization.

In this paper, a highly robust evolutionary computing-based AI technique for natural terrain classification and traversability analysis has been proposed. Our proposed scheme exploits various textural features of the natural terrain and derives significant statistical and transform-based analysis approaches for terrain analysis. Implementing two-dimensional discrete wavelet packet decomposition (2D-DWPD) algorithm, gray-level co-occurrence metrics (GLCM) has been retrieved that provides four key features of the terrain. These features are energy, contrast, entropy/homogeneity, and correlation. Here, an enhanced evolutionary computing algorithm named adaptive genetic algorithm (AGA)-based artificial neural network (ANN) algorithm has been proposed that significantly enhances the terrain classification and traversability analysis under uncertain terrain conditions. The performance analysis of the proposed system with different MARS mountain datasets has revealed that the proposed scheme can perform better for real-time traversability analysis and path planning.

The other sections of the presented manuscript are divided as follows. Section 2 represents the discussion of the related works, while our proposed methodology is discussed in Sect. 3. Results obtained are presented in Sect. 4, which has been followed by conclusion and future work discussion in Sect. 5. Reference section presents the cited reference considered in this research.

Related Work

Traversibality analysis and terrain classification enables semantic characterization of certain natural terrain region. Such characterization can be allied with numeric physical constraints or parameters or traversability estimates that can significantly enhance the traversability classification and prediction. Numerous techniques have been introduced for terrain assessment and traversability analysis on the basis of features extracted from remote sensing devices. Some of the key features applied are the textural features, color, surface geometry, etc. In addition, terrain characterization using vibration from different sensors, ultrasonic range finders, and vision sensors has also been explored by researchers. Researchers have derived path planning strategies based on these significant terrain features. A fuzzy traversability index estimation algorithm was developed in [1], where the visual intensity levels were used to estimate terrain characteristics like slopes, surface quality or roughness, and path disruption. An optimization effort was made in [2], where authors developed a classification model intending to avoid the influence of the variation in the light intensity. They used three-dimensional ladar sensor data for terrain feature extraction. In fact, they implemented it as the offline statistical analysis paradigm so as to segment certain given terrain region into three classes, terrain rock, vegetation, and tree branches or instrumental wire. Researchers [3] applied the two-dimensional (2D) information such as geometric constraints and classified the terrain as

navigable or non-navigable. They applied hidden markov models (HMM) for terrain prediction. In [4] researchers applied local features extracted from a three-dimensional (3D) ladar scans. To classify the terrain features, they applied Naïve Bayes decision tree algorithm. Considering the significance to incorporate multiple features, researchers in [5] used both images and ladar data information for obstacle detection. Further, the laser range finder based range data was processed using Hough transform in [6], where they represented feature planes into three dimensions and used decision tree algorithm for terrain classification. Still, their approach could not identify the non-geometric terrain segment as these cannot be characterized in terms of geometrical variations. An AI-based scheme was developed in [8], where researchers applied rule-based and ANN for visual terrain modeling. An effort was made in [9] where researchers applied visual stereo imaging fusion scheme for terrain mapping and robotic perception modeling. To enable terrain classification and traversability analysis, the interaction between locomotive system and natural terrain was proposed in [10], where researchers suggested vibration generated from the robotic wheel–terrain interaction for terrain quality assessment. A similar effort was made in [11], where frequency response of the vehicle vibrations was used to characterize the terrain. These approaches assume that vibration in the vehicle is correlated with the present terrain conditions. They employed the correlation concept between offline data recorded for each terrain and the real-time data recorded through camera. However, these approaches cannot be considered as effective scheme because of its limited future prediction ability and mobility control under uncertain conditions.

Color as the terrain characterizing feature was used in [12], where the results were found satisfactory. Researchers explored the multispectral imaging features, such as color spaces and its distribution for traversability assessment. Color as the significant terrain signifier was applauded by [14] while assuming that different terrains might have distinct color signatures. Similarly, the textural features were suggested in [15] for traversability index estimation and terrain classification. In [15], researchers applied co-occurrence features along with clustering algorithm for terrain classification. They implemented linear discriminate classifier and ANN-based clustering for terrain classification. Since conventional ANN is found to be suffering due to the issues like local minima and convergence and hence for critical utilities such as autonomous rover navigation, it demands further optimization. In [16], textural gray-level co-occurrence features along with crisp rule-based classifier were applied for terrain classification. To enhance classification, an effort was made by introducing K-means clustering for terrain classification [17]. The implementation of efficient terrain characteristics or features along with enhanced AI-based classifier can enable better traversability assessment and path planning for safe autonomous rover navigation.

Our Contribution

In this paper, a robust terrain classification and efficient traversability analysis scheme have been developed for autonomous rover navigation under uncertain terrain conditions. To classify terrains, the four significant features, energy, contrast, correlation, and entropy or homogeneity, have been considered. The proposed scheme encompasses two sequential phases. The first phase deals with feature extraction followed by the implementation of evolutionary computing scheme, AGA for terrain classification and prediction in the second phase. In our proposed model, we have applied textural analysis paradigm that exploits visual imagery terrain data. In fact, texture signifies the local spatial variation in the intensity of the terrain image. Various approaches like statistical means and transform techniques can be used for textural analysis and its characterization. Statistical approaches employ various cumulative gray value probabilities. Here, the gray-level co-occurrence matrices (GLCM) have been considered that estimates the second-order statistics by means of estimating the iterations or the occurrences of all the gray values and the dislocations in the terrain data (image). The major texture attributes such as energy, entropy or homogeneity, and contrast are derived from GLCM matrix. On the other hand, model-based approach encompasses model fitting schemes such as Markov random field, autoregressive, fractal, etc. The retrieved parameters are further applied to perform segmentation and classification of the textural features. The transform-based scheme applies different transform domains using wavelet decomposition or Gabor filter techniques to analyze the texture.

In this paper, we have applied wavelet transform technique to explore the textural features, as stated above. Retrieving the textural features, the proposed AGA-based ANN (AGA-ANN) has been executed for terrain classification. In our proposed method, at first the terrain image is divided into certain finite sets (here 400) of sub-frames where the individual frame signifies a segment of given natural terrain, also called sub-terrain region. It is then followed by the feature extraction from each frame using two-dimensional wavelet packet decomposition (2D-DWPD) techniques. The statistical and wavelet transform-based approach enables feature extraction that yields four feature vectors in terms of energy, contrast, correlation, and entropy. These extracted feature vectors are fed as the input to the classification model for traversability analysis, whether the terrain region is navigable or not. It should be noted that the above-discussed process is implemented for each sub-terrain region (400) that eventually classifies that terrain region as traversable or non-traversable. Unlike conventional approaches, we have applied AGA-ANN classifier for terrain classification and traversability analysis. The overall proposed terrain assessment and traversability analysis scheme are presented in Fig. 10.1.

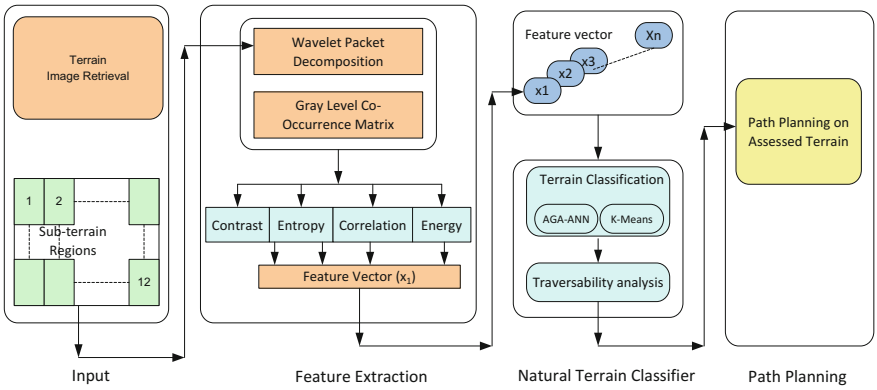


Fig. 10.1 Proposed traversability analysis model

A brief of the proposed traversability analysis and terrain classification process is discussed in the following sections.

A. Terrain Image Retrieval

Nowadays, internet of things (IoT) has given a new broad dimension for advanced computation techniques enabling an array of applications. For natural terrain assessment, the terrain image can be taken from camera or sensor-enabled IoT device setup. In this paper, the standard NASA's Mars Mountain data (Fig. 10.9a) has been applied, which has been obtained from NASA Mars mission information center.

B. Feature Extraction

In our method, the wavelet decomposition method has been applied for feature extraction, which has been implemented in each sub-terrain region of the image. Implementing the wavelet transform, the gray-level co-occurrence metrics (GLCM) has been retrieved for each decomposed sub-band of the terrain region. In our approach, we have extracted four textural features: energy, contrast, entropy/homogeneity, and correlation, which have been further used for terrain classification. A brief of the implemented wavelet decomposition method and GLCM estimation is given as follows:

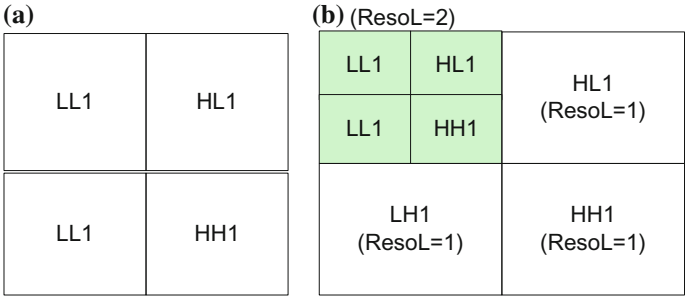


Fig. 10.2 Terrain image decomposition in different sub-bands: **a** single level **b** two level

(a) *DWT-Based Wavelet Packet Decomposition*

For numerous signals, the low-frequency components encompass major significant information. In such cases, it is inevitable to emphasize on the bands having higher energy rather exploring low-frequency bandwidth or low-energy bandwidths. It results in efficient wavelet packet adoption. We have implemented the 2D discrete wavelet packet decomposition (2D-DWPD) algorithm for feature extraction. In our model, at first, the signal has been passed through multiple filters. The proposed 2D-DWPD algorithm decomposes the terrain image into four sub-bands which are defined in terms of LH1, HL1, HH1, and LL1. LH1, HL1, and HH1sub-bands signify the granular level wavelet coefficients. On contrary, LL1 represents the coarse level coefficients (Figs. 10.1 and 10.3).

2D-DWPT scheme divides terrain image into three information (details) image components and one approximation image. Here, the approximation image is

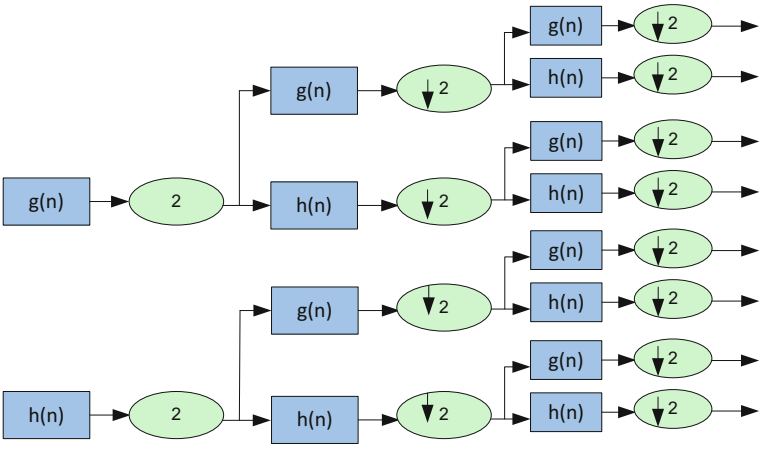


Fig. 10.3 Discrete wavelet packet decomposition

further divided into a second-level estimate and detail images. This mechanism continues iteratively. In such manner, there are $(n+1)$ feasible ways for decomposing the terrain image for n-level decomposition. Performing this iterative process, the terrain features have been obtained.

C. Gray-Level Co-occurrence Matrix

Gray-level co-occurrence matrices (GLCM) scheme that counts the frequencies for all the gray value pairs and all dislocations in the input terrain data (image) has been used to estimate the second-order statistics. Generally, an image can be represented in terms of the matrix of pixel intensities $I(x,y)$ and therefore, we have estimated the co-occurrence $P_d(i,j)$ of the input image in such a way that for each input value $P_d(i,j)$ is the difference in the intensity between a pair of image pixel i and j. In fact, this difference is equivalent to the d pixel apart in the terrain image in certain defined direction. We have estimated four textural features, energy, contrast, entropy, and correlation, as feature vectors to be used for terrain classification.

We define texture contrast as the amount of variations in the local pixel intensities in the terrain image. Mathematically, it is presented by

$$Contrast = \sum_i \sum_j (i-j)^2 P_d(i,j) \tag{1}$$

In this paper, the textual feature energy has been estimated that signifies the textural uniformity of the terrain image data. Mathematically, it can be represented as

$$Energy = \sum_i \sum_j P_d^2(i,j) \tag{2}$$

Entropy, which signifies the disorder in the image, has been considered as inversely proportional to the energy. Mathematically, it is presented as

$$Entropy = -\sum_i \sum_j P_d(i,j) \log P_d(i,j) \tag{3}$$

It should be noted that homogeneity also refers the uniform pixel intensity distribution and hence entropy is related to the homogeneity of the terrain image. Hence, in this paper, both the entropy and homogeneity have been applied for terrain feature mapping and further classification. To enable optimal terrain assessment, it is vital to examine the dependency or the relation between pixels of the terrain image. With this objective, we have estimated a parameter called correlation that estimates the linear dependency of the gray-level values in GLCM statistics. In fact, it exhibits correlation between reference pixels with its neighbor. Mathematically, correlation feature is estimated by

$$Correlation = \left\{ \sum_i \sum_j (i,j) P_d(i,j) \right\} - \mu_x \mu y \ / \sigma_x \sigma_y, \tag{4}$$

where μ_x, μ_y and σ_x, σ_y represent the mean and the standard deviation of $P(i,j)$, respectively.

The aforementioned features have been estimated for all the sub-terrain sectional regions that functions as the feature vector for the proposed terrain classifier. In addition to the aforementioned four terrain features, we have employed two additional features, which are the statistical features of the wavelet comprising energy level of the leaf nodes of the wavelet packet and features estimated from the original sub-terrain regions of the input natural terrain image. Employing these all feature vectors, we have performed classification to classify the terrain region as traversable or non-traversable. The discussion of the terrain classifiers is given in the following sections.

D. Natural Terrain Classification

Clustering has been a potential approach for pattern- or feature-based classification. It classifies the data based on similarity and dissimilarity between data elements or respective features. It has been considered as a robust technique for data analysis, prediction, system modeling, etc. Considering robustness of this approach, in this paper, two different clustering-based classifier models have been developed. These are

(a) *K-means clustering*
(b) *Adaptive genetic algorithm based ANN* (*AGA-ANN*)

(a) ***K-Means Clustering***

In the first step of terrain feature classification, we have performed K-means clustering-based feature classification that partitions n terrain features into k clusters, where $k < n$. Our implemented K-means model selects k initial cluster centers randomly as well as by estimating the means. In process, the cluster centers are updated in such manner that after certain iterations they signify the clusters in the data as the optimal cluster. Considering functional characterization, K-means algorithm initiates with the selection of initial k cluster centers, also called centroides. The initial centroid selection can be done randomly obtained from certain priori information. The individual data index or point is allocated to the nearest centroids. Eventually, the centers are re-estimated as per allied data points. This process continues iterating till convergence. In general, within a cluster, the Euclidean distances between data points are used to be small and are supposed to be

Algorithm for K-Means Algorithm based feature classification

1. Randomly select C data points and estimate the initial cluster center or the centroid.
2. Estimate the membership matrix$U$, where the data element$U_{i,j} = 1$ if the jth data index x_j belongs to the group 1 and 0 otherwise.
3. Estimate the cost function using following equation:
$$J = \sum_{i=1}^{c} J_1 = \sum_{i=1}^{c} \left(\sum_{k, X_k \in C_i} \|X_k - C_i\|^2 \right)$$
4. Stop iteration if current cost function is below a defined threshold level.
5. Assign centroid to each data points.
6. Update the centroid of the clusters C_i by re-estimating it as the mean of all comprising data points within individual cluster and update U matrix.

Fig. 10.4 Pseudocode for K-means based terrain classification

allied with one centriod vector that signifies the "midpoint" of that cluster. In this research work, the mean of the data indexes belonging to the associated cluster has been used as the centriod vector. The pseudocode of the implemented K-means clustering-based feature classification is given as follows: (Fig. 10.4).

This is the matter of fact that K-means clustering has been a potential alternative for data clustering and prediction, and still the robustness of ANN cannot be ignored. Considering the efficiency as well as limitations of ANN, in this paper, an adaptive genetic algorithm (AGA)-based neural network classifier (AGA-ANN) model has been developed. A brief discussion of the proposed AGA-ANN has been developed.

(b) ***Evolutionary Computing-Based Terrain Classification***

The artificial neural networks (ANN) algorithm has been a potential technique serving numerous processes ranging prediction, classification to complex decision processes. However, it suffers from local minima and convergence issues which significantly confine its applicability. ANN applies different learning algorithms like Gradient Descent (GD), Gauss–Newton (GN), and Levenberg–Marquardt (LM) for learning and classification. Interestingly, most of these algorithms are unable to avoid issues like local minima and convergence. To deal with these issues, the selection of optimal weights can be of great significance. With this objective, we have developed an AGA algorithm to estimate the dynamic weights of the ANN-based terrain classifier. To classify the terrain region as traversable or non-traversable, the multivariate regression technique has been applied that intends

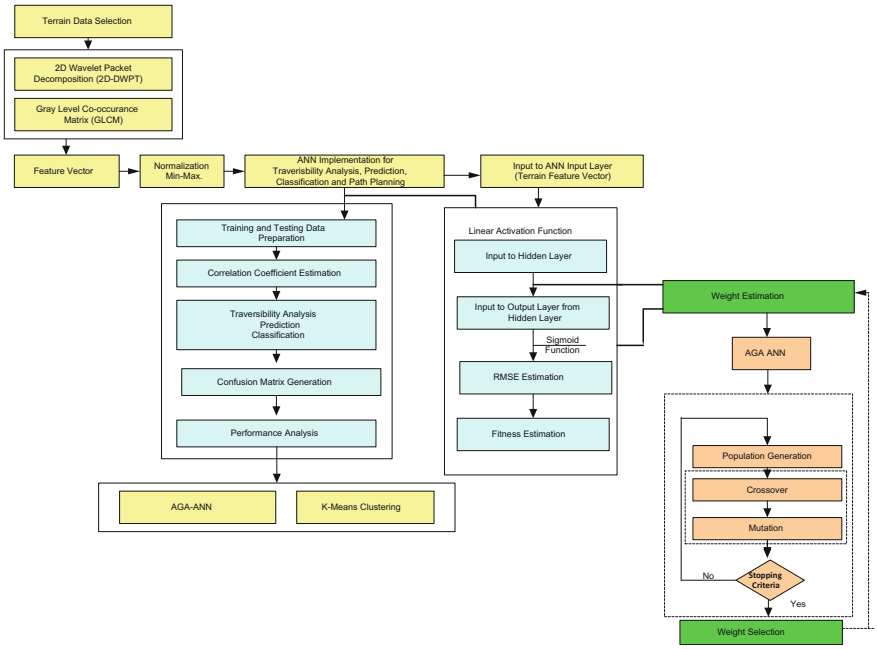


Fig. 10.5 AGA-ANN based terrain classification

to estimate relation between terrain features and the possibility of the traversability. Here, terrain features (contrast, energy, correlation, and entropy) are the independent variable while traversability index represents dependent variable. The overall implementation flow of the proposed algorithm is given in Fig. 10.5.

In the proposed approach, four terrain features retrieved from feature extraction and GLCM metrics have been employed as the input to the ANN model. The ANN model applied is illustrated in Fig. 10.6. Here, three layers of ANN model with four input nodes, six hidden nodes, and one output node have been considered for classification. The input nodes receive terrain features (energy, entropy, correlation, and contrast) as the input continuously from the terrain image feature extraction model. Since we intend to classify the terrain as traversable or non-traversable, therefore, only one output node has been used. Considering the applied ANN configuration $(4 \times 6 \times 1)$, the total of 30 weights is required $(\text{input node} + \text{Output Node}) * \text{hidden node} = 30)$. The proposed AGA algorithm intends to estimate the optimal weight parameter so as to enable most efficient and accurate feature classification and traversability prediction. In general, ANN mode is defined by the function $X^{'} = f(A, B)$, where $X^{'}$ represents output with A represents the weight vector, while B signifies the input vector. During learning, A is updated continuously so as to minimize the error called root-mean-square error (RMSE), which is calculated as

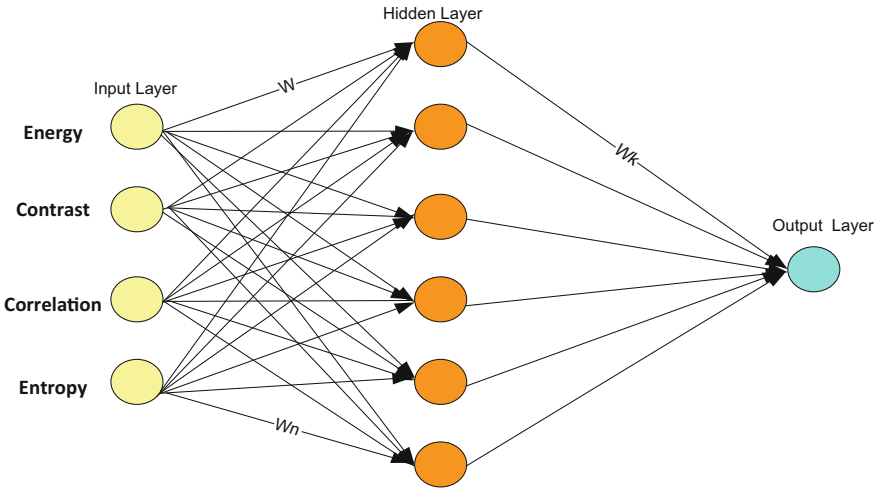


Fig. 10.6 ANN model for terrain classification

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{'} - x_i\right)^2, \tag{5}$$

where x represents the real outcome and $x_i^{'}$ signifies the targeted or expected results. To perform classification, ANN intends to reduce MSE till the targeted results are achieved.

In our proposed ANN architecture, we have employed linear activation function at the input layer that gives output same as input provided (i.e., $O_o = I_i$). Similarly, at hidden layer, a sigmoid function has been applied that gives output O_h with input I_h as $O_h = \frac{1}{1+e^{-I_h}}$. The output layer gives final output as $O_o = \frac{1}{1+e^{-O_h}}$.

Data normalization

To make computation efficient, we have applied min-max normalization approach on the terrain features that significantly enhance the proposed terrain assessment and classification. The min-max normalization scheme normalizes input feature data over the range of [0, 1]. The applied approach of data normalization exhibits linear transformation over the input data elements (original terrain features). It then performs mapping of the data p_i of P to the normalized output $p_i^{''}$ over the range of [0, 1]. The normalized output is obtained as

$$(p_{i\_NOrm}) = p_{i\_Norm}^{'} = \frac{p_i - min(P)}{max(P) - min(P)}, \tag{6}$$

where $max(P)$ and $min(P)$ represent the maximum and minimum values of the data input P, respectively. Thus, applied data normalization scheme significantly avoids premature neuron saturation.

Once performing feature normalization, the normalized data is fed as the input to the ANN mode that applies multivariate regression approach to classify the terrain as traversable or non-traversable. To perform classification, the optimal weight estimation is must. With this objective, in this paper, AGA-based weight estimation has been performed.

1. ***AGA-Based ANN Weight Estimation***

For a conventional ANN algorithm, weight estimation is of great significance to perform efficient learning. Considering existing limitations in GD, GN, and LM learning (weight estimation) algorithm, in this paper we have applied an enhanced GA scheme named adaptive genetic algorithm (AGA) that strengthens ANN model by providing optimal weight values for terrain classification.

Genetic algorithm (GA) has been applauded across scientific horizon for its effectiveness toward optimization and sub-optimal solution retrieval. It functions on the basis of Darwin's principle of natural selection which states that only the candidate (population) having higher fitness value can survive for next generation. At first, GA performs initial population generation, where each candidate represents a solution. These solutions represent a chromosome having binary strings of the associated parameters which are meant to be encoded. The evolutionary algorithm, GA, estimates the fitness value for each chromosome or population candidate. Typically, fitness value signifies certain objective-oriented function that assesses each chromosome whether they are capable of being processed for the next generation or not. Higher fitness value signifies higher survivability for next generation and sub-optimal solution. Based on respective fitness values, the genetic operators, crossover probability (P_c) and mutation probability (P_m), are applied, which decides the selection of the chromosomes for the next generation. With P_c and P_m, the generation continues till the optimal or sub-optimal solutions are obtained. In conventional GA, the process continues for a predefined number of generations (stopping criteria) that significantly introduce computational overheads and time consumption. This is one of the vital limitations in conventional genetic algorithms. To alleviate such issues, we have applied adaptive genetic algorithm (AGA) that updates P_c and P_m adaptively and this process continues till 95% of chromosomes have unique fitness value. It enables reduced computational complexities (due to reduced generation counts) and minimal time consumption. Path planning over unknown terrain being a critical and transient decision process demands swift and transient classification or prediction and hence the proposed AGA-ANN can be of great significance. Here, AGA functions as a supplementary model to ANN and calculates the optimal weight values for efficient learning.

As presented in Fig. 10.5, the proposed ANN model contains single hidden layer configuration with i input, h hidden and O output nodes. Since, in the proposed terrain classification or traversability assessment model, there are four features to be

used for classification and hence here only four input nodes have been defined (Fig. 10.5). The four features, energy, contrast, correlation, and homogeneity or entropy, have been assigned as input to these individual nodes (Fig. 10.6). To avoid huge computational complexities, we have applied only six hidden nodes and one output node that gives eventual classified outcome, whether the respective sub-terrain region of terrain segment is traversable or non-traversable. The number of weights required for ANN learning can be estimated by

$$N = (i + O)^{*}h \tag{7}$$

Now, for $(4 \times 6 \times 1)$ ANN configuration, there are 30 weight values required for learning. Here, each weight has been applied as a gene in the chromosome, which is a real positive value. Let, l be the length of gene; then, the chromosome length L_{Chrom} can be estimated by

$$C_{Length} = N^{*}l = (i + O)^{*}h^{*}l \tag{8}$$

In AGA-ANN based terrain classification model, all the chromosomes have been considered as the GA population and the fitness value, and associated weight for each chromosome has been estimated. Weight A_k has been calculated as

$$A_k = \begin{cases} if\, 0 \le A_{kl+1} < 5 \\ -\frac{A_{kl+2}{}^{*}10^{l-2} + A_{kl+3}{}^{*}10^{l-3} + \cdots + A_{(k+1)l}}{10^{l-2}} \\ if\, 5 <= A_{kl+l} <= 9 \\ +\frac{A_{kl+2}{}^{*}10^{l-2} + A_{kl+3}{}^{*}10^{l-3} + \cdots + A_{(k+1)l}}{10^{l-2}} \end{cases} \tag{9}$$

In order to estimate the weight A_k, at first, the fitness value for each chromosome is required to be estimated. The following figure represents the proposed fitness estimation algorithm.

To alleviate the issues of computational complexity and time consumption, our proposed AGA approach updated the GA parameters P_c and P_m dynamically. In our proposed system, the values of P_c and P_m are updated using the following equations:

$$\begin{aligned} (P_c)_{k+1} &= (P_c)_k - \frac{C_1{}^{*}n}{7} \\ (P_m)_{k+1} &= (P_m)_k - \frac{C_2{}^{*}n}{7} \end{aligned} \tag{10}$$

In above equations, $(P_c)_k$ and $(P_m)_k$ represent the current probability of cross-over and mutation, and $C_1 = 0.01$ and $C_2 = 0.001$ are the (can be any) positive constant. Here, n states the number of chromosome having same fitness value. The proposed AGA process continues till 95% of chromosomes attain same fitness value. After that, the system starts getting saturated. Once the stopping criteria have been attained, the respective optimal weights are assigned to the ANN model,

Algorithm for Fitness Estimation

Input: $\bar{X}_i = (B_{1i}, B_{2i}, B_{3i}, \cdots, B_{li})$

Output: $\bar{X}_i = (X_{1i}, X_{2i}, X_{3i}, \cdots, X_{li})$

where $\bar{B}_i$ represents the input vector, and $\bar{X}_i$ signifies output pairs of the applied ANN configuration.

Step-1: Estimate the weight for each edge of ANN, $\bar{A}_i$ for chromosomes C_i using equation (9).

Step-2: Considering $\bar{A}_i$ as a retrieved weight factor, initiate ANN learning with N input instances and estimate ANN output O_o

Step-3: Estimate RMSE E_j for each input j , $Error_j = \left(Target_{ji} - Observed_{ji}\right)$

Step-4: Estimate RMSE as

$$Error_i = \sqrt{\frac{\sum_{j=1}^{j=N} E_j}{N}}$$

where N is the number of training data

Step-5: Estimate the fitness value for each chromosome

$$Fitness_i = \frac{1}{Error_i} = \frac{1}{\sqrt{\frac{\sum_{j=1}^{j=N} Error_j}{N}}}$$

Step-6 Execute AGA for further process, till stopping criteria is reached.

Fig. 10.7 Pseudocode for candidate fitness estimation

which then starts classification of terrain features using multivariate regression analysis. With the correlation coefficient of value higher than 0.5, the proposed system affirms sub-terrain region or terrain segment as traversable. Thus, performing overall classification, the final output is obtained at the output layer in terms of the terrain segment as traversable or non-traversable and thus the confusion matrix is obtained for performance analysis. The overall proposed AGA algorithm for ANN learning and terrain classification scheme is presented in Fig. 10.7.

The results obtained for the proposed traversability analysis scheme are presented in the following section.

Results and Analysis

In this paper, a novel evolutionary computing algorithm, AGA-based ANN scheme, has been developed for terrain classification. To assess the effectiveness of the proposed system, in this paper, NASA's Mars Exploration Rover mission images

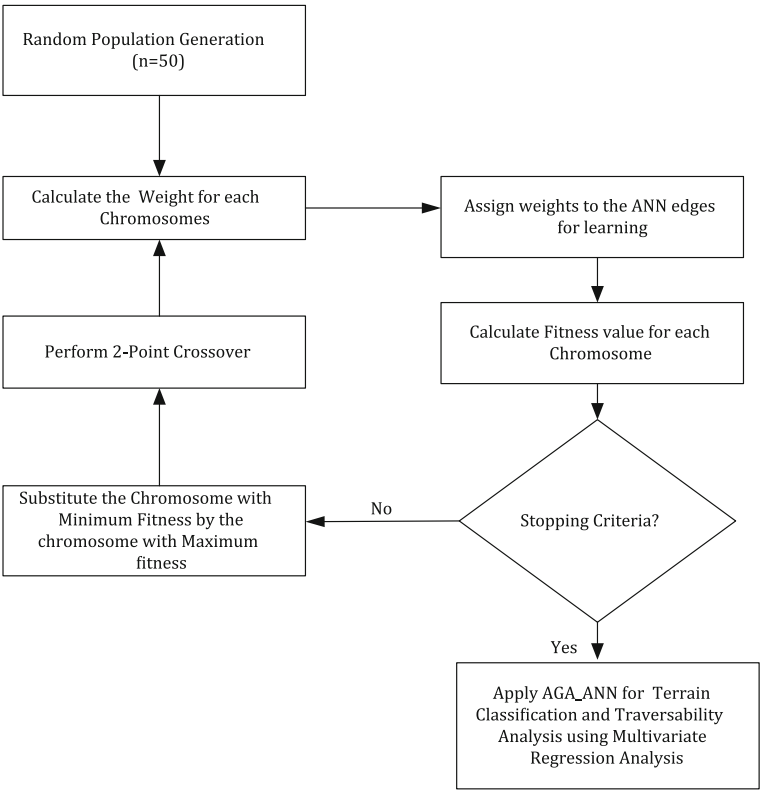


Fig. 10.8 AGA algorithm for ANN learning and terrain classification

(Mars surface scenes, Fig. 10.8a) have been used. These images have been obtained from high-resolution panoramic camera assembled onto rovers. The selected image data has been divided into finite sub-frames, where individual frame signifies a small terrain region. Here, the overall image has been divided into 400 sub-terrain regions. The size of the sub-terrain regions was selected as 320×320. To perform terrain sampling, sub-window frames have been selected with the size of 32×32. Here, we assume that the Mars surface scene represents the terrain with uncertain traversability conditions. We considered rocky, sandy terrain, and loose drift materials as the traversability characteristics. Implementing the 2D discrete wavelet packet decomposition scheme, the features were extracted, which was then followed by GLCM metrics estimation. Using GLCM statistics outcome, we estimated the four features: energy, contrast, correlation, and entropy or homogeneity. These features were fed as the input to the terrain classifier models. In addition to the proposed AGA-ANN based terrain classification and traversability analysis algorithm, K-means algorithm was also applied to classify terrain. The overall

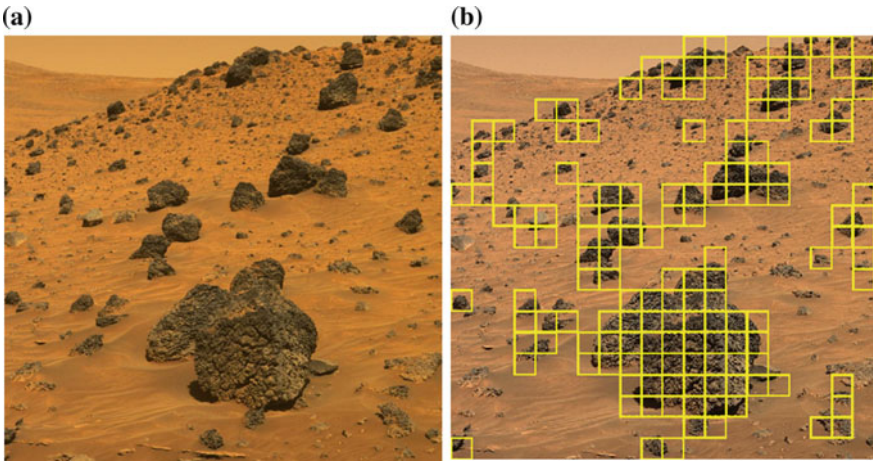


Fig. 10.9 **a** NASA's Mars exploration rover mission image (Mars surface image), **b** sampled and classified terrain region

algorithms have been developed using MATLAB2015b software tool. Figure 10.9b represents the sampled and classified regions over terrain surface image.

In order to perform performance evaluation, the respective results have been obtained using confusion matrix, where the values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) have been retrieved. Based on these variables, the performance analysis has been made in terms of terrain classification (traversability analysis) accuracy, precision, recall, F-measure, and specifications. The applied performance parameters and their mathematical expression are presented in Table 10.1.

As discussed, we have divided the overall terrain regions into 400 sub-terrain regions, which have been further classified as traversable or non-traversable. To examine performance, the confusion matrix for both classification models has been obtained.

Table 10.1 Performance parameters

Parameter	Mathematical expression
Accuracy	$\frac{(TN+TP)}{(TN+FN+FP+TP)}$
Precision	$\frac{TP}{(TP+FP)}$
F-measure	$2.\frac{Recall.Precision}{Recall+Precision}$
Recall	$\frac{TP}{(TP+FN)}$
Specification	$\frac{TN}{(TN+FP)}$

Table 10.2 Confusion matrix after classification by K-means classifier

	Traversable	Non-traversable
Traversable	208	45
Non-traversable	26	121

Table 10.3 Performance of K-means clustering-based terrain classification

Parameters	Output (%)
Accuracy	82.25
Precision	82.21
F-measure	85.42
Recall	88.90
Specification	72.90

A. K-Means Clustering-based Terrain classification

At first, we have applied the K-means clustering scheme for terrain classification by feeding four features as the input data. The confusion matrix obtained after classification is presented in Table 10.2.

Table 10.3 represents the results for the derived classification performance parameters.

B. AGA-ANN based Terrain classification

The confusion matrix obtained for the proposed adaptive genetic algorithm (AGA)-based ANN classifier is presented in Table 10.4. The proposed AGA scheme not only eliminates the existing conventional ANN issues like local minima and convergence but also optimizes generic GA by enhancing optimal minima for swift processing (fast execution) and higher accuracy (Table 10.5).

Table 10.4 Confusion matrix after classification by AGA-ANN classifier

	Traversable	Non-traversable
Traversable	241	12
Non-traversable	30	117

Table 10.5 Performance of AGA-ANN clustering-based terrain classification

Parameters	Output (%)
Accuracy	95.26
Precision	90.50
F-measure	92.00
Recall	88.93
Specification	90.70

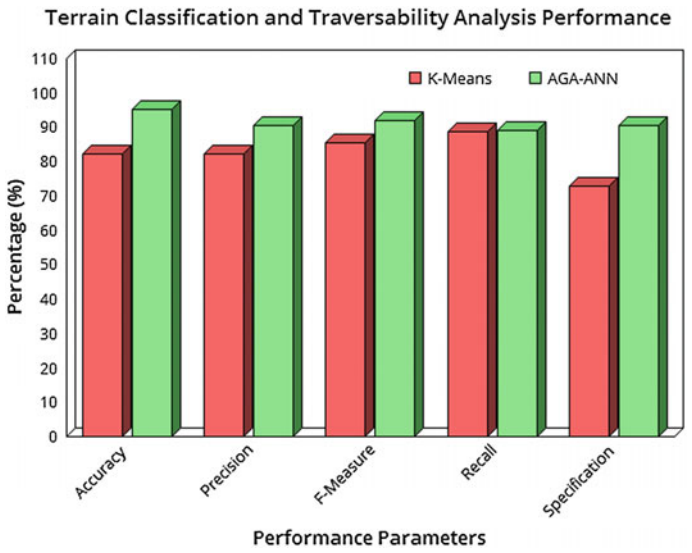


Fig. 10.10 Comparative terrain classification and traversability analysis

Figure 10.9 represents the comparative performance of the existing K-means algorithm based terrain classification and our proposed AGA-ANN based classifier.

Considering overall results (Fig. 10.10), it can be found that our proposed evolutionary computing algorithm AGA-based ANN outperforms existing K-means clustering-based classification for terrain classification and traversability analysis. On the basis of traversability analysis outcome, the path planning can be done to enable safe and efficient navigation under uncertain environmental conditions.

Conclusion

In this paper, an enhanced evolutionary computing-based terrain classification and traversability analysis model has been developed. The proposed traversability analysis process encompasses two phases, where the first phase deals with terrain feature extraction and gray-level co-occurrence metrics (GLCM) estimation, while the second exhibits terrain classification. To perform feature extraction, an enhanced 2D discrete wavelet packet decomposition (2D-DWPD) algorithm has been developed that extracts terrain features, which was further used to compute GLCM. Four terrain features including energy, contrast, entropy or homogeneity, and correlation have been retrieved from GLCM, which were further used to classify a sub-terrain region as traversable or non-traversable. To perform terrain classification, adaptive genetic algorithm (AGA)-based ANN (AGA-ANN) classifier was developed that enhances weight estimation and learning to enable optimal

classification. To evaluate the performance, an additional K-means clustering-based terrain classifier was developed. AGA-ANN algorithm has exhibited the accuracy of 95.26%, precision of 90.5%, recall of 92.0%, F-measure of 88.93%, and specification of 90.70%. The performance evaluation with Mars surface image (retrieved from NASA Mars Mission database) has affirmed that the proposed AGA-ANN based terrain classification model can be a potential technique for terrain's traversability analysis under uncertain conditions. In future, the effectiveness of the proposed scheme can be assessed with real-time data retrieved though camera mounted onto mobile robot roaming across uncertain terrain surface conditions. Based on our traversability results, certain path planning algorithms, such as floodfill algorithm, can be explored for safe autonomous navigation system.

References

1. Howard A., Tunstel E., Edwards D., "An Intelligent Terrain Based Navigation System for Planetary Rovers" in the Joint 9th IFSA World Congress and 20th NAFIPS Int. Conf., pp. 7–12, Vancouver, B.C., Canada, July 2001.
2. Olson C. F., Matthies L. H., Wright J. R., "Visual Terrain Mapping for Mars Exploration," IEEE Aerospace Conference, paper#1176, 2003.
3. Iagnemma K., A Brooks, "Self-Supervised Classification for Planetary Rover Terrain Sensing" IEEE Aerospace Conference, March 2007.
4. Vandapel N., Huber, Kapuria D.F., and Hebert, M. "Natural Terrain Classification using 3-D Ladar Data". Proceedings of the 2003 IEEE International Conference on Robotics and Automation. Pp. 5117–5122.
5. Manduchi R.,. Castano, A., Talukder, A., and Matthies, L "Obstacle Detection and Terrain Classification for Autonomous Off-Road Navigation". Autonomous Robots Vol. 18. Pp. 81–102.
6. Shirkhodaie A., Amrani R., and Chawla N., "Traversable Terrain Modeling and Performance Measurement of Mobile Robots," Performance Metrics for Intelligent Systems Workshop, NIST, August 24–26, 2004.
7. Wolf D F., Dieter Fox and Burgard W., "Autonomous Terrain Classification Using Hidden Markov Models" Proceedings of the 2005 IEEE International Conference on Robotics and Automation [8]. A. Kelly, A., et al. (2006, June). "Toward Reliable Off Road Autonomous Vehicles Operating in Challenging Environments," The International Journal of Robotics Research. 25(5/6).
8. Dima C.S, Vandapel, N., and Hebert, M. (2004). "Classifier Fusion for outdoor obstacle detection," Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)1, 665–671, https://doi.org/10.1109/ROBOT.2004.1307225.
9. Birk A., Stoyanov Nevatia T, Y., "Terrain Classification for Autonomous Robot Mobility: from Safety, Security Rescue Robotics to Planetory Explorations", Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2008.
10. Soennekar J, "Machine Learning and Terrain Classification from LADAR Data".
11. Poppinga J., A Birk and K. Pathak, "Hough Based Terrain Classification for real-time detection of derivable ground," Journal of field Robotics, 2007.
12. Amrani R., "Visual Terrain Traversability Analysis and Assessment Based on Soft Computing Techniques," Master Thesis, Dept. of Mechanical and Manufacturing Engr., Tennessee State University, August 2004.

13. Edmond. M. DuPont, Moore C.A and Jr. Eric Coyle, "Frequency Response method for Terrain Classification in Autonomous ground vehicle", Journal of Field robotics ACM Potal, 2008.
14. Mars Analyst's Notebook 2006 from http://anserverl.eprsl.wustl.edu/.
15. Halatci I, Christopher a. Brooks, "Terrain Classification and Classifier Fusion for Planetory Exploration Rovers," IEEE Aerospace Conference (2007).
16. Mathur P. and Pandian K.K.S, "Traversability Assessment of Terrain for Autonomous Robot Navigation", Proc. of International Multi-Conference of Engineers and Computer Scientists—IMECS, Hong Kong, Vol II, pp 1286–1289, March 2010.
17. Mathur P. and Pandian K., "Terrain Classification for Traversability Analysis for Autonomous Robot avigation in Unknown Natural Terrain", International Journal of Engineering Science and Technology (IJEST), Vol. 4 No. 01, 38–49, January 2012.

Chapter 11
Automatic Brain Tumor Detection and Classification of Grades of Astrocytoma

Nilakshi Devi and Kaustubh Bhattacharyya

Abstract Brain tumor is a deadly disease that medical science has ever seen over the years. Among several brain tumors, astrocytoma is a brain tumor that arises from the astrocytes cells present within the brain. With the event of modern medical imaging modalities, the presence of any abnormality in the human brain has completely been achieved. Among these modalities, MRI holds a special position in detection of brain tumor owing to its many advantages. But, it has been observed that manual detection of tumor, which is the current scenario in medical science, is a delaying process. The manual detection delays the further treatment of the patient which acts as a risk to the patient's life. Also the medical procedure of biopsy, which involves insertion of medical instrument in the human brain, performed to know about the status of the tumor, also leaves its post-surgerical effects on the patient. Thus, to overcome these limitations, automation of tumor detection and creation of noninvasive technology to identify the status of the tumor has become the very need of the hour. Taking a small step to overcome these limitations, we had proposed for a system which uses artificial neural network for automation of brain tumor detection and radial basis function neural network for classification the grades of astrocytoma noninvasively. The results obtained of the grades were also being clinically correlated with the biopsy reports.

Keywords Astrocytoma · Brain tumor · MRI · ANN · RBFN

N. Devi (✉) · K. Bhattacharyya
School of Technology, ADBU, Electronics and Communication, Guwahati, India
e-mail: nilakshidevi4@gmail.com

K. Bhattacharyya
e-mail: kaustubh.bhattacharyya@dbuniversity.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_11

Introduction

Among different types of tumor, brain tumor is a very deadly disease that has been responsible for many deaths over the years. The exact cause of brain tumor is yet to be known to medical science. Owing to the seriousness of the disease and its capability of causing death in no time, proper detection of this disease is very important to save the patient's life. There are several types of brain tumors like astrocytoma, glioma, ependymoma, etc., which occur in different location of the human brain in different shapes and sizes. Among these, astrocytoma is a brain tumor that can occur to any person at any age. In recent medical imaging, magnetic resonance imaging (MRI) is gaining importance because of its ability to generate 3-D views of images and capability to produce high-resolution images [1]. The main challenge lies in separating the tumor part from the normal brain area as well as its surrounding areas like edema, necrosis, etc., after the MRI process is completed. Further, proper detection is important because the future treatment depends on the result of it. Thus, researchers from all corners of the world have proposed many algorithms for detection of brain tumor using many methods. Use of soft computing tools has emerged as an important area of research in tumor detection owing to its many advantages. These intelligent tools have the capability to analyze a problem and provide solutions comparable to a human mind.

Literature Review

A review of literature reveals the fact that considering the numerous disadvantages of manual detection of brain tumor, the researchers have now turned their attention toward using soft computing tools in the detection process to make the detection more robust and less time consuming. Too much of human interaction in the manual detection leads to chaotic results in the detection process which becomes a risk to the patient's life. Intelligent tools like ANN, fuzzy, genetic algorithm (GA), etc., in detection of brain tumor have helped in providing efficient results. Artificial neural networks have been used extensively in classification of normal and tumor images [2]. The advantage of ANN is the requirement of less time consumption in detection of a large number of images. Gray-level co-occurrence matrix (GLCM) has been used extensively in process of feature extraction by many researchers [3]. Fuzzy cognitive maps (FCM) are also being exploited by the researchers as state-of-the-art method in detection of brain tumor. FCM is a soft computing technique that follows a reasoning approach similar to the human reasoning and decision making [4]. The use of fuzzy has emerged as a promising tool in tumor detection. Hybrid techniques in tumor detection have also been evolved as a cutting edge research technology in brain tumor detection using algorithms like GA and particle swarm optimization (PSO) [5].

Brain Tumor and MRI Image Processing

The brain is the most important organ present in the central nervous system of the human body which controls all the functions of the body.

Brain Tumor

Brain tumor is the excessive growth of unwanted cells within the brain randomly in any direction. According to World Health Organization (WHO) brain tumor is of mainly two types: Benign and malignant of which the former is considered to be less aggressive than the latter.

Astrocytoma

Astrocytoma is a brain tumor arising from the "Astrocytes" cells of the human brain. The World Health Organization has classified Astrocytoma into four grades ranging from I to IV. Medically the four grades I, II, III, and IV are known as "Pilocytic Astrocytoma", "Low Grade Astrocytoma", "Anaplastic Astrocytoma", and "Astrocytoma Grade IV", respectively [6].

Magnetic Resonance Imaging (MRI)

The current imaging modality of detection of brain tumor is MRI. MRI uses a high magnetic field to obtain images of the human body. The human body contains randomly arranged protons that get aligned in the direction of the magnetic field when applied under it. A radiofrequency pulse of about 60–120 MHz is pulsed through the patient, which makes the protons to move out of their equilibrium state. In this process, they emit a signal which is captured by the RF receiver. Again when the RF frequency is stopped, the protons go back to their relaxed state and again they emit a signal in this process. Thus, signals produced from millions of protons are combined together to produce a detail image of the internal brain.

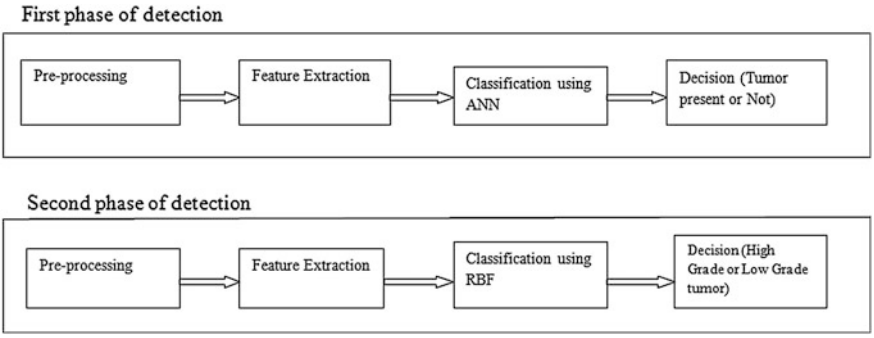


Fig. 11.1 Methodology of the system

Methodology

Our proposed system consists of two phases: detection of astrocytoma tumor using ANN and classification of grades of astrocytoma using RBFN as shown in Fig. 11.1.

Using gray-level co-occurrence matrix (GLCM), two feature values "Energy" and "Homogeneity" were extracted. The feature values were then divided into two sets, the training set and the testing set. The values of the training were then used to train the neural network and the values of the testing set were used to check the validation of our proposed system. The second phase is the detection of the grades of astrocytoma using radial basis function neural network. In this phase, MRI images of the four grades of astrocytoma of different cancer patients were considered and feature extraction was performed extracting three features namely energy, homogeneity, and contrast, followed by optimization using particle swarm optimization to obtain the best variable of each grade. These best variables were then used to train the radial basis neural network to classify the grades of the astrocytoma. The grades classified were then finally correlated with the biopsy reports collected from the hospital.

Preprocessing of the MRI Image

Preprocessing is the process performed to prepare the image for further processing steps like feature extraction and classification. The following steps are applied in preprocessing of the images.

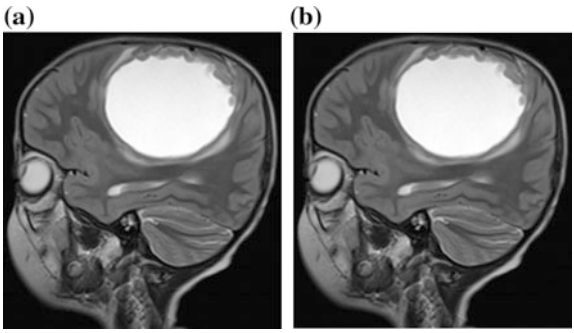


Fig. 11.2 An RGB image converted to a grayscale image

Conversion of the Image to Gray Scale

The input image is in RGB format as shown in Fig. 11.2a, which is converted to grayscale as shown in Fig. 11.2b, before further processing.

Threshold Segmentation

After applying threshold segmentation in the input image Fig. 11.3a, the white portion appeared depicts the tumor part in Fig. 11.3b.

Histogram Equalization

Histogram equalization is a process performed to enhance the image. The enhanced image is as shown in Fig. 11.4b, which was obtained after histogram equalization of the input image as shown in Fig. 11.4a.

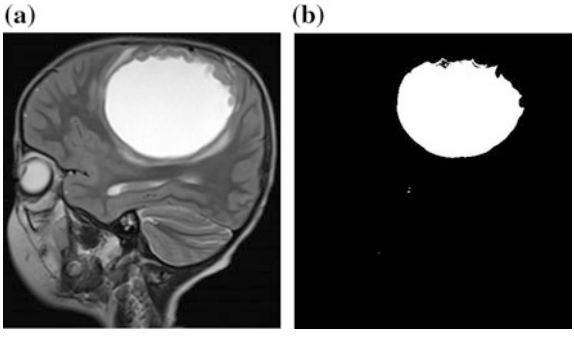


Fig. 11.3 Thresholding operation on the input image

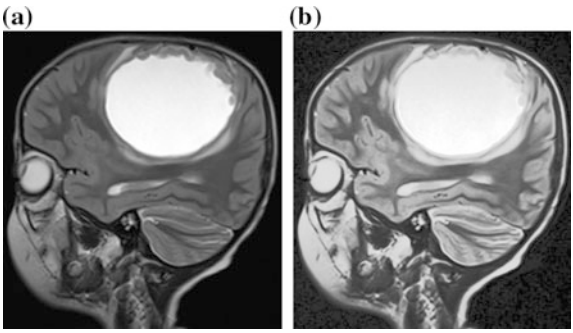


Fig. 11.4 Histogram equalization of the input image

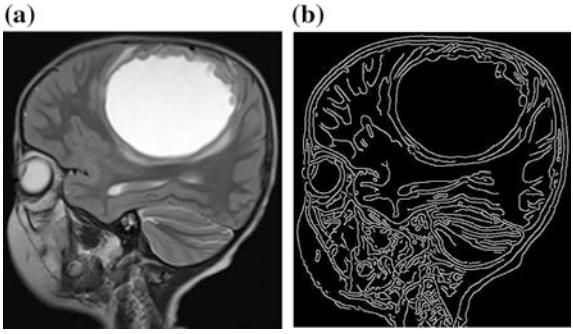


Fig. 11.5 Edge detection of the input image

Edge Detection

Canny edge detection has been used here to find the edge of the tumor. The edges of the tumor can be identified more clearly after applying canny edge detection method as shown in Fig. 11.5b, with input image shown in Fig. 11.5a.

Feature Extraction

Feature extraction is the process by which relevant features are being extracted from the image and the extracted features are then used to train the neural network. In our proposed method, we have used gray-level co-occurrence matrix (GLCM) to extract two features "Energy" and "Homogeneity" for performing the first level of detection for confirmation of presence of tumor. A total of 24 MRI images of three normal patients and 25 images of four tumor patients were collected and feature extraction was performed. Some of the values obtained of normal and tumor images are shown in Tables 11.1 and 11.2, respectively.

Table 11.1 Feature values of images without tumor

No. of images	Energy	Homogeneity
Image1	0.1802	0.9095
Image2	0.1713	0.9168
Image3	0.1825	0.9172
Image4	0.1850	0.9295

Table 11.2 Feature values of images with tumor

No. of image	Energy	Homogeneity
Image1	0.2229	0.9472
Image2	0.2708	0.9495
Image3	0.2521	0.9421
Image4	0.2544	0.9523

ANN-Based Classification of Normal and Tumor Brain Images

After the feature values of both the types of images were obtained, the classification was performed using artificial neural network. ANN is the mathematical network of the human brain. The concept of multilayer perception (MLP) is being also exploited here in configuring the neural network. Levenberg–Marquardt back-propagation algorithm was used in training of the MLP network due to its fastest computational complexity.

Configuration of ANN

The ANN has been configured using 2 nodes in input and 1 node in output layer. The other parameters have been clearly defined in Table 11.3.

Table 11.3 Configuartion of ANN

Network parameters	Values
Number of nodes in input layers	2
Number of nodes in output layers	1
Number of nodes in hidden layers	5
Activation functions	tansig, purelin
Maximum number of iterations	1000
Learning rate	0.05
Learning Algorithm	Levenberg–Marquardt
Mean Square Error	10^{-8}

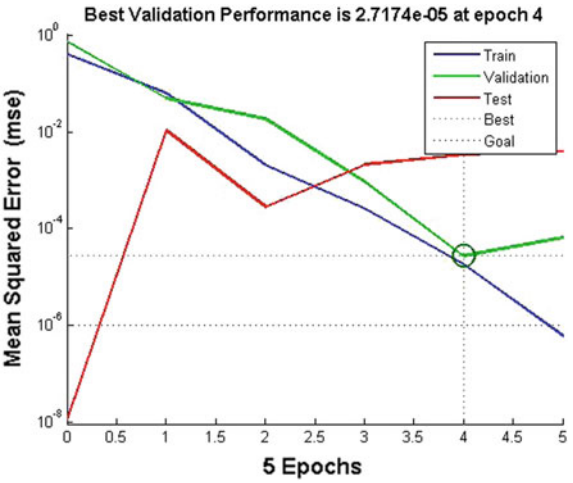


Fig. 11.6 Performance graph of ANN

Results of ANN

From the performance graph, as shown in Fig. 11.6, it was observed that with a total number of 5 iterations and a minimum square error of 10^{-8}, the best validation performance is 2.7174 at 4th iteration with a minimum square error of 10^{-4} with a total elapsed time of 0.504 s. And it was observed that our proposed model of neural network was able to classify the images into normal brain image and tumor brain image with an accuracy of 99%.

Validation of ANN

The validation was carried out using 11 MRI images of which a few are shown in Table 11.4.

Table 11.4 Validation of ANN shown using only a few images

Image	Energy	Homogeneity	ANN output	Remarks
Image 1	0.1989	0.9258	0	Normal brain
Image 2	0.2304	0.9390	1	Tumor brain
Image 3	0.2001	0.9194	0	Normal brain
Image 4	0.2603	0.9582	1	Tumor brain

Table 11.5 Typical values obtained for each grade

Type of images	Energy	Homogeneity	Contrast
Grade I	0.1563	0.9196	0.1683
Grade II	0.1978	0.9217	0.2440
Grade III	0.2219	0.9465	0.1072
Grade IV	0.2603	0.9582	0.2156

RBFN Based Classification of Astrocytoma

After the detection of tumor was successfully completed using ANN, the next step was classification of grades of astrocytoma using RBFN. GLCM was used to extract three feature values "Energy", "Homogeneity", and "Contrast" from the images of the four grades of astrocytoma. 4 images of Grade I, 19 images of Grade II, 14 images of Grade III, and 15 images of Grade IV of different patients were taken for feature extraction. Table 11.5 shows the typical values obtained for each grade.

Particle Swarm Optimization (PSO)

PSO is a search-based algorithm based on the intelligent movement of swarms. We have used the concept of PSO to find the optimal variable for each grade which was then used to train the RBFN for classification of the four grades of astrocytoma. The values obtained after PSO are shown in Table 11.6.

Classification of Grades Using RBFN

The RBFN was trained by the optimized variables obtained after PSO. The classification of grades using RBFN was performed with a total of 48 tumor images of different grades and finally correlating the results obtained with the biopsy report. A few are shown in Table 11.7.

Table 11.6 Optimized values obtained after PSO

Grades	Energy	Homogeneity	Contrast
Grade I	0.1563	0.9190	0.1654
Grade II	0.1771	0.9314	0.1732
Grade III	0.2163	0.9464	0.1073
Grade IV	0.2537	0.9540	0.2202

Table 11.7 Validation of RBFN

Grades	Energy	Homogeneity	Contrast	OutputRBFN
Grade I	0.1743	0.9300	0.2450	I
Grade II	0.1831	0.9214	0.1845	II
Grade III	0.2229	0.9472	0.1059	III
Grade IV	0.2304	0.9390	0.1619	IV

Conclusion and Future Scope

Proper brain tumor detection is vital to carry out further treatment procedures. The limitations of manual detection delay the process of tumor detection which appears to be a risk in the patient's life. Thus, automation of the entire process using intelligent tools has become the need of the hour. In our proposed system, the artificial neural network was trained to detect the normal and tumor human brain. Also radial basis function neural network was used to classify the grades of the tumor which would overcome the side effects of the medical procedure biopsy. Biopsy is an invasive technology which requires the insertion of medical instruments in the human body leaving behind several post-surgerical effects. In this regard, our proposed system which uses RBFN tool to detect the grades of the tumor would emerge as a noninvasive technology to identify the grades without the need of any instrument insertion. Since, our proposed system is a noninvasive method it cannot take decisions regarding the progressive growth of the tumor cells, hence repeated MRI testing has to be performed. Also, the model identifies the four grades for astrocytoma brain tumor so it might lead to incorrect results if tested for any other type of tumor. The proposed technique can be made more robust for tumor detection using neuro-fuzzy and fuzzy-neuro system. Also, in modern medical imaging, a number of imaging modalities like functional magnetic resonance imaging (fMRI), positron emission tomography (PET) Scan, etc., have been emerged to detect any kind of abnormalities present in the brain. Thus, with MRI, these modalities can also be incorporated following optimization of the best imaging technique for noninvasive detection of astrocytoma.

Acknowledgements The authors would like to thank a team of doctors Dr. Chandana Kalita, Dr. Prasanta Kr. Baruah and Dr. Nirod Medi for their valuable suggestions in the medical related issues of the subject. The authors would also like to express sincere thanks to Dr. Jitendra Parmar, Dept. of Interventional Radiology, Dr. B.L. Kapur Hospital, New Delhi, for his help and guidance throughout the research.

References

1. Shan, S., Willam, S.: MRI Fuzzy Segmentation of Brain Tissue Using Neighborhood Attraction With Neural-Network Optimization. In: IEEE Transactions on Information Technology in Biomedicine, vol. 9. IEEE Press (2005).
2. Safaa, A.: Brain tumor diagnosis systems based on artificial neural networks and segmentation using MRI. In: Informatics and Systems International Conference. IEEE Press, New York.
3. Revett, K., Mohsen, M.: Computer aided diagnosis of human brain tumor through MRI. In: Expert systems with Applications. Elsevier Press (2014).
4. Mohamed, N., Sameh, M.: A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data. In: IEEE Transactions on Medical Imaging, vol. 21. (2002).
5. Nandha, G., Kaman, M.: Diagnose Brain Tumor Through MRI Using Image Processing Clustering Algorithms Such As Fuzzy C Means Along With Intelligent Optimization Techniques. IEEE Press, New York (2010).
6. American Brain Tumor Association, http://www.abta.org/brain-tumor-information/types-of-tumor/astrocytoma.html.

Chapter 12
A Fast Adaptive Fuzzy Unsymmetric Trimmed Mean Filter for Removal of Impulse Noise from Digital Images

S. Vijaya Kumar and C. Nagaraju

Abstract Impulse noise reduction from images is an indispensable research issue in the area of image processing. In this paper, we are proposing a fast adaptive fuzzy filter for restoring the pixels that are damaged through salt and pepper noise or impulse noise. The process of filtering consists of three stages. In first stage, we are concentrating on identifying the window dimension for processing pixel using fuzzy detector. In second stage, noise effected pixel can be identified with help of mathematical 3N rule and in the last stage, we restored noise pixel with unsymmetric trimmed mean value. In restoration process, identification of noise pixel is a complex task. This can be simplified in our proposed filtering method with effective performance of 3N rule. Experiments on standard image and medical image sets were conducted to compare our restoration algorithm with two previous competitors. The results show that our method is superior to existing algorithms considering PSNR and elapsed time. The proposed method also indicates to be strong to high ranges of noise, as excessive as 90% with conserving the key details of the image. And it is useful in many applications.

Keywords Adaptive fuzzy filters · Alpha-trimmed mean · Cloud model
High-density salt and pepper noise

Introduction

Image pixels are subjected to salt and pepper noise due to bit errors in channel transmission, ideal medium between the picture machine and scene imperfections in picture sensors, and fault memory location in hardware [1]. Removing the

S. Vijaya Kumar (✉)
Department of CSE, JNTUH, Hyderabad, India
e-mail: svksr105@gmail.com

C. Nagaraju
Department of CSE, Yogi Vemana University College of Engineering, Proddatur, Andhra Pradesh, India
e-mail: cnrcse@yahoo.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_12

high-density impulse noise by way of keeping the image important details with much less computational time is most vital in all the image analysis and processing applications. In the literature, huge numbers of methods have been recommended to address this problem. Linear systems show low performance than nonlinear systems because these systems are not considered rank ordering of pixels for given masks. The classical median filter (MF) is extensively applied filter for eliminating the impulsive noise [2–4]. The MF substitutes the current pixel with the median of the chosen mask. It is easy to implement. However, when the noise is eliminated the usage of MF filter, some of the key details in the image might also be filtered, additionally the quantity of the noise reduced from the image depends on the dimension of filtering mask. If dimension of the processing window is small, the noise eliminated is less, if the window dimension is bigger, more key information are eliminated from the image and similarly it fails at high level of impulse noise. To triumph over these drawbacks, a significant variety of changes to the MF filters proposed in literature, which consists of the weighted and center weighted median filters [5, 6]. These filters provide extra weight to some image pixel values in the mask (filtering window). They preserve details better than MF filter, but fails at the high-density impulse noise. Switching filters are developed to address the high-density impulse noise [7–17]. Two phase techniques BDND [13] and CM [16] are the two filters prominent in expulsion of high thickness salt and pepper noise with better detail saving capacities, however they require more computational time to process the image. This paper presents the new filtering method based on the fuzzy logic, which mimics human thinking. The proposed method works in three phases. In first stage, it concentrating on identifying the window dimension for processing pixel using fuzzy detector. In second stage, noise effected pixel can be identified with help of mathematical 3N rule and in the last stage, the method restored noise pixel with un-symmetric trimmed mean value.

Proposed Method for Noise Detection and Removal Process

In MF, all pixels are processed and replaced by way of the median value of its neighborhood. Unlike the MF filter, the proposed algorithm identified the window size for detection of corrupted pixel with fuzzy rule, and then identifies the noise pixel by using mathematical 3N rule.

In the Table 12.1, P(1,1), P(1,2), P(1,3), and P(1,4) are already processed pixel values and P(2,3), P(3,1), P(3,2), and P(3,3) indicate pixels yet to be processed. Here, minimum half of the window pixels are used to replace the pixel value if the

Table 12.1 3 × 3 processing window

P(1,1)	P(1,2)	P(1,3)
P(2,1)	P(2,2)	P(2,3)
P(3,l)	P(3,2)	P(3,3)

central pixel is corrupted because they are already processed. In the noise detection scheme, first window selection can be performed based on give fuzzy rule

Rule1: If $M_{min} < M_{med} < M_{max}$ then select the given window for the processing.
Else
Increment the window.

Where M_{min}, M_{med} and M_{max} minimum, median, and maximum values are in the given window. The above rule can be selecting the window such that it does not contain the values like examples give below

Pattern 1

255	255	255
255	255	255
255	255	255

Pattern 2

0	0	0
0	0	0
0	0	0

Pattern 3

0	255	0
255	0	255
0	255	0

The above fuzzy rule can be satisfied only if the processed window contains absolute deviation of salt and pepper noise less than to one. After the window has been identified for pixel processing, the proposed method uses the selected window to compute boundaries b1 and b2 using the mathematical 3n rule as given below.

$$\mathrm{b1} = \min(M_{max}, Ex + 3\ En) \tag{1}$$

$$\mathrm{b2} = \max(M_{min}, Ex - 3\ En), \tag{2}$$

where Ex and En defined as

$$Ex = \frac{\sum_{x_{i+s,j+t} \in w_{i,j}^{2N+1}} x_{i+s,j+t}}{n} \tag{3}$$

$$En = \sqrt{\frac{\pi}{2}} \times \frac{1}{n} \sum_{x_{i+s,j+t} \in W_{i,j}^{2N+1}} |x_{i+s,j+t} - Ex| \tag{4}$$

Then the fuzzy selection rule can be written as follows to notice whether the pixel is corrupted or uncorrupted:

Rule2: if $\mathrm{b1} < x_{i,j} < \mathrm{b2}$ then $x_{i,j}$ is asserted as an uncorrupted pixel and its values is left unchanged in any other case $x_{i,j}$ is a corrupted pixel.

After identification of corrupted pixel, the proposed method uses the identical window (used for detection technique) of neighboring pixels around principal pixel for the filtering process. The simple concept in a trimmed filter is to discard the corrupted intensity values in the chosen window. Alpha trimmed mean filtering [1] is symmetrical filter, the place the trimming is symmetric at both the ends. In this technique, even informative pixels are additionally trimmed off. This leads to loss of key photo details and blurring of the image. To overcome these drawbacks, a fuzzy unsymmetric trimmed mean filter [FUTM] is proposed. FUTM filter takes solely uncorrupted pixels of the window and computes the mean of those pixels and that suggest value is used to replace the corrupted central pixel. This filter is known as trimmed mean filter due to the fact the corrupted pixel values are eliminated from the chosen window. The FUTM filter calculates the mean value using the Eq. 5.

$$Y(i,j) = \frac{\sum_{i=1}^{N} D(i)}{N}, \tag{5}$$

where D is set that comprises the pixels which are now not corrupted around the central pixel, N characterizes the number of the uncorrupted pixels around the central pixel and $Y(i,j)$is processed image.

Results and Discussions

The overall performance of the FUTM method has been assessed qualitatively and quantitatively through experimental analysis. Standard images such as Lena, Mandrill and mammogram images of dimension 256 × 256 have been used to check the performance of the various algorithms. To consider detailed preservation, the edge map of the enhanced Lena image is additionally computed, which has been compared with an edge map of CM filtered image and original images. The execution of the recovery procedure is additionally measured through assessing the peak signal-to-noise ratio (PSNR). The PSNR value in decibels can be expressed mathematically with the use of the following equation.

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right)Db, \tag{6}$$

where

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(B_{(i,j)} - I_{(i,j)}\right) \tag{7}$$

Table 12.2 Comparison of Restored Mandrill Image in PSNR (in decibels)

Image-Mandrill/Noise%	50%	60%	70%	80%	90%
CM	31.7522	31.2915	30.7103	30.0861	29.2608
BDND	27.6358	26.1564	24.8796	22.4563	21.9863
FUTM	31.6467	31.2468	30.8303	30.3525	29.5887

Table 12.3 Comparison of Restored Mammogram Image in PSNR (in decibels)

Image-Mammogram/Noise%	50%	60%	70%	80%	90%
CM	42.519	41.9896	41.259	39.667	37.196
BDND	40.263	38.6359	37.568	36.589	35.896
FUTM	44.033	43.3492	41.673	40.142	38.116

Table 12.4 Elapsed time (in seconds) values for various filters operating on Mandrill image at various noise levels

Image-Mandrill	50%	60%	70%	80%	90%
CM	7.714474	7.926351	8.421032	8.788081	9.498124
BDND	135.7792	133.5172	127.2414	108.3322	113.1052
FUTM	7.086073	7.68062	7.922673	8.485249	9.075118

M Number of rows of the image
N Number of columns of the image
$B(i, j)$ The filtered image and
$I(i,j)$ The original image.

The PSNR values of the proposed method, CM, and BDND filters for different images (“Lena”, “Mandrill”, and “Mammogram”) across 50–90% noise levels are shown in Tables 12.2, 12.3, and 12.4, and are also plotted in Fig. 12.1. As can be seen from Tables 12.2, 12.3, 12.4, 12.5, and Fig. 12.1, the FUTM method significantly outperforms the CM and BDND filters, in PNSR measurement. The results indicate that the FUTM method is robust across wide range of the salt and pepper noise densities. The subjective visual comparison of the noise removal and

Table 12.5 Elapsed time (in seconds) values for various filters operating on Mammogram image at various noise levels

Image-Mammogram	50%	60%	70%	80%	90%
CM	9.815891	9.598926	9.605507	9.803099	10.2467
BDND	172.7654	161.6913	145.1387	114.6809	112.494
FUTM	9.089958	9.404306	9.563266	9.688486	10.0433

Table 12.6 Comparison of selection of processed window for Lena image by CM and proposed methods

Processed window	Lena image	50%	60%	70%	80%	90%
3 × 3	Proposed	64835	64254	63303	61774	59611
	CM	49903	41710	31461	19696	8375
5 × 5	Proposed	558	1133	1988	3364	5239
	CM	15621	23788	33982	45663	56938
7 × 7	Proposed	16	16	49	80	185
	CM	7	34	81	146	80
9 × 9	Proposed	3	9	7	17	37
	CM	0	0	9	28	106
11 × 11	Proposed	1	3	1	6	12
	CM	0	0	0	1	32
13 × 13	Proposed	1	1	1	3	8
	CM	0	1	1	0	0
15 × 15	Proposed	1	0	1	2	3
	CM	0	0	0	0	0
17 × 17	Proposed	0	0	1	2	3
	CM	0	0	0	0	0
19 × 19	Proposed	121	120	185	288	438
	CM	5	3	2	2	5

perception of the image details for the test images “Lena” and “Mandrill” are presented in Figs. 12.2 and 12.3, respectively. The FUTM filter is clearly superior in terms of retaining critical details of the image compared with CM and BDND filters.

Runtime Analysis

The computational time (in seconds) for the different filters using INTEL(R), CORE(TM)2DUO CPU, 2.93 GHz processor, 2 GB RAM with MATLAB 7.9.0 (R2009b) are shown in the Tables 12.5 through 12.6 for different images across 50–90% noise range, and are also plotted in Fig. 12.4. The elapsed time for the FUFC method is low compared to the CM and BDND. CM filter replaces the corrupted pixel with a weighted average of uncorrupted pixels which require more

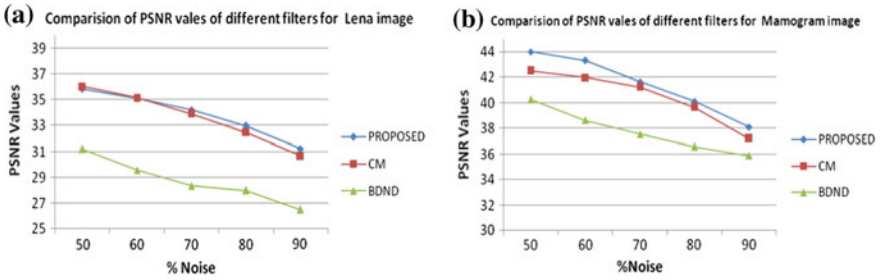


Fig. 12.1 **a** Comparision of PSNR values of different filters for Lena image, **b** Comparision of PSNR values of different filters for Mammogram image

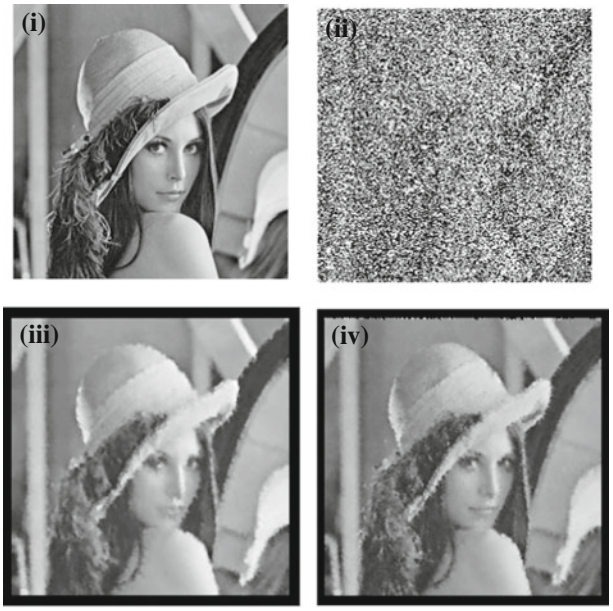


Fig. 12.2 **i** Lena image without noise, **ii** 70% corrupted Lena Image, **iii** CM filtered image and FUTM method filtered image, **iv** Restoration details of Lena image corrupted with 70% Noise

computational time. Proposed filter considers only the pixels which are not corrupted and takes the mean of those pixels without calculating the weights for the uncorrupted pixels like CM filter.

To identifying the window size for processing the central pixel, the proposed method uses the fuzzy rule. The FUTM method uses the 91.05% of the times the 3×3 window for processing of 90% impulse noise corrupted images, 93.92% of the times it uses the 3×3 window for processing the 80% of corrupted images, 96.48% of times it uses 3×3 window for processing the 70% corrupted images,

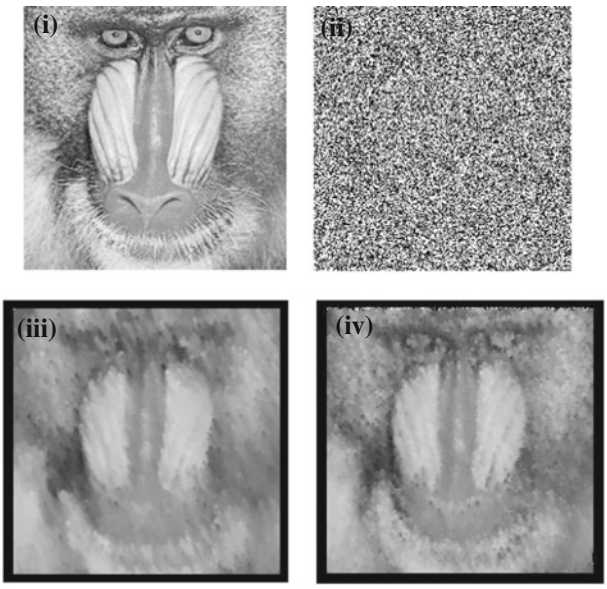


Fig. 12.3 **i** Mandrill image without noise Fig. 12.4, **ii** 90% corrupted Mandrill image, **iii** CM filtered image and **iv** FUTM method filtered image. Restoration details of Mandrill image corrupted with 90% Noise

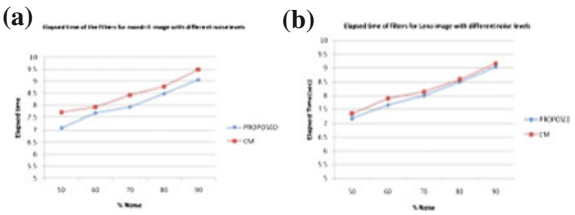


Fig. 12.4 **a**, **b** Elapsed time of the methods for Mandrill image and Lena image with various noise levels

7.83% of the times it uses the 3 × 3 window for processing the 60% corrupted images, and 98.57% of times it uses the 3 × 3 window for processing the 50% corrupted images. Whereas the CM filter uses 12.77%, 30.05%, 48%, 63.64%, and 76.14% of the times the 3 × 3 window for processing 90%, 80%, 70%, 60%, and 50% of corrupted images, respectively. The CM filter uses 23.83%, 36.29%, 51.89%, 69.67%, and 86.88% of times 5 × 5 window as processed window, it shows that CM filter already consider that many number of times 3 × 3 window as processed window, but fail to take a decision so that it incrementing the window size, which causes the increasing of computational time. The details of window selection for processing pixel for the image Lena by CM and FUTM methods are presented in Table 12.5 (Fig. 12.5).

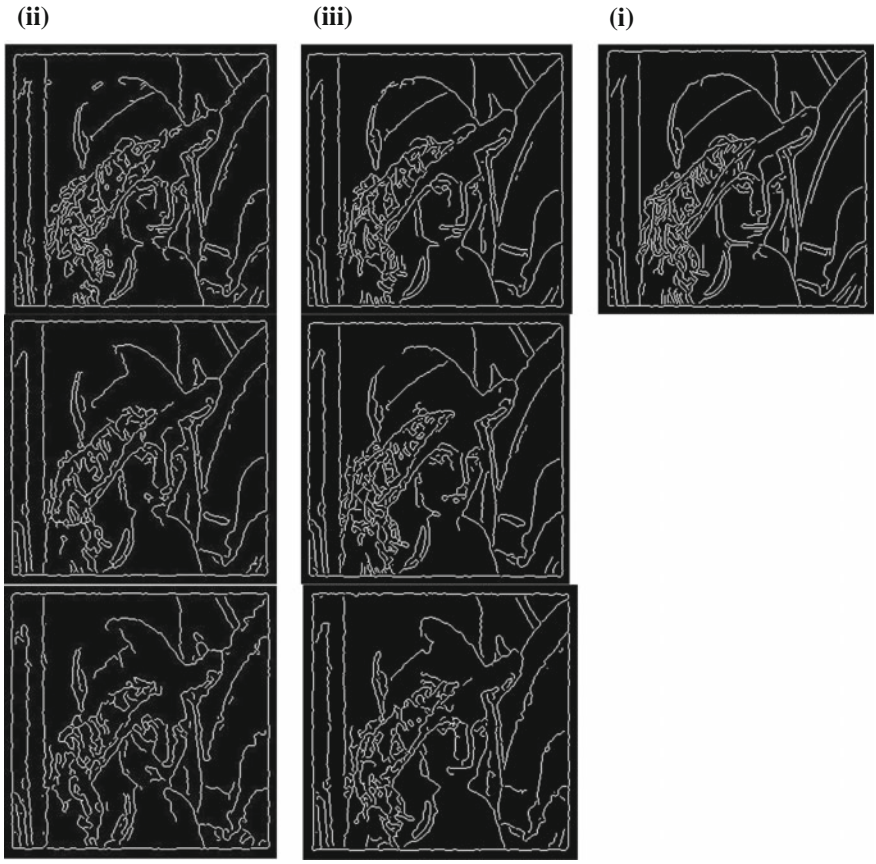


Fig. 12.5 **i** Edge map of image without noise, **ii** Edge map of CM filtered images with 60, 70, and 80% noise and **iii** Edge map of FUTM filtered images with 60, 70, and 80% noise (Top to Bottom)

Conclusions

We are judging the filter performance based on how filter improves the subjective visual quality, preserving the key details of the image and based on the computational time. This paper concludes with all the three factors, respectively, the proposed filter gives the better results compared to the remaining tested filters, but failing in keeping the edge details of the images in some areas, which would be addressed in next paper.

References

1. R. C. Gonzalez and R. E. Woods, "Digital Image Processing", Prentice Hall 2002.
2. W. K. Pratt, "Median filtering," Tech. Rep., Image Proc. Inst., Univ. Southern California, Los Angeles, Sep. 1975.
3. I. Pitas and A. Venetsanopou, Nonlinear Digital Filters: Principles and Application. Norwell, MA: Kluwer, 1990.
4. J. Astola and P. Kuosmanen, Fundamentals of Nonlinear Digital Filtering. Boca Raton, FL: CRC, 1997.
5. D. R. K. Brownrigg, "The weighted median filter," ACM Commun., vol. 27, no. 8, pp. 807–818, Aug. 1984.
6. S. J. Ko and Y. H. Lee, "Center weighted median filters and their applications to image enhancement," IEEE Trans. Circuits Syst., vol. 38, no. 9, pp. 984–993, Sep. 1991.
7. T. Sun and Y. Neuvo, "Detail-preserving median based filters in image Processing," Pattern Recognit. Lett., vol. 15, pp. 341–347, 1994.
8. D. Florencio and R. Schafer, "Decision-based median filter using local signal statistics," in Proc. SPIE Int. Symp. Visual Communications Image Processing, Chicago, Sept. 1994.
9. S. Zhang and M. A. Karim, "A new impulse detector for switching median filters," IEEE Signal Process. Lett., vol. 9, no. 11, pp. 360–363, Nov. 2002.
10. How-Lung Eng and Kai-Kuang Ma, "Noise adaptive soft-switching median filter", IEEE Transactions on Image Processing, Vol. 10, No 2, August 2002.
11. P. E. Ng and K. K. Ma, "A switching median filter with boundary discriminative noise detection for extremely corrupted images", IEEE Transactions on Image Processing, Vol. 15, No. 6, pp. 1506–1516, 2006.
12. K. S. Srinivasan and D. Ebenezer "A new fast and efficient decision-based algorithm for removal of high-density impulse noises", IEEE Signal Process. Lett., vol. 14, no. 3, pp. 189–192 2007.
13. D. Y. Li and Y. Du Artificial Intelligent With Uncertainty, 2007: CRC Press.
14. Zhe Zhou, "Cognition and Removal of Impulse Noise With Uncertainty", IEEE Transactions on image processing, vol. 21, no. 7, pp. 3157–3167, July 2012.
15. Sema Koc Kayhan, "An effective 2-stage method for removing impulse noise in images", J. Vis Commun. Image R, Vol 25, pp. 478–486 2014.
16. P. S. Windyga, "Fast impulsive noise removal," IEEE Trans. Image Process., vol. 10, no. 1, pp. 173–179, Jan. 2001.
17. I. Aizenberg and C. Butakoff, "Effective impulse detectors based on rank-order criteria," IEEE Signal Process. Lett., vol. 11, no. 3, pp. 363–366, Mar. 2004.
18. H. Hwang and R. A. Haddad, "Adaptive Median Filters: New Algorithms and Results", IEEE Transactions on Image Processing, Vol. 4, No. 4, April 1995.
19. Z. Wang and D. Zhang, "Progressive switching median filter for the removal of impulse noise from highly corrupted images," IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process., vol. 46, no. 1, pp. 78–80, Jan. 1999.

Chapter 13
Mining Fuzzy Classification Rules with Exceptions: A Comparative Study

Amarnath Pathak, Dhruv Goel and Somen Debnath

Abstract Adding fuzziness to normal classification rules enables the rules to adapt to the real-life decision-making process. Besides, it also adds to the classification accuracy of the obtained model and the rules look more accurate and reasonable. Further improvement in classification accuracy can be achieved by discovering exceptions corresponding to these fuzzy rules. Fuzzy rules augmented with exceptions (censors) are termed as Fuzzy Censored Classification Rules (FCCRs) and such kind of rules are best at handling uncertainties like vagueness and ambiguity. These rules, being very efficient, have been widely used under exceptional circumstances. In this paper, we have investigated all the algorithms used in past for discovering FCCRs. Based on review of literature, we have also proposed possible modifications to existing algorithms and techniques.

Keywords CPR · Fuzzy rules · FCCRs · Nature-inspired algorithms

Introduction

Data mining is a step in Knowledge Discovery from Databases (KDD) which aims at finding knowledge in the form of interesting patterns from the databases. Pattern discovery varies from one user to another depending on users' requirement. Classification is an important data mining task which extracts knowledge from the database in the form of If-Then Rules. Training data is used to create a classification

A. Pathak
Department of Computer Science and Engineering, National Institute of Technology, Aizawl 796012, Mizoram, India
e-mail: amar4gate@gmail.com

D. Goel · S. Debnath (✉)
Department of Information Technology, Mizoram University, Aizawl 796004, India
e-mail: somen@mzu.edu.in

D. Goel
e-mail: thisisdg@rediffmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_13

model which is later applied to the test data to classify an unseen instance. A good classification model comprises of several If-Then production rules (PRs) which are simple, comprehensible, accurate, and interesting. Interestingness is usually present in the form of exceptional instances which deviate from normal behavior but the traditional methods of classification rule discovery are least concerned with interestingness and leave out such instances during the rule discovery process. Several attempts have been made in past to modify these traditional techniques to incorporate exceptional instances for making the model more interesting and accurate and modified rules are often represented using If-Then-Unless form, better known as Censored Production Rules (CPRs) framework.

However, nature of PRs or CPRs is such that they impose sharp cutoff on values that predicting attributes can take, resulting in a model which is incapable of dealing with uncertainty and vagueness. Imposing sharp boundaries on predicting attributes' values is impractical and unreasonable in real life situations. Consider a classification rule of the form:

$$\text{If } (\text{Humidity} \leq 10) \text{ and } (\text{Temperature} \leq 25) \text{ then } (\text{Play\_outside} = \text{'yes'})$$

Rule says that a child is allowed to play outside provided humidity level is less than 10 and temperature is below 25. However, as per this rule, the child is not allowed to play if humidity is 10.1 and temperature is 25.1 which definitely does not seem to be a justified decision. Augmenting fuzziness to the classification rules may improve real-life decision-making process and can make the rules more effective. A Fuzzy Classification Rule (FCR) takes the form:

$$\text{If } (\text{X is } x_i) \text{ and } (\text{Y is } y_i) \text{ then } (\text{Class} = c_i),$$

where "*X*" and "*Y*" are predicting attributes and "Class" is class attribute and the attribute values of the predicting attributes such as x_i and y_i are fuzzy rather than crisp, meaning that predicting attributes take fuzzy values such as poor, excellent, etc., instead of numerical values. A sample FCR may be given as

$$\text{If (Humidity is moderate) and (Temperature is low) then (Play\_outside = 'yes')}$$

Such a rule is certainly more relevant in context of daily life decision-making process. However, augmenting exceptions with FCRs may further improve the decision-making process and make the rules more interesting. Consider an FCR with exception:

$$\text{If (Humidity is moderate) and (Temperature is low) then Play\_outside} = \text{'yes' unless Rain fall} = \text{'false'}.$$

Rule says that a child is allowed to play outside if humidity is moderate and temperature is low provided there is no rainfall. In presence of exceptional condition, *Rainfall* = "*True*", decision of playing outside changes from "*yes*" to "*no*".

FCRs with exception conditions are termed as Fuzzy Censored Classification Rules (FCCRs) which integrates fuzzy rules with CPRs and possess more interestingness and information than the FCRs.

In this paper, we have tried to investigate all the past works related to mining Fuzzy Censored Classification Rules. We have also proposed possible modifications to existing techniques of mining FCCRs.

Rest of the paper is organized as follows: Section "Fuzzy Censored Classification Rule (FCCR)" discusses different past works related to mining FCCRs. Section "Future Directions" points direction for future research and suggests possible modifications in existing algorithms for mining FCCRs. Section "Conclusion" concludes the paper.

Fuzzy Censored Classification Rule (FCCR)

To the best of our knowledge, first attempt to discover FCCRs dates back to 1992 when Dimiter et al. proposed a way to represent fuzzy rule with unless condition [1]. Authors have argued that fuzzy conditional statements of the form "If X is A then Y is B unless Z is C" can be represented by $R^* = R^+ \cap R^-$ where R^+ is equivalent to $A(x) \wedge \overline{C(z)} \Rightarrow B(y)$ and R^- is equivalent to $A(x) \wedge \overline{C(z)} \Rightarrow B(y)$. Such a rule helps in reasoning in the situations of incomplete information and resource constraints. Decision that Y is B can be concluded in case no information about Z is available. Later, in 1998, fuzzy rule based classification system with reject option was proposed [2]. However, such a system is limited to the situations where cost of misclassification is higher than the cost of rejection. $\alpha_{Classk}(X_p)$ is output value for class k when the new pattern X_p is presented as input vector to the fuzzy classification model and X_p is assigned the class with maximal value of $\alpha_{Classk}(X_p)$. Model rejects the input pattern provided the maximum value of $\alpha_{Classk}(X_p)$ is less than a minimum specified threshold, θ. Difference between largest and second largest value of $\alpha_{Classk}(X_p)$ is also sometimes used as rejection criteria.

Later in 2001, fuzzy exception learning was used to detect noise trading [3]. The proposed algorithm, Competitive Fuzzy Exception Learning Algorithm (CELA), aims at discovering special circumstances, regimes, under which noise trading takes place. Discovering such special circumstances can be achieved by unmasking the fuzzy part (deterministic part). In context of financial market, regimes correspond to exceptional price developments and they are undesirable. An important modification to existing fuzzy exception discovery technique was proposed in 2004 [4]. Authors have proposed a method to identify fuzzy model having maximal rules. Proposed model is multi-objective as it aims at achieving multiple goals, viz., maximum rule, accuracy, and interpretability. Extraction of the rules is followed by finding conflicts caused by them and these conflicts are resolved by including

exceptions to the rules. Further, several strategies like rule reduction, rule merging, and exception merging have been adopted to ensure interpretability of the model.

Rule plus exception format, which is key to exception discovery, has been investigated by the authors in [5] in different contexts. As per authors, there exist two types of exceptions to a rule—the incorrectly covered exceptions and uncovered exceptions. Further, the discovered rules must be simple and number of exceptions discovered should be smallest possible. In 2007, an Ant Colony Optimization plug-in was proposed to enhance the interpretability of fuzzy rule bases with exceptions [6]. In 2008, authors extended their work to build a model which is enhancement of their previous work [7]. In particular, authors have worked on improving the interpretability and computational cost of the previous model. Restriction that defines feasible step of an ant has been relaxed which leads to increase in flexibility of rule base configurations thus providing interpretability improvement. Further, a local search technique has been incorporated to refine each solution provided by an ant. Attempts have been made to simplify the solution provided by ant by applying the concept of subsumption of a rule in rule base. Besides, computational cost of the previous algorithm has been diminished by pruning the construction graph which discards steps with low probability in selection procedure. Pheromone pruning has been used which sets a threshold for the pheromone level and all the edges of construction graph lying below the threshold are considered as infeasible transitions taken by an ant.

In [8], authors proposed an approach for the discovery of quantified rules with exceptions in the form of censored production rules (CPR) from the large set of discovered if-then rules.

A proper framework for discovering FCCRs has been proposed in [9]. Authors have proposed different parameters based on which exceptions can be discovered and also justified the needs and benefits of discovering FCCRs. In particular, the parameters γ_1 and γ_2 are used for exception discovery and are given by

$$\gamma_1 = \frac{P \wedge D}{P}$$

$$\gamma_2 = \frac{P \wedge C}{P}$$

where P is the predicting part of an FCCR, D is the consequent part which holds frequently and C is the censor part which holds rarely. Exceptions to a Fuzzy Classification Rule (FCR) exist subject to conditions: $\gamma_1 + \gamma_2 \leq 1$ && $\gamma_1 \gg \gamma_2$.

Authors have devised a genetic algorithm approach for discovery of FCCRs from the datasets and argued that proposed discovery will enhance the capabilities of automated and expert system and prove its worthiness in fuzzy control applications by predicting the behavior of system in rare circumstances. In [10], Bala et al. have proposed an extension to their work and used genetic algorithm approach for discovering tuned fuzzy classification rules with intra and inter-class exceptions. A three-phased approach has been used for discovery of (FCRs) augmented with

intra- and inter-class exceptions. A preprocessing algorithm is suggested to tune DB in terms of the membership functions and number of linguistic terms for each attribute of a data set in the first phase. The second phase discovers FCRs employing a genetic algorithm approach. Subsequently, intra- and inter-class exceptions are added to the rules in the third phase.

Exceptions, i.e., censors add interestingness to the rule and rule out any possibility of misclassification even under rare circumstances. Intra-class exception refers to the attribute–value pair which, if appended to a decision rule, makes no change in the decision class whereas augmenting inter-class exception to antecedent part alters the class of decision rule [11, 12]. A genetic algorithm based approach to discover intra-class exceptions has been proposed in [11]. It comprises of two stages: In first stage, generalized rules are discovered and in second stage, exceptions are appended to the rules. Michigan approach has been used in rule discovery meaning that the algorithm generates list of best rules and not the best rule list. Further, discovery of intra- and inter-class exceptions using nature-inspired algorithms such as genetic algorithms and ACO has been investigated by the authors in [12]. Need for the exception discovery has been emphasized by the authors in [13, 14]. Suzuki et al. [13] have discovered exceptions in the form of rule pair. Rule triplet structure, which is an extension of rule pair structure, has been discussed in [14]. A brief summary of some of the appraised papers of this section is given in Table 13.1.

Table 13.1 List of appraised papers

Paper	Technique(s) used (if any)	Major finding(s)
[1]		Proposed a different representation for representing fuzzy rules plus exception
[2]		Proposed a fuzzy rule based classification system with reject option but limiting its use to the situations where cost of misclassification is higher than the cost of rejection
[3]	CELA	Use of fuzzy exception learning to detect noise trading
[4]		Multi-objective model to discover exceptions for fuzzy rules
[6]	Ant Colony Optimization (ACO)	Use of nature-inspired algorithms, for the first time, for discovering interpretable fuzzy rule bases with exceptions
[9]	Genetic Algorithm	(i) Fuzzy rules plus exception was given a new name, Fuzzy Censored Classification Rules (FCCRs) where a censor corresponds to exception (ii) Use of parameters γ_1 and γ_2 for discovery of FCCRs
[10]	Genetic Algorithm	(i) Use of GA for discovering tuned fuzzy classification rules with intra- and inter-class exceptions (ii) Use of three-phased approach for discovering FCCRs

Future Directions

From the review of past works, we can identify a number of possible modifications that are worth applying to all the existing algorithms and techniques of mining FCCRs. First could be improvement in interpretability of obtained rules. Interpretability of the model can be improved by modifying our algorithms to generate more number of rules with fewer numbers of exceptions rather than fewer numbers of rules with larger number of exceptions. Excessive number of exceptions usually adds to the complexity of rules making the model less interpretable. Second, nature-inspired algorithms like GA and ACO have been limitedly used in past for mining FCCRs and hence it could be interesting to further investigate these algorithms because these algorithms are best at exploring search space and avoiding convergence to local optima. Third, a common step in discovering fuzzy rules is to fuzzify the predicting attributes using some membership function. In past, same membership function has been used to fuzzify all the predicting attributes irrespective of their nature. Accuracy of the obtained model can be enhanced if different membership functions are used for different predicting attributes.

Conclusion

Several algorithms have been proposed in past to discover fuzzy rules plus exceptions, also called FCCRs. In this paper, we have investigated all such algorithms and suggested possible modifications to existing approaches which, if adopted, may enhance the classification performance.

References

1. Driankov, D., Hellendoorn, H.: Fuzzy logic with unless-rules. In: IEEE International Conference on Fuzzy Systems, pp. 255–262. IEEE (1992)
2. Ishibuchi, H., Nakshima T.: Fuzzy classification with reject options by fuzzy if-then rules. In: IEEE World Congress on Computational Intelligence, pp. 1452–1457. IEEE, Anchorage, AK (1998)
3. Van den Bergh, W.M., Van den Berg, J., Kaymak, U.: Detecting noise trading using fuzzy exception learning. In: IFSA World Congress and 20th NAFIPS International Conference, pp. 946–951. IEEE, Vancouver (2001)
4. Carmona, P., Castro, J.L., Zurita, J.M.: Fuzzy rule identification with exceptions. IEEE Transactions on Fuzzy Systems, 12(1), 140–151 (2004)
5. Yao, Y., Wang, F.Y., Wang, J.: "Rule + exception" strategies for knowledge management and discovery. In: International Workshop on Rough Sets, Fuzzy Sets, Data Mining and Granular-Soft Computing, pp. 69–78. Springer Berlin Heidelberg (2005)
6. Carmona, P., Castro, J.L.: An Ant Colony Optimization plug-into enhance the interpretability of fuzzy rule bases with exceptions. In: Analysis and Design of Intelligent Systems using Soft Computing Techniques, pp. 436–444. Springer Berlin Heidelberg (2007)

7. Carmona, P., Castro, J.L.: Improvements in the identification of interpretable fuzzy models with exceptions based on ant colony optimization. In: 4th IEEE International Conference on Intelligent Systems, pp. 2–39. IEEE, Varna (2008)
8. Siddiqui, T., Alam, M.A.: Discovery of Fuzzy Censored Production Rules from Large Set of Discovered Fuzzy if then Rules. In: Proceedings of World Academy of Science, Engineering and Technology [PWASET], pp. 158–161 (2009)
9. Bala, R., Ratnoo, S.: Discovering Fuzzy Censored Classification Rules (FCCRS): A Genetic Algorithm Approach. International Journal of Artificial Intelligence & Applications, 3(4), 175–188 (2012)
10. Bala, R., Ratnoo, S.: A Genetic Algorithm Approach for Discovering Tuned Fuzzy Classification Rules with Intra-and Inter-Class Exceptions. Journal of Intelligent Systems, 25(2), 263–282 (2016)
11. Vashishtha, J., Kumar, D., Ratnoo, S.: An evolutionary approach to discover intra–and inter–class exceptions in databases. International Journal of Intelligent Systems Technologies and Applications, 12(3–4), 283–300 (2013)
12. Pathak, A., Vashistha, J.: Classification Rule and Exception Mining Using Nature Inspired Algorithms. International Journal of Computer Science and Information Technologies, 6(3), 3023–3030 (2015)
13. Suzuki, E.: Undirected discovery of interesting exception rules. International Journal of Pattern Recognition and Artificial Intelligence, 16(08), 1065–1086 (2002)
14. Suzuki, E., Żytkow, J.M.: Unified algorithm for undirected discovery of exception rules. International journal of intelligent systems, 20(7), 673–691 (2005)

Chapter 14
Impulse Noise Reduction in Digital Images Using Fuzzy Logic and Artificial Neural Network

Khwairakpam Amitab, Kishore Medhi, Debdatta Kandar and Babu Sena Paul

Abstract Impulse noise is the most common types of noise; it degrades the quality of images and must be removed before performing any high level image processing. In this work, we have proposed a hybrid impulse noise filter, it is implemented in two phases, in the first phase, fuzzy rules are used to detect the pixels affected by impulse noise and in the second phase, artificial neural network is used to remove noise from the affected pixel. The proposed filter is comparatively evaluated with some of the popular impulse noise filter based on peak signal-to-noise ratio and edge preservative factor, it was found that the proposed filter reduces impulse noise and simultaneously preserves image details. For highly corrupted images, the proposed filter can be used, recursively.

Keywords Impulse noise · Noise filter · Fuzzy logic · Artificial neural network · Multilayer perceptron · Image processing

Introduction

Impulse noises are caused by error in recording and transmission of images. It is important to reduce impulse noise from images since it deteriorates the quality. In this work, we have considered grayscale images, and each pixel are represented

K. Amitab (✉) · K. Medhi · D. Kandar
Department of Information Technology, North Eastern Hill University, Shillong, India
e-mail: khamitab@gmail.com

K. Medhi
e-mail: kishoremedhi2015@gmail.com

D. Kandar
e-mail: kdebdatta@gmail.com

B. S. Paul
Department of Electrical and Electronic Engineering, University of Johannesburg, Johannesburg, South Africa
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_14

using one byte, the possible intensity values are in the range of 0–255. A two-dimensional matrix I(i, j) can be used to represent a grayscale image, where (i, j) is the location of a pixel in the image I and I(i, j) represents intensity level at location (i, j). An image corrupted by impulse noise can be defined as in Eq. 1 [1]

$$\mathrm{I(i,j)} = \begin{cases} X(i,j) \textit{ with probability } p \\ Y(i,j) \textit{ with probability } (1-p) \end{cases} \tag{1}$$

Here, X(i, j) represents the value of impulse noise with probability "p" and Y(i, j) represents the original pixel value with probability (1 − p). Based upon the value at X(i, j), there are mainly two type of noise model: random and fixed impulse noise model. In fixed impulse, noise model X(i, j) can have value either 0 or 255 [1, 2]. But in case of random impulse noise model X(i, j) can have any value within the range of 0–255.

The goal of impulse noise filter is to reduce the amount of noise and simultaneously preserve fine details (edges) present in the images [3]. The edges contain most of the valuable information present in the images. It is a challenging task to reduce noise and preserve edges in an image effectively. The intensity of noise and distribution of noises are unpredictable. Therefore, a noise filter must be capable of resolving uncertain information.

Fuzzy logic is a very useful tool for decision-making and it is capable of handling uncertainty. Fuzzy logic is an extension of classical set theory; an element can partially belong to a set [4]. The if-then-else rule in fuzzy system can be used for taking expert decision. On the other hand, artificial neural network (ANN) consists of a large number of interconnected neurons, the neurons work collectively to solve a specific problem. ANN can be trained; and during training, it learns from the previous experience and uses this knowledge to achieve the goal. ANN filters are robust, it can even reduce unknown noise [5]. The proposed filter is implemented in two phase: noise detection and reduction. In noise detection phase, pixels corrupted by impulse noise are identified by using fuzzy logic. And in noise reduction phase, noisy pixels are filtered by using multilayer perceptron (MLP) keeping the noise free pixels unchanged. MLP is one of the most commonly used ANNs. The proposed filter is expected to suppress noise and preserve edges, since healthy edge pixels will be kept unchanged and the noisy pixels will be reconstructed.

Proposed Impulse Noise Filter

The proposed filter filters impulse noise in two phases: noise detection phase and noise removal phase. Fuzzy rules are used for detecting impulse noise affected pixel. And MLP is use to reduce impulse noise from the corrupted pixel.

Fuzzy Noise Detection Phase

Most of the filtering technique filters noisy as well as noise free pixels which hamper the quality of result. The main objective of this phase is to detect noisy pixels so that filtering can be performed only on the noisy pixel. To do this, a moving window of size 3 × 3 is used as shown in the Fig. 14.1.

In Fig. 14.1, X(i, j) represents the central pixel and the rest are neighbors of the central pixel. The difference between each neighboring pixel and the X(i, j) is denoted by Δi, where i ∈ {1, 2, 3, 4, 6, 7, 8, 9}. Let Δs1 and Δs2 represent the two smallest values in Δi. The average of the two most similar neighbor is computed as Δa = (Δs1 + Δs2)/2.

`Rule 1.1: If Δa is LARGE then impulse noise is present in X(i,j).`

Let us assume that there is a line whose width is just one pixel, and X(i, j) is also located in the line with no impulse noise then Δa will have very small value, since the other two neighboring pixel having similar intensity value will be in the line, and Δa will have large value if X(i, j) is noisy. However, Rule 1.1 may fail if central pixel is not noisy but neighboring pixels are noisy. To solve this problem, we determine the average of eight differences Δi. $\Delta avg = \frac{\sum \Delta i}{8}$.

`Rule 1.2: If the Δavg is LARGE then impulse noise is present in X(i,j).`

Rule 1.1 and Rule 1.2 are combined in Rule 1.3.

`Rule 1.3: If Δa is LARGE and Δavg is LARGE then impulse noise is present in X (i,j).`

LARGE is a fuzzy membership function, it is defined as LARGE = 1 − SMALL. And SMALL is defined by using triangular shaped membership function as shown in the Fig. 14.2.

Fig. 14.1 Pixels under 3 × 3 window

1 X(i-1,j-1)	2 X(i-1,j)	3 X(i-1,j+1)
4 X(i,j-1)	5 X(i ,j)	6 X(i,j+1)
7 X(i+1,j-1)	8 X(i+1,j)	9 X(i+1,j+1)

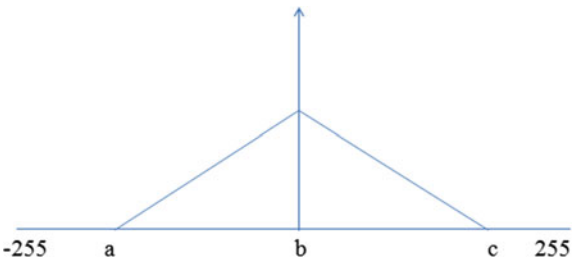


Fig. 14.2 Fuzzy membership function SMALL

The values of a, b, and c are chosen such that $-255 < a < b < c < 255$. The accuracy of the noise detection depends on the value of a, b, and c. Smaller value of "a" and larger value of "c" decreases the noise detection capability as only the high valued noise will be detected.

Noise Reduction Using Artificial Neural Network

If the central pixel is labeled as noisy then, it will be replaced by a value computed using MLP otherwise it will kept unchanged. ANN is implemented in two stages: learning and testing. During learning/training stage, network gains the capability to make accurate generalization [6]. In testing stage, the trained network is applied to solve problems. Training of ANN can be performed by using back propagation algorithm. Back propagation algorithm is composed of two sub-phases: (a) Forward computation and (b) back propagation of error.

The architecture of designed MLP is shown in Fig. 14.3. It consists of three layers: input layer with N neurons, hidden layer with N neurons, and the output layer consisting of single neuron. In case of 3×3 window, $N = 9$. We have used log-sigmoid transfer function in hidden layer and output layer. In forward phase, an input is supplied to the network and output is computed by using random weight. And in the back propagation of error phase, weights are modified to minimize total error of the network.

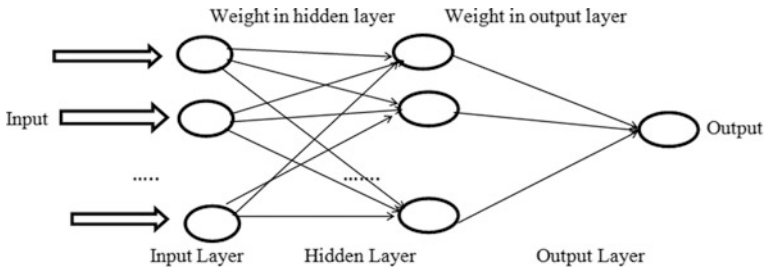


Fig. 14.3 Architecture of multilayer perceptron

Pixels under the window are converted into vector and supplied as an input to the MLP. The mean square error (MSE) between the actual output "ST" and desired output "OT" is computed. The steps involved in training phase are as follows" [6]:

```
Weights (win and wout) are initialized with small random values.
H_out = I * win
ST = H_out * wout
Until termination condition met
(i) For each input "I" compute network simulated output "ST"
(ii) Compute Δk for output unit K
 Δk = OT(1-OT)(ST-OT)
 Δwi = η*Δj*Δk
(iii) Compute Δh for hidden unit h
 Δh = OH(1-OH) ∑ wh.Δk
 Δwi = η*Δj*Δh
(iv) For each weight wi
 wi = wi + Δwi
End
```

where η and Δj represent the learning rate and gradient, respectively. In testing phase, the output "Out" is computed by using the weights determined in training phase. Let "I" be the input to the MLP, "win" and "wout" represent the weight matrix of input layer and output layer, respectively. The output of the hidden layer "H_out" is obtained by multiplying input with the input layer weight matrix, i.e., "H_out" = "I" * "win". The actual output is calculated by multiplying "H_out" with the weight matrix of output layer, i.e., "out" = "H_out" * "wout". The noisy pixels are filtered by replacing with "out".

Experimental Results

To evaluate the accuracy of the proposed filter, we have used a grayscale image added with different percentages of fixed and random impulse noise, peak signal-to-noise ratio (PSNR) and edge preserving factor (EPF) are used for evaluating. PSNR is defined as the ratio between the maximum gray intensity (MAX = 255) and the intensity of impulse noise which reduces the image quality.

$$\text{PSNR} = 10 \log \left(\frac{MAX^2}{MSE} \right) \tag{2}$$

where MSE, measure the average of square difference between the filtered image and the desired target image.

$$MSE = \frac{1}{MN} \tag{3}$$

The edge preserving factor (EPF) is computed using the formula [7, 8]:

$$EPF = \frac{\sum \left(\Delta x - \Delta x'\right)\left(\Delta y - \Delta y'\right)}{\sqrt{\sum \left(\Delta x - \Delta x'\right)^2 \left(\Delta y - \Delta y'\right)^2}}, \tag{4}$$

where Δx and Δy represent the edges present in the original image x and filtered image y, $\Delta x'$ and $\Delta y'$ are the mean value of Δx and Δy. The edges are extracted by using Laplacian operator [9].

In noise detection phase of the proposed approach, we have used fuzzy membership function SMALL and it is associated with three parameters a, b, and c. For Rule 1.1, the parameter values are taken as $a = -20$, $b = 0$, and $c = 20$ and for Rule 1.2, parameter values are $a = -20$, $b = 0$, and $c = 20$. The output of noise detection gives noise level "L" affecting the pixels in the range [0, 1], if $L = 0.3$ it is assumed that no noise is present, if $0.3 < L \leq 0.5$, we have used 3×3 window for filtering, if $0.5 < L \leq 0.9$ we have used 5×5 window and for $L > 0.9$ we have used 7×7 window.

In noise filtering phase, the termination criteria of the training algorithm is set to MSE = 0.0 or 250 iteration and bias = 0.7. To comparatively evaluate the filtering performance, we have used mean, median, FIRE, and MLP filter. Mean filter replaces the center of the window by mean of the neighboring pixels [9]. In median filter, center pixel value is replaced with the median of its neighborhood [10]. Fuzzy Inference Ruled by Else action (FIRE) [11] uses the difference between central pixel and its neighbors Δx. Using fuzzy rules, the correction parameter Δy is determine. The output pixel value y(i, j) is computed by adding Δy with original pixel value x(i, j). The FIRE filter uses two fuzzy triangular shaped membership function named as negative and positive. The MLP-based filter replaces all the pixels in the image by a value computed by using the neighboring pixel value; it is trained by using back propagation algorithm discussed in the preceding section. It is observed that with increase in window size for mean, median and MLP filter, EPF value reduces. So, we have used 3×3 window for all the filters except the proposed filter.

Tables 14.1 and 14.2 show the PSNR performance of the implemented filters on images corrupted by different percentages of fixed and random impulse noise. Tables 14.3 and 14.4 show the EPF performance of the implemented filters on images corrupted by different percentages of fixed and random impulse noise.

From our experimental result, it is clear that the proposed filter has the highest PSNR value on image affected with fixed impulse noise (Fig. 14.4). In Fig. 14.5, we can see that the proposed filter and median filter perform almost same in image affected with random impulse noise. In Figs. 14.6 and 14.7, the proposed filter has the highest EPF for fixed impulse and random impulse noise. It is also observed that

Table 14.1 PSNR performance of filters on image by different levels of fixed impulse noise

% of noise	Proposed filter	Mean filter	Median filter	ANN	Fire
40	14.60	12.01	13.22	11.64	13.69
20	17.20	15.23	16.45	11.70	16.80
10	22.11	20.04	20.50	11.77	19.63
9	22.40	20.71	20.93	11.89	19.80
7	23.17	21.20	22.34	12.17	20.17
5	24.91	21.63	22.85	12.26	21.43
2	26.79	23.01	24.77	13.01	23.27

Table 14.2 PSNR performance of filters on image affected by different percentages of random impulse noise

% of noise	Proposed filter	Mean filter	Median filter	ANN	Fire
40	17.79	17.20	17.30	10.37	17.43
20	18.93	17.59	17.45	12.45	17.23
10	23.19	22.63	26.34	16.07	22.36
9	26.27	22.91	26.39	19.14	26.02
7	27.62	23.37	26.60	20.00	26.21
5	27.96	23.89	26.82	22.14	26.33
2	28.83	24.68	27.14	24.57	27.48

Table 14.3 EPF performance of filters on image affected by different percentage of fixed impulse noise

% of noise	Proposed filter	Mean filter	Median filter	ANN	Fire
40	2.33	1.6	1.6	1.14	1.84
20	3.36	2.55	3.03	2.12	3.10
10	4.23	3.08	4.19	3.17	4.17
9	4.79	3.10	4.27	3.45	4.30
7	4.87	3.39	4.49	4.23	4.40
5	4.95	3.63	4.72	4.45	4.59
2	4.99	3.96	4.82	4.55	4.83

Table 14.4 EPF performance of filters on image affected by different percentages of random impulse noise

% of noise	Proposed filter	Mean filter	Median filter	ANN	Fire
40	2.96	2.60	2.42	1.14	2.80
20	4.08	3.12	3.91	1.35	2.90
10	4.63	3.65	4.49	1.42	3.88
9	4.66	3.73	4.60	1.65	4.51
7	4.81	3.87	4.51	1.72	4.58
5	4.87	4.05	4.73	1.80	4.42
2	5.18	4.23	4.87	1.92	4.76

PSNR and EPF performance of all the filter reduces with increase in amount of noise.

For highly corrupted images, the proposed filter can be applied recursively to improve the result. Figure 14.8 shows an image affected with 40% impulse noise, after filtering one time with the proposed approach, the resulting image has

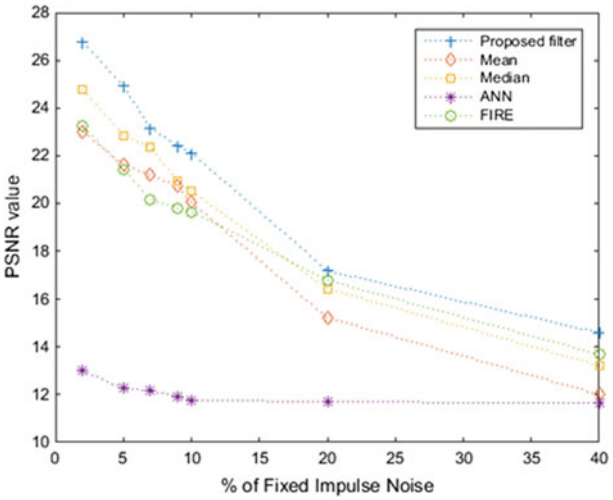


Fig. 14.4 Plot of PSNR value w.r.t random impulse

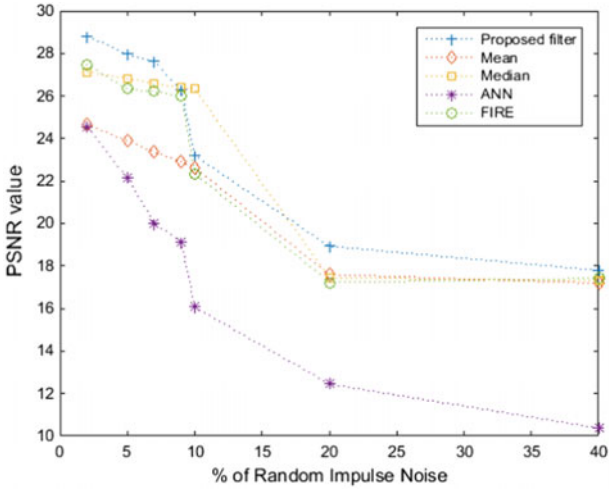


Fig. 14.5 Plot of PSNR value w.r.t fixed impulse noise

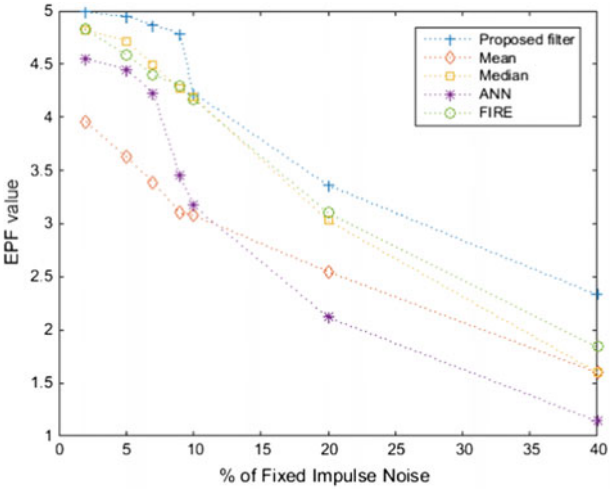


Fig. 14.6 Plot of EPF value w.r.t fixed impulse noise

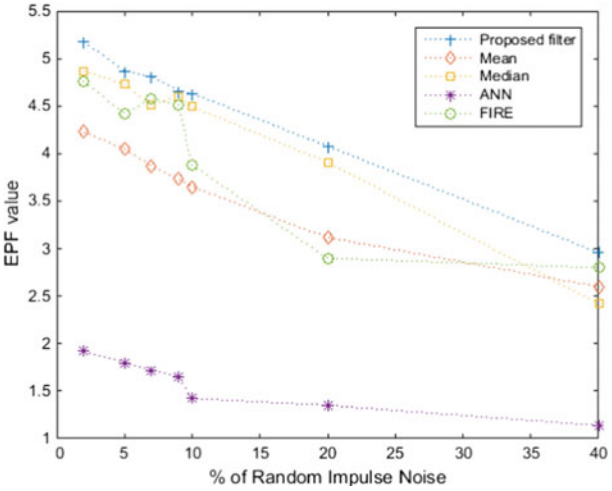


Fig. 14.7 Plot of EPF value w.r.t random impulse noise

PSNR = 15.51 and EPF = 2.76, and after filtering twice recursively by using the proposed filter, the result is increased to PSNR = 15.9 and EPF = 2.84. The filter can be applied recursively four or five times depending on the amount of noise.

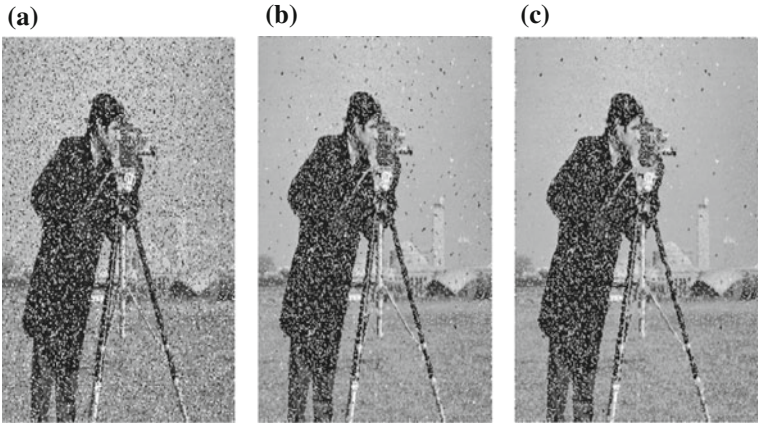


Fig. 14.8 **a** Noisy image, **b** After first filtering, **c** After second filtering

Conclusion

This paper presents a hybrid impulse noise filter. The filter is implemented in two phase: noise detection and reduction. Fuzzy rules are used for detecting the noisy pixel and MLP is used for reducing the noise. The accuracy of the proposed filter is comparatively evaluated with mean, median, FIRE, and ANN filter based on PSNR and EPF. It was found that the proposed filter is capable of reducing impulse noise and simultaneously preserves edges present in the images. We have tested on images affected by different percentages of fixed and random impulse noise. Among the implemented filters, the proposed filter has the highest PSNR value in images affected by fixed impulse noise, and highest EPF value in images affected by fixed and random impulse noise. For highly corrupted images, the proposed filter can be used recursively.

References

1. Roy, Amarjit, Salam Shuleenda Devi, and R. H. Laskar. "Impulse noise removal from gray scale images based on ANN classification based Fuzzy filter." *Computational Intelligence and Networks (CINE), 2016 2nd International Conference on*. IEEE, 2016.
2. Schulte, Stefan, Mike Nachtegael, Valérie De Witte, Dietrich Van der Weken, and Etienne E. Kerre. A fuzzy impulse noise detection and reduction method. *IEEE Transactions on Image Processing* 15(5): 1153–1162, 2006.
3. Zimmermann, H-J. Fuzzy set theory. *Wiley Interdisciplinary Reviews: Computational Statistics 2(3):* 317–332, 2010.
4. Amitab, Khwairakpam, DebdattaKandar, and Arnab K. Maji. "Comparative Evaluation of Radial Basis Function Network Transfer Function for Filtering Speckle Noise in Synthetic

Aperture Radar Images." Emerging Research in Computing, Information, Communication and Applications. Springer Singapore, pages: 243–252, 2016.
5. Budak, Cafer, Mustafa Türk, and Abdullah Toprak (2015). "Reduction in impulse noise in digital images through a new adaptive artificial neural network model." *Neural Computing and Applications,* 26(4): 835–843, 2015.
6. Ghosh, A., & Chakraborty, M. Hybrid Optimized Back propagation Learning Algorithm for Multi-layer Perceptron. *International Journal of Computer Applications*, *57*(1): 1–6, 2012.
7. Wang, Guobao, and Jinyi Qi. "Penalized likelihood PET image reconstruction using patch-based edge-preserving regularization." *IEEE transactions on medical imaging* 31.12 (2012): 2194–2204.
8. Umamaheswari, G., & Vanithamani, R., An adaptive window hybrid median filter for despeckling of medical ultrasound images. *International journal of scientific and industrial research*, 73(1): 100–102, 2014.
9. Gonzalez, R. C. *Digital image processing. Pearson Education, India*, 2009.
10. Chan, R. H., Ho, C. W., & Nikolova, M. (2005). Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on image processing*, *14*(10), 1479–1485.
11. Qiu, Tian, Yong Yan, and Gang Lu. "An autoadaptive edge-detection algorithm for flame and fire image processing." *IEEE Transactions on instrumentation and measurement* 61.5 (2012): 1486–1493.

Chapter 15
Utility Aware Cooperative Spectrum Sensing Using Coalitional Game Theory

Prakash Chauhan, Meenakshi Sharma, Sanjib K. Deka and Nityananda Sarma

Abstract Cooperative spectrum sensing (CSS) emerges as an efficient technique for cognitive radio overcoming the issues of independent sensing. In this work, a distributed cooperative spectrum sensing scheme using non-transferable coalitional game model has been proposed for cognitive radio networks (CRNs) which contributes to improve the opportunity detection and throughput. To determine the cooperative decision, a novel algorithm for selecting a fusion center is also proposed. The performance evaluation of the proposed CSS scheme is carried out through simulation study.

Keywords Cognitive radio (CR) · Coalitional game theory · Cooperative spectrum sensing (CSS) · Utility function

Introduction

With the advent of wireless communication demand, the traditional spectrum allocation policy results in spectrum holes [4]. To utilize the spectrum holes efficiently, J. Mitola proposed the concept of cognitive radio (CR) [8]. CR emerges as an effective technique for dynamic spectrum access opportunistically. The two types of users in CRN are the CR user or secondary user (SU) and the licensed user known as primary user (PU). The core functionalities of the CR are spectrum sensing, spectrum management, spectrum sharing, and spectrum mobility. Spectrum sensing has been

P. Chauhan (✉) · M. Sharma · S. K. Deka · N. Sarma
Department of Computer Science and Engineering, Tezpur University, Tezpur 784028, Assam, India
e-mail: prakashc@tezu.ernet.in

M. Sharma
e-mail: amee@tezu.ernet.in

S. K. Deka
e-mail: sdeka@tezu.ernet.in

N. Sarma
e-mail: nitya@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_15

the key to discover the opportunities by detecting the spectrum holes. It has been revealed [1] that the detection performance of local spectrum sensing suffers from several environmental issues. In order to improve the spectrum sensing performance, cooperative sensing techniques have been proved to be efficient. Modeling of cooperative spectrum sensing (CSS) framework is performed using: Game theory [9], parallel fusion [12], particle swarm optimization [10], random matrix theory [3], and machine learning [11] techniques. Amongst all, game theory has been proven as an efficient and robust technique for modeling the CSS framework, which can incorporate dynamic interactions during the cooperation. Game theory can broadly be classified into two categories: cooperative and noncooperative. In this work, cooperative game theory based on coalitional game has been used to model the proposed CSS framework. In the literature [2, 5–7, 9], a number of approaches have been proposed which formulate CSS for CRN using cooperative game theory. However, a mechanism to model a CSS framework considering the impact of coalition size on the average utility achieved by the SUs during cooperation is important for effective usage of the spectrum holes. Motivated with this fact, we propose a cooperative spectrum sensing (CSS) scheme using coalitional game theory which considers the impact of coalition size on the average utility achieved by the SUs during coalition formation. We propose a distributed algorithm for CSS by coalition formation based on merge and split operations. A technique to select a fusion center for a coalition is also proposed to assist the CSS framework depending upon merge and split operations. A simulation study has been carried out to evaluate the performance of the proposed CSS scheme.

This paper is organized as follows: the system model and assumptions are described in section "System Model and Assumptions". Section "Game Theoretic Formulation of the Proposed Distributed CSS" presents the game theoretic formulation of the proposed CSS scheme. Simulation results and analysis are demonstrated in section "Simulation Results and Analysis" followed by conclusion and future work in section "Conclusion and Future Works".

System Model and Assumptions

A CRN consisting of N number of SUs and a PU is considered. A time-slotted system is assumed where the entire time is divided into K number of slots of duration of T seconds each in which the SUs and the PU synchronize perfectly. It is assumed that if PU is absent at the beginning of any time slot, it will remain absent for that entire slot. Again each time slot of duration T is divided into three sub-slots as s, r and t which represent sensing, reporting, and transmission time, respectively. To perform local/noncooperative sensing, energy detection technique has been used.

We assume that SUs within a coalition sense same channel at a time and report their local sensing decisions to the fusion center (FC) using a dedicated channel in TDMA fashion. The determination of FC within a coalition is done using Algorithms 1 and 2. The FC combines the local sensing decision and propagates the final decision to each member of the coalition. For any coalition S, the probability of false alarm

and probability of detection can be expressed by

$$P_{f,S} = 1 - \prod_{i \in S} (1 - P_{f,i}) \tag{15.1}$$

$$P_{d,S} = 1 - \prod_{i \in S} (1 - P_{d,i}), \tag{15.2}$$

where $P_{f,i}$ and $P_{d,i}$ represent probability of false alarm and probability of detection for an SU i, which can be computed by the equations (Eqs. (15.1), (15.2)) as given in [6].

For a licensed channel at a time slot T_k, where k = 1…p (p represents number of time slot taking into consideration), the probability of presence and absence of PU are given by P_{H1} and P_{H0}, respectively where H1 and H0 indicates the binary hypothesis of presence and absence of PU signal. For the sake of simplicity, we assume that FC uses OR fusion rule on local sensing decisions to make a global decision. The suitability of selecting a fusion rule is out of the scope of this paper.

Game Theoretic Formulation of the Proposed Distributed CSS

The proposed CSS scheme is modeled as a non-transferable coalitional game given by $G(N,U)$, where N and U represent the set of SUs in the network (i.e., players) and their corresponding utility, respectively. The proposed game partitions the CRN into multiple nonoverlapping subsets called coalitions. Following are the definitions used to describe the characteristics of the game given by

Definition 1 Partition: Let N be the set of SUs in a CRN. Then the set S = $\{S_1,S_2,S_3,S_4,\ldots,S_k\}$ is a partition of N if $S_i \cap S_j = \phi$, $\forall(i,j)$ $1 \leq i,j \leq k$, i ≠ j and $\sum_{i=1}^{k}(|S_i|) = \mathrm{N}$, where N represent the cardinality of set N and $|S_i|$ represents the cardinality of coalition S_i.

Definition 2 Stability of Partition: A partition S = $\{S_1,S_2,\ldots,S_k\}$ of a set N is a stable partition if S_i does not have any incentive to merge with S_j nor has an incentive to split into S_g and S_h ($\forall\{i,j,g,h\}, i \neq j, g \neq h, S_g, S_h, S_i, S_j \in \mathrm{S}$) and ($0 < |S_g|, |S_h|, |S_i|, |S_j| < \mathrm{N}$).

Definition 3 Preference of Coalition: A coalition Q is said to be preferable than coalition R, i.e., $Q \rhd R$ iff $U(Q) \geq U(R)$, where $\rhd$ is the preference operator.

Definition 4 Merge Operation: A coalition Q will try to form a joint coalition with R if both Q and R be the member of an unstable partition say U i.e., both Q and R gain the mutual benefits by forming a joint coalition.

Definition 5 Split Operation: A coalition S will split into sub-coalitions Q and R if S be the member of a unstable partition say U i.e., Q or R have higher utility than S ($Q \rhd S$ or $R \rhd S$).

Definition 6 Grand Coalition: A coalition Q is said to be a grand coalition if it is the only member of the partition S, i.e., the coalition Q contains all the SUs of network N.

Definition 7 Singleton coalition: A coalition Q is said to be a singleton coalition if it contains only one element. A partition S=$\{S_1, S_2, \ldots, S_k\}$ of network N is said to be a collection of singleton coalition if $|S| = |N|$

Utility Function Design

The utility function incorporates overall gain achieved through coalition formation considering cost accrued in terms of the total overhead/penalty incurred/received during the game. The gain component in the utility function represents the improvement of probability of detection, throughput that can be achieved, while the cost incurred represents the penalty due to miss detection, energy consumption, and time spent during coalition formation. We design the utility function for G in order to be aware of average utility of each users within coalitions, i.e., even if the gain is higher than the cost for the newly formed coalitions but if the average utility of them is lower than the previously formed coalitions, then coalition formation will not take place. Inspired by [6], the utility function $U(S)$ for a coalition S can be modeled as

$$U(S) = \frac{Gain(S) - Cost(S)}{|S|}, \tag{15.3}$$

where $Gain(S)$, $Cost(S)$ and $|S|$ represent the gain achieved, loss incurred, and size of the coalition S, respectively, which are given by $Gain(S) = P_{H0}(1 - P_{f,S})(T - s - r|S|) \sum_{i=1}^{|S|}(\bar{r}_i)$ and $Cost(S) = P_{H1}(T - s - r|S|)\delta(1 - P_{d,S})$, where $|S|$ indicate size of the coalition S, i.e., total number of SUs in S, $\bar{r}_i$ represents data rate of SU i and δ represents penalty factor for miss detection.

The absolute throughput of each SU within a coalition depends on parameters like bandwidth of the channel, detection as well as false alarm probability, length of transmission duration, channel access mechanism, encoding scheme, input power, etc. Therefore, designing a weighted throughput equation could be the provision to check the efficacy of the proposed CSS model against noncooperative counterpart, which is given by Eq. (15.4). The same can be given by Eq. (15.5) for noncooperative spectrum sensing.

$$R_{i,C} = \frac{P_{H0}(1 - P_{f,S})(T - s - r|S|)\bar{r}_i}{1 - P_{d,S}} \tag{15.4}$$

$$R_{i,NC} = \frac{P_{H0}(1 - P_{f,i})(T - s)\bar{r}_i}{1 - P_{d,i}}, \tag{15.5}$$

where $R_{i,NC}$ and $R_{i,C}$ represent weighted throughput of SU i for noncooperative and the proposed cooperative model of spectrum sensing, respectively. It is assumed that $(1 - P_{d,i}) \neq 0$ and $(1 - P_{d,S}) \neq 0$ which is true for many real-time scenarios having with very low received *SNR*.

In the next section, we proposed an algorithm for Fusion Center(FC) selection for a coalition S. Initially singleton coalitions containing an SU will be formed, where the respective SU itself will be the FC for that particular coalition. While two singleton coalitions merge to form a larger coalition or a large coalition splits into subcoalitions and the SU satisfying certain criteria will be chosen as the FC as described in Algorithm 1 and 2.

Fusion Center selection algorithm

Algorithm 1: :FC selection during merge operation

1 **Input**: Any two coalitions S_i and S_j with their respective FC, i.e., FC_{S_i} and FC_{S_j}
2 **Result**: FC of the merged coalition S_{ij}
3 ***Steps***:
1. If $(P_{d,FC_{S_i}})^2 - (P_{f,FC_{S_i}})^2 \leq (P_{d,FC_{S_j}})^2 - (P_{f,FC_{S_j}})^2$ goto *Step 2* Else goto *Step 3*
2. $FC_{S_{ij}}$=FC_{S_j}
3. $FC_{S_{ij}}$=FC_{S_i}
4. Exit.

Property 1: Proposed coalition formation game has nontransferable utility.

Proof For an SU i in coalition S, the individual probability of detection and probability of false alarm is same as probability of detection and probability of false alarm of the coalition S respectively, i.e., $P_{f,i} = P_{f,S}$ and $P_{d,i} = P_{d,S}$. This is due to the fact that within a coalition, the final decision about the presence of PU on a channel is made by the FC who broadcasts the final decision to every member of that particular coalition. That means, the utility of each SU in a coalition is equal to the utility of that coalition. Therefore, the utility of coalition S cannot be arbitrarily distributed amongst its members. Hence, it can be established that the game has nontransferable utility.

Time complexity of Algorithm 1: At any stage of coalition formation, the algorithm takes one comparison to decide the FC for the newly formed coalition. So, the time complexity of the algorithm would be $O(1)$.
Time complexity of Algorithm 2: During split operation, a coalition splits into smaller

Algorithm 2: :FC selection during split operation

Input: Any coalition S
Result: FCs of the newly formed sub-coalitions
Steps:

1. If $|S|$=1 then the only SU of coalition S acts as FC and goto *Step 3*.
2. If $|S| \neq 1$ then goto *Step 2(a)*
   (a) FC=SU_1
   (b) For $i = 2$ to $|S|$ repeat *Step 2(c) to 2(d)*
   (c) If $(P_{d,FC})^2 - (P_{f,FC})^2 \leq (P_{d,SU_i})^2 - (P_{f,SU_i})^2$ goto *Step 2(d)*
   (d) FC=SU_i
3. Exit.

coalitions. For each of the smaller coalitions, the total number of comparisons required to decide a FC will depend on the size of that coalition. Therefore, a coalition of size m requires at most m number of comparison to decide the FC. Hence, the proposed algorithm has $O(m)$ worst time complexity.

The Distributed CSS Algorithm (DCSS)

Algorithm 3 realizes the proposed coalitional game theoretic CSS model for CRN.

Algorithm 3: :Distributed CSS Algorithm

Input: The information of P_f and *SNR* of each SU within a coalition S
Result: The final coalitional decision
Steps:

1. Each SU forms a singleton coalition by considering itself as only member and acts as FC for that coalition.
2. Every SU performs local sensing and makes a local decision themselves about the existence of the PU in the licensed channel.
3. Any two coalitions S_i and S_j combine to form a larger coalition S_{ij} iff
   (a) $S_{ij} \rhd S_i$ and $S_{ij} \rhd S_j$
4. If coalition is formed then FC is decided using Algorithm 1.
5. Any existing coalition S may split into smaller coalitions S_i and S_j iff
   (a) $S_i \rhd S$ or $S_j \rhd S$
6. If split operation occur then FC for each new sub-coalitions can be decided using Algorithm 2.
7. *Step* 3 to 4 are repeated until a stable partition is formed.
8. Once a stable partition (according to *Definition* 2) is formed each SU of the coalition S reports their local sensing results to the FC.
9. FC makes a global decision by applying OR-rule on the received local sensing decisions.
10. FC broadcast the final decision to each of the SUs within the coalition using dedicated control channel.

Time complexity of Algorithm 3: The worst case complexity of the algorithm is decided mainly by the occurrences of merge and split operations during coalition formation. So, the complexity of the algorithm can computed as follows:

Merge Operation: Suppose there are n number of singleton coalitions in the network represented by $S=\{S_1, S_2, S_3, \ldots, S_n\}$. At first iteration, each S_i tries to form coalition with every other S_j, $\forall(i,j), i \neq j, S_i, S_j \in S$. At worst case, only one coalition of size 2 is formed which takes total $n(n-1)$ comparisons resulting with $(n-1)$ numbers of coalitions. This process is repeated until the partition of the network becomes stable. In worst case, a grand coalition may get formed with total number of comparison given by $T_1(n) = n(n-1) + \cdots + 2 \times 1 = O(n^3)$.

Split Operation: The same worst case complexity occurs for split operation while a large coalition of size n gets split into n singleton coalitions. In this case, at the first iteration, a coalition of size n gets split into two smaller coalitions of size 1 and $n-1$. Similarly in second iteration, the coalition of size $n-1$ is further split into two coalition of size 1 and $n-2$ and the entire process repeats until n singleton coalitions are formed. Therefore, the worst case time complexity can be given by $T_2(n) = n + (n-1) + \cdots + 1 = O(n^2)$ Therefore, the overall time complexity of the proposed CSS algorithm can be computed as: $T(n) = T_1(n) + T_2(n) = O(n^3)$

Simulation Results and Analysis

In this section, we presented the results of simulation study to evaluate the efficacy of the proposed CSS model with respect to the noncooperative model for spectrum sensing. During the simulation using MATLAB, the values of the parameters taken for bandwidth is 100 KHz, $P_{H,0} = 0.6$, $T = 1$ s, $s = 10$ ms, $r = 1$ ms, *SNR* in range of −40 to 30 db and N=30.

Figure 15.1a demonstrates about how the P_d changes with *SNR* for both the proposed and noncooperative model of spectrum sensing. It can be observed that even at very low *SNR* of the PU signal, the proposed model results in higher value of P_d than the noncooperative model. This improvement in detection is achieved because of the collaborative sensing performed by the SUs.

Figure 15.1b, c present the achieved utility and the weighted throughput against *SNR* for both the models, respectively. Figure 15.1b shows that with increase in *SNR*, the utility for proposed model improves significantly compared to the noncooperative model. This happens because of the improvement in terms of $P_{d,S}$ which eventually minimizes the cost incurred due to the penalty for miss detection. Similarly, Fig. 15.1c shows that with increase in *SNR* the weighted throughput improvement can be observed for the proposed model in comparison to the noncooperative one.

Figure 15.1d shows the improvement of weighted throughput achieved against the slot duration. It reveals that with increase in the time slot duration the weighted throughput also get increased for both the models. But the rate of improvement in

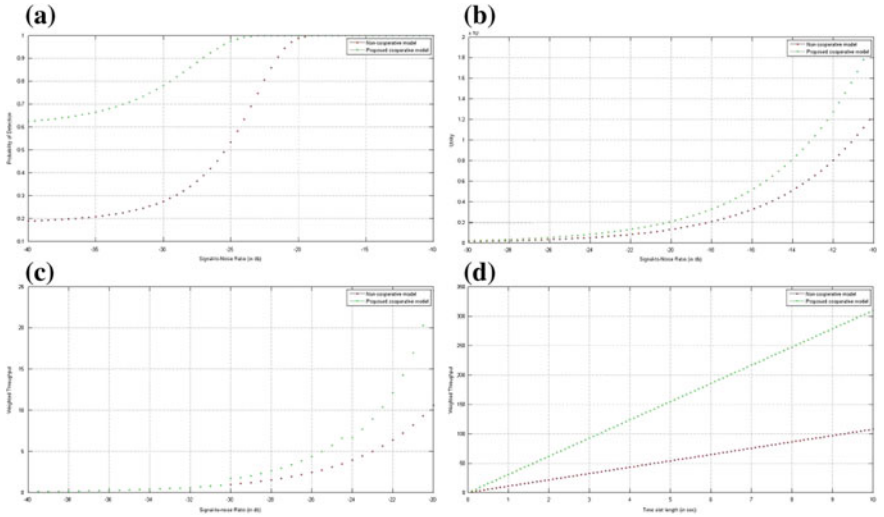


Fig. 15.1 **a** Signal-to-noise ratio (in db) versus probability of detection, **b** Signal-to-noise ratio (in db) versus utility, **c** Signal-to-noise ratio (in db) versus weighted throughput and **d** Time slot length (in s) versus weighted throughput

throughput in the proposed model is reasonably higher due to the improvement of $P_{d,S}$ in comparison to the noncooperative model.

Conclusion and Future Works

In this paper a distributed cooperative spectrum sensing scheme using the coalitional game theory is proposed. The proposed scheme aims to overcome the drawback of individual spectrum sensing. The distributed algorithm incorporates the utility aware sensing by SUs during coalition formation. The game model creates stable coalitions resulting with nontransferable utility. The simulation results validate the performance of the proposed CSS scheme. Incorporation of multichannel sensing with heterogeneous SUs into the scheme is quite interesting and is left for future work.

Acknowledgements This work is supported by MHRD, under FAST proposal and UGC, Government of India under SAP level-II. The authors are thankful to both the funding agencies.

References

1. Akyildiz, I.F., Lo, B.F., Balakrishnan, R.: Cooperative spectrum sensing in cognitive radio networks: A survey. Physical communication 4(1), 40–62 (2011)
2. Balaji, V., Hota, C.: Efficient cooperative spectrum sensing in cognitive radio using coalitional game model. In: Contemporary Computing and Informatics (IC3I), 2014 International Conference on. pp. 899–907. IEEE (2014)
3. Cardoso, L.S., Debbah, M., Bianchi, P., Najim, J.: Cooperative spectrum sensing using random matrix theory. In: Wireless Pervasive Computing, 2008. ISWPC 2008. 3rd International Symposium on. pp. 334–338. IEEE (2008)
4. Commission, F.C.: Spectrum policy task force report. Report ET Docket no. 02-135 (2002)
5. Deka, S.K., Chauhan, P., Sarma, N.: Constraint based cooperative spectrum sensing for cognitive radio network. In: Information Technology (ICIT), 2014 International Conference on. pp. 63–68. IEEE (2014)
6. Hao, X., Cheung, M.H., Wong, V.W., Leung, V.C.: A coalition formation game for energy-efficient cooperative spectrum sensing in cognitive radio networks with multiple channels. In: Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE. pp. 1–6. IEEE (2011)
7. Lu, Y., Duel-Hallen, A.: Game-theoretic framework for cooperative sensing and fair spectrum access in multichannel cognitive radio networks. arXiv:1605.01627 (2016)
8. Mitola, J., Maguire, G.Q.: Cognitive radio: making software radios more personal. IEEE personal communications 6(4), 13–18 (1999)
9. Saad, W., Han, Z., Debbah, M., Hjorungnes, A., Basar, T.: Coalitional games for distributed collaborative spectrum sensing in cognitive radio networks. In: INFOCOM 2009, IEEE. pp. 2114–2122. IEEE (2009)
10. Shami, T.M., El-Saleh, A.A., Kareem, A.M.: On the detection performance of cooperative spectrum sensing using particle swarm optimization algorithms. In: Telecommunication Technologies (ISTT), 2014 IEEE 2nd International Symposium on. pp. 110–114. IEEE (2014)
11. Thilina, K.M., Choi, K.W., Saquib, N., Hossain, E.: Machine learning techniques for cooperative spectrum sensing in cognitive radio networks. IEEE Journal on selected areas in communications 31(11), 2209–2221 (2013)
12. Varshney, P.K.: Distributed bayesian detection: Parallel fusion network. In: Distributed Detection and Data Fusion, pp. 36–118. Springer (1997)

Chapter 16
Codon-Based Analysis of Alzheimer's Disease (AD) Using Soft Computational Tool

Hemashree Bordoloi and S. R. Nirmala

Abstract Alzheimer's disease is a most common disease seen in today's world. It is one form of dementia where a person loses his memory progressively and finally loses his life. It is often seen in people above 60 but it may occur early. This disease destroys memory cell of brain. Till now, it is a disease without any treatment and proper way of diagnosis. Research shows that most often it occurs due to the deposition of defective structure of amyloid protein in the brain. In this paper, we have proposed a work to detect the defective amyloid protein using two classifiers. Secondary structure of amyloid protein is detected and analyzed in our work which provides a way to predict the cause of Alzheimer.

Keywords Alzheimer's disease · Codon · Amyloid protein
Artificial neural network

Introduction

Alzheimer's disease (AD) is a neurodegenerative type of neurological disorder. It involves death of brain cells which causes memory loss and cognitive decline. This disease starts slowly and progressively gets worse that is why it is also known as progressive disease. It destroys memory and other important functions of brain. It also represents one type of dementia. This disease normally occurs in elderly people. It starts with memory loss and finally patient loses their conversation as well

H. Bordoloi (✉)
Department of Electronics & Communication Engineering, ADBU, Assam, India
e-mail: hemashree.bordoloi@dbuniversity.ac.in

S. R. Nirmala
Department of Electronics & Communication Engineering,
Gauhati University, Assam, India
e-mail: nirmalasr3@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_16

as responding ability. Death of brain cell occurs over a course of time in case of Alzheimer's patient. It has been seen that in the brain of a person suffering from Alzheimer, there is an existence of plaques and tangles. These are tiny inclusions seen in the nerve tissue of brain [1].

From clinical assessment, deposition of beta amyloid protein on the outside of the brain, primarily, between the spaces in the nerve cells causes AD. This type is known as Plaques. It destroys normal brain cell as well as cell-to-cell communication [2].

Plaques also occur in elderly people but in case of a person which has a chance of Alzheimer, these plaques directly effect the cells responsible for memory. Other symptoms of Alzheimer's disease are inability to receive and remember new information, difficulty in reasoning, complex tasking, impaired speaking, reading and writing, behavioral changes, etc. [1].

Related Works

Alzheimer's disease was named after German psychiatrist and neuropathologist Alois Alzheimer. He had described Alzheimer's disease in 1906 [1]. After that, several works have been completed on AD including detection of various factors that causes AD. Most of the work emphasizes on clinical aspect of AD but very less work has been done on soft computing based analysis of AD. Basically, beta amyloid protein is responsible for Alzheimer. Each protein has unique structure. In case of AD-affected brain, there is a change in the structure of this protein. In our proposed work, we have emphasized on the structure of protein. Proteins are made up of amino acids and these amino acids are composed of codons. We are analyzing these two proteins by considering the codons. A detailed literature review is presented which includes both clinical and soft computational treatment of AD.

Anoop et al. in their paper describe about various cerebrospinal fluid (CSF) based biomarkers of Alzheimer's disease. Biomarkers play vital role in the early development of AD before significant cognitive dysfunction [3].

Barber in his paper also described biomarker of AD such as beta amyloid protein. These biomarkers facilitate improved detection and diagnosis of AD [4].

Autoantibodies are used in the early detection of AD in the paper of DeMarshall et al. They have described various types of blood-based biomarkers of AD and various methods used to detect these biomarkers as a treatment of AD [5].

A backpropagation artificial neural network approach to classify neurodegenerative diseases is proposed by Chaudhury et al. in their paper. They have considered three neurodegenerative diseases which include Alzheimer. Using artificial neural network they have classified these three types of diseases with an accuracy of 96.42% [6].

Proposed Methodology

Beta amyloid protein is the biomarker of Alzheimer's disease. Mutation in the primary structure of both these two proteins leads to Alzheimer's disease. This protein leads to improper functioning of brain cells by destroying the cells [7]. Beta amyloid protein forms plaques outside the brain cells. Deposition of both these proteins leads to death of brain cells progressively. As a result person suffering from AD losses memory [8].

Detection of Alzheimer is primarily done either by considering the brain images or by considering the protein structure. Proteins are considered from the clinical aspect. In our work, we have proposed a technique based on codon concentration. Proteins are formed by the combination of amino acids and amino acids are formed by the codons. Hence, codons are the basic unit of proteins. So in our work, detection of Alzheimer is done by considering the abnormality in the codon sequence.

Proposed methodology for the detection of AD consists of the following steps:

- Sample collection
- Analysis of data using proper algorithm
- Use of soft computational tool to develop an algorithm for further analysis
- Testing with clinical data for validation of the proposed algorithm

Sample Collection

Beta amyloid protein is the biomarker of AD [9]. In our proposed work, we have considered the normal as well as abnormal codon sequence of both these two proteins. These sequences are treated by a ANN-based simulation model to detect the abnormality in the codon sequence.

The codon sequence of normal amyloid protein is—GAT GCT GAA TTT AGA CAT GAT AGT GGT TAT GAA GTT CAT CAT CAA AAA CTT GTT TTT TTT GCT GAA GAT GTT GGT AGT AAT AAA GGT GCT ATT ATT GGT CTT ATG GTT GGT GGT GTT GTT ATT GCT.

The codon sequence of amyloid protein of a person suffering from AD is—GAT GCT GAA TTT AGA GCT GCT AGT GGT TAT GAA GTT GCT GCT CAA GCT CTT GTT TTT TTT GCT GAA GAT GTT GGT AGT AAT AAA GGT GCT ATT ATT GGT CTT ATG GTT GGT GGT GTT GTT ATT GCT.

Analysis of Data Using Proper Algorithm

The proposed algorithm for the collected samples is given below

Step 1: Calculation of codon frequency
Step 2: Detection of codon concentration

Use of Soft Computational Tool to Develop a Algorithm for Further Analysis

In our approach, we have used two ANN classifiers. First classifier detects the four nucleotides and second classifier classifies 64 codons. Based on the appearance of codons, codon frequency is calculated. Frequency of codons determines whether there is abnormality or not. The system model for our proposed work is as follows

Four nucleotides are coded and fit to the first classifier. Based on the coding of nucleotides, 64 codons are coded and fit to the second classifier.

The two classifiers are designed using backpropagation algorithm. In both the classifiers, two hidden layers are used. To classify codons, networks are designed as multilayer perceptron feedforward network [10].

The first network is trained with four nucleotides. With normal and abnormal codon sequence of beta amyloid protein second network is trained. For testing coded data are used. The proposed model is as follows (Tables 16.1 and 16.2).

Two classifiers are used in our approach which is shown in Fig. 16.1. The entire system is designed by two level classifiers where four nucleotides sequences are given as input to the first classifier and codon sequences are given as input to the second classifier. First classifier identifies four nucleotides and second classifiers identify codon sequences as codon is nothing but the combination of three nucleotides (Fig. 16.2).

Table 16.1 Parameters of proposed classifier

ANN	MLP
Data set size	Training: 50
	Testing:: 40
Training type	Back propagation with Levenberg–Marquardt optimization
Maximum No. of Epochs	4500
Variance in training data	50%

Table 16.2 Performance of ANN setup

Epoch	Time (s)	MSE	% rate
1000	25	10^{-3}	90
1800	40	10^{-4}	93
2300	70	10^{-5}	95
3500	80	10^{-7}	100

Fig. 16.1 Proposed system model

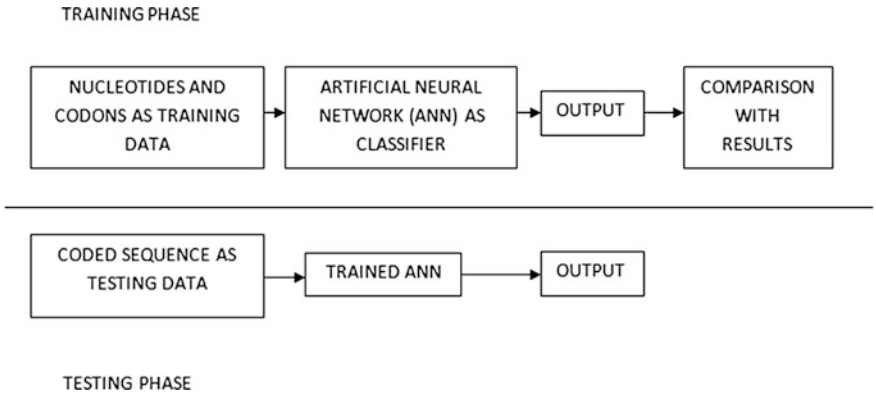


Fig. 16.2 Training and testing phases of the designed classifier

Results

Normal and abnormal codon sequence of beta amyloid protein is fed to the ANN classifier in the training phase. Error back propagation algorithm is used for training. 50 data are used during training phase. Designed ANN classifiers give the result with 100% accuracy.

10^{-7} performance goal is attained in around 80 s by the designed ANN. 50 samples are used for training and 40 samples are used for testing. 100% accuracy is shown by the designed ANN under best training conditions with a MSE convergence of about 10^{-7}. Training time is varied to check the robustness of the system. Both the classifiers give good performance. Codon frequency of both the normal and abnormal beta amyloid protein is shown in Figs. 16.3 and 16.4.

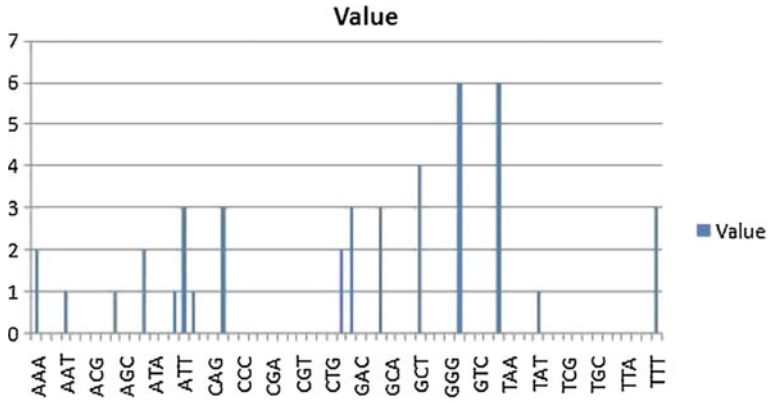


Fig. 16.3 Codon frequency for normal beta amyloid protein (without AD)

From the Figs. 16.3 and 16.4, we can observe the frequency of codons present in the amyloid protein and from this, we can conclude that the amino acid sequence of beta amyloid protein of a normal person as: D-A-E-F-R-H-D-S-G-Y-E-V-H-H-Q-K-L-V-F-F-A-E-D-V-G-S-N-K-G-A-I-I-G-L-M-V-G-G-V-V-I-A. This is the primary structure of amyloid protein in case of normal brain. Here, each letter signifies amino acids. For example, D means aspartic acid, A means alanine, and so on. In case of AD-affected brain, primary structure of amyloid protein is changed.

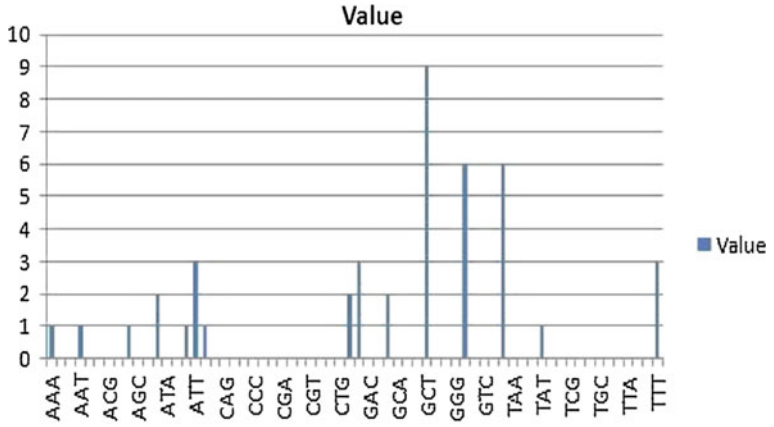


Fig. 16.4 Codon frequency for abnormal beta amyloid protein (with AD)

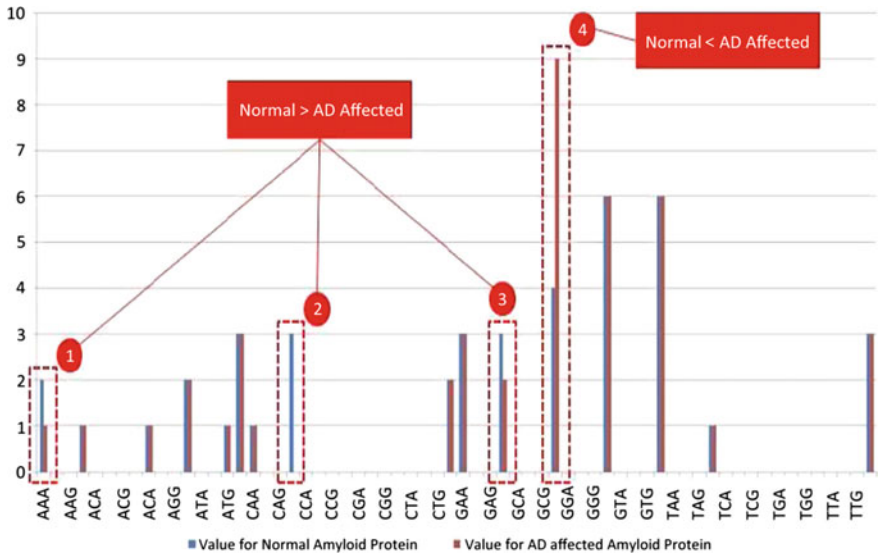


Fig. 16.5 Codon frequency comparison for normal and abnormal beta amyloid protein

In that case, some amino acids in the sequence are substituted by some other amino acids.

The amino acid sequence of amyloid protein in case of AD-affected brain is D-A-E-F-**A-A**-D-S-G-Y-E-V-**A-A**-Q-**A**-L-V-F-F-A-E-D-V-G-S-N-K-G-A-I-I-G-L-M-V-G-G-V-V-I-A. Finally, we can represent the difference in the codon frequency in both the cases by Fig. 16.5.

A simulation-based approach is described in our proposed work which detects the abnormality in the codon sequence of the protein structure. With appropriate configuration of the classifiers accuracy can be increased by rectifying the difficulties.

Conclusion

AD is the most common neurodegenrative disease seen in elderly people. With age progressively it appears. Life of a person suffering from AD detoriates day by day and it is also very difficult. There is no such medicine till now which can cure AD. So early detection of AD is necessary as a preventive measure. In our proposed work, we have considered the codon sequence of beta amyloid protein. These codons are primary unit of protein. So if the mutation or chances of mutation in the beta amyloid protein can be detected early then occurrence of AD can be minimized by following proper life style and preventive measures. This work can also help in the preparation of drugs for AD.

References

1. Bordoloi H, Sarma K K. "A Soft Computational Framework to Predict Alzheimer's Disease (AD) from Protein Structure", *Recent Trends in Intelligent and Emerging Systems.* Springer India 2015.
2. Berchtold NC, Cotman CW. "Evolution in the Conceptualization of Dementia and Alzheimer's Disease", *Greco-Roman Period to the 1960 s. Neurobiol.-Aging*. 1998; 19(3): 173–89.
3. Anoop A, Singh P K, Jacob R S and Maji S K. "CSF Biomarkers for Alzheimer's Disease Diagnosis", *International Journal of Alzheimer's Disease. PMC, NCBI.*
4. Barber R C, "Biomarkers for Early Detection of Alzheimer Disease", *The Journal of the American Osteopathic Association,* Vol 110.
5. DeMarshall C A, Nagele E P, Sarkar A, Acharya N K, Godsey G, Goldwaser E L, Kosciuk M, Thayasivam U, Han M, Belinka B, Nagele R G. "Detection of Alzheimer's disease at mild cognitive impairment and disease progression using autoantibodies as blood-based biomarkers. *Alzheimer's & Dementia Diagnosis, Assessment & Disease Monitoring,* Vol 3, 2016.
6. Chaudhary N, Aggarwal Y, Sinha R K. "Artificial Neural Network based Classification of Neurodegenerative Diseases ANN Classifies Neurodegeneration." *Advances in Biomedical Engineering Research (ABER), Volume 1, Issue 1,* March 2013.

7. Davies L, Wolska B, Hilbich C, Multhaup G, Martins R, Simms G, Beyreuther K, Masters CL. "A4 amyloid protein deposition and the diagnosis of Alzheimer's disease: prevalence in aged brains determined by immunocytochemistry compared with conventional neuropathologic techniques", *Neurology*. 1988 Nov; 38(11):1688-93, Department of Neuropathology, Royal Perth Hospital, Australia.
8. R Perneczky, L-H Guo, S M Kagerbauer, L Werle, A Kurz, J Martin and P Alexopoulos. "Soluble amyloid precursor protein β as blood-based biomarker of Alzheimer's disease" *Transl Psychiatry*, 2013 February, 3(2), e227.
9. James D. Doecke, Simon M. Laws, Noel G. Faux, William Wilson, Samantha C. Burnham, Chiou-Peng Lam, Alinda Mondal, Justin Bedo, Ashley I. Bush, Belinda Brown, Karl De Ruyck, Kathryn A. Ellis, Christopher Fowler, Veer B. Gupta, Richard Head, S. Lance Macaulay, Kelly Pertile, Christopher C. Rowe, Alan Rembach, Mark Rodrigues, Rebecca Rumble, Cassandra Szoeke, Kevin Taddei, Tania Taddei, Brett Trounson, David Ames, Colin L. Masters, Ralph N. Martins, "Blood-Based Protein Biomarkers for Diagnosis of Alzheimer Disease", *Alzheimer's Disease Neuroimaging Initiative and Australian Imaging Biomarker and Lifestyle Research Group, Arch Neurol.*2012, 69(10):1318-1325. https://doi.org/10.1001/archneurol.2012.1282.
10. S. Haykin. "Neural Networks A Comprehensive Foundation", *2nd ed., Pearson Education, New Delhi, 2003.*

Chapter 17
An Improvised Backpropagation Neural Network Model Based on Gravitational Search Algorithm for Multinomial Classification

R. Priyadarshini, M. R. Panda and N. Dash

Abstract Backpropagation (BP) algorithms are widely used for propagating adaptive learning to an artificial neural network (ANN), which uses a gradient descent learning approach to train an ANN. But choosing an efficient and appropriate training methodology is still a challenge. The gravitational search algorithm (GSA) is a recently developed metaheuristic approach which tries to find out a global solution by balancing both exploration and exploitation. Though GSA always is guaranteed to provide a better solution, some time it takes more time to generate a solution. In contrary to this, gradient descent-based algorithms have the limitation of getting stuck into local optima. In view of this, the proposed work has tried to do an amalgamation of both the techniques to avoid the individual deficiencies lying in each of the techniques. A synergy of both the techniques is validated by some of the benchmark datasets along with using some benchmark functions. The performance is measured and compared by running the algorithms individually and by the hybrid approach. The generated results establish the validity of the proposed model.

Keywords Gravitational search algorithm · Gradient descent
Mean squared error · Back propagation algorithm

R. Priyadarshini (✉) · N. Dash
Department of IT, C V Raman College of Engineering, Khordha, Odisha, India
e-mail: priyadarshini.rojalina@gmail.com

N. Dash
e-mail: nilamadhab04@gmail.com

M. R. Panda
Department of CSE, C V Raman College of Engineering, Khordha, Odisha, India
e-mail: mohit1146@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_17

Introduction

Data mining is the analysis of data for relationships that have not previously been discovered. It is accomplished by building models. A model uses an algorithm to act on a set of data. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. In this paper, we are working on several benchmark datasets to predict the classes for which we have used a hybrid model called as GSA-based backpropagation algorithm (G-BPNN) by making amalgamation of two techniques. They are artificial neural network (ANN) and gravitational search algorithm (GSA) to build up a classification model. GSA is a new heuristic search-based algorithm and is a new optimization algorithm based on the law of gravity and mass interactions. In this algorithm, the search agents are a collection of masses which interact with each other based on the Newtonian gravity and the laws of motion. Neural networks (NN) are an effective tool in the field of pattern classification. The training algorithm used in the field of pattern classification is backpropagation neural network. The basic idea behind it is to set weights randomly, apply the inputs to the NN and perform the learning with gradient descent learning and test output with learnt NN. We have tested and validated the model by implementing the above two algorithms individually on the same dataset and also by the proposed hybridized approach. A comparison has been made on the outcomes of the algorithm basing on accuracy of result by k-cross validation scheme.

Literature Survey

Research for effective heuristic algorithms in optimization areas remains an open issue. Lot of work has been done in the field of pattern classification with optimization techniques but no universal algorithm is capable of obtaining efficient and high performance for all problems. A naïve optimization algorithm has been developed in 2009 [1] based on the Newtonians law of gravity named as gravitational search algorithm (GSA), which has been successfully applied in many problem domains associated with optimization issues. The experimental results prove that this algorithm is performing at par with its contemporary techniques. This technique is well exploited in the field of pattern classification and data mining and shows the potentiality of the method in the said domain [2, 3]. GSA sometimes suffers from premature convergence and loses the ability of exploration and exploitation when solving complex problems [4]. So many researchers do the hybridization of GSA with others to overcome the deficiencies in individual techniques. In the paper [5], merging of simple operator with GSA is done to improve its performance. It also proposes a new linear time-invariant system theory-based convergence analysis of GSA. In the paper [6], high complexity of

multi-objective problems which affect the efficiency is taken into consideration. Another work is based on the hybridization of GSA and backpropagation algorithm for training of feedforward neural network [7]. It makes advantage of both the proposed algorithms. Experimental results show that the proposed hybrid algorithm outperforms in training feedforward neural network and the standard deviation calculated the values as 0.151317 which was better than the previous results.

C. Purcaru et al. used another hybridization employing GSA and PSO where path planning algorithm is used which minimizes the objective function and transform the solution to provide optimal results for mobile robot path planning [8]. The path generated is collision free in static environment with position of obstacles known in advance. To improve power system stability hybridization of GSA-PSO is used [9] to find controller parameters which concluded that for a multi-machine power system the modal oscillation is very dangerous and can be damped out with the proposed approach. With PSO having the ability of social thinking and GSA being the local search capability the GSA-PSO approach is used for the coordinated PSS and damping controller structure parameter tuning.

The paper [10] analyses the research done previously on the occurrence of diabetes using 'pima' dataset. Joint implementation of Support Vector Machine [11] and Naïve Bayes statistical modelling in the data set acquired from the medical examination of 402 patients were diagnosed. 201 instances were taken for training and 201 for testing which improves the overall reliability of the system outcome. If both the classifiers classify the records as being positive or negative, then the given output of a classifier is having higher reliability. The work done in [12] proposed a glucose tracking system based on Android platform and a post-prandial glucose prediction model on machine learning techniques was conceived where the sample data was collected from the diabetics' diet and analyzed for predicting postprandial glucose level of user. The experimental result show that this prediction can be used an auxiliary tool to control glucose and diet. The function in this experiment was needed to be optimized. Lavenberg–Marquardt algorithm has given optimal training results in [13]. It provides a comparative performance evaluation of various training algorithms in artificial neural network. The system was implemented on pima Indian of diabetes data and was observed that the algorithm achieves faster convergence compared to other conjugate gradient algorithms.

Proposed Model

The proposed model worked in two phases. In the first phase, the GSA has been applied to each of the class to obtain the centroids of each class. In the second phase, the obtained centroids are taken as inputs to the training phase of ANN. The algorithm for the proposed work is divided into several phases as given below:

Algorithm: The algorithm of the proposed model is divided into four phases, consisting of some steps explained below.

Phase-1: Initialization

Step 1: Divide the whole dataset S_i into two sets $S_{train(i)}$ and $S_{test(i)}$. Where 'i' is the number of datasets examined by the model. On $S_{train(i)}$ execute steps from 1 to 9.

Step 2: Initialize the GSA parameters and the structure of G-BPNN. The G-BPNN architecture contains one hidden layer. The initial weight sets are set randomly in between [0, 1]. The number of nodes are taken as 10 in the used architecture on a hit and trial basis.

Phase-2: Encoding Strategy: There are three encoding techniques available for representing weights and biases in feedforward neural network [14], which are vector, matrix and binary encoding schemes

Step 3: Perform the suitable encoding for G-BPNN structure. In this work, we have used matrix encoding scheme. An example of this strategy is given below:
If W_{H1} represents the weight matrix from hidden to output layer an B_1 is the bias matrix, they can be given by:

$$W_{H1} = \begin{pmatrix} W_{11} & W_{n1} \\ W_{11} & W_{nm} \end{pmatrix}$$

where 'n' is the number of nodes in the hidden layer and 'm' is the number of nodes in the input layers.

Phase-3: Fitness Evaluation and Finding the centroid for each class of $S_{train(i)}$: The fitness function is taken as the Euclidean distance function [15] and for training the G-BPNN, the activation function used for calculating the output of input layer is a linear function where as to compute the output of hidden layer, a nonlinear sigmoidal function is used, for weight updation gradient descent [16] method is used.

Step 4: Assess the fitness value of all candidate solutions with respect to the initially set candidate solution. The distance function is chosen as the fitness function here.

Step 5: Set the instance with maximum distance as worst(t), the minimum distance as best(t) and Mi(t) for all the agents.

Step 6: Compute mass (M), forces (F) and acceleration (a) of all agents.

Step 7: Update the agents' position and velocity.

Step 8: If the stop condition is satisfied, then go to Step 9; else go to Step 3.

Step 9: Collect the centroids for each class of training data. For example if the dataset contains two classes, then it gives two centriods and if a dataset contains 'c' number of classes, it will generate 'c' number of centroids.

Step 10: Use the BPNN algorithm to train the G-BPN where the target data are the centroids found from step-8 till convergence.

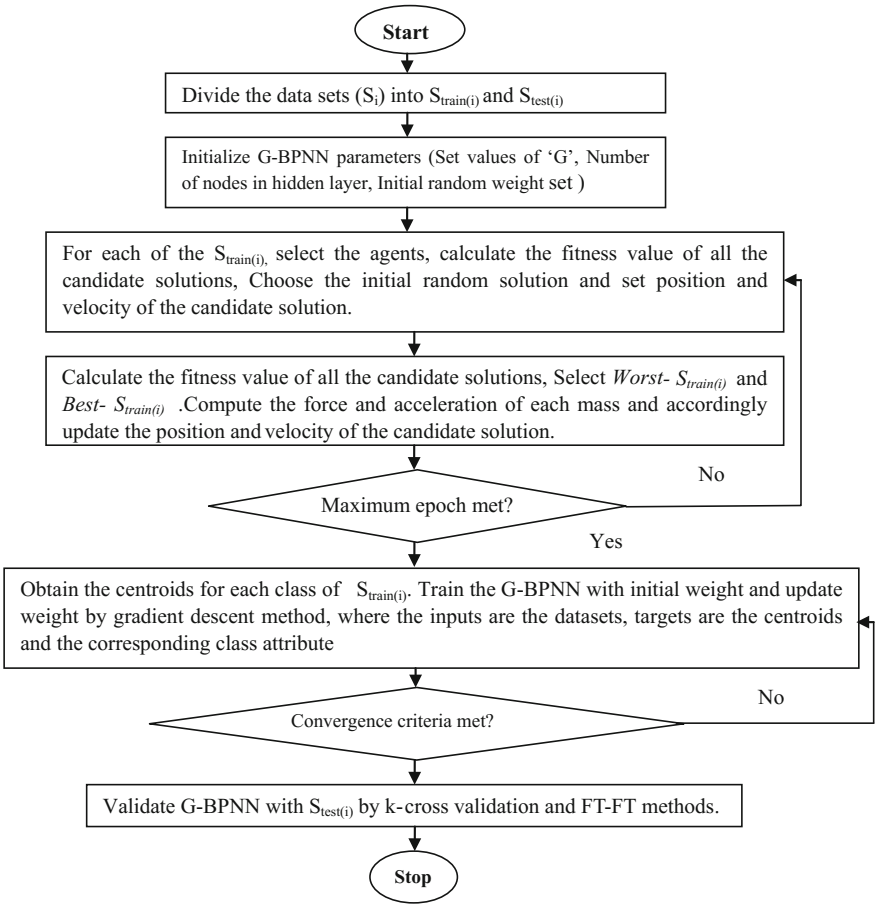


Fig. 17.1 Flow chart representation of proposed model-G-BPNN

Phase-4: Testing and validation: After testing the model with the test data, the accuracy is obtained by using K-cross validation [16] and full-train and full-test (FT-FT) [17] method. The percentage of accuracy is calculated by

$$Accuracy\ percentage\ Accuracy\ percentage = \frac{Number\ of\ instances\ correctly\ by\ the\ classifier}{Number\ of\ actual\ instances\ present\ in\ the\ dataset}$$

Step11: Test the model using the fractional data: $S_{test.}$

Step-12: valid the model by using k-cross validation [18] (here, k = 10) and FT-FT method. The flowchart for the given algorithm is given by Fig. 17.1

About GSA

In the year 2009 Rashedi et al. [1] came up with a new metaheuristic algorithm inspired from the Newton's theory of gravity. This theory was proposed by Newton in 1687, which states that states that '*every particle attracts every other particle in this Universe with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between them'*. GSA is considered as a system of all possible solutions, which can be considered as individual agents containing masses which are proportional to their values of the fitness function [16]. In further generations, the agent having heaviest mass exhibits more force on other masses, so it attracts all other masses towards itself and thereby finding a global solution.

Mathematically the problem can be stated as if there exist a system with 'N' number of possible solutions or masses, then their positions are defined as

$$X_i = x_i^1, x_i^2, \ldots, x_i^d, \ldots, x_i^N i = 1, 2, \ldots, N$$

where *'d'* represents the dimension and x_i^d represents the *i*th solution in *d*th dimension.

In our approach we have created the agents by treating a single data sample as an individual data agent. It means that, the number of data samples present in the training data set represent the number of candidate solutions.

Initially the masses and velocity of first agent is set to 0. Then fitness of each agent is calculated at each iteration. The masses and velocity are updated by the masses and velocities of the agent having the optimum fitness value by Eqs. 1, 2 and 3

$$M_{ab} = M_{pb} = M_{bb} = M_b, b = 1, 2, \ldots, N \tag{1}$$

$$M_b(t) = (fitness_b(t) - worst(t))/(best(t) - worst(t)) \tag{2}$$

$$M_b(t) = m_b(t)/\sum_{j=1}^{N} m_c(t) \tag{3}$$

In Eq. 1, the active mass, passive mass and inertia mass are treated as the same for the given problem.

In Eq. 2, the *fitness* $_b$ *(t)* indicates the fitness value of *b*th mass at time 't'. The highest fitness value is assigned as the *worst(t)* and likewise the least value of fitness is assigned to best(t), in a particular epoch for a minimization problem. The reverse is the case for a maximization problem.

The value of gravitational constant G(t) is taken as ($6.671 * 10^{-11}$).

The gravitational strength of each agent and the total force can be calculated by the following formulas.

$$F_{b,c}(t) = G(t) * \frac{(Mpb(t) * Mac(t))}{Rb, c(t) + \in} \left(x_c^d(t) - x_b^d(t)\right) \tag{4}$$

where '$F_{b,c}(t)$' is the gravitational force acting on $mass_b$ on $mass_c$, m_p and m_a are the active and passive masses of object b and object c, $\in$ is a tiny positive constant, $R_{b,c}$ is the Euclidean distance between object 'b' and 'c' and 'G' is the gravitational constant.

The total force exerted on an individual agent is given by

$$F_b(t) = \sum_{b=1,\, b!=c}^{l} rand_b F_{b,c}(t) \tag{5}$$

Here $rand_b$ is a random number in between 0 and 1.

The acceleration can be computed as the following equation:

$$a_b(t) = \frac{F_b(t)}{Mb(t)} \tag{6}$$

The initial set velocity and position can be updated by Eqs. (7) and (8).

$$v_b(t+1) = v_b(t) * rand_b + a_b(t) \tag{7}$$

$$x_b(t+1) = x_b(t) * rand_b + v_b(t+1) \tag{8}$$

In these cases, the next velocity is found by the current velocity summed with the acceleration, and the next position is found by the current position and the new velocity.

Experimental Results

The above algorithm is executed for two types of sets of problems: (1) Function approximation (2) Data Classification [19]. For the function approximation problem, we have used the function f = sin (x) $2.e^{-x}$ in which both the training and test data are generated between the range of [0, 2Π] in the interval of 0.02 in total '315' data are generated, from which the training set is built on 300 data and the testing set includes 15 number of instances. After going through fivefold cross validation, the averaged results are shown in Table 17.1. The worst and best values for minimum and maximum are taken after running among all iterations. The mean squared error has been taken as the evaluation criteria for function approximation and classification accuracy has been taken as performance criteria of classification problem. The function has been approximated by the proposed algorithm to generate the minimum and maximum values for the function and successfully yield the

Table 17.1 Result of function approximation

Method type	Worst for minimum	Best for minimum	Average (5 iterations) for minimum	Worst for maximum	Best for maximum	Average (5 iterations) for maximum
BPNN	0.0377	0.0361	0.0363	0.0699	0.0711	0.0707
GSA	0.0375	0.0360	0.0362	0.0700	0.0713	0.0708
G-BPNN	0.03789	0.0361	0.0363	0.0700	0.0713	0.0708

Table 17.2 Dataset details

Dataset name and type	Number of total samples	Number of predictive attributes	Number of class attributes
(1) Flower data—Iris	150	4	3
(2) Chemical analysis data—Wine	178	13	3
(3) Medical data—pima	768	8	2

results. The results are shown in Table 17.1. There is no significant improvement in case of function approximation, the result behind this is that there is no variation in case of data inputs. For the classification problem, we have taken three benchmark data sets from UCI learning repository [20], whose details are given in Table 17.2.

Table 17.3 represents the result of a random chosen from a set of five overlapping data samples. As in the model verification process, we are using k-fold cross validation, the number of k sets are chosen from the entire data set in such a manner that, each of the data samples for some iterations are trained and as well as tested. Table 17.4 represents the results obtained from all the five dataset samples. For all the sets the algorithm is run for ten iterations and then average of all has been shown in the table. That table contains the result of worst and best performance among all the iterations and run over all the datasets. Table 17.5 compares the result of the same data samples run by BPNN algorithm, simple GSA and the proposed G-BPNN. In BPNN, a continuous sigmoid function has been used as an activation function and the errors are adjusted by gradient descent methodology. For the later, simple GSA is being used to calculate the centroid for each class of the data set. After that the samples are tested by computing the Euclidian distance among testing sample and centroid.

Table 17.3 Result of a random training and testing samples for G-BPNN

Dataset	Ratio of train and test instances	Result of FT—FT	Training accuracy	Testing accuracy
1	75:25	100:100	100.00	95.311
2	80:20	100:100	100.00	98.301
3	75:25	100:100	82.00	76.563

Table 17.4 Result of 5-cross validation on G-BPNN averaged over 10 iterations

Dataset number	Worst testing performance	Best testing performance	Average testing performance
1	86.672	100.00	95.331
2	94.122	100.000	98.302
3	68.833	81.823	76.563

Table 17.5 Comparison with BPNN, GSA and G-BPNN

BPNN with sigmoidal activation function and gradient descent weight updation		
Dataset	Average testing accuracy by fivefold cross validation	Average testing accuracy by FT-FT
1	95.271	100
2	95.302	99.981
3	72.561	72.731
GSA with Centroid to calculation and distance-based method		
1	Not applicable	94.103
2	Not applicable	97.065
3	Not applicable	69.092
Proposed G-BPNN Algorithm		
1	95.331	100.00
2	98.302	99.931
3	76.561	77.371

Conclusion and Future Work

In this chapter, we propose a hybrid G-BPNN algorithm founded on the GSA and BPNN amalgamation. This algorithm collects and synergies the GSA algorithm's sturdy capability in exploring and exploiting the search space for performing a global search and the BPNN's power to do local searching; as a result, it shows superior performance in terms of classifier-convergence and prediction accuracy. Which validates the proposed model and at the same time the results shown in the work proves the hybrid model outperforms than the individual algorithms. The model is being used to solve the function approximation problem and also for pattern classification problem on the three benchmark data sets. The common conclusion can be drawn from this model is that G-BPNN could be applied as a training algorithm for ANN for most of the classification problems. In future research, we might think of using a parallel version of this algorithm which can accommodate big data and unstructured data.

References

1. Esmat Rashedi, Hossein Nezamabidi-pour, Saeid Saryazdi, GSA: A Gravitational Search Algorithm, Journal of Information of Sceinces 179 (2009) 2232–2248.
2. YugaL kumar, G. Sahoo: A Review on Gravitational Search Algorithm and its Applications to Data Clustering & Classification, I.J. Intelligent Systems and Applications (2014), 6,79–93.
3. Andrew Ng. CS229 lecture notes. http://cs229.stanford.edu/notes/cs229-notes5.pdf, (2012).
4. Schutz, B., Gravity from the ground up, Cambridge University Press, (2003).
5. Shanhe Jiang, Yan Wang, Zhicheng Ji: Convergence analysis and performance of improved gravitational search algorithm, Applied Soft Computing 24(2014) 363–384.
6. Hamid Reza Hassanzadeh, Modjtaba Rouhani: A Multi-Objective Gravitational Search Algorithm, Second International Conference on Computational Intelligence, Communication Systems and Networks (2010).
7. V.-H. Nguyen et al. (eds.), *Knowledge and Systems Engineering,* Advances in Intelligent Systems and Computing 326, https://doi.org/10.1007/978-3-319-11680-8_30.
8. Constantin Purcaru, Radu-Emil Precup, Daniel Iercan, Lucian-Ovidiu Fedorovici, Emil M. Petriu, Emil-Ioan Voisan, Multi-Robot GSA- and PSO Based Optimal Path Planning in Static Environments, Proceedings of the 9th International Workshop on Robot Motion and Control, Wasowo Palace, Wasowo Poland, (2013).
9. Rajendra Ku Khadanga, Jitendriya Ku Satapathy, Time delay approach for PSS and SSSC based coordinated controller design using hybrid PSO-GSA algorithm, Electrical Power and Energy Systems 71 (2015) 262–273.
10. Dr. B.L. Shivakumar, S. Alby, A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes, International Conference on Intelligent Computing Applications (2014).
11. Zhilbert Tafa, Nerxhivane Pervetica Bertran Karahoda, An Intelligent System for Diabetes Prediction, 4th Mediterranean Conference on Embedded Computing MECO-2015, Budva Montenegro (2015).
12. Ge Shi, Shihong Zou, Anpeng Huang, Glucose Tracking: A Postprandial Glucose Prediction System for Diabetic Self-Management.
13. Sumi Alice Saji, Balachandran K, Performance Analysis of Training Algorithms of Multilayer Perceptrons in Diabetes Prediction, International Conference on Advances in Computer Engineering and Applications (2015).
14. Zhang, J.R., Zhang, J., Lock, T.M., Lyu, M.R.: A hybrid particle swarm optimization back-propagation algorithm for feedforward neural network training. Applied Mathematics and Computation 185(2), (2007) 1026–1037.
15. Jos´e M. Valls, Ricardo Aler, Oscar Fern´andez, Evolving generalized euclidean distances for training rbnn, Computing and Informatics, Vol. 26, (2007)33–43.
16. Quang Hung Do, A Hybrid Gravitational Search Algorithm and Back-Propagation for Training Feed forward Neural Networks, Springer International Publishing Switzerland (2015).
17. Y.J. Oyang, S.C. Hwang, Y.Y. Ou, C.Y. Clen: Data classification at radial basis function network based on a novel kernel density estimation algorithm, IEEE Transaction on Neural Network 16(1) (2005) 225–236.
18. M.J. Zolghadri, E.G. Masoori, Weighing fuzzy classification rulesusing receive operating characteristic (ORC) analysis, Information Science 177(2007) 2296–2307.
19. Zahra Beheshti, Siti Mariyam Hj. Shamsuddin: A Review of Population-based Meta-Heuristic Algorithm, Int. J. Advance. Soft Comput. Appl., Vol. 5, No. 1, (2013) 1–35.
20. UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science, last accessed on 3.07.2016.

Part III
Communication Engineering

Chapter 18
Probability of Detection Analysis in Fading and Nonfading Scenario Using Cooperative Sensing Technique

Meenakshi Sharma, Prakash Chauhan and Nityananda Sarma

Abstract In the field of cognitive radio network (CRN), spectrum sensing is one of the major function to detect the free spectrum band, for opportunistic use by cognitive radio (CR) users. However, major challenges arise when detection in the presence of licensed (primary) users fail due to various issues like noise uncertainty, multipath fading, shadowing, receiver uncertainty. Use of cooperative spectrum sensing (CSS) technique minimizes the impact of these issues and improve the detection performance by exploiting spatial diversity. The aim of this paper is to analyze the detection performance of individual CR users and their cooperation based performance detected by fusion center (FC) for AWGN and Rayleigh fading channel. For cooperation, a majority-based voting scheme is proposed to collect reliable sensing information from each CR user. Simulation results show that in both channel types, cooperative detection at fusion center (FC) performed better than individual detection by CR user. To achieve higher detection accuracy, the optimal range of thresholds are analyzed by simulation. Further the performance of energy detector is analyzed through ROC curves by changing the signal-to-noise ratio (SNR) value from low to high.

Keywords Cognitive radio · Cooperative spectrum sensing · Probability of detection · Overall detection probability · Overall false detection probability

M. Sharma (✉) · P. Chauhan · N. Sarma
Department of Computer Science and Engineering, Tezpur University, Assam 784028, India
e-mail: amee@tezu.ernet.in

P. Chauhan
e-mail: prakashc@tezu.ernet.in

N. Sarma
e-mail: nitya@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_18

Introduction

With the increasing requirement of wireless applications and rapid development of various radio access technologies, the demand for radio spectrum is drastically boosted everyday [1]. However, a large portion of the assigned spectrum is actually remained unused with respect to time and space. To solve such problem of radio spectrum scarcity and underutilization, cognitive radio technology has been proposed which can solve the spectrum scarcity problem by opportunist utilization of unoccupied spectrum band without harmful interference to primary users (PU) [2]. The process of finding unoccupied spectrum is achieved by sensing the radio spectrum continuously without interfering PUs communication. However, continuous sensing by secondary user (SU) faces problems like fading, shadowing, receiver uncertainty which can degrade the sensing performance adversely. To overcome the problems, cooperative spectrum sensing technique has emerged as a solution where sensing information from multiple SUs are incorporated for PU detection and exploits spatial diversity during the sensing process [1, 3].

Two important matrices that determines the performance of spectrum sensing are: probability of detection (P_d) which indicates the likelihood of a SU announcing that a PU is available when a PU is really present [3], and the probability of false alarm (P_f), i.e., the likelihood of a SU proclaiming that a PU is available when a PU is actually absent [3]. The probability of missed detection (P_m) is denoted as ($1 - P_d$), and it can cause a SU to interfere with a PU by wrongly identifying it [4]. From the SU perspective, higher the P_d better will be the primary user protection and less interference. On the other hand, lower the P_f better will be the opportunity for unlicensed access and high data rate (throughput) [4]. In a perfect scenario, a high Pd is required, subject to the requirement of Pf.

In this paper, we have introduced a centralized CSS technique to analyze the detection behavior of individual SUs and their collaboration at FC in two channel models. The uniqueness of our work in this paper is analyzing effectiveness of energy detector by tuning the collected samples over some range of P_f values at a defined time period and send reliable sensing information to FC by discarding unreliable ones by SUs. This technique helps to establish the optimal range of P_f and threshold of the energy detector to calculate the high P_d value among SUs. The overall detection performance of FC is examined with the help of different fusion rules and investigates the cooperative result with individual user's detection.

The rest of the paper is organized as follows. In section "System Model and Assumptions" the system model with various assumptions are discussed. Section "Implementation Details" consists of implementation details with the proposed algorithm and used parameters list. Simulation results are discussed and analyzed in section "Results and Performance Analysis" and finally the paper is concluded with future work in section "Conclusion and Future Aspect".

System Model and Assumptions

System Model

The system model of the proposed CSS scheme as shown in Fig. 18.1a is considered with n nodes and following assumptions: (1) All SUs are plotted randomly in the network. (2) An initial amount of energy is eventually distributed among them and each SU can sense only one channel at a time. (3) Total time is divided in time slots and each SU is allowed to sense the channel at each time slot. (4) The SUs use a common dedicated control channel in sequential/TDMA fashion for reporting to FC. (5) FC is assumed to be a centralized base station. In our work, we consider a situation where prior learning of the PU is inaccessible. This model consists of two phases.

Decision-Forward Phase

According to the signal detection framework, on basis of test hypothesis (T_s), the sensing signal energy over the collected sample (L) is compared with some threshold values (obtained from range of P_f over tuning) as shown in Eqs. 18.1 and 18.2 [1–5].

$$T_s = \frac{1}{L} \sum_{t=1}^{L} |x(t)|^2 \tag{18.1}$$

where x(t) is the primary user signal.

$$Di = \begin{cases} 1, & Ts \leq \lambda \\ 0, & Ts > \lambda \end{cases} \tag{18.2}$$

Each SU makes a binary decision Di (Di = 1 stands for the presence of the PU as detected by the *i*th SU, and Di = 0 stands for the absence of the PU as detected by the *i*th SU). The collected samples are simulated for given P_f values and for each P_f every SU stores their binary decision in a local list (data structure) that is associated with them as shown in Fig. 18.1b. Each SU can access their own list only. Then each SU applies majority voting technique to select the binary decision Di with majority and send their individual final decision (1/0) to the FC. The idea of introducing the list data structure helps to select those decisions that appear maximum time during observation for each SU. It refrains SUs from sending unreliable decisions to FC and helps to transmit reliable decisions only. Though this scheme increases the sensing overhead while computing local decisions for all P_f values, it increases the accuracy of the energy detector for sending correct sensing decisions to FC.

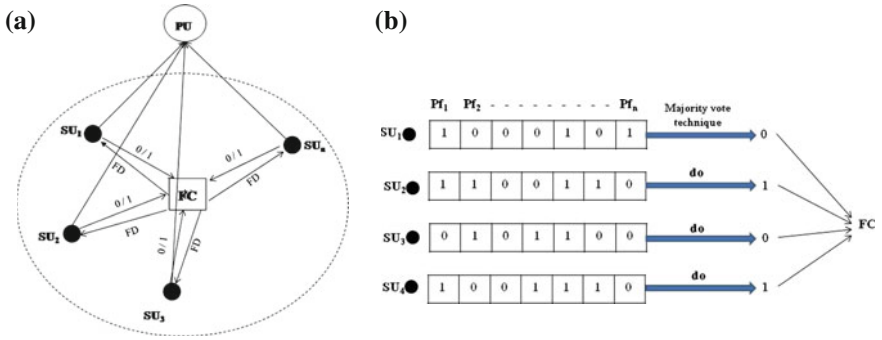


Fig. 18.1 **a** System model structure, **b** Local list structure of each SU

Data Fusion Phase

While accepting the binary decisions from each SU, FC settles on final decision (FD) (either $H0$ or $H1$) on the premise of the no. of SUs guaranteeing the presence of the PU signal. Let d1 be the number of SUs claiming the presence of PU. Therefore, the final decision strategy is given as in Eq. 18.3 [5, 6]:

$$FD = \begin{cases} H0, & d1 < k \\ H1, & d1 \geq k \end{cases} \tag{18.3}$$

where $k = 1, k = N, k > N/2$ and $1 < k < N$ are corresponding to OR rule, AND rule, majority rule, and k-out-of-N rule, respectively, with N being the total no. of SUs. In this chapter, we determine k for k-out-of-N rule using the following formula:

$$k = ceil\left(\frac{N}{2}\right) + 1 \tag{18.4}$$

Overall Detection Probability by the Fusion Center

Let D be the quantity of 1's (indicating the presence of PU signal) in the n local decisions reported by the N no. of SUs. These SUs are autonomous and subsequently the distribution is binomial as follows [6, 7]:

$$\mathrm{P(D = x)} = \binom{n}{x} \mathrm{P_1}^{x}(1 - \mathrm{P_1})^{n-x} \tag{18.5}$$

where P_1 is the probability of reporting 1 by the SUs in the local decisions sequence sent to the FC and x denotes the value of k which is obtained from different fusion rules. Therefore, the overall detection probability P_{od} by FC is calculated as follows

[7]:

$$P_{od} = \sum_{x=k}^{n} \binom{n}{x} P_1{}^x (1 - P_1)^{n-x} \tag{18.6}$$

Overall False Detection Probability by the Fusion Center

Let E be the number of 0s (indicating the absence of PU signal) in the n local decisions reported by the N SUs. Therefore, binomial distribution for the SUs is as follows [6, 7]:

$$P(E = x) = \binom{n}{x} P_0{}^x (1 - P_0)^{n-x} \tag{18.7}$$

where P_0 is the probability of reporting 0 addressing PU signal is absent and we can write $P_0 = 1 - P_f$ [7]. Therefore, the overall true negative probability P_{otn}, i.e., primary signal is actually absent and is calculated as follows [7]:

$$P_{otn} = \sum_{x=k}^{n} \binom{n}{x} P_0{}^x (1 - P_0)^{n-x} \tag{18.8}$$

And overall false detection probability (P_{od}) is calculated by [7]:

$$P_{ofd} = 1 - P_{otn} \tag{18.9}$$

Implementation Details

The algorithm for the proposed model is discussed below. Symbols and notations used in the algorithm are given in Table 18.1. The simulation work of the proposed algorithm is carried by MATLAB 7 (R2013a) in a 64-bit PC with a core i5 processor and 4 GB RAM.

Results and Performance Analysis

Parameters

We consider that SUs participating in CSS are located in a grid topology of a 200 m × 300 m area. The parameters used in the simulation are as follows: No. of SUs are 10, time–bandwidth product (m) is 5, sensing duration (τ) is 100 μs, total no. of samples are 1000, SNR varies from −2 to −25 db (to receive various signals of different

Algorithm 1: Centralized Cooperative Spectrum Sensing Algorithm

input : P_f, λ, SNR, k
output: P_d, P_m

```
for P_f do
    for λ do
        for user s do
            G = list(different values of P_f)
            Sense the spectrum ⟶ T_s
            if T_s > λ then
                Store 1 to G
            else
                Store 0 to G
            end
        end
    end
end
for user s do
    Calculate the majority vote in G and send to FC
end
for user s do
    Calculate k
    Calculate final decision at FC using various fusion rules (AND, OR, majority,
    k-out-of-N)
end
```

Table 18.1 Notation used and description

Notations	Descriptions
λ	Threshold for the energy detector
G	List to store each decision of individual SU
T_s	Test hypothesis
k	Intermediate integer variable within range [1 to N], where N = total no of SU
H_0	Absence of PU as hypothesis 0
H_1	Presence of PU as hypothesis 1

strength), P_f range from 0.01 to 1 (consider a range from minimum to maximum value).

Figure 18.2a plots individual SU's P_d w.r.t increased P_f at fixed SNR $= -10$ db (since at -10 db low signal value is received). P_f value gradually increases from 0.1 to 1 and observes the detection performance of each SU. As P_f increases the P_d of each SU too increases in both the scenarios, but finally in Rayleigh fading the P_d reaches up to 0.5 whereas in AWGN, it improves to 0.9. In Fig. 18.2b individual SUs P_d with increased thresholds is analyzed. For the lower threshold, the P_d of SUs is found to be better in AWGN, but while threshold value increases, P_d degrades in AWGN and after reaching 1.2 it degrades faster than in Rayleigh fading. Therefore,

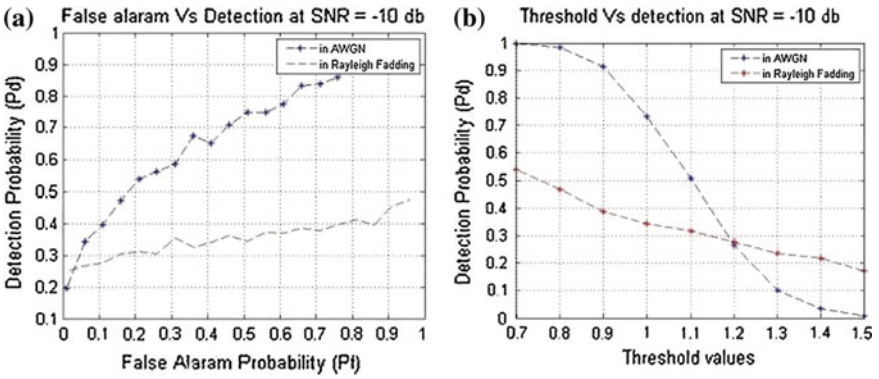


Fig. 18.2 **a** P_d versus P_f curves, **b** P_f versus Threshold curves (in AWGN and Rayleigh fading)

the simulation results that improved P_d can be achieved by keeping the threshold value of the energy detector less than 1.2 in both cases.

Next, in Fig. 18.3a, b the overall detection probability (P_{od}) at FC using various fusion rules (OR, AND, Majority, K-out-of-N) are depicted. In FC, the P_1 (probability of reporting 1 by SUs) is calculated from the decision sequences sent by each SU for each threshold value. For OR rule, the *ROC* curves increase rapidly *for both the channels* as P_1 of SU gradually increases from 0 to 1. For AND rule the curve remain almost zero (up to $P_1 = 0.6$ and $P_1 = 0.7$ in AWGN and in Rayleigh respectively) but as P_1 increments to 0.7, all SUs start sending local decision 1 and the P_{od} increases fastly. The obtained curves for k-out-of-N and majority rule falls between the curves of the OR and the AND rule.

In Fig. 18.3c the overall false detection rate (P_{ofd}) by the FC is depicted for both the channels. In case of P_{od} as the threshold decreases, the value of P_1 in each decision sequence improves and the P_{od} rate also gets improves. On the other hand, with a decrease in threshold the value of P_0 as well as true negative rate decreases which eventually increases the rate of P_{ofd}. In AWGN channel, the rate of overall false detection is found to be high (almost 1) than Rayleigh fading (nearly 0.5).

The ROC curves for different SNR are described in Fig. 18.4. In both the channels, for lower values of SNR, the missed detection probability increases and as the SNR value increases, missed detection decreases. Although, for AWGN channel, the ROC curves found to be better than in Rayleigh fading as less missed detection occurs at varying SNR.

Conclusion and Future Aspect

In this chapter, a centralized cooperative spectrum sensing technique is proposed and its detection performance has been analyzed in terms of P_d, P_{od}, P_{ofd} w.r.t. different

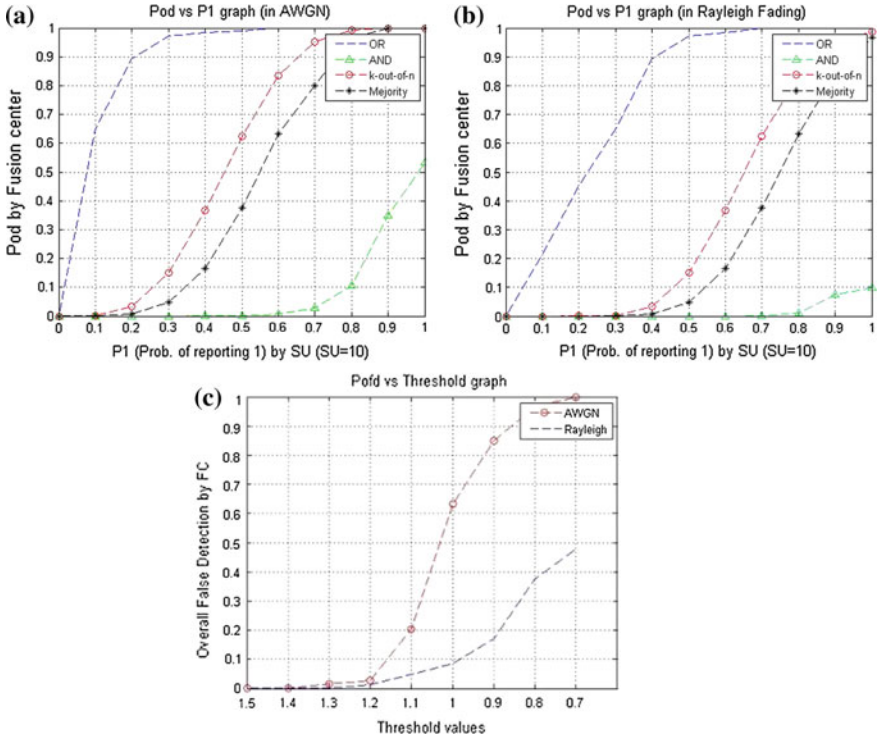


Fig. 18.3 **a** P_{od} versus P_1 curves in AWGN, **b** P_{od} versus P_1 curves in Rayleigh fading, **c** P_{ofd} versus threshold curves in AWGN and Rayleigh fading

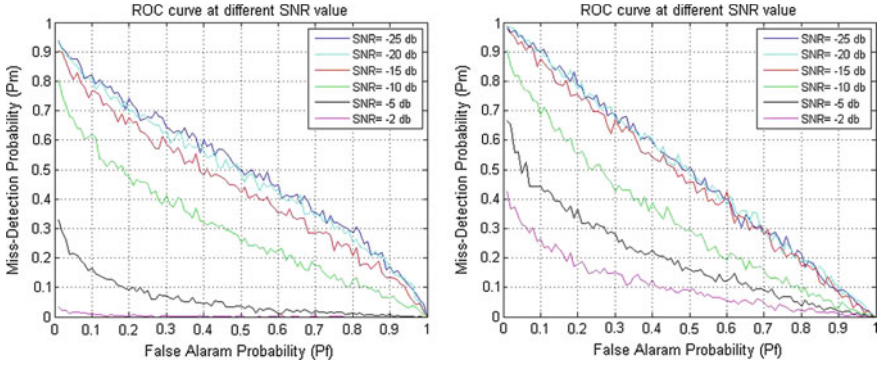


Fig. 18.4 ROC curves in AWGN and Rayleigh fading

thresholds for two channel models. Simulation results show that for both the models the P_d of individual SUs get enhanced at low range of threshold (less than 1.2) and degrades with the increase in threshold value. Along side, analyzing P_{od} at FC, except OR rule, the k-out-of-N rule enhances the performance in both types of channel. Finally, with the ROC curves it is observed that in low SNR the performance of energy detector degrades in Rayleigh fading and produces poor sensing results. Hence, based on the simulation results it can be stated that the overall performance of the proposed CSS technique is found to be better in AWGN channel.

The presence of Nakagami, Rician fading, shadowing in channels, it would be interesting to analyze the performance statistics of the proposed cooperative model and left for future work.

Acknowledgements This work has been supported by MHRD sponsored Center of Excellence under FAST MLRBDA project for the period of 2014–2019.

References

1. Akyildiz, I. F., Lee. W. Y., Vuran, M. C., and Mohanty, S. (2006): NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Computer Networks, 50(13), (2127–2159).
2. Akyildiz I. F., F. L. Brandon and Balakrishnan R. (2010): Cooperative spectrum sensing in cognitive radio networks: A survey, Physical Communication, Elsevier Publication, Vol. 4, (40–62).
3. A. Ghasemi and E. S. Sousa (2005) : Collaborative spectrum sensing for opportunistic access in fading environments. First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, USA, (131–136).
4. A. Ghasemi and E. S. Sousa (2008): Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs, IEEE Communications Magazine, Vol. 46, (32–39).
5. Omar Altrad and Sami Muhaidat (2013): A new mathematical analysis of the probability of detection in cognitive radio over fading channels. EURASIP Journal on Wireless Communications and Networking, https://doi.org/10.1186/1687-1499-2013-159, (1–11).
6. Srinivas Nallagonda, Sanjay Dhar Roy, Sumit Kundu (2013): Performance Evaluation of Cooperative Spectrum Sensing Scheme with Censoring of Cognitive Radios in Rayleigh Fading Channel, Wireless Personal Communications, Vol. 70, (1409–1424).
7. N. Marchang, R. Rajkumari, S. B. Brahmachary and A. Taggu (2015): Dynamic decision rule for cooperative spectrum sensing in Cognitive Radio Networks IEEE International Conference on Electrical, Computer and Communication Technologies, Coimbatore, https://doi.org/10.1109/ICECCT.2015.7226198, (1–5).

Chapter 19
Selection Diversity-Based Energy Detection over Weibull Fading Channels

Niharika Barman, Parag Moni Patangia and Rupaban Subadar

Abstract The most difficult task in the cognitive radio scheme is the detection of spectrum availability. The system should continuously monitor the channel so that the primary user signal is not interfered. Out of the different spectrum sensing techniques, energy detection is considered as the most appropriate for practical applications because its complexity is less, computational cost is low and can be easily implemented. This technique can be performed without having any information of the primary user signals thus making it more suitable for any kind of communication environment. The detection of primary users depends on whether energy detector's output is greater or less than a predefined threshold value. This paper deals with the study and analysis of selection diversity-based spectrum monitoring over Weibull fading channels based on energy. The system is also replicated with Monte Carlo simulation for verification of the results.

Keywords Weibull fading channel · Probability of detection · Energy spectrum sensing · Meijer G-function · Selection combining

Introduction

Wireless communication has been experiencing exponential growth around the world due to its immense usefulness both in licensed and unlicensed applications. These lead to the scarcity of spectrum particularly in the unlicensed band thereby making it extremely difficult to accommodate new services in this band. This issue in

N. Barman (✉) · P. M. Patangia · R. Subadar
Department of Electronics and Communication Engineering, North-Eastern Hill University, Shillong, India
e-mail: neyhabarman@gmail.com

P. M. Patangia
e-mail: mail.parag19@gmail.com

R. Subadar
e-mail: rupaban.subadar@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_19

wireless services is surmounted by means of cognitive radio (CR). CR makes use of the licensed band in absence of the primary user (PU) [1]. This technique opens a new window for new wireless applications. The crucial task of a CR is to sense the availability of spectrum. This is done through different detection methods out of which energy detection is studied here. Energy-based detection method is the best suitable among all spectrum sensing methods based on implementation simplicity. In this method, the received energy of the signal is evaluated and compared with threshold decision of presence or absence of a PU is taken.

The unknown deterministic signal was detected first by Urkowitz in [2] in a white Gaussian noise channel, where the author has exploited the central chi-square distribution to get an expression of Probability of detection (P_d) and Probability of false alarm (P_{fa}). A few decades later, Kostylev [3] derived the same expression of detection probability for signals having random amplitude by the method of decision statistics. In [4], P_d over different fading channels is derived and the system performances are investigated with diversity schemes employed. Among the different diversity schemes, selection combining (SC) diversity is the simplest scheme which is used to reduce the effects of fading, it selects the signal with highest SNR from the available diversity path. An infinite series solution of P_d, P_{fa} is derived in [5] for selection diversity over Nakagami-m fading. In [6], MGF and PDF approach are exploited to derive the P_d over Nakagami-m fading by considering both diversity and without diversity. The authors have derived the expressions considering SC diversity case when $L = 2$ and $L \geq 2$. In [7], P_d has been presented considering Weibull fading. Derivation of expression for miss-detection probability is presented in [8] over Rayleigh and Nakagami-m fading. The expression of P_d is given for K and $K - G$ fading in closed form considering both no-diversity case and including both MRC and SC [9].

From the literature study, we came to a conclusion that the fading environment deeply influences the performance of energy detector and its detection probability can be enhanced and by employing diversity scheme along with the detector. The performance of diversity detector based on the energy is not known for Weibull fading. This inspired us to analyze the probability of detection of energy sensing based diversity detector over Weibull fading channels.

The structure of the rest of the paper is in the following manner. In section "System and Channel Model", system and channel model have been described in brief. Derivation of average probability of detection has been presented in section "Detection Probability". Analyses have been carried out in sections "Average Detection of Probability" and "Conclusions" consists of the conclusion.

System and Channel Model

The figure of a diversity-based energy detector with main blocks are shown in Fig. 19.1. It is similar to the ordinary energy detector except a diversity combining block is added in front to enhance the quality of signal from the primary user.

We have considered a SC diversity in this case, which is simple to implement and have low computational complexity. Hence, it will not increase the time delay of the detector much. We are considering the diversity-based detector is working in a slow and flat Weibull fading environment. Expression of the received signal by the lth antenna, over one- bit duration T_b can be given as

$$y_l(t) = \alpha_l e^{j\phi_l} x(t) + n_l(t), \tag{19.1}$$

where $x(t)$ represents the transmitted signal containing energy E_b, $n_l(t)$ is the complex Gaussian noise with power spectral density $2N_0$ with zero mean and ϕ_l denotes the instantaneous phase. The RV α_l represents the amplitude with Weibull distribution and the PDF is [10]

$$P_{\alpha_l}(\alpha_l) = c\left(\frac{\Gamma\left(1+\frac{2}{c}\right)}{\Omega}\right)^{c/2} \alpha_l^{c-1} e^{-\left(\frac{\alpha_l^2}{\Omega}\Gamma\left(1+\frac{2}{c}\right)\right)^{c/2}}, \alpha_l > 0. \tag{19.2}$$

where $\Omega_l = E[\alpha_l^2]$, and $c \geq 0$ is the Weibull fading parameter. Weibull distribution named after the Swedish Mathematician, Walodi Weibull, in 1951 is a continuous probability distribution. Weibull distribution can be employed in both indoor and outdoor scenarios including the nonhomogeneous environment.

In Fig. 19.1 the output of the integrator represents the energy of the received signal, which has been given as the input to the decision device. Decision device based on the comparing the energy with a threshold, finds out the presence or absence of

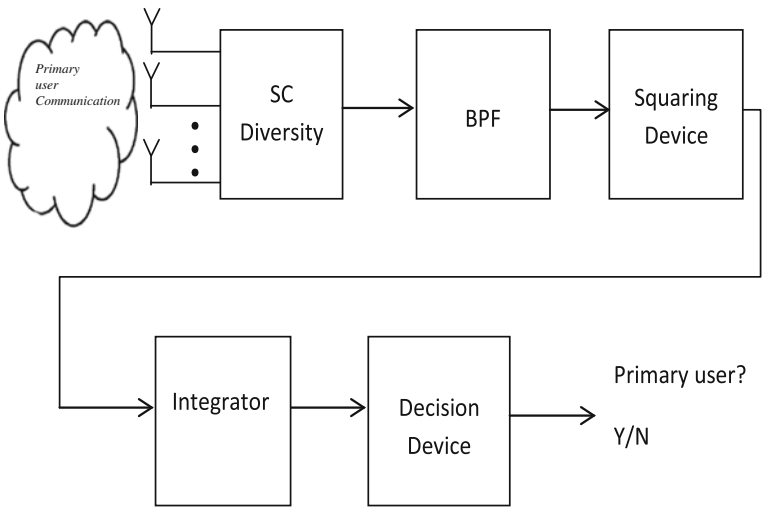


Fig. 19.1 Energy detector-based spectrum sensing

primary user. Mathematical expression of the decision device can be given as follows:

$$y(t) = \left\{ \begin{array}{c} n(t) : \quad Primary\ user\ absent \\ s(t) + n(t) : Primary\ user\ present \end{array} \right\} \tag{19.3}$$

where y(t) is the received signal, s(t) is primary user's transmitted signal, and n(t) is additive white Gaussian noise.

In the SC receiver, the signal with the highest SNR is processed for detection. The mathematical expression of the selected signal is

$$\gamma = \text{Maximum}(\gamma_1, \gamma_2, \ldots, \gamma_L), \tag{19.4}$$

where $\gamma_l = \frac{E_b}{N_0}\alpha_l^2$ is the SNR of lth branch of the SC receiver. The average of which can be obtained by taking statistical average, and for convenience of presentation we assume that the average SNRs are equal, i.e., $E[\gamma_l] = \bar{\gamma}, \forall l$.

Detection Probability

To obtain average probability of detection expression in a SC diversity-based energy detector, we need the information of the SNR pdf at the output of the SC receiver. For Weibull fading channels this PDF is not known. Hence, we have divided the work into two different parts. In the first part, we explained the steps that are used to derive the SNR PDF of SC receiver. In the second part, we derive the expression for the probability of detection.

PDF of SNR for SC Receiver

The SNR PDF for Weibull distribution can be derived from (19.2) by applying the rule of transformation random variable and it is given as [10]

$$f_{\gamma_l}(\gamma_l) = \frac{c}{2}\left(\frac{\Gamma\left(1+\frac{2}{c}\right)}{\bar{\gamma}}\right)^{c/2} \gamma_l^{c/2-1} e^{-\left(\frac{\gamma_l}{\bar{\gamma}}\Gamma\left(1+\frac{2}{c}\right)\right)^{c/2}}, \gamma_l > 0 \tag{19.5}$$

CDF of which is [9]

$$F_{\gamma_l}(\gamma_l) = 1 - \exp\left(-\frac{\Gamma_l^{c/2}}{\beta}\right), \tag{19.6}$$

For L independent and identical fading signals, the joint CDF of all input SNRs can be obtained by the product of CDF of all the branches. CDF for output SNR of SC receiver can be obtained (by substituting $\Gamma_L = \gamma_{sc} \forall l$) as

$$F_{\gamma_{sc}}(\gamma_{sc}) = \left[1 - \exp\left(-\frac{\gamma_{sc}^{c/2}}{\beta}\right)\right]^L, \tag{19.7}$$

Taking the differentiation of (19.7) and then expanding the expression by applying Binomial theorem, the final expression of output SNR for SC receiver with Weibull fading can be expressed as

$$f_{\gamma_{sc}}(\gamma_{sc}) = \sum_{k=0}^{L-1} (-1)^k \binom{L-1}{k} \frac{Lc}{2\beta} \times \gamma_{sc}^{\frac{c}{2}-1} e^{\left(-\frac{(1+k)\gamma_{sc}^{\frac{c}{2}}}{\beta}\right)} \tag{19.8}$$

where $\binom{a}{b}$ is known as the binomial coefficient [11].

Average Detection of Probability

The general expression of P_d for energy detection based on selection diversity working over fading environment can be given as [12]

$$\overline{P_d} = \int_0^\infty Q_u(\sqrt{2\gamma}, \sqrt{\lambda}) f_{\gamma_{sc}}(\gamma_{sc}) d\gamma_{sc} \tag{19.9}$$

where $f_{\gamma_{sc}}$ is the density function of output SNR of the SC which is given in (19.8) and putting this value in (19.9) the expression of detection probability can be given as

$$\begin{aligned} \overline{P_d} = \int_0^\infty Q_u\left(\sqrt{2\gamma_{sc}}, \sqrt{\lambda}\right) \sum_{k=0}^{L-1} (-1)^k \binom{L-1}{k} \frac{Lc}{2\beta} \\ \times \gamma_{sc}^{\frac{c}{2}-1} \exp\left(-\frac{(1+k)\gamma_{sc}^{\frac{c}{2}}}{\beta}\right) d\gamma_{sc} \end{aligned} \tag{19.10}$$

The solution of involved integral in (19.10) is not known, however the same can be solved using the infinite series expansion of Marcum Q function as given in [13]

$$Q_v(\sqrt{2a}, \sqrt{2b}) = e^{-a} \sum_{n\geq 0} \frac{a^n}{n!} \frac{\Gamma(v+n, b)}{\Gamma(v+n)} \tag{19.11}$$

Rewriting the expression of probability of detection, we obtain

$$\begin{aligned} \overline{P_d} = & \sum_{k=0}^{L-1} (-1)^k \sum_{n\geq 0} \binom{L-1}{k} \frac{Lc}{2\beta} \frac{\Gamma\left(1+n, \frac{\lambda}{2}\right)}{n!^2} \\ & \times \int_0^{\infty} \gamma_{sc}^{n+c/2-1} e^{-\gamma_{sc} - \frac{1+k}{\beta}\gamma_{sc}^{c/2}} d\gamma_{sc} \end{aligned} \tag{19.12}$$

Using [14, (4)], we can solve the integral and the final expression is obtained as

$$\begin{aligned} \overline{P_d} = & \sum_{k=0}^{L-1} (-1)^k \sum_{n\geq 0} \binom{L-1}{k} \frac{Lc}{2\beta} \frac{\Gamma\left(1+n, \frac{\lambda}{2}\right)}{n!^2} \\ & \times (2\pi)^{\frac{1-c/2}{2}} \left(\frac{c}{2}\right)^{n+c/2-1} \\ & \times G_{1,c/2}^{c/2,1}\left(\frac{1}{\frac{(1+k)}{\beta}(c/2)^{c/2}} \Big|_{\frac{n+c/2}{c/2}, \frac{n+c/2+1}{c/2}, \ldots\ldots \frac{n+c/2+c/2-1}{c/2}}^{1} \right) \end{aligned} \tag{19.13}$$

Numerical Results and Discussion

Numerical evaluation has been carried out for the derived expression for different values of diversity order, average SNR and the fading parameter. In Figs. 19.2 and 19.3 P_d versus $\bar{\gamma}$ have been plotted considering various values of L and c. From Fig. 19.2 it is observed that the detection probability improves with increase in diversity order as expected. It can also be observed from the figure that for a detection probability of above 0.99 dual diversity scheme needed an average branch SNR of 20 dB but for $L = 6$, for the same probability of detection average branch SNR reduces to less than 10 dB. Figure 19.3 shows the detection probability for different values of c. As expected, with increase in c probability of detection badly effected due to severe fading in the channels. Figure 19.4 shows the comparison of dual diversity and

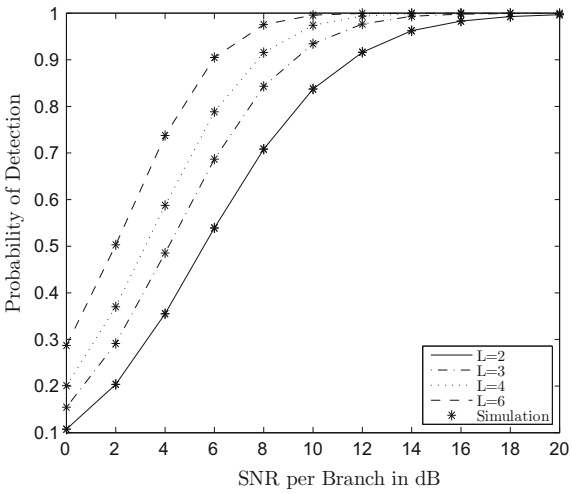


Fig. 19.2 P_d for different L

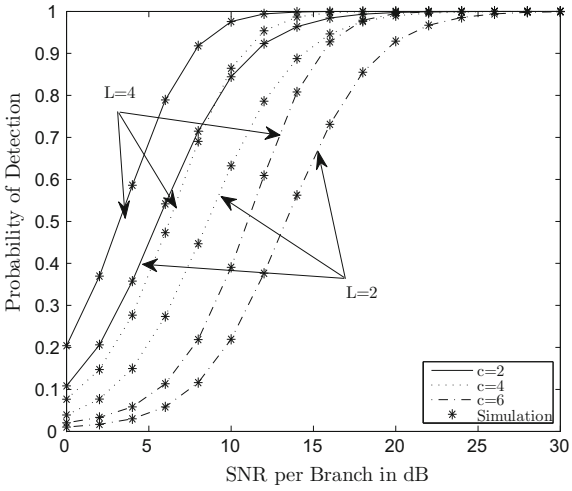


Fig. 19.3 Probability of detection for different L and c

no- diversity energy detector. It is clear from Fig. 19.4 that energy detector without diversity is not practically possible for Weibull fading environment, since it requires very high average branch SNR of the transmitted primary signal. A close agreement of the results have been observed with the Monte Carlo simulation.

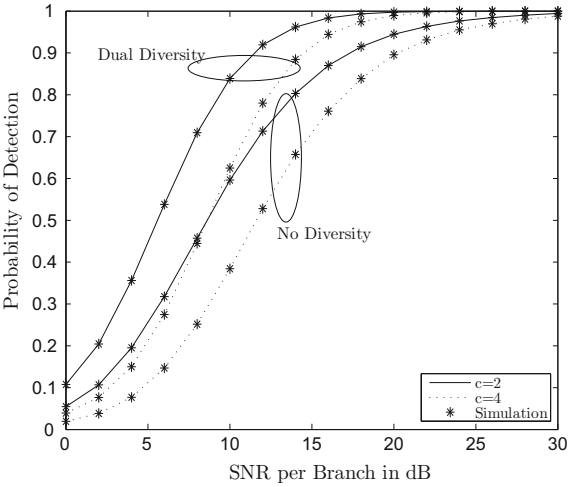


Fig. 19.4 Comparison of energy detector with and without diversity

Conclusions

In this paper, we present an analytical expression for the average probability of detection in Weibull fading channels employing SC diversity-based technique. The obtained expressions are found in terms of Meiger-G functions. The results have been numerically evaluated and analyzed for different scenarios. Also, a comparison of energy detector without diversity has been presented. Simulated results are in close agreement with the obtained results.

References

1. Federal Communications Commission, "Notice of proposed rule making and order: facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies," ET Docket No. 03-108, Dec. 2003.
2. H. Urkowitz, "Energy detection of unknown deterministic signals," *in Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, April 1967.
3. V. I. Kostylev, "Energy detection of a signal with random amplitude," *Communications, 2002. ICC 2002. IEEE International Conference on*, 2002, pp. 1606–1610 vol. 3.
4. F. F. Digham, M. S. Alouini, M. K. Simon, "On the energy detection of unknown signals over fading channels," *IEEE Trans. Commun.* vol. 55, no. 1, pp. 21–24, Jan. 2007.
5. S. P. Herath, N. Rajatheva and C. Tellambura, "On the energy detection of unknown deterministic signal over Nakagami channelswith selection combining," *Electrical and Computer Engineering, 2009. CCECE'09. Canadian Conference on*, St. John's, NL, 2009, pp. 745–749.
6. S. P. Herath, N. Rajatheva, C. Tellambura, "Energy detection of unknown signals in fading and diversity reception," *IEEE Trans. Commun.*, vol. 59, no. 9, pp. 2443–2453, 2011.
7. P. C. Sofotasios, M. K. Fikadu, Khuong Ho-Van and M. Valkama, "Energy detection sensing of unknown signals over Weibull fading channels," *Advanced Technologies for Communications (ATC), 2013 International Conference on*, Ho Chi Minh City, 2013, pp. 414–419.

8. S. Atapattu, C. Tellambura and H. Jiang, "Spectrum Sensing via Energy Detector in Low SNR," *Communications (ICC), 2011 IEEE International Conference on,* Kyoto, 2011, pp. 1–5.
9. S. Atapattu, C. Tellambura and H. Jiang, "Performance of an Energy Detector over Channels with Both Multipath Fading and Shadowing," *in IEEE Transactions on Wireless Communications*, vol. 9, no. 12, pp. 3662–3670, 2010.
10. M. K. Simon and M. -S. Alouini, *Digital Communications Over Fading Channels*, 2nd ed. New York: Wiley, 2005.
11. I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York: Academic, 2007.
12. Eleftherios Chatziantoniou, B. Allen, Vladan Velisavljevic, "Performance analysis of energy detection over hyper-Rayleigh fading channels" *The Journal of Engineering* , January 2014.
13. Andrs S, Baricz, Sun Y, The generalized Marcum Q-function: an orthogonal polynomial approach. *Acta Universitatis Sapientiae Mathematica.* 3(1), 6076 (2011).
14. Julian Cheng, Chintha Tellambura and N. C. Beaulieu, "Performance analysis of digital modulations on Weibull fading channels," *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE* 58th, 2003, pp. 236–240 Vol. 1.

Chapter 20
Design of Dense Wavelength Division Multiplexing System Using DQPSK Modulation Format

Dhiman Kakati and Subhash C. Arya

Abstract In this chapter, a Dense Wavelength Division Multiplexed (DWDM) system is presented for a transmission distance of 1000 km. A maximum data rate of 160 Gbps is achieved using Differential Quaternary Phase-Shift Keying (DQPSK) modulation format with direct detection at the receiver. The fiber attenuation and dispersion are compensated using Erbium-Doped Fiber Amplifier (EDFA) and Dispersion Compensating Fiber (DCF) respectively. The 16 × 10 Gbps system is evaluated in terms of Bit Error Rate (BER $\leq 10^{-9}$) and Quality factor (Q-factor $\geq$ 6) based on the simulation platform. The proposed design shall find application in Passive Optical Network (PON) architecture and Optical Add-Drop Multiplexing (OADM) transmission system.

Keywords DWDM · PON · BER · OADM

Introduction

The exponentially growing demand for higher bandwidth has led to the development of various network architectures to achieve spectral efficient transmission. There are various multiplexing techniques such as Electronic Time-Division Multiplexing (ETDM), Optical Time-Division Multiplexing (OTDM), Wavelength Division Multiplexing (WDM), Spatial Division Multiplexing (SDM), Polarization Division Multiplexing (PDM) and Orthogonal Frequency-Division Multiplexing (OFDM) to exploit the bandwidth available in the optical domain and each of them has their own merits and demerits [1, 2]. DWDM was developed to enhance the data rate by implementing the spectral components (i.e., wavelengths) to carry the

D. Kakati (✉) · S. C. Arya
Lightwave Communication Systems Laboratory, Department of Electronics and Communication Engineering, NEHU, Shillong 793022, India
e-mail: dhiman.kakati@gmail.com

S. C. Arya
e-mail: aryasubh@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_20

bit streams. In the C-band (1525–1565 nm) by implementing capability and cost-effectiveness of EDFA, it is possible to transmit data up to thousands of km at an acceptable BER level. The designed DWDM system can also be modified with Optical Add-Drop Multiplexer (OADM) and one of the data bearing carriers can be extracted at any position from the link. Recently, an offset Quadrature Amplitude Modulation (QAM) WDM super-channel is reported and achieved a Spectral Efficiency (SE) of 7.44 bits/s/Hz for a transmission distance of 968 km as in [3]. Multi-gigabit WDM system implementing chirped OFDM has been experimentally verified for Visible Light Communication (VLC) and reported a 1.05 Gbps for 30 cm in [4]. Coherent PON system transmitting 12 × 10 Gbps bidirectional data was experimentally evaluated over hybrid 80 km SSMF [5]. Cost-effective transmission at a higher transmission rate (i.e., above 100 Gbps) and maintenance of the Base Stations (BSs) is a challenging issue [6, 7]. In the present work, a 16-channel DWDM system is implemented using DQPSK modulation format with direct detection and is evaluated for faster and robust transmission systems.

Design Schematic of the DWDM Transmission Link

Figure 20.1 shows the setup of the proposed system based on OptiSystem v.13 simulation platform.

The lightwave emitted from the Continuous Wave (CW) laser array has uniform launch power of −10 dBm and frequency separation of 100 GHz (i.e., ITU standard G.694.1 [8]) and are fed to the respective DQPSK transmitter blocks. In each

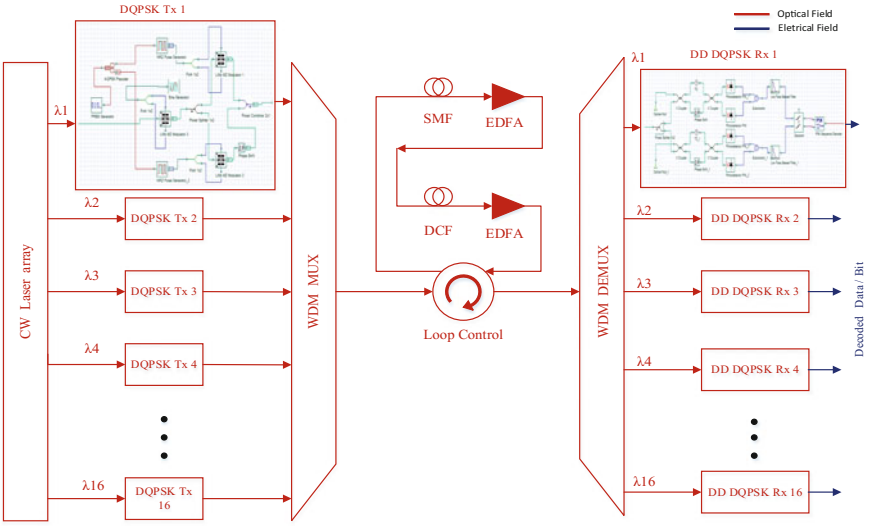


Fig. 20.1 Design schematic of the proposed DWDM transmission link

DQPSK transmitter, the Pseudorandom Bit Sequence (PRBS) generator data is fed to Lithium Niobate ($LiNbO_3$) Mach Zehnder Modulator (LN-MZM) and the two optical signals are modulated accordingly. The lower LN-MZM output signal is 90° phase shifted by the phase shifter and the required DQPSK signal is generated by combining these optical signals. The function of WDM-MUX is equivalent to an ideal adder and there is no power splitting or filtering. The fiber dispersion is compensated by the DCF of 5 km length having a negative dispersion coefficient of −80 ps/nm-km and the fiber attenuation is compensated by the EDFA of 5 dB and 2.5 dB at regular intervals of 25 km SMF and 5 km DCF. The received optical signal is split, filtered, and then fed to the related Direct Detection (DD) DQPSK receivers. Three factors (i.e., bandwidth, ripple and depth of the optical filter) of the MUX and DEMUX decides the crosstalk in the system and also the amount of power transferred from the neighbouring channels. The mathematical expression of the electrical and optical signals at various positions of the system can be found out from [1].

Simulation Results and Discussion

Figure 20.2 shows the constellation and the eye diagrams for 500 and 1000 km fiber lengths. The two constellation diagrams, i.e., Fig. 20.2a and c show the signal points at the four quadrants and noise at the center.

A clear eye opening, Q-factor ≥ 6, BER $\leq 10^{-9}$ and timing jitter $\leq$ (0.15 $\times$ bit period), i.e., 15 ps is observed for both the cases as shown in Fig. 20.2b and d for the two considered transmission lengths.

Figure 20.3 shows the change in Q-factor and BER for various frequencies after transmission over 1000 km SSMF. It is observed that the BER and Q-factor are not same due to the fact that there is power transfer from one channel to its neighboring channels while transmission and the peak are noticed at 193.814 THz (∼1546.806 nm).

The optical spectrum at the output of the transmitter and the receiver is as shown in Fig. 20.4. The frequency difference between consecutive wavelengths is set at 100 GHz and can be observed from Fig. 20.4a. Then the WDM signal is transmitted over the fiber and at the WDM-Demux input, the signals are as shown in Fig. 20.4b. Signal attenuation is compensated using EDFA of gain 5 dB and 2.5 dB after the SSMF and DCF respectively.

The unnecessary side band power of the WDM spectrum can be filtered out using a bandpass filter (BPF) of required bandwidth.

Figure 20.5 shows the plot of Q-factor for different transmission distances and from the constellation diagram it is observed that with increasing transmission length, the noise power goes high degrading the Signal-to-Noise Ratio (SNR) of the system and a maximum transmission distance of 1000 km is achieved for a −10 dBm launch power at each channel.

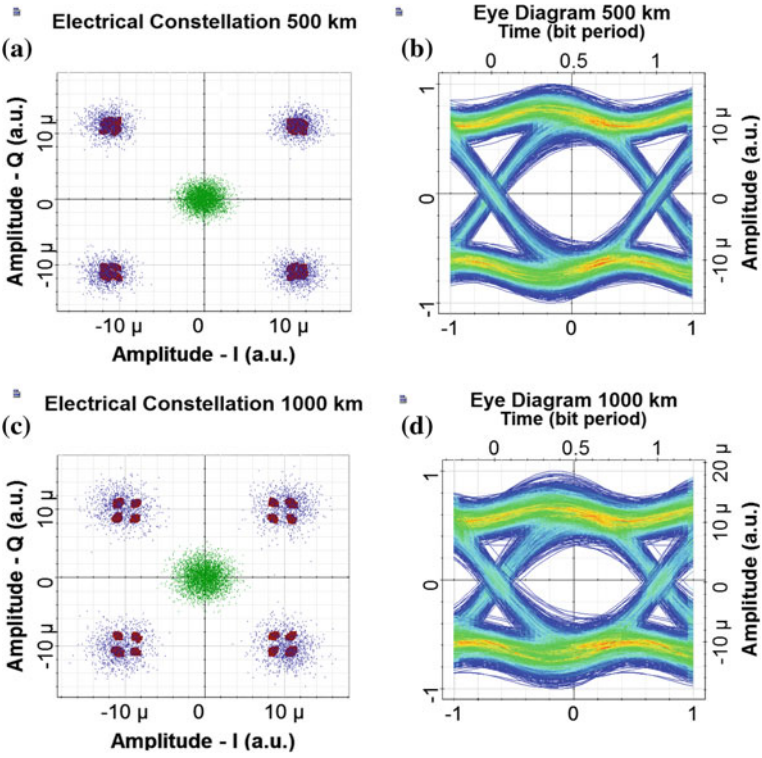


Fig. 20.2 Constellation diagrams and eye diagrams of the system. Here **a**, **b** represent 500 km and **c**, **d** represent 1000 km transmission distances respectively

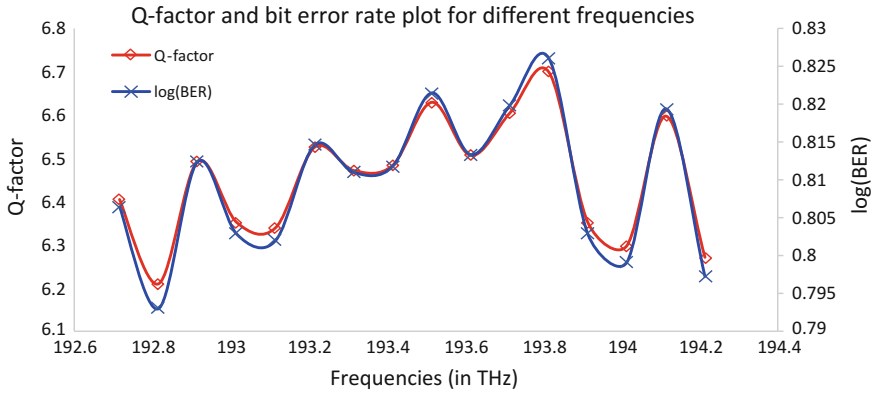
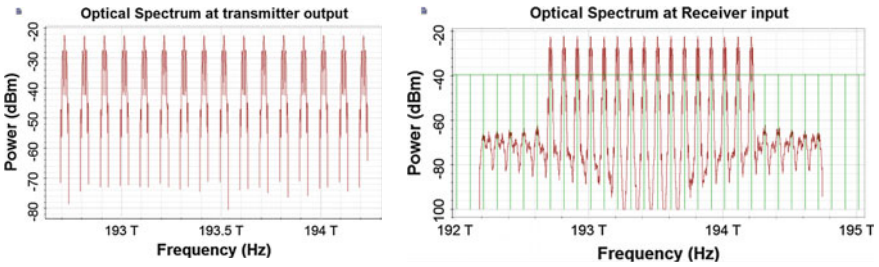


Fig. 20.3 Q-factor and BER versus frequency plot for −10 dBm launch power for all transmitted wavelengths (i.e., frequencies) after transmission over 1000 km SSMF



(a) WDM-MUX output (b) WDM-DEMUX input

Fig. 20.4 Optical spectrum at **a** the WDM-MUX output (i.e., before transmission) and at **b** the WDM-Demux input (i.e., after transmission over the SSMF)

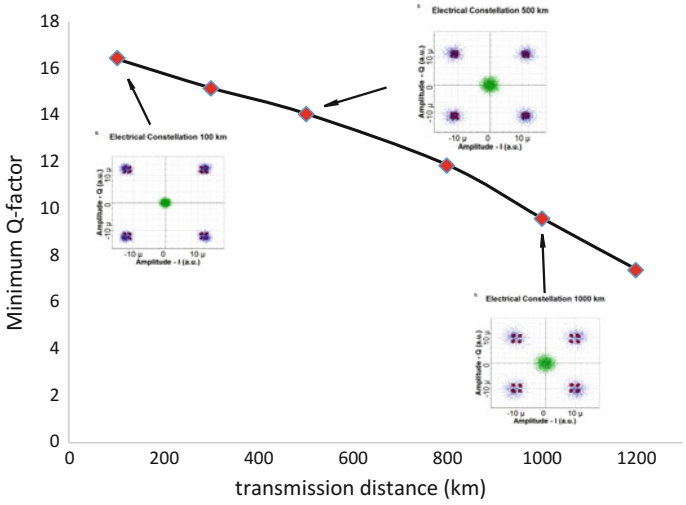


Fig. 20.5 Minimum Q-factor versus transmission distance for various fiber lengths

To investigate the system performance, a graph is plotted as in Fig. 20.6 considering Q-factor for various values of optical power at each channel, as shown in Fig. 20.3. The 500 km system shows that, to maintain a Q-factor of $\geq$ 6 (i.e., in electrical domain), the required transmit power is -13 to 7 dBm and for 1000 km the power range becomes -10 to 4 dBm. It is observed that the system performance degrades after a certain value of launch power, this is due to the fact that at higher power the fiber behaves nonlinearly leading to signal distortion often termed as Kerr nonlinearity.

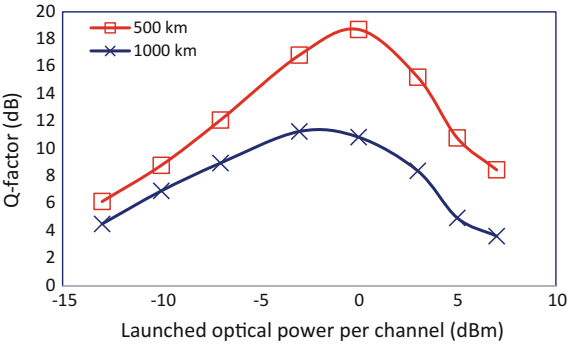


Fig. 20.6 Quality factor versus optical launch power of each channel graph for fiber lengths of 500 km and 1000 km

Conclusion

A DWDM system is evaluated for a maximum transmission distance of 1000 km with lightwave launch power of −10 dBm at each wavelength. The transmission distance can also be increased to 1500 km for a launch power of 0 dBm. The maximum achievable bit rate is limited by the modulation format, i.e., 10 Gbps using DQPSK scheme and for 16 channels, a total bit rate of 160 Gbps is achieved in simulation platform. The fiber chromatic dispersion is compensated using DCF in inline configuration. It has been observed that the designed system maintains an acceptable BER for the launch power range −10 to 4 dBm. The future work is to implement polarization multiplexing technique to achieve twice the data rate and implement more number of carriers and error correction methods.

References

1. Zhang, R., Ma, J., and Xin, X., "Full-duplex fiber-wireless link for alternative wired and 40 GHz wireless access based on differential quaternary phase shift optical single sideband millimeter wave signal," Optical Engineering, Vol. 54, No. 2, Feb 2015.
2. Chen, A. and Murphy, E. J., Broadband optical modulators: science, technology and application. CRC Press, Taylor & Francis Group, 2012.
3. Xiang, M., Fu, S., Tang, M., Tang, H., Shum, P. and Liu, D., "Nyquist WDM super channel using offset-16QAM and receiver-side digital spectral shaping," Optics Express, Vol. 22, No. 14, pp. 17448–17557, Jul 2014.
4. Wang, Y., Wang, Y. and Chi, N., "Experimental verification of performance improvement for a gigabit wavelength division multiplexing visible light communication system utilizing asymmetrically clipped optical orthogonal frequency division multiplexing," Photon. Res., Vol. 2, No. 5, Oct 2014.
5. Shahpari, A., Ferreira, R., Ribeiro, V., Sousa, A., Ziaie, S., Tavares, A., Vujicic, Z., Guiomar, F. P., Reis, J. D., Pinto, A. N. and Teixeira, A., "Coherent ultra-dense wavelength division multiplexing passive optical networks," Optical Fiber Technology, Vol. 26, pp. 100–107, Jul 2015.

6. Shim, H. K., Kim, H. and Chung, Y. C., "28-Gb/s Upstream Transmission in RSOA-based WDM PON Using Polar RZ PAM-N Format and Direct Detection," Optical Fiber Communication conference, OSA, 2016.
7. Mitchel, S., "Why complex modulated optical signals?" Agilent Technologies, Technical Article Vol. 30, No. 4, July 2013.
8. ITU Standard G.694.1 (06/02), "Spectral grids for WDM applications: DWDM frequency grid," International Telecommunication Union (2002).

Chapter 21
An Approach to Convergence Between LTE and DSRC

Raymond Star Jyrwa, Debdatta Kandar and Babu Sena Paul

Abstract With the rapid advancements of technologies and parallel development of intelligent machines in the field of wireless communication, customer needs also arises and ubiquitous network services are expected. In order to support customer's demand there is huge need of hybrid heterogeneous network. In this paper, we propose the collaboration of cellular technology like Long Term Evolution (LTE) with traditional vehicular technology i.e., the Dedicated Short Range Communication (DSRC). In this paper we proposed a LTE/DSRC hybrid architecture which can establish a bridge between LTE/DSRC communication. We proposed no changes to the already available standards but we modify some of the configurable parameters in the PHY layers. We concentrate on the OFDM modulation which is the common feature for both LTE and DSRC.

Keywords LTE · DSRC · OFDM · VANETs · Convergence

Introduction

The number of vehicles in the road has increased significantly throughout the years and will increase even more in the years to come. Therefore Intelligent Transportation Systems (ITS) have become the need of the hour with the increasing number of accidents and severe traffic jams. Researchers around the world are still

R. S. Jyrwa (✉) · D. Kandar
Department of Information Technology, North Eastern Hill University, Shillong, India
e-mail: rayj9121@gmail.com

D. Kandar
e-mail: kdebdatta@gmail.com

B. S. Paul
Department of Electrical and Electronic Engineering, University of Johannesburg, Johannesburg, South Africa
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_21

striving to come up with suitable solutions for efficient traffic routing and accident predictable systems. To ensure communication between the vehicles, the IEEE 802.11p WLAN standards have been formulated. According to the standards which fall under the Wireless Access Vehicular Environment(WAVE) communications the radio frequency spectrum from 5.85 to 5.925 GHz have been reserved for Vehicular communications. Although the Dedicated Short Range Communication(DSRC) protocol of 802.11p ensures high capacity for Vehicular-to-Vehicular (V2V) communication, its lacks scalability and infrastructure as well. LTE supports high mobility therefore it is very suitable for vehicular communication. From an economic point of view this cellular network infrastructure is already present and deployed on the ground and its devices are easily available in the market. In DSRC, vehicles are connected to each others through an adhoc wireless network called "Vehicular AdHoc Network" or VANET. VANETs are a subset of Mobile Adhoc Networks (MANET). It includes V2V communications and V2I communications and is an important components of ITS. VANETs follow the standards of IEEE 802.11p, they have less coverage range, hence we consider the support of LTE cellular technology to enhance the coverage range of VANETs. LTE offers wide range communication but costly backhaul infrastructure and licensed spectrum. Related background study is mentioned in section "Background Study". VANETs and LTE architecture is mentioned in section "Proposed LTE-DSRC Architecture". Section "Results and Discussions" Presents the proposed hybrid architecture of LTE-DSRC.

Background Study

VANET applications play an important role in the world of anywhere, anytime, any communication. Applications may be of the following types (1) Driver Service applications: Here the car can inform the driver about the surroundings, temperature, tyre's air pressure in case if there is a puncture and the nearest gas station where it can be fixed, while making a reverse turn or a u-turn etc. (2) Road Security Applications: Providing information to the driver like speed limit, or informing of the surrounding environment, and sending messages to distant cars of approaching vehicles, especially in blind curves where the driver is unable to see the approaching vehicles, information like when to over take other cars where other cars are fully aware of each other. (3) Comfort applications: Like various mode of driving for example cruise driving where the vehicle can automatically drive by sensing the car in front and make a stop if the car in the front stops. Vehicular networks by using DSRC standards provide communication in the range of 1000 m. Base station or Road side units have to be placed at regular intervals in order to provide seamless connection to the vehicles. LTE on the other hand provides communication in the range of 10 Kms thereby

reducing the number of required base stations. In [1] the author mentioned that LTE satisfy many of the application's demand like reliability, scalability and mobility support. The convergence of VANETS and cellular communication will enable a variety of applications in road safety. Over a long range LTE can enable VANETS to improve the scalability, throughput and lower latencies. The objective is to take advantage of the already available LTE infrastructure where LTE can support vehicular networking through the use of eNodeB in place of Road Side Units and using LTE enable On Board Units. The challenging task is delivering of real time data from one wireless technology to another efficiently without any delay and distortion in the data frames. Amir et al. [1] analyse the performance and compared both DSRC and LTE separately with various parameters such as vehicle density, vehicle average speed and beacon transmission frequency. Both DSRC and LTE are compared in terms of delay, reliability, scalability, and mobility. The authors concluded that DSRC provides delays less than 10 ms with throughput of 10 kbps. Nabih Jaber in his paper [2] shows the possibility of combining both DSRC and Worldwide Interoperability for Microwave Access (WiMAX). They designed WiMAX tunnel using DSRC with specific changes like network entry and handover processes in the WiMAX protocol. The authors claimed to improve WiMAX communication by decreasing the number of WiMAX connections thereby reducing the traffic congestion at WiMAX base station. The road maps for future development and their challenges related to advances in VANETs is mentioned in [3], where an overview on current research state, challenges, potentials of VANETs are mentioned. The need to deploy a heterogeneous network topology and model is mentioned in [4], based on the concept of improving the spectral efficiency per unit area. In this paper authors have proved a cost effective approach of heterogeneous networks that enhanced the capacity of LTE cellular networks. An Ubiquitous heterogeneous wireless network architecture is envisioned in [5] by integrating DSRC standards with 3GPP LTE cellular networks in order to achieve constant connectivity for ideal multimedia sessions among the clustered vehicles. Abinader et al. [6] mentioned the radio frequency spectrum of both LTE and WiFi in addition to the impacts of LTE transmissions on the WiFi, if cellular networks were to be allowed on the same unlicensed bands. It also proposes some mechanism where we can use LTE with WiFi such as the flexible spectrum access and Absolute Blank Frames of LTE. Atat et al., proposes cooperative techniques for short range (SR) collaborations to complement long range (LR) [7]. They wanted to ensure the transfer of data in the shortest time possible with the consideration of LTE network for long range links and WLAN 802.11p for short range links. The proposed cooperative schemes adapted assisted the LR networks to complement the SR networks. Beacon transmission probability in LTE and DSRC is performed in [8], the authors concluded that LTE to support beaconing for vehicular safety applications is poor according to their simulations. In LTE, the network becomes easily overloaded. Vinel et al., compared and analyzed LTE with 802.11p and suggested that ad-hoc WAVE architecture looks more promising for vehicular safety. A framework for a vehicular network using LTE namely LTE4V2X is proposed in [9] where

vehicles having both LTE and 802.11p interfaces are considered. The authors proposed a new protocol architecture and a unique frame. They proposed to use the eNodeB of LTE network to manage the clusters created in VANETs. A technique for real time video streaming of scalable video coded(SVC) over vehicle to infrastructure (V2I) are examined in [10] in which RSUs or road side units communicate with moving vehicles using IEEE 802.11p, and LTE system is used to communicate through cellular base stations. Here, three scenarios are investigated in this paper and different video transmission schemes were compared using QoS/QoE metrics for different video sequences. A heterogeneous vehicular network where LTE is the backhaul network and V2I communication is proposed in [11] in which WiFi is used for V2V communication and Ethernet for on board communication within the vehicle. A multi-interface router is introduced which has three network interfaces, LTE, WiFi, and Ethernet. The network is measured in terms of data rate, data loss ratio, delay and jitter. Imadali et al., envisioned the use of IPv6 Internet access over LTE backhaul infrastructure in vehicular networks [12]. They considered the V2V2I communication paradigm. They also proposed that in the near future there will be two types of vehicles one equipped only with 802.11p technology called as leaf vehicles and another containing both LTE and 802.11p called as Internet Vehicles. Al Sherbaz et al., proposed a Wimax-WiFi Convergence using an OFDM bridge [33]. The authors proposed not to change not to change any of the standards but to modify some of the parameters in the physical layer which are configurable.We proceed with our study using the concept mentioned in [13] and analyzing if by modifying the parameters we can bridged the gap between LTE and DSRC.

Proposed LTE-DSRC Architecture

The proposed architecture scenario for establishing communication in this convergent system of DSRC and LTE comes into play by transforming or adapting to the traditional scenario of Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I), where downlink communication is established between the LTE Road Side Units (RSU's) and DSRC On Board Unit (OBU). Whereas DSRC embedded vehicles can communicate with each other through DSRC-DSRC standards if they are within the coverage of 802.11p standard. The communication is such that LTE transmitter communicate with the 802.11p On board Unit in the vehicles. In our research we have considered common radio frequency spectrum in order to established communication. LTE operating in unlicensed bands is still an ongoing research work which requires careful investigation as mentioned in [6]. All the paper that have mention so far [1, 2, 14, 15], have omitted the radio frequency (RF) difference between the two heterogenous networks. In our research we consider other physical layer parameters apart from RF since this is still an ongoing debate to allow LTE in unlicensed bands. The proposed architecture for this particular communication is shown in Fig. 21.1.

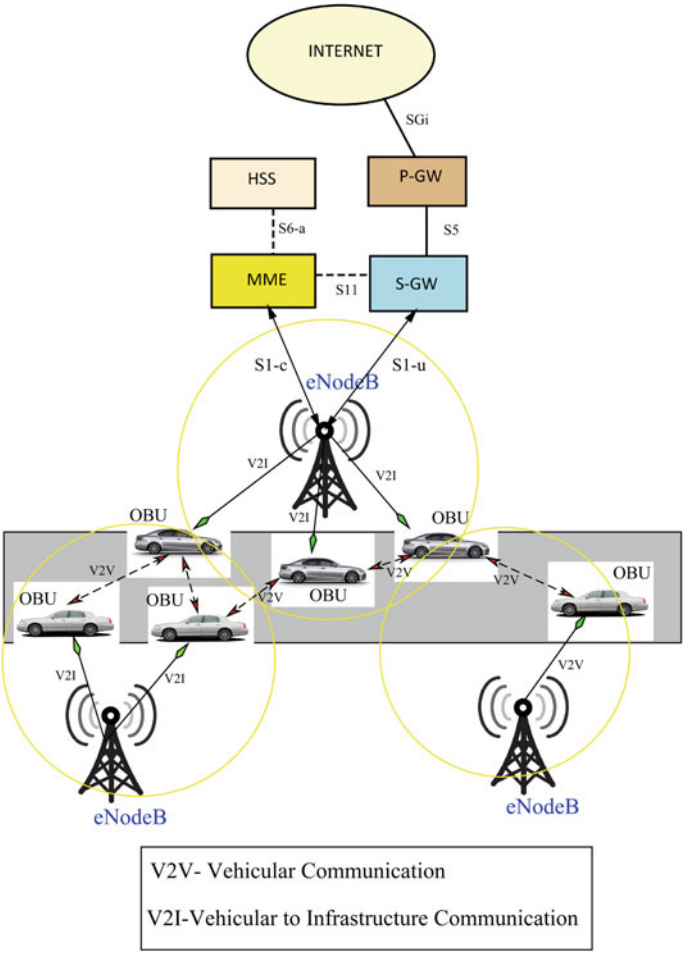


Fig. 21.1 LTE-DSRC converge architecture

In our proposal, we consider the works that have been done in [13] where the author proposed a convergence bridge between WiFi and WiMAX. Based on their work, we undertake some of the physical layer parameters of LTE and DSRC and made a few comparisons in Table 21.1. LTE uses OFDM modulation in the downlink and Single Carrier-Frequency Division Multiple Access (SC-FDMA) in the uplink whereas DSRC uses OFDM modulation both in the uplink as well as the downlink. In our research work, we considered only downlink channel as OFDM modulation is common in both technologies.

Table 21.1 Physical layer parameters

Features	DSRC	LTE
IFFT/FFT	64	1024
Bandwidth (MHz)	10	10
Subcarrier spacing (KHz)	156.25	15
Subcarriers	52	600
OFDM symbol (ms)	8	66.7
Modulation	OFDM	OFDM (downlink only)

OFDM symbols are produced by the following equation:

$$x(t) = \sum^{k} X_k e^{j2\pi ktF_0} \tag{21.1}$$

where X_k is the data to be modulated. $e^{j2\pi ktF_0}$ is the subcarrier where X_k is modulated on. $F_0 = \frac{B}{N}$ is the subcarrier frequency, B is the bandwidth and N is number of subacrriers. k is the number of the subcarrier on which X_k is modulated. Generating this signal $x(t) = \sum^{k} X_k e^{j2\pi ktF_0}$ is difficult due to the large number of subcarriers. So the signal is sampled at:
Sampling interval $= \frac{1}{sampling\,frequency} = \frac{1}{B} = T$
Lth sampling instant $= LT = \frac{L}{B}$
Now $x(t) = \sum^{k} X_k e^{j2\pi ktF_0}$

$$\begin{aligned} L\text{th sample } x(L) &= x(LT) = \sum^{k} X_k e^{j2\pi kLTF_0} \\ L\text{th sample } x(L) &= \sum^{k} X_k e^{j2\pi kL\frac{1}{B}\frac{B}{N}} \\ L\text{th sample } x(L) &= \sum^{k} X_k e^{j2\pi k\frac{L}{N}} \ \ldots \end{aligned} \tag{21.2}$$

Equation (21.2) gives the expression for the Inverse Fast Fourier Transform.

Results and Discussions

Whether it is LTE or DSRC the OFDM symbols are produced by Eq. (21.2). According to [13] convergence is a physical layer issue that cannot be solved in the MAC layer problem. We also propose not to change the standards but to modify some functions which are configurable. Simulations are performed in Matlab software where a transmitter, receiver and an ideal channel are considered. An audio signal is transmitted, where coding and modulation is performed. The data is modulated with an IFFT size of 1024 and on the receiver side the data is demodulated with an 64-point FFT. This modulation is performed on a small part of the received data. This modulation

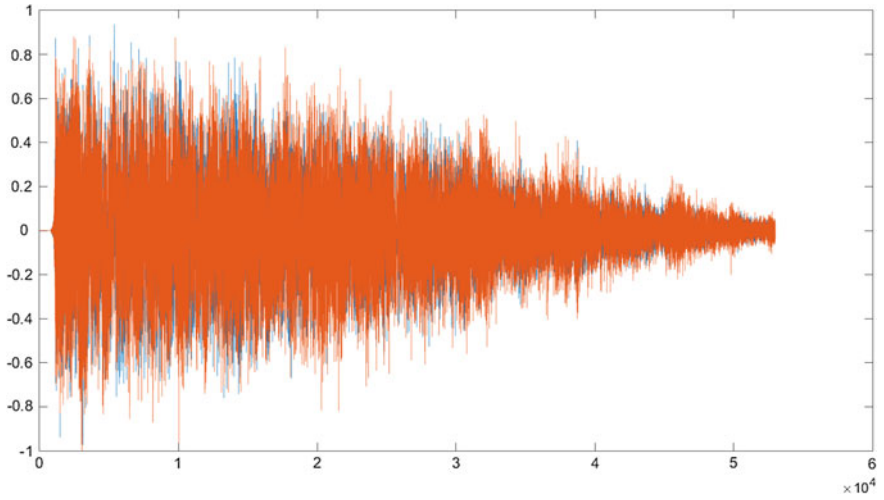


Fig. 21.2 Input audio data samples

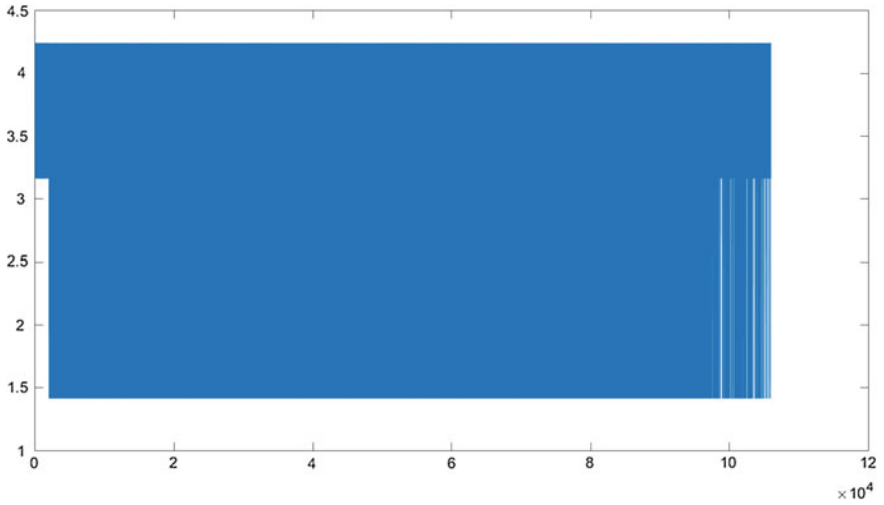


Fig. 21.3 Modulated samples

is done 16 times on all the parts in order to retrieve the entire signal and modulated using 16QAM. Figure 21.2 plots the data samples of the audio signal. Figure 21.3 plots the modulated samples. Figure 21.4 plots the ideal channel. Figure 21.5 plots the output data samples after demodulation with an FFT of size 64.

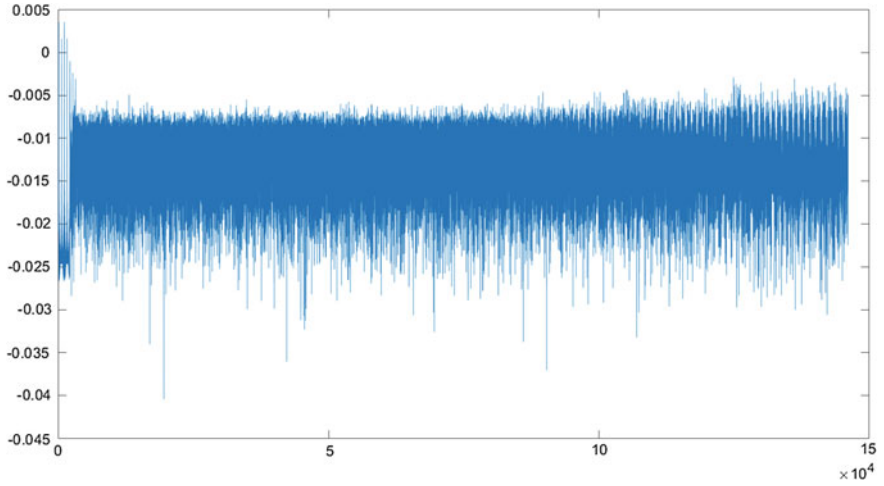


Fig. 21.4 Ideal channel plot

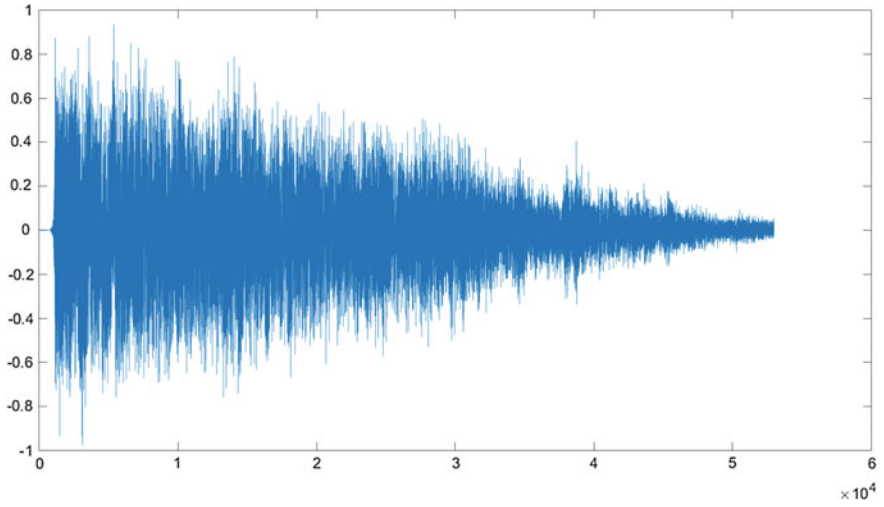


Fig. 21.5 Output audio data samples

Conclusion

Various authors compared LTE with DSRC as mentioned earlier but in this work, we addressed issues on the physical layer. Based on our study of the physical layers of both LTE and DSRC, we observed various challenging issues and differences for researchers to converge the two wireless standards. Subcarrier Spacing of LTE and DSRC is a challenging factor for such convergence approach. In conclusion we

would like to consider further research on physical layer of wireless networks and to come up with better solution. We also hope that in the near future radio frequency issues of LTE in unlicensed band will be addressed.

References

1. Zeeshan Hameed Amir, Fethi Filali, LTE and IEEE 802.11p for vehicular networking: A performance evaluation, EURASIP Journal on Wireless Communications and Networking, 2014.
2. Nabih Jaber, Nicholas C Doyle, Keal E Tepe, New Combined WiMAX/DSRC infrastructure design for efficient vehicular networking, EURASIP Journal on Wireless Communications and Networking, 2012.
3. Elias C. Eze, Si- Jing Z, En -Jie L, Joy C. Eze, Advances in Vehicular Ad-Hoc Networks (VANETs): Challenges and Road - map for Future Development, doi: https://doi.org/10.1007/s11633-015-0913-y, pp.1–18, February 2016.
4. Khandekar, A., Bhushan, N., Tingfang, J. and Vanghi, V., 2010, April. LTE-advanced: Heterogeneous networks. In Wireless Conference (EW), 2010 European (pp. 978–982). IEEE.
5. Sivaraj, R., Gopalakrishna, A.K., Chandra, M.G. and Balamuralidhar, P., 2011, October. QoS-enabled group communication in integrated VANET-LTE heterogeneous wireless networks. In Wireless and Mobile Computing, Networking and Communications (WiMob), 2011 IEEE 7th International Conference on (pp. 17–24). IEEE
6. Abinader, F.M., Almeida, E.P., Chaves, F.S., Cavalcante, A.M., Vieira, R.D., Paiva, R.C., Sobrinho, A.M., Choudhury, S., Tuomaala, E., Doppler, K. and Sousa, V.A. Enabling the coexistence of LTE and Wi-Fi in unlicensed bands. Communications Magazine, IEEE, 52(11), 2014.
7. Atat, Rachad, Elias Yaacoub, Mohamed-Slim Alouini, and Fethi Filali. Delay efficient cooperation in public safety vehicular networks using LTE and IEEE 802.11 p. In Consumer Communications and Networking Conference (CCNC), IEEE, 2012.
8. Vinel, Alexey. 3GPP LTE versus IEEE 802.11 p/WAVE: which technology is able to support cooperative vehicular safety applications? Wireless Communications Letters, IEEE 1.2 (2012): 125–128.
9. Guillaume Remy, Sidi-Mohammed Senouci, Francois Jan, Yvon Gourhant. LTE4V2X: LTE for a Centralized VANET Organisation, 2011 IEEE Globecom proceedings, 2011.
10. Yaacoub, Elias, Fethi Filali, and Adnan Abu-Dayya. SVC video streaming over cooperative LTE/802.11 p vehicle-to-infrastructure communications. Computer and Information Technology (WCCIT), 2013 World Congress on. IEEE, 2013.
11. Sadek, N. M., Halawa, H. H., Daoud, R. M., Amer, H. H., and Ali, N. A. Heterogeneous LTE/Wi-Fi architecture for ITS traffic control and infotainment. In Electrical Systems for Aircraft, Railway, Ship Propulsion and Road Vehicles (ESARS), 2015 International Conference on (pp. 1–6). IEEE. 2015.
12. Imadali, S., Kaiser, A., Decremps, S., Petrescu, A., Veque, V. V2V2I: Extended inter-vehicles to infrastructure communication paradigm. In Global Information Infrastructure Symposium. IEEE. 2013.
13. Al-Sherbaz, Ali, Chris Adams, and Sabah Jassim. WiMAX-WiFi convergence with OFDM bridge. SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, 2009.

14. George K, Onur A, Eylem E, Geert H, Boangoat J, Kenneth L, Timothy W, Vehicular Networking: A Survey and Tutorial on requirements, Architectures, Challenges, Standards and Solutions, ISSN: 1553-877X, doi: https://doi.org/10.1109/SURV.2011.061411.00019 IEEE, 2011.
15. Ian F. A, David M. Gutierrez-Estevez , Elias C Reyes, The evolution to 4G cellular systems: LTE- Advanced, Physical Communication, pp. 217–244, ISSN: 1874-4907, doi: https://doi.org/10.1016/j.phycom.2010.08.001, 2010.

Chapter 22
A Novel Weighted Vehicular Network Clustering Scheme

Phidahunlang Chyne, Debdatta Kandar and Babu Sena Paul

Abstract In today's world of emerging success in trending technologies, one of the most open and vast area of wireless communication research is the Vehicular Ad-hoc Network (VANET). Incorporating the trending technologies with the growing internet services, and deploying these into Intelligent Transportation System (ITS) is one of the open challenges. Therefore, for efficient vehicular network in ITS, in the urban areas as the nodes are mobile therefore clustering of these nodes comes into the picture. Many clustering algorithms are proposed based on different metrics and different results are obtained. This paper attempts to proposed a new clustering procedure where direction, distance, speed and transmission range are considered so as to obtain stable clusters and a weighted value is computed to select cluster head. The simulation results were simulated in NS-2.

Keywords VANET · Cluster · Mobility · DSRC · Cluster head

Introduction

There have been huge improvements in the field of wireless communication in the last decade. Vehicular ad-hoc network (VANET) communications has been a recent research area where the researchers all around the world are excavating. In vehicular communication as the nodes are movable and illustrate a dynamic topology, com-

P. Chyne (✉) · D. Kandar
Department of Information Technology, North Eastern Hill University, Shillong, India
e-mail: phida.chyne3991@gmail.com

D. Kandar
e-mail: kdebdatta@gmail.com

B. S. Paul
Department of Electrical and Electronic Engineering, University of Johannesburg, Johannesburg, South Africa
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_22

munication may be feeble at times when there is huge density especially in urban areas and due to frequent node movements frequent link failures are observed. One of the solution to enhance communication in VANET (Vehicular Network) is to cluster vehicles based on some metrics or parameters such as speed, distance, direction, velocity etc. In this paper, our main motive is to employ clustering procedure in the vehicular network. By adapting clustering into the vehicular network scenario, scalability is obtained, longer range communication in addition with less overhead at the base stations due to the presence of cluster head in each cluster which can communicate directly with the cellular infrastructure and relay the information to the cluster members. Through clustering the following advantages are achieved like the increased in network connectivity which reduced message counts, less routing overhead, message delivery is augmented and appropriate usage of network bandwidth.

In real time traffic scenarios, the vehicles move at random speed and this movement affects the stability of cluster and reorganization of clusters has to be carried out. The cluster head has to be selected in such a way as to maintain stable clusters. The rest of the paper is organized as follows. The next section reviews the related work based on clustering followed by a brief summary on the state of art focusing on clustering of mobile nodes with the system's design. Results are discussed and presented in section "Results and Discussions" and the paper is concluded in section "Conclusion".

Related Works

A few recent related papers on clustering schemes are studied which are described in this section. Clustering can be categorized based on the parameters considered for selecting a cluster head as different authors proposed various methods and chose different parameters for choosing of a cluster head.

Comparison Table of Different Clustering Algorithms

The Table 22.1 shows the comparisons between different Clustering Algorithms based on their different metrics considered along with their merits and demerits.

Proposed Approach

Our proposed algorithm is based on Fig. 22.1. The depicted scenario presented the proposed architecture for VANET communication consisting of Road Side Units (RSUs) or infrastructure based communication and on board units (OBUs) or mobile nodes.

Table 22.1 Comparisons of different clustering algorithms

Algorithm	Metrics considered	Advantages	Disadvantages
Maslekar et al. [1]	Based on the direction which the vehicle will take after crossing intersection	Overcome the influence of overtaking within the clusters and formed stable clusters	Overhead is added due to the exchange of messages and the density information amongst the nodes due to cluster-head switching
Almalag et al. [2]	The vehicle holding the highest cluster head level (CHL) value	Able to select the most stable and longer lifetime cluster-head	It applies the knowledge of each vehicles lane and the flow direction of each lane
LORA-DCBF [3]	Makes use of member direction to form the cluster in which cluster head and cluster members moves in the same direction	Minimizes flooding of its control traffic, more effective and results in more stable clusters	Only gateway nodes can transmit information to neighbour clusters within a specified period of time
AMACAD [4]	Destination, location, speed, relative destination and final destination of vehicles as a parameter to arrange the clusters	Decrease the delay of communications, bottlenecks, overhead and temporary fragmentation	the performance may decrease due to the speed difference among the members
APROVE [5, 6]	Elects cluster heads periodically, by using affinity propagation	Decreases relative mobility and distance between the cluster head and its member nodes	Performance may degrades in terms of number of clusters formed
MOBIC [7]	Uses mobility metric for selection of cluster heads	More stable clusters and thus better performance	If few neighbour nodes move differently, this method results in dramatic increase in the variance
Zhang et al. [8]	Vehicle nodes having low aggregate mobility can be selected as the cluster head nodes	It avoids unnecessary re-clustering	It does not provide secure authentication and balancing security
CBLR [9]	A new node entering the network can either join an existing cluster or create a new cluster by acting as a cluster head	Enables dynamic, self-starting, and multi-hop muting between participating mobile nodes	Performance is better when speed is less than 30 m/s, therefore limited

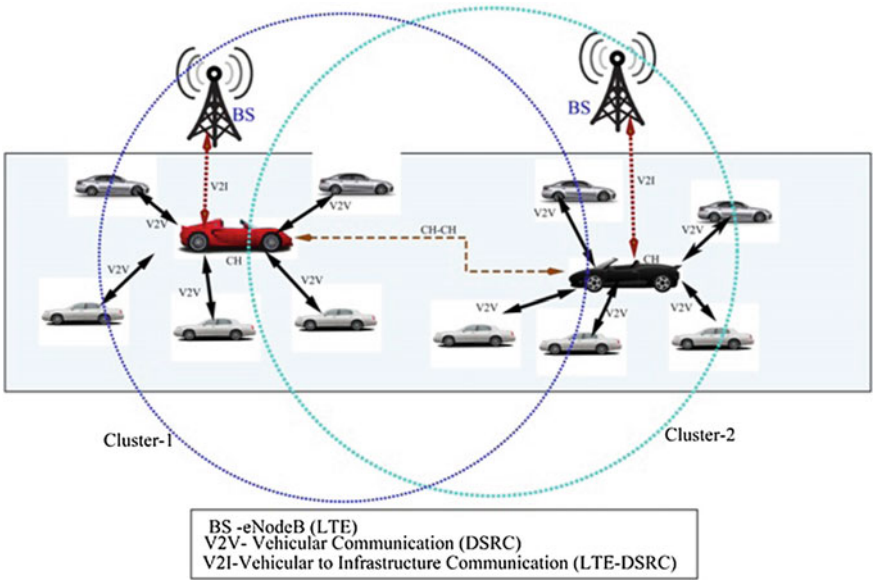


Fig. 22.1 Vehicular network scenario

Clustering in VANET means systematizing vehicles into small sets based on some common traits or characteristics such as vehicle's direction, speed, velocity etc. Due to the dynamic topology changes, cluster's stability is weak and cluster head changes are unavoidable. The main criterion for cluster formation is to minimize the overhead and increase cluster's stability. Cluster's stability can be achieved only if the cluster head duration is augmented and cluster reorganization is minimized. The formation of clusters and usage of bandwidth are the most important tasks in any clustering algorithm.

System Design

A novel architectural scenario is proposed in this paper for vehicular communication pictured in Fig. 22.1. It is assumed that all vehicles are fitted with GPS (Global Positioning System) for locating a vehicle.

The Base Stations can be any cellular infrastructure based technology and the OBUs are the vehicles with DRSC capabilities antennas. There are specific terms during cluster formation and in this paper some terms are discussed as follows for cluster formation.

1. Cluster Head (CH)
   A cluster head will act as a relay for disseminating data to the cluster members

from the base station and from the CMs to the BS. CH can also communicate with other clusters through the cluster head of another cluster which can be termed as inter-cluster communication.

2. Cluster Member (CM)
   Cluster members are the vehicles belonging to a particular cluster which are under the proximity of a particular BS. CMs can communicate with the BS through the CH, where the CH will relay the information. In addition to this, the cluster member can also communicates with each other in 1-hop communication but when the same message needs to be relayed in 2-hop the messages should not be delayed else the second CM might get stale messages.

From Fig. 22.1 it is clearly seen that the clusters might overlap. The coverage range of DSRC is 1000 m; hence vehicles within this proximity can communicate with each other. Within each cluster there is one vehicle which acts as a cluster head (CH) and as discussed earlier can relay information to the cluster members within the cluster. For a vehicle which falls into two cluster, the decision as to which clusters will it belong will be determined by its distance with the cluster heads of both the clusters and it will be with the one which is minimum.

Formation of Clusters

In real life traffic scenarios, vehicles in the same lane moves in the same or in a particular direction to the vehicles moving in the opposite lane. Therefore based on mobility and direction clusters are formed. The question arises as to how the vehicles formed clusters? As the nodes are mobile, when a vehicle enters a particular area that vehicle will first calculate it's distance with the base station to make sure whether it is under the proximity of the base station or not. If yes, then the vehicles belong to that particular cluster and then proceed in calculating its distance with the cluster head which is depicted in the flowchart in Fig. 22.2. If the new vehicle receives an acknowledgement within 5 s then a CH is already present in the proximity and the new vehicle will send a request message to join, then the CH will update the BS about the new cluster member. Otherwise, the new vehicle will send HELLO message to BS and it will elect itself as the cluster head.

During cluster formation, cluster head acts as access points for the cluster members belonging to the same cluster hence formed intra-cluster communication and cluster head of one cluster communicates with the CH of another cluster therefore, inter-cluster communication takes place.

Algorithm for Forming Clusters and Election of Cluster Head:
The following procedure is carried out for forming clusters and cluster head election. Let V = V1,V2, V3,Vn … Vn be a class or set of vehicles moving in a road, where these vehicles have DSRC transceivers. Let C = C1, C2, C3, … Cn be the cluster head classes.

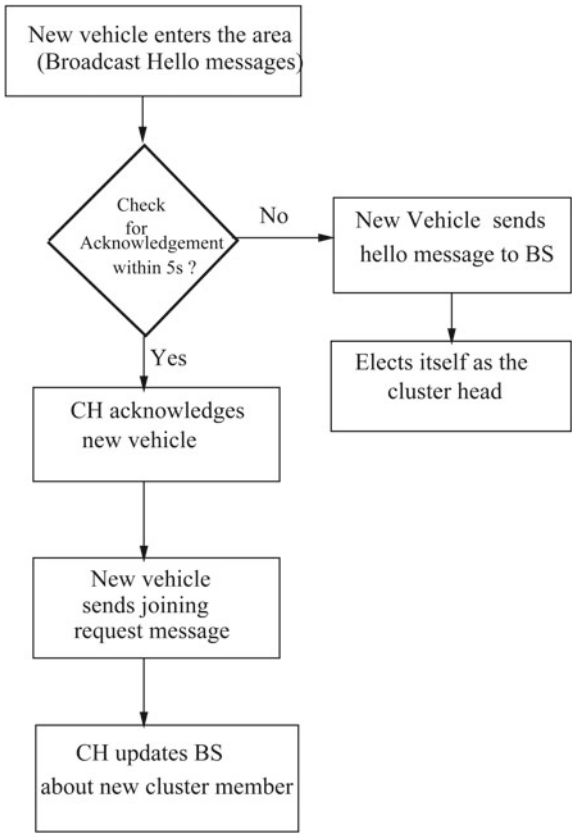


Fig. 22.2 Flowchart for cluster formation

The first step is to calculate the distance and the direction of a particular vehicle when it enters a particular area. Before calculating the direction, distance should be calculated first which can be computed as follows.

1. Calculating the relative distance between the vehicles:
   The distances between the vehicles are computed considering one vehicle with its neighboring vehicle which can also be termed as 1-hop neighbour. The most popular distance algorithm is adapted i.e., the Euclidean distance which can be computed by the following equation.

$$D_v = \sqrt{\sum (X_i - Y_j)^2} \tag{22.1}$$

   where, D_v is the distance between vehicles X_i, Y_j are the coordinates of the X and Y axis.
2. Calculating direction between the vehicles:
   For calculating the direction of the vehicles (V1, V2), beaconing message trans-

mission is taken into consideration. A vehicle V1 will broadcast two beaconing messages (M1, T) and (M2, T + 1) to its neighbor vehicle V2 in time T and T + 1, and the distances (D1, D2) are computed with these respective timing. The following algorithm is carried out

```
If D2 > D1
{
Then V2 is moving in the opposite
direction,
Cluster_members= false;
}
else
{
V2 is moving in the same
direction as V1
Cluster_member= true;
}
```

3. Transmission Range:
   The next step is to calculate the transmission range for determining the distance of the vehicles with the base station.

$$R = C_0 * t/2 \tag{22.2}$$

From the above calculated distance and transmission range a weighted value is computed for all the vehicles and the vehicle with the minimum weighted value is considered to be the cluster head. The weighted value can be computed as follows:

$$W_0 = \sum (D_v + R) \tag{22.3}$$

Hence, from the above computations clusters can be established and cluster head can be selected.

Results and Discussions

In this section the simulated results are described which was simulated in network simulator using NS-2 and SUMO which is a traffic simulator using Open Street Maps for real time environment taking up Guwahati city in Assam. Figure 22.3 depicts the broadcast of messages from the cluster head to the cluster members. The vehicles within the cluster can communicate among themselves in 1000 m range as specified for DSRC. Here, the vehicles joining the cluster as well the vehicles leaving the cluster are being noted.

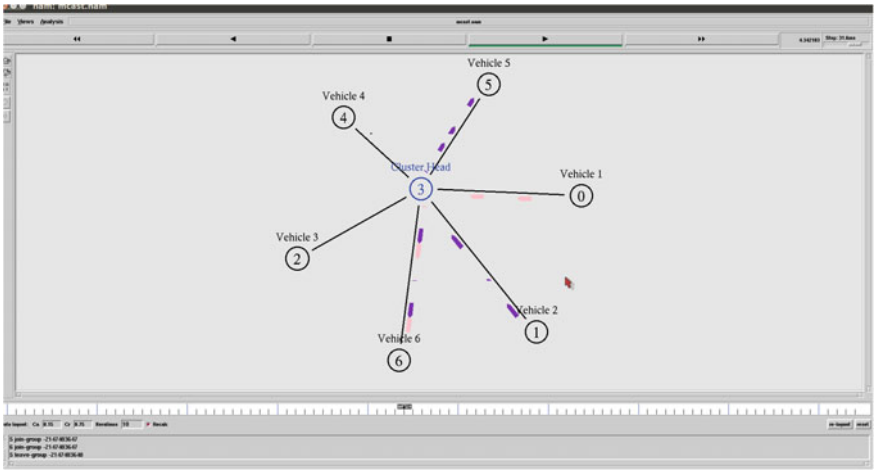


Fig. 22.3 Cluster head communication with cluster members

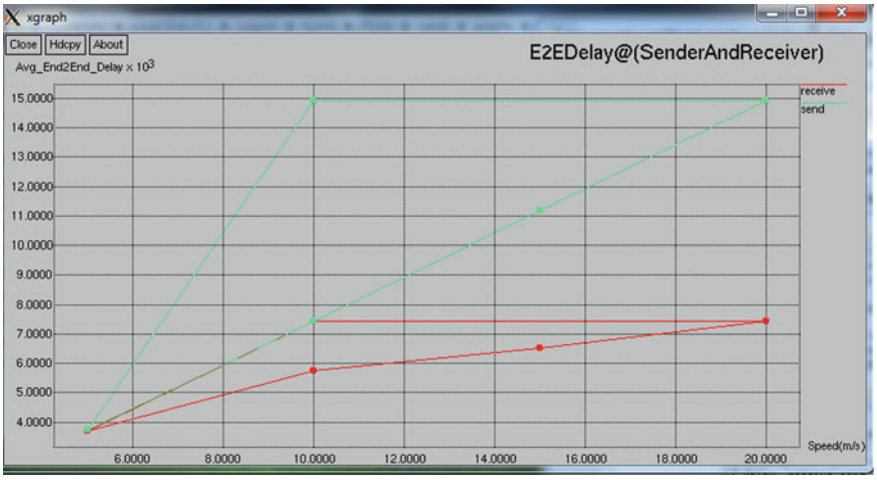


Fig. 22.4 End-to-end delay between receiver and sender

From Fig. 22.4 end-to-end delay between receiver and sender is portrayed and in Fig. 22.5 Packet Delivery Ratio (PDR) and packet drops are shown. With this approach, data dissemination is increased as clusters formed reduced congestion and bottleneck towards the base station. The improvements also include the reliability of message broadcast from the cluster head to the cluster members. In our approach clusters are formed such that any pair of nodes can either communicate directly with its 1-hop neighbor also termed as 1-hop cluster.

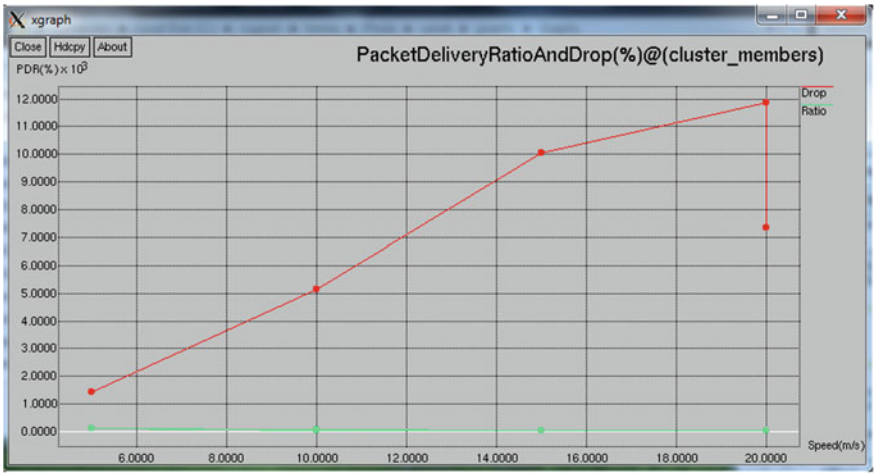


Fig. 22.5 Packet delivery ratio rate and packets drop

Conclusion

VANETs becomes a widely used promulgation technology in the recent years. In this paper, a brief review has been done on the various vehicular communication clustering algorithms proposed so far. A novel clustering procedure based on direction and distance the clusters are formed and based on a weighted value the cluster head is selected. Future works includes computation of performances with the previous proposed related clustering algorithms and to improve security as well as reliability of safety messages being delivered to vehicles.

References

1. Maslekar N, Boussedjra M, Mouzna J, Houda L, " Direction based clustering algorithm for data dissemination in vehicular networks," *In 2009 IEEE Vehicular Networking Conference (VNC)*, pp. 1–6, IEEE, 2009.
2. Mohammad, S. Almalag, and C. Weigle Michele, "Using traffic flow for cluster formation in vehicular ad-hoc networks," *In Local Computer Networks (LCN), 2010 IEEE 35th Conference*, pp. 631–636, IEEE, 2010.
3. Sadaf Momeni, Mahmood Fathy and Nasser Mozayani, "Vehicular Ad-Hoc Networks Clustering," *In: proc. Of Advanced Satellite Mobile Systems,* pp. 299–304, 2008.
4. Mildred M. Caballeros Morales, Choong Seon Hong and Young-Cheol Bang, "An Adaptable Mobility-Aware Clustering Algorithm in Vehicular Networks," *In: proc. Of Network Operations and Management Symposium (APNOMS),* pp. 1–6, 2011.
5. Hassanabadi, Behnam, Christine Shea, L. Zhang, and Shahrokh Valaee,"Clustering in vehicular ad hoc networks using affinity propagation," *Ad Hoc Networks*, Vol. 1, No. 13, pp. 535–548, 2014.

6. Shea, Christine, Behnam Hassanabadi, and Shahrokh Valaee, "Mobility-based clustering in VANETs using affinity propagation," *In Global Telecommunications Conference, GLOBECOM 2009*, pp. 1–6, IEEE, 2009.
7. P. Basu, "A Mobility Based Metric for Clustering in Mobile Ad Hoc Networks," *International Conference on Distributed Computing Systems Workshop,* pp. 413–418, 2011.
8. Zhenxia Zhang, Azzedine Boukerche and Richard W. Pazzi, "A Novel Multi-hop Clustering Scheme for Vehicular Ad-Hoc Networks," *In: Proceedings of the 9th ACM international symposium on mobility management and wireless access,* pp. 19–26, 2011.
9. R. E. R. A. Santos and N. Seed, "Inter vehicular data exchange between fast moving road traffic using ad-hoc cluster based location algorithm and 802.11b direct sequence spread spectrum radio," *PostGraduate Networking Conference,* 2003.

Chapter 23
Uplink Communication Between WiMAX and IEEE 802.11P Using Sub-channelization

V. Dhilip Kumar, D. Kandar and Babu Sena Paul

Abstract In this chapter, an integrated architecture of Vehicular standards and Mobile standards for Efficient Intelligent Transportation system is proposed. By combining these two technologies, many research challenges have been adopted to establish a communication between different standards. In this chapter, we considered the hybrid scenario of Uplink and Downlink between IEEE 802.11p and IEEE 802.16d systems. We addressed the problem of sub-channelization by establishing communication between different subcarriers. We provide an analytical solution based on the Modulation schemes as well as AMC (Adaptive and Modulation Coding). The simulation results are provided to compare the performance of efficiency of payload, coverage range and data capacity with different modulation schemes and sub-channelization.

Keywords DSRC · WiMAX · 64FFT · 256 FFT · AMC

Introduction

The Fixed Environment of Cellular standard, i.e., IEEE 802.16d (Fixed WiMAX) connection establishment is based on physical layer components such as Orthogonal Frequency Division Multiplexing(OFDM) and Medium Access Control (MAC) for providing communication to vehicular standards such as IEEE 802.11p

V. Dhilip Kumar (✉) · D. Kandar
Department of Information Technology, North Eastern Hill University, Shillong, Meghalaya, India
e-mail: dhilipkumardoc@gmail.com

D. Kandar
e-mail: kdebdatta@gmail.com

B. S. Paul
Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg, South Africa
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_23

(Dedicated Short-Range Communication) also called as DSRC. However, in some high-speed and longer range dynamic environment signals may not be transmitted or received using DSRC technology. The solution to overcome this drawback of DSRC is by incorporating Fixed WiMAX with DSRC to convey the frequency signal in multiuser DSRC environment for ITS.

Overview of ITSs

Intelligent Transportation System (ITS) are being deployed in various VANET research challenges such as inter-vehicle communication for safety application in a high mobility dynamic environment and establishing communication between different networks for longer range communication (i.e., DSRC to WiMAX). Each vehicle takes on different realm equipped with VANETs standards such as DSRC, WAVE, CALM, or IEEE 802.16e (Mobile WiMAX) also acted as On Board Unit (OBU). These standards are employed to communicate from vehicle to vehicle (V–to–V) [1, 2]. In a realistic situation, vehicles require to communicate with Road Side Infrastructure Unit (RSU) with the use of possible configuration parameters and communication protocol to deliver the information or communication between the OBUs and RSUs.

VANETs Standards

The standard IEEE 802.11 p was considered for DSRC and WAVE architecture which provides a short-to-medium range communication service in high-speed data transfers with low latency communication between vehicle to vehicle (V-2-V) and vehicle to Roadside Infrastructure (V-2-I) as well as it is deployed in the various safety applications such as Navigation system, Automatic Toll Collection, Traffic Management, etc. In 1999, the US Federal Communication Commission allocated 5.850–5.925 GHz Radio frequency band at 75 MHz spectrum [3]. DSRC is organized into 10 MHz band which consists of seven channels out of which six channels are reserved for service channels used for non-safety communications and the remaining one channel is reserved for control channel used for safety [4, 5]. DSRC are installed in vehicles and acts as OBU for gathering the sensor's information from vehicle and transmit the gathered information to other vehicles within the range of 1000 m (1 km) [6, 7].

WiMAX Standards

The WiMAX (Worldwide Interoperability for Microwave Access) is a wireless broadband access standard. The IEEE 802.16-2004 standard comprises of the static

and dynamic environment, i.e., Fixed WiMAX (IEEE 802.16d) and Mobile WiMAX (IEEE 802.16e). WiMAX governs the air interface dependent on physical layer OFDM. In Fixed WiMAX, the OFDM consists of 256 FFT with subcarrier frequencies at 3–10 MHz channel bandwidth with data capacity up to 3.5–5.8 GHz frequency band. Therefore, fixed WiMAX offers wide coverage range up to 30 Miles (around 50 km) and also act as a Road side Infrastructure. In Mobile WiMAX, physical layer contains the Orthogonal Frequency Division Multiple Access (OFDMA) and it supports multi-carrier FDMA and it supports both Frequency and Time domain duplexing (i.e., it depicts the time and frequency domain allocated to the subcarriers into various user slot and transmitting signal to multiple OFDM) [8, 9]. The physical layer of dynamic WiMAX supports 128, 512, and 1024 FFT subcarriers at 2–11 GHz radio frequency band for Non-Line of Sight (NLOS) applications and operates on the RF band from 2.3 to 3.5 GHz [10]. It also supports point to multipoint communication using MAC layer and it covers up to the range of 3–10 Miles (5–15 km) whereby it can act as an RSU or an OBU.

Limitation of Standalone Technologies:

By utilizing the existing technologies such as VANET (DSRC) and Cellular standards (WiMAX) to construct a new combined architecture framework for providing an efficient vehicular communication, modulation schemes and sub-channelization is one of the major issues to combine the different wireless communication standards. In 2013, Nicholas Doyle have proposed the combined architecture of DSRC and Mobile WiMAX by evaluating the payload calculation between DSRC physical layer OFDM and Mobile WiMAX OFDMA by using clustering and PUSC and FUSC sub-channelization. By evaluation of the performance of combined architecture, it will increase the throughput as well as decrease the control overhead [11–13].

Proposed Mechanism (Possibilities of Convergence Between DSRC and Fixed WiMAX)

In this chapter, we proposed a new combined scenario by establishing communication between DSRC and Fixed WiMAX for longer range communication. By considering the existing scenario of the combined architecture of Mobile WiMAX and DSRC, the payload can be calculated, and communication using 1024 FFT and 64 FFT subcarrier by using Partial Usage and Full Usage of Sub Channels (PUSC and FUSC) sub-channelization can be carried out [14]. Fixed WiMAX does not support the PUSC and FUSC sub-channelization scheme in 256 FFT subcarriers [15]. So we brought into picture AMC sub-channelization to evaluate the system

efficiency by influencing the existing mathematical model for solving the efficiency issue by utilizing Fixed WiMAX and DSRC, thereby providing longer range communication with high bandwidth and low latency [16]. The proposed scenario will aim for better efficiency as well as increase the coverage range and data capacity and it helps an establish communication to the huge number of users using the modulation scheme such as 16 QAM.

Sub-channelization of IEEE 802.16d

IEEE 802.16d (Fixed WiMAX) standard defines two modes of sub-channelization are Adjacent and Distributed (Diversity) Sub-channelization mode. In distributed sub-channelization mode, the sub-channels are summarized with distributed sub-carriers over all the available channel bandwidth [17], and it is further divided into two modes: Partial usage and Full Usage FUSC (Full Usage of Sub-Channels) mode and PUSC (Partial Usage of Sub-Channels) mode.

- In FUSC mode, subcarriers are dedicated to a full usage of the total formatted sub-channels in one cell.
- In PUSC mode, subcarriers are dedicated to the partial usage of the total formatted sub-channels, and it can be splits into sector/cell.

In Adjacent Sub-channelization mode, adjacent subcarriers are used to form sub-channels and it can rapidly assign a modulation and coding combination per sub-channel. Adjacent sub-channelization is also further divided into two modes: AMC (Adaptive and Modulation Coding) mode and Adjacent PUSC mode.

AMC mode is strongly suggested as it increases the spectral efficiency of the system. In Fixed WiMAX, Physical layer OFDM requires 256 FFT subcarriers and DSRC-OFDM requires 64 FFT subcarrier and supports up to 10 MHz channel Bandwidth. Therefore, with the parameters already available, in the implementation part, the following formula is adapted for calculating the Inverse Fast Fourier Transform.

$$\mathrm{N}\ -\mathrm{IFFT}:\ \mathrm{X(t)} = \sum \mathrm{NJ}\ -1\,\mathrm{X(j)} \cdot \mathrm{e}\,\mathrm{j}2\pi\mathrm{fjt}$$

The Supported Size in IEEE 802.16d is given as, [256 FFT; 192 + 8 + 56], Where, 192 data subcarriers are for data transmission, 8 pilot subcarriers for synchronization and estimation purposes and the remaining 56 Null Subcarriers are for guard bands and DC carriers.

In DSRC, the supported size is, [64FFT; 48 + 4 + 12], Where, 48 data subcarriers are for data transmission, 4 pilot subcarriers for synchronization and estimation purposes and remaining 12 Null subcarriers for guard bands and DC carriers. While establishing the communication between DSRC and Fixed WiMAX (64 to 256 FFT) sub-channelization and modulation schemes are necessary. Sub-channelization is used for grouping of subcarriers into different permutations

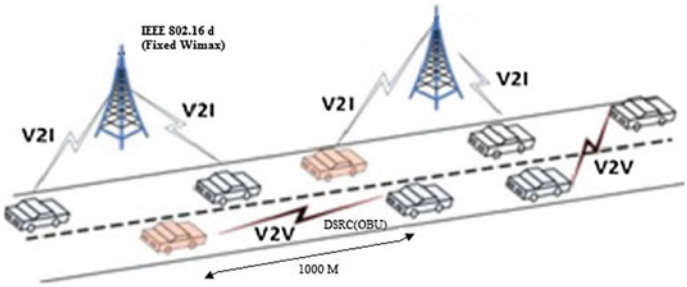


Fig. 23.1 DSRC—fixed WiMAX architecture

such as distributed subcarrier (PUSC and FUSC) and Adjacent Subcarrier (AMC). Many radio communications systems nowadays use the dynamic adaptive modulation techniques such as (BPSK, QPSK, 8PSK, 16QAM, 32QAM, 64QAM). They sense the channel conditions and adapt the modulation scheme to obtain the highest data rate for the given conditions. As signal-to-noise ratios decrease, errors will increase along with re-sends of the data, thereby slowing throughput.

Figure 23.1 shows the combined architecture scenario with DSRC radio vehicles as OBUs and Fixed WiMAX towers as Base Stations or RSUs.

Downlink Scenario

The modulation schemes for Downlink Scenario, i.e., Fixed WiMAX to DSRC is given in Table 23.1.

Fixed WiMAX to DSRC (256 FFT to 64 FFT).

While establishing the communication between same standards (Example: DSRC to DSRC) the physical layer OFDM requires 64 FFT for modulation as well as demodulation for one OFDM symbol. Therefore, BPSK requires 1 bit per OFDM symbol (i.e., 1 * Bit rate = 1* 64 FFT subcarrier).

In DSRC-OFDM, the subcarrier frequency spacing Δf is given as (Δf = BW/Nfft = 10 MHz/64 = 156.25 kHz), Table 23.2. Depicts the Parameters for 256 FFT OFDM AMC carrier allocations. Where Tg is the guard time (Tg = 8 μs). For the downlink scenario, 10 MHz system Bandwidth is required with 5.8 GHz Radiofrequency, and the Number of Subcarriers (Δf) is 256FFT and 16QAM Modulation Scheme is used, i.e., 16QAM Modulation Scheme is used for

Table 23.1 Modulation schemes for WiMAX—DSRC

Modulation	Bits per symbol	Symbol rate
BPSK	1	1* bit rate
16 QAM	4	¼ bit rate

Table 23.2 Parameter table for 256 FFT OFDM AMC carrier allocations

Parameters	Values
Number of subcarriers	256 FFT
Number of data subcarriers	192
Number of data carriers per subchannel	48
Number of bins per second	4
Number of bands	6
Number of used subcarriers	217
Number of pilot subcarriers	24
Number of DC subcarriers	1
Number of guard subcarriers, left	20
Number of guard subcarriers, right	19

modulation and demodulation of the signals from Fixed WiMAX to DSRC. Therefore, Δf * Symbol rate = 256 FFT * ¼ = 64 FFT.

Uplink Scenario—DSRC (64 FFT) to Fixed WiMAX (256 FFT)

IEEE 802.16d does not support Partial Usage of Sub-channels (PUSC) or Full Usage of Sub-channels (FUSC) channel allocations [18]. So we have considered AMC sub-channelization for the uplink scenario. Figure 23.2 depicts the block diagram of

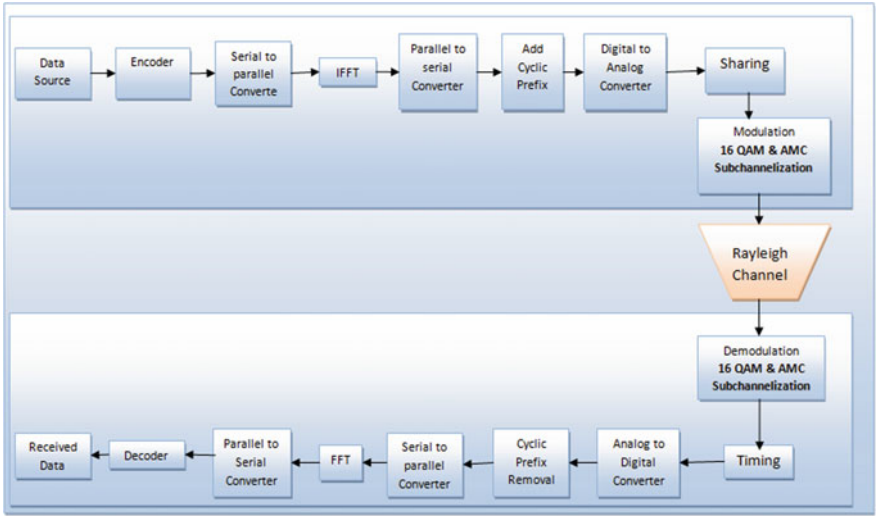


Fig. 23.2 Block diagram for WiMAX DSRC system at 16 QAM and AMC sub-channelization

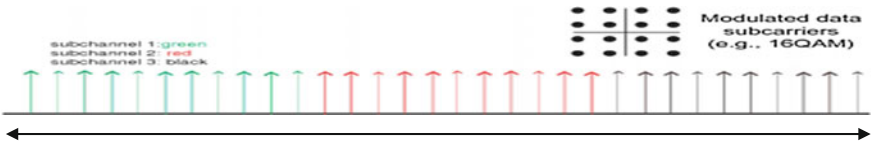


Fig. 23.3 AMC sub-channelization in 256 FFT subcarriers

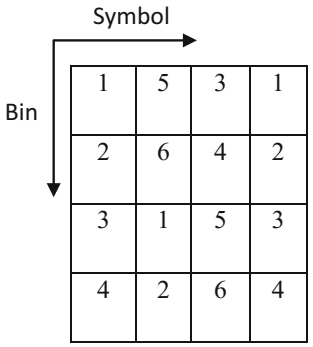
WiMAX DSRC system using 16 QAM and AMC sub-channelization. Figure 23.3 shows how the subcarriers are spaced using the AMC sub-channelization and modulated data carriers using 16QAM modulation technique.

AMC OFDM Symbol

The value of pilot subcarrier index is 9 k + 3 M + 1 for K = 0,1,…. 23, where, M = [symbol index] mod 3., therefore, the symbol of index 0 is the first AMC data symbol in the downlink or uplink. There are four types of AMC sub-channels which are different in the collection of 6 bins in a band. In the first type(default type), the available bins in a band are enumerated by starting from the lowest bin in the first symbol to the last bin in the symbol and then going to the lowest bin in the next symbol and so on. A sub-channel consists of 6 consecutive bins in this enumeration. In the other types, the shapes of an AMC sub-channel are shown in Fig. 23.3. In all the types, the index of the sub-channels in a band is increased along bins and then symbols. In Fig. 23.4, we depict the default type of AMC sub-channelization.

AMC uses adjacent subcarriers to build a sub-channel for establishing a communication for 256 FFT subcarriers into 64 FFT subcarrier. This sub-channelization consists of 6 contiguous bins within the same band. Each bin contains 6 consecutive symbols. For 256 FFT subcarriers r4 bins per second are needed, so it supports AMC by default. In AMC permutation, there are 9 adjacent subcarriers, 8 data subcarriers + 1 pilot subcarrier to form 1 Bin [19]. In Fig. 23.4, we depict 6 bands

Fig. 23.4 AMC default type



in 256 FFT AMC type, out of which 4 bins can be processed every symbol time. In the Uplink and Downlink Scenario of Fixed WiMAX and DSRC 16 QAM Modulation scheme is used to establish communication from 256 FFT to 64 FFT subcarriers (i.e., 256 * ¼ symbol rate = 64 FFT). While in the Uplink, AMC sub-channelization is based on segmentation of sub-channelization and is dependent on time (n, n + 1, n + 2, etc.). Therefore, AMC requires 1 sub-channel * (1, 2, or 3 OFDM symbols) for calculating Payload, and to examine the data capacity, range and as well as spectral efficiency.

The number of data carriers (N_{DC}) for Uplink and Downlink is given as follows:

$$N_{DC} = 48\,\text{subchannels} * \underset{1\,\text{Bin/tile}}{8\,\text{datacarriers / Bin}} * 2\,\text{Bins/symbol} = [N_{DC} = 768\,\text{Data Subcarriers}].$$

Simulation Parameters (Fixed WiMAX and DSRC)

The parameter table for finding the data capacity of uplink and downlink scenario between fixed WiMax and DSRC using AMC sub-channelization with 16 QAM modulation scheme with 10 MHz bandwidth is mentioned below in Table 23.3 [19, 20].

The graph below describes uplink and downlink communication data capacity of AMC sub-channelization at 10 MHz Bandwidth. The X-Axis denoted the number of users can achieve the signal from WiMAX and the Y-Axis denotes the data rates in kbps (Fig. 23.5).

In this plot, the red line indicates Minimum Demand for data rates to establish communication to users. It initializes the data signal, so the data rates starts from 0 and ends with 9860.4 Kbps in downlink scenario. For uplink concept, it will end

Table 23.3 AMC sub-channelization parameters

Parameter	DL AMC	UL AMC
System channel bandwidth	10 MHz	
FFT size(N)	1024	
Null subcarriers	159	
Pilot subcarriers	96	
Data subcarriers	768	
Used subcarriers	865	
Number of subchannels	48	
Subcarrier spacing	10.94 kHz	
Sampling frequency	11.2 MHz	
Cyclic prefix length	¼ FFT Size = 22.8	
Modulation scheme	16 QAM Code rate ½, 2/3, ¾, 5/6	16 QAM ½, 2/3, 5/6

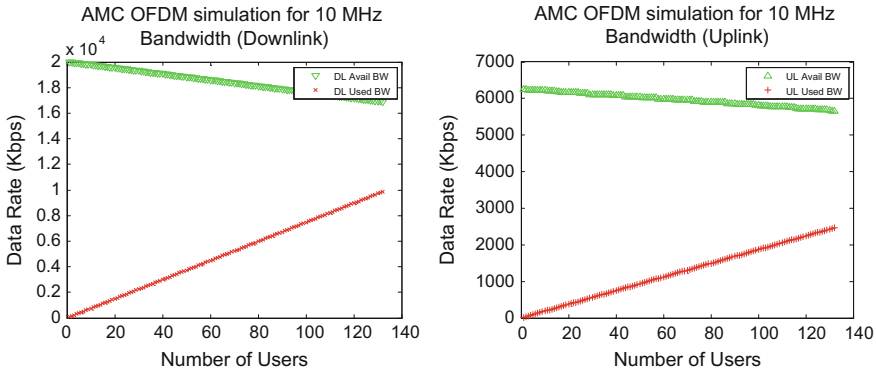


Fig. 23.5 AMC OFDM simulation for 10 MHz bandwidth

with 2465.1 Kbps. The Green line indicates the available bandwidth of uplink and downlink scenario at AMC sub-channelization. In downlink concept, the peak rate starts from 19988 Kbps and ends with 16882.9 Kbps available bandwidth. In the uplink scenario, it starts from 6246.2 Kbps and ends with 5667.41 Kbps. By evaluating these downlink and uplink communication between DSRC and Fixed WiMAX, we can achieve data capacity and we can also handle communication for 132 subscribers simultaneously.

Conclusion

The authors have proposed an uplink and downlink scenario for Fixed WiMAX and DSRC by using 16 QAM modulation scheme and AMC Sub-channelization technique to establish communication between Fixed WiMAX and DSRC, i.e., 256 FFT to 64 FFT subcarriers in 10 MHz channel bandwidth as well as 5.8 GHz radiofrequency. Using AMC Sub-channelization, we may observe some loss of frequency diversity, however, we observe an increase in the system capacity and throughput because a sub-channel is allocated to the user at any given time with the highest signal-to-noise ratio as well as capacity in the sub-channels.

References

1. IEEE Std. 802.16-2009 – IEEE Standard for Local and metropolitan area networks Part 16, “Air Interface for Broadband Wireless Access Systems”. WG802.16 - Broadband Wireless Access Working Group, 2009, pp. pp. C1–pp. C2004. https://doi.org/10.1109/IEEESTD.2009.5062485.

2. [19] Pontes, A.B., Silva, D.D.P., Jailton Jr, J., Rodrigues Jr, O. and Dias, K.L., Handover management in integrated WLAN and mobile WiMAX networks. Wireless Communications, IEEE, 15(5), pp. 86–95. 2008.
3. Dedicated Short Range Communications (DSRC) Message Set Dictionary, SAE Std. J2735, SAE Int., DSRC Committee, Nov. 2009.
4. Federal Communications Commission, "Amendment of parts 2 and 90 of the commission's rules to allocate the 5.850–5.925 GHz band to the mobile service for dedicated short range communications of intelligent transportation services", Oct. 1999.
5. Draft DSRC Message Communication Minimum Performance Requirements VBasic Safety Message for Vehicle Safety Applications, SAE Draft Std. J2945.1 Revision 2.2, SAE Int., DSRC Committee, Apr. 2011.
6. "Paul Alexander, David Haley and Alex Grant", Alexander et al.: Cooperative Intelligent Transport Systems: 5.9-GHz Field Trials, IEEE Transaction 2011.
7. Altera. An OFDM FFT Kernel for WiMAX. SanJose, CA: Altera Corporation, 2007. Application Note.
8. IEEE. Standard 802.16–2004, Part 16: Air interface for fixed broadband wireless access systems, June 2004.
9. ASTM International, "Standard Specification for Telecommunications and Information Exchange Between Vehicle and roadside Systems of 5 GHz Band for Medium Access Control (MAC), WAVE, Dedicated Short Range Communications (DSRC) and Physical Layer (PHY) Specifications, ASTM E2213-03, Aug. 2003.
10. Ran, Moshe. A mixed OFDM downlink and single carrier uplink for the 2–11 GHz licensed bands. s.l.: IEEE 802.16a, 2002.
11. "Nicholas C Doyle, Nabih Jaber, Kemal E Tepe", Improvement in vehicular networking efficiency using a new combined WiMAX and DSRC system design, Communications, Computers and Signal Processing (PacRim), IEEE Pacific Rim Conference, 2011.
12. "Nabih Jaber, Nicholas C Doyle and Kemal E Tepe", New combined WiMAX/DSRC infrastructure design for efficient vehicular networking, EURASIP Journal on Wireless Communications and Networking 2012.
13. Bhakthavathsalam, R., & Nayak. S, "Operational inferences on VANETs in 802.16e and 802.11p with improved performance by Congestion Alert", In Consumer Communications and Networking Conference (CCNC), pp. 467–471, IEEE 2011.
14. "D. Kandar, V. Dhilip kumar, P.Chyne", An Adaptive broadcasting scenario for Intelligent Transportation system, TCIFES, CLRI, Chennai 2015.
15. . "Marcel O. Odhiambo and Amimo P.O. Rayolla", The WiMAX PHY Layer, Digital Communication, ISBN: 978–953-51-0215-1, InTech, 2012.
16. . Matos, Ricardo, Bruno Sousa, Pedro Neves, Susana Sargento, and Marília Curado. "Advanced mobility in broadband wireless access scenarios." In Wireless and Mobile Computing, Networking and Communications, 2009. WIMOB 2009. IEEE International Conference on, pp. 214–220. IEEE, 2009.
17. "Carle Lengoumbi, Philippe Godlewski, Philippe Martins", Subchannelization Performance for the Downlink of a Multi-Cell OFDMA System, Wireless and Mobile Computing, Networking and Communications, IEEE 2007.
18. "Ramjee Prasad and Fernando J. Velez", OFDMA WiMAX Physical Layer, WiMAX Networks, Springer Science, 2010.
19. "Farrukh Sahar, Nawab Zada Nasrullah Khan & Arzak khan", WIMAX OFDM ADAPTIVE MODULATION WITH BROADBAND WIRELESS TECHNOLOGY HSDPA, WiMAX Forum. WiMAX and IMT-2000. (2007).
20. "Mohamed Melki, Soumaya Hamouda, Sidi Ould Mohamed, Sami Tabbane", WiMAX Subchannelization Modes: Performance Analysis and Spectrum Efficiency Enhancement", IEEE, 2009.

Chapter 24
The Time of Arrival Statistics for Cellular Communication Using Multiple Concentric Annular Rings of Uniformly Distributed Scatterers

Babu Sena Paul

Abstract Mobile communication has become popular over the past couple of decades. The channel between the base station and the mobile device depends on the environment in which the mobile device is placed. The time of arrival statistics of the signal arriving from the base station to the mobile device depends on the channel. The time of arrival statistics is an important parameter as it assists in determining the maximum width of the pulse that should be transmitted over the channel. Geometrical channel model is one of the widely used channel modeling techniques for determining the time of arrival statistics of the signals. In this chapter, we approximate the Gaussian distribution of scatterers around the mobile device by concentric annular rings of uniformly distributed scatterers to evaluate the time of arrival statistics. We subsequently compare the statistics obtained by both the methods for validation of the results.

Keywords Time of arrival · Geometrical channel model · One-ring channel model

Introduction

Mobile communication has gained popularity and brought in a revolution to the communication industry. The end to end link in cellular communication is achieved by routing the signal through a base station. The base station is generally located on top of a high rise building or at the highest point in the locality it serves. Thus, the base station is devoid of scatterers around it. Whereas, the mobile device is located at the ground level and surrounded by objects like buildings, furniture, etc. The objects around the mobile device can be considered as scatterers. The signal from

B. S. Paul (✉)
Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg, South Africa
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_24

the base station reaches the mobile device after being reflected, refracted and dispersed by the scatterers. The time required for the different signals to reach the mobile device depends on the path it traverses from the base station to the mobile device via the scatterers. The signal may reach the mobile device from the base station after being scattered by a single scatterer or multiple scatterers. If the signal undergoes multiple scattering before reaching the mobile device, then the power of the received signal degrades drastically and contributes negligible power to the detectable signal. Hence in this chapter, we consider a signal that reaches the mobile device after being scattered by a single scatterer.

Model Description

The geometrical based models are a widely used technique for modeling channels between the transmitter and receiver of wireless communication [1, 2]. In this modeling technique, as shown in Fig. 24.1, the transmitter and the receiver are separated by a distance D. The receiver is surrounded by scatterers around it inside a circle of radius R. The distance between the transmitter and the receiver (D) is considered to be much greater than the radius (R) of the scattering circle, thus justifying the use of geometric optics and representation of the waves by rays. The scatterers around the mobile device can be considered to be Gaussian or uniformly distributed depending on the environment it models. The number of scatterers does not affect the time of arrival statistics, as the time of arrival statistics is dependent on the distribution of the scatterers in the circle. But the number of scatterers should be adequately large to represent the distribution in the circle.

In the present work, we consider the scatterers to have a normal distribution around the mobile device. Similar work has been done before by Janaswamy in [3, 4]. The analytical approach in [3] is involved and cumbersome. In this chapter, we propose a much easier approach to find the time of arrival statistics of the signal at the mobile device. This is a more generalized approach and may be extended to any kind of scatterer distribution around the mobile device.

As mentioned earlier, we consider the scatterers to have a normal distribution around the mobile device confined in a circle of radius R. The probability density

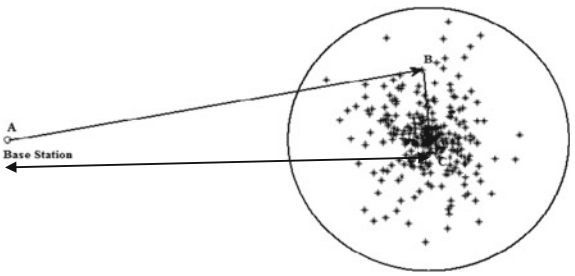


Fig. 24.1 Geometrical-based model of cellular channel

function and the cumulative density function of Gaussian distribution is given by Eqs. 24.1 and 24.2 respectively,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{24.1}$$

$$F_X(x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{24.2}$$

where,

x	*random variable*
μ	*mean value of the normal distribution*
σ^2	*variance*
$erf\,(x)$	*error function* $= \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt$

Figure 24.2 shows a plot of the Gaussian density function. It may be observed both from the plot and Eq. 24.2 that 99.7% of the data are within the -3σ *to* 3σ limits. The histograms in Fig. 24.2, gives the relative number of items in the bins. The items in each bin can be assumed to be containing uniformly distributed data selected over the bin width.

In this chapter, we approximate the Gaussian-distributed scatterer in the scattering circle by concentric rings of uniformly distributed scatterers as shown in Fig. 24.3. The number of scatterers in a particular ring is determined by the mean, standard deviation of the Gaussian distribution and the distance of the ring from the center of the circle.

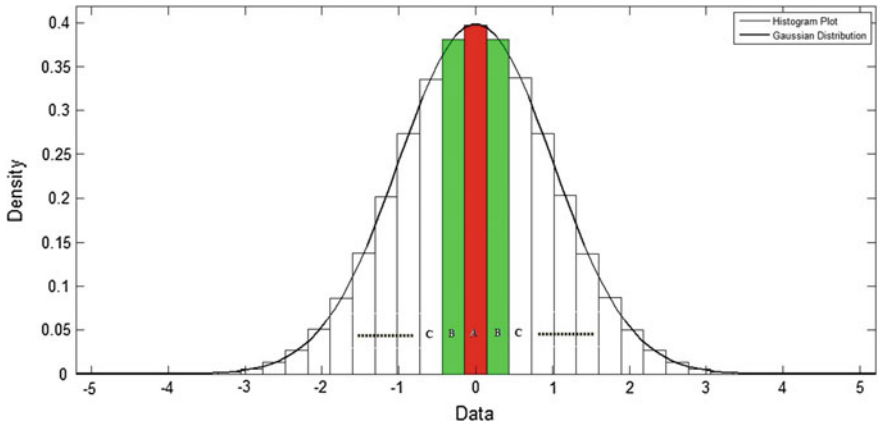


Fig. 24.2 A plot of the Gaussian density function with zero mean and unity variance

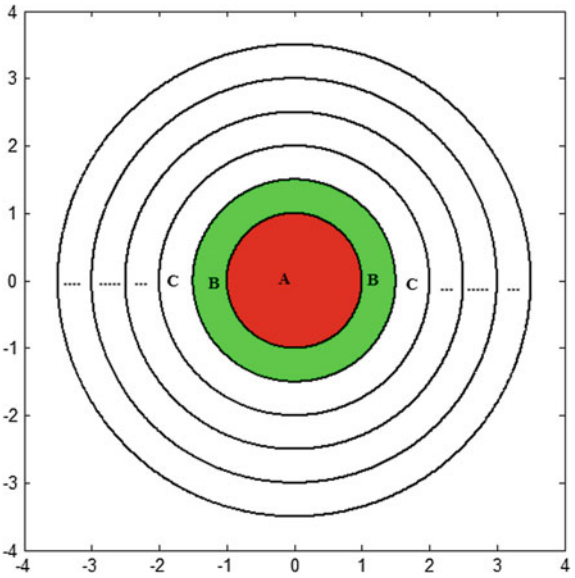


Fig. 24.3 Concentric annular rings of scatterers

Let us assume that the radius of the scattering circle is R and the standard deviation of the Gaussian distribution of the scatterers is σ. If we equate the radius of the circle to 3σ, then we can make sure that 99.7% of the scatterers are within the circle of radius R. Let us assume that the circle of radius R is divided into N concentric circles of equal width, then the width of each annular ring would be R/N. Larger the value of N, better is the approximation. Now, we are required to put the scatterers in the annular rings in such a way that in each annular ring the scatterers are uniformly distributed whereas at the same time the scatterers should be Gaussian distributed over the circle [5]. Figure 24.3 shows the scattering circle being divided into annular rings. Correlating Figs. 24.2 and 24.3, the histogram A, B,C,…. of Fig. 24.2 corresponds to the annular rings A, B, C … of Fig. 24.3. The number of scatterers in an annular ring of Fig. 24.3 should be proportional to the area under the corresponding histogram in Fig. 24.2. The area under the histogram can be obtained by putting the lower and upper limits to the integral of Eq. 24.1. It may be noted here that the integral of Eq. 24.1 is given by the cumulative density function represented in Eq. 24.2.

The time of arrival is evaluated by calculating the time taken by the signal from the base station to reach a scatterer and then to the mobile device. The statistics of the time of arrival can be obtained after calculating the time of arrival from each scatterer. It is evident that the pattern of time of arrival statistics depends on the way the scatterers are distributed. Whereas, the actual time of arrival also depends on the separation distance between the base station and the mobile device [6].

Results and Discussion

Figure 24.4 above shows 1000 Gaussian distributed scatterers generated in a circle of radius hundred meters and at whose center the mobile device is located. The position of the scatterers is generated by using the randn function of MATLAB. This generates a set of points Gaussianly distributed over the x-axis. Then to each point, an angle is associated which is generated by using a uniform random generator which generates angles between 0 and 2π radians.

Wheras Fig. 24.5 above shows one thousand scatterers generated in a circle of radius hundred meters and at whose center the mobile device is located. Here the circle is first divided into thirty annular rings of equal width. Then the scatterers are uniformly distributed in these annular rings. The total number of scatterers were considered to be thousand. The number of scatterers in each ring decreases with its distance from the center and depends on the area of the histogram it corresponds to in the Gaussian distribution.

As the coordinates of the scatterers generated by above two methods are known, the time of arrival of the signal to the mobile from the base station placed at 500 meters away from the mobile device can be calculated. The time of arrival was calculated for the signal arriving the mobile device from each scatterer in both the cases. The probability density of the time of arrival from both the cases were then plotted, as shown in Fig. 24.6 below. It can be observed that the probability density function obtained by both the methods match each other.

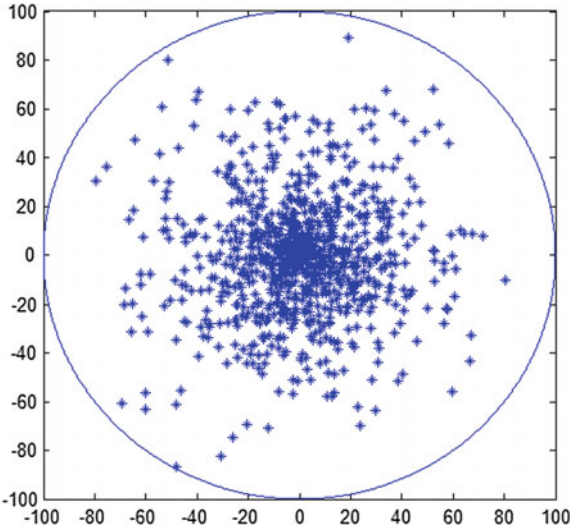


Fig. 24.4 Sactterers generated using Gaussian distribution directly

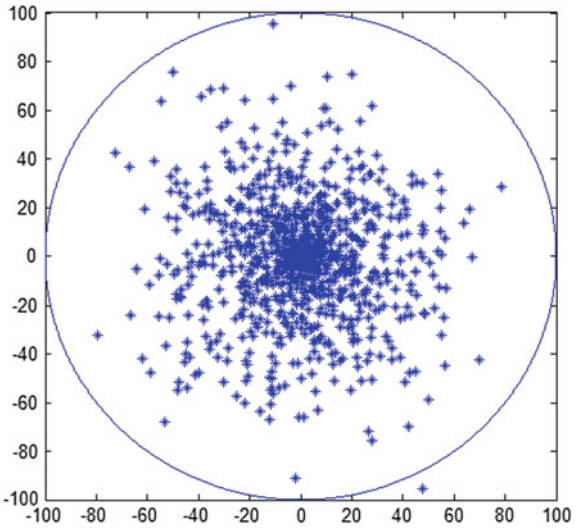


Fig. 24.5 Gaussian distribution of scatterers generated using annular rings with uniform distribution approach

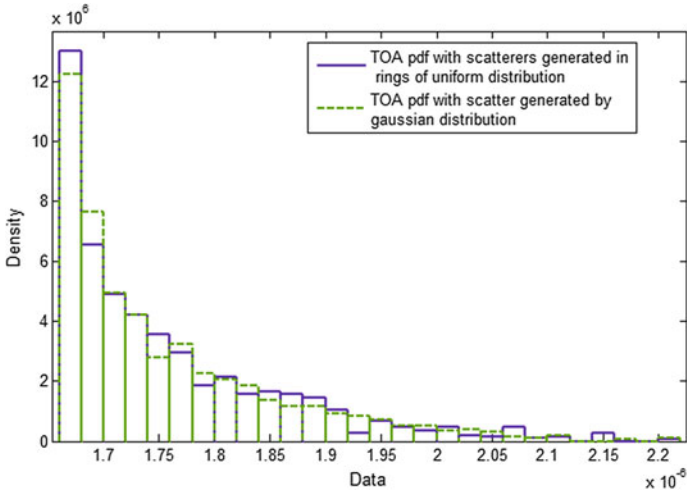


Fig. 24.6 Probability density function of time of arrival obtained by scatterers generated by Gaussian distribution and annular rings with uniform distribution

Conclusion

The time of arrival of the signal and its statistics is an important parameter in wireless communication system. The time of arrival is one of the parameters which determine the pulse width of the communication system. In this paper, we propose an alternative way to generate the Gaussian-distributed scatterers to evaluate the time of arrival statistics between the base station and the mobile device. This method of generation divides the scattering circle into equal width annular rings. Then the scatterers are uniformly distributed in the annular rings. The number of uniformly distributed scatterers in annular rings depends on its distance from the center of the ring, the standard deviation of the Gaussian distribution it needs to replicate. This method of generation of scatterers makes the mathematical formulation for the time of arrival more tractable. The proposed method was then validated with the standard method by evaluating the time of arrival statistics by both the approaches.

References

1. P. Almers, E. Bonek, A. Burr, N. Czink, M. Debbah, V. Degli-Esposti, H. Hofstetter, P. Kyosti, D. Laurenson, G. Matz, A. F. Molisch, C. Oestges, and H. Ozcelik. Survey of channel and radio propagation models for wireless MIMO systems. EURASIP Journal on Wireless Communications and Networking, 2007(1):56–75, January 2007.
2. M Patzold, and B O Hogstad, A space-time channel simulator for MIMO channels based on the geometrical one-ringscattering model, In Proceedings of the Vehicular Technology Conference (VTC 2004), Vol. 1, pages 144–9, Sep. 2004.
3. R. Janaswamy. Angle and time of arrival statistics for the gaussian scatter density model. IEEE Transactions on Wireless Communications, 1(3):488–497, July 2002.
4. R. Janaswamy. Radiowave Propagation And Smart Antennas For Wireless Communications. Kluwer Academic Publishers, 2001.
5. B. S. Paul and R. Bhattacharjee, Tine and angle of arrival statistics of mobile-to-mobile communication channel employing dual annular strip model IETE Journal of Research, vol. 56, issue 6, pp 327–332 Nov–Dec. 2010.
6. R. B. Ertel and J. H. Reed. Angle and time of arrival statistics for circular and elliptical scattering models. IEEE Journal on Selected Areas in Commnunication, 17(11):1829–1840, November 1999.

Part IV
Multimedia and Speech Processing

Chapter 25
A Study on Variation of Suprasegmental Phonetic Appearance Considered for Prosody Design with Respect to Assamese Language

Parismita Sarma and Shikhar Kumar Sarma

Abstract This paper deals with a major issue for designing a Text-to-Speech synthesizer. To design a speech synthesizer, we need speech prosody where all significant and important utterance-related information are systematically stored. An utterance can be divided into the segmental level as well as suprasegmental level. Suprasegmental level deals with syllable, word, and sentence. We are experimentally studying the behaviors of these segments with respect to Northeast Indian language Assamese. To design intonation model of any language, a clear understanding of a prominent portion of an utterance, prosodic boundary as well as tunes of the concerned sentences are necessary. In this paper, we are discussing all these features with the help of some selected speech items from our Assamese speech database. We are trying to explain the suprasegmental behavior of utterances with the help of tables and graphs prepared from our experiment.

Keywords Prosody · Intonation model · Prosodic boundary
Syllable

Introduction

In the age of information technology, the speech signal and artificial speech producing tools have a huge impact on technology as well as on society. Computers can generate artificial sound or mimic like a human with the help of an efficient speech synthesizer. The quality of a synthesizer basically depends on its ability to

P. Sarma (✉)
Department of Information Technology, Gauhati University, Guwahati, Assam, India
e-mail: parismita.sarma@gmail.com

S. K. Sarma
Cotton College State University, Guwahati, Assam, India
e-mail: sks001@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_25

understand the text and utter almost like a human with appropriate emotions or tone.

Prosody design is the core architecture of any speech synthesizer [1]. For a Text-to-Speech synthesizer, its success depends on the efficiency of the prosody. Prosody deals with phrasing, loudness, pitch, duration, and intonation in suprasegmental level. Suprasegmental phonetic appearance or features mean segments which are beyond phoneme level. Phonemes are the smallest component considered for speech synthesis [2]. Suprasegmental features may be syllables, words or sentence at best. These characteristics are very much dependent on the semantic behavior of the language. Depending upon the concerned language, the range of the said features changes, still few of the parameters shows similar behavior when prosody design is concerned. The basic suprasegmental properties we will discuss here are a variation of pitch or fundamental frequency (F0), accent in the syllables or words, tune prediction and prosodic phrasing. The evaluation of proper feature values for these acoustic parameters enriches the prosody.

Assamese Language

Assam is a state of Northeast India. The inhabitants of Assam is called Asomiya and Assamese is the native language for the people of Assam. Literally, we can define a language as sequences of symbols capable of expressing our thoughts. Speech is a structured arrangement of meaningful smaller acoustic units. Every language is unique in terms of semantics and scripts used to write. The Assamese language is basically spoken by the inhabitants of Brahmaputra valley. The broad Assamese language has totally three allied languages [3]. All of them have some similarity with Assamese language. That is why there is dielectric variation found in the Assamese language. This variation exists due to the geographical variance of the state. Hazong is the one used by some of the people of Goalpara district of Assam. Second allied language Bishnupriya is spoken by people of Barak valley and it is a language of Bangladesh. The other language Sakma was formed at Sattagram of Bangladesh. Some people of Barak valley of Assam, Arunachal, Mizoram and Tripura district use this Sakma language. In spite of these vast variations in spoken Assamese language there is only one written script officially recognized till now. In this chapter, we are analyzing suprasegmental behavior which is necessary for intonation study of the language with respect to officially recognized Assamese language. It is basically spoken by middle and upper Assam people and also used in any official or social conversation. The text of our concerned language is written in Assamese script and the language inherits phonetic character set and behavior from Sanskrit [4]. Assamese is an Indo-Aryan language, written from left to right and has no capital letters. Assamese has totally 8 phoneme vowel, 21 consonants and totally 15 diphthongs [4]. Vowel phonemes can occur in all the three positions, beginning, ending, or middle of a word. A few important phonological characteristics of Assamese language are mentioned here. The language has no any retroflex sounds

[4]. But it has a series of alveolar sounds. In Assamese language extensive use of velar nasal/ŋ/is found. [4] Voiceless velar fricative/x/(শ,ষ,স) is also used in most of the Assamese words. This phoneme is not found in any other Indian languages.

Introduction to Text-to-Speech Synthesizer

There are two basic phases of speech synthesis process [5]. Every phase consists of a few separate modules which carry out specific tasks necessary for that phase. The first phase is called as natural language processing (NLP) phase or front end of the synthesis system. Front end does preprocessing of raw text which generally includes text normalization, digit identification whether date, time, phone number or rupees, expansion of abbreviation to meaningful units. As a whole tokenization of whole text to identifiable units are done in this phase. A linguistic module is used for syllabification; which is important for our suprasegmental analysis, some morphological psychoanalysis, grapheme to phoneme conversion, etc., are done in the first phase. The second phase which is called as back end, make the desired sound file taking input from Front end phase. Second phase finally gives the speech waveform after processing a few steps. In the whole process of speech synthesis phonetic and acoustic models are used for intonation and prosody design. In this chapter, we are showing results of few intonation events for speech units on suprasegmental level for Assamese language.

Related Works Done with Other Indian Languages

Suprasegmental analysis is found for other Indian languages also. A lot of works are already done with south Indian languages like Tamil, Telugu, etc. Suprasegmental speech features are extracted for word boundary recognition for the Hindi language [6]. This is done for Bangla and Punjabi languages too. Duration prediction and F0 detection in syllable level are done for Hindi, Tamil, and Telugu languages with the help of neural network application [7]. As these features are most useful for building intonation model for the particular language, every Indian language where speech recognition and synthesis are done, suprasegmental analysis is carried out accordingly. Hindi, Marathi, Gujarati languages also have their own speech synthesizer and their speech also has different suprasegmental behavior according to the structure of the language. All these works are already done for most of the Indian languages.

Intonation as a Suprasegmental Occurrence

The term intonation in a broad sense means a variation of pitch in the period of utterance [8]. Pitch is the fundamental frequency of an utterance. In this chapter, we are discussing the movement of pitch and type of movement. To build a good speech prosody, we need a multilevel data structure. To construct this multi-level data structure we need intonation model of that language. For languages such as Assamese, Bangla, Tamil, etc., and English sentence-level pitch variation is critically observed instead of a word or syllable-level stress. A study of total pitch variation in the whole sentence is called as intonation. Intonation model is used to convey sentence-level meaning to an utterance. It is not concerned with lexical stress or lexical tone. In this paper, we are discussing three most important suprasegmental phonetic event called as prominence, boundaries, and tune.

Prominence

We can see most of the sentences have some words, phrase, or syllables that are more prominent than the others. The prominent portion is the focus point, it leads the meaning of the whole sentence. For English language, the prominent portion may be a word or syllable of the sentence. It is louder and longer. The same sentence carries different meanings depending on the prominent phrase. In this chapter, we are experimenting prominence characteristics for Assamese sentence. Some language has prominent portion only in louder and longer syllables or words [9]. Some others have high pitch movement instead of loudness or long segment length. Pitch accent is another feature of prominent segments of utterance. We have made a few tables and graphs to determine prominent characteristics of some selected Assamese sentences. We have compared acoustic features of the emotional sentences with their affirmative or normal form.

Prosodic Boundary

Any spoken sentence can be divided into one or more chunks. This procedure is called prosodic phrasing [1]. Delimiters generally identify phrasing breaks. All prosodic phrases should be pronounced with proper and accurate duration. Accurate duration of these prosodic phrasing is necessary to understand the correct meaning of the sentence. For a spontaneous speech, it is quite difficult to imagine the necessary pause. There are some techniques with the help of which proper prosodic boundary can be assigned [9]. This is reasonably significant for Assamese language. There are some words in this language with more than one meaning. One example is কলা (kola). This word has two meanings ‘deaf’ and another meaning ‘art’. For

example an Assamese sentence "ল'ৰাটো কলা কুশলী হতঁৰ বৰ আদৰৰ" meaning "The boy is adorable to the artists". কলা কুশলী should be pronounced together to mean *the artists*. The time gap between the words should be negligible compared to others when it is uttered in a sentence. It is just like one word. Without proper duration between the two words, meaning of the whole sentence will be "The boy is deaf, he is adorable to Kushali and all". The Kushali will mean a woman's name. Normally, a pause is imposed on punctuation marks like comma, semicolon, and full stop. These are explicit identifiers to assert pause. Though there are many techniques to determine prosodic phrasing, we tried in Festival framework with a statistical model called CART decision tree. This approach uses a probabilistic predictor. Prosodic phrases are predicted on features like surrounding word's parts of speech, length of utterance counted in number of words, potential boundary distance from starting or ending of an utterance, surrounding words are accented or not [1]. This procedure is used in Festival with the wagon-predictor tool. When we designed an Assamese speech synthesizer with the festival framework, we used this tool and evaluated different results. Between two adjacent syllables, the duration can be calculated using four features. Each invocation uses a feature vector in the form shown below.

((syl_name name) (syl_type type id) (p_syl_name name) (p_syl_type type id)).

The function return a duration value based on those features. For example, ((syl_name যাম)(syl_type 8) (p_syl_name খাম)(p_syl_type 6)).

Prosodic Tunes

In a single way, tune is called as the melody of utterance. Some languages are called tunic language. The sentences of this type of language express different meanings depending upon the tune of the utterance. We are experimenting in search of this feature for Assamese language also. The position and distribution of the words in all the sentences remain the same, but the meaning becomes different due to different tunes. The same string of text can express joy, sorrow, grief, or question on diverse speaking tune. In English language, we can see this feature. In case of English interrogative (yes/no answered) sentences, verb occupies the first position. But this structure is not applicable to Assamese sentence. For example "তুমি বজাৰলৈ যাবা" is an Assamese sentence, in the affirmative or normal mode the sentence means "You will go to market". But the same sentence can be uttered with an angry, inquiry, or surprise mood. In asking mood sentence will be tuned like "Will you go to market?". Its Assamese, structure of the sentence is same in both the moods. The sentence in interrogative mood is enquiring about whether he is going to market. On the other hand in surprise mood, exclamation is asserted like *you* will go to *market!* The person, who is speaking, is surprised to know that you are going to market. All these differences in tunes are obvious from the pitch or fundamental frequency contour.

Sample Selection and Recording Procedure

We have collected almost 500 phonetically rich Assamese sentences. The recording was done on a multimedia computer with inherent noise cancelation property. The room is sound proof and quiet. We used Audacity and Cool Edit Pro to record and process the speech segments. The speakers were both male and female in the age group of 25–30 years old. We have taken this age group to get proper pronunciation of every utterance, as formant, pitch and other phonetic as well as acoustic features depend upon dimension and shape of the vocal tract. Both of the persons were adequately trained, such that they can incorporate the proper stress to appropriate syllable and tune to the whole sentence. This is the most important work have to be taken care of when recording is carried out. Next, the recorded pronunciation was examined by an Assamese phonetic expert. The microphone we used with 48000 Hz. frequency response.

Experiment and Analysis

We have selected some sound Assamese sentences for prominent accent detection. One of them is "চিনেমা খন ভাল নালাগিল" in English "The movie was not good" and other is "তুমি বজাৰলৈ যাবা" ("You will go to market"). Prosodic prominence can be distinguished with the help of pitch accent. In this chapter, we are trying to establish this experimentally for Assamese speeches. Syllables and words are the segments, where pitch accent is found. Detection of prominent accent in syllable or words is important for designing intonation model. From a general study, we observe that Assamese language is influenced by pitch accent. This feature is more prominent in case of syllables compared to words. Using an efficient Assamese syllabification technique we syllabified both of the above mentioned sentences. Every language has a few different syllabification rules specific to that language. But overall syllabification theory for all language is same. A sonority is identified for a group of phonemes and it forms a syllable. Already we have motioned that to study accented segment of an utterance for prosody design we need a good syllabification or segmentation technique. It should be such that which is able to extract required units of utterance for proper study of accented or stressed parts in it.

Table 25.1 shows F0 values for individual syllable in the sentence "চিনেমা খন ভাল নালাগিল". The sentence was uttered by a female radio announcer. We have considered normal and angry mood for the sentence. After the recording is over, we divided the sentence to its constituent syllable manually in Cool Edit Pro software. Then we calculated F0 values for every syllable. Thus, we have prepared the table. From this table, the following graph is constructed (Fig. 25.1).

From the graph, it is obvious that angry mood is significantly different from normal mood if we compare them in parallel. We have plotted fundamental frequency (F0) for every syllable in the sentence and connecting these points got these

Table 25.1 F0 values in Hertz for individual syllable in normal and angry mood

	চি	নে	মা	খন	ভাল	না	লা	গিল
Normal	315.91	275.25	334.2	312.78	293.07	265.78	196.9	164.49
Angry	357.89	313.88	350.77	317.43	351.8	458.86	219.58	171.04

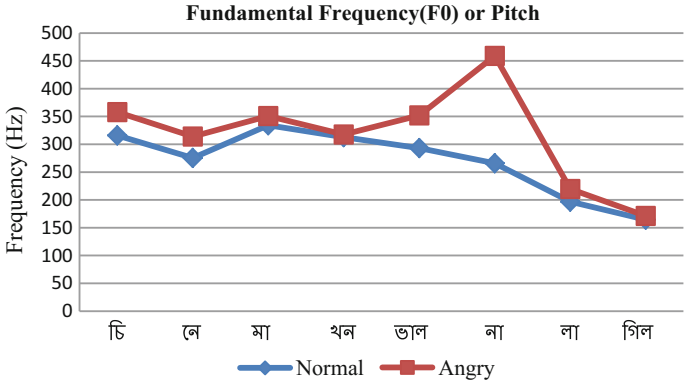


Fig. 25.1 Graph for syllable pitch values for "চিনেমা খন ভাল নালাগিল" in normal and angry mood

two graphs. Significantly most of the syllables in both the modes with values ranging from 5 to 250. A huge difference is observed in two syllables only. In angry mood an accented syllable is there with high value, that syllable is stressed to express anger. The syllable is না (naa). and in Assamese it has a negative meaning. This syllable which makes the sentence emotional, a sharp change in pitch at that point, and high value is noticed.

Another example is "তুমি বজাৰলৈ যাবা". English meaning of the sentence is "You will go to market". We are considering normal, interrogative and surprise mood for the sentence (Table 25.2).

With the same procedure as above, this sentence is also divided into its constituent syllables. We have had to syllabify all the three sentences as they are uttered in three different moods, normal, interrogative, and surprise. Next, we have plotted them with a syllable in x-axis and F0 values in Hz. on the Y-axis. To make the

Table 25.2 F0 values in Hertz for individual syllables in normal, interrogative, and surprise mood

	তু	মি	বজা	ৰলৈ	যা	বা
Normal	137.04	157.73	109.93	113.4	99.26	128.72
Interrogative	145.6	154.367	105.7	141.31	97.1	132.56
Surprise	146.19	167.35	94.52	182.43	216.14	113.79

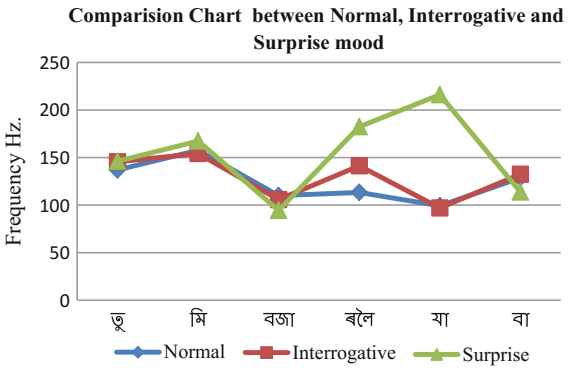


Fig. 25.2 Graph of F0 values for the syllables of "তুমি বজাৰলৈ যাবা" in three moods

comparison easy, we had plotted them on the same graph. The graph is as shown below and explanation is given in the following paragraph.

Figure 25.2 is the plot of F0 for Assamese sentence "তুমি বজাৰলৈ যাবা". F0 values of an individual syllable are marked and we got the graphs. In the interrogative sentence, the speaker is asking, where the person will go. Is it *market*? In Assamese *বজাৰলৈ?*. When the same sentence was uttered in surprise mood meaning was converted to "Is that true you will *go to market*". Stress was given on *বজাৰলৈ যাবা*. Syllable accent is prominent at syllable "ৰলৈ" in all three moods and at "যা" for surprise mood. At syllable "বা" for the same surprise tune, a lower value is noticed. It is because the speaker is rising down its tune to express surprise at end of the utterance. "যাবা" (go) is single word, yet accented part is prominent only in one syllable of the word. The same experiment was done in many sentences and the same type of results were obtained. So we can say that emphasis is given on syllabic level when the prominent accent is concerned for Assamese emotional sentences. The speaker decides in which word or phrase more emphasis should be given to express the meaning. The accented part is highly depended on context. When designing the prosody all these factors should be considered.

Tune means melody. In phonology, this term is related to F0 contour or pitch contour. A sentence can be uttered in different rhythms according to speaker's intention. For different emotions, the same sentence will have different utterances with different pitch contours. To analyze the shapes of pitch contour, we have selected some other prominent and clear speaking Assamese sentences from our database. Among many emotionally pronounced sentences, we have selected most frequently uttered emotional tunes. One is the interrogative type of sentence, where answer is only yes or no. Another interrogative type sentences have wh- starting words. Such as what, when, where, and why. In Assamese language sentences started with, কিয়(why),কত(where),কেতিয়া(when), কোনে(who), etc., belongs to this category. Pitch contour rising statements surprise or contradiction tune all are important and we are analyzing these tunes with their F0 contour.

As shown in Fig. 25.3 a, yes or no answer questions have tune gradually rising at end. Here, the example sentence is "*মইঘৰত যাম?*". Such type of question making

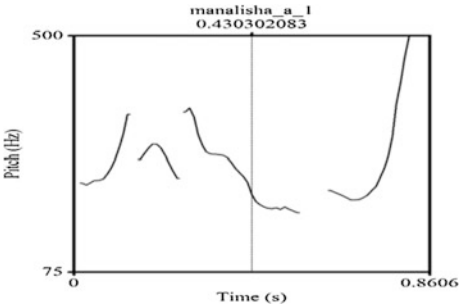


Fig. 25.3 F0 contour for "*মই ঘৰত যাম?*" (should *I go* home?)

sentence generally has one or more accented words. The contour of this type of sentence gradually rises up from the accented word. As shown in this example, the speaker is asking someone whether he will go home. The answer will be yes or no. For this type of sentence, contour rises from the main accent and ends at end of the sentence. If a sentence starts with the words where, when, what, it means WH question asking words, pitch contour is as shown in Fig. 25.4. Sentence with these starting words has gradually falling contour. As stress is imposed on the particular WH word to get the answer what is asked. Another type of contour is shown in Fig. 25.5. For declarative and informative type of sentences, beginning and ending words become more prominent. The focus of the sentences becomes broad. So broad accent is seen in the contour of these sentences. As shown in the example "*খাবলৈ মন যোৱা নাই*" (I *do not* want to have *food)* is an informative sentence. It has a flat and broad pitch contour. Steep or high rise in pitch contour at end of the sentence signifies seeking approval. As shown in Fig. 25.6, an Assamese sentence "*তেওঁ পৰীক্ষাত উত্তীর্ণ হ'ব?*". In the above sentence, the speaker is asking whether the fact will really qualify. The meaning of the sentence according to utterance is "is it possible that he will pass the *examination?*". Can the statement be approved? This type of sentences has the meaning that speaker comes to know the fact just now only. It has high rising F0 contour at the end of the sentence. Figure 25.7 shows another type of pitch contour. The example sentence here is "কি কৈছা *তুমি*!!". It is a surprise or exclamatory tune. The speaker will be wondered to know what the other person is speaking. Inherent meaning of the sentence is "Do not you know! that is

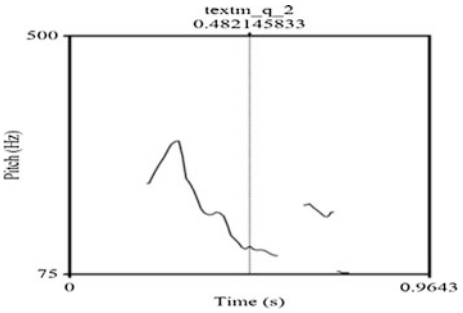


Fig. 25.4 F0 contour for "*ক'ত যাবা তুমি?*" (*where* will *you* go?)

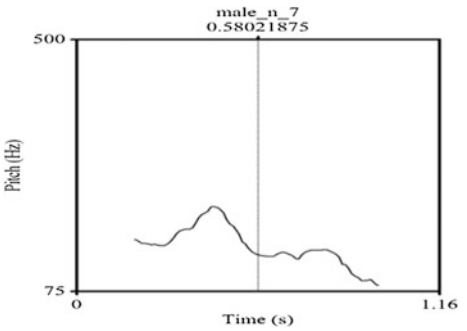


Fig. 25.5 F0 contour for "*খাবলৈ* মন যোৱা *নাই*" (I *do not* want to have *food*)

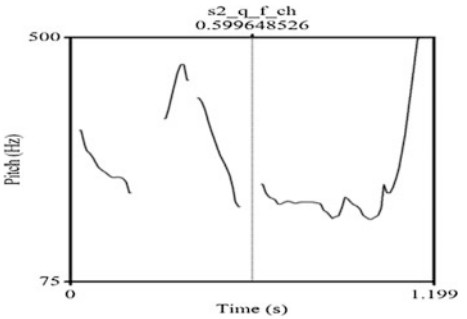


Fig. 25.6 F0 contour "*তেওঁ পৰীক্ষাত* উত্তীর্ণ *হ'ব*?" (will he *pass in exam*?)

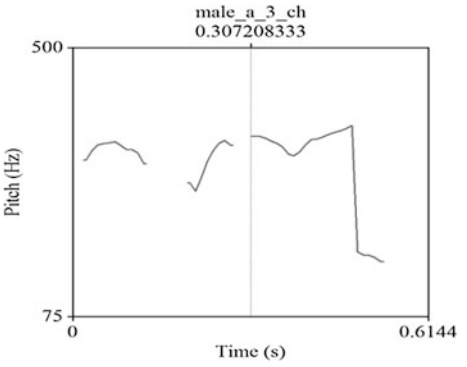


Fig. 25.7 F0 contour for "কি কৈছা *তুমি*!" (what are *you* saying!)

not true". This contour has similarity with the example discussed in prominence paragraph of the sentence "তুমি বজাৰলৈ যাবা". The last syllable of that sentence had a lower pitch in surprise mood. Sometimes, we utter this type of tune to express

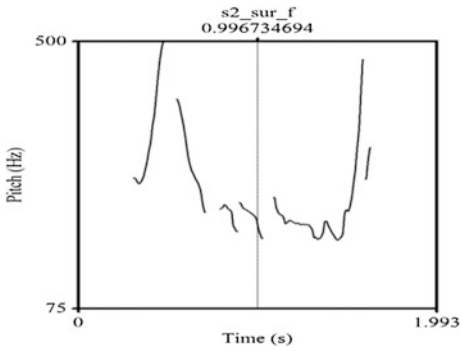


Fig. 25.8 F0 contour "*তেওঁপৰীক্ষাত উত্তীর্ণ হ'ব!*" (will *he pass* in exam! In surprise mood)

over confidence. There are many sentences in Assamese language of this type. According to pitch contour shown in Fig. 25.7, the sentence already starts with a high accent at the beginning. It maintains that accent and falls down with high velocity at the end. The last contour shown in Fig. 25.8 is called as a contradictory or opposing tune. At the beginning, there is a steep or quick fall, then slow and flat motion of contour, at last rising. Our example sentence is "*তেওঁপৰীক্ষাত উত্তীর্ণ হ'ব!*". It means "it is so surprise to know that he will pass the exam". Starting accent is on *তেওঁ* means *he*. In middle flat falling down accent, rising suddenly at the end. It is a contradiction type of contour.

In all the above graphs, Figs. 25.6 and 25.8 show pitch contours of the same sentence, but they are uttered in two different moods. That is why we can see some differences in their contour also. Figure 25.6 is a contour of an interrogative mood, expressing doubt over the situation, will he *pass in exam*. The accent is noticeable in *পৰীক্ষাত(exam)* and (a verb means complete the work). On the other hand contour in Fig. 25.8, is a contradictory curve seen with meaning "is that *he*, who will pass the *exam*".

Conclusion

In this chapter, we have experimentally analyzed different Assamese sentences and their pitch variation, prosodic phrasing, and prominence patterns in a larger unit of utterance. We have examined individual syllable of a sentence pronounced in different moods and studied the accented syllable. A syllable or word of a sentence is accented to express the meaning more accurately. A sentence will have a different meaning if the position of prominence segment changes. In the process of building a speech synthesizer to get natural sounds, every syllable or word should be uttered with appropriate accent and break. Pause between two words is also variable depending upon the context of the sentence. We have mentioned about prosodic

phrasing for a better understanding of pause. In this chapter, we are experimenting only with officially spoken Assamese tune. The pitch variation for people of lower Assam seems to be a bit different, as they speak in some different tune. We are experimenting with those dialects and planning to work with those datasets also. Pitch contour is related to the whole sentence. There are different types of pitch movement along with the sentence like rising, falling, and flat. Every type has its own significance. At last, we can conclude that prominence is attached to syllable, prosodic boundary is related with words and phrase of an Assamese utterance, the tune is with the whole sentence. This type of analysis is done for most of the languages in the world. To extract and understand all these prosodic features speech recording is most important. We need proper and sophisticated recording tools and expert persons to record the sentences with appropriate loud tunes. Moreover, at the time of recording the person must be able to incorporate the exact emotion to the whole sentence. We cannot expect actual scenario without appropriate data.

References

1. Thomas S.: Natural Sounding Text-To-Speech Synthesis Based On Syllable-Like Units, (2007).
2. Shattuck-Hufnagel, S. and Turk A. E.: A prosody tutorial for investigators of auditory sentence processing. Journal of sycholinguistic Research 25 (2), (1996), 56–80.
3. Kakati Banikanta.:Assamese its formation and development, 5th ed.. Guwahati, India, LBS publication, 2007.
4. Goswami G.C.:Structure of Assamese, FIRST EDITION, Department of Publication, GAUHATI UNIVERSITY, 1982.
5. Kishore S.P., Sangal R. and Srinivas M., Building Hindi and Telugu voices using Festvox, in Proceedings of International Conference on Natural Language Processing, (2002), 18–21.
6. Agrawal A, Jain A, Prakash N, Agrawal S.S.: Word boundary detection in continuous speech based on suprasegmental features for Hindi language, Signal Processing System, 2nd International Conference, IEEE (2012).
7. Shreenivasa Rao. K: Modeling Supra-Segmental Features of Syllables Using Neural Networks, Book Chapter "Speech, Audio, Image and Biomedical Signal Processing using Neural Networks" By Prasnna Prasad, volume 83, pp 71– 95.
8. Alam F.:kotha The First Text to Speech Synthesis for Bangla Language, (2006), 17–21.
9. S. Lee, Y.H. Oh,:Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems, Speech Communication, Vol. 28, 1999, 283–300.

Chapter 26
A Hybrid Particle Swarm Optimization and Artificial Bee Colony Algorithm for Image Contrast Enhancement

Saorabh Kumar Mondal, Arpitam Chatterjee and Bipan Tudu

Abstract Image contrast enhancement is a vital part of image processing application for improving visual and informational quality of a distorted image. For this purpose, Conventional Histogram Equalization techniques are most common approaches for both the purpose of enhancing the image contrast and preserving its main characteristics. But conventional HE techniques are not suitable all the times for preserving all the image characteristics to improve the overall quality of an image. In this regard, optimization techniques provide better results by controlling proper parameters for different methods. This paper shows the implementation of a hybrid optimization technique comprising of the search dynamics of Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC). The effective output from PSO search algorithm has been implemented with the ABC techniques to get better contrast enhancement while optimizing the objective function designed towards preserving the important characteristics of the low contrast images. The method is tested with different test images. The output is compared with the conventional techniques in both visually and against different image quality metrics. The visual results as well as the metric-based comparisons show the potential of the presented method over the conventional techniques.

Keywords Image contrast enhancement · Histogram equalization (HE) Particle swarm optimization (PSO) · Artificial bee colony (ABC) Hybrid optimization

S. K. Mondal (✉)
Haldia Institute of Technology, Haldia, India
e-mail: sm_751@yahoo.com

A. Chatterjee · B. Tudu
Jadavpur University, Kolkata, India
e-mail: arpitamchatterjee@gmail.com

B. Tudu
e-mail: bip_123@rediffmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_26

Introduction

Image enhancement is an imperative process in different image processing applications. By this process, the original images are basically converted to other modified images, which will improve the image interpretability of human viewers. Contrast enhancement is one of the most vital part of image enhancement process. A contrast enhancement technique is one for getting a better quality image for the purpose of all image processing applications [1].

Classical Histogram Equalization (CHE) is a common way to improve image contrast, which probabilistically remaps the existing image intensity levels to the available intensity levels [2]. Apart from CHE, many algorithms have been presented in this regard, such as Dual Sub-Image Histogram Equalization (DSIHE), Brightness Preserving Bi-Histogram Equalization (BBHE), etc. [3]. However, many such techniques suffer from problems like over enhancement, whitening of the image, non-preservation of image brightness, false contouring, etc.

Optimization-based HE may be a possible alternative to such problems with conventional techniques. Such types of techniques are Genetic Algorithm [4], Particle Swarm Optimization [5], Artificial Bee Colony [6], etc. These optimization techniques basically search an optimal solution, which will preserve the image characteristics as well as enhance the image contrast to improve the overall quality of an original image. This paper presents a hybrid optimization approach, where the search dynamics of Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) are combined. The approach in this paper performs the search using PSO algorithm while performing the ABC operations on the solutions found with different objective measures to find one solution that can be considered as optimal for a better quality of the image.

PSO-ABC Hybrid Optimization

PSO and ABC hybrid methodology is totally based on the basic operation of Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC) optimization techniques.

Particle Swarm Optimization

PSO was proposed by Eberhart and Kennedy [7]. This method is a population-based optimization algorithm. In this method, a number of particles are used for searching the best solution around in search area by comprising the swarm moving. During the search, each particle tracks it best position around in the search area. This is denoted by Personal best, in short Pbest. From all the Pbest values, PSO tracks the best Pbest

value achieved by any particle in the neighborhoods. This is denoted by global best, in short Gbest. Every particle moves with a certain velocity and reached the final location using Pbest and Gbest, calculated by Eqs. (26.1) and (26.2) respectively [8].

$$V_c = wV_p + c1 \times r1 \times (pbest - L_p) + c2 \times r2 \times (gbest - L_p) \tag{26.1}$$

$$L_c = L_p + V_c \tag{26.2}$$

where V_c is the current velocity, V_p is the previous velocity, L_c is the current location of the particle, L_p is the previous location of the particle, r1 and r2 denote randomly generated numbers in between 0 and 1. c1 and c2 are stochastic factors or learning factors, which are normally, set to 1.3 each [9]. w is inertia weight, which affects the convergence and exploration in PSO process. Normally, it is set from 0.6 to 0.9 [9]. By proper tuning of process parameters, the final position has been achieved, which will provide the better improvement of an image.

Artificial Bee Colony

ABC algorithm was introduced by Karaboga and Basturk [10]. This algorithm is a newly invented optimization technique, which works with the basic activities of natural bees in bee colony. In ABC algorithm, three kinds of bees are there: employed bees responsible for searching nectar sources, onlooker bees for choosing the food sources by employed bees and scouts, which search the new food sources randomly.

In ABC algorithm, consider N number of employed bees are there. These N number of employed bees' search N different food sources. Next, each onlooker chooses a food source searched from those N sources depending on the fitness probability and generates a new food source based on that food source. This probability function can be calculated as in Eq. (26.3) [11].

$$p_i = fit_i / \sum_{k=1}^{N} fit_k \tag{26.3}$$

where p_i denotes probability function. fit_i denotes fitness of food source X_i (i = 1, 2, … ,N). fit_k is the total fitness of N food sources. The food source, which is not accepted according to the fitness value, acting as scout and starts searching new food sources randomly by Eq. (26.4) [11].

$$Y_{ij} = X_j^{\min} + r(X_j^{\max} - X_j^{\min}) \tag{26.4}$$

where r denotes random real number in between 0 and 1. X_j^{min} is the lower limit and X_j^{max} denotes upper limit of that problem. j indicates the dimension of the problem.

PSO-ABC Hybrid Algorithm

In this method, PSO algorithm is executed first by initializing particles with random position and velocity. For each iteration process, it will generate one solution through the calculation of updated particle position (L_c). At the end of the final iteration, various solutions are achieved based on the updated particle position. These solutions will act as the initializing solutions for ABC operation. Then, ABC operation is executing for calculating the final solution called optimal solution by memorizing greatest solution achieved so far, which will provide much better quality of image for all aspects. The pseudo code of this method is shown below.

- Take low contrast image.
- Compute histogram and unique intensity levels.
- Initialize particles with random position and velocity for PSO algorithm.
- Evaluate Pbest for each particle.
- Evaluate Gbest among all the particles.
- Update particle velocity.
- Iteration wise update particle position and set as the final solution.
- Generate various solutions for maximum number of iterations.
- Take those solutions for initializing ABC operation.
- Execute ABC operation.
- Memorizing greatest solution achieved so far.
- Stop once the set target is reached.
- Reconstruct the image.

Objective Function Formulations

The objective function of this paper is designed using three important image quality metrics, namely, Absolute Mean Brightness Error (AMBE), Entropy and Peak Signal-to-Noise Ratio (PSNR). Apart from these three metrics, other types of metrics are also available, based on which the overall image quality is dependent. The brief mathematical descriptions of these three metrics are shown below.

- The absolute difference between the input image mean and the output image mean is known as AMBE. It is used to preserve the input image's original brightness and to maintain the overall image quality. Mathematically, it is calculated by Eq. (26.5) [12].

$$AMBE = |E(I) - E(O)| \tag{26.5}$$

where I denotes the input image and O denotes the output image. E(.) represents the expected statistical mean value. To get better preservation of input image brightness, AMBE should be low.

- Image entropy is a measurement, which indicates the prosperity of an image. Quality of an image will be proportional to the entropy value. Entropy is calculated by Eq. (26.6) [13].

$$Ent[P] = -\sum_{k=0} P(k) \log_2 P(k) \tag{26.6}$$

where P denotes probability for measure the difference between 2 adjacent pixels of value k.

- PSNR is defined as the ratio between the maximum signal power and the existing noise power. It will indicate the efficiency of the output. Here, the original image data represents the signal and noise is the error in the output image.
  To calculate PSNR of an image, first Mean Square Error (MSE) calculation has to be performed. MSE denotes the average mean error of input image pixels intensity values and output image pixels intensity values. Lower MSE result will provide better image quality.
  Assuming that N is the total number of pixels for the input image (X) and output image (Y). Then, MSE (Mean Square Error) is calculated by Eq. (26.7) [14].

$$\text{MSE} = \sum_{\text{i}} \sum_{\text{j}} |X(i,j) - Y(i,j)|^2 / N \tag{26.7}$$

where i and j denotes the pixels position.

Based on MSE, PSNR is then calculated by Eq. (26.8) [14].

$$PSNR = 10 \log_{10} (L-1)^2 / MSE \tag{26.8}$$

where L denotes the total intensity levels. The higher value of PSNR indicates the better output image quality.

The objective function of this paper is designed as Eq. (26.9), where PSNR and entropy are used to divide the AMBE since both of them indicates betterment with higher values while AMBE shows betterment with lower values. It also enables the optimization toward minimization of the objective function.

$$\varphi(x,y) = AMBE(x,y)/PSNR(x,y) + Entropy(y) \tag{26.9}$$

where x denotes input image and y denotes output image.

Results and Discussions

Various standard test images have been taken for verifying the presented method. Three of the results are presented here with the results of conventional techniques for visual comparison purpose. Figures 26.1, 26.2, and 26.3 represent the enhanced results using various methods including the presented hybrid PSO-ABC optimization method. All the images taken here are in grayscale format using standard database [15]. The results show that the presented method shows comparatively better visual improvement rather than other techniques. It is observed that the presented method not only improves contrast but also tries to preserve the other image characteristics for visual information and thus the gray tones are better retained while the other techniques tend towards more black and white images with lesser gray tonal effects. Also, the presented technique overcomes the false contouring and patches effects which are present in conventional techniques and give a sort of synthesized image appearance. The objective comparison using different standard metrics are shown in Table 26.1.

In Table 26.1, three different types of image quality metrics PSNR, AMBE, and Entropy are indicated here for comparison against conventional image contrast enhancement method (CHE, BBHE, and DSIHE respectively) along with presented PSO-ABC hybrid optimization method. Each metric value indicated here is the

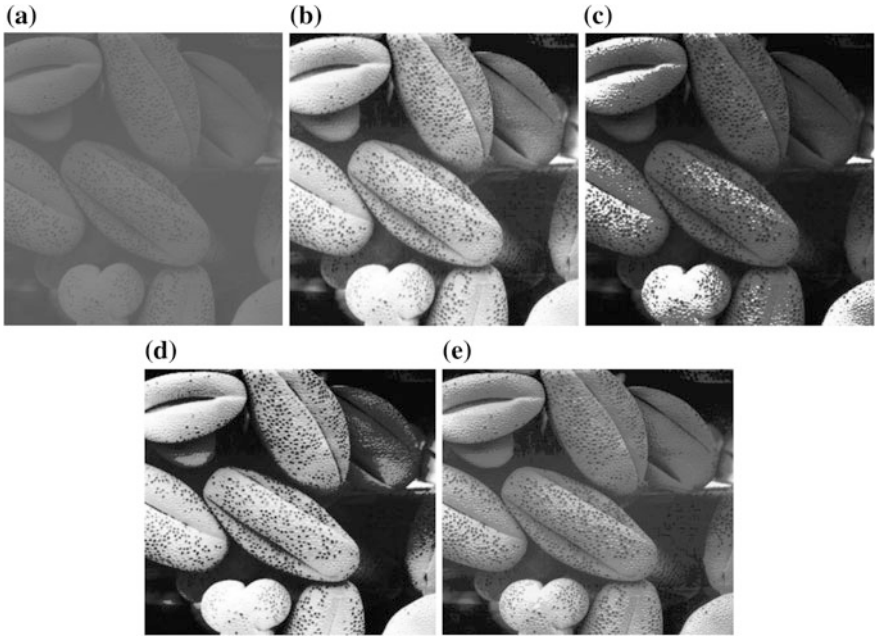


Fig. 26.1 Nut image: **a** Original image, **b** CHE-based enhancement, **c** BBHE-based enhancement, **d** DSIHE-based enhancement, and **e** PSO-ABC hybrid optimization-based enhancement

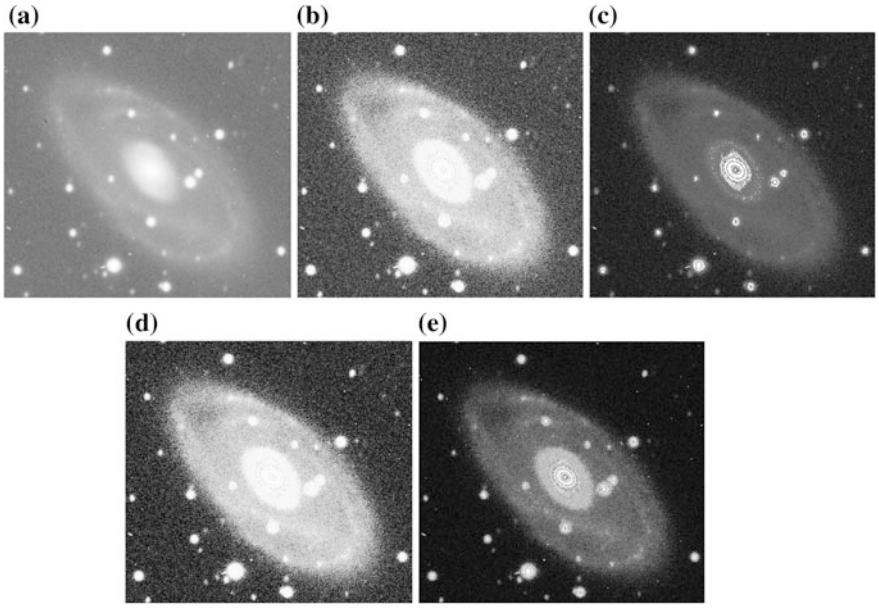


Fig. 26.2 Solar image: **a** Original Image, **b** CHE-based enhancement, **c** BBHE-based enhancement, **d** DSIHE-based enhancement, **and e** PSO-ABC hybrid optimization-based enhancement

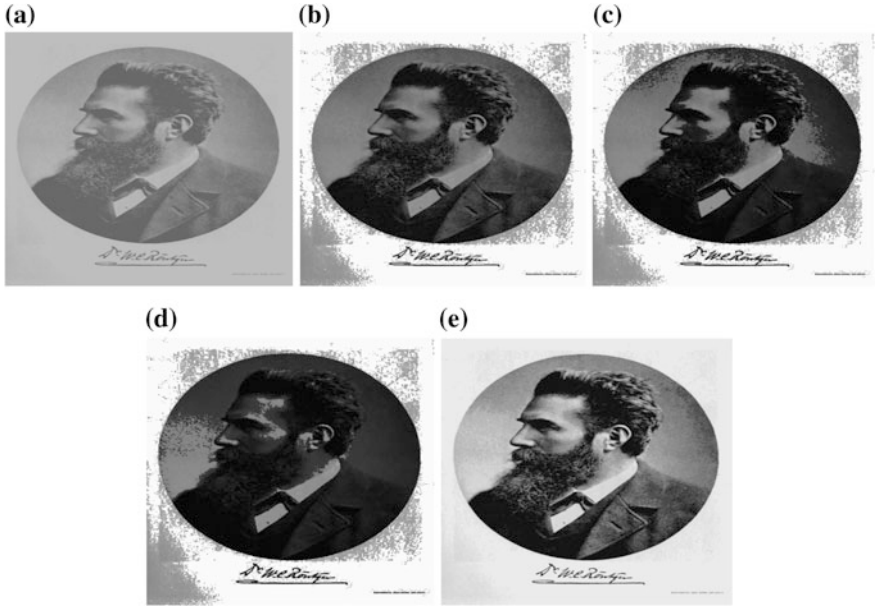


Fig. 26.3 Man image: **a** Original Image, **b** CHE-based enhancement, **c** BBHE-based enhancement, **d** DSIHE-based enhancement, **and e** PSO-ABC hybrid optimization-based enhancement

Table 26.1 Average mean values of PSNR, AMBE, and entropy for different methodology

Metrics	Techniques			
	CHE	BBHE	DSIHE	PSO-ABC hybrid
PSNR (dB)	11.12	14.1	10.21	17.82
AMBE	42.96	24.01	32.73	14.6
Entropy	5.44	5.36	5.38	5.56

average mean value for different types of grayscale test images. To obtain high quality of the image, the modified output images from different methodology should have low AMBE, high value of Entropy and high value of PSNR. It is observed that for the case of the presented hybrid optimization method, the output values comparatively much better with respect to other methodology indicated here.

Conclusions

This paper presents a hybrid PSO-ABC technique that integrates swarm operation from PSO with search dynamics of ABC algorithm. The optimal outputs of PSO methodology through iteration process are again executed by efficient ABC optimization process to produce a better quality of the image in all aspects. This hybrid technique is implemented to minimize the objective function designed as well as preserve the important characteristics of an image. Minimum objective function value provides better image quality. The method is tested with different grayscale images and found to be competitive in comparison with CHE, BBHE, and DSIHE. The work can be further extended to the implementation to enhance the color images, implementation with multi-objective optimization algorithms, designing a more robust optimization function, etc. One important drawback of this hybrid method is the executing time which is a little bit higher than the conventional contrast enhancement techniques. This happens due to the iterative nature of the presented technique to attain the desired output of the tested images. The visual as well as objective comparison shows that this hybrid technique is the suitable substitute for enhancing the contrast of an image in comparison with conventional methods.

References

1. Patil P and Patil AM (2015) Contrast enhancement Technique for Remote Sensing Images. International Journal of Emerging Trends & Technology in Computer Science 4(4): 57–61.
2. Gonzalez RC and Woods RE (2002) Digital Image Processing. 2nd Edition, Prentice Hall.
3. Chen SD (2012) A new image quality measure for assessment of histogram equalization-based contrast enhancement techniques. Digital Signal Processing: 640–647.

4. Gupta P (2016) Contrast Enhancement for Retinal Images using Multi-Objective Genetic Algorithm. IJETED 1(6):8–10.
5. Singh N, Kaur M, Singh KVP (2013) Parameter Optimization In Image Enhancement Using PSO. AJER e-ISSN: 2320-0847 p-ISSN: 2320-0936 2(5): 84–90.
6. Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony algorithm. J. Global Optim. 39 (3): 459–471.
7. Kennedy J, Eberhart RC et al (1995) Particle swarm optimization. IEEE international conference on neural networks, Perth, Australia 4: 1942–1948.
8. Tiedong Z, Lei W, Yuru X et al (2008) Sonar Image Enhancement Based on Particle Swarm Optimization. Industrial Electronics and Applications, 3rd IEEE Conference: 2216–2221.
9. Barik M, Sheta A, Ayesh A (2007) Image Enhancement Using Particle Swarm Optimization. Proceedings of the WCE London, U.K. ISBN: 978-988-98671-5-7 Vol. I.
10. Zhu G et al (2010) Gbest-guided artificial bee colony algorithm for numerical function optimization. Applied Mathematics and Computation 217(7):3166–3173.
11. Draa A and Bouaziz A (2014) An artificial bee colony algorithm for image contrast enhancement. Swarm and Evolutionary Computation 16: 69–84.
12. Raju A, Dwarakish GS, Venkat Reddy D (2013) A Comparative Analysis of Histogram Equalization based Techniques for Contrast Enhancement and Brightness Preserving. International Journal of Signal Processing, Image Processing and Pattern Recognition 6(5): 353–366.
13. Wang C and Ye Z (2005) Brightness Preserving Histogram Equalization with Maximum Entropy: A Variational Perspective. IEEE Transactions on Consumer Electronics 51(4): 1326–1334.
14. Kaur J and Chand O (2012) Comparative analysis for contrast enhancement using histogram equalization techniques. JBRCS, ISSN: 2229–371X 3(5).
15. http://www.sipi.usc.edu/database. Sipi Image Database.

Chapter 27
A Novel Approach for Segmenting WBCs in Smartphones Using Color-Based Segmentation

Ratnadeep Dey, Kaushiki Roy, Debotosh Bhattacharjee, Mita Nasipuri and Pramit Ghosh

Abstract White Blood cell (WBC) is one of the most important blood cells in our body. This cell controls human immune system. The high count and low count of WBC reflect different kinds of disease in our body. Therefore, WBC counting is very important to diagnose several diseases. The pathological process of WBC counting is very time consuming, tedious, and prone to human error. Thus, an automated WBC counting system is highly needed to obtain accurate results. WBC segmentation is very important to make an automated WBC counting system. In our work, we are proposing a color-based WBC segmentation algorithm, which segments WBCs from microscopic images of peripheral blood smear in smart phones. Experimental results show the system to be robust and effective for WBC segmentation. The proposed technique will be very helpful to make an automated WBC counting system and could potentially overcome the errors arising due to manual inspection.

Keywords WBC segmentation · WBC counting · App based segmentation Color-based segmentation · L*a*b color space

R. Dey (✉) · K. Roy · D. Bhattacharjee · M. Nasipuri
Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
e-mail: ratnadipdey@gmail.com

K. Roy
e-mail: kaushiki.cse@gmail.com

D. Bhattacharjee
e-mail: debotoshb@hotmail.com

M. Nasipuri
e-mail: mitanasipuri@gmail.com

P. Ghosh
Department of Computer Science and Engineering, RCC Institute of Information Technology, Kolkata 700015, India
e-mail: pramitghosh2002@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_27

Introduction

White blood cells (WBCs) [1] also known as leucocytes are very important part of our immune system. They help us fight infections by attacking any foreign substances that invade our body. WBC count is a laboratory test, which informs us about the total number of white blood cells present in our body. The normal WBC count [2] in our body is 4,500–11,000 white blood cells per microliter of blood. Abnormally low as well as abnormally high WBC count indicates certain abnormal conditions in our body. A doctor recommends WBC count when he suspects that the patient is suffering from some local or systemic infection. A patient complaining of persistent fever, chills, body aches, sore throat, boils, sepsis, cholera, malaria, etc., is recommended for obtaining the WBC count to find out the hidden infections in the body. Normally WBC count is done in pathological labs. However, the problem arises when huge number of blood samples is to be checked by the pathologists. This manual inspection process is tedious, time-consuming, and erroneous results might incur due to boredom and fatigue. In addition, the manual inspection process also requires expert lab technicians.

In order to eliminate these problems we have developed a smartphone-based app, which takes as input the microscopic images of blood smear, segments WBCs from other blood cells using image processing operations and gives the number of WBCs present in each input peripheral image of blood smear.

Many research works has been done on WBC segmentation from blood smear images. Dorini et al. [3] segmented WBC nucleus and cytoplasm both. In another work [4], WBC was identified but the nucleus and cytoplasm could not be identified separately. Nemane et al. [5] proposed WBC segmentation processes from RGB image of blood smear. The size of the RBCs was determined by granulometry function and depending on the size of RBC the size of WBC was determined. However, the cytoplasm area of the segmented image was greatly over-segmented. The human visual perception simulation cannot be done on RGB images [6]. The color information of WBC nucleus may be an attractive property for color-based segmentation. In the work [7], WBC was segmented using color-based segmentation from HIS images. In another approach, WBC was segmented using shape analysis [8]. The shape analysis was done by roughly boundary detection and the approximate boundary was determined by image thresholding. Another interesting WBC segmentation was done [9] using Teager energy operator. The cytoplasm of WBC was determined by morphological operation. In another work, WBC [10] has been segmented by extracting shape, size and texture features. An automated system for counting WBCs has been done in [11]. Not only WBC segmentation, many more works have been done on identifying several diseases [12–15].

In our work, we have done WBC segmentation using color-based segmentation. The shape of WBC nucleus is not same always as there are five types of different nucleus. In addition, the size of WBC nucleus varies from an immature WBC to a matured one. Therefore, WBC nucleus cannot be segmented properly using the

shape and size. The color of WBC nucleus is always same in Leishman staining. Therefore, color can be a good identifying measure for WBC segmentation.

Methodology

We have collected Leishman stained blood slides of approximately 100 different patients from a nearby laboratory. In our laboratory, we captured microscopic images of these slides using an optical light microscope (OLYMPUSCX21I). The slides were placed under 100x objective of the microscope. These captured microscopic images were then stored in our database. Figure 27.1 shows some of the images in our database.

Algorithm 1 WBC segmentation procedure

1: **function** WBC segmentation technique

2: given an input RGB image I;

3: RGB image is converted to L*a*b color space;

4: chromaticity layers a* and b* are extracted;

5: primitive mathematical operations are performed on the extracted chromaticity layer and resultant image I_r is obtained;

6: image I_r is binarized with a small threshold to obtain binary image I_b consisting of only platelets and WBCs;

7: morphological operations are performed on I_b to remove minute objects like platelets and obtain final image I_w consisting of only WBCs.

8: **end function**

These images were transferred from the desktop to the Android phone and could be stored either in the internal memory or in the secure digital card (sdcard) of the phone. As clear from Fig. 27.1, the platelets and the WBCs got a bluish color due to the Leishman stain whereas the RBCs got a reddish appearance. Due to this drastic color difference amongst the three types of blood cells, we have used color-based segmentation to extract WBCs. The detailed description of this work is presented in

Algorithm 1. The entire code was written in Android using a software named Android Development tool kit.

Input Image

As mentioned earlier microscopic image of Leishman stained peripheral blood smear were used as input and stored in our database as in Fig. 27.1. The images are in RGB color space.

Color Space Conversion

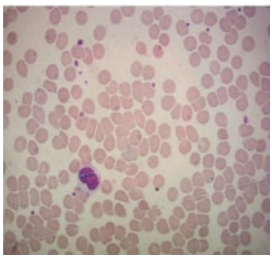
The images captured in digital microscope are usually in RGB color space whose visual segmentation is difficult. For better segmentation, they have been converted to L*a*b [16] color space. Another advantage of using L*a*b color space over RGB color space is the former is device independent unlike the latter which is dependent on the device used (scanner, camera, printer etc.).

In this color space, L* layer represents the lightness information whereas the actual color information is contained in chromaticity layers a* and b* layer of the color space. The a* layer represents the red, green information of a pixel. Positive values in a* layer represents red whereas negative values in this layer represents green color. The b* layer represents blue and yellow information of a pixel. Positive values in this layer represent yellow and negative values represent blue color respectively [17].

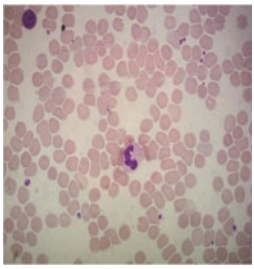
Chromaticity Layer Extraction

The L* layer contains only the brightness information whereas the actual color information is contained only in a* and b* layer. Hence, we extracted only the chromaticity layers a* and b* for further analysis. Figure 27.2 shows the image obtained after extracting the chromaticity layers a* and b* corresponding to the input image Fig. 27.1a.

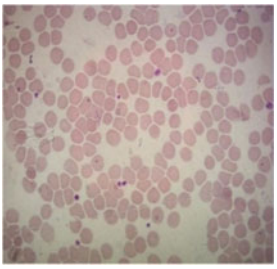
More clear from Fig. 27.2, in a* layer, platelets and WBCs are much whitish in appearance than RBCs whereas in b* layer the platelets and WBCs are much darker than RBCs. Thus, the intensity difference of platelets and WBCs in these two layers is much more than RBCs. This difference in intensity becomes much highlighted for platelets and WBCs when one layer is subtracted from the other. However, for



(a) Image containing RBCs, Platelets and one WBC

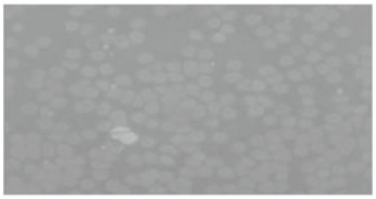


(b) Image containing RBCs, Platelets and two WBCs

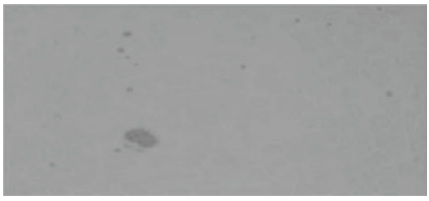


(c) Image containing RBCs and Platelets.

Fig. 27.1 Microscopic images of blood smear



(a) Image obtained after extracting chromaticity layer a*



(b) Image obtained after extracting chromaticity layer b*

Fig. 27.2 Images obtained after extracting chromaticity layers a* and b*

RBCs the intensity difference in a* layer and b* layer was not much. So on subtracting the chromaticity layers, the RBCs get faded out because their intensity difference approximates to zero and are thus eliminated with the other background pixels. Figure 27.3 shows the image obtained after subtracting the chromaticity layer.

Fig. 27.3 Image obtained after performing primitive mathematical operation on chromaticity layers

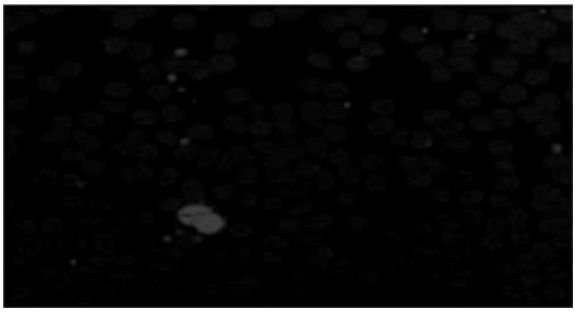


Image Binarization

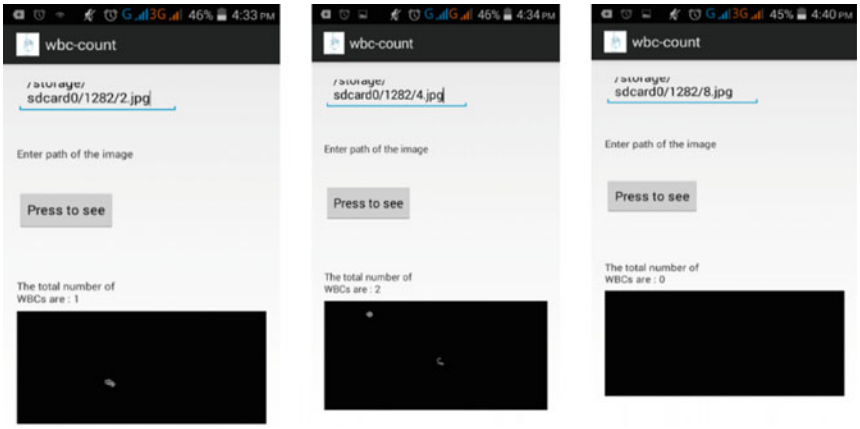
The former step was followed by image Binarization. A binary image is a digital image where there are only two possible intensities for each pixel. The required pixels are set to intensity value 1 whereas the background and other unnecessary pixels are set to 0.

In our work, the main object of interest is WBCs. As clear from Fig. 27.3, the WBCs and platelets are prominently visible while others are not. Therefore, we extracted these two objects by image binarization.

The images obtained from previous step were binarized with a very small threshold value approximately 0.144 to eliminate RBCs. This small threshold value suffices in eliminating RBCs since the RBCs gets a very dimmer appearance as a result of previous step which is clear from Fig. 27.3. Thus, the RBCs were eliminated along with other background pixels leaving behind image consisting of platelets and WBCs.

Platelet Removal

Platelets are very minute structures and they were eliminated using morphological [18] operations along with other artifacts leaving behind the desired WBCs whose size are much larger than platelets. Figure 27.4 shows the outputs generated by our app corresponding to the input image Fig. 27.1a, b and c after performing all the previously mentioned image processing operations.



(a) App output corresponding to figure 1a (b) App output corresponding to figure 1b (c) App output corresponding to figure 1c

Fig. 27.4 Output given by our developed app

Result and Discussion

Our developed app consists of four major components namely text-field, text-view, button, and image-view. The path of the image to be analyzed is given in the text-field named "**Enter path of the image**". The button named "**Press to see**" is clicked next. When this button is pressed, the image whose path is mentioned in the text-field is fetched, further segmentation takes place on it, and the final image consisting of only WBCs is displayed in the image-view along with its count.

We have tested the app output on 100 different input images and checked whether it segments the WBC correctly. We got appropriate results in most of the cases. Table 27.1 shows the WBC count per input image obtained by manual inspection and the one given by our app for 20 different input images. Our system gives an overall accuracy of 91.66%.

Table 27.1 Comparison between WBC counts obtained by manual inspection and the one given by our app

Serial no.	WBC count per input image obtained by manual inspection	WBC count per input image given by the developed app
1	1	1
2	2	2
3	0	0
4	3	3
5	1	1
6	0	0
7	1	1
8	3	5
9	1	1
10	4	4
11	1	1
12	1	1
13	3	3
14	1	1
15	3	4
16	2	2
17	2	2
18	1	1
19	2	2
20	4	4

Conclusion

White Blood cell counting is very important for the diagnosis of several diseases. The manual processes for counting WBC are error prone. These errors can lead to wrong diagnosis of diseases. An automated system for counting WBC may remove the errors of manual counting processes. WBC segmentation is very important to make an automated system for WBC counting. The shape and size are not always correct identification parameters of WBCs. Because shape and size of WBCs are not always same. WBCs can change their shape and size under different circumstances. However, WBCs are always in bluish appearance. In our proposed work, we have segmented WBC from microscopic image using color-based segmentation. Our developed app correctly segments WBCs from peripheral images of blood smear. This app could eliminate the dependency on expert lab technicians and would help to avoid huge rush to pathology centers. This would be highly beneficial for people residing in outskirts where pathological centers and expert pathologists are not readily available.

Acknowledgements The authors are grateful to DBT, Govt. of India (Letter No. BT/PR8456/MED/29/739/2013) for financial support.

References

1. Martin S. Blumenreich; "The White Blood Cell and Differential Count"; Chapter 153; Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition; Boston: Butterworths; 1990.
2. Hansen LK1, Grimm RH Jr, Neaton JD. "The relationship of white blood cell count to other cardiovascular risk factors." Int J Epidemiol. 1990 Dec; 19(4):881–8.
3. Leyza Baldo Dorini, Rodrigo Minetto and NeucimarJerˆonimoLeite: "Semi-automatic white blood cell segmentation based on multiscale analysis" IEEE Journal Of Biomedical And Health Informatics, Vol. 17, No. 1, January 2013.
4. W. Shitongand, W. Min, "A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 5–10, Jan. 2006.
5. J.B. Nemane, V. A. Chakkarwar: "A Novel Method of White Blood Cell Segmentation and Counting", International Journal on Advanced Computer Engineering and Communication Technology Vol-1 Issue:1: ISSN2278–5140.
6. H. D. Cheng, X. H. Jiang, Y. Sun, J. Wang. Color image segmentation: advances and prospects. Pattern Recognition, 2001, 34(12):2259–2281.
7. Jun Duan, Le Yu; "A WBC SEGMENTATION METHORD BASED ON HSI COLOR SPACE"; Proceedings of IEEE IC-BNMT 2011.
8. Q. Liao and Y. Deng, "An accurate segmentation method for white blood cell images," in Proc. IEEE Int. Symp. Biomed. Imag., 2002, pp. 245–248.
9. B. Kumar and T. Sreenivas, "Teager energy based blood cell segmentation,"in *Proc. Int. Conf. Digital Signal Process.*, 2002, vol. 2, pp. 619–622.
10. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, "Blood smear analyzer for white blood cell counting: A hybrid microscopic image analyzing technique", Applied Soft Computing, Elsevier, January, 2016.

11. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, "An Automatic Non-invasive System for Diagnosis of Tuberculosis", Applied Computation and Security Systems, Springer India, 2015.
12. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, "Automation in Sputum Microscopy: A Hybrid Intelligent Technique in Diagnostic Device Automation", Handbook of Research on Advanced Hybrid Intelligent Techniques and Applications, IGI Global, 2015.
13. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, "Computer Intelligence in Healthcare", Handbook of Research on Computational Intelligence for Engineering, Science, and Business, 2013.
14. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, "Medical Aid for Automatic Detection of Malaria", arXiv preprint arXiv:1312.0940, 2013.
15. Pramit Ghosh, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu, "Automatic White Blood Cell Measuring Aid for Medical Diagnosis", International Conference on Process Automation, Control and Computing (PACC), IEEE, 2011.
16. Application note "Measuring color using Hunter L, a, b versus CIE 1976 L*a*b*". www.hunterlab.com (Accessed on 4/9/2015).
17. Connolly C and Fliess T, "A Study of Efficiency and Accuracy in the Transformation from RGB to CIELAB Color Space", IEEE Transactions on Image Processing, vol. 6, no. 7, July 1997.
18. "Mathematical morphology", Chapter 6; http://www.cs.uu.nl/docs/vakken/ibv/reader/chapter6.pdf (Accessed on 9th September, 2016).

Chapter 28
Some Issues Related to Phone Recognition and Language Identification Using Phonetic Engine

Sushanta Kabir Dutta and L. Joyprakash Singh

Abstract The introduction of a phonetic engine in the literature is that it is a system which transforms a speech signal into some symbolic form. Phone recognition is a primary task of a PE. Among the various other applications of PE, one very useful application is the Language Identification (LID). This chapter discusses some issues pertaining to the use of PE in phone recognition as well as in language identification. Here, we have used PEs for three Indian Languages: Manipuri, Assamese and Bengali, in building the LID system. These languages are widely spoken in the Northeastern region of India. In the development of PEs, the International Phonetic Alphabet (IPA) symbols are used in the data transcription process. In modeling the phonetic units Hidden Markov Models (HMM) have been used in the training phase. These trained HMMs are then used in phone recognitions leading to the identification of language(s) of unknown test utterances. The overall phone recognition accuracies reported by the existing PEs for the above selected languages are 62.11% for standard Manipuri language, 59.0% for Kakching Dialect of Manipuri, 43.28% for standard Assamese and 48.58% for Bengali language. Automatic LID is possible using a set of PEs in testing unknown utterances due to the language bias of these systems. Various levels of identification rates reported in some LID tasks carried out with PEs are discussed here to look into the issues belonging to it.

Keywords Phone recognition · Language dependency · LID · PA · IDR · IPA HMM

S. K. Dutta (✉) · L. J. Singh
North-Eastern Hill University, Shillong, India
e-mail: sushantatzp@gmail.com

L. J. Singh
e-mail: jplairen@nehu.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_28

Introduction

The term Phonetic Engine (PE) was introduced in [4, 5]. The PE transforms acoustic phonetic information present in a speech signal into some symbolic form. Therefore, it produces a sequence of symbols that uses no language constraints such as lexicon, syntax, etc. The detailed description of PE is available in [19]. The symbols are chosen in a way so that they can capture all kinds of phonetic variations in the speech signal. Existing PEs implemented for Indian languages produce syllable-like units as the output such that these units are most basic in sound production system. The International Phonetic Alphabet (IPA) symbols are used as the underlying sub-word units in the process. These sub-word units act as one level of abstraction for continuous information available in the signal. Thus these symbols may be considered to be some form of the quantized version of the information available in the speech signal [14].

The PE is used as a front-end module in many applications, viz. Speech Recognition (SR), information retrieval (IR) and Language Identification(LID). One distinct advantage of PE is that it uses an open vocabulary set unlike conventional SR systems. It incurs a low error for Out Of Vocabulary (OOV) words. While conventional SR systems are used for languages with written form, the PE can be used with a database containing only spoken form of data. Another advantage is the possibility to make a PE for a new language with similar sounds using minimum speech data from the language since it uses very low level of representation like the phonetic units of the speech data. Moreover, the concept of PE is more suitable for speech in Indian languages due to use of phonetic nature of characters used during production as well as writing [19]. The existing PEs developed for Indian languages use syllable like sub-word units and Hidden Markov Model (HMM) is used to model each of the sub-word units in building the system [18, 19]. The motivation for using syllable like units as sub-words is that such units are directly represented in text symbols in most of the Indian languages.The PEs use the IPA symbols since IPA provides one symbol for each distinctive sound. A symbol more often comprises of one or more elements of two basic types, letters and diacritic marks [8]. The letters represent basic sound units while the diacritics are small markings placed around the IPA letter to show a certain alteration or specific description of pronunciation.

In the present study we are considering the existing PEs of the three different languages Manipuri, Assamese and Bengali widely spoken in the North Eastern part of India. In Manipuri, we have two PEs, one for the standard Manipuri language (the officially used version) and another for a dialect known as ‘Kakching’ spoken by the people in Kakching area of Southeastern part of Manipur. However, for Assamese and Bengali languages, we have the PEs available for only the standard version of languages (being officially used). The phone recognition accuracies reported by the Manipuri PEs are 62.11% for the standard version [3] and 59.0% for Kakching dialect [1]. Similarly, the accuracies reported by Assamese and Bengali PEs are 43.28% and 48.58% respectively [3].

The other sections of the chapter are organized in the following way: The Sect. "A Brief Discussion on Language Identification Cues" provides a short discussion on language identification cues. The Sect. "An Overview of Language Identification Using Phonetic Engine" discusses the overview of the process of language identification using PEs. The Sect. "Issues with Phonetic Engine in Phone Recognition and Language Identification" details the issues with PE in language identification. The Sect. "Summary with Conclusion" contains the summary with conclusion for the future directions of the work.

A Brief Discussion on Language Identification Cues

The task of automatic language identification(LID) is the process of identifying a particular language from a set of languages. In the literature, LID is broadly classified into two main categories, Explicit LID and Implicit LID systems [11]. Explicit LID requires a segmented and phonetically labelled speech corpus while Implicit LID uses digitized speech samples with corresponding true identities of the language. The LID system using PEs is an Explicit system [2]. A few LID cues are presented below.

Acoustic: It is a physical characteristics of the speech signal described by frequency, time and intensity information of the speech. These are represented as a sequence of feature vectors where each individual feature vector corresponds to acoustic information for a particular time frame. Acoustic information is one of the most primitive forms of information obtained by speech parameterization [16, 17]. The widely used speech parameterization techniques are Linear prediction coefficients(LPC), Mel-frequency capstral coefficients(MFCC), Perceptual Linear Prediction(PLP) and Linear prediction Cepstral coefficients (LPCC) [12].

Phonotactic: There are various phonological factors governing the distinctiveness of a particular language. Some of these are the phone sets and phonotactic constraints of a language. The word 'phonotactic' refers to the rule that govern the different combination of phones or phonemes in a language [15]. Different languages may have different allowable combination of sequence of phoneme combinations. For example, Japanese has strict phonotactic constraint prohibiting a consonant following another consonant, while English does not have such rules. Thus, phonotactic information may be useful in capturing some dynamic nature of speech signal usually lost in feature extraction [13].

Vocabulary: Conceptually, the most important difference among the languages is the different set of words they use. Therefore, the vocabulary differs [13]. This is an important cue.

Prosodic: The stress, intonation and rhythm are all important elements used within the prosodic structure of a spoken utterance. The manner in which these are incorporated varies from language to language. For example tonal languages like Mandarin has a very different intonation characteristics than a stress language like English [13].

There exposed a significant importance of acoustic-phonetic approaches in language identification [6, 9, 20]. To get accurate estimation of the information sources some form of detailed modeling is necessary [20]. In PE-based language identification systems [2, 7], continuous density HMMs are used to build language-dependent Phone Recognizers (PRs). An acoustic model is then created by using audio recordings of speech with their corresponding transcriptions which are later complied to get statistical representations of the phone units in the PRs.

An Overview of Language Identification Using Phonetic Engine

Acoustic analysis is used as the first step in the development of an LID system using PEs [3]. Acoustic analysis deals with slicing the speech signal into successive frames of around 20–25 msec duration with 10–15 msec overlap, where each frame is later on represented as a set of feature vectors extracted therefrom. In [2, 3], frames with 25 msec duration and 10 msec overlap are used. Each frame is represented by a 39-dimensional feature vector containing one energy coefficient, 12 MFCCs, 13 delta and 13 acceleration co-efficients. Then a 5-state prototype left to right HMM is used for modeling of each phonetic unit. First and last states of the model are non-radiating while the remaining three are radiating states. The global mean and variance of HMMs per state are calculated using the predefined prototype along with acoustic vectors and transcriptions. Once an initial set of models are built, the optimal HMM parameters are re-estimated to get the acoustic model. The acoustic model is the backbone of a PE. A total of 30 phonetic units including a silence are used in Manipuri language PE while 34 phonetic units are used in both Assamese and Bengali PEs [2].

The extracted feature vectors from a test utterance are then compared with the 'acoustic model' to estimate the 'acoustic likelihood' scores. This is illustrated in Fig. 28.1 below. Here the PEs of three different languages, viz. L1, L2 and L3 are used to build the LID system.

The highest likelihood score emanating from a PE is used to identify a language. The system can be extended to contain any number of PEs as required.

The performance of an individual PE is evaluated as [14]

$$PA = \frac{N - D - S - I}{N} \times 100\% \tag{28.1}$$

where, PA is percentage accuracy, N is the number of words in the test set, D is the number of deletions, S is the number of substitutions and I is the number of insertions. Thus, PA provides the accuracy rate of a particular PE. While using in LID tasks, the identification rate (IDR) of the overall system is estimated as

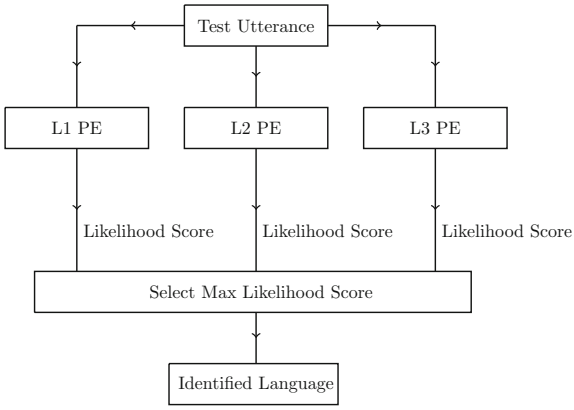


Fig. 28.1 LID system

$$IDR = \frac{n}{N} \tag{28.2}$$

where 'n' is the number of correctly identified utterances and 'N' is the total number of utterances used in a particular language L.

Issues with Phonetic Engine in Phone Recognition and Language Identification

Language dependency of a PE leads to its application in LID tasks [7]. However, the data collection process plays a crucial role in the development of PEs . It is found that the phone recognition accuracy achieved by a PE is less if data are collected using telephone channels, as compared to a phone recognizer (PR) for which collection is done using broadband channels. For example, the phone recognition accuracies achieved by Kakching dialect PE is 59.0% [1], standard Manipuri, Assamese and Bengali language PEs are 62.11%, 43.28% and 48.58% respectively [2]. However, the phone recognition accuracy achieved by a PR in the TIMIT database with same settings of HMM recognizers is 79.0% [10]. This may be summarized as due to the channel noise in the data.

Another observation is that the phone recognition accuracy is also dependent on the mode in which the data are being collected. Some experiments on the Manipuri PE were carried out with the speech data collected from male and female speakers in Read, Conversation and Lecture modes. During this experiment, three types of speech databases were prepared consisting of only male speakers, only female speakers and both male and female speakers together for Read and Lecture modes. However, just one type of database could be prepared in the Conversation mode of speech data which contained only male speakers. The Table 28.1 below shows the various levels of phone recognition accuracies achieved in all these cases.

Table 28.1 Performance analysis of Manipuri PE

Sl no	Mode	Type of speaker	Accuracy
1	Conversation	Male	64.71
2	Lecture	Female	69.71
3	Lecture	Male	72.50
4	Lecture	Both male and female	68.62
5	Read	Female	74.21
6	Read	Male	72.07
7	Read	Both male and female	70.49

Table 28.2 Phone duration measurements

Phone	Standard Manipuri	Kakching Dialect
aa	745530	1093180
ea	380774	590853
e	773230	667588
i	638225	980036
o	731329	922449
u	598389	869011

One observation from here is that with the phone recognition rate is always higher when the database contains only one type of speakers (either male or female) rather than both together. However, the overall accuracy with all the three modes being used is only 62.11%.

Read and Lecture mode data are more formal mode of speech data than the conversation mode. The recognition accuracy is high in these two modes due to more clarity in the utterances.

The transcriptions of the data have significant role in the PE performance and therefore language recognition. For example, to distinguish between two varieties of the Manipuri language—the standard Manipuri and the Kakching Dialect, the durations of phones can be measured with respective PEs. This provides important information about the two varieties of the language. Table 28.2 shows the exact phone durations measured in units of 10^{-7} s for the above two versions of Manipuri Langauge. Only vowels are considered as the consonant durations are insignificant.

The comparison of the phone durations can be visualized as in Fig. 28.2.

The number of speakers used in the database also influences the PE phone recognition tasks. With 16 speakers the Manipuri PE overall accuracy is 62.11%, while for 31 speakers the Assamese PE overall accuracy is 43.28%. Similarly, the overall accuracy of the Bengali PE is also 48.50% for 43 speakers. This is due to speaking variations by the speakers which influence the transcription process.

However, when used for LID tasks the system comprising of a set of PEs the IDR is much higher even though the individual PE's PA is not so good. This is a good sign

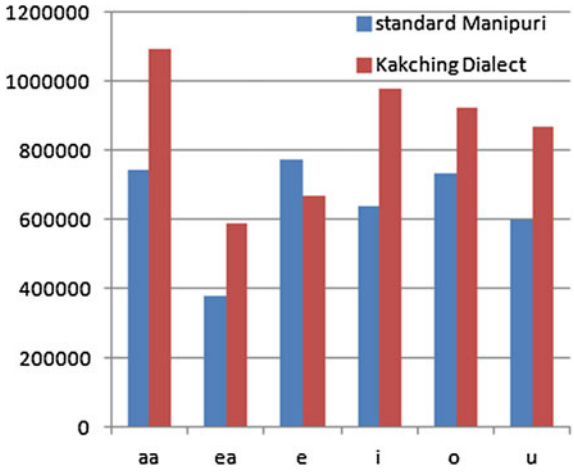


Fig. 28.2 Comparison of phone durations: standard Manipuri and Kakching dialect

Table 28.3 IDR by LID system with 3 PEs

Sl no	Language	IDR by acoustic likelihood
1	Manipuri	99%
2	Assamese	99%
3	Bengali	100%

that the PEs are worth in use for LID tasks. Table 28.3 below shows the IDR for the sytem with three different PEs, Manipuri, Assamese and Bengali. The reason behind this is that the overall system performance is a statistical estimate of the maximum likelihood of an unknown utterance being coming from one of the PEs, which is to some extent independent of the PA of individual PEs. Table 28.3 shows the IDR with 100 utterances each of which is 6–7 s duration from each of Manipuri, Assamese and Bengali languages being used as test utterances.

The native speaker dependency is another issue with the PEs. If the collected database is from non-native speakers of a language the phone recognition accuracies of the PE will change [7, 14]. This is due to the speakers may use some of the phones from their native language. Therefore, the IDR in an LID task will subsequently will vary.

Summary with Conclusion

Here, we have discussed a few issues related to the phone recognition as well as language identification by PEs. The primary task of the PE is phone recognition and the higher the level of accuracy is, better the PR it would be. While using in language identification, a set of good accuracy PRs are usually expected, since the common

understanding is that it will result an LID system with better accuracy. However, our experiments with the present LID system shows that the IDR of the system does not much depend on the accuracies of individual PRs (here the PEs). This is due to considering the maximum acoustic likelihood in case of the identification of the unknown test utterances. Another point is that different languages use different phone sets. Therefore, such an LID system can result in sufficient IDRs.

However, the IDR can be tested with one PE for different varieties of a single language. That is the PE of a single language can be used to observe how it identifies different dialects of the language. For this the transcribed data from all the dialects will be required to train the system.

Acknowledgements The authors like to acknowledge the assistance received from the students Salam Nandakishore and Y Chandrika Devi in carrying out the experimental works.

References

1. Devi, Y.C.: Acoustic-Phonetic Feature based Manipuri Dialect identification: standard and Kakching Dialect. Master's thesis, North eastern Hill University (2016)
2. Dutta, S.K., Nandakishore, S., Singh, L.J.: Development of language identification system using phonetic engine. In: Proceedings of I3CS'15 (2015)
3. Dutta, S.K., Nandakishore, S., Singh, L.J.: Development of manipuri phonetic engine and its application in language identification. IJETR 3(issue 8), 200–203 (August 2015)
4. Eswar, P.: A Rule-based Approach for Spotting Characters from Continuous Speech in Indian Languages. Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras (1990)
5. Gangashetty, S.V.: Neural Network model for Recognition of Consonent-Vowel units in multiple languages. Ph.D. thesis, Department of Computer Science and Engineering, IIT Madras (2004)
6. Hazen, T.J., Zue, V.W.: Automatic language identification using a segment based approach. In: Eurospeech (1993)
7. F IIIT(Hayderabad), IIT(Kanpur), Thapar University, IIT(Guwahati), Tezpur University, North Eastern Hill University, RIT(Kottayam), DA-IICT, IIT(Hyderabad), IIT(Kharagpur): Development of prosodically guide phonetic engine for searching speech database in indian languages. Tech. rep., DEITY, Government of INDIA (2015)
8. International Phonetic Association (ed.): Handbook of International Phonetic Association. Cambridge University Press (1999)
9. Li, K.P.: Automatic language identification using syllabic spectral features. In: International conference on Acoustic, Speech and Signal processing (1994)
10. Lopes, C., Perdigao, F.: Phone Recogniton in TIMIT database. Prof. Ivo ipsid (2011)
11. Nagesh, A., Sadanadam, M.: Language identification using ergodic hidden markov model. International Journal of Advanced research in Computer Science and software Engineering 2(11), 297–301 (2012)
12. Rabiner, L., Juang, B.: Fundamentals of Speech Processing. Prantice Hall (1993)
13. Rong, T.: Automatic Speaker and Language Identification. Ph.D. thesis, Nayang Technical University (2006)
14. Sarma, B.D., Sarma, M., Prasanna, S.R.M.: Development of assamese phonetic engine: Some issues. In: Annual IEEE India Cnference (INDICON) (2013)
15. Schultz, T., Kirchoffe, K.: Multilingual Speech processing. Elsevier (2006)

16. Wang, L.: Automatic Spoken Language Identification. Ph.D. thesis, Faculty of Engineering, The University of New South Wales (2008)
17. Wong, K.Y.E.: Automatic Spoken Language Identification using Acoustic phonetic Speech information. Ph.D. thesis, School of Electrical and Electronics system Engineering, Queensland University of Technology (2001)
18. Yegnanarayana, B., Gangashetty, S.V.: Machine learning for speech recognition- an illustartion of phonetic engine using hidden markov models. In: Frontiers of Interface Between Statistics and Sciences (2009)
19. Yegnanarayana, B., Gangashetty, S.V., Rajendran, S., Murty, K.S.R., Dhananjaya, N., Guruprasad, S.: A phonetic engine for indian languages. In: ICON (2009)
20. Zissman, M.: Comparision of four appraches to automatic language identification of telephone speech. IEEE trans. Speech and Audio Processing 4(1), 33–44 (1996)

Chapter 29
An Image-Based Approach for the Classification of Dementia from Brain Magnetic Resonance Images

Ibanylla Sari, Gaurav V. Bhalerao and Sudip Paul

Abstract White Matter is considered to be an important biomarker for detecting various brain disorders. Many studies have documented the involvement of white matter and its degeneration with brain disorders. An image-based approach is developed to detect the white matter from T1-weighted brain magnetic resonance images. K-means clustering algorithm is used to perform 2D and 3D segmentation of white matter from all the input images. Backpropagation neural network classifier is trained with area, volume features, and Open Access Series of Imaging Studies datasheet features to classify demented and non-demented classes. The proposed approach reports highest classification accuracy of 90% using backpropagation neural network. Extracted features such as area and volume along with datasheet features such as Estimated Total Intracranial Volume, Normalized Whole-brain Volume, Mini-Mental State Examination score and Atlas scaling factor are found to be important for the detection and classification of dementia.

Keywords White matter Dementia · K-means clustering · Magnetic resonance imaging · Backpropagation neural network

I. Sari (✉)
Manipal Institute of Technology, Manipal University, Manipal 576104, Karnataka, India
e-mail: ibasari1991@gmailo.com

G. V. Bhalerao
National Institute of Mental Health and Neurosciences, Bangalore 560029, Karnataka, India
e-mail: gaurav_bhalerao005@rediffmail.com

S. Paul
School of Technology, North-Eastern Hill University, Shillong 793022, Meghalaya, India
e-mail: sudip.paul.bhu@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_29

Introduction

White Matter (WM) is considered to be an important biomarker for detecting various brain disorders because degeneration of white matter is a symbol of brain disorders [1]. WM is particularly required for long distance information transfer between various brain regions. Quantification of anatomical brain tissues such WM is very important for the clinical diagnosis of neurological diseases, as it connects various gray matter structures in the brain to each other and perform selectrical activity between neurons [2]. Dementia is brain abnormality which occurs when brain is damaged by disease. It includes broad category of symptoms which lead to a severe cognitive decline in the patients. It is generally an age-old disease and mostly found in older adults. The symptoms of dementia are different and serious than normal aging, hence not to be confused with normal aging. Dementia patients unlike normal older adults can show severe memory loss. Demented people may not even recognize familiar places, faces, directions, languages, objects, which is not the case with normal aging. According to WHO report in the entire world, there are 7.7 million cases of dementia every year. The report estimates that a number of patients will be tripled by coming 2030, also according to Indian report of dementia, there are on average 3.7 million cases in 2010. It is expected to be doubled by coming 2030 with significant increase in cost [3–5].

Many novel techniques in previous literature have been reported that the 2D as well as 3D segmentation of WM is possible. As manual segmentation of WM is a time consuming, techniques such as conventional thresholding approach was introduced, but it failed to provide good quality segmentation due to the nonuniform distribution of pixel intensities in MR images. K-means clustering algorithm has been popular for performing MR image segmentation and classification for extracting the region of interest. Its implementation is popular and unique because of the simplicity of the algorithm, reduced computational time and higher accuracy [6]. Several studies have been carried out to analyze the effect of various brain abnormalities due to white matter hyperintensities in MRI. WM hyperintensities are found to be in strong association with stroke dementia and cognitive decline. White matter atrophy has always been a vast area of research in cases of traumatic brain injury, depression, dementia, multiple sclerosis, schizophrenia, hydrocephalus, Alzheimer's disease, and healthy aging [7]. Cognitive disorders such as Parkinson's disease, dyslexia, autism, AD, multiple sclerosis are found in association with the structural and functional decline in the largest white matter structure known as Corpus Callosum [8]. MR images of the brain for dementia patients have shown the importance of white matter features to classify healthy controls and demented patients. It has been investigated that demented brains have reduced white matter structure than the normal patients. Since accurate feature extraction depends upon the quality of segmentation from MR images, several supervised and unsupervised algorithms have been proposed. Several neuroimaging tools, such as SPM, FSL, and Freesurfer provide automated segmentation routines for extracting WM from brain MR images [6]. Open Access Series of Imaging Studies (OASIS) is an

open-source database which has been used widely for investigating the neuroimaging studies on dementia [9, 10]. Several feature classification approaches such as VBM, PCA has reported classification of dementia with accuracies more than 80%. Classifiers such as neural network, SVM, multilayer perceptron have reported classification accuracies of more than 85% with OASIS data. Morphological and textural features of Corpus callosum have reported significant contribution in detecting dementia from brain MR images [6].

The increasing threat of brain disorders imposes the need to design automated approaches for early detection. Though imaging modalities are available, misdiagnosis of these disorders can lead to delay in treatment and recovery of patients. Diagnosis with only visual inspection can lead to subjective errors and hence can be supported with automated techniques. Keeping this goal in mind, this work uses white matter features and OASIS dataset features as biomarkers for detecting dementia. An automated approach of 2D and 3D segmentation of WM has been proposed using k-means clustering algorithm. 3D segmentation approach includes semi-automated segmentation routine with MATLAB and 3D slicer interface (MATLAB-bridge) which ensures a better quality of segmentation [11]. This work uses a k-means clustering algorithm for segmentation and backpropagation neural network for classification of input MR images. An image-based approach is developed to detect the white matter from T1 weighted brain magnetic resonance images. Segmentation of white matter is performed for all input images using k-means clustering algorithm. From each segmented white matter, morphological features are extracted. Backpropagation neural network classifier is then trained and tested with the extracted features to evaluate classifier performance.

Methodology

A major analysis of the project including segmentation, feature extraction, and classification is performed using MATLAB [12–14]. Input MR Images are collected from OASIS. They provide a dataset which consists of a longitudinal collection of 150 patients of both normal and abnormal from aged 60 to 96, including both male and female [9, 10] (Fig. 29.1).

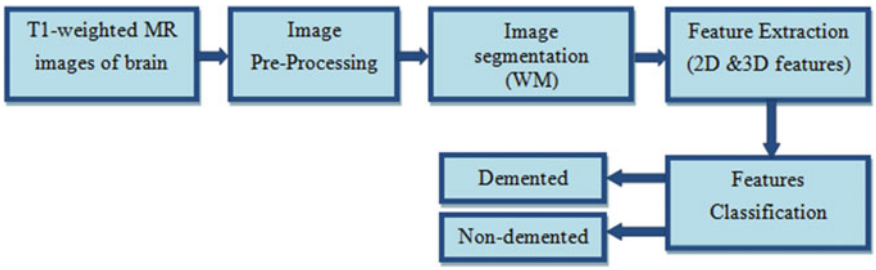


Fig. 29.1 MR Image analysis for classification

Image Preprocessing

Using MATLAB, all images are converted to grayscale and made equal in size. Normalization and median filtering operations are then performed for each brain scan.

2D and 3D WM Segmentation

The 2D morphological feature extracted is the area of WM using MATLAB; whereas in the 3D morphological feature, volume is extracted using 3D slicer. Editor tool and MATLAB interface of 3D slicer (MATLAB-Bridge) are used to perform 3D segmentation. All segmented slices are built together to form a 3D volume of WM. Segmentation task is performed using an unsupervised k-means Clustering algorithm. The algorithm clusters N input data points into K different clusters. The algorithm can be implemented in following steps (Fig. 29.2):

1. Decide the number of K and initialize the centroids
2. Find the distances between N data points and centroids
3. Group the data points into K clusters depending upon the minimum distances
4. Update the centroids until there is no movement of clusters.

As Euclidean distance is used as a distance measure, the objective function J is given as

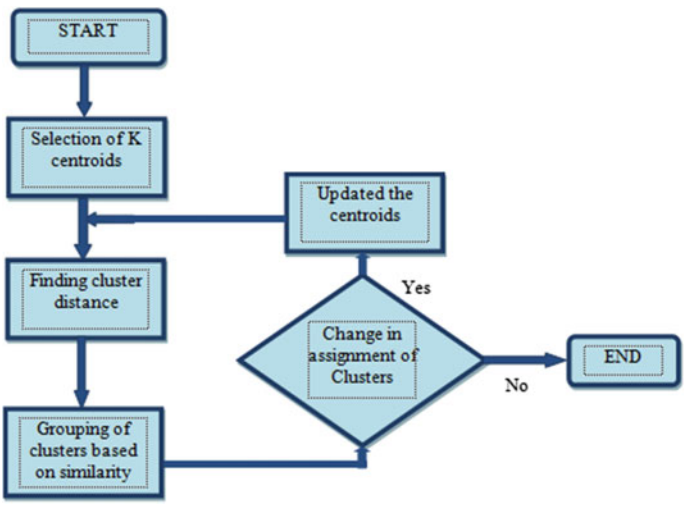


Fig. 29.2 K-means clustering algorithm

$$J = \sum_{q=1}^{k} \sum_{p \in c_q} \left\| X_p - \mu_q \right\|^2 \tag{29.1}$$

where c_q is the qth cluster and μ_q is the centroid of cluster c_q [8].

Extraction of WM Features

Along with area and volume, some more features are also considered to improve classification accuracy. These features provided in the OASIS datasheet, eTIV, nWBV, MMSE, and ASF [15].

(i) Estimated Total Intracranial Volume (eTIV): eTIV was computed by scaling the manually measured intracranial volume of the atlas by computing the determinant of the affine transform connecting each individual's brain to the standard atlas.
(ii) Normalized Whole-brain Volume (nWBV): This is total intracranial volume. The image is initially segmented into 3 different structures such as GM, WM, and CSF. Then by using maximum likelihood estimates, the voxels are classified to each tissue class. The nWBV is then calculated as the amount of all voxels within each tissue class.
(iii) Mini-Mental State Examination score (MMSE): This test is a 30-point questionnaire that is used for the assessment of cognitive decline. A score greater than 24 (out of 30) indicates healthy cognitive performance. Scores below 24 indicate mild, moderate, and severe conditions in dementia. For example, mild dementia has an average score of 19–23 points, moderate has 10–18 points, and severe dementia shows scores less than 9 points.
(iv) Atlas Scaling Factor (ASF): Derived from transformation information generated while converting the native space into atlas space.

Features Classification by Design of Classifier

Backpropagation neural network (BPNN) is used as a classifier for the classification of dementia. It is a supervised learning algorithm which calculates the gradient of the loss function, and then the optimization technique minimizes this loss function to update the weights and biases [16–19]. The network is designed with three different layers, i.e., the input layer, hidden layer, and the output layer. The input layer consists of 6 neurons with 30 neurons in the hidden layer and 2 output neurons in the output layer. This network is trained such that it can it can identify the input data if it is demented or non-demented with respect to the given features [11]. Two target classes, demented and non-demented are given binary values of 1 and 0 respectively. Along with classification accuracy, sensitivity and specificity

values are also estimated. Sensitivity is a measure of true positive rate (demented patients correctly classified in class 1), while the specificity is a measure of true

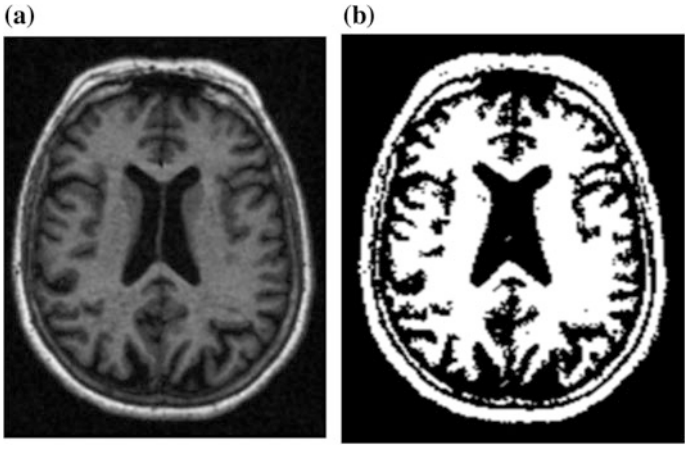


Fig. 29.3 White matter segmented from 2D MR images **a** Input Image, **b** WM segmentation from input image

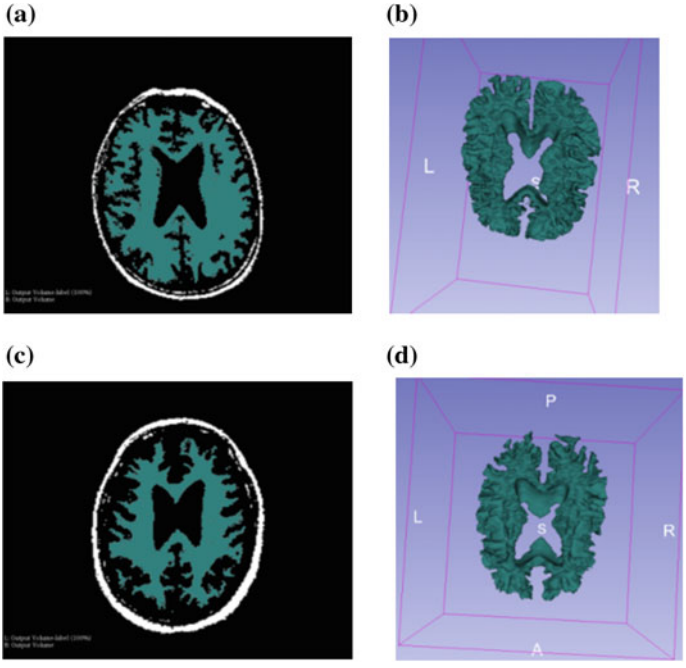


Fig. 29.4 2D and 3D segmentation of WM using MATLAB and 3D slicer interface: **a**, **b** 2D and 3D segmented WM for normal subject, **c**, **d** 2D and 3D segmented WM for demented subject

negative rate (non-demented subjects correctly classified in class 0) of classification. The sensitivity and specificity are computed as

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{29.2}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{29.3}$$

where TP = true positives, TN = true negatives, FN = false negatives (demented patients incorrectly classified as normal), and FP = false positives (normal subjects incorrectly classified as demented). The classifier performance is analyzed by plotting Receiver Operating Characteristics (ROC) plot and confusion matrix. ROC plot is generated by calculating the true positive and false positive rate, while confusion matrix shows the accuracy of the classifier along with a number of correct and incorrect classification of input data [8, 11] (Figs. 29.3, 29.4).

Results

All 42 subjects are quantified and tabulated. The mean values of all the features are given in Table 29.1 From the mean values of WM features area and volume, it is evident that demented subjects have a loss in their WM structure as compared to non-demented or normal subjects [11].

The confusion matrix and ROC plot generated for classifying the data with 35 subjects are shown in Fig. 29.5a and b respectively. These plots are produced for a hybrid feature vector consisting of 2 WM features and 4 features from OASIS datasheet. Out of the 16 normal samples, 14 are correctly detected as normal and 2 samples are misclassified into demented class. While out of the 19 demented samples, only 1 sample is incorrectly classified into normal class, rests of the 18 samples are correctly classified into demented class. As a result, a trained neural network gives out the results with an accuracy of 90% [11]. The ROC plot shows the relative balance between true positive rate and false positive rate. The area under the ROC curve shows the true positive rate (sensitivity) and false positive rate (1-sensitivity) of the classifier. A perfect plot shows points in the upper left corner. For this analysis, ROC plot shows that the network performs well.

Table 29.1 Quantified value of white matter of 42 subjects

Feature (Mean) / class	Area of WM	Volume of WM	MMSE	ASF	nWBV	eTIV
Demented	168814	225230	25	1.186	0.73	1500.982
Non-demented	180175	289045	29	1.193	0.73	1520.399

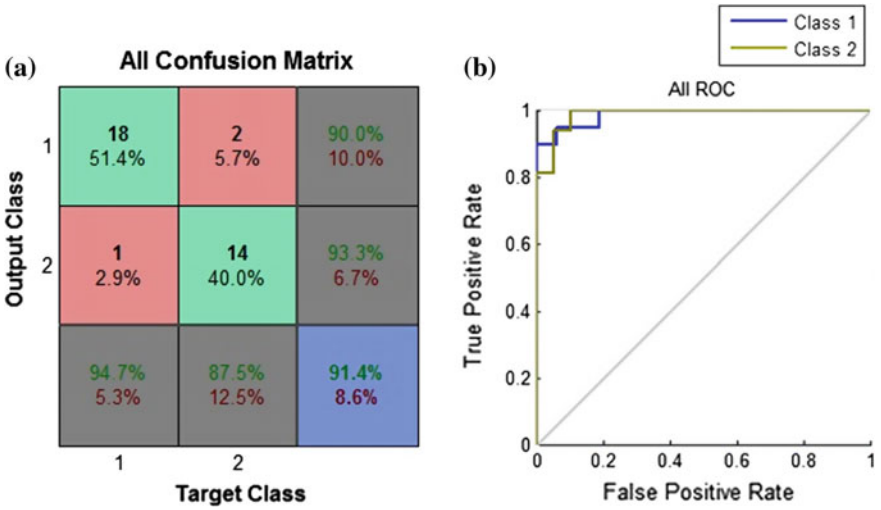


Fig. 29.5 Confusion matrix and ROC plots, **a** Confusion matrix, **b** ROC plot

Conclusion

WM features are found to be promising and important in classifying demented and non-demented subjects. In case of normal and demented patients, the changes in area and volume of WM are observed to be different, because for demented subject, the concentration of WM is less when compared with non-demented subjects. The highest accuracy up to 91% is achieved for the classification of MR images of normal and demented subjects. The extracted features are proven to be a significant insight to discriminate demented and non-demented brains.

References

1. Aurauma Chutinet, and Natalia S. Rost: White matter disease as a biomarker for long-term cerebro vascular disease and dementia, Cardiovascular Medicine, vol. 18, issue 122(2014).
2. Deepa V, Benson C.C, Lajish V.L: Gray Matter and White Matter Segmentation from MRI brain images using clustering methods. International Research Journal of Engineering and Technology (IRJET)|Department of Computer Science, University of Calicut, Kerala, India-673635|Volume: 02 Issue: [08 March 2016].
3. Phoebe ministries Alzheimer's Association. (2015). Types of Dementia. Retrieved from http://www.alz.org/dementia/types-of-dementia.asp. Alzheimer's Foundation of America (2015). [Accessed: 30-Feb-2016].
4. Who.int, WHO|Dementia, 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs362/en/. [Accessed: 16-Feb-2016].
5. Saloni Tanna and Béatrice Duthey: Background Paper 6.11 Alzheimer Disease and other Dementias [Priority Medicines for Europe and the World A Public Health Approach to

Innovation] Update on 2004 Background Paper, BP 6.11 Alzheimer Disease, 20 February 2013.
6. Bhalerao G.V. and SampathilaN.: K-means Clustering Approach for Segmentation of Corpus Callosum from Brain Magnetic Resonance Images. Proceedings of International Conference on Circuits, Communication, Control and Computing (I4C 2014), MSRIT, BANGALORE, India, pp. 434–437 (2014).
7. Brickman A.M., Honig L.S., Scarmeas N., Tatarina O., Sanders L., Albert M.S., Blacker J.B. D., Stern Y.: Measuring Cerebral Atrophy and White Matter Hyper intensity Burden to Predict the Rate of Cognitive Decline in Alzheimer Disease. American Medical Association, ARCH NEUROL / VOL 65 (NO. 9), pp: 1202–1207 (2008).
8. G.V. Bhalerao: M.Tech thesis on Extraction of corpus callosum for the detection of abnormality in brain magnetic resonance images. Manipal Institute of Technology, Manipal University, India. (2015).
9. OASIS brain MRI database. Available: http://www.oasis-brains.org/ [Accessed: 2-February-2016].
10. Marcus D.S.: Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults.J CognNeurosci. Vol. 22(12), pp. 2677–2684 (2010).
11. Sari I.: B.Tech thesis on Segmentation of white matter for the classification of brain MR images. Manipal Institute of Technology, Manipal University, India. 2016 [Submitted].
12. Olszewski R.: Slicer3 Training Tutorial Manual segmentation of orbit structures from isotropic MRI data within 3D Slicer for windows. National Alliance for Medical Image Computing [Accessed: 15 March 2016].
13. Lasso A., Jean-Christophe Fillion Robin, Wang K., Fichtinger G.: MatlabBridge Extension NA-MIC National Alliance for Medical Image Computing 2013, [Accessed: 29 March 2016].
14. Pujol S.: Slicer3 Minute Tutorial, National Alliance for Medical Image Computing http://na-mic.org 2010, [Accessed: 15 March 2016].
15. Marcus D.S.: Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults.NIH Public Access Author Manuscript, J CognNeurosci. Vol. 22(12), pp. 2677—2684 (2010).
16. Egmont-Petersena M., Ridderb D., Handelsc H.: Image processing with neural networks, Article in Pattern Recognition. (2002) [Accessed: 20 June 2016].
17. Savio A., Garcia-Sebastian M., Hernandez C., Grana M., and Villanua J.: Classification results of artificial neural networks for Alzheimer's disease detection. Intelligent Data Engineering and Automated Learning-IDEAL 2009, Springer Berlin Heidelberg, pp. 641–648 (2009).
18. Gavadia G., Lopez R., and Soudah E.: Computer-aided diagnosis of dementia using medical image processing and artificial neural networks. Computational Vision & Medical Image Processing, pp. 51–55 (2011).
19. Riedmiller M. and Braun H.: A direct adaptive method for faster back propagation learning: The RPROP algorithm. Proceedings of the IEEE International Conference on Neural Networks, pp. 586–591 (1993).

Chapter 30
A Deblocking Algorithm for JPEG-Compressed Images Using Wavelet Domain Adaptive Non-Local Means Filtering

Vijay Kumar Nath, Satyabrat Malla Bujar Baruah, Deepika Hazarika and Hilly Gohain Baruah

Abstract This paper proposes a new post-processing scheme for blocking artifact reduction in JPEG-compressed images through wavelet domain adaptive non-local means (NLM) filtering. It is assumed that the degradation due to discrete cosine transform (DCT) domain quantization noise is some additive noise. The good denoising property of undecimated wavelet transform (UDWT) is utilized here. It is shown that there exist a strong relationship between the JPEG image quality factor (Q) and the experimentally observed 'optimal' NLM filter parameters. Based on a subband adaptive threshold, the detail coefficients are first classified into significant and non significant coefficients. The coefficients below the threshold are insignificant coefficients that are treated with adaptive NLM filter using proposed parameter selection strategy. With this strategy, the wavelet coefficients corresponding to edge structures are well preserved and the noisy ones are reduced effectively. The coefficients above the threshold are significant coefficients which are treated with adaptive NLM filter using different parameter selection strategy that yields comparatively less smoothing and avoids oversmoothing of important details. The proposed deblocking algorithm achieves high competitive performance compared to that of existing well-known image deblocking techniques both objectively and subjectively.

Keywords Deblocking · Block discrete cosine transform · Non- local means filter · JPEG · Redundant wavelet decomposition

V. K. Nath (✉) · S. M. Bujar Baruah · D. Hazarika · H. G. Baruah
Department of Electronics and Communication Engineering,
School of Engineering, Tezpur University, Tezpur , Assam, India
e-mail: vknath@tezu.ernet.in

S. M. Bujar Baruah
e-mail: baruah.satyabrat@gmail.com

D. Hazarika
e-mail: deepika@tezu.ernet.in

H. G. Baruah
e-mail: hilly90@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_30

Introduction

The block-based discrete cosine transform (DCT) is widely popular due to its near-optimal energy compaction feature and availability of efficient hardware implementations [1]. Thus block DCT is popularly used in image and video coding standards such as JPEG and MPEG. However in JPEG at high compression ratios, the decoded images exhibit visually disturbing blocking artifacts which appear in the form of gray-level discontinuities along the block boundaries. A significant amount of attention has been paid to the post-processing of DCT-coded images in the literature.

Algorithm described in [2] apply the 3×3 Gaussian filter to pixel elements along the block boundaries to reduce blocking artifacts, however the true edges get blurred in this process. The iterative projection onto convex set (POCS) [3, 4]-based methods produces good deblocking results but are computationally demanding. Nath et al. [5] proposed to apply an adaptive bilateral filter to suppress blockiness in DCT compressed images. Recently Alireza et al. [6] proposed an JPEG artifacts reduction algorithm via two stages—the first stage suppress blocking artifacts via block boundary smoothing and guided filtering. The second stage takes care of blurring and aliasing reduction around the edges via local edge regeneration scheme. In [7], Wang et al. proposed a deblocking method consisting of three steps-thresholding, model classification, and a deblocking filtering. The authors used a directional filter to reduce ringing around the edges.

In [8], Wu et al. described an efficient wavelet thresholding-based post-processing technique where the soft threshold value is made adaptive to various block DCT compressed images using various quantization tables. This method produces very good deblocking results, while preserving important features in an image. The shape adaptive DCT [9]-based image deblocking technique has shown state-of-the-art deblocking results in the literature. In [10], blocking artifacts are suppressed in wavelet domain using local Wiener filtering. The high correlation between the mean of first 3×3 values of quantization table and the noise standard deviation has been demonstrated. In [11], a wavelet domain blocking artifact suppression method based on local Laplace prior was proposed. The wavelet coefficients in this method are modeled using a Laplace pdf that uses local parameters.

In [12], You and Cho proposed an image denoising technique based on wavelet domain NLM filtering. Unlike conventional wavelet domain methods, this method can effectively preserve the small coefficients that are corresponding to important signal features and eliminate the actual noisy ones. Motivated by their approach we propose to suppress blocking artifacts in UDWT domain using improved NLM filtering. It should be noted that the improved NLM filter [13] performs better than simple NLM filter [14] and very efficiently computes the weight not only based on intensity distance but also on spatial distance. Since the 'optimal' selection of domain and range parameters highly affects the filtering performance, we have made these parameters adaptive to the JPEG image quality factor (Q).

The rest of the paper is organized as follows: section "Proposed Method" describes the proposed method and provide details on adaptive parameter selection for

improved wavelet domain NLM. In section "Experimental Results" we present the experimental results of many blocking artifact suppression techniques. Section "Conclusion" concludes this paper.

Proposed Method

Although various block DCT domain quantization noise models exist in the literature, the degradation due to DCT domain quantization noise in this chapter is modeled as some additive noise. The model of a degraded image has the form $Y = X + N$ where X denotes the original non-compressed image and Y represents the image that is obtained after block DCT domain quantization and N denotes some additive noise with standard deviation σ_n. After applying undecimated DWT, we have $y = x + n$ where y, x, and n denotes the wavelet domain observed image, uncompressed image, and noise respectively.

Buades et al. [14] introduced the concept of NLM filtering that has attracted significant attention in the recent years. The conventional NLM filter modifies its pixels as the weighted sum of its neighborhood pixels where the weights are computed based on self- similarity of the neighbors. The improved NLM filter [13] used in this paper is defined by the following expression:

$$\hat{x}(m,n) = \frac{1}{Z_{m,n}} \sum_{i',j' \in \Omega} e^{-\left(\frac{\|(m,n)-(m',n')\|_2^2}{2h_d^2}\right)} e^{-\left(\frac{\|y(s_{m,n})-y(s_{m',n'})\|_{2,a}^2}{2h_r^2}\right)} y(m',n') \tag{30.1}$$

where $y(s_{m,n})$ is the image patch centered at spatial position (m, n). h_d and h_r controls the weights in spatial and intensity domains respectively. $Z_{m,n}$ is the normalization factor given by the following expression:

$$Z_{m,n} = \sum_{m',n' \in \Omega} W_d(m',n') W_r(m',n') \tag{30.2}$$

$$W_d(m',n') = e^{-\left(\frac{\|(m,n)-(m',n')\|_2^2}{2h_d^2}\right)} \tag{30.3}$$

$$W_r(m',n') = e^{-\left(\frac{\|y(s_{m,n})-y(s_{m',n'})\|_{2,a}^2}{2h_r^2}\right)} \tag{30.4}$$

(30.1) can be treated as a bilateral filter when the image patch reduces to the size of a pixel element and also can be treated as a conventional NLM filter when the parameter h_d equals to $+\infty$. Therefore expression in (30.1) can be viewed as a combined form of basic NLM filter and bilateral filter. The very low values of h_d and h_r may leave the noise unfiltered and very high values of these parameters may oversmooth

the input image. Therefore, the proper adaptive estimation of these two parameters is highly important.

In the past few decades, a good amount of work has been reported on DWT domain image denoising methods because of its sparsity property and multiresolution structure. The orthogonal DWT lacks shift invariance that makes it unsuitable for various image processing applications. The removal of downsamplers from the analysis stage and upsamplers from the synthesis stage can still provide perfect reconstruction and translation invariance is also achieved. This type of decompositions is known as undecimated DWT. In traditional wavelet domain denoising while thresholding the large wavelet coefficients which correspond to important signal features are usually preserved well but small wavelet coefficients which are due to important signal features are often removed along with the unimportant small coefficients which are due to noise.

We exploit the effective denoising property of UDWT and improved NLM together to suppress the blocking artifacts in JPEG-compressed images. We propose to classify the wavelet coefficients in a detailed subband into significant and non-significant coefficients using Donoho's global universal threshold [15].

$$Th = \sigma\sqrt{\log M \times N} \tag{30.5}$$

where $M \times N$ is the image size. When $M = N$, the threshold is given by

$$Th = \sigma\sqrt{2\log M} \tag{30.6}$$

where σ is estimated from the finest diagonal subband using

$$\widehat{\sigma}^2 = \frac{Median\left(|y(m,n)|\right)}{0.6745}, y(m,n) \in HH_1 \tag{30.7}$$

We propose to apply improved NLM filtering on insignificant wavelet coefficients which is expected to preserve all the important structural features and also the small coefficients that are important while eliminating the actual noisy ones. In order to select the parameters of improved NLM, we first perform experiments using ten number of JPEG-compressed images (of different characteristics) to find a relation between these parameters and JPEG image quality factor (Q). It should be noted that the Q value can be calculated easily from the quantization table available at the decoder. Through extensive deblocking experiments, we found that the 'optimal' h_r is very difficult to relate with optimal h_d. To study the relationship between 'optimal' h_r and Q, the parameter h_d is set to some constant value and we record the best values of h_r that offers the least mean square error for each JPEG image (with quality factor Q). The experimentally obtained best values of h_r are then plotted against Q (Fig. 30.1a). It is clearly observed from the scatter plot that there exists a relationship between h_r and Q. After fitting the scatter plot data with a sixth-degree polynomial, the h_r equation is expressed as

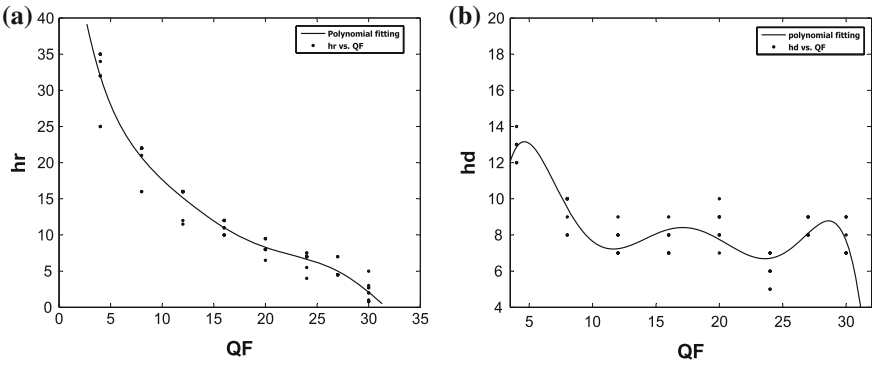


Fig. 30.1 Scatter plot with curve fitting of **a** h_r and Q **b** h_d and Q

$$h_r = P1Q^6 + P2Q^5 + P3Q^4 + P4Q^3 + P5Q^2 + P6Q + P7 \tag{30.8}$$

where $P1 = 1.179 \times 10^{-6}$, $P2 = -1.372 \times 10^{-4}$, $P3 = 0.006299$, $P4 = -0.1469$, $P5 = 1.885$, $P6 = -14.03$ and $P7 = 65.88$.

Similarly, the relationship between 'optimal' h_d and Q is studied by setting the parameter h_r to some constant value and we record the best values of h_d that provides the least mean square error for each JPEG image (with quality factor Q). The values of h_d are then plotted against Q (Fig. 30.1b). The scatter plot data is fitted with a sixth-degree polynomial and we have h_d as

$$h_d = P1Q^6 + P2Q^5 + P3Q^4 + P4Q^3 + P5Q^2 + P6Q + P7 \tag{30.9}$$

where $P1 = -5.876 \times 10^{-6}$, $P2 = 6.024 \times 10^{-4}$, $P3 = -0.02414$, $P4 = 0.4762$, $P5 = -4.752$, $P6 = 21.61$ and $P7 = -22.55$. The parameters of improved NLM, applied on insignificant wavelet coefficients, are calculated using (30.8) and (30.9). We propose to apply the same NLM filter on approximation subband because the coefficients in this subband are also contaminated by the aforementioned noise.

The classified significant wavelet coefficients which are mainly due to important image features like edges, etc., are filtered using NLM which provides comparatively less smoothing so that any oversmoothing of important details can be avoided. The parameters of improved NLM that are computed using (30.8) and (30.9) are scaled by half in the case of significant coefficients. Any coefficient which is due to noise and is classified as significant coefficient is reduced in magnitude.

In the next section, it is shown that the deblocking results obtained using (30.8) and (30.9) are not very sensitive to fitting errors.

Experimental Results

In order to evaluate the deblocking performance of proposed scheme, we have carried out extensive experiments using test images of different characteristics. The test images Barbara and Goldhill are examined in this section.

The proposed implementation uses UDWT with one level of decomposition and Daubechies length-4 wavelet. We have used MATLAB compression toolbox to obtain JPEG images compressed at various quality factors. The performance of proposed algorithm is compared with shape-adaptive DCT-based deblocking method (SADCT) [9], one very recent deblocking method based on local edge regeneration (LER) [6] and spatial domain improved NLM filter (INLM) [13]. The results of SADCT and LER were obtained using MATLAB code made available by its authors in their official websites. We implemented the INLM method ourselves. The blocking artifact suppression performance of our method is assessed using two popular metrics: peak signal-to-noise ratio (PSNR) [16] and structural similarity (SSIM) index [17]. As the PSNR measure alone cannot give a decision on visual quality of an image, we have also considered SSIM index. The structural similarity between two images can be measured using structural similarity index. It has a value between 0 and 1. The SSIM index of 1 indicates that the original image and filtered image are exactly the same.

In Table 30.1, we present the PSNR results of our scheme, SADCT, LER, and INLM schemes. The proposed method provides best PSNR results compared to those obtained by the SADCT and LER methods and consistently outperforms the spatial domain INLM method by a large margin (Table 30.1).

The proposed method provides better SSIM index results compared to SADCT, LER, and INLM methods in most of the situations (Table 30.2).

The blocking artifacts have been effectively reduced in almost all the cases but the oversmoothing of important details is clearly observed in Fig. 30.2c and e. The details are best preserved in the images provided by proposed method (Fig. 30.2f).

Table 30.1 PSNR (in dB) performance of different deblocking methods in comparison with our method

Barbara (512 × 512)						Goldhill (512 × 512)				
Q	JPEG	INLM [13]	LER [6]	SADCT [9]	Proposed	JPEG	INLM [13]	LER [6]	SADCT [9]	Proposed
4	23.48	23.71	24.18	24.66	24.74	25.14	25.71	26.12	26.21	26.23
8	25.19	25.46	25.41	26.14	26.37	25.60	27.62	28.05	28.24	28.39
10	25.79	26.14	25.89	26.63	26.93	28.28	28.32	28.61	28.87	29.02
12	26.33	26.85	26.44	27.11	27.49	28.84	28.86	29.09	29.33	29.50
14	26.90	27.69	26.88	27.67	28.10	29.42	29.42	29.44	29.75	29.93
16	27.41	28.47	27.29	28.15	28.64	29.71	29.99	29.82	30.11	30.31

Table 30.2 SSIM index of different deblocking methods in comparison with our method

Barbara (512 × 512)						Goldhill (512 × 512)				
Q	JPEG	INLM [13]	LER [6]	SADCT [9]	Proposed	JPEG	INLM [13]	LER [6]	SADCT [9]	Proposed
4	0.7290	0.7669	0.7767	0.7902	0.7876	0.7051	0.7002	0.7502	0.7457	0.7481
8	0.8423	0.8482	0.8543	0.8729	0.8725	0.8406	0.7988	0.8542	0.8479	0.8539
10	0.8763	0.8761	0.8771	0.8990	0.8990	0.8710	0.8299	0.8792	0.8733	0.8797
12	0.8972	0.8962	0.8916	0.9138	0.9158	0.8923	0.8533	0.8968	0.8902	0.8983
14	0.9144	0.9164	0.9038	0.9279	0.9311	0.9087	0.8787	0.9110	0.9057	0.9144
16	0.9264	0.9327	0.9128	0.9364	0.9411	0.9215	0.9031	0.9224	0.9164	0.9269

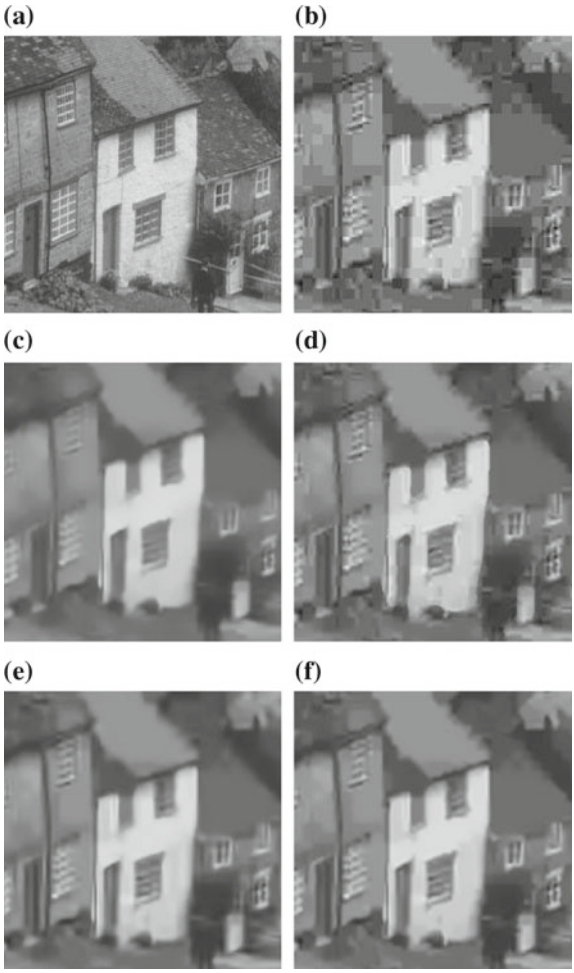


Fig. 30.2 Deblocked results of Goldhill image **a** A segment of original Goldhill image **b** JPEG compressed (Q = 5) **c** INLM [13] **d** LER [6] **e** SADCT [9] and **f** Proposed method

Conclusion

This paper presented a new wavelet domain adaptive NLM filtering algorithm for suppressing blocking artifacts in JPEG-coded images. The wavelet coefficients in detailed subbands are classified into significant and insignificant coefficients which are then filtered using adaptive NLM employing different parameter estimation strategies. It is shown that the parameters of improved NLM have got strong correlation with JPEG image quality factor Q. The proposed method while eliminating the noisy wavelet coefficients preserves the important large and also small coefficients corresponding to important signal features. The proposed scheme is compared with

many existing deblocking schemes which include one very recent scheme. Experimental results show that the blocking artifact suppression performance of proposed technique is highly encouraging and in almost all the situations outperforms the very recent one.

References

1. Jain, A.K.: Fundamental of Digital image processing. Englewoods Cliffs, NJ:Prentice hall (1989)
2. Reeve, H.C., Lim, J.S.: Reduction of blocking artifacts in image coding. Optical Engineeering 23(1), 34–37 (1984)
3. Paek, H., Kim, R.C., Lee, S.U.: On the pocs based postprocessing technique to reduce the blocking artifacts in transform coded images. IEEE Transactions on Image Processing 8(3), 358–367 (1998)
4. Yang, Y., Galatsanos, N.P., Katsaggelos, A.K.: Regularized reconstruction to reduce blocking artifacts of block discrete cosine transform compressed images. IEEE Transactions on Circuits and Systems for Video Technology 3(6), 421–432 (1993)
5. Nath, V.K., Hazarika, D., Mahanta, A.: Blocking artifacts reduction using adaptive bilateral filtering. International Conference on Signal processing and communications 6, 243–250 (2010)
6. Golestaneh, S.A., Chandler, D.M.: Algorithm for jpeg artifact reduction via local edge regeneration. Journal of Electronic Imaging 23(1), 1–13 (2014)
7. Wang, J., Wu, Z., Jeon, G., Jeong, J.: An efficient spatial deblocking of images with dct compression. Digital Signal Processing 42, 80–88 (2015)
8. Wu, S., Yan, H., Tan, Z.: An efficient wavelet-based deblocking algorithm for highly compressed images. IEEE Transactions on Circuits and Systems for Video Technology 11(11), 1193–1198 (2001)
9. Foi, A., Katkovnik, V., Egiazarian, K.: Pointwise shape adaptive dct for high quality denoising and deb-locking of grayscale and color images. IEEE Transactions on Image Processing 16(5), 1395–1411 (2007)
10. Nath, V.K., Hazarika, D.: Blocking artifacts suppression in wavelet transform domain using local wiener filtering. In: IEEE National coference on Emerging trends and applications in Computer Science (NCETACS). pp. 93–97 (2012)
11. Nath, V.K., Hazarika, D.: Image deblocking in wavelet domain based on local laplace prior. International Journal of Multimedia and its Applications 4(1), 39–46 (2012)
12. You, S.J., Cho, N.I.: A new image denoising method based on the wavelet domain non local means filtering. In: IEEE International conference on Accoustics, Speech and Signal Processing. pp. 1141–1144 (May 2011)
13. Chen, F., Huang, X., Chen, W.: Texture-preserving image deblurring. IEEE Signal Processing Letters 17(12), 1018–1021 (2010)
14. Buades, A., Coll, B., Morel, J.M.: A non local algorithm for image denoising. In: IEEE confrence on Computer vision and Pattern recognition. pp. 60–65 (2005)
15. Donoho, D.L.: De-noising by soft-thresholding. IEEE Transactions on Information Theory 41, 613–627 (1995)
16. Chen, T., Wu, H.R., Qiu, B.: Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts. IEEE Transactions on Circuits and Systems for Video Technology 11(5), 594–602 (2001)
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment : From error visibility to structural similarity. IEEE Transactions on Image processing 13(4), 600–612 (2004)

Chapter 31
Extracting Qualitative Spatiotemporal Relations for Objects in a Video

Juwesh Binong and Shyamanta M. Hazarika

Abstract Appropriate qualitative abstraction of a given domain is essential for qualitative reasoning. This chapter presents a method of extracting episodes from a video, where an episode is a maximal temporal interval within which qualitative spatial relations between objects do not change. Movement patterns of objects during an episode are also estimated. Experimental results with vignettes taken from the Mind's Eye Dataset shows encouraging results.

Keywords Qualitative representation · Spatial relation · Constraint graph · Episode · Movement pattern

Introduction

Commonsense reasoning is predominantly driven by qualitative abstractions rather than quantitative knowledge a priori. Providing ways and means of imitating such phenomenon within a computational framework has given birth to the area of Qualitative Reasoning [1] within Artificial Intelligence. Qualitative reasoning is motivated by the claim of it being close to how human perform everyday reasoning. Qualitative reasoning is able to deal with incomplete knowledge.

Qualitative spatial reasoning refers to qualitative reasoning on spatial configurations among distinct spatial entities without incorporating information such as size, shape, texture, or color of spatial entities [2]. Similarly, qualitative temporal reasoning deals with reasoning that exploits temporal relations between spatial entities. Qualitative Spatiotemporal Reasoning [3] is concerned with constraint-based

J. Binong (✉)
Department of Electronics & Communication Engineering,
North Eastern Hill University, Shillong 793022, India
e-mail: binong.j@gmail.com

S. M. Hazarika
Biomimetic and Cognitive Robotics Lab, Department of Computer
Science & Engineering, Tezpur University, Tezpur 784028, India

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_31

formalisms for representing, and reasoning with spatial and temporal information over infinite domains. Given a fixed granularity, one identifies a set of relations to be used in disjunctive qualitative statements, which allow for expressing indefinite and imprecise knowledge about the application domain [4].

The remainder of this paper is structured as follows. Section "Conceptual Basics" provides basic concepts related to qualitative representations and reasoning, and also concepts necessary for formulating the proposed method. Section "Generation of Qualitative Representation" then discusses the methods of generating qualitative representation.

Conceptual Basics

Qualitative Spatiotemporal Representation

Qualitative representations employ symbols that abstract from quantitative details irrelevant for the task at hand to represent semantically meaningful concepts of a domain [5]. A large variety of qualitative representations are found in the literature (See [6] for an overview). Reasoning tasks are benefited from the specificity of the qualitative representation. Designing a qualitative representation of a domain starts with two design decisions [5]:

1. To identify a finite set of concepts and
2. To fix a knowledge representation (language) to compose statements.

Spatial Relations

In this work, the term *spatial relation* [7] means a topological relation between two spatial entities. An abstracted version of a spatial representation language Region Connection Calculus (RCC-8) is used to represent spatial configurations of the objects in the video.

RCC-6: Abstraction of RCC-8

Fine granularity as provided by RCC-8 is not required for the work described in this chapter. Some of the spatial relations of RCC-8 are combined thereby reducing the number of spatial relations to six.

The relations TPP and NTPP are combined together into a single relation, *PP*, i.e., *proper-part of* as these relations represents an object being proper-part of another object, and similarly their inverses (i.e., TPPi and NTPPi respectively) to *PPi*. This

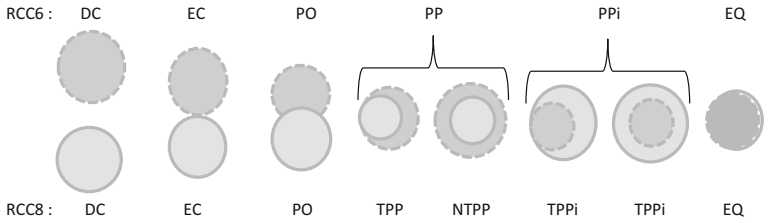


Fig. 31.1 Abstracting RCC-8 to RCC-6

set of six spatial relations is termed *RCC-6*. Figure 31.1 shows the abstraction of RCC-8 to RCC-6.

Temporal Relations

The term *temporal relation* [8] means a qualitative relation between two temporal intervals. We denote a time instant by t_i and temporal interval by $< t_1, t_2 >$. The temporal relation *meet* is defined as follows:

Definition 1 *(Meet)*: We say that a time interval $< t_1, t_2 >$ meets another time interval $< t_3, t_4 >$; if and only if $t_2 = t_3$.

Movement Patterns

The movement pattern of an object x in a spatial environment can be *stationary*, *cyclicity*, *non-cyclicity*, *rotational*, etc. In the context of a change of location, an object can have only two qualitative movement patterns, i.e., mobile or immobile (or *stationary*).

Definition 2 *(Immobile Pattern)*: The movement pattern of an object is termed as *immobile*, if it is occupying the same location through out a given time interval.

Definition 3 *(Mobile Pattern)*: The movement pattern of an object is termed as *mobile*, if it is not occupying the same location through out a given time interval.

IMB(x) and MOB(x) are used to denote that object 'x' is immobile and mobile respectively.

Constraint Graph

A spatial configuration of objects is represented using *constraint network* [9]. When the spatial relations are binary, i.e., involve pairs of objects, the constraint network becomes a *binary constraint network*. A binary constraint network is represented by an edge-labeled directed graph termed a *constraint graph*. As a frame of a video represents a spatial configuration, it can be represented as a *constraint graph* G.

Definition 4 *(Constraint Graph)*: Given a set of vertices V, a set of edges E, and a set of possible values R for an arbitrary edge, a constraint graph $G = (V, E, \psi)$ is an weighted directed graph, where ψ is a mapping from a pair $(v_i, v_j) \in V$ to an element $r_i \in R$.

Episode

An episode represents a maximal temporal slice of a video during which there is no change of spatial relations between any pair of objects. Formally, an episode can be defined as follows:

Definition 5 *(Episode)*: An episode is the maximal part of the observation time period within which there is no change of qualitative spatial relation between the observed objects.

A video consists of a sequence of frames and an episode is a temporal slice of a video. A video can be represented as a sequence of episodes. Each episode consists of qualitative binary spatial relations between objects and movement patterns of individual objects.

Generation of Qualitative Representation

To extract qualitative spatiotemporal representation, first, key frames from the given video are extracted, followed by extraction of the binary spatial relations between objects in each key frames and transforming it into constraint graph. Then, episodes are determined by comparing consecutive constraint graphs, and finally for each object in each episode the movement pattern is estimated.

Figure 31.2 depicts the general architecture for generating qualitative spatiotemporal representation from a given video.

The following assumptions are made while transforming the given video into qualitative representation:

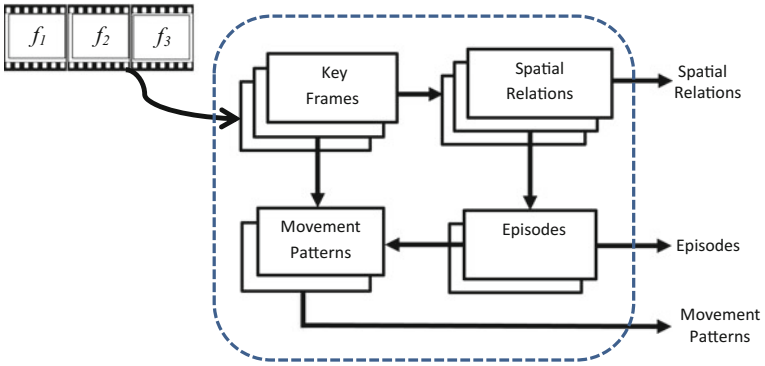


Fig. 31.2 Block Diagram of qualitative spatiotemporal representation generation method

1. The actual size of objects remains unchanged across frames.
2. The object lost during full occlusion, if any, still exists within the field of view (*though not visible).*

Algorithm 1 describes how the given video is transformed into qualitative representation consisting of episodes.

Input: Video V
Output: Qualitative Spatial Relations, Movement Patterns & Episodes
1 Generate Key Frames $\{k_f\}_{f=1}^{N}$ for video V
2 **forall** *key frames* $f = 1 \cdots N$ **do**
3 | Generate constraint graph G_c using algorithm 2.
4 **end**
5 **forall** *constraint graph* $G_c = 1 \cdots N$ **do**
6 | Determine episodes e_i using algorithm 3.
7 **end**
8 **forall** *episode* $e_i = 1 \cdots M$ **do**
9 | Estimate movement patterns $\{P_j^i\}, j = (1 \cdots K)$ objects in e_i using algorithm 4.
10 **end**

Algorithm 1: Extracting Episodes containing Qualitative Spatial Relations between objects and their individual Movement Patterns from Video

Extracting Spatial Relations

A frame represents a spatial configuration of objects at an instant of time. Spatial configuration of objects is represented using constraint graph.

Algorithm 2 describes how binary spatial relations between pairs of objects in a frame is determined and the frame is transformed into a constraint graph.

Input: Video V
Output: Constraint Graph Set $\{R_{ij}\}_{f=1}^{N}$
Generate Key Frames $\{k_f\}_{f=1}^{N}$ for video V
forall $f = 1 \cdots N$ **do**
    $\{B_i\} \leftarrow$ BoundingBox of the i^{th} object
    $\{x_i, y_i, b_i, h_i\} \leftarrow$ Coordinates of $\{B_i\}$
end
forall $i = 1 \cdots m$ **do**
    Read $\{x_i, y_i, b_i, h_i\}$
    forall $j = 1 \cdots m$ and $i \neq j$ **do**
        Read $\{x_j, y_j, b_j, h_j\}$
        $R_{ij} \leftarrow$ *Spatial Relation*$(\{x_i, y_i, b_i, h_i\}, \{x_j, y_j, b_j, h_j\})$
        $\{R_{ij}\}_{f=1}^{N} \leftarrow \{R_{ij}\}_{f=1}^{N} + R_{ij}$
    end
end

Algorithm 2: Pseudocode for Extracting Spatial Relations from Key Frames.

Note: The arguments (x_i, y_i) in the coordinates $\{x_i, y_i, b_i, h_i\}$ of a bounding box $\{B_i\}$ represent the coordinates of the top-left corner of $\{B_i\}$, and the argument b_i and h_i represent the width and height of B_i respectively.

Slicing the Video into Episodes

An episode is a temporal slice, i.e., a sequence of frames, each of which when transformed constitutes the same constraint graph. Episode can be determined from a sequence of video frames by transforming each frames into constraint graph and checking the equality of the constraint graphs.

Algorithm 3 describes how the given video can be sliced into episodes.

Determination of Movement Patterns

The movement pattern of an object during an episode is determined by comparing the coordinates of bounding boxes of an object across different key frames belonging to the given episode.

Algorithm 4 describes how the movement patterns of objects in each episode are estimated.

Input: Constraint Graph Set $\{R_{ij}\}_{f=1}^{N}$
Output: Episode Set $\{e_{[i]}\}_{=1}^{n}$

```
i ← 1;
Read g[i];
e[i]← g[i];
forall j = 2 ··· (n − 1) do
    Read g[j];
    if (g[i] ≠ g[j]) then
        i ← i + 1;
        e[i] ← g[i];
    end
end
```

Algorithm 3: Pseudocode for Determining Episodes from Constraint Graphs.

Input: Episode set $\{e[l]\}_{l=1}^{n}$, Key frames set $\{k_f\}_{f=1}^{N}$ for video V
Output: Movement Pattern set $\{M_i^l\}$

```
i ← 1;
l ← 1;
f ← 1;
while (i ≤ number of objects) do
    if (k_f ∈ e[l]) then
        {x_i, y_i, b_i, h_i} ← Coordinates of BoundingBox {B_i^f}
        forall f = 2 ··· N and i = j do
            if (k_f ∈ e[l]) then
                {x_j, y_j, b_j, h_j} ← Coordinates of BoundingBox{B_j^f}
                if ({x_i, y_i, b_i, h_i} = {x_j, y_j, b_j, h_j}) then
                    M_i^l ← }Immobile'
                end
                else
                    M_i^l ← }Mobile'
                end
            end
            else
                l ← l+1
                {x_i, y_i, b_i, h_i} ← Coordinates of BoundingBox{B_j^f}
            end
        end
    end
    i ←i+1
end
```

Algorithm 4: Pseudocode for Estimating Movement Pattern of Objects

Experimental Results

The proposed methods have been evaluated on 13 number of action verbs vignettes taken from the Mind's Eye dataset. Figure 31.3 shows a sequence of frames extracted

Fig. 31.3 Sequence of frames showing qualitative spatial relations between two bounding boxes

Table 31.1 Success rate against ground truth on 13 vignettes from mind's eye dataset

Parameters	Episodes	Spatial relations	Movement patterns
Ground truth	13	73	73
Proposed qualification	11	70	71
Success rate (in %)	84.62	95.89	97.26

from one vignette. The sequence of frames has three episodes. Each of frame Fr#00192 and Fr#00200 represents an episode. Frames Fr#00160 and Fr#00170 represent a single episode as they possess the same spatial relation 'dc'. The movement pattern of the taller person (say P_1) is MOB(P_1) in all episodes, while shorter person (say P_2) has immobile movement pattern IMB(P_2).

Each of the 13 vignettes considered represents some form of occlusion and posses multiple episodes. Outputs of the algorithms that matched with the ground truths and the corresponding success rates are recorded in Table 31.1. Ground truths are provided by human experts.

Conclusion

In this chapter, we propose a method of transforming a video into qualitative spatiotemporal representation. The representation can be used in applications exploiting symbolic reasoning. The experimental results with real-world video show encouraging results.

References

1. Renz J., Nebel B.: Qualitative spatial reasoning using constraint calculi. Handbook of Spatial Logics, p. 161–215, Springer (2007)
2. Chen J., Cohn A.G., Liu D., Wang S., Ouyang J., Yu Q. : A survey of qualitative spatial representations. The Knowledge Engineering Review, Vol.30(1), 106–136 (2015)

3. Hazarika, S.M.: Qualitative spatial change : space-time histories and continuity. PhD Thesis, The University of Leeds (2005).
4. Westphal M., Dornhege C., Wolfl S., Gissler M., Nebel B.: Guiding the Generation of Manipulation Plans by Qualitative Spatial Reasoning. Spatial Cognition and Computation. 11(1), 75–102 (2011)
5. Wolter D., Kirsch A.: Qualitative Spatial Reasoning for Boosting Learning-Based Robotics. In IROS Workshop of Machine Learning in Planning and Control of Robot Motion. Hamburg (2015)
6. Cohn A. G., Hazarika S. M.: Qualitative spatial representation and reasoning: An overview. Fundamenta Informaticae. 46(1-2):1–29 (2001)
7. Randell D. A., Cui Z., Cohn A. G.: A spatial logic based on regions and connection. In (KR-92), pp. 165–176. Morgan Kaufmann (1992)
8. Allen J. F. : Maintaining knowledge about temporal intervals. Communications of the ACM. 26(11), 832–843 (1983)
9. Ligozat G, Renz J.: What is a qualitative calculus? A general framework. In Proceeding of the 8^{th} Pacific Rim International Conference on Artificial Intelligence. pp. 53–64 (2004)

Chapter 32
A Spell Correction Method for Query-Based Text Summarization

Nazreena Rahman and Bhogeswar Borah

Abstract Finding and correcting incorrect spelling from text documents always plays an important role particularly in information retrieval. There are many approaches to tackle spell checking problem. Here, one spell checking and correcting method is proposed particularly for query-based text summarization purpose. Our method mainly works on non-word errors. Query-based text summarization finds the summary based on the given query. This method refines the query by replacing each misspelled word with exact match of that word with the dictionary. We use TAC 2009 dataset for our experiments and validation purpose and obtain encouraging results.

Keywords Information retrieval · Spell checking · Non-word errors · Query-based text summarization

Introduction

The rapid and continuous growth of text databases makes it difficult to retrieve the required information. Therefore, query- based text summarization can be used for finding the summarized answer according to the user's need. Text summarization defines a text that is constructed from single or multiple texts and length of the summarized text is just half of the main text documents [1]. Text summarization can be obtained on a single text document or multiple text documents. It can be of extraction based or abstraction based. In extractive method, sentences can be extracted from original texts. Abstractive summarization needs information integration, sentence

N. Rahman (✉) · B. Borah
Department of Computer Science and Engineering, Tezpur University, Tezpur, Sonitpur 784028, Assam, India
e-mail: naz1912@tezu.ernet.in

B. Borah
e-mail: bgb@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_32

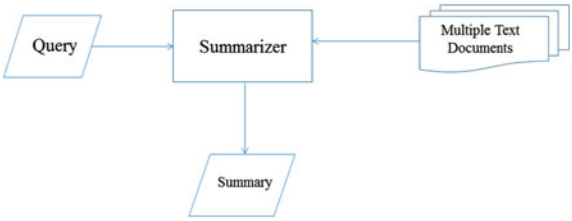


Fig. 32.1 Query-based summarization overview

compression, and reformulation. Summarization can be of generic or query-based type. In generic summarization, summary can be achieved based on the document, no query is required but in case of query-based technique, summary depends on the user's query. Based on details, indicative and informative summary can be constructed. Indicative summary indicates whether the reader should go through the text document or not and informative summary gives most important information of the texts. Query-based text summarization plays a vital role in information retrieval. It gives tremendous contribution to natural language processing, specially in the areas of information extraction, question answering, text summarization, and text analysis.

Query-based summarization can be applied for complex question answering. A complex question can be defined as such type of question for which answer can be obtained by integrating and interpreting knowledge from single or multiple text documents. Here, answer can be given to the complex question by combining important information from each document in a cohesive and redundancy-free manner. The query-based text summarization process requires the query and one or more text documents to be given as input and the summary is obtained as output. The Fig. 32.1 shows an overview of query-based text summarization.

Spelling correction in query always plays an important role in information retrieval. We can consider spell checking as a preprocessing part of query-based text summarization. This indeed helps in finding more relevant and user focused query. Two types of errors can be found: one is nonword errors, these are not found in dictionary and the other is real-word errors, which are present in the dictionary. Real-word errors can be typographical and cognitive. In typographical error, the errors are introduced by mistake. For cognitive error, this error occurs when spelling of the word is not known. This happens for homophones. Example of such errors are piece and peace. The first word 'piece' means slice and the second word 'peace' means silence [2]. Spell checking and correcting help to get the summarized text document more accurately. In fact, correct words help in extracting more semantically similar sentences to get useful summary.

Related Work

Spelling task can be divided into error detection and correction. From the survey, it is found that 26% of spelling errors are done in web queries (https://web.stanford.edu/class/cs124/lec/spelling.pdf). For efficient retrieval, revised n-gram-based technique has been put forward, where n-gram statistics and lexical resources are used [3]. In this method, the authors try to generate ranked list of correction candidates to derive the most suitable candidate word for the incorrect word. Garaas et al. [4] suggest a personalized error correction system using neural networks. They train a simple feedforward neural network so that if the same error happens it can detect and give the proper word.

Islam et al. [5] use Google Web 1T 3-grams dataset for detecting and correcting incorrect real words. For string similarity measure, they use the longest common subsequence string matching algorithm with different normalizations and modifications. Their method tries to improve the detection and correction recall value of incorrect words. Detection recall value means detecting fraction of errors correctly and correction recall value means modifying fraction of errors correctly.

Duan et al. [6] propose a model to find out the incorrect words in online spelling correction. They try to provide all spell-corrected complete suggestions while the query words are adding. They try to train the Markov n-gram transformation model to analyze the user's spelling behavior. To enhance the efficiency of this transformation model, they study different techniques. Finally, for searching correct words, they use A* informed search algorithm. Different pruning and thresholding methods are being used to increase the result of A* algorithm.

A dictionary-based approach has been proposed by Amorim et al. [7] using an unsupervised method. They integrate anomalous pattern initialization and partition around medoids (PAM) clustering algorithms. Their result shows 88.42% success rate for challenging datasets.

Sharma et al. [8] propose a system to correct confused words if they are found to be contextually wrong. To identify and correct real-word errors, one phase of the algorithm applies trigram approach and the other phase applies Bayesian technique. Commonly used confused words set and brown corpus are applied for this system.

Though different spell checking and correcting methods have been found for query completion in information retrieval, our spell checking and correcting method is applicable particularly for query-based summarization purpose. Here, input text documents are used for correction of misspelled words.

An Approach for Spell Checking and Correcting

In this chapter, a method for correcting misspelled nonword errors is presented. In this approach, a dictionary-based spelling correction is done which also depends on input text documents.

Overview of Spell Correction Method for Query-Based Text Summarization

First, this method finds out the incorrect words using dictionary. After finding incorrect words, it searches for candidate words for these incorrect words. Candidate words are those words whose spellings are quite similar to the incorrect words and they are present in dictionary. These words are real words. Then the candidate words are filtered out by finding the highest n-gram character matching with the incorrect words. We use all these candidate words for finding matching words from the input text documents. If we find the matching words, then we will see if the highest n-gram character matching word and input text matching word is a unique word. If it is a unique word, then we consider that unique word as the correct word. If it is not a unique word, then we take both the highest n-gram character matching words and the input text matching words and finds the scores of those words on the basis of following similarity scores.

1. Minimum Edit Distance Score (MEDS): Edit Distance finds dissimilarity between two words. It calculates least number of operations necessary for transforming from one to another word. Here, Levenshtein distance [9] is used for finding the edit distance value. Levenshtein distance between two strings i, j is given by $D_{(i,j)}$ where
   Initialization:
   $D_{i,0} = i$
   $D_{0,j} = j$
   Recurrence Relation:

$$
\begin{aligned}
&\textit{For each } i = 1 \ldots M \\
&\textit{For each } j = 1 \ldots N \\
&D_{(i,j)} = min \begin{cases} D_{i-1,j} + 1 \\ D_{i,j-1} + 1 \\ D_{i-1,j-1} + 2; \begin{cases} \textit{if } X(i) \neq Y(j) \end{cases} \\ D_{i-1,j-1} + 0; \begin{cases} \textit{if } X(i) = Y(j) \end{cases} \end{cases} \\
&\textit{Termination}: \\
&D(N, M) \textit{ is the minimum edit distance value.}
\end{aligned}
\tag{32.1}
$$

   Now, the equation for edit distance score is

$$
s1 = \frac{\textit{no. of characters in the longest word}}{\textit{minimum edit distance value}} \tag{32.2}
$$

2. Character Similarity Measure Score (CSMS): It finds similar characters between two words. The percentage of similarity can be calculated as follows:

$$s2 = \frac{no.\ of\ similar\ characters}{no.\ of\ characters\ in\ the\ longest\ word} \tag{32.3}$$

3. Longest common substring score (LCSS): Here, longest common substring among words are found. The substring similarity can be found as follows: At first it finds out the longest common suffix. The required equation is as follows:

$$LCSuff(S_{1..p}, T_{1..q}) = \begin{cases} LCSuff(S_{1..p-1}, T_{1..q-1}) + 1 & if\ S_{|p|} = T_{|q|} \\ 0 & otherwise. \end{cases} \tag{32.4}$$

Now, this following equation is used to find the maximum longest common substring

$$LCSubstr(S, T) = \max_{1 \le i \le m, 1 \le j \le n} LCSuff(S_{1..i}, T_{1..j})$$

Finally, longest common substring score between two words can be found by the following equation:

$$s3 = \frac{length\ of\ longest\ common\ substring}{no.\ of\ characters\ in\ the\ longest\ word} \tag{32.5}$$

4. First Letter Weighting (FLW): Yannakoudakis and Fawthrop [10] surveyed that usually people do not make mistakes at first letter of a word while writing. Therefore, an extra weight is given to the words with similar first letter. Here, weight given to the first letter matching is 0.5.

The Steps of Spell Correction Method for Query-Based Text Summarization (SCQBT)

The pseudo-code for finding incorrect words and replacing with a correct word is as follows:

Experimental Data and Results

Experiments are performed on the datasets provided by Text Analysis Conference (TAC) (http://www.nist.gov/tac/data/). Here, text documents are taken from TAC 2009 datasets. There are 44 documents each having 2 topics. For each topic, there are 10 text documents.

Data: Query (Q_i) and Input Text (I)
Result: Correct Query ($Q_{correct}$)
Find the incorrect words ($Q_{incorrect}$) using dictionary
for *each incorrect word* In_w *in* $Q_{incorrect}$ **do**
  Generate the candidate words (C_w)
  Do n-gram character matching between In_w and C_w
  Take highest n-gram character matching words (N_w) from C_w
  if $C_w \in I$ **then**
    take those words (I_w);
  else
    do not take any C_w words;
  end
  if $N_w \cap I_w$ *has one unique word* **then**
    Take the unique word and replace with In_w ;
  else
    Take $N_w \cup I_w$
    while *each word* $W \in (N_w \cup I_w)$ **do**
      Calculate *MEDS* (In_w, W)
      Calculate *CSMS* (In_w, W)
      Calculate *LCSS* (In_w, W)
      Calculate *FLW* (In_w, W)
      Sum up all scores (*score*)
      Replace In_w with W having highest (*score*)
    end
  end
end

Algorithm 1: Steps of Spell Correction method for Query-Based Text Summarization (SCQBT)

Table 32.1 Candidate words list

Incorrect words	Candidate words
detal	'deal', 'dental', 'detail', 'dealt', 'delta',
	'metal', 'petal', 'fetal', 'decal'
chna	'tuna', 'china'
accidnt	'accident', 'accidence', 'acidness', 'accordant',
	'account'

Below, experimental process is described with a sample query and a input text file. Let us consider the sample query is "detail china accident". When the query is entered, it is written with spelling mistake as "detal chna accidnt". Now, dictionary indicates the incorrect words along with possible candidate words shown in Table 32.1.

Now, n-gram character matching is done for the candidate words. Here, value of n is 3. The list of values of n-gram character matching are shown in Table 32.2.

Each of the candidate words are searched in the text document provided. Here for incorrect word 'chna', we get one unique word which has highest n-gram character matching score and is present in input text document and the unique word is

Table 32.2 n-gram character matching values of candidate words

Incorrect words	n-gram value of candidate words
detal	'dental': 0.5, 'detail': 0.5, 'deal': 0.44, 'petal': 0.4,
	'metal': 0.4, 'fetal': 0.4, 'decal': 0.4, 'delta': 0.16,
	'dealt': 0.16
chna	'china: 0.4', 'tuna: 0.2'
accidnt	'accident', 0.58, 'account', 0.38, 'accordant', 0.33,
	'accidence', 0.33, 'acidness', 0.26

Table 32.3 Scores of filtered candidate words

Filtered candidate words	Score value
metal	6.6
dental	7.83
detail	8.0

Table 32.4 Comparison of performance with baseline systems

Method name	Accuracy (%)	Recall (%)	Precision (%)	F (%)
SCQBT	70	89.7	87.5	89.1
Microsoftword spell corrector	64	89.2	89	89.1
Character similarity	48	85	82.8	83.9
Minimum edit distance	42	84	80.8	82.4
Longest common subsequence	18	69	64.3	66.6

'china'. Similarly, for incorrect word 'accidnt' , we get the one unique words which is 'accident'. But for 'detal', we get input text matching word as 'metal' and highest n-gram character matching words as 'dental' and 'detail'. Hence, according to our method, total scores of these three filtered candidate words are calculated. Scores are given in Table 32.3. Finally, word replacement is done on the basis of highest score. Therefore, the word 'detal' will be replaced by 'detail'.

Here, we take 50 queries for evaluation purpose. Comparison of our method is done with baseline methods like Microsoft Word Spell Corrector, Character Similarity, Minimum Edit Distance, and Longest Common Subsequence. Following Table 32.4 shows the detailed results.

From the above Table 32.4, it is observed that our proposed algorithm SCQBT works well in-terms of accuracy. In fact, SCQBT performs well for various precision,

Table 32.5 Candidate words list

Incorrect words	Candidate words
polici	'polis', 'police', 'policy'

Table 32.6 n-gram character matching values of candidate words

Incorrect words	n-gram value of candidate words
polici	'police': 0.45, 'policy': 0.45, 'polis': 0.36

recall, and F-measures of Character Similarity, Minimum Edit Distance, and Longest Common Subsequence methods. Similarly, the recall and F-measure values of SCQBT method are higher than Microsoft Word Spell Corrector but its precision value is low.

Discussion

Here, our method is limited to nonword errors. Sometimes, we get the highest score for the incorrect words also. For example, we consider the sample query as 'India Pakistan peace polici'. Hence, for this query, incorrect word is polici. Initially, these following candidate words are found in Table 32.5.

Hence, we calculate the n-gram character matching values for the all candidate words shown in Table 32.6.

Now, we find 'police' word from the input text matching. Finally, 'police' unique word is found as the correct word from both the highest n-gram character matching word and the input text matching word, but this is not the correct word for the above query. Therefore, our method does not find the correct word for this incorrect query.

Conclusion and Future Work

This spell checking technique is mainly applied in query-based text summarization purpose. Here, the proposed method is based on dictionary. Our method tries to select most appropriate word from the candidate words provided by dictionary. Experimental results show better performance, however this work can be extended by adding corpus-based semantic similarity to get more improved results.

References

1. Hovy, E., Lin, C.Y.: Automated text summarization and the summarist system. In: Proceedings of a workshop on held at Baltimore, Maryland: October 13–15, 1998, Association for Computational Linguistics (1998) 197–214
2. Martin, J.H., Jurafsky, D.: Speech and language processing. International Edition **710** (2000)
3. Ahmed, F., Luca, E.W.D., Nürnberger, A.: Revised n-gram based automatic spelling correction tool to improve retrieval effectiveness. Polibits (40) (2009) 39–48
4. Garaas, T., Xiao, M., Pomplun, M.: Personalized spell checking using neural networks
5. Islam, A., Inkpen, D.: Real-word spelling correction using google web it 3-grams. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, Association for Computational Linguistics (2009) 1241–1249
6. Duan, H., Hsu, B.J.P.: Online spelling correction for query completion. In: Proceedings of the 20th international conference on World wide web, ACM (2011) 117–126
7. Cordeiro De Amorim, R., Zampieri, M.: Effective spell checking methods using clustering algorithms. In: Proceedings of Recent Advances in Natural Language Processing, Association for Computational Linguistics (2013)
8. Gupta, S., et al.: A correction model for real-word errors. Procedia Computer Science **70** (2015) 99–106
9. Haldar, R., Mukhopadhyay, D.: Levenshtein distance technique in dictionary lookup methods: An improved approach. arXiv preprint arXiv:1101.1232 (2011)
10. Yannakoudakis, E.J., Fawthrop, D.: An intelligent spelling error corrector. Information Processing & Management **19**(2) (1983) 101–108

Chapter 33
Study and Analysis of Different Face Recognition Techniques Based on Graph

S. Warjri, I. Wahlang and A. K. Maji

Abstract Human face recognition is very interesting as well as very challenging area of research. In our study of different existing face recognition schemes, we have seen that it has very important role in many applications. For this paper, we have studied and worked on the various steps of face recognition and analyzed a method to work on them. Where the face is detected from an input image and Gabor filter is applied. Then, two approaches are used (1) Delaunay triangulation and (2) Euclidean distance which are apply to the extracted feature points (which is referred as node set graph) and they are stored in the database. These store node set is used for checking the similarity with the input image. We have also compared both of the proposed approach in the system and we found that node set with Euclidean distance gives a better recognition rate.

Keywords Recognition · Gabor wavelet · Node set graph · Euclidean distance · Delaunay triangulation

Introduction

From the name "Face Recognition", we can understand that this is all about identifying or verifying the face of a particular person by the created system. Thus face recognition system is a security system based on computer. This system comprises of many face features which are stored in the database and the system is capable to decide that the input image is matched or not matched (Recognized/Not recognized),

S. Warjri (✉) · I. Wahlang · A. K. Maji
Department of Information Technology, North Eastern Hill University, Umshing, Shillong 793022, India
e-mail: sunitawarjri@gmail.com

I. Wahlang
e-mail: imayanwahlang@gmail.com

A. K. Maji
e-mail: arnab.maji@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_33

by comparing the feature of the image given as input with the features which is stored in the database. Human face recognition is interesting and very challenging area of research [1–4] in computer vision.

Steps for Face Recognition

Basically for face recognition system, a picture is captured from a camera, this picture acts as an input image where we have to detect a person's face from that image. Toward this goal, the procedure for face recognition is of three steps [5]:

1. Face detection.
2. Face feature extraction.
3. Face recognition.

Face Detection

This is the very first step toward any face recognition system, the main function is to find whether human faces appear in an input image, and if present, extract the face location. This step is useful for analyzing the information contained in faces. Several techniques are there for face detection, they are Viola and Jones face detector [6], neural network [7], support vector machine [8], Gabor wavelet [9], skin color method [10], etc. (Fig. 33.1).

Feature Extraction

The main aim of feature extraction is to extract relevant information from the detected face after face detection. Most of the feature extraction techniques used are local and global features: (1) Local features—It is used to automatically extract the local information from the face (like eyes, mouth, nose, etc.) and (2) Global

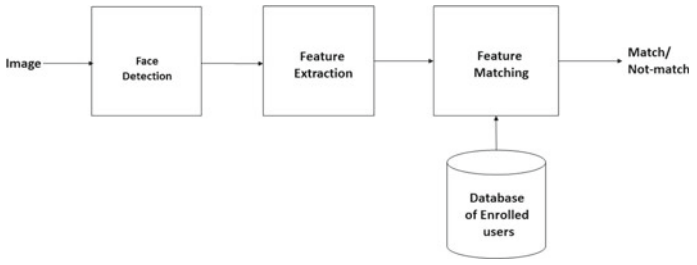


Fig. 33.1 The face recognition basic procedure

features—It is used extract the global information from the face like the shape of face, color of face, hair color, etc. The image features are divided into four categories [11]: visual features, statistical pixel features, transform coefficient features, and algebraic features. The approaches used for feature extraction are geometry-based [12], template-based [13], color segmentation [14], and appearance-based approaches [15].

Face Recognition

After extracting the features of the input image, the last step is to identify/verify the faces. For this step, a face database is required to build. The face database consists of several images of different persons and their features that are extracted and it is stored in that database. In this step, comparison of the features of input face image takes place with the features which are stored in the database. For comparing these features, the features matching algorithm is used.

A Review on Face Recognition

There had been many works based on the different stages and processes for face recognition. For face detection, according to [16], a captured image is given as input, the main aim of face detection algorithm is to recognize the location and scale of the face in the input image. From several studies, we found that there are many challenges for face detections such as posture and scale, face orientation, facial expression, ethnicity, and skin color [17, 18]. According to [19], the general methods for face detection can be classified into feature-based approach and image-based approach. For feature extraction, according to [20], facial feature extraction is used to find the most appropriate points of the face images for face recognition. The features that are extracted can be color of face, eyes, mouth, nose, chin angle, etc. For the facial feature, one can also extract the facial expressions to have a good recognition for a system, the six basic facial expressions are sadness, anger, surprise, happiness, fear, and disgust [21]. For face recognition techniques, we have studied several algorithms which can be used for matching the face like elastic bunch graph matching [22–26], eigenface method [4, 27, 28], fisherface method [29, 30], etc. The main difficulty in face recognition includes [22] physical changes, acquisition geometry changes, and imaging changes. There are several methods which are used for these problems like feature-based, image-based approaches, and appearance-based algorithms [14]. Table 33.1 represents the feature points which are extracted and the methods which are used for their extraction.

Table 33.1 A review on some methods which are used for feature extraction and the facial features which are extracted

Author	Method	No. of features	Advantage or Disadvantage
Y. Benjemaa and S. Khanfir,	Gabor wavelets	Eyes, mouth and nose	Method achieves good recognition rate (99.98%)
B. Gupta, S. Gupta, A. K.Tiwari	Gabor feature extraction and artificial neural network	Eyes, nose, mouth and order characteristics such as dimples, moles, etc.	Have a good face detection rate which can be used for face recognition
Dr. S. Vijayarani and S. Priyatharsini	Face part detection (FPD) algorithm and gray level cooccurrence matrix (GLCM) algorithm	Eyes, eyebrows, and lips	GLCM feature extraction algorithm performance is comparatively better than FPD
M. Pantic, M. Tomc and L. J. M. Rothkrantz	A hybrid technique and knowledge based approach	Upper lip lower lip, left and right mouth corners	It can deal only with limited out of plane head rotations and sequences starting with an expressionless mouth appearance
Eli Saber and A. M. Tekalp	Color, shape, and symmetry-based cost functions	Tip of nose, eyes, and center of mouth	This can be employed as a preprocessing step to a face recognition system

Face Recognition Techniques Based on Graph

A face recognition system based on the local feature extraction by using Gabor wavelet transform and a simple face node graph matching algorithm is used for matching the similarity of the image. The Gabor wavelet transform is used for facial features extraction in the region of interest (ROI) like eyes, eyebrow, nose, and mouth. Where the face node set graph is collection of these extracted feature points with some information about their relative location which constitutes an image graph [24].

Node set: Node set is the extracted feature set points from the ROIs, where each node represents an individual fiducial point. These extracted fiducial points are stored in a database and is referred as a node set graph.

Procedure of the Proposed Work

1. Perform face detection to the input image (I).
2. Perform preprocessing to the detected region of interest (ROI).
3. Apply Gabor wavelet transform to ROI and generate the node set.
4. Generate the node set graph (i.e., by performing 1. Delaunay triangulation to the node set and 2. Euclidean distance to the node set) and store it in database.
5. For node set graph of the input image (I) determine the most similar node set graph from the database where the node set graph is stored.
6. For checking the similarity of the image, we have computed simple matching algorithm by using superimposition and by putting some threshold value to consider the input image match or not match with the image feature stored in the database.

Implementation Details

First, we capture the image by digital camera, then we register the image in the database where we have used Mysql database. For the face detection from the input image, we have used the traditional Viola–Jones object detection framework. After face detection, the same Viola–Jones framework is used to detect the facial regions of interest (e.g., eyes, nose, mouth, etc.). The eyebrows region was estimated from the eye region position. Then several image processing methods are used to improve the image quality (e.g., histogram equalization, intensity adjustment, Gaussian filtering, etc.).

Then Gabor wavelets are computed in order to extract the local information from ROI and then fiducial points are extracted. In our work, the facial features that were extracted from the face are eyes, eyebrows, nose, and mouth. These extracted feature points are referred as a node set graph which constitutes an image graph and it is stored in the database.

For example, below some of the preprocess images of the regions of interest (ROI) and its Gabor magnitude are shown in Fig. 33.2.

Then, we have used two methods to get the node set graph which is stored in the database by computing the Delaunay triangulation to the node set and Euclidean distance to the node set. Then, the node set graph of the input image is compared with the node set graph that is stored in database by using the graph matching algorithm.

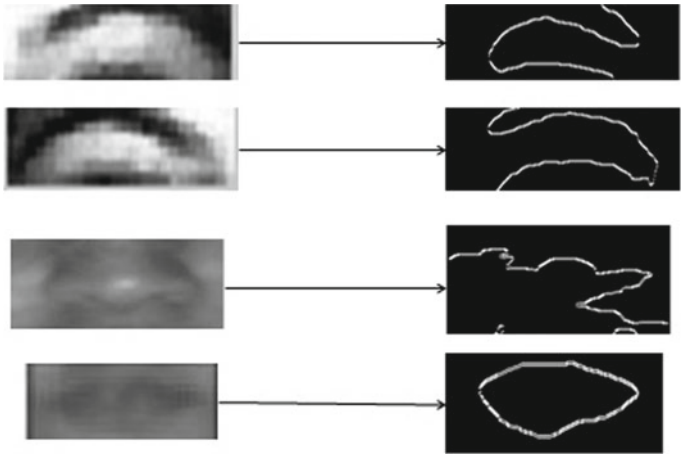


Fig. 33.2 The preprocess image of the region of interest and its Gabor magnitude

Experimental Result of the Proposed Work

For testing the system, we have register 32 number of faces images of 32 different persons. Then, we have stored the extracted node set graph of all the images and then we perform a matching algorithm to the node sets graph. And it is found that Euclidean distance node set graph gives a better result after matching than the Delaunay triangulation node sets. The result is shown in Table 33.2 (Fig. 33.3).

Recognition Rate

For checking the recognition rate of the system, we have also tested the matching algorithm on both of the extracted node set graph, with the other existing technique.

Table 33.2 The table shows the performance analysis of the stored images

Matching the features	Number of images	Recognized image	Non-recognized image	Running time (s)
Delaunay triangulation node sets graph	32	26	6	30.514
Euclidean distance node sets graph	32	32	0	29.116

Fig. 33.3 These are 32 face images that are stored in database

Table 33.3 Comparison analysis of recognition rate with other existing system

Technique	Recognized face	Unrecognized face	Recognition rate (%)	Computational cost (s)
Delaunay triangulation node set graph	7 out of 8	1 out of 8	87.50	30.514
Euclidean distance node set graph	8 out 8	0 out of 8	99.98	29.816
Gabor and contour	8 out of 8	0 out of 8	100	37.087
Linear discriminant analysis	6 out of 8	2 out of 8	75	92.124
Principal component analysis	5 out of 8	3 out of 8	62.50	71.867

For testing, we have registered several face images with different poses, expression, and illuminations of a person. The comparison table is shown below. For testing the system, we have registered eight different faces of the same person (Table 33.3).

Performance Analysis of False Rejection Rate

We have also tested both of the proposed system for the false rejection rate (this is an error rate for comparing the face image with various posture of the same person). We have stored the different frontal faces of different persons with different poses and expressions for the test. Tables below represent the recognition rate of both of the proposed methods where we have used 26 faces of four different people and we found that the Delaunay triangulation rejection rate is 37.16%, whereas the Euclidean distance of the rejection rate is 5.07%.

A. Delaunay Triangulation Node Set Graph

See Table 33.4.

B. Euclidean Distance Node Set Graph

See Table 33.5.

Table 33.4 The performance analysis false rejection rate using Delaunay triangulation node set graph

Frontal face database	No. of faces tested	Detection rate (%)	Tested face	False rejection rate (%)
Julie face	7 faces	43	Julie face no.1	57
Michel face	8 faces	75	Michel face no.1	25
Jacinta face	5 faces	66.33	Jacinta face no.1	33.33
Cilia face	6 face	66.33	Cilia face no.1	33.33
Average rate	26	62.66	4 people	37.16

Table 33.5 The performance analysis false rejection rate using Gabor with Euclidean distance node set graph

Frontal face database	No. of faces tested	Detection rate (%)	Tested face	False rejection rate (%)
Julie face	7 faces	100	Julie face no. 1	0
Michel face	8 faces	100	Michel face no. 1	0
Jacinta face	5 faces	97	Jacinta face no. 1	17.3
Cilia face	6 face	100	Cilia face no. 1	3
Average rate	26 faces	99.25	4 people	5.07

Summary and Conclusions

In this work, we have used Viola–Jones object detection for face detection, then Gabor filter is utilized to extract the local feature from the ROI of the input face image. Then, we have used two approaches to get the node sets graph and these node sets are stored in database and a matching is done. After the matching of both the methods, we found that the Euclidean distance note set graph gives more better result than Delaunay triangulation note set graph. Hence, we can also see that the proposed system has a better recognition rate when comparing with the other existing system.

References

1. Belhumeour, P. N., Hespanha, J. P., Kriegman, D. J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19(7), pp. 711–720, (1997).
2. Tacacs, B., Wechsler, H.: Face recognition using binary image metrics. In Pro. Third International Conference on Automatic Face and Gesture Recognition, pp. 294–299, (1998).
3. Turk, M. A., Pentland, A. P.: Face recognition using eigenfaces. In Pro., Eleventh International Conference on Pattern Recognition, pp. 586–591, (1991).
4. Gokmen, A., Yilmaz, M.: Eigenhill vs. Eigenface and Eigenedge. Pattern Recognition, Vol. 34, No. 1, pp. 181-184,(2001).
5. Riddhi, P., Y.S.B.: A Literature Survey on Face Recognition Techniques. International Journal of computer Trends and Technology (IJCTT), vol. 5, No. 4 pp. 189–195, (2013).
6. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Proc. IEEE Conf. computer Vision and Pattern Recognition, pp. 511–518, (2001).
7. Rowley, H.A., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. Computer Vision and Pattern Recognition, (1996).
8. Guo-yun, Z., Long-yuan, G., Jian-hui, W., Hong-min, L., Guan-qi, G.: Face Detection Using Gabor Wavelet and SVM. International Conference on Computer Application and System Modeling (ICCASM), (2010).
9. Sahoolizadeh, H., Sarikhanimoghadam, D., Dehghani, H.: Face Detection using Gabor Wavelets and Neural Networks. International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, Vol. 2(9), (2008).
10. Tayal, Y., Lamba, R., Padhee, S.: Automatic Face Detection Using Color Based Segmentation, International Journal of Scientific and Research Publications, Vol. 2(6), (2012).
11. Ismaila, W.O., Adetunji, A.B., Falohun, A.S.: A Study of Feature Extraction Algorithms for Human Face Recognition. Transnational Journal of Science and Technology, Vol. 2(6), (2012).
12. Kanade, T.: Computer Recognition of Human Face. Basel, Stuttgan and Birkhauser, (1977).
13. Yuille, A., Cohen, D., Hallinan, P.: Facial feature extraction from face using defomable templates. Proc. IEEE Conf. computer Vision and Pattern Recognition, pp. 104–109, (1989).
14. Chang, T.C., Huang, T.S., Novak, C.: Facial Extraction from color images. Proc. 12th ICPR International Conference on Pattern Recognition, vol. 2, pp. 39-43, (1994).
15. Tian, Y., Kanade, T., Cohn, J.F.: Evaluation of Gabor-wavelet-based facial action until recognition in image sequences of increasing complexity. Proc. 5th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 218–223, (2002).
16. Mounika, B.R., Reddy, N.J., Reddy, V.B.D.: A Neural Network Based Face Detection Using Gabor Filter Response. International Journal of Neural Networks (ISSN: 2249-2763 & E-ISSN: 2249-2771), Vol. 2, Issue 1, pp. 06–09, (2012).

17. Phung, S. L., Chai, D., Bouzerdoum, A.: A universal and robust human skin color model using neural networks. Proc. IJCNN'01, pp. 2844–2849, (2001).
18. Phung, S. L., Bouzerdoum, A., Chai, D.: Adaptive skin segmentation in color images. Proc. ICASSP'03, (2003).
19. Hatem, H., Beiji, Z., Majeed, R.: A Survey of Feature Base Methods for Human Face Detection, International Journal of Control and Automation, Vol. 8, No. 5, pp. 61–78, (2015).
20. Jemaa, Y. B., Khanfir, S.: Automatic local Gabor features extraction for face recognition. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 3, No. 1, (2009).
21. Kotsia, I., Pitas, I.: Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. IEEE Transactions On Image Processing, Vol. 16(1), (2007).
22. Tiwari, M.: Gabor Based Face Recognition Using EBGM and PCA. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3), 4322–4326, (2012).
23. Jayakumar, D., Sandeep, R.: Face Recognition Using Elastic Bunch Graph Matching. International Journal For Technological Research In Engineering, Vol. 2, Issue 11, (2015).
24. Wiskott, L., Fellous, J.M., Kruger, N., Malsburg, C.V.: Face Recognition By Elastic Bunch Graph Matching. Intelligent Biometric Technique in Fingerprint and Face Recognition, pp. 355–396, (1999).
25. Zafeiriou, S., Petrou, M.: 2.5D Elastic graph matching. Computer Vision and Image Understanding, 115, pp. 1062–1072, (2011).
26. Senaratne, R., Halgamuge, S., Hsu, A.: Face Recognition by Extending Elastic Bunch Graph Matching with Particle Swarm Optimization. Journal Of Multimedia, Vol. 4, (2009).
27. Levine, Z. M., Hafed, M.: Face Recognition Using the Discrete Cosine Transform. International Journal of Computer Vision, Vol. 43, No. 3, pp. 167–188, (2001).
28. Venkatesh, D., Ramasubramanian, Y.V. : Encoding and Recognition of Faces Based on the Human Visual Model and DCT. Pattern Recognition, Vol. 34(8), pp. 2447–2458, (2001).
29. Wechsler, C., Liu, H.: Enhanced Fisher Linear Discriminant Models for Face Recognition. Proceedings of International Conference on Pattern recognition, (1998).
30. Kak,A. M., Martinez, A. C.: PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23(2), pp. 228–233, (2001).

Chapter 34
Extraction and Analysis of Expression from a Captured Face

Imayanmosha Wahlang, Sunita Warjri and Arnab Kumar Maji

Abstract With the help of computer science in today's world many advances have been made in the field of facial expression recognition. It is considered as an important measure of social interactions and there are many applications that can be helpful to human in many ways as facial expressions not only show the emotional status of an individual but also other mental activities as well. To develop such a system is a challenging task as it involves learning and understanding of expressions of human which varies from time to time. In this paper, a method on facial expression analysis has been proposed where face detection was done using RGB color model and feature classification using Gabor filter.

Keywords Face detection · Feature extraction · Proposed facial expression analysis

Introduction

Overview

Facial expression recognition is important to identify facial expressions of human and in human–machine interfaces. It is one of the latest research topics whereby a face is detected and its expression could be extracted and analyzed which can be used as a tool against crime, psychology, and other related fields. Even robots that can detect human expressions were being made. Therefore, it can prove to be a vital

I. Wahlang (✉) · S. Warjri · A. K. Maji
Department of Information Technology, North Eastern Hill University, Shillong 793022, India
e-mail: imayanwahlang@gmail.com

S. Warjri
e-mail: sunitawarjri@gmail.com

A. K. Maji
e-mail: arnab.maji@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_34

and essential field. There are many other applications as well like in behavioral science and medical science [1–3].

Facial Expression Classification

Human communication is said to be of two types: verbal where communication is through conversation between two or more individuals and nonverbal where communication is through sign language or some other mode of communication. Expression of a person is considered to be nonverbal and it is said to be the main focus during a conversation. It is the source of information where we can predict about the mood or even character of a person. From an expression, we can know the feelings of a person which can be positive or negative in nature. A positive expression is the one that brings in peace and joy to the surrounding and a negative expression is believe to bring about sadness, grieve, pain, suffering, and all sorts of negativity. In most of the research, researchers have considered the six basic emotions which are happy, sad, anger, disgust, fear, and surprise [3, 4]. In addition to the above-mentioned expressions, other expressions are also available. Here, two more familiar expression were studied which are wink and sleepy.

The existing procedure in facial expression classification involves face detection, feature extraction, and facial expression classification [5, 6]. Other steps involved are image enhancement and other methods of improving the image quality. Facial expression analysis was studied and given by many researchers using different techniques. One of the existing facial expression analysis technique involves two phases which are training and testing phase where in training phase, the different classes of expressions were taken and feature vectors were obtained using local binary pattern (LBP) and features were reduced using principal component analysis (PCA) and were stored in a database. In testing phase, LBP features were obtained with its histogram and some error calculation were performed and classification of expression was done using LBP features. The drawbacks that were found are that it uses only frontal image and works on 2D system and cannot work properly in presence of occlusions [7]. Another existing method was done using multi-class AdaBoost learning algorithm and was found that the problems are due to low resolution and it cannot correctly identify some complex expression [8]. Some other techniques are also there where 3D method was used as in [9].

In this paper, the things that are included are face detection where an input image was taken and a face was detected followed by feature extraction and then analysis of facial expression.

Related Work

Face Detection

Face detection is the first step toward facial expression analysis. It is a process in which a face is detected. It gives the location of a face in an image. Some of the algorithms for face detection are Viola–Jones algorithm [10] and AdaBoost algorithm [5, 10]. Techniques that can be used are local binary pattern (LBP) [5], skin color modeling like RGB, HSV, YCbCr [10], and others.

Here, we used skin color model as a technique for face detection. Skin color provides an effective result and it is much more simpler than any other method. Many skin color techniques have been developed in the past few years. The default color space is RGB and other color model can be obtained from it with some modifications [11]. Methods like RGB, YCbCr, and HSV were studied and implemented.

RGB

In RGB color model, the image was taken and was classified into skin and nonskin region based on the values given in Eq. 34.1 [12, 13].

$$\begin{aligned} \text{Skin} = \text{R} > 95 \text{ and } \text{G} > 40 \text{ and } \text{B} > 20 \text{ and } \max\text{R, G, B} > 15 \\ \text{and } |\text{R} - \text{G}| > 15 \text{ and } \text{R} > \text{G} \text{ and } \text{R} > \text{B} \end{aligned} \tag{34.1}$$

YCbCr

In this color model, RGB color space was first converted into YCbCr color space. Y denotes luminance component and Cb and Cr denote chrominance component. An image is classified into skin region based on the values of Cr and Cb as given in Eq. 34.2 [14].

$$\begin{aligned} 132 < \text{Cr} < 17 \\ 76 < \text{Cb} < 127 \end{aligned} \tag{34.2}$$

HSV

This color model includes H which is hue, S is saturation, and V is value. RGB color space was converted into HSV color space. The H is set as in Eq. 34.3 [15] whereby if it lies between the given values it is classified as skin region otherwise it is a nonskin region.

$$0 \leq \text{H} \leq 50, 0.20 \leq \text{H} \leq 0.68, 0.35 \leq \text{H} \leq 1.0 \tag{34.3}$$

Facial Feature Extraction

There are varieties of features that are available in an image like color and shape [16]. Based on these features, many techniques have come into existence. The techniques available for extracting of features are principal component analysis (PCA) [5], Gabor filter [17], and least discriminant analysis (LDA) [5].

After the face was detected using RGB color model, the next step is to extract the features. Gabor filters were used to extract the features. Gabor filters are known for their invariance to rotation, scale, and translation [16]. In this technique, an image was taken both in color and grayscale image. If an image is a color image, it is converted into grayscale image. A two-dimensional Gabor filter was used and given as shown in Eq. 34.4. It is a Gaussian kernel function modulated by a complex sinusoidal wave. Gabor filter helps in filtering of an image by applying different frequencies, scales, and orientations. Here, five scales and eight orientations were used. Downsampling was done to remove and reduce redundancy. After downsampling, since the feature vectors that were obtained were numerous, reduction can be done using dimensionality reduction techniques [17, 18]. This method is simple and it is known for using in texture analysis but it is not that accurate.

$$g(x, y) = \frac{f^2}{\pi\gamma\eta}\exp\left(-\frac{x'^2 + \eta^2 y'^2}{2\pi\sigma^2}\right)\exp\left(j^2\pi x' + \varphi\right) \tag{34.4}$$

$$x' = x\cos\theta + y\sin\theta$$
$$y' = x\sin\theta + y\cos\theta$$

Proposed Facial Expression Extraction and Analysis

Facial Expression Overview

Facial expression analysis has gained a lot of interest in recent days. Neutral, anger, disgust, fear, happiness, sadness, and surprise [7] are the common facial expressions that were being used mostly. Depending on different researchers, some uses other expressions like wink, sleepy, sick, and so on. Facial expression analysis is the last step of classifying expressions where an image will be classified into different expressions which are happy, neutral, sad, sleepy, surprise, and wink in our case as shown in Fig. 34.1.

Fig. 34.1 Different facial expressions

Proposed Method

In our proposed method expressions are classified using Yale database. It is a database consisting of 165 grayscale images in GIF format of 15 individuals where there are 11 images for each individual showing the different facial expressions which are happy, sad, surprise, sleepy, wink, normal, center light, left light, right light, with glasses, and with no glasses. The method used is based on file comparison. Yale database has been used for testing and classifying of expressions. The proposed method is shown in Fig. 34.2.

The proposed method involves two stages as described below:

In the first stage, all images of Yale database were tested where extraction of the feature vectors was done using Gabor filter and the feature vectors are stored in a separate files for each image. The feature vectors are reduced by keeping only unique values and are sorted. Comparison of files of the same expression was done and thresholding was applied. Similarity between files was found using Jaccard coefficient and by file matching using intersection between the two files. A threshold is applied depending on the difference of the first and last values of the feature vectors. Feature vectors obtained for different image of same expression were compared to classify into different expressions and in Fig. 34.3 Happy expression is shown where all other expressions follow the same procedure. Here, n is the total number of images used for happy expression. Depending on the

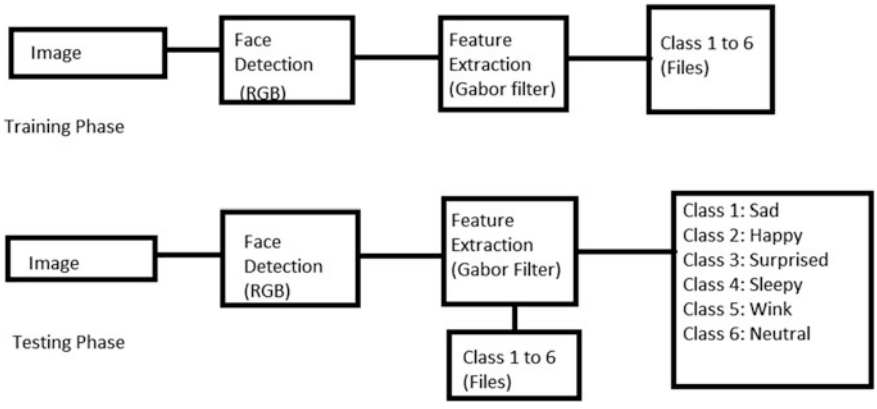


Fig. 34.2 Proposed expression classification method

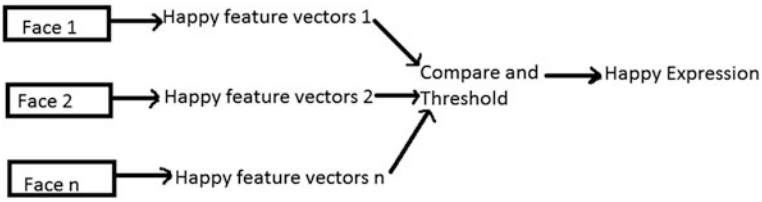


Fig. 34.3 Process involving in classifying of happy expression

threshold that was applied, expression was classified into happy, sad, surprise, sleepy, wink, and neutral. The threshold that was used for happy was $2.15 \leq t \leq 2.19$, sad was $1.76 \leq t \leq 1.90$ and $2.05 \leq t \leq 2.10$, surprise was $1.95 \leq t \leq 2.04$, sleepy was $2.20 \leq t \leq 2.34$, wink was $1.91 \leq t \leq 1.94$ and $2.35 \leq t \leq 2.69$, and neutral was $2.10 \leq t \leq 2.15$ and $2.70 \leq t \leq 2.80$.

In the second stage, an image that has been captured was taken where face detection using RGB was done and the same step was repeated as in stage one. Depending on the threshold values, the image will be classified accordingly into different expressions.

After successfully implementing the proposed method, it was found that for the same face it is easier to classify expressions as some values varies for every expression of the same individual but for different faces it is hard to classify correctly as the values obtained are of the same for two or three expressions.

Advantages:

- It is a simple approach.
- Computationally simple than any other method or techniques.

Disadvantages:

- The proposed methodology is limited to classify frontal image only.
- Using larger database will consume a lot of time for checking similarities and differences between files.
- Not accurate.

Experimental Result

Output of RGB showing a detected face is shown in Fig. 34.4a which shows the original image of an input and Fig. 34.4b shows the detected face after Face Detection.

Feature extraction was done and the result obtained is shown in Fig. 34.5 where it shows Gabor filter having five scales and eight orientations. Figure 34.5a shows the magnitude of Gabor filter and Fig. 34.5b shows the real part of the Gabor filter.

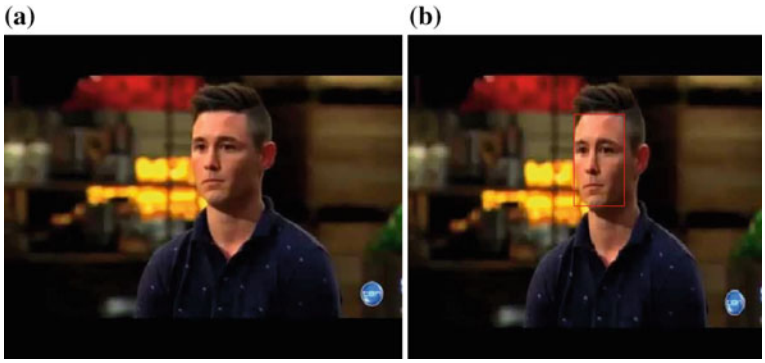


Fig. 34.4 **a** Original image. **b** Detected face

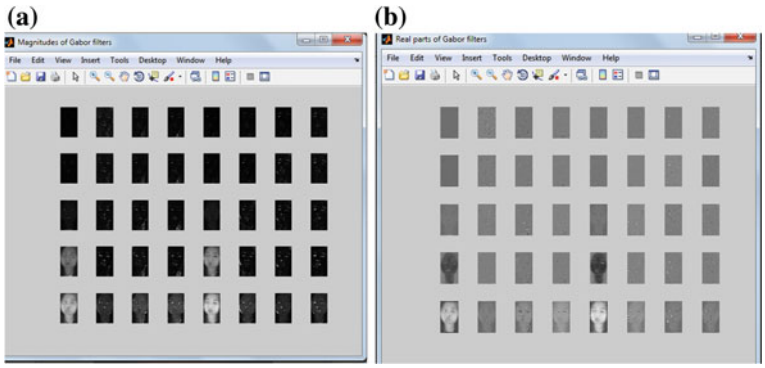


Fig. 34.5 **a** Magnitude of Gabor filter. **b** Real part of Gabor filter

The table showing rate of recognition for different expressions namely sad, happy, surprise, sleepy, wink, and neutral is shown in Table 34.1 where it can be seen that for neutral expression recognition rate is the least and for sad and sleepy expression recognition rate is maximum.

Table 34.1 Recognition rate for different expressions

Expression	Recognition rate (%)
Sad	86.66
Happy	73
Surprise	80
Sleepy	86.66
Wink	80
Neutral	66

Table 34.2 Overall classification rate of different expressions

Authors	Methods	Average
P. Lucey et al. (2010)	AAM + SVM	66.7
S. Khandait et al. (2011)	AFER	95.26
S. W. Chew et al. (2011)	CLM + SVM	74.4
L. A. Jeni et al. (2012)	CLM + SVM (AU0 norm.)	77.6
M. Islam et al. (2013)	LM2 + SVM	87.0
H. Boughrara (2014)	MLP	96.66
S. Mahto et al. (2015)	LBP + KNN	96
N. Perveen et al. (2016)	RST + KNN	90
Proposed method	Gabor filter + file based	77.0

Fig. 34.6 Obtained expression showing sad and happy expression

In Table 34.2, comparison showing the overall accuracy between different existing techniques, including the proposed method, is given.

42 faces were captured and used for classification of different expressions. Some of the result are shown in Fig. 34.6 which show sad and happy expressions.

Conclusion and Future Work

As the objective suggests that the goal of this project is to extract the different facial expressions of human so as to make use of it in some of the related field. Here, face was detected using RGB color model and features were extracted using Gabor filter and classification of expression was done using simple file comparison. The result

of proposed method is not that accurate, as only 77% average has been obtained. It has been found that for every expression, the obtained values are more or less the same. Therefore, it is difficult to obtain a very good result by only keeping a threshold value. More research and analysis should be done relating to this topic using some other methods where result could be optimum.

References

1. Dam, A.V.: Beyond Wimp. In: IEEE Computer Graphics and Applications, Vol. 20, No. 1, pp. 50–51, IEEE press, CA (2000).
2. Bruce, V.: What the Human Face Tells the Human Mind: Some Challenges for the Robot-Human Interface. In: Proc. IEEE International Workshop on Robot and Human Communication, pp. 44–51, IEEE Press, Tokyo (1992).
3. Rani, J., Garh, K.: Emotion Detection Using Facial Expressions-A Review, In: International Journal of Advanced research in Computer Science and Software Engineering, vol. 4, pp. 465–466 (2014).
4. Tatepamulwar, C.B., Pawar, V.P., Fadewar, H.S.: Techniques for Facial Expression Recognition. In: International Journal of Advanced Research in Computer Science and Software Engineering Paper, Vol. 4, Issue 3, pp. 56 (2014).
5. Hemalatha, G., Sumathi, C.P.: A Study of Techniques for Facial Detection and Expression Classification. In: International Journal of Computer Science & Engineering Survey (IJCSES) Vol. 5, No. 2, pp. 30–31 (2014).
6. Surbhi, M., Arora, V.: A Face Identification Technique for Human Facial Image. In: International Journal of Computer Science and Information Technologies (IJCSIT), vol. 3, pp. 5391 (2012).
7. Happy, S.L., George, Routray. A.: A real time facial expression classification system using Local Binary Patterns. In: Intelligent Human Computer Interaction (IHCI), 4th International Conference on IEEE. (2012).
8. Wang, Y., Ai, H., Wu, B., Huang, C.: Real time facial expression recognition with adaboost. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference, Vol. 3, pp. 926–929, IEEE Press, Cambridge (2014).
9. Soyel., Hamit., Demirel, H.; Optimal feature selection for 3D facial expression recognition using coarse-to- fine classification. In: Turkish Journal of Electrical Engineering & Computer Sciences, vol. 18, Turkey (2010).
10. Gupta, V., Sharma, D.: A Study of Various Face Detection Methods: In: International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6694 (2014).
11. Kakumanu, P., Makrogiannis, S., Bourbakis, N.; A survey of skin-color modeling and detection Methods. In: Pattern recognition, Vol. 40, No. 2, pp. 110–117 (2007).
12. Hsieh, C.C., Liou, D.H, Lai, W.R.: Enhanced Face Based Adaptive Skin Color Model. In: Journal of Applied Science and Engineering, Vol. 15, No. 2, pp. 167–168 (2012).
13. Hajraoui, A., Sabri, M.: Face Detection Algorithm based on Skin Detection, Watershed Method and Gabor Filter. In: International Journal of Computer Applications, Vol. 94, No 6, Morocco, pp. 34 (2014).
14. Kaur, A., Kranthi, B.V.: Comparison between YCbCr Color Space and CIELab Color Space for Skin Color Segmentation. In: International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, Vol. 3– No. 4, pp. 30, New York (2013).

15. Xiang, F.H., Suandi, S.A.: Fusion of Multi Color Space for Human Skin Region Segmentation. In: International Journal of Information and Electronics Engineering, Vol. 3, No. 2, pp. 172 (2013).
16. Tian, D. Ping.: A review on image feature extraction and representation techniques. In: International Journal of Multimedia and Ubiquitous Engineering, vol 8, pp. 385, 387–388. (2013).
17. Jema, Y.B., Khanfir, S.: Automatic local Gabor features extraction for face recognition. In: (IJCSIS) International Journal of Computer Science and Information Security, Vol. 3, No. 1 (2009).
18. Haghighat, M., Zonouz, S. Mottaleb. M.: CloudID Trustworthy cloud-based and cross-enterprise biometric Identification. In: Expert Systems with Applications, vol. 42, pp. 7909 (2015).

Part V
Machine Learning and Natural Language Processing

Chapter 35
Machine Learning Based Comparative Analysis for the Classification of Earthquake Signals

D. S. Parihar, Ripul Ghosh, Aparna Akula, Satish Kumar and H. K. Sardana

Abstract This research aims at classifying earthquake signals from seismic noises caused due to anthropogenic activities. We aim at designing a seismic classifier for classifying true earthquake signals so as to reduce the false alarms thereby avoiding excessive data logging due to cultural noise. Based on theoretical and experimental consideration, a set of time and frequency domain features are extracted and used as features to train the supervised classifier network, viz., *k*-nearest neighbor (*k*-NN), maximum likelihood (ML), artificial neural network (ANN), and support vector machine (SVM). Two datasets were used in this research work K-NET (Kyoshin Network), Japan and strong motion seismic data recorded at CSIR-CSIO, Chandigarh using BASALT accelerograph of Kinemetrics Inc. Comparative analysis of the classifiers shows that SVM outperforms the other methods with an accuracy of 99.60%.

Keywords ANN · Classification · Earthquake · Feature extraction *k*-NN · ML · Noise · SVM

D. S. Parihar · R. Ghosh (✉) · A. Akula · S. Kumar · H. K. Sardana
CSIR-Central Scientific Instruments Organisation (CSIR-CSIO), Chandigarh 160030, India
e-mail: ripul.ghosh@csio.res.in

D. S. Parihar
e-mail: devendra.jec@gmail.com

A. Akula
e-mail: aparna.akula@csio.res.in

S. Kumar
e-mail: satish@csio.res.in

H. K. Sardana
e-mail: hk_sardana@csio.res.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_35

Introduction

Earthquake is a natural disaster, costing loss of lives and properties. The increasing number of skyscrapers and growing infrastructure in earthquake-prone urban areas are vastly affected by the tremors. Earthquake occurrence is sensed using seismic sensors by detecting the primary waves which travel faster as compared to disastrous secondary wave [1]. To detect the seismic signal and classify them as an earthquake or local noise is one of the problems in an earthquake warning system (EWS). EWS is designed to disseminate the information related to ongoing or impending earthquakes in the form of public alerts in real time. The external noise added to signal reduces the signal-to-noise ratio (SNR), thereby degrading the performance of the system by increasing the false alarms. Hence to make EWS robust, optimal features should be extracted from the signals which will have discriminative information.

In recent years, researchers have addressed the issues in the discrimination of seismic wave and man-made explosion [2]. In earthquake and underwater fisherman's explosion [3], classification is performed using supervised neural algorithms based on multilayer perceptron (MLP). For classifying local seismic signals from volcano-tectonic earthquake [4], predictor coding technique is used to extract the features and neural network based MLP as classifier. An automatic discrimination among landslide, explosion and microtremor has been analyzed using both supervise and unsupervised NN in [5]. However, these works have reported results on limited datasets and to comment on their generalized performances they are needed to be evaluated on larger datasets.

In this work, a comparative analysis has been devised among four existing classifiers (k-NN, ML, ANN, SVM) using larger datasets to discriminate the patterns of earthquake and seismic noise data. Preprocessing techniques have been used to denoise the signals and obtained features from larger data set to train the existing classifier networks. A comparative analysis is put together in investigating improved accuracy of classifiers.

Related Work

The use of integrated classification machine (ICM), a hierarchy of ANN has been reported for classification of earthquake and man-made explosion [2]. For the robust classification of noisy nonstationary signals, redundant classification environment (RCE) was created. The main factor influencing the classification performance was time-frequency resolution and smoothening level. The use of linear prediction coding (LPC) as a feature extractor instead of the conventional Fourier spectral analysis for discrimination of earthquake and underwater explosion have been reported in [3], whereas in [4], both LPC and function of amplitude are used for feature extraction and realized better performance due to amplitude features.

Using multilayer perceptron, over 343 data set of seismic signals demonstrated the accuracy of classifier up to 94% [6]. In volcano seismic signal detection and classification, Hidden Markov models are used over optimal features of seismic signals and obtained 80% accuracy [7]. Machine learning based classifiers such as SVM, *k*-NN, and ANN classifiers have been also used for classification of vehicles using seismic and acoustic signals by extracting short time Fourier transform (STFT) features [8]. This type of feature extraction method evolved a combination of both time, frequency, and joint time-frequency based features such as spectral statistics and wavelet coefficients characterization (SSWCC) for classification of military vehicles [9]. Similar research works related to feature extraction have been reported in EEG signal [10], audio radar signal [11], and image recognition [12].

Experimental Dataset

Two sources of datasets are used in this study. First, data set comprises of data recorded at CSIR-CSIO, Chandigarh using BASALT accelerograph which has an internal tri-axial Episensor accelerometer of sensitivity 2.5114 mV g^{-1}. The sampling rate was kept at 250 samples per second per channel with bandwidth ranging from DC to 200 Hz. Second, data set comprises of strong motion seismograph (K-NET) earthquake data recorded in the year 2005 at National Research Institute for Earth Science and Disaster Prevention [13, 14]. The sampling rate was kept at 100 samples per second per channel. A total 25000 events have been used in the study, out of which 13000 earthquake events from K-NET dataset and 12000 noise events from BASALT accelerograph are used.

Methodology

The methodology for the classification of earthquake and noise signal is shown in Fig. 35.1. The analysis has been performed in MATLAB® R2015b platform running on an Intel i5 processor with 4 GB RAM installed. The earthquake and noise signals are taken as input to the feature extractor and classification algorithm. The seismic signals are then passed through a series of preprocessing blocks. The external noise gets coupled with the actual signal in electronic form and reduces the SNR levels. Hence, preprocessing of these signals is requisite to filter out the higher frequency bands and harmonics of line frequencies present in the signals.

Since seismic signals are acquired at different geographically located stations with different sensors with different sensitivity levels, so it is required to convert the voltage to acceleration unit for unified reference. The input seismic signal and converted signal are shown in Fig. 35.2a and Fig. 35.2b, respectively. Denoising methods have been used to restore the smoothness of signal after thresholding the wavelet coefficients of a noisy raw signal [15]. The events are analyzed using

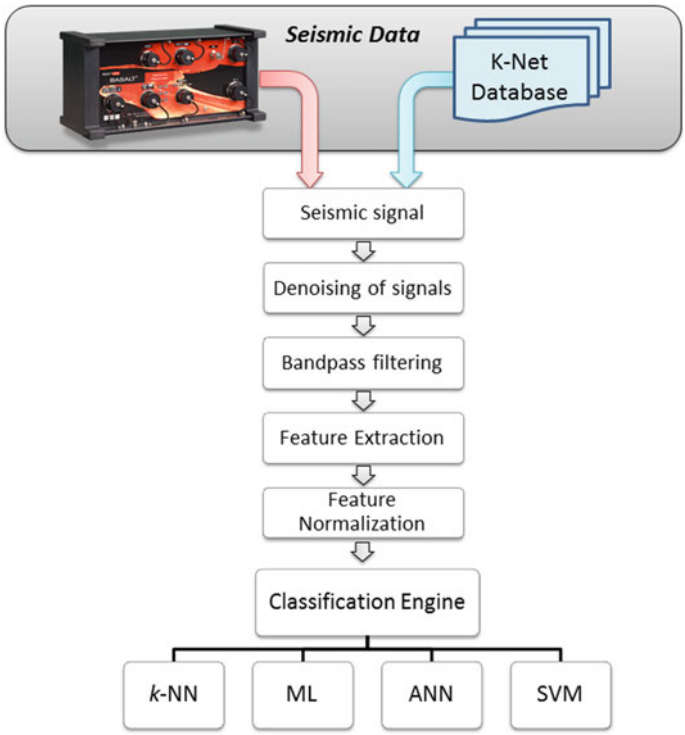


Fig. 35.1 Schema of experimentation methodology

Daubechies 3 wavelet (*db3*) with soft thresholding and the denoised signal is shown in Fig. 35.2c. Due to flat frequency response and more linear response in passband, a Butterworth bandpass filter [16] of fourth order with bandpass frequency 0.1–25 Hz has been applied to filter out higher frequency component present in signal. The fast Fourier transform (FFT) spectrum of filtered signal is shown in Fig. 35.2d.

The feature bag comprises of ten statistical time-domain features and three frequency domain features. Statistical time domain features comprised of mean, mode, median, maxima, minima, variance, standard deviation, root mean square, kurtosis, and Skewness of the time-series signal. Frequency domain features consist of frequency component of dominant energy, amplitude of the corresponding energy, and total energy content. These features are derived after computation of FFT. A total 13-dimensional feature vector are analyzed corresponding to earthquake events and noise events. The extracted features are then normalized which is shown in Figs. 35.3 and 35.4.

To generate the training and testing sets, all normalized features are randomized before they can be used to train the classifier networks; *k*-NN, ML, ANN and SVM. *k*-NN is evaluated for three nearest neighbors ($k = 3$) with distance metric as Euclidean. The distance is calculated between test data and each example of training data. This distance depends upon the number of features used for the

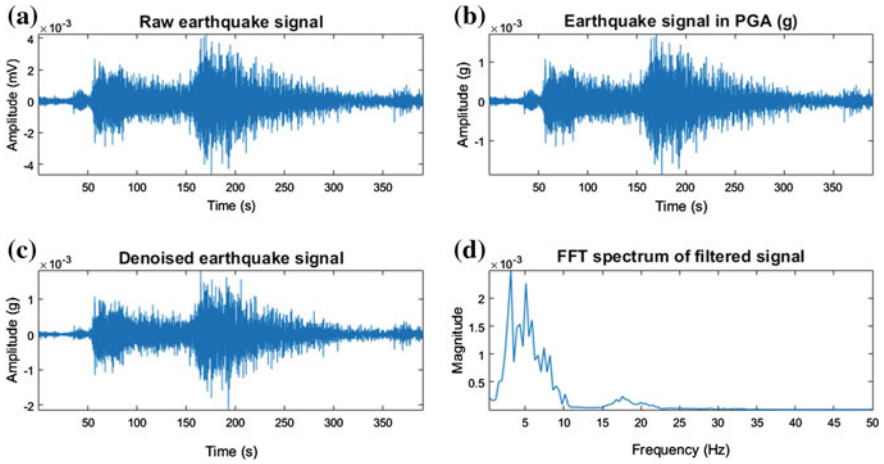


Fig. 35.2 Seismic signal preprocessing and frequency spectrum **a** input signal in voltage **b** converted signal in peak ground acceleration (PGA) **c** denoised signal **d** FFT spectrum containing 0.1–25 Hz frequency

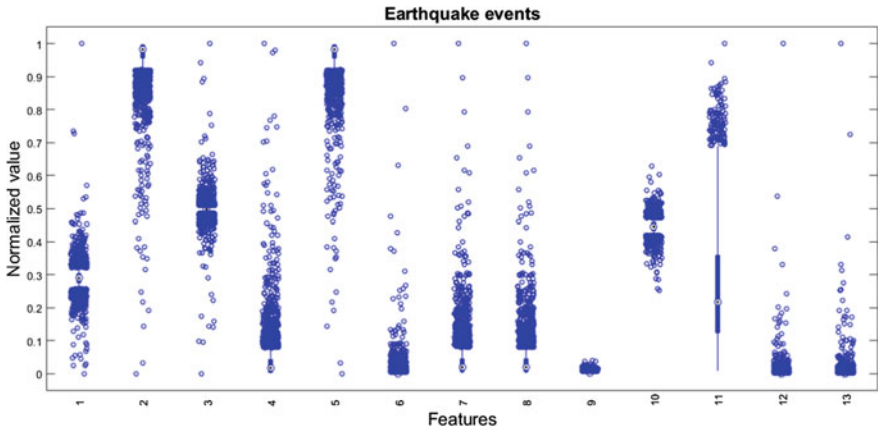


Fig. 35.3 Normalized features of all true earthquake events

classification. These distances along with the supervised classes are sorted in an ascending order, then depending upon the value of k, the k-nearest neighbors are taken out and maximum class allotted to those samples is selected as the class of the test data. ML-based classification is performed by computing the probabilities of the test events and is assigned a class for which the probability is highest.

ANN model is generated using 40 hidden neurons in single hidden layer, with TAN-Sigmoid as activation function and back propagation algorithm is used to train the network. The input feature vector is randomly divided into training set (70%), testing set (15%), and validation set (15%). SVM with radial basis function

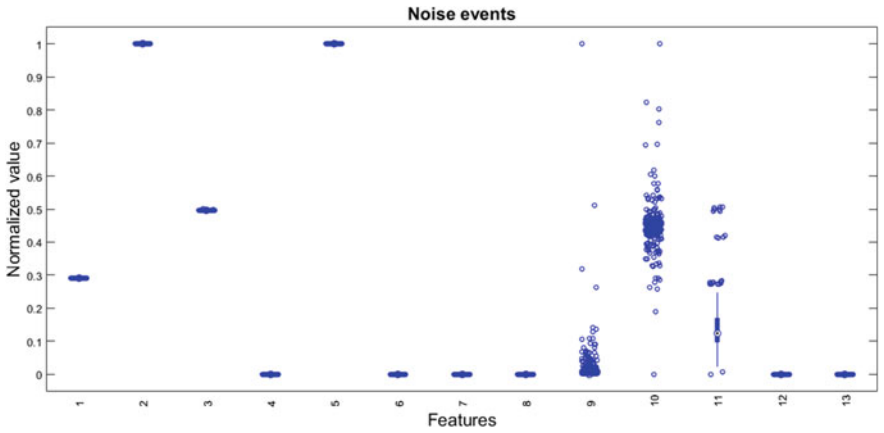


Fig. 35.4 Normalized features of all noise events

(RBF) kernel is used to train and test the network using *svmtrain* and *svmclassify* function, respectively. The accuracy of the individual classifier is calculated using confusion matrix.

Result and Discussion

To validate and evaluate the performance of classifiers, fivefold cross-validation method with 25000 events is used. The accuracy of each fold is computed and the classifiers models with highest accuracy are used for testing with a new set of 2475 events which have been not used during any of the training cases. The performance of the classifiers has been validated by using statistical evaluation metrics, viz., accuracy, sensitivity, and positive predictive value (PPV). In case of *k*-NN and ML, an accuracy of 86.91 and 85.89% was obtained on the new test events. With increasing size of training events in *k*-NN and ML, an increased computation time was observed whereas in model based approaches such as ANN and SVM the computation time remains same. In ANN, the accuracy was 88.36% and its performance was observed with varying number of hidden neurons. It was observed that the performance increased by increasing the number of neurons, but when the number of hidden neurons exceeded 50, the performance of the classifier did not change much. An accuracy of 99.60% was achieved in case of SVM classifier. Confusion matrix on the new test data in Table 35.1 shows correctly classified and misclassified events.

The results show that the PPV of strong motion earthquake data is above 98% whereas in case of noise data, the classification rate is more than 65% for each of the classifiers. The performance metric is computed using the below formulas:

Table 35.1 Confudion matrix of *k*-NN, ML, ANN, and SVM

	k-NN		ML		SVM		ANN	
	Earthquake	Noise	Earthquake	Noise	Earthquake	Noise	Earthquake	Noise
Earthquake	1409	16	1402	23	1420	5	1399	26
Noise	308	742	326	724	5	1045	262	788

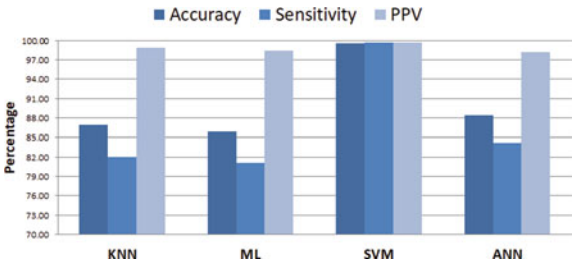


Fig. 35.5 Performance metric showing accuracy, sensitivity, PPV of the classifiers

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$
$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$
$$\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$$

where TP = true positive (Correctly identified), TN = true negative (Correctly rejected), FP = false positive (Incorrectly identified), FN = false negative (Incorrectly rejected). The performance metrics of different classifiers is presented graphically in Fig. 35.5. *k*-NN, ML, and ANN have nearly similar performance, whereas SVM outperforms the other three in terms of all measures.

Conclusion

The realized off line seismic classifiers for the classification of earthquake events from noise utilize robust features obtained from time and frequency domain analysis of the data. The comparative analysis of *k*-NN, ML, SVM, and ANN has been performed and the performance metric shows that SVM performs better than the other analyzed classifiers.

The results definitely institute the conception that the features required for the training of the model for classifier should be robust and distinctive, so that various other techniques of supervised and unsupervised learning can be explored to enhance the performance. Also, work needs to be extended for classification of earthquake on real-time seismic data.

Acknowledgements The work is supported in part by funds of Council of Scientific and Industrial Research (CSIR), India under the project OMEGA PSC0202, Task 2.3.1: *Inferential Imaging—Photo, Acoustics and Seismic*.

References

1. Shearer, P.M.: Introduction to seismology. Cambridge University Press (2009).
2. Shimshoni, Y., Intrator, N.: Classification of seismic signals by integrating ensembles of neural networks. IEEE Transactions on Signal Processing 46, 1194–1201 (1998).
3. Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M., Scarpetta, S.: Discrimination of earthquakes and underwater explosions using neural networks. Bulletin of the Seismological Society of America 93, 215–223 (2003).
4. Scarpetta, S., Giudicepietro, F., Ezin, E., Petrosino, S., Del Pezzo, E., Martini, M., Marinaro, M.: Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. Bulletin of the Seismological Society of America 95, 185–196 (2005).
5. Esposito, A., Giudicepietro, F., Scarpetta, S., D'Auria, L., Marinaro, M., Martini, M.: Automatic discrimination among landslide, explosion-quake, and microtremor seismic signals at Stromboli Volcano using neural networks. Bulletin of the Seismological Society of America 96, 1230–1240 (2006).
6. Agliz, D., Atmani, A.: Seismic signal classification using multi-layer perceptron neural network. International Journal of Computer Applications 79, (2013).
7. Gutiérrez, L., Ibañez, J., Cortés, G., Ramírez, J., Benítez, C., Tenorio, V., Isaac, Á.: Volcano-seismic signal detection and classification processing using hidden Markov models. Application to San Cristóbal volcano, Nicaragua. In: 2009 IEEE International Geoscience and Remote Sensing Symposium, pp. IV-522-IV-525.
8. Changjun, Z., Yuzong, C.: The Research of Vehicle Classification Using SVM and KNN in a ramp. In: Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on, pp. 391–394 (2009).
9. Tian, Y., Qi, H., Wang, X.: Target detection and classification using seismic signal processing in unattended ground sensor systems. In: IEEE International Conference on Acoustics Speech and Signal Processing, pp. 4172–4172 (2002).
10. Mehmood, R.M., Lee, H.J.: Emotion classification of EEG brain signal using SVM and KNN. In: Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on, pp. 1–5 (2015).
11. McConaghy, T., Leung, H., Bosse, E., Varadan, V.: Classification of audio radar signals using radial basis function neural networks. IEEE Transactions on Instrumentation and Measurement 52, 1771–1779 (2003).
12. Javidi, B.: Image recognition and classification: algorithms, systems, and applications. CRC Press (2002).
13. Aoi, S., Kunugi, T., Fujiwara, H.: Strong-motion seismograph network operated by NIED: K-NET and KiK-net. Journal of Japan Association for Earthquake Engineering 4, 65–74 (2004).
14. http://www.kyoshin.bosai.go.jp/.
15. Donoho, D.L.: De-noising by soft-thresholding. IEEE transactions on information theory 41, 613–627 (1995).
16. Douglas, A.: Bandpass filtering to reduce noise on seismograms: Is there a better way? Bulletin of the Seismological Society of America 87, 770–777 (1997).

Chapter 36
Temporal Representation of Vowels in Khasi Dialect

Bronson Syiem, Fairriky Rynjah and Lairenlakpam Joyprakash Singh

Abstract Some speech parameters such as short-time energy, zero crossing rate, and minimum amplitude of the signal can be estimated directly by temporal analysis. This analysis is the easiest method of representing a speech signal since temporal features are easy to extract and have a very easy physical interpretation. An efficient automatic speech recognition system can be achieved by preserving the information needed to determine phonetic identity of the speech signal. Representation of speech in appropriate form yields a more efficient coding system as it can improve the quality of speech synthesis and also increases the performance of speech recognizers. In this paper, phonemic vowels found in Khasi dialect such as /a/, /i/, /I/, /e/, /ɛ/, /ɔ/, /o/, and /u/ are represented purely based on temporal features. Depending on the presence of periodicity information on vowel sounds, fundamental frequency (F_0) of each vowel has been calculated using autocorrelation function.

Keywords Envelope · Periodicity · Fine structure · Formant · Pitch

Introduction

Speech can be defined as a communicating signal between the individuals. Temporal properties of speech sound play a vital role in perception of speech in humans. These properties are related to time or lasting for a relatively short time. Speech is a fast time-varying signal, therefore it is necessary to analyze in short interval of time and so temporal properties can be shown or represented as how they are associated with

B. Syiem (✉) · F. Rynjah · L. J. Singh
Speech and Image Processing Laboratory, Department of ECE,
NEHU, Shillong 793022, India
e-mail: bronzoe12@gmail.com

F. Rynjah
e-mail: phernaky.rynjah@gmail.com

L. J. Singh
e-mail: jplairen@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_36

the speech. Based on temporal aspects, speech signal can be said to be comprised of three main temporal features, namely envelope, periodicity and fine structures [1]. In this paper, vowels in Khasi dialect will be represented based on purely temporal aspects. As discussed by Benedetto [2], vowel sounds can be classified into several dimensions such as height, backs, tenseness, etc., and each formant frequency can be used as acoustic parameters representative of the different dimensions. Vowel representation can be used to obtain accent-adapted features and so it is suitable to identify accents [3]. Generally, Khasi does not have its own script until Thomas Jones initiated the process of writing Khasi language in Roman script. Bareh [4] described some phonemic vowels that can be found in Khasi dialects such as /a/,/i/, /I/, /e/, /ɛ/, /ɔ/, /o/, and /u/. Research had been going on for the past decades about the temporal properties of speech sounds. From Nourski and Brugge [5], periodic nonspeech sounds, like trains of acoustic pulses and bursts of amplitude modulated noise or tone, can produce different percepts with respect to rate of pulse repetition or frequency of modulation. These sounds are useful for the study of timing information (temporal properties) associated with them. Temporal properties play a great role in determining for similarity and classification of the sound. However, most of the current approaches ignore temporal information [6].

Temporal Representation

Speech sounds can be represented in different forms. However, in this paper, vowel sounds of Khasi dialect will be represented temporally using temporal features such as envelope, periodicity, and fine structures. Figure 36.1 shows the waveforms of eight phonemic vowels found in Khasi dialect. Figures 36.2, 36.3, and 36.4 will be used to illustrate envelope, periodicity, and fine structure information contained in the vowels from Fig. 36.1.

Envelope

The envelope of the speech signal can be defined as the boundary within which the signal is contained when viewed in time domain. Among the temporal features, envelope peaks are more perceptually important [7]. From an acoustic point of view, the envelope may be defined by four different parameters such as intensity, duration, rise time, and fall time and these are auditory correlated with loudness, length, attack, and decay [1]. It is shown that how silent gaps can be used to indicate the presence of voiceless plosive or affricative sounds. In this paper, since we are representing only vowel sounds, some cues to identify vowels can be considered. Figure 36.2 shows how the amplitude envelope of one vowel differs from the other. The duration of the vowel depends from vowel to vowel. Thus, this cue can be used to identify vowels.

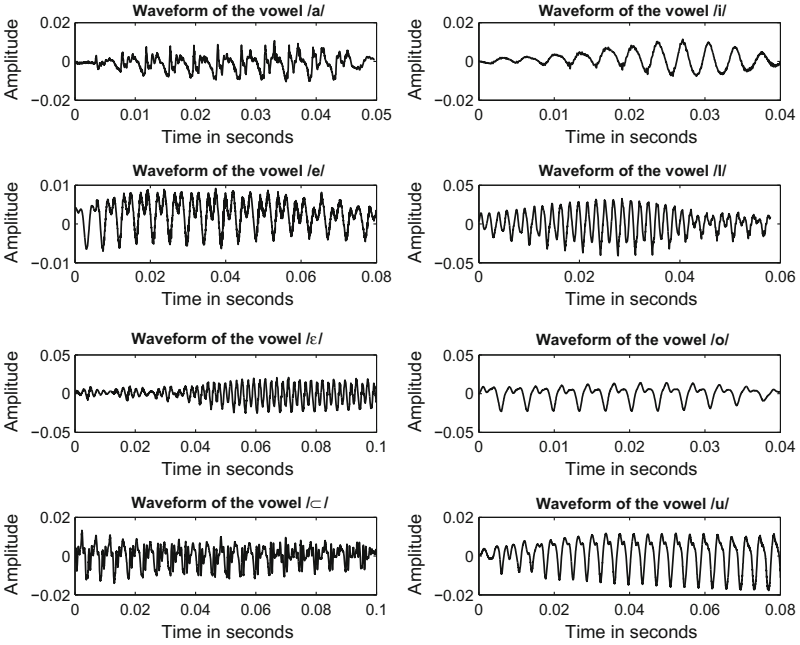


Fig. 36.1 Waveforms of the vowels /a/, /i/, /e/, /I/, /ɛ/, /o/, /ɔ/ and /u/ extracted from some of the Khasi words "ale", "itynnad", "kumne", "u ei", "tyngeh", "shano", "baroh", "u briew"

Periodicity

We defined periodicity as the tendency to repeat at intervals. The speech signal is a complex signal and it contains both periodic and aperiodic information. Periodicity of speech comes from the fact that it consists of quasi-periodic vibration due to vocal chords. By quasi-periodic, we mean repetition of non-periodic stimulation as shown in Fig. 36.3. Generally, periodic sounds are low-frequency signals and are fluctuated at the rates of 50–500 Hz, however aperiodic sounds are high-frequency signals fluctuating at the rates ranging from a few kHz to 5–10 kHz [1]. The presence of these different stimulations makes up periodicity information in speech sound. From the shape of the waveform, it can be seen clearly that each of the eight vowels consists of some low-frequency quasi-periodic vibration. From an acoustic point of view, this cue can be used to identify voice or voiceless speech sound. This property can be considered as one of the most important features in phonology. Another cue of periodicity is that it gives basic information for pitch contouring which is important in lexical function. In this paper, we have also calculated fundamental frequency (F_0) for each vowel, where we have recorded sounds uttered by a female speaker (Table 36.1).

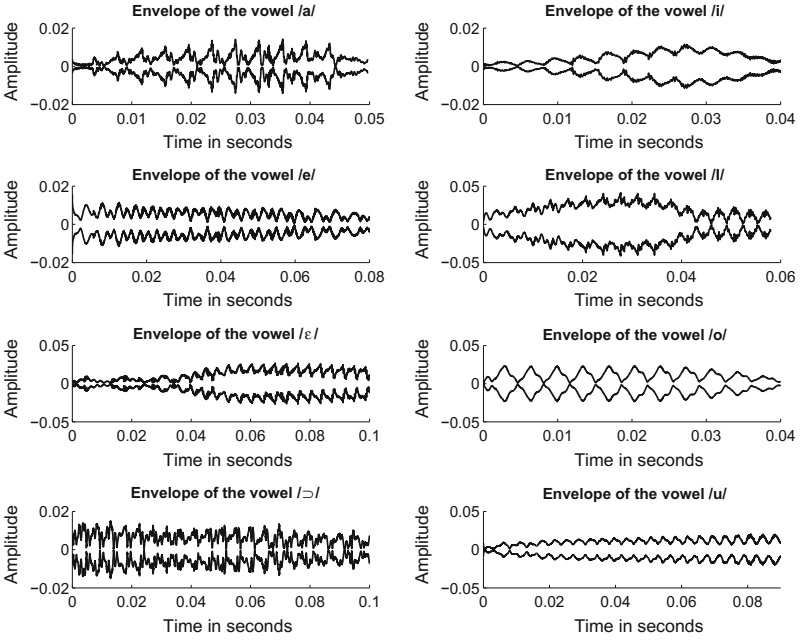


Fig. 36.2 Waveforms obtained by Hilbert transforming the original signals to preserve the envelopes

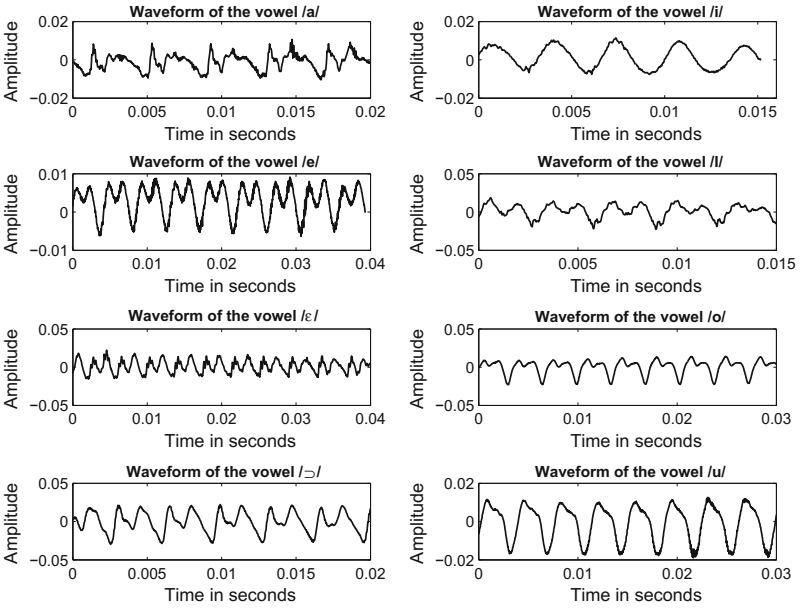


Fig. 36.3 Waveform of the vowels /a/, /i/, /e/, /I/, /ɛ/, /o/, /ɔ/ and /u/, chosen so as to illustrate periodicity information

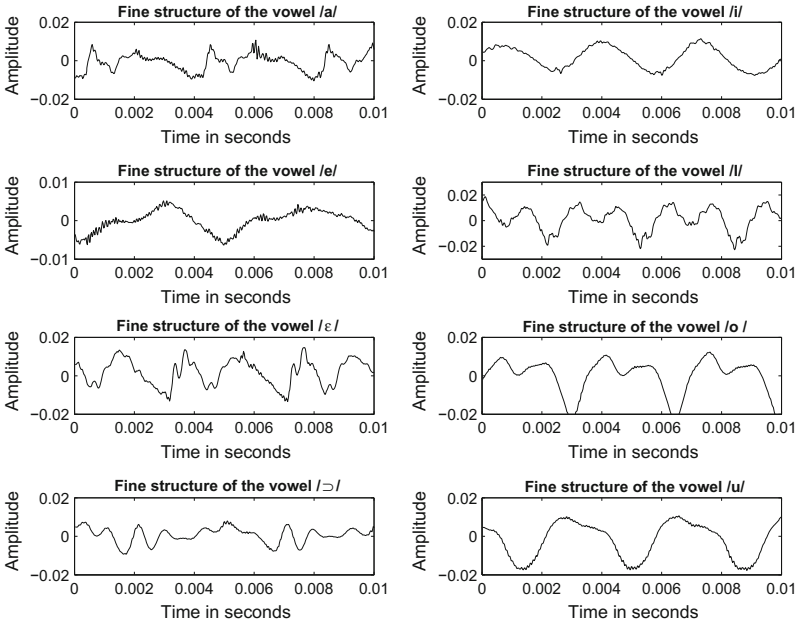


Fig. 36.4 Fine structures information contained in vowels /a/, /i/, /e/, /I/, /ϵ/, /o/, /ɔ/ and /u/

Table 36.1 Fundamental frequency associated with each vowel

Vowel	a	i	e	I	ϵ	o	ɔ	u
F_0 in Hz	252	295.97	214.07	288.23	131.64	284.51	215.12	268.90

Fine Structure

In the production of voice speech sound not only vocal chords acts as barrier for air flow, but in the tract also the air stream used to get resonate. This gives the information about the fine structure in speech sound. By fine structure, we mean the fluctuation of the shape of the sound wave when we analyzed in a very short time interval or within the voice pitch. Being temporal property, fine structure can be used to detect formant frequency as it contains formant pattern. Generally, fine structures are high-frequency stimulation and provide information about the spectrum of the sound and also spectral shape is a main cue for vowel identity [1]. Acoustically fine structures can be considered as the most important temporal features to distinguish vowels in Khasi dialect. Figure 36.4 shows fine structures of eight vowels and how they are distributed in each vowel. Another cue is that it may distinguish between voice and voiceless sound, since voiceless sound has high fluctuation rate as compare to voice sound.

Conclusion

In this paper, vowels found in Khasi dialect have been represented using various temporal features. The representation is purely based on temporal features and no other features have been used. The proposed representation has been carried out to understand how temporal properties are associated with the vowel sounds. We have also discussed how temporal envelope, periodicity, and fine structure can be used as the cues for identifying various vowels found in Khasi dialect and also to distinguish between voice and voiceless sounds.

References

1. S. Rosen: Temporal information in speech: Acoustic, auditory and linguistic aspects. Philosophical transactions: Biological sciences 336(1278), 73–79 (1992)
2. M.D. Benedetto: Vowel representation: Some observations on temporal and spectral properties of the first formant frequency. Acoustical society of America 86(1), 56–66 (1989)
3. Z. Ge, *et al.*: Accent classification with phonetic vowel representation. pp. 529–533 (2015)
4. S. Bareh: Khasi proverbs: Analysing the ethnography of speaking folklore. Ph.D. thesis (2004)
5. K.V. Nourski, J.F. Brugge: Representation of temporal sound features in the human auditory cortex. Rev. Neurosci 22(2), 187–203 (2011)
6. J. Reed, C.H. Lee: On the importance of modeling temporal information in music tag annotation. pp. 1873–1876 (2009)
7. R. Drullman: Temporal envelope and fine-structure cues for speech intelligibility. Acoustical society of America 97(1), 585–592 (1995)

Chapter 37
Hemodynamic Analysis on Human Cerebrovascular Phantoms with and Without Aneurysm

Pranati Rakshit, Nirmal Das, Mita Nasipuri and Subhadip Basu

Abstract Assessment of cerebral aneurysm rupture risk is very crucial for clinicians, because it causes a number of deaths worldwide. Hemodynamics is commonly thought to play a significant role in the mechanisms of development, maturity, and rupture of aneurysm. It is reasonable to assume that rupture risk assessment can be improved by incorporating hemodynamic analysis on the parameters like wall shear stress, velocity, static pressure information, etc. So to compare the hemodynamic parameters on different cerebrovascular phantoms with and without aneurysm, carries a significant role in prediction of occurrence of aneurysm and subsequently rupture risk of the same. Computational fluid dynamics is one conventional approaches to determine the hemodynamic parameter.

Keywords Aneurysm · Hemodynamic analysis · Computational fluid dynamics Cerebrovascular phantoms · Wall shear stress

Introduction

Aneurysms, basically are tiny blood-filled bulges formed in the blood vessel wall and are responsible for nearly 500,000 deaths a year worldwide if they burst before being treated. Several hemodynamic parameters are alleged to be major factors related to the genesis and progression of aneurysm. So it is of enormous importance for treatment to recognize the hemodynamic factors that play a role in the formation and development of disease like aneurysms [1]. The flow pattern will be changed in the vascular structure if there exits an aneurysm and also the hemodynamic parameter values differ from the one without aneurysm. So it is very crucial to analyze the vascular structure with and without aneurysm and hemodynamics there in.

P. Rakshit · N. Das · M. Nasipuri · S. Basu (✉)
Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
e-mail: subhadip@cse.jdvu.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_37

Carotid vasculature is the one of the major parts of the human vasculature. In this work, we focus on the major arterial structure present in the human carotid region. To do the hemodynamic analysis it is important to understand the carotid blood flow patterns and so structural analysis of carotid arteries is needed for determination of abnormal out pouching, detection of possible shrinkage or blockage in the blood flow etc. To depict the network of carotid arteries, we have to know about the structure of the Circle of Willis which is present at the base of the brain. This is a circular vasculature formed by the left and right Anterior cerebral artery (left and right), Internal Carotid Arteries (ICA), Anterior Communicating Artery (ACA), Posterior cerebral artery (left and right), Posterior communicating artery (left and right). The basilar artery and middle cerebral arteries are also considered as part of the circle [2].

In vivo or ex vivo analysis of carotid arterial system is a time consuming job and need to be taken care very minutely by the expert physician. Collecting the medical image is really a troublesome job. Patient permission and participation is equally important to capture and collect the in vivo images. So as much it is important to analyze real medical data, it is also important to construct synthetic structures or phantoms. Therefore, the generation of the phantoms or synthetic structures are often valuable for evaluation and application of new computational techniques. For the construction of 3D digital phantom, one can find many existing methods in the current state of the art [3–5]. It also helps the patients to be escaped from repeated harmful cerebral scans. Phantom based simulation experiments are most popular due to the apparent simplicity of the design.

In general, there are two aspects of the study; (1) using physical vascular phantoms with and without aneurysm, generated by casting a replica of the actual vasculature, and (2) digital modeling of the vasculature, using mathematical models. Here in the present study we have opted the second one.

Cerebral aneurysm is a vascular disease which is a bulge in the vessel wall of an artery to be found in the brain. The reason behind the aneurysm development and progress is due to weakness of the arterial wall. It can rupture to leak blood into the neighboring tissue, if it becomes bulky enough. Generally there are three types of aneurysm—Saccular, Fusiform, and Giant. Among these three, the most common one is Saccular aneurysm. Its shape is almost round and the part which helps it to attach with an artery is called neck of the aneurysm. It is also called berry aneurysm because the shape is like berry. Fusiform aneurysm is formed due to the widening of the vessel wall. Because of the widening of the vessel wall it outlines a spindle-shape. It is generally a less common type of aneurysm. The third type of aneurysm is a giant aneurysm. It is actually a berry aneurysm but the size is large, and it is seen mainly at the bifurcation of an artery.

The prevention of aneurysm is very difficult job as the reason behind the initiation of this vascular disease is not well understood. As hemodynamic information is supposed to be one of the major reasons behind aneurysm initiation, several studies have been performed to analyze the thorough hemodynamic status [6] of the artery. Identifying the appropriate hemodynamic parameter related to the aneurysm progression and initiation is a crucial task. Hemodynamic parameters—the blood

pressure, velocity of blood flow, and the wall shear stress (wss) are linked to the progression of the aneurysm [7, 8]. Wall shear stress is an important hemodynamic factor which is considered to be highly allied with the growth of the cerebral aneurysm. The wall shear stress (wss) acts directly on the endothelium cell of an aneurysm. However, magnitude of high wall shear stress or high spatial and temporal disparity of wall shear stress (wss) may damage the inner wall artery [9, 10].

In this connection, the purpose of the work presented here is to study the CFD-based flow analysis on the cerebrovascular phantoms with and without aneurysm and the comparison of the hemodynamic parameters therein. Digital topology can be used for further study related to this [11–16].

3D Reconstruction of Cerebrovascular Phantom with and Without Aneurysm

Development of appropriate 3D mathematical model is one of the crucial tasks to be performed for hemodynamic analysis of cerebral vasculature. For hemodynamic analysis through the finite element modeling we need synthetic 3D phantom which will resemble human carotid arteries. Here we have designed 3 cerebrovascular phantom with and without aneurysm which are (1) complex bifurcation structure with and without aneurysm (2) ICA (internal carotid artery) with and without aneurysm (3) *circle of willis* with and without aneurysm. It is informative to note that for the generation of 3D digital phantoms and reconstruction of cerebrovascular phantoms from patients' CTA, we have used the method stated in [4, 17] and then get a 3D surface reconstruction for hemodynamic analysis. During the second phase of the 3D reconstruction process 3D surface is converted to a 3D solid mesh. In case of complex structures a series of preprocessing methods are applied to discard spurious edges and vertices. MeshLab_64bit_v1.3.4BETA [18] has been used for generation of the initial surface mesh and for the rendering purpose. Rhinoceros 5.0 [19] has been used to convert the surface mesh to solid mesh. Figure 37.1a–f shows the solid geometry of some cerebrovascular phantom which are now prepared for flow analysis.

Flow Analysis

ANSYS [20] is a well-known computational fluid dynamics software and here it is used for hemodynamic analysis on some cerebrovascular phantoms. For analysis of different hemodynamic parameters the fluent module of ANSYS Workbench 16.0 is used. This CFD-based analysis may be used as a benchmark to compare the digital flow-based model, to be developed as a future direction of the current work.

Velocity and wall shear stress, pressure are the major hemodynamic parameter which are analyzed in the present work.

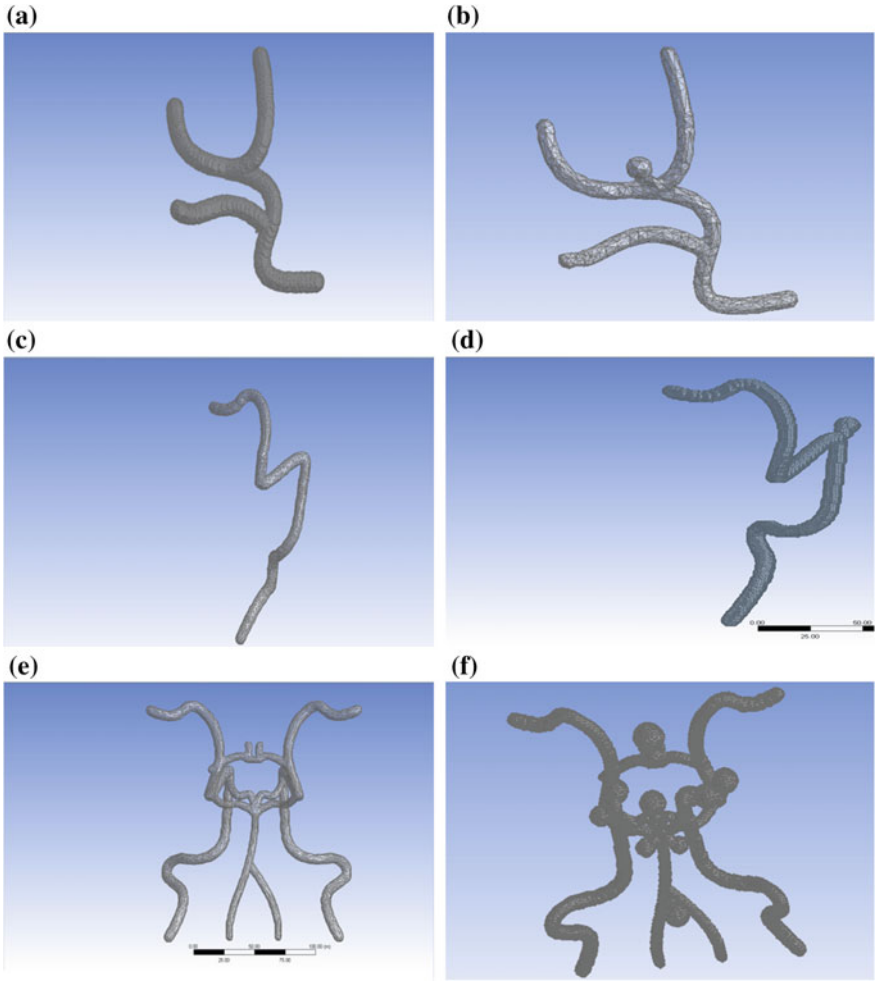


Fig. 37.1 **a** Complex bifurcation structure without aneurysm **b** complex bifurcation structure with aneurysm **c** ICA (internal carotid artery) without aneurysm **d** ICA (internal carotid artery) with aneurysm **e** *circle of willis* without aneurysm **f** *circle of willis* with aneurysm

In Fig. 37.2a, one complex bifurcation phantom geometry without aneurysm is shown and it underwent through flow analysis to get velocity vector (in Fig. 37.2b), wall shear stress(in Fig. 37.2c) and static pressure (in Fig. 37.2d).

In Fig. 37.3, the same complex bifurcation phantom but with aneurysm is considered. Here we have shown the mesh structure of the geometry of Fig. 37.3a in Fig. 37.3b. Because mesh generation is one of the important step before doing the flow analysis. We have shown solid geometry or mesh or both of the vascular phantom in different figures. Velocity and wall shear stress are shown in Fig. 37.3c and Fig. 37.3d respectively.

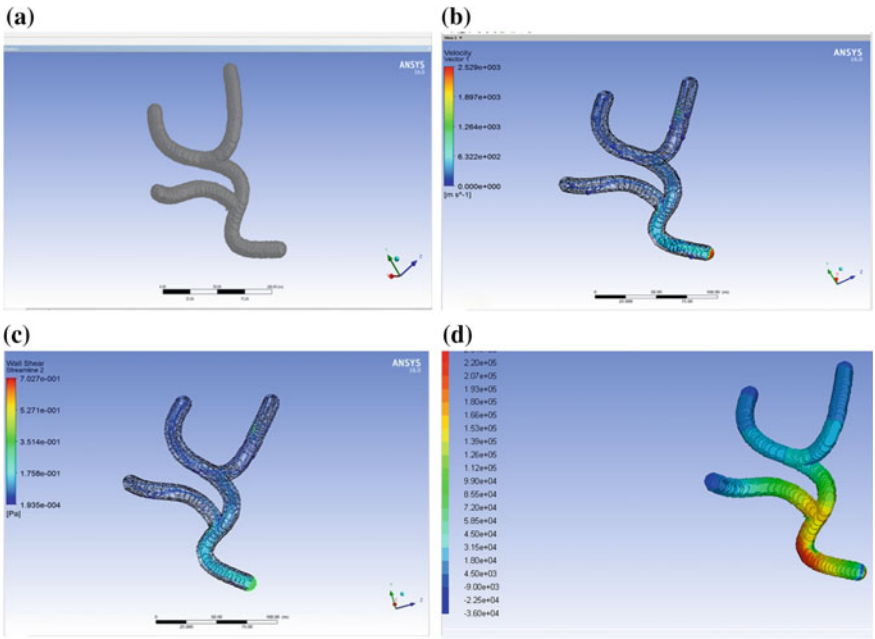


Fig. 37.2 Complex bifurcation without aneurysm **a** geometry **b** velocity vector **c** wall shear stress **d** static pressure

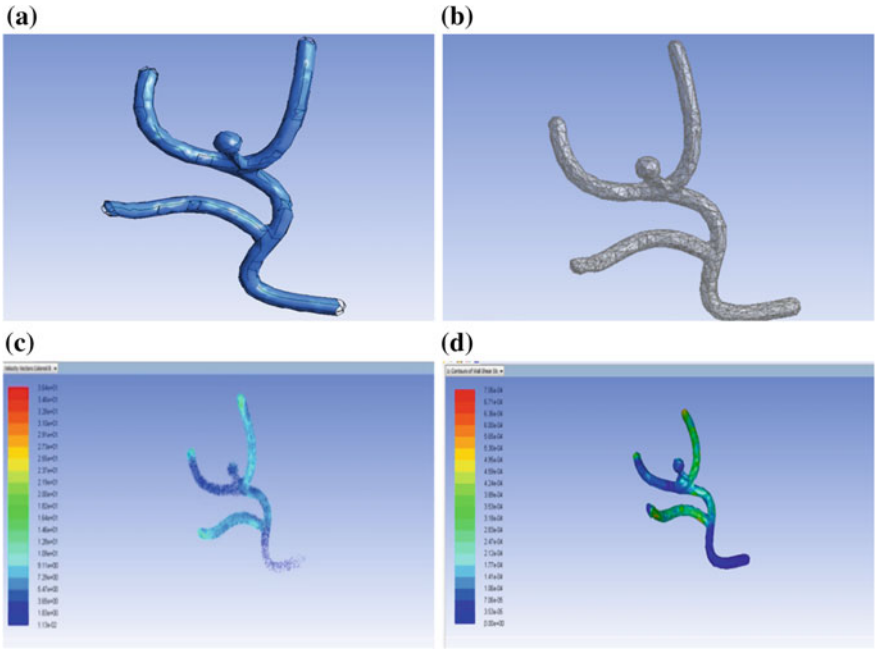


Fig. 37.3 Complex bifurcation with aneurysm **a** geometry **b** mesh **c** velocity vector **d** wall shear stress

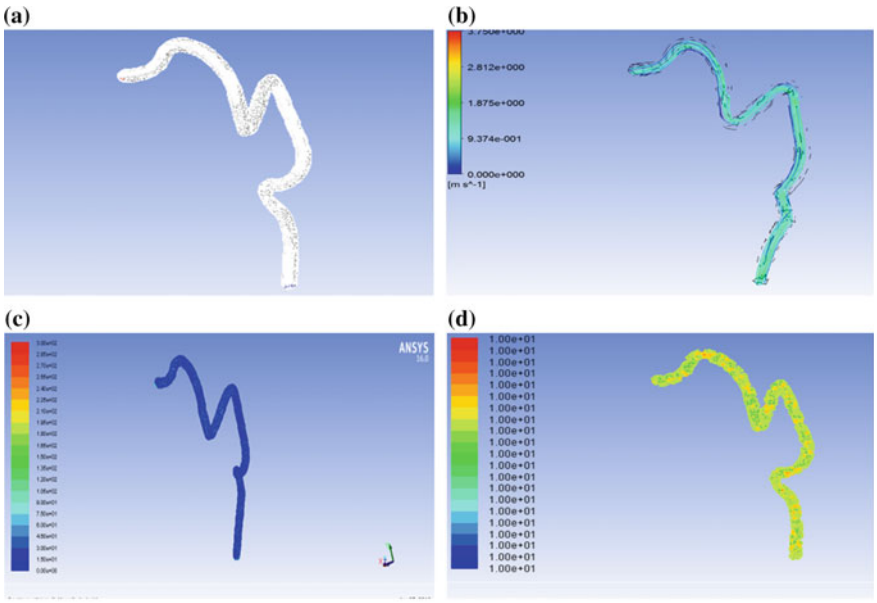


Fig. 37.4 ICA (internal carotid artery) without aneurysm **a** mesh **b** flow streamline **c** velocity **d** static pressure

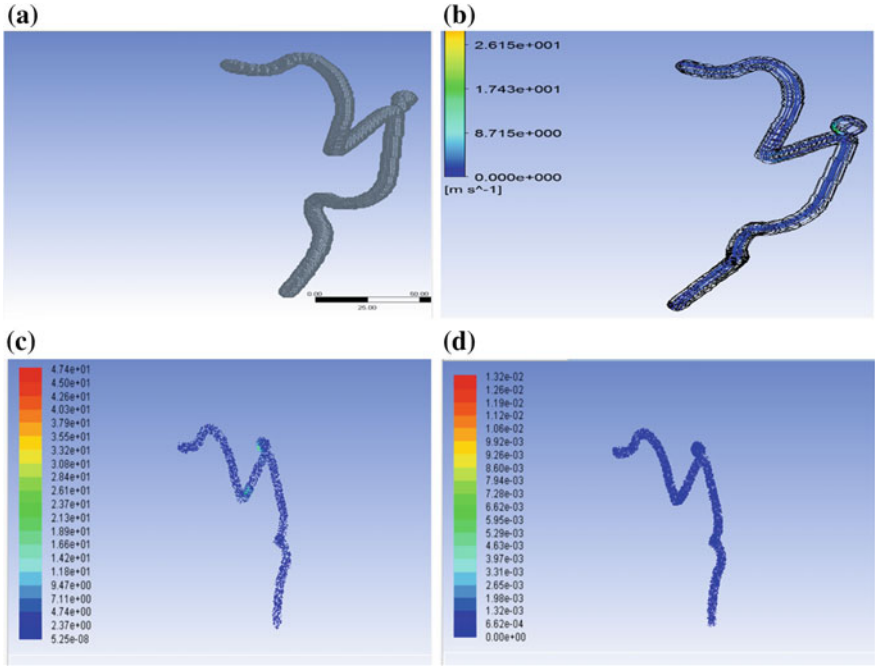


Fig. 37.5 ICA (internal carotid artery) with aneurysm **a** geometry **b** flow streamline **c** velocity **d** wall shear stress

We have analyzed the ICA (Internal carotid artery) without aneurysm in Fig. 37.4. In Fig. 37.4, (a) mesh (b) flow streamline (c) velocity (d) static pressure are developed and shown respectively. In Fig. 37.5 the same phantom geometry (ICA) with aneurysm is analyzed.

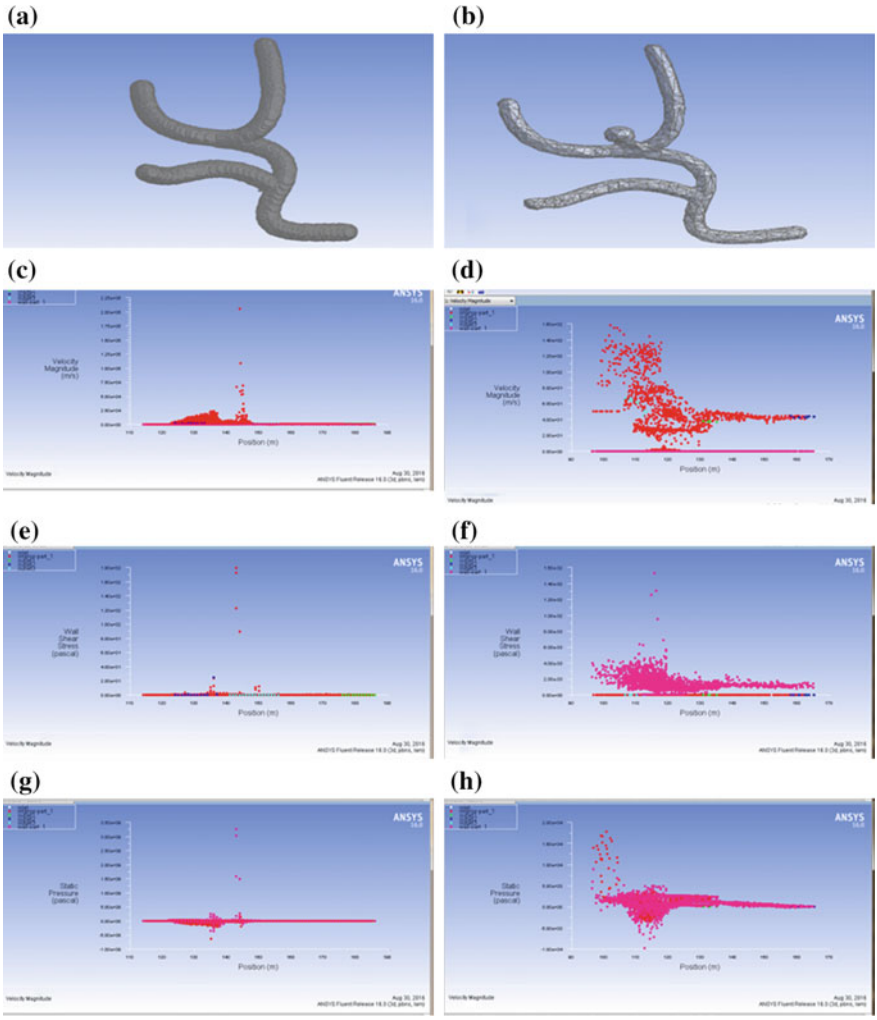


Fig. 37.6 Comparison of hemodynamic parameter through flow analysis on complex bifurcation phantom with and without aneurysm **a** the phantom without aneurysm **b** the phantom with aneurysm **c** velocity of the phantom without aneurysm **d** velocity of the phantom with aneurysm **e** wall shear stress of the phantom without aneurysm **f** wall shear stress of the phantom with aneurysm **g** static pressure of the phantom without aneurysm **h** static pressure of the phantom with aneurysm

In Fig. 37.5 (a) geometry (b) flow streamline (c) velocity (d) wall shear stress are developed and shown respectively.

From the Figs. 37.4 and 37.5, it is very clear that there are several changes in hemodynamic parameter values if the phantom structure includes any aneurysm.

Results and Discussion

We have done the comparison of hemodynamic parameters like velocity, wall shear stress and static pressure on complex bifurcation phantoms with and without aneurysm in the present work. The result is shown in Fig. 37.6a–h. From the Fig. 37.6 we can easily say that there is a turbulence in the flow on the phantom which is with the aneurysm. We can say this statement because we can observe that there is a variation in the ranges of values for the hemodynamic parameters. For the numerical simulation of hemodynamic parameter in finite element method we have used ANSYS 16.0.

Conclusion

In this present study, we have emphasized on the comparison of cerebrovascular phantoms with and without aneurysm and hemodynamics therein. In the flow analysis of the phantom with aneurysm, we have found that there is a turbulent nature in the flow which differs from the flow in the phantom without aneurysm. These results are useful for understanding of fluid flows through cerebral vasculature with and without aneurysm and serve as benchmarks for 3D digital flow models. Computational studies on digital flows on phantom with and without aneurysm are important future research directions for the current work. 2D digital flows have already been used successfully for the study of structural/plastic changes in hippocampal dendritic spines [21]. The work presented may be included in the future study on the 3D digital flow-based hemodynamic analysis in both the patients' CTA images with and without aneurysm as well as on complex mathematical phantoms.

Acknowledgements This project is supported by the DST PURSE-II, Government of India, project of Computer Science and Engineering Department of Jadavpur University.

References

1. C. F. Gonzalez, Y. I. Cho, H. V. Ortega, and J. Moret, "Intracranial aneurysms: Flow analysis of their origin and progression," Am. J. Neuroradiol., vol. 13, no. 1, pp. 181–188, 1992.
2. W. J. Vankan, J. M. Huyghe, J. D. Janssen, a. Huson, W. J. G. Hacking, and W. Schreiner, "Finite element analysis of blood flow through biological tissue," Int. J. Eng. Sci., vol. 35, no. 4, pp. 375–385, 1997.
3. A. Banerjee, S. Dey, S. Parui, M. Nasipuri, and S. Basu, "Design of 3-D phantoms for human carotid vasculature," Proc. - 2013 3rd Int. Conf. Adv. Comput. Commun. ICACC 2013, pp. 347–350, 2013.
4. A. Banerjee, S. Dey, S. Parui, M. Nasipuri, and S. Basu, "Synthetic reconstruction of human carotid vasculature using a 2-D/3-D interface," in Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, 2013, pp. 60–65.
5. Guha I., Das N., Rakshit P., Nasipuri M., Saha P.K., Basu S. (2018) Design of Cerebrovascular Phantoms Using Fuzzy Distance Transform-based Geodesic Paths. In: Sa P., Sahoo M., Murugappan M., Wu Y., Majhi B. (eds) Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Advances in Intelligent Systems and Computing, vol 518. Springer, Singapore.
6. C. K. Zarins, D. P. Giddens, B. K. Bharadvaj, V. S. Sottiurai, R. F. Mabon, S. Gladov, and S. Glagov, "Carotid Bifurcation Atherosclerosis: Quantitative Correlation of Plaque Localization with Flow Velocity Profiles and Wall Shear Stress," Circ. Res., vol. 53, no. 4, pp. 502–514, 1983.
7. M. F. Fillinger, S. P. Marra, M. L. Raghavan, and F. E. Kennedy, "Prediction of rupture risk in abdominal aortic aneurysm during observation: Wall stress versus diameter," J. Vasc. Surg., vol. 37, no. 4, pp. 724–732, 2003.
8. M. L. Raghavan and D. A. Vorp, "Toward a biomechanical tool to evaluate rupture potential of abdominal aortic aneurysm: identification of a finite strain constitutive model and evaluation of its applicability," J. Biomech. no.4, pp. 475–482, 2000, vol. 33, 2000.
9. M. L. Raghavan and D. A. Vorp, "Toward a biomechanical tool to evaluate rupture potential of abdominal aortic aneurysm: identification of a finite strain constitutive model and evaluation of its applicability," J. Biomech. no. 4, pp. 475–482, 2000, vol. 33, 2000.
10. D. A. Vorp, M. L. Raghavan, and M. W. Webster, "Mechanical wall stress in abdominal aortic aneurysm: Influence of diameter and asymmetry," J. Vasc. Surg., vol. 27, no. 4, pp. 632–639, 1998.
11. P. K. Saha, F. W. Wehrli, and B. R. Gomberg, "Fuzzy Distance Transform: Theory, Algorithms, and Applications," Comput. Vis. Image Underst., vol. 86, no. 3, pp. 171–190, 2002.
12. P. K. Saha, J. K. Udupa, and D. Odhner, "Scale-Based Fuzzy Connected Image Segmentation: Theory, Algorithms, and Validation," Comput. Vis. Image Underst., vol. 77, no. 2, pp. 145–174, 2000.
13. S. Basu, E. Hoffman, and P. K. Saha, "Multi-scale Opening–A New Morphological Operator," in Image Analysis and Processing—ICIAP 2015, Springer, 2015, pp. 417–427.
14. P. K. Saha, R. Strand, and G. Borgefors, "Digital Topology and Geometry in Medical Imaging: A Survey," IEEE Trans. Med. Imaging, vol. 34, no. 9, pp. 1940–1964, 2015.
15. S. Basu, M. L. Raghavan, E. A. Hoffman, and P. K. Saha, "Multi-scale opening of conjoined structures with shared intensities: methods and applications," in Intelligent Computation and Bio-Medical Instrumentation (ICBMI), 2011 International Conference on, 2011, pp. 128–131.
16. P. Saha, S. Basu, and E. Hoffman, "Multi-Scale Opening of Conjoined Fuzzy Objects: Theory and Applications," IEEE Trans. Fuzzy Syst., 2015.[In press].
17. A. Banerjee, S. Dey, S. Parui, M. Nasipuri, and S. Basu, "Synthetic reconstruction of human carotid vasculature using a 2-D/3-D interface," in Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, 2013, pp. 60–65.

18. “MeshLab.” [Online]. Available: http://www.3dcoform.eu/index.php/tools/meshlab [Accessed: 12-Aug-2016].
19. “Rhino - Downloads.” [Online]. Available: https://www.rhino3d.com/download. [Accessed: 12-Aug-2016].
20. “ANSYS - Simulation Driven Product Development.” [Online]. Available: http://www.ansys.com/.
21. S. Basu, D. Plewczynski, S. Saha, M. Roszkowska, M. Magnowska, E. Baczynska, and J. Wlodarczyk, “2dSpAn: semiautomated 2-d segmentation, classification and analysis of hippocampal dendritic spine plasticity,” Bioinformatics, p. btw172, Apr. 2016.

Chapter 38
Problems and Issues in Parsing Manipuri Text

Yumnam Nirmal and Utpal Sharma

Abstract This paper discusses the issues and challenges that are faced during parsing Manipuri text. Parsing normally follows two approaches: rule-based and statistical. For a resource-poor language like Manipuri, statistical approaches have not been possible so far due to the non-availability of annotated corpora. Different types of issues like lexical ambiguity, attachment ambiguity arising due to the constituents of noun phrases, also add to the difficulty of parsing Manipuri text. The large number of possible variations of word order is another major problem.

Keywords Parsing · Manipuri · Free word order · Tibeto-Burman

Introduction

Today, it is very important for a language to be used through a computer. The number of speakers being small, Manipuri has been a less studied language for natural language processing (NLP) as compared to other Indian languages like Hindi, Bengali, etc. The lack of research has also contributed to the non-availability of language processing tools for the language, which would have otherwise benefited its speakers and others. Today, language processing tools have moved to statistical methods which are highly accurate and reliable. For Manipuri, due to the lack of annotated data in any form, statistical methods have not been so far applicable. On the other hand, manual annotation takes a huge amount of time and manpower.

Y. Nirmal (✉) · U. Sharma
Tezpur Univerity, Tezpur 784028, Assam, India
e-mail: ynirmal@tezu.ernet.in

U. Sharma
e-mail: utpal@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_38

Manipuri Language

Manipuri (Meitei-lon), the official language of the northeastern Indian state Manipur, is one among the 22 scheduled languages of the Indian Constitution. It is also spoken in neighbouring states Assam and Tripura, and also in the neighbouring countries Bangladesh and Myanmar. It is the lingua-franca (bridge language, trade language) of 33 different tribal communities of Manipur, whose mother tongues are different. Among the different native languages spoken in Manipur, it alone has its own script known as the Meitei Mayek. It has been currently classified as a vulnerable language by UNESCO [17].

Manipuri belongs to the family of Tibeto-Burman languages [16]. Its tonal and highly agglutinative nature are distinctive features with respect to Indian languages. A verbal root may take as many as ten suffixes [18] as in Example 1.

Example 1 "ꯄꯨꯁꯤꯟꯍꯟꯖꯥꯡꯒꯥꯗꯥꯕꯤꯗꯣ"
("pusinhənɟərəmgədəbənidəko")
(I wish I) myself would have caused to bring in (the articles)

has "ꯄꯨ" ('pu') as root, and "ꯁꯤꯟ" ('sin'), "ꯍꯟ" ("hən"), "ꯖ" ("ɟə"), "ꯡꯒꯥ" ("rəm"), "ꯒ" ("gə"), "ꯗ" ("də"), "ꯕ" ("bə"), "ꯇꯤ" ('ni'), "ꯗ" ("də"), "ꯣ" ('ko') as the suffixes.

Parsing

In computational linguistics, parsing refers to the formal analysis of a sentence or a string of words into its constituents, leading to a parse tree indicating the syntactic relation between its constituents. The resultant parse tree may conjointly contain semantic as well as other useful linguistic information.

One of the most important and common use of parse trees is grammar checking in word processing systems. If a sentence cannot be parsed, then it may have grammatical errors or the sentence is hard to read [15]. Parsing is an immediate vital stage for semantic analysis, and also plays an important role in applications such as machine translation [6], speech recognition [7], question answering [22], text summarisation [28], information extraction [13], etc.

Natural language parsing has a wide range of approaches, which employ different search strategies (top-down or bottom-up), or use different grammar formalisms, or implementation models. Among the most successful grammar formalisms are Context-Free Grammars (CFG), Tree Adjoining Grammar (TAG) [14], Lexical Functional Grammar (LFG) [4], Combinatory Categorial Grammar (CCG) [23], Head-Driven Phrase Structure Grammar (HSPG) [19], Dependency Grammar (DG) [24], etc. On the other hand, implementation models may be non-probabilistic, probabilistic (generative, discriminative, etc.) or hybrid.

Related Works

There is no significant work reported on parsing of Manipuri. Few works regarding parsing of Myanmar language which also belongs to Tibeto-Burman family can be found. Manipuri shares many features with Tibeto-Burman languages such as tone, widespread stem homophony, agglutinative verb morphology, verb derivational suffixes, duplication or elaboration, evidentiality and emotional attitudes signalled through sentence final particles, aspect rather than tense marking, lack of gender marking, verb final order and lack of grammatical relations such as 'subject' and 'object' [9].

Win Win Thant et al. [25, 27] proposed a Context-Free Grammar based top-down parser for Myanmar sentences. The parser uses function tags[1] proposed by Win Win Thant et al. [26] for producing the grammar rules and a total of 183 rules for grammatical relations of the phrases have been used. The work focuses on parsing of simple and complex sentences. However, there has been no report of handling ambiguity of any type and simple top-down parsers are not much efficient in handling ambiguity.

Bharti et al. [2] proposed a constraint-based parsing method for free word order languages using the Paninian framework that has been successfully applied to Indian languages. It uses the notion of *karaka* relations between verbs and nouns in a sentence. The *karaka* relations are syntactico-semantic (or semantico-syntactic) relations between the verbals and other related constituents in a sentence. Manipuri is a role-dominated language where there is no distinction between grammatical relations such as subject and object. But, in case of reference-dominated language, distinctions between grammatical relations such as subject and object are significant for syntactic operations and structure [11]. Due to this reason and other differences (not discussed in this paper), application of Paninian framework to Manipuri language structure is still questionable. Attempts by Manipuri Sanskrit scholars in the 1960s and 1970s to apply Sanskrit grammatical principles in the analysis of Manipuri led to an imperfect analysis of the language [5].

Another work on parsing of an Indian language Assamese can be found in the work of Navanath Saharia et al. [20] wherein they describe about the word order freedom. The work is preliminary and does not discuss about ambiguity and other problems.

Dependency parsing methods for free word order languages were also proposed by Bharti et al. [1] and Falavarjani et al. [10]. These systems employ two well-known dependency parsers, namely MST parser and Malt parser and uses dependency Treebank for their training. But, due to the non-availability of annotated corpora, these methods cannot be applied to Manipuri yet.

[1]Function tags are a context-sensitive annotations (e.g. syntactic categories like subject, object, time, location, etc.) applied to natural language text, marking the syntactic or semantic role of words and phrases within the text.

Issues in Parsing Manipuri Text

Most of the existing techniques and approaches cannot be applied to Manipuri due to its language structure and lack of crucial resources. Manipuri is tonal, highly agglutinative, morphologically very rich and has a high freedom of word order.

Corpus

Suitable raw and annotated corpora for Manipuri are not available. A small amount of un-annotated corpus may be obtained from TDIL (Technology Development of Indian Language) Program, DeitY, MC & IT, Govt. of India, under Research license, which is very limited in terms of its domain. A balanced corpus should consist of texts from various domains. Annotated data of Manipuri in any form is not available till date. This is one of the major setbacks in applying statistical methods for Manipuri language processing. And natural language parsing, be it rule-based or statistical model, normally requires resources like a Lexicon or a Treebank, that are well structured and rich with linguistic annotations.

Ambiguity

Ambiguity has always been a major problem and hindrance in parsing of natural languages. A sentence is said to be ambiguous if it has more than one parses. A few causes of ambiguity encountered in Manipuri are as follows:

Lexical Ambiguity: Manipuri being a highly agglutinative and tonal language there are plenty of words whose POS (parts of speech) category and meaning are ambiguous, and are determined in accordance to tone as well as the context.

Example 2 ꯂꯩ (lei) can have two POS categories, viz. noun and verb, and can have at least six different meanings: ‘flower’, ‘tongue’, ‘buy’, ‘have’, ‘stay’, ‘rotate’.

Coordination Ambiguity: Coordination ambiguity arises when different sets of phrases are conjoined by a conjunction.

Example 3 ꯑꯄꯪꯕ ꯅꯨꯄꯥ ꯑꯃꯁꯨꯡ ꯅꯨꯄꯤ ꯑꯗꯨ ꯆꯠꯈ꯭ꯔꯦ ॥
(əpəŋbə nupa əməsuŋ nupi ədu cət-k^{h}re)
foolish - man - and - woman - that - gone

In Example 3, the phrase ꯑꯄꯪꯕ ꯅꯨꯄꯥ ꯑꯃꯁꯨꯡ ꯅꯨꯄꯤ (əpəŋbə nupa əməsuŋ nupi) can be bracketed as [ꯑꯄꯪꯕ [ꯅꯨꯄꯥ ꯑꯃꯁꯨꯡ ꯅꯨꯄꯤ]] ([əpəŋbə [nupa əməsuŋ nupi]]), or as [ꯑꯄꯪꯕ ꯅꯨꯄꯥ] ꯑꯃꯁꯨꯡ [ꯅꯨꯄꯤ] ([əpəŋbə nupa] əməsuŋ [nupi]).

Attachment Ambiguity: If a particular constituent of a sentence can be attached at more than one place in the parse tree of the sentence, then the sentence is said to have an attachment ambiguity. It is one of the main reasons for a sentence to have a large number of parses. According to Church and Patil [8], for English language, the prepositional phrase (PP) attachment ambiguity leads to an exponential growth in the number of analyses for a sentence.

Manipuri is a post-positional language [5] opposed to prepositioned; all the case markers are suffixed to the noun and to the adverb of place. Unlike English where prepositional phrase (PP) attachment ambiguity exists, post-positional phrase (PsP) attachment ambiguity does not exist in Manipuri. Rather, attachment ambiguity exists within noun phrases where the constituents can be attached at different places.

Example 4 ꯑꯃꯟꯕ ꯈꯣꯡꯎꯞ ꯌꯣꯟꯕ ꯗꯨꯀꯥꯟ ꯑꯗꯨ ꯀꯥꯢꯔꯦ ॥
(əmənbə kʰoŋŋup yon-bə dukan ədu kaire)
old - shoe - sell - shop - that - broken

The entity ꯑꯃꯟꯕ (əmənbə) in Example 4 may either refer to ꯈꯣꯡꯎꯞ (kʰoŋŋup) or ꯗꯨꯀꯥꯟ (dukan) (Fig. 38.1).

Two different parse trees for the sentence in Example 4 where the entity ꯑꯃꯟꯕ (əmənbə) being attached at different places have been illustrated in Fig. 38.2. Figure 38.2a depicts the meaning of the sentence as ‘a shop that sells old shoes’, whereas, Fig. 38.2b depicts the meaning as ‘an old shop that sells shoes’.

Word Order

Word order freedom (position of words) is also one of the problems faced during parsing. Grammar formalisms like CFG are positional and tend to capture only the position of words in a phrase or sentence. In order to accommodate the freedom of word order in a phrase or sentence, the number of grammar rules has to be increased. As a result, the grammar may become too bulky with a huge number of rules and its usefulness may be reduced.

The vast majority of Tibeto-Burman languages including Manipuri, and all Indian languages except Kashmiri and Khasi have the basic word order of subject–object–verb (SOV) [5]. However, the interchange of subject and object is quite an acceptable common phenomenon in Manipuri, which agrees with Greenberg’s Universal 7 [12], according to which the only alternative of a dominant SOV order is OSV.

Example 5

ꯇꯣꯝꯕꯅ - ꯕꯣꯜꯗꯨ - ꯀꯧꯏ ॥ (S - O - V)	or,	ꯕꯣꯜꯗꯨ - ꯇꯣꯝꯕꯅ - ꯀꯧꯏ ॥ (O - S - V)
tomb-nə bol-du kəu-i		bol-du tomb-nə kəu-i
by Tomba - that ball - kick		that ball - by Tomba - kick

Example 5 illustrates the interchange of subject and object. The first form is that of subject–object–verb, whereas the later is of the form object–subject–verb.

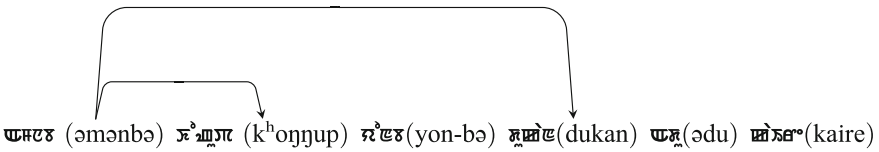


Fig. 38.1 Ambiguous reference for entity ꯑꯃꯅꯕ (əmənbə) in Example 4

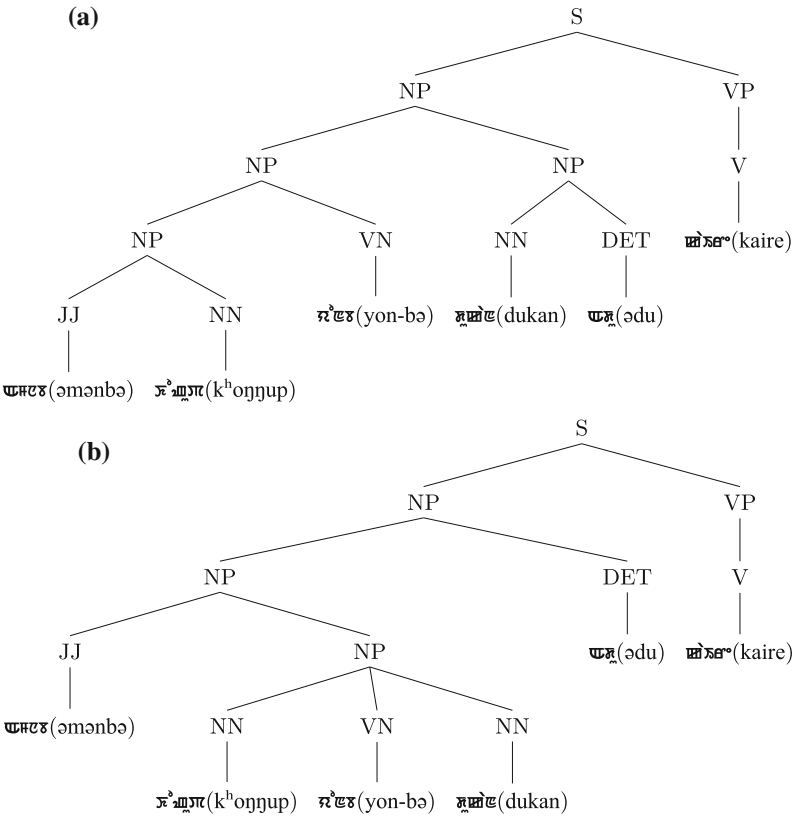


Fig. 38.2 Parse trees for the sentence in Example 4. **a** 'a shop that sells old shoes' and **b** 'an old shop that sells shoes'

Bhat [3] also observed that Manipuri is a verb final language and the relative order of arguments to the left of the verb is quite free. The arguments (to the left of verb) can be arranged in any of the several possible ways, if all the arguments are properly marked by the relevant case suffixes; such changes would not affect the semantic roles of the arguments.

Table 38.1 Word order variations for Example 6

Manipuri sentence	Word order
ꯃꯍꯥꯛꯅ ꯉꯁꯤ ꯑꯅꯥꯕꯁꯤꯡ ꯌꯦꯡꯂꯔꯣꯏ ॥	subject–adverbial of time–object–verb
ꯃꯍꯥꯛꯅ ꯑꯅꯥꯕꯁꯤꯡ ꯉꯁꯤ ꯌꯦꯡꯂꯔꯣꯏ ॥	subject–object–adverbial of time–verb
ꯉꯁꯤ ꯃꯍꯥꯛꯅ ꯑꯅꯥꯕꯁꯤꯡ ꯌꯦꯡꯂꯔꯣꯏ ॥	adverbial of time–subject–object–verb
ꯉꯁꯤ ꯑꯅꯥꯕꯁꯤꯡ ꯃꯍꯥꯛꯅ ꯌꯦꯡꯂꯔꯣꯏ ॥	adverbial of time–object–subject–verb
ꯑꯅꯥꯕꯁꯤꯡ ꯃꯍꯥꯛꯅ ꯉꯁꯤ ꯌꯦꯡꯂꯔꯣꯏ ॥	object–subject–adverbial of time–verb
ꯑꯅꯥꯕꯁꯤꯡ ꯉꯁꯤ ꯃꯍꯥꯛꯅ ꯌꯦꯡꯂꯔꯣꯏ ॥	object–adverbial of time–subject–verb

Example 6 ꯃꯍꯥꯛꯅ - ꯉꯁꯤ - ꯑꯅꯥꯕꯁꯤꯡ - ꯌꯦꯡꯂꯔꯣꯏ ॥
(məhak-nə ŋəsi ənabə-siŋ yeŋ-lə-roy)
by him - today - ill people - will not treat

The sentence in Example 6 can have as many as six forms by rearranging the position of the arguments of the verb into 3! ways. It has been illustrated in Table 38.1. The arguments ꯃꯍꯥꯛꯅ (məhak-nə), ꯉꯁꯤ (ŋəsi) and ꯑꯅꯥꯕꯁꯤꯡ (ənabə-siŋ) on the left of the verb ꯌꯦꯡꯂꯔꯣꯏ (yeŋ-lə-roy) can be arranged in any order without affecting the semantics of the sentence.

In Manipuri, the freedom of order is not confined to the arguments of verb. The descriptive adjective can also change its order by preceding or following the noun [21]. Examples 7 and 8 present such cases where the position of descriptive adjective is changed.

Example 7 ꯑꯆꯧꯕ(JJ) ꯇꯦꯕꯜ(NN)
(əcəubə tebəl)
big - table

or,

ꯇꯦꯕꯜ(NN) ꯑꯆꯧꯕ(JJ)
(tebəl əcəubə)
table - big

Example 8 ꯑꯃꯨꯕ(JJ) ꯍꯨꯏ(NN)
(əmubə hui)
black - dog

or,

ꯍꯨꯏ(NN) ꯑꯃꯨꯕ(JJ)
(hui əmubə)
dog - black

Conclusion

Parsing natural language texts requires certain resources, which are not available for Manipuri. Effective statistical parsing requires a Treebank and it is neither available for Manipuri nor can be easily created. So, rule-based methods using grammar formalisms like CFG, TAG, etc., look better suited for Manipuri. For such methods, a suitable grammar must be defined, and resources such as a Lexicon and POS annotated corpus must be developed. Moreover, issues like ambiguity and word order freedom shall have to be addressed.

References

1. Bharati, A., Husain, S., Misra, D., Sangal, R.: Two stage constraint based hybrid approach to free word order language dependency parsing. In: Proceedings of the 11th International Conference on Parsing Technologies. pp. 77–80. Association for Computational Linguistics (2009)
2. Bharati, A., Sangal, R.: Parsing free word order languages in the paninian framework. In: Proceedings of the 31st annual meeting on Association for Computational Linguistics. pp. 105–111. Association for Computational Linguistics (1993)
3. Bhat, D.S.: Grammatical relations: the evidence against their necessity and universality. Routledge (2002)
4. Bresnan, J., Asudeh, A., Toivonen, I., Wechsler, S.: Lexical-functional syntax, vol. 16. John Wiley & Sons (2015)
5. Ch Yashawanta, S.: Manipuri Grammar. Rajesh Publications, New Delhi (2000)
6. Charniak, E., Knight, K., Yamada, K.: Syntax-based language models for statistical machine translation. In: Proceedings of MT Summit IX. pp. 40–46. Citeseer (2003)
7. Chelba, C., Jelinek, F.: Exploiting syntactic structure for language modeling. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. pp. 225–231. Association for Computational Linguistics (1998)
8. Church, K., Patil, R.: Coping with syntactic ambiguity or how to put the block in the box on the table. Computational Linguistics 8(3-4), 139–149 (1982)
9. DeLancey, S.: Sino-tibetan languages. The world's major languages pp. 797–810 (1987)
10. Falavarjani, S.A.M., Ghassem-Sani, G.: Advantages of dependency parsing for free word order natural languages. In: International Conference on Current Trends in Theory and Practice of Informatics. pp. 511–518. Springer Berlin Heidelberg (2015)
11. Foley, W.A., Van Valin Jr, R.D.: Functional syntax and universal grammar. Cambridge Studies in Linguistics London (38) (1984)
12. Greenberg, J.H.: Some universals of grammar with particular reference to the order of meaningful elements. Universals of language 2, 73–113 (1963)
13. Grishman, R.: The role of syntax in information extraction. In: Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996. pp. 139–142. Association for Computational Linguistics (1996)
14. Joshi, A.K., Schabes, Y.: Tree-adjoining grammars. In: Handbook of formal languages, pp. 69–123. Springer (1997)
15. Jurafsky, D.: Speech & language processing. Pearson Education India (2000)
16. Matisoff, J.A., Baron, S.P., Lowe, J.B.: Languages and dialects of tibeto-burman (1996)
17. Moseley, C.: Atlas of the World's Languages in Danger. UNESCO, Paris (2010)
18. Nonigopal Singh, N.: A Meitei Grammar of Roots and Affixes, A Thesis. Ph.D. thesis (1987)

19. Pollard, C., Sag, I.A.: Head-driven phrase structure grammar. University of Chicago Press (1994)
20. Saharia, N., Sharma, U., Kalita, J.: A first step towards parsing of assamese text. Special Volume: Problems of Parsing in Indian Languages (2011)
21. Sharma, H.S.: A comparison between khasi and manipuri word order. Linguistics of the Tibeto-Burman Area 22(1), 139–48 (1999)
22. Shen, D., Lapata, M.: Using semantic roles to improve question answering. In: EMNLP-CoNLL. pp. 12–21. Association for Computational Linguistics (2007)
23. Steedman, M., Baldridge, J.: Combinatory categorial grammar. Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell (2011)
24. Tesnière, L.: Eléments de syntaxe structurale. Librairie C. Klincksieck (1959)
25. Thant, W.W., Htwe, T.M., Thein, N.L.: Context free grammar based top-down parsing of myanmar sentences. In: International conference on computer science and information technology, Pattaya. pp. 71–75 (2011)
26. Thant, W.W., Htwe, T.M., Thein, N.L.: Function tagging for myanmar language. International Journal of Computer Applications 26(2) (2011)
27. Thant, W.W., Htwe, T.M., Thein, N.L.: Parsing of myanmar sentences with function tagging. arXiv preprint arXiv:1205.1603 (2012)
28. Turner, J., Charniak, E.: Supervised and unsupervised learning for sentence compression. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 290–297. Association for Computational Linguistics (2005)

Chapter 39
Classification of Bangla Compound Characters Using a HOG-CNN Hybrid Model

S. M. A. Sharif, Nabeel Mohammed, Sifat Momen and Nafees Mansoor

Abstract Automatic handwriting recognition is challenging task due to its sheer variety of acceptable stylistic differences. This is especially true for scripts with large character sets. Bangla, the sixth most widely spoken language in the world has a complex, large and rich set of compound characters. In this study, a hybrid deep learning model is proposed which combines the use of the manually designed feature Histogram of Oriented Gradients (HOG), with the adaptively learned features of a Convolutional Neural Networks (CNN). The proposed hybrid model was trained on the CMATERDB 3.1.3.3, a Bangla compound character data set which divides Bangla compound characters into 177 broad classes and 199 specific classes. The results demonstrate that CNN-only models achieve over 91% and 92% test accuracy respectively. Furthermore, it is shown that the proposed model, which incorporates HOG features with a CNN, achieves over 92.50% test accuracy on each division. While there is still room for improvement, these results are significantly better than currently published state of art on this data set.

Keywords Bangla handwriting recognition · Bangla compound characters Convolutional Neural Networks (CNN) · Deep learning · Histogram of Oriented Gradients (HOG) · CMATERDB 3.1.3.3

S. M. A. Sharif (✉) · N. Mohammed · S. Momen · N. Mansoor
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh
e-mail: sma.sharif.cse@ulab.edu.bd

N. Mohammed
e-mail: nabeel.mohammed@ulab.edu.bd

S. Momen
e-mail: sifat.momen@ulab.edu.bd

N. Mansoor
e-mail: nafees.mansoor@ulab.edu.bd

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_39

Introduction

Bangla language has a rich set of compound characters which are formed by combining multiple individual characters. While they do not appear as frequently as individual vowels and consonants, these compound characters have an important role in day to day verbal communication as well as in writing scripts [5]. An exhaustive study [4] identified over 334 compound characters with 171 pattern classes. Some of these pattern classes even contain multiple shapes. Such large number of pattern classes, some with subtle shape changes, make Bangla compound character classification a challenging task.

Handwriting recognition is a part of the image classification domain. Two dominant approaches to image classification can be observed from recent studies. The first approach uses predefined and/or hand-crafted features, e.g. SIFT, HOG etc., extracted either locally or globally (or both) which are then used to train a classifier e.g. SVM, K-Means, GA-based etc. The other approach is popularly known as the deep learning approach, where convolutional neural networks (CNN) are trained with images and their corresponding labels. These networks are remarkable because they learn useful features during the training phase.

Both approaches have been employed successfully to the task of handwritten character classification. This paper reports on the application of a Hybrid model, which combines CNNs and the popular Histogram of Oriented Gradients (HOG) image feature, to the task of classifying Bangla compound characters. It is shown that the proposed model achieves over 92.5% test accuracy in classifying Bangla compound characters, which appreciably better than the best reported results so far.

The rest of the paper is organised as follows: Section "Background" discusses the CMATERDB 3.1.3.3 data set, CNN, and HOG. Section "Proposed Method" describes the proposed hybrid model. Section "Experimental Setup" details the experiments performed for this study. Section "Results" presents the results and section "Conclusion" concludes the paper.

Background

More than 230 million people speak Bangla all over the world. It is the sixth most widely spoken language and is the second most widely spoken language in the Indian subcontinent. The Bangla writing script has a set of vowels, consonants as well as compound characters. Compound characters are composed of two or more consonants and at times also a vowel. The vocalisation of these characters is typically done by the simultaneous pronunciation of the individual characters. The shapes of these compound characters are necessarily complex and in some cases, the individual original characters are not recognisable in the final shape. Some examples compound characters are shown in Fig. 39.1.

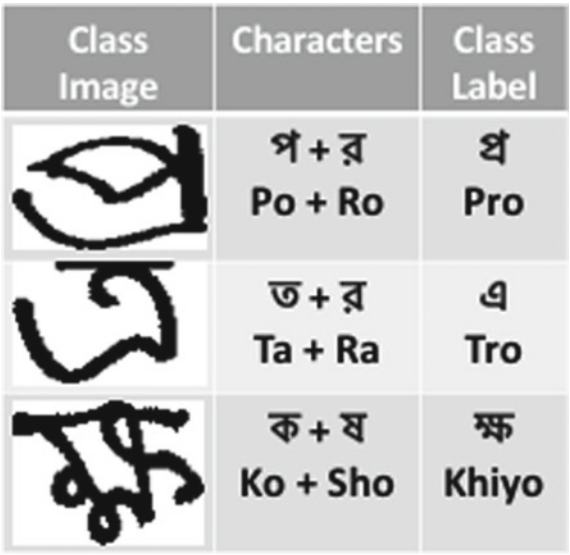


Fig. 39.1 Examples of Bangla compound characters from CMATERDB 3.1.3.3 model structure

The work in [4] documents the CMATERDB 3.1.3.3 handwritten Bangla compound character data set. The study performed a thorough automated analysis of 2.4 million words collected from three Bangla newspapers to identify 171 different compound character classes. These 171 classes do no include individual consonants with vowel allographs. Some compound characters have multiple visually distinct writing patterns. If these individual patterns are separated, then the total number of compound character classes rise to 199. Currently, the data set contains 44,152 training images and 11,126 testing images with annotations for 171 and 199 classes. The number of samples per character class is not equal and vary between 125 and 474 samples. Das et al. [3] applied a GA-and SVM-based approach to classifying this data set and achieved an impressive test accuracy of 78.93%. Das et al. [4] used a Quad-tree-based approach and reported an improved test accuracy of 79.35%. In both cases, the features extracted from the images were predetermined and not learnt during the training phase, as convolutional neural networks are designed to do.

A breakthrough study [6] showed that convolutional neural networks (CNN) [7] can be successfully applied to image classifications tasks with a very large number of classes. These networks have multiple layers of filters, and pooling/subsampling layers. The filters are typically two or three dimensional, depending on whether the image is gray scale or colour, and are used to extract image features by convolving them with the image. Unlike traditional systems, these filters are not predefined by the researcher/engineer but are learnt by minute adjustments during the training phase. Nonlinearities (i.e. tanh, RelU, sigmoid), are usually applied to the output of the convolutional layers to allow these networks to learn complex nonlinear functions. The output of the convolutional layers are typically two/three dimensional,

which are then flattened and passed through one or more fully connected layers for image classification.

There has been some recent CNN-based work on Bangla handwriting recognition . For instance, [10] employed a five-layer network similar to the popular LeNet architecture, which has four convolutional and pooling layers. The last layer is labelled as the F5 layer which extracts a feature vector of 300 dimensions. The work aimed to achieve a generalised character recognition feature extractor by training this network on a data set of 50 character classes. The features extracted by the trained model was then used to classify characters from other data sets by classifying the features using an SVM. This method yielded an accuracy rate of 98.375% on the test partition of the ISI Bangla numeral data set. Even more recently, a smaller CNN was used by Akhand and Mahtab Ahmad [9] to train on and classify the ISI Bangla numeral data set. This study reported accuracy rates of 98.98% on the test partition of the data set. The training set of 19, 392 images was augmented by rotating the images by fixed angles. Different networks were trained for different augmentation angles ($5°, 10°, 20°$ and $30°$). The best testing accuracy rate of 98.98% was reported when a $10°$ rotation angle was used. However, this method is equivalent to fine-tuning the training process to learn features which are applicable on the validation set, instead of generalised learning.

Most of these studies rely entirely on "learnt features" extracted by CNNs, and do not take into account the role of manually designed features. This paper presents the use of a hybrid approach to designing a network, which aims to bridge this gap. The proposed approach combines, in a simple way, the "learnt features" of a CNN with the very effective, manually designed, Histogram of Oriented Gradients (HOG) feature.

Histogram of Oriented Gradients (HOG) [2], is a feature which emanates from a different philosophy. Before the recent popularity of deep learning, features such as HOG, SIFT [8], etc.... were considered state of the art. These features were designed through the hard work and ingenuity of the researchers and were typically used in conjunction with linear classifiers, e.g. SVM. HOG features are particularly interesting because of their simplicity and effectiveness at distinguishing shapes [1]. It utilises image gradient information extracted from images after subdividing the image into small regions which are referred to as "cells". The gradients at sampled pixels within a cell are used to compute a histogram. The combined histogram of all the cells becomes the image descriptor. For gradient calculation, applying the simple masks: $[-1, 0, 1]$ and $[-1, 0, 1]^T$ on a non-smoothed image works best. The image is subdivided into cells of configurable size and an orientation histogram of the gradient values is calculated for every cell. For contrast normalisation, particularly, when cells sizes are small, multiple cell histograms can be normalised together.

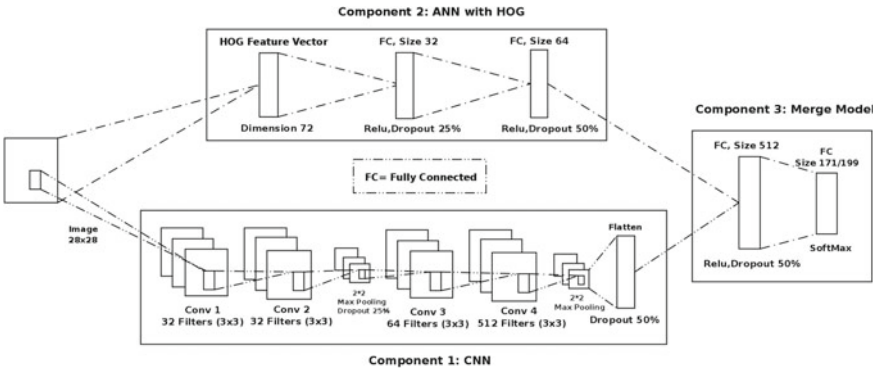


Fig. 39.2 Proposed hybrid model

Proposed Method

The aim is to integrate HOG features in the learning and classification process of a deep learning model. Figure 39.2 shows the scheme of the proposed model which has three components and two input layers.

The first component is comprised of a single CNN, which accepts a 28×28 grayscale image as input and successively applies convolution and pooling operations. The network has two convolution layers, each with 32 3×3 filters. The two convolutional layers are followed by a max-pooling layer, which is in turn followed by two more convolutional layer with 64 and 512 3×3 filters, respectively. The output of this last convolutional layer is a stack of two-dimensional filter outputs. These values of the filter outputs are unrolled, to form a vector input to further fully connected layers.

A fully connected layer which expects a HOG image descriptor as input forms the second part of the model. HOG features are extracted from each image with a cell size of 8 pixels, with a 8 bin histogram of edges created at each cell. The histograms are then concatenated to form the final HOG image descriptor. This results in a feature vector of 72 dimensions, which is what this part of the proposed structure expects as input. This component has two hidden layers of 32 and 64 dimensions, respectively. The output of the last layer as a feature vector of 32 dimensions and is concatenated with the output of the first component.

The concatenated feature vector is then used as input to a third fully connected network which has a hidden layer of 512 dimensions. The output layer has 171 or 199 dimensions depending on which classification scheme is used in a particular experiment.

All three components are created as a part of a single network and trained end-to-end together. As shown in Fig. 39.2, the Dropout [12] rates used are quite high; this was done to ensure that the network does not over train as it has a large number of parameters.

Experimental Setup

Data Set Preparation

This study uses the aforementioned CMATERDB 3.1.3.3 Bangla compound character collection. The training set of the collection is augmented [11] by randomly rotating between $-50°$ and $50°$. The original images have black writing on a white background; these are processed so that the background is black and the character is in white. All images are resized to be 28×28. The aspect ratio of the written characters is not preserved.

Models Used

In total, four different models were used to facilitate comparison using either a CNN-only approach or the proposed hybrid approach.

The two CNN-only models are labelled CNN-171 and CNN-199, respectively, to indicate whether they are classifying 171 classes or 199 classes. The networks have a similar structure to the first component of the hybrid model, except the output of the flattened layer is directly used in a fully connected network whose structure is similar to the third component.

The Hybrid models are labelled Hybrid-Hog-171 and Hybrid-HOG-199, respectively, to indicate whether they are classifying 171 classes or 199 classes. The only place where they differ is in the dimensionality of the last softmax layer (of the third component).

Model Training

Each model was trained on the augmented training set for 100 epochs. At the end of each epoch, the models performance on the test data set was measured and the best performing model, over the 100 epochs, was kept. The training batch size was set to 100. The Adadelta [13] optimiser was used to minimise the categorical cross entropy loss values. All training was done on an a machine running Ubuntu 14.04 while taking advantage of an NVIDIA GTX 980 TI GPU to speed up the experiments.

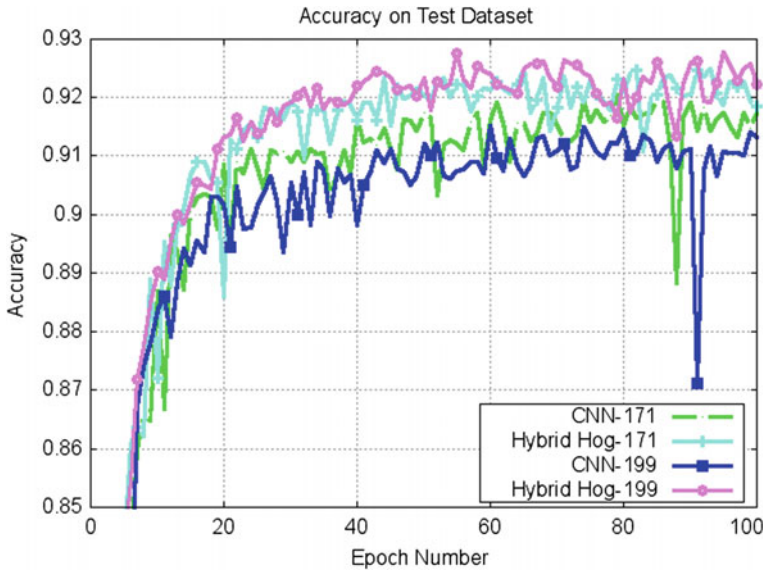


Fig. 39.3 Accuracy of the four models on the test data set during training

Table 39.1 Best validation accuracy and training epochs required

Experiment name	Accuracy (%)
CNN-171	92.05
CNN-199	91.53
Hybrid-Hog-171	*92.57*
Hybrid-Hog-199	*92.77*

Results

Figure 39.3 shows how the accuracy of the test set varied during training. Although only the best performing models were saved, it is evident from the comparison that the hybrid models consistently outperform the CNN-only models.

Table 39.1 shows the best performing test accuracy rates of the four models used in the experiments, and the epoch at which the rates were achieved. The hybrid models achieve improvements of 0.52 and 1.24% on the 171 and 199 classifications when compared to the CNN-only models.

Table 39.2 compares the performance of the proposed method with those which have previously been published. The proposed approach achieves a significant improvement over previous work, with improvements of 13.22% and 14.1% over the 171 and 199 classes respectively.

The proposed models actually achieve a higher accuracy on the 199 classes compared to the 171 classes. While this might seem counter-intuitive initially; the

Table 39.2 Comparison with published results

Work	Year	Feature selection	Classification	Number of classes	Test accuracy (%)
Nibaran Das et al. [3]	2012	GA	SVM	171	78.93
Nibaran Das et al. [4]	2014	CH,QTLR	SVM	171	79.35
Nibaran Das et al. [4]	2014	CH,QTLR	SVM	199	78.67
Hybrid-Hog-171 (Proposed)	2016	Hog-Adaptive	CNN-ANN	171	*92.57*
Hybrid-Hog-199 (Proposed)	2016	Hog-Adaptive	CNN-ANN	199	*92.77*

behaviour is easily explained when taking into account the nature of the class separations. The 171 classification problem includes cases, where the different patterns for the same shape is assigned to the same class. For the 199 classification problem, this added difficulty is not present.

Conclusion

This paper presents a Hybrid deep learning model which combines a convolutional neural network with HOG features. The proposed model was trained end-to-end using an augmented version of the training image set provided by the CMATERDB 3.1.3.3 collection. After training, the model achieves test accuracy rates of 92.57% and 92.77% on the 171 and 199 classes of the data set, respectively. These results are an improvement of over 13% and 14%, respectively, in accuracy rates compared to previous studies, which is significant. In future, multiple features are planned to be incorporated in such Hybrid models to study their effect.

Acknowledgements This work was supported by the ICT division of Ministry of ICT, Bangladesh [Grant number 56.00.0000.028.33.066.16-731].

References

1. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC. vol. 2, p. 8 (2011)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893. IEEE (2005)
3. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A novel ga-svm based multistage approach for recognition of handwritten bangla compound characters. In: Proceedings of the International Conference on Information Systems Design and Intelligent

Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. pp. 145–152. Springer (2012)
4. Das, N., Acharya, K., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: A benchmark image database of isolated bangla handwritten compound characters. International Journal on Document Analysis and Recognition (IJDAR) 17(4), 413–431 (2014)
5. Das, N., Basu, S., Sarkar, R., Kundu, M., Nasipuri, M., Basu, D.: Handwritten bangla compound character recognition: Potential challenges and probable solution. In: IICAI. pp. 1901–1913 (2009)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
7. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361(10), 1995 (1995)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
9. M. A. H. Akhand, Mahtab Ahmad, M.M.H.R.: Convolutional neural network training with artificial pattern for bangla handwritten numeral recognition. ICIEB 1(1), 1–6 (2016)
10. Maitra, D.S., Bhattacharya, U., Parui, S.K.: Cnn based common approach to handwritten character recognition of multiple scripts. In: Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. pp. 1021–1025. IEEE (2015)
11. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 806–813 (2014)
12. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1), 1929–1958 (2014)
13. Zeiler, M.D.: Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701 (2012)

Chapter 40
GMM-Based Formant Transformation of the Vowels/Diphthongs of One Assamese Variety to Another

Nath Sanghamitra and Sharma Utpal

Abstract Formants, considered to be responsible for differences in vowel quality, also represent regional variations in the vowel/diphthong sounds of a language. In this paper, three approaches based on the Gaussian Mixture Models (GMMs) are used to develop mapping functions to map the most informative formants, F1, F2, & F3 of vowels/diphthongs of one variety of Assamese to another. The first is based on a single GMM for vowel/diphthong formants in training data. The second, maps the formants at four equidistant temporal points of vowel/diphthong duration. The third approach trains separate GMMs for the formants of each vowel/diphthong. In objective evaluation, all three approaches bring the vowel/diphthong formants of the source variety closer to the target variety. The third, outperforms the previous two.

Keywords Dialect · Formants · Vowel space · GMMs · RMSE · Assamese Nalbaria

Introduction

The Assamese language is an Indo-European language spoken in the Indian state of Assam and the adjoining regions. The language has a number of regional variations such as the Eastern variety, the Central variety, the Kamrupi variety, and the Goalpariya variety. Our study focuses on transformation of the vowels/diphthongs in All India Radio (AIR) (source) variant to those in Nalbaria (target) variant. The AIR variant is spoken by the readers of Assamese news of AIR. It belongs to the Central variety and is also referred to as the 'likʰitɔ bʰaxa' (or written form). The Nalbaria variant is spoken by the people in and around the district of Nalbari in Assam.

N. Sanghamitra (✉) · S. Utpal
Department of Computer Science and Engineering,
Tezpur University, Tezpur, India
e-mail: s.nath@tezu.ernet.in

S. Utpal
e-mail: utpal@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_40

It falls under the Kamrupi group. The current study is limited to the six vowels /ax/ (/axmrit/- nectar, IPA: ɔ), /aa/(/aam/- mango, IPA: a), /i/ (/itaa/- brick, IPA: i), /u/ (/uraa /- fly, IPA: u) /e/ (/etaa/- one, IPA: ɛ), /o/ (/mor/- mine, IPA: ʊ) and diphthongs such as /eaa/,/uaa/, /iaa/, /eu/, /iu/, /oi/, /ou/, etc. of the Assamese phoneme inventory.

Regional variations in vowels/diphthongs have been identified and documented for a number of languages in many works [1, 2]. Nath and Sharma [3] report that the position of vowels in the F1-F2 plane is significantly different for the two varieties of Assamese under study. Furthermore, dynamic formant plots of most Nalbaria diphthongs indicate incomplete movement from the first vowel to the second vowel making it difficult to perceive the diphthong.

If the vowel/diphthong formants of the source variety are transformed to match that of the target variety, it is very likely that the vowel/diphthong sounds of the source variety would sound more like their counterparts in the target variety. This study aims to transform formants of the AIR variety to that of Nalbaria variety using three different approaches based on the GMMs.[1]

The rest of the paper is organized as follows. Section "Literature Review and Related Works" presents a brief review of related works, section "Building the Speech Corpus" describes building of the speech corpus, section "Methodology" describes the GMM-based methods developed for formant transformation. In Section "Experimental Results" experimental results and a comparison of the methods are presented. Finally, conclusions and plans for future work are presented in section "Conclusions and Future Work".

Literature Review and Related Works

A number of studies relating formant frequencies to dialect variation have been reported in the literature. Fox and Jacewicz [1] report the nature of dynamic spectral change in the vowels of three distinct varieties of American English. Hagiwara [2] in his work reports considerable variation in the vowel spaces of contemporaneous regional variants of American English. Labov [4] reported that the correlation between social factors and vowels, almost entirely concentrated in F2, and F1 frequency is mainly used for cognitive differentiation of vowels. However, Grama's [2] findings run counter to Labov's claims and report that social meaning can lie in both F1 and F2. His findings are supported further by Teutenberg and Watson [5] who also infer that F1, F2 formants have a significant contribution to vowel quality and points on the F1-F2 plane often represent the pronunciation of a speaker. Therefore, they attempt to modify the source accent to match the target accent by mapping the vowel space of the source to the target. Although vowel quality alone is not sufficient for accent modification, they state that even a simple

[1] https://www.ll.mit.edu/mission/cybersec/publications/publication-files/full_papers/0802_Reynolds_Biometrics-GMM.pdf.

[2] http://www.ling.hawaii.edu/research/WorkingPapers/wp-Grama.pdf.

transformation can yield a significant shift in the perceived accent. Since a dialect is almost always associated with an accent, the transformation of vowel formant space is very likely to improve the quality of synthesis of target dialect from source dialect.

Formants are commonly used to represent features of the vocal tract system and therefore formant transformation is used to transform the vocal tract system from source to target speaker. In one method for formant transformation [6], artificial neural networks are used to capture the nonlinear relation between source and target formants. In another work, the statistical distributions of formants are modeled by Rentzos et al. [7] with a two-dimensional Hidden Markov model. However, almost all the works on formant transformation have been applied to Voice Conversion (VC), which aims to change the source speaker's voice to sound like that of the target speaker. Our study capitalizes on these observations and attempts to transform vowel/diphthong formants of one dialect to another using three approaches based on the GMMs.

Building the Speech Corpus

For our study, speech data containing vowels/diphthongs in both varieties are required. A set of 60 text prompts, P, of short sentences in Nalbaria, is prepared containing at least 10–20 occurrences of each vowel/diphthong. These are recorded from speaker S, fluent in both varieties of Assamese, at a sampling rate of 48 kHz with 16-bit resolution, in a sound proof room with a Zoom H4Next recorder. This forms our target set C_T out of which 10 sentences are kept aside for testing. Utterances in C_T are manually transcripted by a person and cross checked by another, both well versed in phonetic transcription. This is P_{TR}.

The source set of utterances, C_S, is generated by a HMM based TTS system S_{HTS} with P_{TR} as input. S_{HTS} has been trained with speech data from S in the AIR variety with pronunciation and syllabification rules pertaining to the AIR variety. Hence C_S is of AIR variety. C_S and C_T are down sampled to 16 kHz and cleaned to remove unwanted pauses and noise. The main reason for generating C_S from a TTS and not directly recording from the speaker is because of a future goal of incorporating dialectal features into synthesized speech.

C_S and C_T are annotated with vowel/diphthong labels in Praat.[3] A Praat script extracts F1, F2, F3 from the vowels/diphthongs at four equidistant temporal locations corresponding to 20%, 40%, 60% and 80% points of vowel/diphthong duration, to eliminate the effects of adjacent consonants on formant transitions and also to capture the dynamics of formant trajectories. The formants are manually corrected for improved modeling.

[3] http://www.fon.hum.uva.nl/praat/.

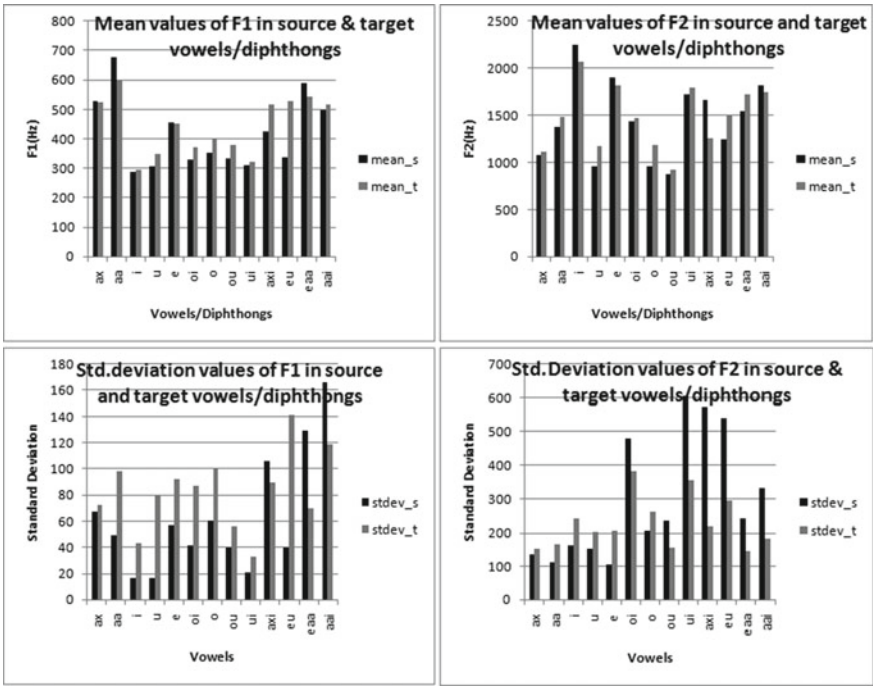


Fig. 40.1 Mean & Std.deviation of F1, F2 formant frequencies in source and target Vowels/Diphthongs

Methodology

A statistical analysis is carried out on the formants F1, F2, F3 of source and target vowels/diphthongs. It is observed that in both varieties of Assamese, F3 does not change much with vowel quality and therefore, its transformation is not of much significance. Results presented in Fig. 40.1, show a considerable difference in F1, F2 mean for most vowels/diphthongs. F1, F2 values deviate more from their means in target vowels than in source vowels while F2 values of source diphthongs deviates more from their means than in target diphthongs. In short, the source vowel space is different from that of target and our study attempts to carry out this transformation of formants.

In a GMM-based transformation, initially proposed for VC by Stylianou et al. [8], the system learns by fitting a GMM model to the augmented source and target feature vectors. During training, a GMM is adopted to model the distribution of the pairedfeature sequence z_t, representing the joint feature vector of source

speech vector x_t and target speech vector y_t at frame t. The Expectation Maximization algorithm[4] trains the GMM with the joint source and target vectors which are already aligned since the formants are extracted at four equidistant points in every vowel/diphthong segment.

The mapping function as described in [9], converting the formant frequencies, of source to target speech data is given by

$$F(x_t) = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E_{m,t}^{(y)} \tag{40.1}$$

where

$$P(m|x_t, \lambda^{(z)}) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})} \tag{40.2}$$

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (x_t - \mu_m^{(x)}) \tag{40.3}$$

The total number of mixture components is M and the weight of the mth mixture component is w_m. $\mu_m^{(x)}$ and $\mu_m^{(y)}$ are the mean vectors, $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$ are the covariance matrices, $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$ are the cross-covariance matrices of the mth mixture component of source and target feature vectors.

The three approaches, based on the GMMs, used to carry out the transformation of formants, are implemented in MATLAB. The mapping functions are developed to transform F1, F2, and F3 formants. However, since F3 transformation is not significant, transformation results presented in section "Experimental Results", are confined to F1, F2 only. Following subsections describe the three approaches in brief.

Approach 1: A Single GMM for F1, F2, F3 Formants

The first three formant frequencies, F1, F2, and F3, extracted from all the vowels/diphthongs of source and target speech data, at 20%, 40%, 60%, and 80% of vowel/diphthong duration, form the feature vector. The augmented feature vector consists of F1, F2, and F3 values of vowel/diphthong segments in both source and target training speech data and a GMM is trained to model this data. A mapping function is then developed to map the formants from source to target.

[4] http://www.cse.iitm.ac.in/~vplab/courses/DVP/PDF/gmm.pdf.

Approach 2: Separate GMMs for F1, F2, F3 at Equidistant Temporal Points of Vowel/Diphthong Segments

The second approach uses four separate mapping functions for the temporal points 20%, 40%, 60%, and 80% of vowel/diphthong duration. In order to estimate the parameters of the mapping functions, the probability distributions of the joint vectors $z_i = [x_i^T, y_i^T]$ where x_i and y_i are source and target formant frequencies F1, F2, F3 at i = 20%, 40%, 60%, and 80% of vowel/diphthong duration, are represented by separate GMMs. Each mapping function maps the formants at the specified time instant.

Approach 3: A Separate GMM for the Formants of Each Vowel/Diphthong Segment

In this approach, label files consisting of vowel/diphthong labels are assigned to the utterances, which are used to cluster extracted formants from training speech data into separate vowel/diphthong groups. A separate GMM-based mapping function is built for the formants of each vowel/diphthong. The number of mixture components (M) for each vowel/diphthong GMM is selected after experimenting with different values. During transformation, the label file associated with the test feature vector is used to select the appropriate mapping function for formant transformation. Transformed formant contours for the vowels/diphthongs in the test utterances are further smoothed using a Moving Average filter.

Experimental Results

The mapping functions are tested with different values of M, to transform formants F1, F2 of the vowels/diphthongs of 10 test utterances. For the first approach best results are obtained with M = 16, for the second with M = 8, while for the third, M is set to different values (4, 8, 16) for individual vowels/diphthongs. The transformation is evaluated objectively in terms of root-mean square error (RMSE) between test and target formants and between mapped and target formants. Transformation results for the vowels/diphthongs, in order of occurrence in a test utterance using the three approaches are presented in Figs. 40.2 and 40.3. To provide a quantitative representation of the transformation results, percentage improvement of RMSE values before and after transformation, with the three approaches, are calculated for all vowel tokens in the test utterances and presented in Fig. 40.4.

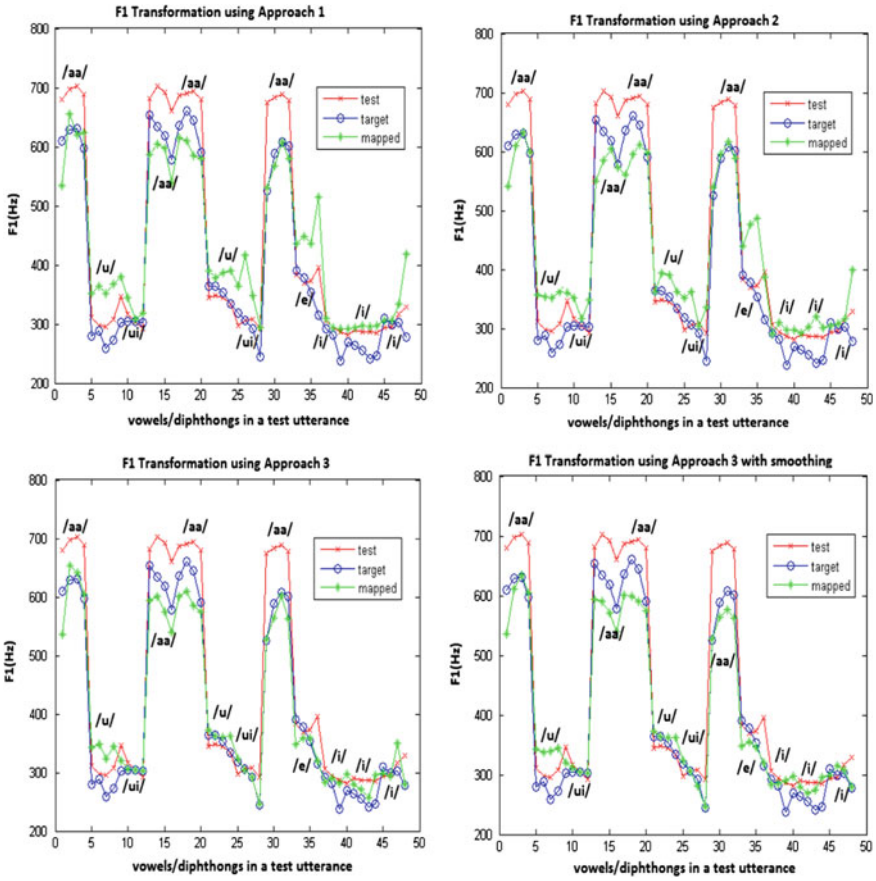


Fig. 40.2 Transformation of F1 formant frequencies

Conclusions and Future Work

Objective evaluation results indicate that in most of the transformations, with any of the three approaches, the source formants are brought closer to the target formants. In most cases, the third approach, where a separate mapping function is built for each of the vowels/diphthongs, outperforms the other two. The disadvantage of this approach is the additional requirement of label files for each of the utterances. Therefore, automatic vowel/diphthong labeling may be considered as a future task. Furthermore, to find out whether the transformation has been successful to some extent in incorporating regional variations into the test utterances, perceptual tests need to be carried out, after resynthesis of the test utterances with transformed formants.

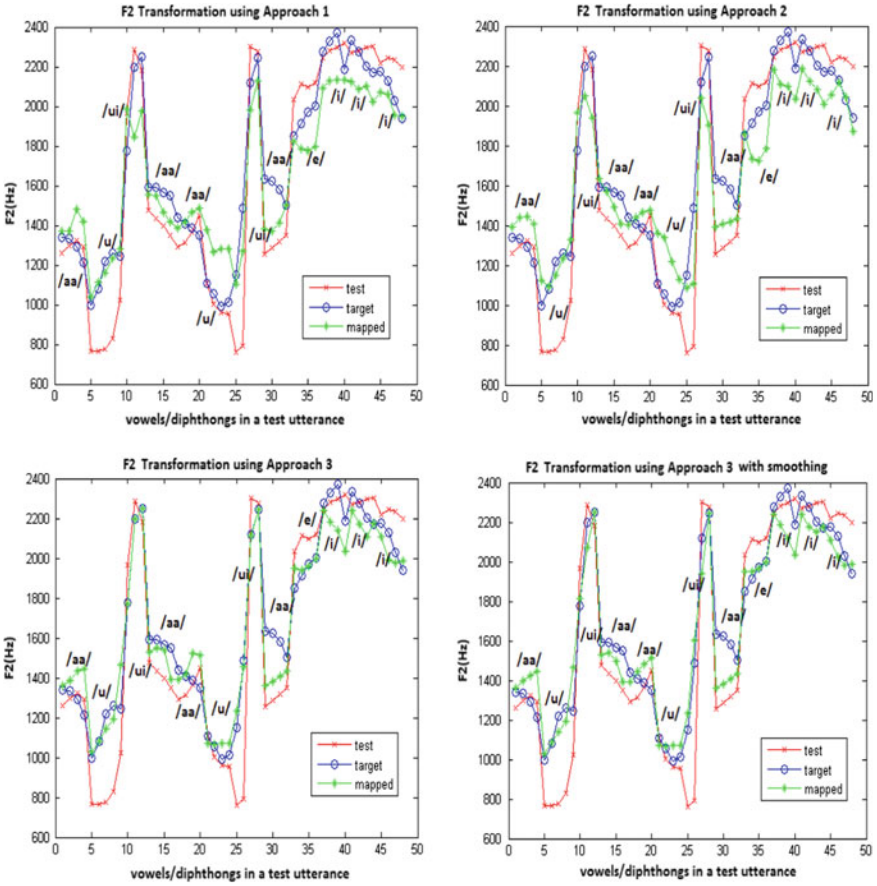


Fig. 40.3 Transformation of F2 formant frequencies

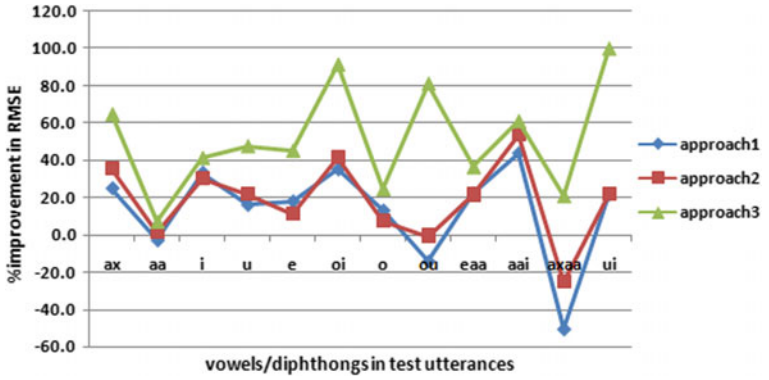


Fig. 40.4 Plot showing % improvement in RMSE values after transformation

Acknowledgements The authors are thankful to Trideep Baruah and Mancha J. Malakar for their help in corpus building and to MHRD Centre of Excellence for financial support.

References

1. Fox, R.A., Jacewicz, E.: Cross-dialectal variation in formant dynamics of american english vowels. The Journal of the Acoustical Society of America **126**(5) (2009) 2603–2618
2. Hagiwara, R.: Dialect variation and formant frequency: The american english vowels revisited. The Journal of the Acoustical Society of America **102**(1) (1997) 655–658
3. Nath, S., Sharma, U.: An analysis of the vowels and diphthongs of the assamese language and its nalbaria variety. In: Computing and Communication Systems(I3CS), 2015 International Conference on. (2015)
4. Labov, W.: Principles of linguistic change, cognitive and cultural factors. Volume 3. John Wiley & Sons (2011)
5. Teutenberg, J., Watson, C.: Vowel quality in accent modification. In: Proc. Australian International Conference on Speech Science & Technology, University of Auckland, New Zealand. (2006) 292–295
6. Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B.: Voice conversion using artificial neural networks. In: Automatic Speaker Recognition, Identification and Verification. (1994)
7. Rentzos, D., Vaseghi, S., Yan, Q., Ho, C.H.: Parametric formant modelling and transformation in voice conversion. International Journal of Speech Technology **8**(3) (2005) 227–245
8. Stylianou, Y., Cappé, O., Moulines, E.: Continuous probabilistic transform for voice conversion. Speech and Audio Processing, IEEE Transactions on **6**(2) (1998) 131–142
9. Toda, T., Black, A.W., Tokuda, K.: Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Transactions on Audio, Speech, and Language Processing **15**(8) (2007) 2222–2235

Chapter 41
English to Nepali Statistical Machine Translation System

Abhijit Paul and Bipul Syam Purkayastha

Abstract Machine Translation (MT), perhaps the earliest NLP applications, is the method of translating one human language sentence into another, using computer or any kind of machine. The aim of this research paper is to develop an MT system for Nepali language which can translate an English sentence to its most probable Nepali sentence using Statistical Machine Translation (SMT) approach. The system is implemented using three different tools like MOSES for decoding, GIZA++ for generating translation model and IRSTLM for estimating target model probability. Also for training the system, English-Nepali parallel corpus is used and for testing, English raw corpus is used. Both these two corpora are collected from TDIL (Technology Development for Indian Languages). The system has been manually evaluated using two parameters viz. fluency and adequacy and it gives an average accuracy of 2.7 out of 4 (level no), i.e., approximately 68%. Though the implemented system achieves an accuracy of 68% but for OoV (Out of Vocabulary) words the research still continuing. A small comparison has also been made with exiting English-Nepali MT system.

Keywords NLP · SMT · Computational model

Introduction

The IT revolution has led to an explosive growth of information and services in the web. This flood of information is available only a section of the society and beyond the reach of a significant portion in a multilingual country [1]. This has led a division in the society. This division in the multilingual country has triggered a

A. Paul (✉) · B. S. Purkayastha
Department of Computer Science, Assam University Silchar, Silchar, India
e-mail: abhijitpaul16@gmail.com

B. S. Purkayastha
e-mail: bipul_sh@hotmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_41

need for retrieving the information and services in a language of own choice. Machine Translation can help to overcome this technological barrier.

Today a number of MT systems are available for translating one natural language text into another natural language text, but they are not perfect; limited success has been achieved within a specific domain [2]. From the literature survey and related works, it has been found that not much work has been done previously on MT for Nepali to any language or any language to Nepali in India due to lack of a comprehensive set of the parallel corpus or a correct set of handwritten grammatical rules. Different approaches are there for MT, among them Statistical Machine Translation (SMT) system has earned special attention in recent years. The advantages of this system over traditional rule-based approach are that it is completely automatic and require less human effort. However, the system requires sentence-aligned parallel corpus for each language pair. Also, it cannot be used for language pairs for which such corpora do not exist. SMT translates sentence based on the information or statistics what it has trained or collected from a parallel corpus. This corpus consists of two texts, each of which is the translation of the other. Parallel corpus is also useful for other NLP tasks, such as information retrieval and extraction, bilingual dictionary, word sense disambiguation, etc. Training the translation model component in SMT requires large parallel corpora for the parameters to be estimated. The attempt of this work is to build a statistical model which can translate an English sentence to its equivalent Nepali sentence.

The rest of the paper is structured as follows: section "Related Works" discusses some of the existing MT systems for Indian languages. Section "Statistical Approach for Machine Translation" explains the fundamental SMT components. Section "Preprocessing Steps on Bilingual Corpus for SMT" mentions important preprocessing steps to develop the MT system. Section "Implementation and Result Analysis" describes the process of implementation and evaluation of English-Nepali MT system.

Related Works

Machine Translation is considered as one of the earliest NLP applications. It started its journey in 1959 abroad. But in India, it begins in the year 1980. Since then, many institutions and organizations are working on MT. Among them, the eminent institutes are [3] IIT Kanpur, C-DAC Mumbai, University of Hyderabad, C-DAC Pune, TDIL, etc.

Major popular MT systems that are available in India for Indian languages are as follows:-

(i) A domain free MT system, named "**Anusaaraka**" is designed in the year 1995 at IIT Kanpur. Though it is a domain free system but the system mainly used for translating children's stories [4].

(ii) In the year 1999 at C-DAC, Bangalore, "**Mantra**"—a machine translation system was developed mainly for the Rajya Sabha Secretariat. Nowadays, it also works for any Indian language pairs [5].
(iii) "**Matra**", another MT system has been developed in the year 2004 at C-DAC. It translates an English sentence to its equivalent Hindi sentence. It is a domain free system [6].
iv) Using the concept of rule-based approach and the generalized form of the lexicon, a MT system named "**AnglaBharti**" was designed. It has been developed in the year 1991 at IIT Kanpur. After "**AnglaBharti**", another MT system named '**AnuBharti**' was developed in the 2004 at same institution [7].
(v) **Shiva and Shakti,** the MT system "**Shiva**" mainly works for automatic translation of English sentences to different Indian languages. But the system "**Shakti**", works from English to Hindi, Marathi, and Telugu language only. It was developed in the year 2004 at Carneige Mellon University USA and IIIT Hyderabad [8].
(vi) In the year 2004 at Jadavpur University Kolkata, a machine translation system "**Anubaad**" was developed for common Bengali people those who don't know English. The system mainly translates English news to its equivalent Bengali news [9].
(vii) Sampark machine translation team of Consortium of Institutions has been designed a machine translation system named "**Sampark**" in the year 2009. It works for almost all language pairs [3].

Statistical Approach for Machine Translation

SMT approach is the most popular approach for translating natural language sentences. The SMT system is based on the fact that there will be many possible ways of translation for every sentence in one language to another. The choice among these possible translations is a matter of translator's preference [10]. Let us take the case of English to Nepali translations. This means that every Nepali sentence, 'n', is a possible translation of an English sentence, e. Assume that every pair of sentences (e,n) a probability, P(n|e), which is the probability that a translator when presented with an English sentence 'e', will produce 'n' as its Nepali translation. Also assume that when native speakers of Nepali speak English, they have in mind some Nepali sentence which they translate mentally. The goal of the translation system is to find the sentence 'n' that the native speaker has in mind when he produces 'e'. Minimizes the chance of error by selecting that Nepali sentence 'n' for which P(n|e) is maximum. Using Bayes's theorem, it can be written as:-

$$P(n|e) = P(e,n)/P(e) = P(n)XP(e|n)/P(e)$$

n can be calculated as

$$n = \arg \max_{n} \mathrm{P(n|e)} = \arg \max_{n} \mathrm{P(n)XP(e|n)/P(e)}$$

Since the denominator is independent of h, we can write the preceding expression as

$$n = \arg \max_{n} \mathrm{P(n)XP(e|n)},$$

where P(n) is prior to Nepali text and P(e|n) is conditional probability of the English text given the Nepali text. The idea is not intuitively appealing as it is hard to think of going through a list of sentences computing the product of prior and conditional probability while translating them.

Basically, there are two components in the SMT system. One is training pipeline and another is the decoder component. The input of the training pipeline component is the English-Nepali preprocessed file and the output will be two machine translation models, one is Language Model (LM) and another is Translation Model (TM). The input of the decoder will be two translation models and a source language sentence and output will be the best possible translated target language sentence. The architecture of the SMT system is shown in the Fig. 41.1.

The TM and LM are the responsible for the adequacy and fluency of the translated sentence respectively. If the translation is good, then adequacy of the translated sentence will be good, i.e., P(e|n) will be high. And if translation sentence is well formed, then its fluency will be good, i.e., P(n) will be high. In the implementation section, the working principles of LM, TM, and DECODER are discussed.

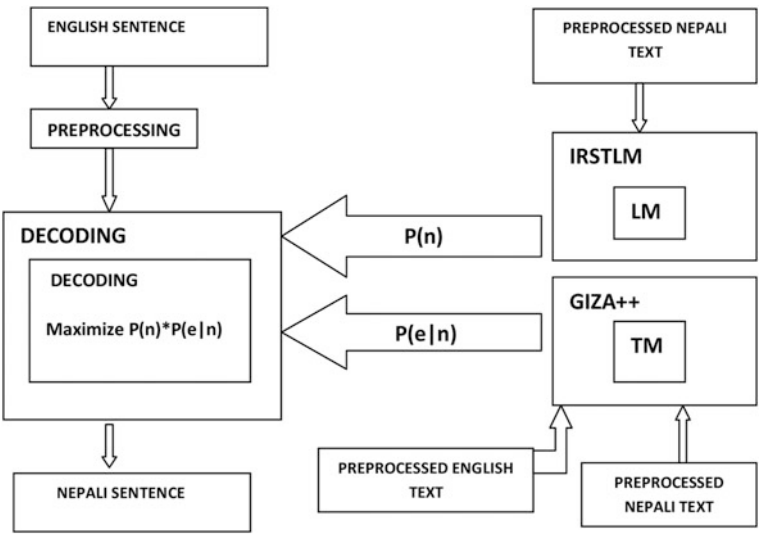


Fig. 41.1 Architecture of English-Nepali SMT system

Preprocessing Steps on Bilingual Corpus for SMT

Correct preprocessing techniques on the bilingual parallel corpus (text translated into two different languages) play an important role in NLP tasks, also in statistical machine translation systems. Preprocessing is the process of something before it is being processed by something else. In NLP, a preprocessor is a module which consists of a collection of programs. All the programs work in a sequential manner; the output of one program will be the input of another program. Development of English to Nepali SMT system is the main aim of this preprocessing works. Preprocessing module includes the following steps and is shown in the Fig. 41.2.

Step1. **Sentence Splitter:** The parallel corpus which has been collected from TDIL [11] consists of 51 files in .XL format and total number of sentences in these files are 8830 (4415 English sentences and 4415 Nepali sentences). The **sentence splitter** is a small JAVA which takes these 51 files as input and produces two separate text files, one file consists of English sentences and other consists of translated Nepali sentences.

Step2. **Tokenization:** Tokenization is the important text processing steps. It is the process of the breaking up of text into words. This means that spaces have to be inserted between words and punctuation. Though English language Tokenizer is easily available on the web for Nepali language or other Indian languages, it is not. Also, English Tokenizer does not work properly

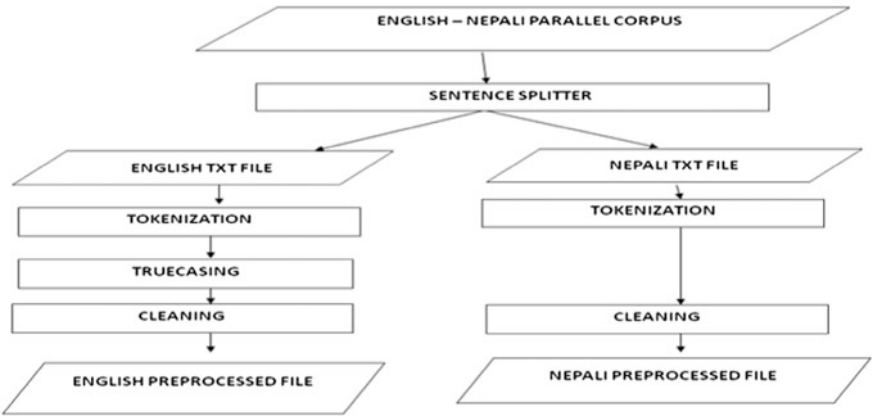


Fig. 41.2 Preprocessing module for SMT systems

for Indo-Aryan languages (e.g.,: Nepali) due to its multiple characteristics and compound words. So a Perl scripts is written for tokenization of Nepali sentences. And for English sentences, Moses inbuilt Tokenizer scripts have been used.

Step3. **Truecasing:** Related to tokenization, truecasing is the issue of words that occur in the lowercased or uppercased form in the text. The items house, House, and HOUSE may occur in a file, in the middle of a sentence, at the beginning of a sentence, or in the headline, respectively. Meaning of these words is same, so it is useful to normalize the case, i.e., convert to their most probable casing, typically by lowercase or truecasing. This helps the system to reduce data sparsity. The letter preserves uppercase in names, allowing the distinction between Mr. Fisher and a fisher. Since there is no concept of uppercase and lowercase in the Nepali language so applying truecasing on the Nepali language is meaningless. The output of English Tokenizer will be the input file of English truecasing.

Step4. **Cleaning:** Cleaning is the process of removing the long sentences, empty sentences and misaligned sentences from truecasing files. Long sentences, empty sentences, and misaligned sentences can cause problems in the training pipeline. Along with English truecasing files, Nepali Tokenizer files need to clean for smooth training.

Implementation and Result Analysis

The system is implemented using three different tools viz. (i) MOSES [12], i.e., DECODER, finding the best translating sentence, (ii) GIZA++ [13] for TM, generating conditional probability and (iii) IRSTLM [14] for LM, generating target language sentence probability. The input of the GIZA++ and IRSTLM will be the output of preprocessing module. GIZA++ constructs the phrase table, which contains the English words and its probable Nepali words. IRSTLM gives the probability of the words in 1g, 2g, and 3g method for the Nepali sentence. The system first reads the English sentence as an input word and then finds Nepali meaning of the English phrase (segment) from the phrase table. Once it gets the Nepali words, it goes for alignment. If the length of the English sentence is m and the length of Nepali sentence is n then for alignment total m * n combinations are possible. Then, the system finds the best alignment with the help of the language model, after that the DECODER will choose the best combination which gives the maximum value of P(n|e).

The input and output of the DECODER is shown below (Table 41.1).

The system trains on more than 5K sentences pairs and produces LM and TM. The translation of 100 English sentences was done to produce Nepali sentence. The system takes a file of 100 English sentences as an input. Out of 100 sentences, only 10 sentences are shown in this paper. There are two ways of evaluating any MT

Table 41.1 Input and output of the English-Nepali SMT system

Input sentence	Output sentence
Jaipur is known as a pink city	जयपुर गुलाबी रूपमा परिचित छ।
The best time to visit Jaipur city is October to March	उपदेश जयपुर सहर भ्रमणको नम्रित उचित समय **october** महिनादेखि मार्च महिनासम्म हो।
Shawls in Himachal Pradesh are very famous	हिमाचल प्रदेशका **Shawls** प्रसिद्ध छन्।
Birla Mandir is situated in New Delhi which is the capital of India	बिडला मन्दिर अवस्थित छ अनि नयाँ दिल्ली भारतको राजधानी हो।
Jaipur is known as a pink city	जयपुर गुलाबी रूपमा परिचित छ।
Sahajan built Tajmahal	**Tajmahal Sahajan** निर्माण गरेका थिए।
India gate is in Delhi	भारतको हो युक्त प्रवेशद्वार दिल्लीमा स्थापित गरे।
Kullu is the coolest place in Himachal	कुल्लु **coolest** हिमाचल स्थान हो।
Lalbagh is a botanical garden	**Lalbagh** एउटा उद्भिद उद्यान छ।
The Pushkar fair is one of the most colorful fairs of Jaipur	पुष्कर मेला उपदेश अतिरङ्गीन मेलाहरूमध्ये एउटा हो।
Ayodhya is famous for Lord Ram	अयोध्या भगवान् राम नम्रित प्रसिद्ध छ।

system one is automatic or another manual. In this paper, a manual evaluation method has been used. The translation was evaluated based on the two parameters viz. fluency and adequacy. Adequacy is responsible for convey the meaning in the translation sentence. And the fluency is responsible for the well-grammatical formed of the translation sentence. The levels (4, 3, 2, and 1) are used for parameters adequacy and fluency, which are discussed below [15].

Adequacy 4 means that the translated sentence expresses all the meaning of the source sentence, **Adequacy 3** indicates the translated sentence conveys all the meaning of source sentence, **Adequacy 2** indicates that the translated sentence conveys some meaning and **Adequacy 1** means that the translated sentence has no meaning conveyed. And for **Fluency 4** indicates that good grammar of sentence that is translated, **Fluency 3** indicates that the translated sentence is easy to understand but has lack correct grammar, **Fluency 2** means that the translated sentence is broken, but is understandable with efforts and **Fluency 1** indicates that **t**he translated sentence is not clear.

The system was evaluated by three linguistics persons. The geometric average of the individual parameters was taken and the results are shown in the Table 41.2.

Table 41.2 Accuracy of the proposed system

Geometric average (out of 4)		Accuracy (%)	Overall accuracy
Fluency	2.5	62.5	67.5%
Adequacy	2.9	72.5	

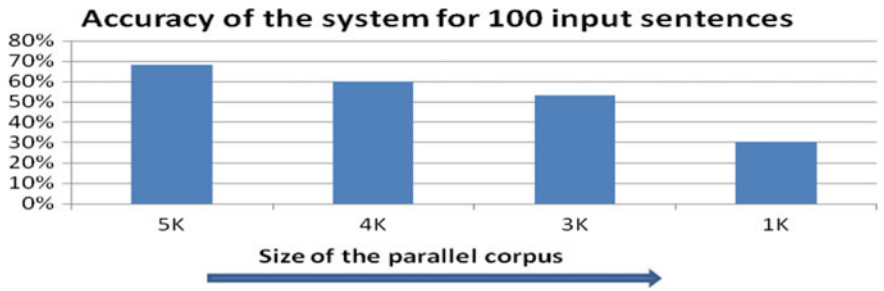


Fig. 41.3 Accuracy of the proposed system on various sizes parallel corpus

One thing it has been observed in the overall evaluation process that the system gives poor results if it reduces the size of the parallel corpus and trains the system. Following figure clearly shows the behavior of system and variation in the accuracy (Fig. 41.3).

Comparison

The proposed system has been compared with existing English-Nepali MT system, which was developed by TDIL [16]. The existing system was implemented using rule-based approach and it uses the concept of morphological analysis, part of speech tagging, word sense disambiguation, and a bilingual English-Nepali dictionary. But the proposed system was implemented using statistical approach; it uses few preprocessing steps along with generated language model and translation model. The existing system was tested with same input sentences, which were used for testing the implemented system and it achieves accuracy over 80%. Though it achieves accuracy over 80%, but there is no provision for a file as an input, it takes only sentences as an input. Also, the time it takes to generate the output is little high compare to the proposed system.

Conclusion and Future Direction

This paper discussed the statistical approach for developing English to Nepali MT system. The statistical approach requires a reasonable amount of parallel corpus for training. In this research work, a small amount of parallel corpus of approx. 5k sentence pairs is used for training the system. This is very small compared to a good SMT system which uses millions of sentence pairs for training. Though the system gives an accuracy of approximately 68%, but if one can increase the size as well as improve the quality of the parallel corpus, the system may produce better results.

Also, if one can do the transliteration after translation for OoV (Out of Vocabulary) words and Name Entity words, then the system may produce better results. Automatic evaluation technique can also be considered as one of the future work.

References

1. Siddiqui, T., Tiwary, U.S.: Natural Language Processing and Information Retrieval, New Delhi, Oxford Press (2010).
2. Taghipour, K., Afhami, N., Khadivi, S., Shiry, S.: A discriminative approach to filter out noisy sentence pairs from bilingual corpora, Telecommunications (IST), 5th International Symposium. pp 537–541 (2010).
3. Dwivedi, S.K., Sukhadeve, P.P.: Machine Translation System in Indian Perspectives, J. Computer Science, vo 6, Issue 10, pp 1111–1116.
4. Bharati, A., Chaitanya, V., Kulkarni, A.P., Sangal, R.: Anusaaraka: Machine translation in stages. pp 22–25.
5. Bandyopadhyay, S.: Use of machine translation in India. AAMT J., 36: pp 25–31.
6. Rao, D., 2001. Machine translation in India: A brief survey. Proceedings of SCALLA 2001 Conference, (SCALLA'01), National Centre for Software Technology. Bangalore, India, pp: 1–6.
7. Sinha, R.M.K., Jain, R., Jain, A.: Translation from English to Indian languages, Anglabharti approach. Proceeding of the Symposium on Translation Support System. pp 15–17. IIT, Kanpur, India (2001).
8. Bharati, A., R. Moona., P. Reddy., B. Sankar., Sharma, D.M.: Machine translation: The Shakti approach. Proceeding of the 19th International Conference on Natural Language Processing. MT-Archive, India, pp: 1–7 (2003).
9. Bandyopadhyay, S.: ANUBAAD - the translator from English to Indian languages. proceedings of the 7th State Science and Technology Congress, (SSTC'00), Calcutta, India, pp 1–9 (2000).
10. Rao, D.D.: Machine Translation A Gentle Introduction. RESONANCE (1998).
11. Technology Development for Indian Languages, http://www.tdildc.in.
12. Statistical Machine Translation System User Manual and Code Guide, http://www.statmt.goes.org/moses/manual/manual.pdf/.
13. GIZA++, https://github.com/moses-smt/giza-pp/tree/master/GIZA++.
14. IRSTLM, https://sourceforge.net/projects/irstlm/.
15. Brown, P.F., De, S., Pietra, V.D., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. J. Computational Linguistics, vol. 19, no3, (1993).
16. Technology Development for Indian Languages, http://tdil-dc.in/mt/common/php#.

Chapter 42
Language Identification on Code-Mix Social Text

Nayan Jyoti Kalita and Navanath Saharia

Abstract Social media/network has become one of the comfortable medium for people to share their feelings instinctively. With the increasing use of social media, language identification of code-mixed text is a new problem to the linguistics as they influence the growing use of informal languages. Most of the traditional language detectors fail to identify the language. This paper describes a study to detect languages at the word level in English-Assamese code-mixed text. For the work, we have collected texts from Facebook groups and pages. We develop a system to evaluate the level of mixing between different languages in our corpus and to detect the languages. Our experiments have been carried out using a linear kernel support vector machine.

Keywords Language identification · Code-mixed text · SVM
English · Assamese

Introduction

The focus of this chapter is to identify the name of the language in a given text. Language identification as a research problem is addressed many times and solutions with accuracy 90–99% are already devised for mono script/lingual texts/documents. In the advent of social media/network, texting (henceforth social text) is reached in such a height that only the agents associated with the texts understand the meaning, though they are public. Most of the cases, texts are bi/trilingual, unstructured, full of nonstandard abbreviations, and phonetically biased. A huge amount of texts are generating as a part of day-to-day communication, which

N. J. Kalita (✉)
Royal School of Engineering and Technology, Guwahati 781035, India
e-mail: nayan.jk.123@gmail.com

N. Saharia
Indian Institute of Information Technology Manipur, Imphal 789002, India
e-mail: nsaharia@iiitmanipur.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_42

increases the importance of extracting information from such user-generated content. Being dependent on languages, the devised language detection techniques often fail to identify the languages in a mixed language texts. Among the reasons informal use/selection of language lexicon, a high percentage of spelling errors, and mixing the language properties are main.

Fischer [9] provides an interesting insight on the usage of language in Twitter in different geographical regions. He states that Europe and Southeast Asia are the most language-diverse areas of the ones currently exhibiting high Twitter usage. Language identification in code switching text is different from identifying multiple languages in document [21], as the different languages present in a single document might not necessarily be due to instances of code mixing.

This paper is organized as follows. The next chapter discusses the previous work on code mixing in a social media text. Section "Code-Mixed Corpus" discusses the code-mixed corpus collected for our experiment. The support vector machine-based language detection technique is explained in section "SVM-Based Word Language Detection." Our experiments along with results are discussed in section "Experiments and Results." We conclude our report with section "Conclusion."

Related Work

Being a well-known source of user-generated content, short messages of social network sites are applied in almost all areas of information extraction research including emotion, sentiment, and attitude detection [20, 27], stock market prediction [3, 19], disaster management [23, 25, 26], and disease surveillance [17]. Being so widespread, the language identification problem needs new means to identify language from these short messages. Though the problem of identifying the language in code-mix data is not new [2, 10, 15], researchers are still working on it because of the dynamic nature of the data and changing media.

Several approaches have been proposed to identify code-mixed language; it includes Markov models [7], Monte Carlo methods [22], dot products of word frequency vectors [5], morphological and phonological analysis [8]. Many researchers use the n-gram feature, first suggested by [4], which classifies documents by comparing n-gram profiles between a document and a global language profile. More recently, support vector machines (SVM) have been applied to language identification as a new popular paradigm [1, 6, 14, 16].

The reasons for and types of code mixing has also been studied by several researchers. These include the surveys on Cantonese-English code switching in Hong Kong [18] and Macao [24]. They stated that linguistic motivations were sparking code switching in those highly bilingual societies. The paper reported in [13] showed that, in Facebook comments, all the types of code switching, such as intra-sentential, tag and inter-sentential switching appeared.

Code-Mixed Corpus

Most of the researches on Social Media analysis has so far based on the English language. India has around approx languages along with 22 official languages. Code mixing is very common in this region. The English-Assamese language pairs were selected for our study. The Assamese language is of Indo-Aryan origin and is the major language in the North Eastern India. We have collected the code-mixed text through various posts and comments from one Facebook page.[1] Our corpus contains 1012 utterances, 26,444 words, and 5,977 unique tokens. Since Hindi language works as a bridge between different tribes, it has great influence on all the other regional languages. In the corpus, we found a small amount of Hindi words (approx 3–5%).

Annotation

We tagged the corpus at the word-level with eight different tags. The tag-set is presented in Table 42.1. The English, Assamese, and Hindi words are tagged with EM, AS, HI tags, respectively. To denote the emoticons or special characters, such as ”, ’, !, @, we used UNIV. WE used NE tag to denote the Named Entity and UNDEF to the tokens that are hard to categorize.

Level of Code Mixing in the Corpus

We have different types of code switching categories, such as inter-and intra-sentential, or intra-word and tag. The distribution of different code switching present in our corpus is as follows: 26.69% from intra-sentential, 69.26% from inter-sentential and 4.05% from intra-word. If all the words of a sentence are completely tagged with either English or Assamese, then that sentence was considered as an inter-sentential code switching.

1. **Inter-sentential**
   Pohi/AS hosak/AS val/AS lagile/AS …/UNIV
2. **Intra-sentential**
   eman/AS short/EN hand/EN moi/AS eku/AS buji/AS npalu/AS

[1] https://www.facebook.com/GhyMetroConfessions.

Table 42.1 Code mixing tag-set

Tag	Description	Count (%)
EN	English word	64.6
AS	Assamese word	16.3
UNIV	Emoticons, universal characters	13.1
NE	Named entity	2.0
HI	Hindi word	3.0
ACRO	Acronym	0.5
WM	Word-mixing	0.2
UNDEF	Tokens that are hard to categorize	0.1

3. **Intra-word**
   Etya/AS nu/AS kaunsa/AS bhaal/AS servicetu/WM d/AS ase/AS :p/UNIV
4. **Tag**
   Oh/UNIV usrt/AS a/AS sun.AS ./UNIV moi/AS ghy/NE t/AS thka/AS bili/AS vab-slu/AS

Some sentences collected from our corpus is given above as examples. In the intra-word case, servicetu is the mixing of two words from two different languages, service (English) and tu (Assamese). Similarly, in the tag switching case, one tag from English word, ghy is inserted into the sentence.

SVM-Based Word Language Detection

SVM are learning machines that can perform binary classification. They are one of the most popular methods for a new paradigm of classification and learning. SVM is used in various domains including language identification. SVMs minimize the expected error instead of minimizing the classification error. SVM include several classes of methods which differ among themselves based on the kernel function and the linear or nonlinear separating surfaces between classes. For our experiment, we used Weka v3.6.12 [11] for SVM with the default parameters. The linear kernel SVM system was trained with the following features:

n-gram with weights: This feature is carried out using the bag-of-words principle. If there are n unique n-grams for a language pair, then consider them as n unique features. For example, that in is the *i*th bigram in the list. In a given word w (e.g., raining), a particular n-gram occurs k times (twice for in raining). If the precalculated weight of in is t_w^i, the feature vector is 1; 2; :::; (t_w^i k); :::; (n 2); (n 1); n.

Acronym check: For acronym and named entity, we use their boolean features. In general, named entity begins with a capital letter. Again, acronyms are written in all capital letters. Therefore, we have checked whether all the letters are capitalized, only the first letter is capitalized or any letter of the word is capitalized.

Dictionary-based: Three dictionaries are used in our experiment as a binary features. For English, a Lexical Normalization Dictionary [12] and ABC

(Australian Broad-casting Commission 2006)[2] are used. Since no dictionary is available for Assamese, we used an Assamese corpus, collected from Indian Language Technology Proliferation and Deployment Centre,[3] as our dictionary. This corpus is written in Unicode form. Hence, it is transliterated into Romanized text so that it can be used with our phonetic corpus data.

Word Position: We use the position of the words as a feature value. For this, the position of a word in particular utterances is directly used.

Experiments and Results

As mentioned earlier, we trained the SVM with n-gram and dictionary-based features. Our experiment is many folded with different features, as shown in Table 42.2. In word-level evaluation, we obtained 89.51% accuracy from the code-mixed corpus.

In order to analyze the errors made by the SVM classifiers, we performed an error analysis. Table 42.3 reports the F-scores for each of the feature sets along with the accuracy. Table 42.4 shows the confusion matrix of our experimental setup for all the tags over our code-mixed data. The confusion matrix contains information about actual and predicted classifications done by the system. The diagonal cells of the matrix correspond to the correctly classified instances. For example, in row number 5 and column number 3, the cell contains actually an English word, that is 'en' tag, incorrectly classified as 'univ'.

From the table (see Table 42.4), it is clear that highest confusion is in between 'en' and 'as' tag class. Though both language English and Assamese are highly dissimilar in structure and syntax, our system misclassified some of the tokens tagged with 'en' and 'as'. Many English words, like 'table', are used for a long time in Assamese that they too become native words.

Discussion

The code-mixed text is a new challenge to the NLP. Everyday social media users create new forms of words. For example, many times they create new words by adding suffixes coming from Assamese to the English root words. eg, timeor, familyk, and customere. Sometime they create new words just for fun. Another problem we faced is the Assamese words written in different transliterated forms. For example, social media users may use 'ghor', 'ghr' or 'ghar' to indicate house, because there is the standard form for transliteration.

[2] http://www.abc.net.au/.

[3] http://www.tdil-dc.in/index.php.

Table 42.2 Test set results for language detection from code-mixed test

Features	Precision	Recall	F-score
n-gram	0.798	0.792	0.765
n-gram + emnlp dict	0.798	0.792	0.765
n-gram + emnlp dict + AS Dict	0.798	0.792	0.766
n-gram + emnlp dict + AS Dict + ABC Dict	0.891	0.895	0.889

Table 42.3 Token level results on the test data

Description	Tag	Precision	Recall	F-score
Acronym	acro	0.000	0.000	0.000
Universal characters	univ	0.778	0.958	0.859
Named-entity	ne	0.750	0.333	0.462
English word	en	0.935	0.950	0.942
Assamese word	as	0.760	0.670	0.712
Hindi word	hi	0.538	0.184	0.275
Word-mixing	wm	0.000	0.000	0.000
Undefined	undef	0.000	0.000	0.000

Table 42.4 Word-level confusion matrix for all the 8 tags

		acro	univ	ne	en	hn	as	wm	un
Acronym	(acro)	0	0	0	1	0	0	0	0
Universal characters	(univ)	0	249	0	4	0	7	0	0
Named-entity	(ne)	0	1	6	9	0	2	0	0
English word	(en)	0	22	2	1208	1	39	0	0
Hindi word	(hn)	0	4	0	5	7	22	0	0
Assamese word	(as)	0	44	0	65	5	231	0	0
Mixed word	(wm)	0	0	0	0	0	3	0	0
Undefined	(un)	0	0	0	0	0	0	0	0

The northeastern part of India is home to different tribes. They live very closely with each other. Therefore, they have a great effect on their cultures, languages from the neighboring tribes. We can find this effect on the social media also. Again, we have some words which are present in both Assamese and Hindi dictionaries. For example, ‘Bhai’ (Brother in English) have same meanings both in Assamese and Hindi.

Conclusion

In this digital world, the number of social media users is growing exponentially. Code mixing is a new problem to the linguistic works. In this paper, we have reported an experiment on the language identification of code mixing in a social media text. We have used SVM-based classification with n-gram and dictionary features. In our experiment, we have focused only in Facebook posts or comments written in English or Assamese language. Generally, data collected from different social media exhibits different characteristics. For example, Facebook posts tend to be longer than Twitters' tweet as there is no any limitation on lengths. In near future, we would like to experiment on other social media texts as they might carry some special linguistic characteristics.

References

1. Barman, U., Wagner, J., Chrupała, G., Foster, J.: Dcuuvt: Word-level language classification with code-mixed data. EMNLP 2014 p. 127 (2014)
2. Beesley, K.R.: Language identifier: A computer program for automatic natural-language identification of on-line text. In: Proceedings of the 29th Annual Conference of the American Translators Association, vol. 47, p. 54. Citeseer (1988)
3. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science 2(1), 1–8 (2011)
4. Cavnar, W.B., Trenkle, J.M., et al.: N-gram-based text categorization. Ann Arbor MI 48113 (2), 161–175 (1994)
5. Damashek, M., et al.: Gauging similarity with n-grams: Language-independent categorization of text. Science 267(5199), 843–848 (1995)
6. Das, A., Gamback,¨ B.: Identifying languages at the word level in code-mixed indian social media text. In: Proceedings of the 11th International Conference on Natural Language Processing, Goa, India, pp. 169–178 (2014)
7. Dunning, T.: Statistical identification of language. Computing Research Laboratory, New Mexico State University (1994)
8. Elfardy, H., Diab, M.T.: Sentence level dialect identification in arabic. In: ACL (2), pp. 456–461 (2013)
9. Fischer, E.: Language communities of twitter (2013)
10. Gold, E.M.: Language identification in the limit. Information and control 10(5), 447–474 (1967)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
12. Han, B., Cook, P., Baldwin, T.: Automatically constructing a normalisation dictionary for microblogs. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 421–432. Association for Computational Linguistics (2012)
13. Hidayat, T.: An analysis of code switching used by facebookers (a case study in a social network site). Ph.D. thesis, BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia, October (2012)
14. Jalam, R., Teytaud, O.: Kernel-based text categorisation. In: Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on, vol. 3, pp. 1891–1896. IEEE (2001)

15. Joshi, A.K.: Processing of sentences with intra-sentential code-switching. In: Proceedings of the 9th conference on Computational linguistics-Volume 1, pp. 145–150. Academia Praha (1982)
16. Kruengkrai, C., Srichaivattana, P., Sornlertlamvanich, V., Isahara, H.: Language identification based on string kernels. In: Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on, vol. 2, pp. 926–929. IEEE (2005)
17. Lamb, A., Paul, M.J., Dredze, M.: Separating fact from fear: Tracking flu infections on twitter. In: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 789–795. Atlanta, USA (2013)
18. Li, D.C.: Cantonese-english code-switching research in Hong Kong: a y2k review. World Englishes 19(3), 305–322 (2000)
19. Li, Q., Zhou, B., Liu, Q.: Can twitter posts predict stock behavior?: A study of stock market with twitter social emotion. In: Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on, pp. 359–364. IEEE (2016)
20. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1–167 (2012)
21. Nguyen, D., Dogruoz, A.S.: Word level language identification in online multilingual communication. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (2014)
22. Poutsma, A.: Applying monte carlo techniques to language identification. Language and Computers 45(1), 179–189 (2002)
23. Saharia, N.: Detecting emotion from short messages on nepal earthquake. In: Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on, pp. 1–5. IEEE (2015)
24. San, H.K.: Chinese-English code-switching in blogs by macao young people (2009)
25. Takahashi, B., Tandoc, E.C., Carmichael, C.: Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the Philippines. Computers in Human Behavior 50, 392–398 (2015)
26. Teodorescu, H.N.: Using analytics and social media for monitoring and mitigation of social disasters. Procedia Engineering 107, 325–334 (2015)
27. Teodorescu, H.N., Saharia, N.: A semantic analyzer for detecting attitudes on sns. In: 2016 International Conference on Communications (COMM), pp. 47–50. IEEE (2016)

Chapter 43
Automatic Visualization of Product Features Using LDA and Word2Vec

Thomas N. T and Rajeev Mullakkara Azhuvath

Abstract In an e-commerce environment, the product features and specifications may not be intuitive to a customer. The customers are often forced to go through the whole product description to find products that match their purpose. Since customer satisfaction is the success of any company, it is necessary that we present our product features in short and simple, yet catchy way. Also, it is very difficult to manually arrange product features from a corpus of data. Hence, we propose a system which provides the users with a hierarchy of salient features from a corpus of the retail product description. First, we abstract major topics from the input corpus using Latent Dirichlet Allocation (LDA). Next, we find the feature vectors for the words in the input corpus. Continuous skip-gram model of Word2Vec is used to predict the context words surrounding each topic. A hierarchy of salient features is then formed based on the extracted topics and corresponding context words. The output is visualized by forming a collapsible tree structure.

Keywords Latent Dirichlet allocation · Internet shopping · Word2Vec

Introduction

E-shopping or online shopping is a part of electronic commerce (e-commerce) that allows the users to purchase and sell products over the internet. The online shopping has increased tremendously in our country and has made a revolutionary change in the way we sell and buy. It helps the customers to search and buy products at the cheapest rate by sitting anywhere, at any time. E-commerce companies consider customer's time, security concern, and trust as the main aspects of internet shopping.

Thomas N. T (✉) · R. M. Azhuvath (✉)
Artificial Intelligence Lab, Tata Consultancy Services, Kochi, Kerala, India
e-mail: thomas.nt@tcs.com

R. M. Azhuvath
e-mail: rajeev.azhuvath@tcs.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_43

The operational cost is very low for e-shopping. Plus, there are different types of products to attract and satisfy customers. However, the customer may have to go to multiple pages of the website to buy a product and it is not very easy to find products that match their purpose from this large set of items. To solve this problem, we propose a system which helps the customers to find their items easily and in less time. To achieve this goal we use certain algorithms.

Latent Dirichlet Allocation (LDA) [1] algorithm is used to abstract the topics from the input dataset of that particular product. It is a probabilistic generative method used for topic modeling. In this model, each document is considered as a mixture of a small number of topics and each word's creation as an attribute to one of the document's topics. An algorithm called RAKE (Rapid algorithm for Keyword Extraction) is also used to extract concepts, phrases, and keywords from a document. Next, the words similar to the extracted topics are identified using Word2Vec from the phrases and considers that these keywords represent the main ideas expressed in that document.

Word2Vec produces word vectors as output. This can be achieved using either of the two models such as CBOW (Continuous Bag-Of-Words) or skip-gram to carryout word embedding. It is a bag-of-words model.

Related Work

Schmitt and Bergmann [2], in their work use case-based reasoning technology for intelligent sales support. In their system, cases are compared using similarity measures and then using adaptation techniques further customization of retrieved data. However, the paper does not take of the optimization and maintenance aspects of CBR.

Katawetawaraks and Wang [3] in their paper explains the different factors that influence the decisions of an online shopper.

Yu et al. [4] in their paper explains how the drawbacks of product search on e-commerce sites are addressed by developing a Latent Dirichlet Allocation-based retrieval approach along with a Multivariate Bernoulli LDA model.

Schmitt and Bergmann [5] in their paper explains the factors that make the users leave the site in frustration, e.g., there were

- too many questions,
- too difficult questions,
- redundant questions,
- wrong or no search results, etc.

The paper also says that e-commerce sites implement search improperly, and the shoppers struggle with poor search interfaces. Paper explains that this can be time-consuming, boring, and/or an even unsolvable task for the customer, especially if she is not familiar with the complexity of the products (e.g., personal

computers) and that it would be worst case for both, vendors and potential buyers, if the system returned wrong results or no results after the acquisition phase.

Proposed Method

The input dataset that is given to the system is a detailed description of products. From the dataset, the proposed system extracts the most relevant words to show to the user. Figure 43.1 shows the workflow of the system. In order to provide the users with only relevant details of the product, from a large set of product details, the proposed system follows two main steps:

1. LDA (Latent Dirichlet Allocation)
2. Word2Vec

LDA (Latent Dirichlet Allocation): LDA posits that, by analyzing the set of words a large corpus of data uses, an outline of that text can be obtained. This is because the set of words used in document carry very strong semantic information. In LDA [1] these words are interpreted as topics and these topics make a short description of the entire corpus of data. It is a joint distribution, given the hyper-parameters, as:

$$\left(\prod_{k=1}^{k} P(\beta_k|\eta)\right)\left(\prod_{n=1}^{D} P(\theta_d|\alpha)\left(\prod_{n=1}^{N} P(Z_{d,n}|\theta_d).P(W_{d,n}|Z_{d,n},\beta)\right)\right),$$

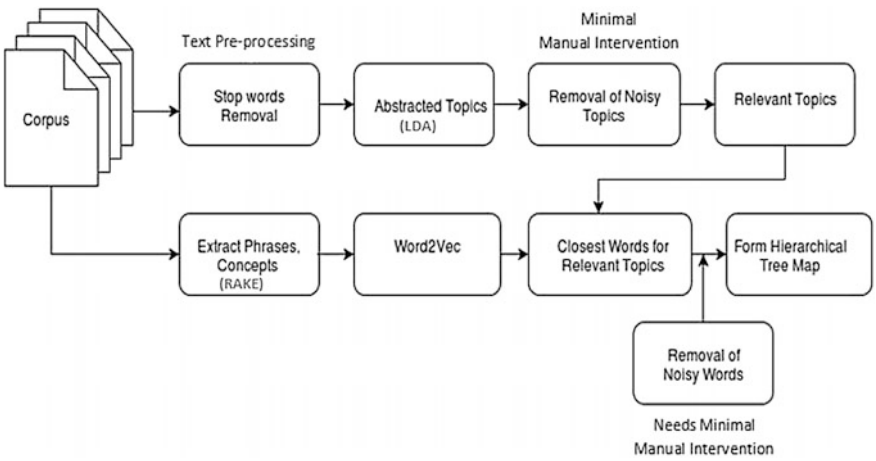


Fig. 43.1 Workflow of the system

where,

α represents the Dirichlet hyperparameter for document topic distribution.
θ_d is the per-document topic proportion in case of a collection of the document.
$Z_{d,n}$ is the per-word topic assignment in case of a collection of the document.
β_k is the per-corpus topic distribution in case of a collection of the document.
$W_{d,n}$ is the observed word.
η is the Dirichlet parameter.
N is the document length.
D is the number of documents.
k is the number of topics.

LDA is applied to a multiple product descriptions using the following steps:

Step 1: Provide the algorithm with the required number of topics needed.
Step 2: Every word will be assigned to a temporary topic using Dirichlet distribution.
Step 3: Algorithm will check and update topics (iterative).
Step 4: To get the most important features, specify the number of top N words.

Irrelevant topics are then removed from the abstracted topics, using minimal manual intervention, to get the most relevant topics.

Word2Vec: From the relevant topics, in order to identify the most relevant words we use Word2Vec method.

First, by using RAKE (Rapid Algorithm for Keyword Extraction), the keywords, phrases, and context from the corpus are extracted. Rake splits the text into sentences by treating punctuation signs as sentence boundaries. For phrases, all words listed in the stop words file is treated as phrase boundaries.

Word2Vec is then applied to the output of Rake, to get the words similar to the input word (context) of that particular product. Here, the extracted topics from LDA are considered as the input words for Word2Vec. It uses a single hidden layer, fully connected neural network, to find a high probability for a target word when a context word is given as input. It provides an efficient implementation of the continuous bag-of-words architecture for computing vector representations of words. It produces the word vectors as output. The output of Word2vec is given and is visualized on screen.

For each abstracted topics from LDA of that particular product, a hierarchical relationship is formed with the most similar words (features) extracted using word2vec. The output is visualized by using a JavaScript library for manipulating documents based on the data called D3.js (Data-Driven Documents) by forming a collapsible tree. The tree structures formed are shown in Figs. 43.2 and 43.3. The tree layouts share a common root.

By clicking on one node, the tree translates to accommodate new nodes. In Figs. 43.2 and 43.3, we used product description of a dataset containing wrist watch data and refrigerator data as input corpus.

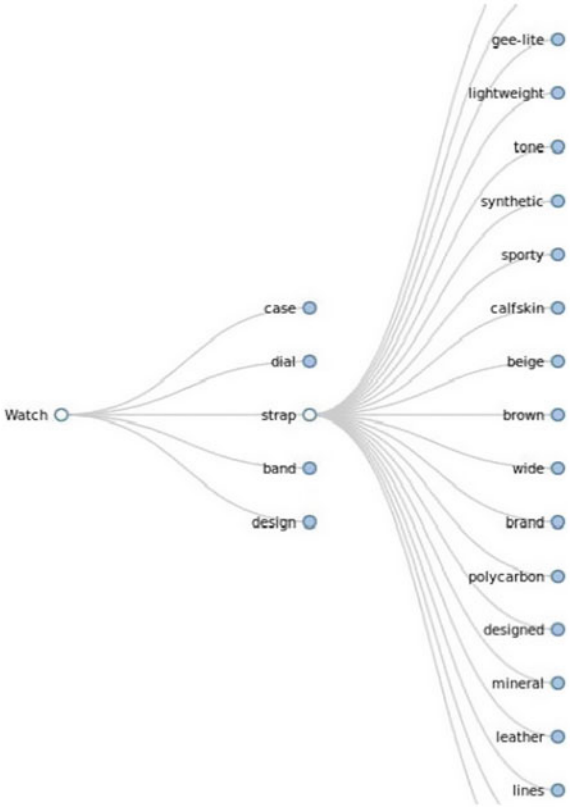


Fig. 43.2 Collapsible tree structure of wrist watch data

As the input dataset, which is the multiple product descriptions, is bigger and complex, the dimensionality of data increases. The high-dimensional input dataset is then visualized on the screen using a dimensionality reduction algorithm called t-SNE (t-distributed stochastic neighbor embedding). It helps to visualize the hidden data unmasking the clusters. Here, we used Barnes-Hut t-SNE and is visualized in a standard scatter plot.

For example Fig. 43.4 depicts the Barnes-Hut t-SNE scatter plot of the watch. Since the system output of watch in Fig. 43.2 shows a clear similarity to the clusters formed by the Barnes-Hut t-SNE scatter plot of the watch in Fig. 43.4, it is understood that the system output is similar to the expected output (Fig. 43.5).

Figure 43.4 depicts the sample UI of the system. The UI clearly shows us that the system is user-friendly. If say the user has a leather strap watch in his mind. He first searches for the term "watch" in the search bar. After he clicks on search, he is given options such as: strap, brand, dial, etc., based on which he can precede his search. Since he wants leather watches he can choose leather and continue with his purchase. The main advantage of this type of system over other shopping sites like Amazon is that, the user need not remember the exact keywords of the item he

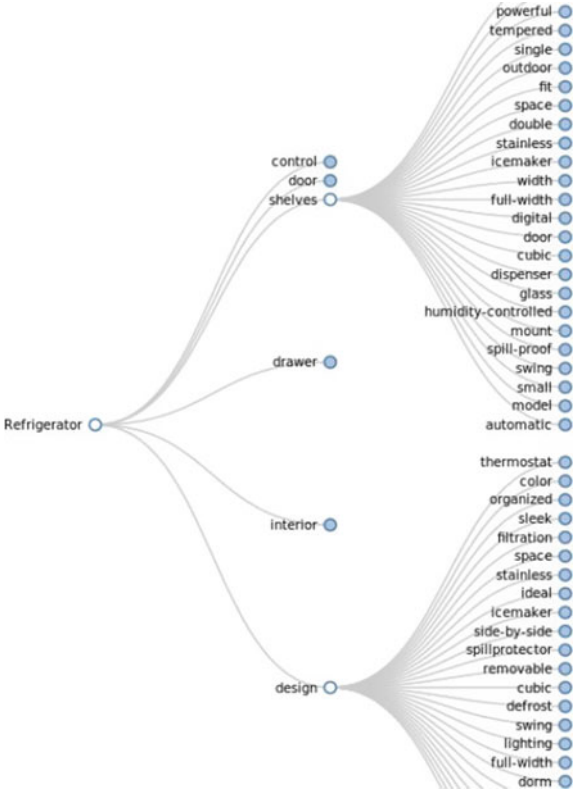


Fig. 43.3 Collapsible tree structure of refrigerator data

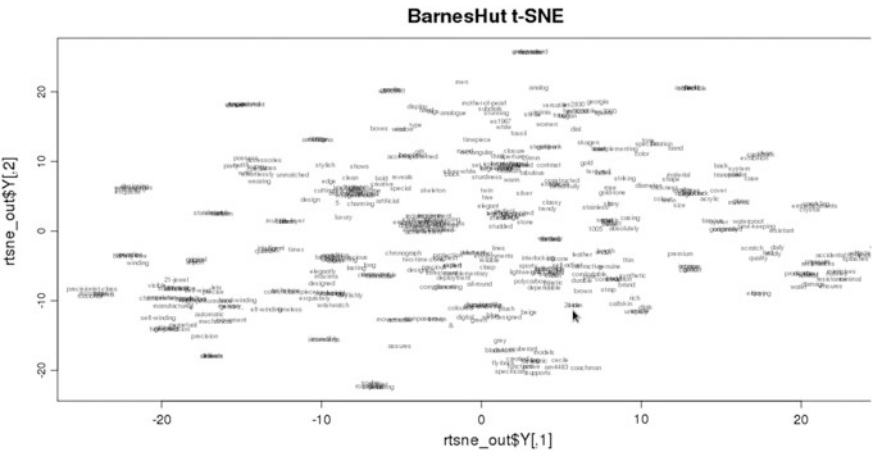


Fig. 43.4 Barnes-Hut t-SNE standard scatter plot of watch data

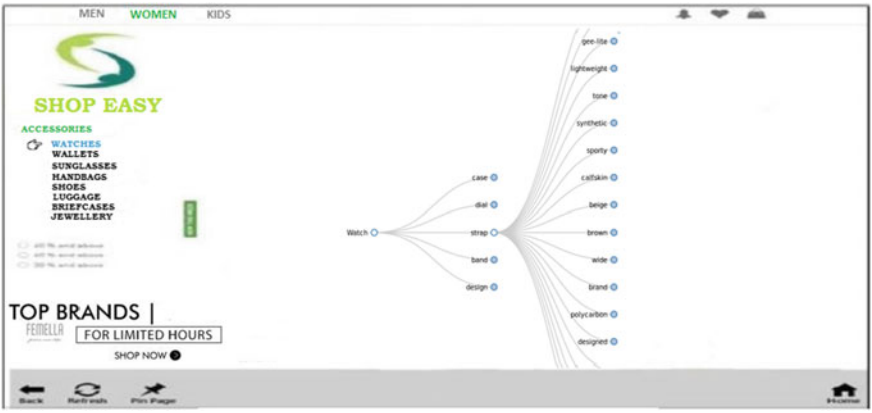


Fig. 43.5 Sample UI of the system

wants to purchase. This system takes the user to his desired item in an easy manner for the effective shopping experience and customer satisfaction.

Experimentation

We conducted an online survey to understand the importance of the idea of extracting only the relevant details from a huge product description. The survey is also used to support the idea of feature-driven search over the random search for user-friendly e-commerce environment. The survey was conducted in social networking site group (Facebook), using SurveyMonkey surveys and their Facebook Collector to gather data. A few of the questions used for the survey include:

1. Have you ever made any purchases online?
2. On average, how often do you buy things online?
3. Do you always find what you were looking for on the e-commerce sites?
4. Would you like to see the product description as a lengthy paragraph with all the details?
5. Do you think it is better to have the product description as a short description of only relevant points?
6. Do you prefer feature-driven search over the random search of products?
7. Would you appreciate if you have to go through a vast number of products, which comes under your conditions, while searching for the product you want?
8. Would you like to reach your desired product by selecting the product features you want, as you proceed, in a step by step manner?

The responses to questions 5, 6, 7, and 8 are given prime importance and are depicted in Fig. 43.6. From the responses of the survey, we can observe that, for

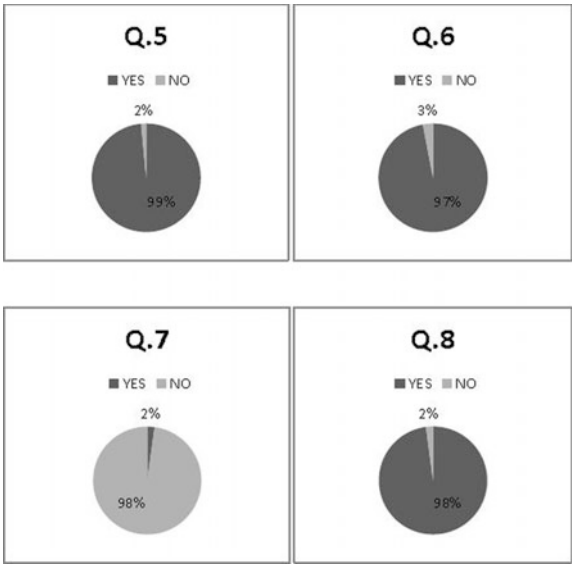


Fig. 43.6 Responses of the survey

customer satisfaction, it is important to have only a short product description that would capsulize the relevant factors of a product rather than providing them with a lengthy description and that feature-driven search is favorable than random search.

Conclusion

In order to make online shopping user-friendly, a system is proposed which will generate only the relevant features from a large set of the product description. It is not manually possible to arrange the entire product list based on the features. To achieve this we used methods, such as: Latent Dirichlet Allocation and Word2Vec. The output is then visualized by forming a collapsible tree structure. Surveys were conducted to understand the necessity of feature-driven search over the random search for user-friendly e-commerce environment and the results supported the same. In this project, we tried for getting relevant topics that covers input data of one product and tested for another product data. We applied LDA on a specific product description (ex:-watch) and topics are abstracted. In future, we are planning to consider many products. By applying text categorization on top of all products to categorize products. Then, to apply LDA to get salient topics of products.

References

1. David M. Blei, Andrew Y. Ng and Michael I. Jordan, "Latent Dirichlet Allocation" The Journal of Machine Learning Research, vol. 3, pp. 993–1022 (2003)
2. Sascha Schmitt and Ralph Bergmann, "Applying Case-Based Reasoning Technology for Product Selection and Customization in Electronic Commerce Environments", Twelfth International Bled Electronic Commerce Conference Bled (1999)
3. Chayapa Katawetawaraks and Cheng Lu Wang, "Online Shopper Behavior: Influence of Online Shopping Decision", Asian Journal of Business Research, vol1, (2011)
4. Jun Yu, Sunil Mohan, Duangmanee PutthiVidya (PEW) and Wengkeen Wong, "Latent Dirichlet Allocation Based Diversified Retrieval for E-Commerce Search", Proceedings ofthe 7th ACM International Conference on WEB Search and Data Mining, pp. 463–472, (2014)
5. Schmitt S, Bergmann R. A Formal Approach to Dialogs with Online Customers. The 14th Bled Electronic Commerce Conference. Bled, Slovenia, 2001:309–28

Chapter 44
A Bengali-Sylheti Rule-Based Dialect Translation System: Proposal and Preliminary System

Saurav Chakraborty, Anup Sinha and Sanghamitra Nath

Abstract This paper proposes a preliminary architecture for text translation between two dialectal variants of Bengali-Sylheti and Chalita Bangla. The proposed system consists of a single main module—the dialect translation module, capable of translating a given word, or group of words in one of the dialects to the other. The outcome of this study is twofold—first, from the linguistic point of view, such a system will help in understanding and analyzing the dialectal differences and second, from a social viewpoint, it will be helpful to people interested in learning/speaking/communicating in either or both the dialects. The paper takes a tour through textual analysis, construction, and morphology of words, grammar, and ambiguity and finally, proposes a methodology for such an intended translation system.

Keywords Bengali · Sylheti · Dialect · Rule-based dialect translation
POS tagging · Suffix matching

Introduction

A *Rule-Based Dialect Translation* (RBDT) system is an application that translates a word or a group of words, phrases, and/or sentences from one dialect of a language to another based on a set of rules. These rules may be related to grammar or morphology or anything else related to natural languages. This interdialect translation will be useful for people who know either dialect or wish to learn both.

S. Chakraborty · A. Sinha (✉) · S. Nath
Department of Computer Science & Engineering, Tezpur University, Tezpur, India
e-mail: anups253@gmail.com

S. Chakraborty
e-mail: sauravc2325@gmail.com

S. Nath
e-mail: s.nath@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_44

The two dialects in contention are Sylheti and the more popular Standard Colloquial Bengali (known as Chalita Bengali or চলিতভাষা). Bengali is one of the top ten largest speaking languages with more than 240 million[1] speakers all over the world. It is one of the 22 official languages of India and is the national language of Bangladesh. Sylheti (sometimes spelt *Sylhetti*, *Siloṭi*, or *Syloti*) is the main language of the Barak valley region of Assam, India, and the Kushiara and Surma valleys of Sylhet Division in Bangladesh.[2] Spoken by over 11 million people,[1] Sylheti is related to both Assamese and the rural dialects of eastern Bengal.

Though a lion's share of words were derived into Sylheti from Persian and Arabic, Sylheti also had its distinct grammar and script known as *Sylheti Nagari*. However, the gradual transition toward Bengali has significantly reduced the usage of its script and grammar and has a good share of vocabulary intertwined with Bengali.[3] As such, not many people are familiar with the script and unavailability of a good number of manuscripts or texts does not help either. Moreover, Sylheti Nagari is almost obsolete now and the script is limited mainly to linguists and researchers. Hence, the proposed system also uses the Bengali script rather than Nagari.

The scope of this paper revolves around the text-based processing of the two dialects. The paper has been organized in the following pattern. Section "Related Works" lists out certain works related to the current study. Section "Proposed RBDT System" puts forward the proposed system and its analysis. Section "Methodology" describes the methodology used for designing the proposed system. The Evaluation Framework is presented in sections "Evaluation Framework" and "Conclusion and Future Scope" follows with a conclusion and future scope of the proposed system.

Related Works

Quite a few works have been reported in the field of dialect translation and rule-based machine translation (RBMT), with a limited number of works in Indian languages and almost none in Sylheti. Tan et al. [1] propose a "Malay Dialect Translation System" which consists of three modules: the Malay Dialect Translation module, the Malay Grapheme to Phoneme module and the Malay Dialect Synthesis module. "A Machine Translation System for Standard Punjabi to Malwai Dialect" [2] has also been reported, which consists of four modules: Preprocessing of the source text, i.e., Standard Punjabi, Grammatical Translation Rules, Lexicon Lookup, and Translation of standard Punjabi text to text in the Malwai dialect. A direct Machine to Machine approach, based on word-to-word translation has been used in this work.

[1] www.wikipedia.org/wiki/Bengali_language.

[2] www.sylheti.org.uk.

[3] Chalmers (1996) reports at least 80% vocabulary overlap.

Regarding the Bengali language, a large amount of work has been done in Bengali to English or English to Bengali machine translation using different approaches. Adak [3] presents an advanced approach for "Rule Based English to Bengali Machine Translation" using Penn-Treebank parts-of-speech (PoS) tags and a Hidden Markov Model tagger. Ashrafi et al. [4] propose an "English to Bangla Machine Translation System using Context-Free Grammars", which uses the *approximate lexical meaning mapping* (ALMM) approach. Similarly, the authors in their work on a "Tense Based English to Bangla Translation" [5] use context-free-grammar for validating the syntactical structure of the input system and a bottom-up approach for parsing.

There are other works partly related to this work such as "Vaasaanubaada: Automatic Machine Translation of Bilingual Bengali-Assamese News Texts" [6] which involves bilingual texts at the sentence level. "A Common Parts-of-Speech Tagset Framework for Indian Languages" [7] covers POS tagging by employing a hierarchical and decomposable tagset schema.

Proposed RBDT System

Proposed Idea for Translation

The proposed system will take a piece of text as input in either dialect, process it according to rules and ambiguity and finally, produce the output in the other dialect. The steps can be exemplified using the following illustration (Fig. 44.1).

Methodology

Text-Based Analysis

Basic Analysis

Bengali alphabet consists of 12 vowels and 36 consonants. Besides, there are a number of conjunct consonants (যুক্তাক্ষর). The differences between the two lie mainly in the way the words are pronounced besides certain words which are totally different. As such, mere text-based processing has its limitations, because people make out the tense and/or other characteristics of the sentence from the tone of the speaker and this cannot be easily incorporated in text-based translation.

Bengali, like any other Indian language, unlike English, is a *free word-order* language. This means that interchanging word positions do not always change the meaning of the sentence. For example, *I am writing with a pen* can be written correctly as either আমিকলমদিয়ালেখিয়ার or as কলমদিয়াআমিলেখিয়ার.

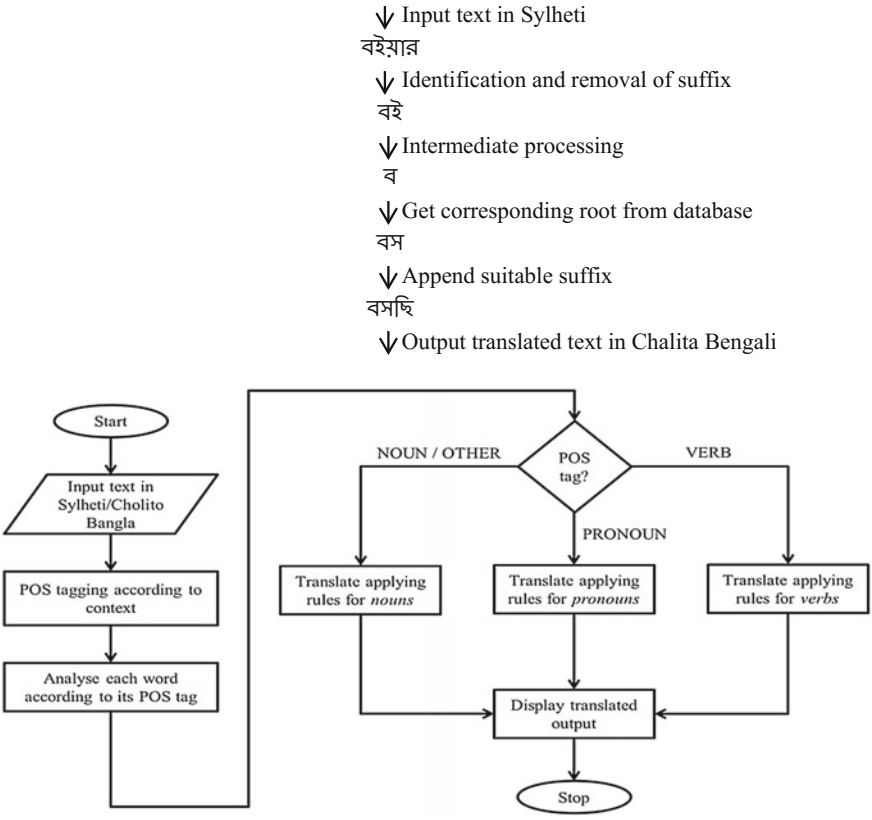


Fig. 44.1 Flowchart of the proposed RBDT system

Grammar and Parts-of-Speech (POS)

The Sylheti grammar was different from Bengali but as stated earlier, the gravitation toward Bengali has reduced such differences to a minimum. Also, the parts-of-speech list of words was made manually according to knowledge, memory, and inputs from local people. The following is a sentence exemplified with POS:

আমি	কলম	দিয়া	লেখিয়ার
Pronoun	*Noun*	*Post-position*	*Verb*

Morphology

Bengali is a morphologically rich language. The main dialectal differences between Sylheti and Chalita Bangla lies in the morphology of words including tense-forms

and parts-of-speech-based ambiguities. Thus, the main target was to formulate a set of morphological rules that pertain to both the dialects and in turn, help in translation.

For example, the sentence *I swam in the river* can be written in Sylheti as আমি নদীত্ সাতার কাট্ছি and in Chalita as আমি নদীতে সাঁতার কেটেছি,, providing evidence that there are clear distinctions in terms of morphology.

Formation of Translation Rules

For nouns and pronouns, basic rules belong to those of *case markers*. These are single letters or group of letters agglutinated at the end of a noun/pronoun to show their functional relationship in the sentence. In Bengali, cases are known as কারক while their markers are known as বিভক্তি and are always appended at the end of a noun or pronoun. Besides, another important feature is the use of marker', which denotes a shorter sound of *O*, and is peculiar to Sylheti, Assamese and some other dialects and languages, but not used in Bengali. Also, a major chunk of morphological differences lies in tense forms of verbs of both the dialects. The following tables lay out all the rules used in the proposed system (Tables 44.1 and 44.2).

Database Design

The dictionary (database), called *Lexicon*, in this implementation, can hold various root-words and certain suffixes of one dialect and their corresponding root-words

Table 44.1 List of rules for nouns and pronouns with examples

Sylheti marker		*Chalita marker*	*Usage examples*
Direct translation from DB, both noun & pronoun			আম → আম
ও,ো,'		য়ে, ে	বাজার'→বাজারে
র		র / ের	নদীর → নদীর
এ / য় / ে		*Null*/ ে	কুত্তায় → কুকুরে
রে		কে	বাঘরে → বাঘকে
◌া *vowel* +	ত	য়	রাস্তাত → রাস্তায় পানীত →জলে
Other vowel +		তে	
Otherwise +		ে	
Noun/Pronoun word ending in a pronoun-suffix must have both root and suffix in DB			ওগু→ এটা

Table 44.2 List of rules for verb tense forms with examples

SylhetiSuffix	*ChalitaSuffix*	*Example*
ই / ি	ই / ি	খাই → খাই
ও / ো	ও / ো	খাও → খাও
য় / ে	য় / ে	খায় → খায়
ইন	ন / েন	খাইন → খান
র / ের	চ্ছে / ছে	খার → খাচ্ছে
(ই / ি) + য়ার	চ্ছি / ছি	খাইয়ার → খাচ্ছি
(ই) + রায়	চ্ছ / ছ	খাইরায় → খাচ্ছ
(ই) + রা	চ্ছেন / ছেন	খাইরা → খাচ্ছেন
(ই) + রে	চ্ছিস / ছিস	খাইরে → খাচ্ছিস
(ই) + মু	বো	খাইমু → খাবো
(ই) + ব / ত	বে	খাইব → খাবে
(ই) + বা / তা	বেন	খাইবা → খাবেন
(ই) + বায় / তায়	বে	খাইবায় → খাবে
(ই) + বে / তে	বি	খাইবে → খাবি
(ই) + তাম	বো	খাইতাম → খাবো
(ই / ি) + ছ্	বি	খাইছ → খাবি
(ই / ি) + য়া	য়ে / ে	খাইয়া → খেয়ে
(ই) + ছি	য়ে / ে + ছি	খাইছি → খেয়েছি
(ই) + ছে	য়ে / ে + ছে	খাইছে → খেয়েছে
(ই) + ছ	য়ে / ে + ছ	খাইছ → খেয়েছ
লা	লেন	খাইলা →খেলেন
ল	লো	খাইল → খেলো
লাম	লাম	খাইলাম →খেলাম
লায়	লে	খাইলায় →খেলে
(ই) + ছস	য়ে / ে + ছিস	খাইছস → খেয়েছিস
(ই) + ছইন	য়ে / ে + ছেন	খাইছইন → খেয়েছেন
(ই) + ছিলাম	য়ে / ে + ছিলাম	খাইছিলাম → খেয়েছিলাম
(ই) + ছিল	য়ে / ে + ছিল	খাইছিল → খেয়েছিল
(ই) + ছিলা	য়ে / ে + ছিলেন	খাইছিলা → খেয়েছিলেন
(ই) + ছিলে	য়ে / ে + ছিলি	খাইছিলে → খেয়েছিলি
(ই) + ছিলে	য়ে / ে + ছিলি	খাইছিলায় → খেয়েছিলে
(ই) + ছস	য়ে / ে + ছিস	খাইছস → খেয়েছিস
(ই) + ছইন	য়ে / ে + ছেন	খাইছইন → খেয়েছেন

Table 44.3 Schema of the database table *Lexicon*

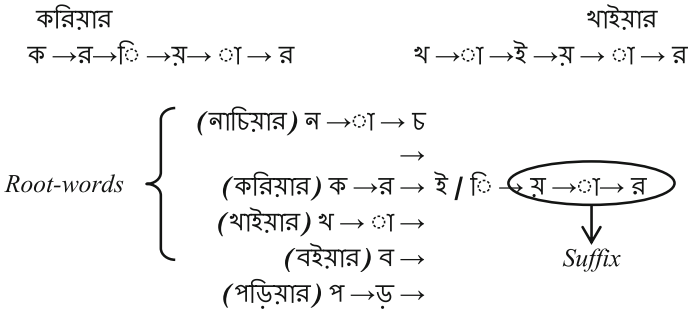
Column name	Description
sylheti_root	The root-word and/or suffix in Sylheti
Chalita_root	The root-word and/or suffix in Chalita Bengali
english_description	The meaning/short description of the word in English
POS	Parts-of-speech tag

and suffixes from the other, besides parts-of-speech tag and a short description of root-words. The following is a schema of *Lexicon* (Table 44.3).

System Implementation

Suffix Matching

The best way to deal with the morphological entity is to analyze the root-word and prefix/suffix of the same. The examples shown in the listed rules show that Bengali and Sylheti append suffixes, i.e., they both follow the rule—*different root-word + similar suffix = different forms of the same tense, case markers,* etc. The following examples illustrate this fact.



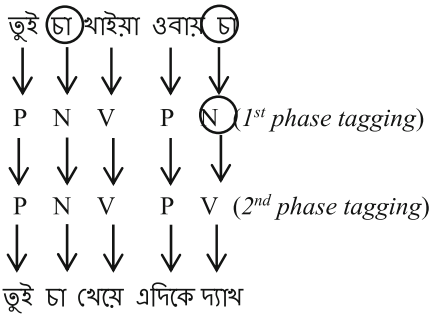
Consequently, the suffix-matching technique proved valuable and efficient. This technique gets the suffix and matches it to the list of rules to arrive at the most suitable one. The root-word of the source dialect is then extracted and its corresponding POS tag is also obtained from the lexicon, which is again exchanged with the root-word of the other dialect in the database and then the suffix is replaced with the target suffix. This is done according to the rules of nouns/pronouns (case marker rules) and/or verbs (tense rules) as obtained from the POS tagging.

Handling Ambiguity

There are certain words which are spelled the same but mean different in different circumstances. Though the human brain is trained to understand the ambiguity of words mainly by using circumstance analysis and tone of the speaker, direct implementation of the above rules in text-based processing becomes problematic and often, erroneous results are obtained. The following example shows ambiguity in the Sylheti word চা.

তুই ওবায় চা → *Used as verb here, meaning look/see*
তুই চা খা → *Used as noun here, meaning tea*

Words which cause ambiguity can be pre-identified manually and kept in a separate list or used directly through the computer program. Ambiguity is handled in the proposed system as part of a twofold POS tagging method. In the first step, the words are looked up in the database and tagged according to the basic tagging mechanism of the words. Then, if an ambiguous word is obtained, using other tags in the sentence and/or conforming to grammatical rules, it is tagged again. The following example illustrates this method:



Evaluation Framework

Although the size of the database is limited, it meets our requirements since the data has been carefully chosen to test all possible rules and various ambiguities. Testing was done by inserting single words as well as a group of words, taken from various speakers of both the dialects. Consequently, the following results were obtained (Table 44.4).

Here, the *only database* refers to the fact that the word(s) in the input text is stored in the database. *In general* refers to any text, which may or may not be contained in the database. An example of the erroneous result is given as follows:

Table 44.4 Results of evaluation

Dataset	Accuracy (%)
Words (only database)	98
Words (in general)	55
Sentences (with words from database)	80
Sentences (in general)	38.5

Input Sylheti text—আমি ছুটো থাকতে বাদাম খাইতাম
Output Chalita text—আমি ছোট থাকতে বাদাম খাবো
Correct Chalita text—আমি ছোট থাকতে বাদাম খেতাম

One reason for low accuracy in the general cases is the lack of vocabulary, i.e., the small size of the database. As this was a simulation project with limited analysis, design, and implementation time, the obtained results can be bettered with a bigger database. Also, since Sylheti is a dialect, the main emphasis is on the tone of the speaker and hence, users might insert text in the manner they speak and not according to the actual word structure, the consequence of which might be an erroneous result. In the above example, খাইতাম can be written as খাবো or খেতাম depending on the tense of the sentence, which in turn depends on the tone of the speaker.

Conclusion and Future Scope

The theme of this paper was to present the translation rules and hence, a system pertaining to text-to-text translation between Sylheti and Chalita Bengali. Translation within the limits of the database is crisp and calls for a bigger database with all possible words used in the language. However, not all rules have been analyzed in detail, because Sylheti has distinct rules as regards speech, tone, and prosody, which are different from other dialects of Bengali, including Chalita Bangla. For example, হাতি (*elephant*) is often pronounced as হাত্তি or আত্তি in Sylheti, but not written in the same way. Hence, apart from enlarging the database to incorporate more words, rules for understanding and generating speech and analysis of basic speech features and prosody of the dialects should be the next step forward.

Acknowledgements We would like to thank Mr. Wahiduzzaman Laskar, Mr. Muktar Hussain Laskar and Mr. Mahbubul Hasan Laskar and Dr. Rajib Das for their valuable advice on various aspects of this paper.

References

1. Tan, Tien-Ping, et al. "A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System." *Asian Language Processing (IALP), 2012 International Conference on*. IEEE (2012).
2. Singh, Harjeet, et al. "Machine Translation system for Standard Punjabi to Malwai Dialect." *Research Cell: An International Journal of Engineering Sciences* (2013).
3. Adak, Chandranath. "An Advanced Approach for Rule Based English to Bengali Machine Translation." *Computer Science & Information Technology* (2013).
4. Ashrafi, Kabir, et al. "English to Bangla Machine Translation System Using Context-free Grammars." *International Journal of Computer Science Issues*, *Vol. 10* (2013).
5. Muntarina, Kanija, et al. "Tense Based English to Bangla Translation Using MT System."*International Journal of Engineering Science Invention* (2013).
6. Vijayanand, Kommaluri, et al. "Vaasaanubaada: Automatic Machine Translation of Bilingual Bengali-Assamese News Texts." *Language Engineering Conference, 2002. Proceedings*. IEEE (2002).
7. Baskaran, Sankaran, et al. "A Common Parts-of-Speech Tagset Framework for Indian Languages." *In Proc. of LREC 2008*. (2008).

Part VI
Devices and Signal Processing

Chapter 45
An Approach Based on Information Theory for Selection of Systems for Efficient Recording of Electrogastrograms

Paramasivam Alagumariappan and Kamalanand Krishnamurthy

Abstract Electrogastrograms (EGG) are electrical patterns or signals which are generated by the stomach muscles and the amplitude of these signals increase after meals. These signals can be used to diagnose several digestive disorders and are recorded noninvasively using surface electrodes. In this work, a two electrode and a three electrode EGG recording system have been designed and developed for measurement of EGG signals. Further, the efficiency and performance of the developed systems are compared using tools based on the Information theory. The information content of the recorded EGG signals has been analyzed using Renyi Entropy calculated at three different α values ($\alpha = 0.2$, $\alpha = 0.5$, and $\alpha = 0.8$). Results demonstrate that the entropy of EGG signals acquired using the three electrode system is higher when compared to the signals acquired using the two electrode system. It is observed that the Information content of EGG signals acquired using three electrode system is higher when compared to the two electrode system. This work appears to be of high clinical relevance, since the accurate measurement of EGG signals without loss in its information content, is highly useful for diagnosis of digestive abnormalities.

Keywords Electrogastrograms · Two electrode and three electrode systems Renyi entropy · Information content

P. Alagumariappan (✉) · K. Krishnamurthy
Department of Instrumentation Engineering, MIT Campus, Anna University, Chennai 600044, India
e-mail: parama.ice@gmail.com

K. Krishnamurthy
e-mail: kamalanand@mitindia.edu

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_45

Introduction

The human digestive system is a highly complex system consisting of several interconnections. The process of digestion has several stages in which the food is broken into smaller parts and the nutrients are absorbed. Further, the nutrients are converted into energy for the functioning of the physiological system [1]. The disorders of the digestive system leading to various problems like bleeding, bloating, constipation, nausea, vomiting, bradygastria, and tachygastria [1, 2].

The Electrogastrograms (EGG) is the electrical signals generated by the stomach muscles. The frequency of these signals is approximately three cycles per minute (cpm) in normal individuals. Bradygastria occurs at frequencies lower than that of the normal EGG signals and Tachygastria occurs at frequencies higher than the normal EGG signals [2]. The EGG signals are acquired noninvasively by placing two or more electrodes onto the abdomen over the stomach, using surface electrodes [3]. The amplitude of the EGG signals increases after consuming food. In patients with digestive abnormalities, there are significant variations in the features of the recorded EGG signals [4]. By analyzing these electrical patterns, several digestive disorders can be diagnosed [1].

Riezzo et al. [5] have compared the electrode placement positions for the two and three electrode EGG systems. Further, the authors have presented the advantages and disadvantages of the two electrode and three electrode systems for recording EGG signals. Also, the authors have shown that a three electrode system has good signal-to-noise ratio when compared to a two electrode system. Gopu et al. [6] have designed and developed a three electrode system with Ag/Agcl electrodes for recording and analysis of electrogastrograms in cases of digestive system disorders. Yin and Chen [7] have analyzed the correlation of the electrogastrograms with gastric contraction. Further, the authors have presented the applications of electrogastrograms such as detection of digestive abnormalities and efficacy assessment of an intervention or therapy.

In recent years, tools from information theory have been utilized for analyzing the quality, information content, and uncertainty of biosignals. Further, the features based on the information theory have been utilized for classification of normal and abnormal biosignals. The entropy associated with the signal is a measure of the information content of the signal [8].

Cohen and Hudson [9] have developed new nonlinear dynamical approaches for the analysis of ECG signals. Further, the authors have discussed the extension of the nonlinear analysis to electrogastrograms. Liu et al. [10] have proposed an algorithm based on a mutual information minimization criterion to recover source signals from a mixture of various biosignals. Xie et al. [11] have introduced a new method, namely, cross fuzzy entropy for analyzing the similarity of patterns between two distinct EMG signals. Further, the authors have demonstrated that the cross fuzzy entropy of EMG decreases significantly during the development of muscle fatigue. Several researchers have utilized the entropy of biosignals for analyzing the functional conditions of physiological and pathological states [11–14].

Further, a high value of entropy indicates large information content in the acquired signal [13].

The objective of this work is to analyze the Information content of EGG signals acquired from the two electrode and three electrode recording systems.

Methodology

Acquisition of EGG Signals

In this work, a two electrode and a three electrode system were designed and developed for the acquisition of EGG signals. Standard surface electrodes were used to noninvasively collect EGG signals originating from the stomach. The signals were amplified using an instrumentation amplifier designed using operational amplifiers. Further, a microcontroller-based data acquisition system was developed for online monitoring and recording of EGG signals using a PC or Laptop, by means of serial communication. LABVIEW (V14.0.1) software was used for the acquisition of EGG signals. Figures 45.1 and 45.2 show the developed circuits for the two electrode and three electrode EGG recording systems, respectively.

The placement of surface electrodes is an important factor for proper acquisition of EGG signals [15–18]. In this study, the standard electrode placement protocol was adopted [4]. In the two electrodes system, one electrode is placed in between the xyphoid process and the umbilicus, and the second electrode is placed on the left side

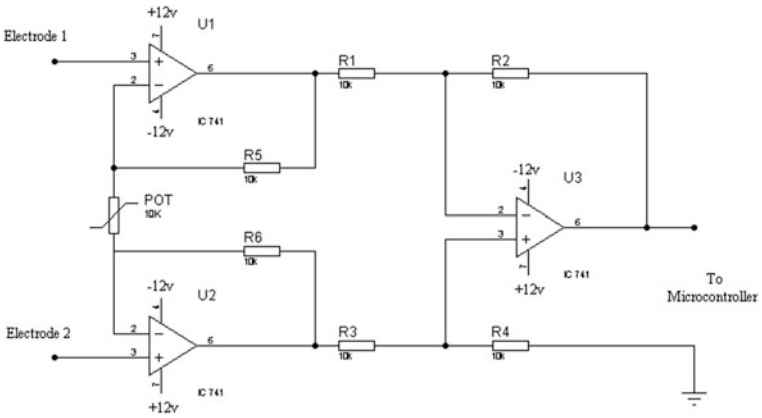


Fig. 45.1 The two electrode system for acquiring EGG signals

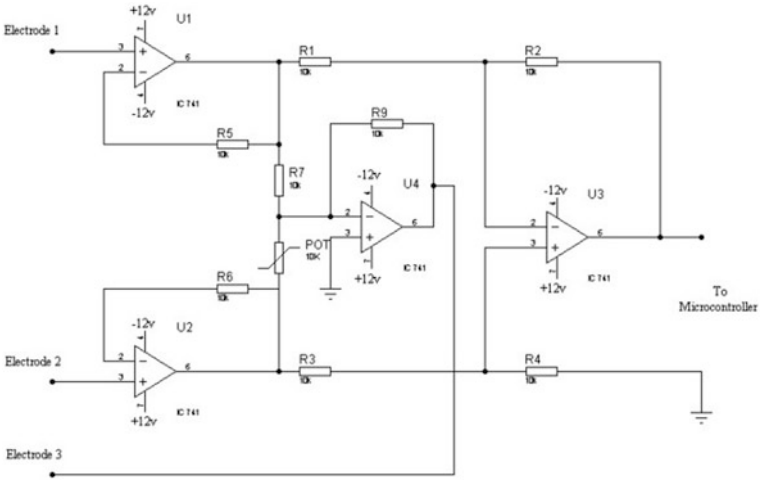


Fig. 45.2 The three electrode system for acquiring EGG signals

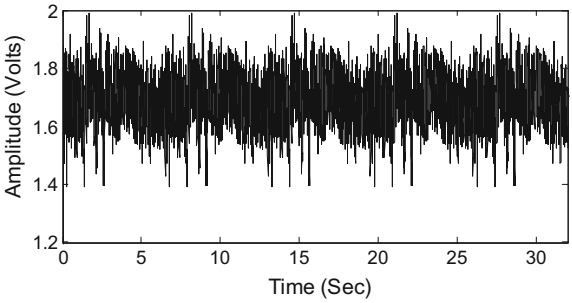


Fig. 45.3 Typical EGG signal recorded for an individual using two electrode system

of the abdomen [19]. In the three electrode system, the first and second electrodes are placed in the fundus and the mid corpus of the stomach, respectively. The third electrode is placed away from the stomach region for isolation purpose [5].

The EGG signals of six normal individuals were recorded using both the two electrode and three electrode systems for a period of 32.5 s. Figures 45.3 and 45.4 show the typical EGG signals recorded for an individual using two electrode and three electrode systems, respectively.

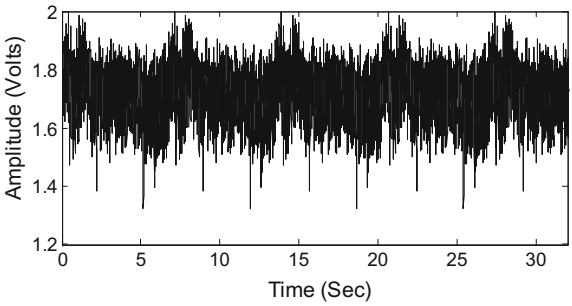


Fig. 45.4 Typical EGG signal recorded for an individual using three electrode system

Analysis of Information Content of Recorded EGG Signals Using Renyi Entropy

Entropy is the disorder associated with a system and it is correlated with the uncertainty associated with the system and hence to the information content of the system. The concept of entropy is derived from the second law of thermodynamics:

$$S = K_B \log D, \tag{45.1}$$

where, S is the entropy, K_B is the Boltzmann constant and D is the atomic disorder, which states that the entropy is a measure of the information content of a given signal [8]. There are several entropic measures such as Shannon's entropy, Tsallis entropy, Renyi entropy, and approximate entropy, etc. [20]. The Renyi entropy is a special generalization of the Shannon entropy. Renyi entropy ($H(a)$) is defined as [21]:

$$H(a) = \frac{1}{1-a} \log_2 \left(\sum_{i=1}^{n} p_i^a \right), \tag{45.2}$$

where, p_i is the probability of a random variable taking a given value out of n values, and α is the order of the entropy measure. As α increases, the measures become more sensitive to the values occurring at higher probability and less sensitive the values occurring at lower probability [21]. Renyi entropy is a measure of the complexity of the signal hence correlates with the information content of the signal [22, 23]. For $\alpha = 1$ the Renyi entropy equals the Shannon entropy. In this work, the Renyi entropy at three different α values ($\alpha = 0.2$, $\alpha = 0.5$ and $\alpha = 0.8$) was calculated for the signals acquired from the two electrode and three electrode systems. Further, the Information content of the signals was compared.

Results and Discussion

Figure 45.5 shows the Renyi entropy of EGG signals acquired from a typical subject, using the developed two electrode and three electrode systems, as a function of α. It is found that there is a significant variation in the Renyi entropy of EGG signals acquired from the two electrode and three electrode systems. It is seen that the Renyi entropy of the signals acquired using both the systems, increases with increase in α. It is observed that the difference in entropy values acquired from both two electrode and three electrode systems is less when α is low, and the difference in entropy values increase as α increases. Further, it is seen that the Renyi entropy of the signals acquired using the three electrode system is higher when compared to the two electrode system.

Figure 45.6 shows the Average Renyi entropy of EGG signals acquired from six subjects, using the developed two electrode and three electrode systems, as a function of order α. It is seen that the average entropy increases with increase in α. Further, it is found that the signals acquired using the three electrode system have higher entropy compared to the two electrode system. Hence, it is observed that the signals acquired with the three electrode system have higher information compared to the two electrode system. Also, the Renyi entropy values at three different orders

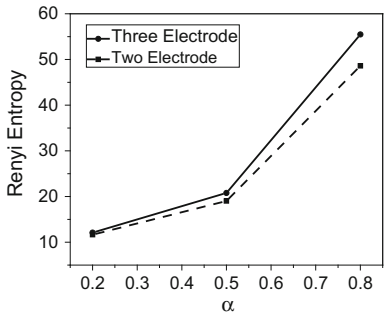


Fig. 45.5 The Renyi entropy shown as a function of α for typical EGG signals acquired using a two and three electrode systems

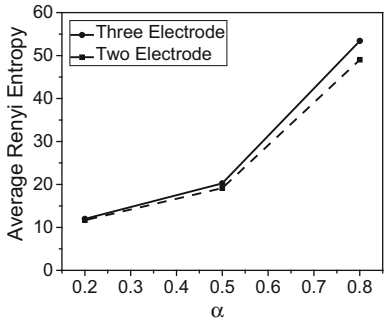


Fig. 45.6 Average Renyi entropy of EGG signals shown as a function of α acquired using two electrode and three electrode systems

Table 45.1 Renyi entropy values for different α values for signals acquired using two electrode and three electrode systems from six normal subjects

Case	Renyi entropy					
	Two electrode system			Three electrode system		
	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
1	11.65	19.01	48.60	12.10	20.78	55.48
2	11.46	18.24	45.46	11.44	18.14	44.93
3	11.77	19.52	50.57	12.11	20.79	55.55
4	12.21	21.22	57.25	12.06	20.59	54.79
5	11.33	17.78	43.70	12.04	20.53	54.48
6	11.65	19.01	48.55	12.08	20.71	55.23

($\alpha = 0.2$, $\alpha = 0.5$ and $\alpha = 0.8$) are presented in Table 45.1. It is found that, in most of the cases, the information content of the EGG signals acquired using the three electrode system is higher when compared to the two electrode system.

Conclusion

Electrogastrography is a noninvasive technique to record gastric electric potentials by means of surface electrodes placed over the stomach. It gains more importance due to its noninvasive nature and simplicity in recording when compared to any other diagnostic methods [5]. In this work, two electrode and three electrode systems for recording EGG signals have been designed and developed. The information content in the gastric electric potentials is analyzed using a tool from Information theory known as the Renyi entropy computed at three different α values ($\alpha = 0.2$, $\alpha = 0.5$, and $\alpha = 0.8$). Results demonstrate that Average Renyi entropy of EGG signals for different α values acquired using the three electrode system is higher when compared to the entropy of EGG signals acquired using the two electrode system. Hence, it is inferred that the three electrode system is more efficient in preserving the information content of the EGG signals acquired from the human digestive system. Further, the three electrode system has the best signal to noise ratio and usually, more than three electrodes are used to pick up propagation. The signals acquired from three electrode systems can be used for diagnosis of several digestive disorders, such as dyspepsia, stomach ulcer, gastric esophagus reflux disease, nausea, cyclic vomiting syndrome, etc. [6]. Also, the cost of the three electrode system is less when compared to the higher electrode systems. Further, interpretation of the EGG signals acquired using the three electrode system is less complex when compared to the analysis of EGG signals acquired from higher electrode systems. This work appears to be clinically significant since the proper recording of EGG signals without loss of information is required for efficient analysis of the physiological and pathological states of the digestive system.

References

1. Gopu, G., Neelaveni, R., Pokumaran, K., & Shekar, M. G. (2010). An Enhanced Technique for Recording and Analysis of Electrogastrogram using Active Electrodes. Sri Lanka Journal of Bio-Medical Informatics, 1(1).
2. Parkman, H. P., Hasler, W. L., Barnett, J. L., & Eaker, E. Y. 2003. Electrogastrography: a document prepared by the gastric section of the American Motility Society Clinical GI Motility Testing Task Force. Neurogastroenterology & Motility, 15(2), pp. 89–102.
3. Ravariu, C., Ursutiu, D., Babarada, F., Arhip, J., Arama, S. S., Radulian, G., & Samoila, C. 2014, February. Remote measurements of the electrical gastric signals-between theory and practice. In Remote Engineering and Virtual Instrumentation (REV), 2014 11th International Conference on (pp. 281–284). IEEE.
4. Kasicka-Jonderko, A., Jonderko, K., Krusiec-Swidergol, B., Obrok, I., & Blonska-Fajfrowska, B. 2006. Comparison of multichannel electrogastrograms obtained with the use of three different electrode types. Journal of Smooth Muscle Research, 42(2, 3), pp. 89–101.
5. Riezzo, G., Russo, F. and Indrio, F., 2013. Electrogastrography in adults and children: the strength, pitfalls, and clinical significance of the cutaneous recording of the gastric electrical activity. BioMed research international, 2013.
6. Gopu, G., Neelaveni, R. and Porkumaran, K., 2008, December. Acquisition and analysis of electrogastrogram for digestive system disorders using a novel approach. In Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on (pp. 65–69). IEEE.
7. Yin, J. and Chen, J.D., 2013. Electrogastrography: methodology, validation and applications. Journal of neurogastroenterology and motility, 19(1), pp. 5–17.
8. Kaufman, M., Zurcher, U. and Sung, P.S., 2007. Entropy of electromyography time series. Physica A: Statistical Mechanics and its Applications, 386(2), pp. 698–707.
9. Cohen, M.E. and Hudson, D.L., 2004, September. Diagnostic potential of nonlinear analysis of biosignals. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE (Vol. 2, pp. 5396–5399). IEEE.
10. Liu, J., He, Z. and Mei, L., 1998. Blind separation of biosignals by a novel ICA algorithm based on information theory. In Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE (Vol. 3, pp. 1653–1656). IEEE.
11. Xie, H.B., Zheng, Y.P. and Jing-Yi, G., 2009, September. Detection of synchrony in biosignals using cross fuzzy entropy. In 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 2971–2974). IEEE.
12. Komorowski, D. and Tkacz, E., 2015, August. A new method for attenuation of respiration artifacts in electrogastrographic (EGG) signals. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6006–6009). IEEE.
13. Cornforth, D.J., Tarvainen, M.P. and Jelinek, H.F., 2013, July. Using renyi entropy to detect early cardiac autonomic neuropathy. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5562–5565). IEEE.
14. Richman, J.S. and Moorman, J.R., 2000. Physiological time-series analysis using approximate entropy and sample entropy. American Journal of Physiology-Heart and Circulatory Physiology, 278(6), pp. H2039–H2049.
15. Chen, J.D.Z., 1998. Non-invasive measurement of gastric myoelectrical activity and its analysis and applications. In Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE (Vol. 6, pp. 2802–2807). IEEE.
16. Patterson, M., Rintala, R., Lloyd, D., Abernethy, L., Houghton, D. and Williams, J., 2001. Validation of electrode placement in neonatal electrogastrography. Digestive diseases and sciences, 46(10), pp. 2245–2249.

17. Sobrinho, Á., Perkusich, A., da Silva, L.D. and Cunha, P., 2014, July. Using Colored Petri Nets for the requirements engineering of a surface electrogastrography system. In 2014 12th IEEE International Conference on Industrial Informatics (INDIN) (pp. 221–226). IEEE.
18. Brown, B.H., Smallwood, R.H., Duthie, H.L. and Stoddard, C.J., 1975. Intestinal smooth muscle electrical potentials recorded from surface electrodes. Medical and biological engineering, 13(1), pp. 97–103.
19. Buist, M.L., Cheng, L.K., Sanders, K.M. and Pullan, A.J., 2006. Multiscale modelling of human gastric electric activity: can the electrogastrogram detect functional electrical uncoupling?. Experimental physiology, 91(2), pp. 383–390.
20. Maszczyk, T. and Duch, W., 2008, June. Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. In International Conference on Artificial Intelligence and Soft Computing (pp. 643–651). Springer Berlin Heidelberg.
21. Cornforth, D.J., Tarvainen, M.P. and Jelinek, H.F., 2014. How to calculate Renyi entropy from heart rate variability, and why it matters for detecting cardiac autonomic neuropathy. Frontiers in bioengineering and biotechnology, 2, p. 34.
22. Gonzalez Andino, S.L., Grave de Peralta Menendez, R., Thut, G., Spinelli, L., Blanke, O., Michel, C.M. and Landis, T., 2000. Measuring the complexity of time series: an application to neurophysiological signals. Human brain mapping, 11(1), pp. 46–57.
23. Bromiley, P.A., Thacker, N.A. and Bouhova-Thacker, E., 2004. Shannon entropy, Renyi entropy, and information. Statistics and Inf. Series (2004–004).

Chapter 46
Pursuit-Evasion: Multiple Pursuer Pursue Multiple Evader Using WaveFront and Hungarian Method

Ayush Mittal, Akshay Jain, Akshay Kumar and Ritu Tiwari

Abstract Traditional pursuit-evasion search algorithms usually search for single evader in a simple environment. In this paper, we have proposed an algorithm that searches for a path from multiple starting points to multiple target points (static/dynamic) in real time in real world. The world consists of a grid of cells comprising of static and dynamic obstacles. The algorithm uses available environmental information to successfully capture the evader in dynamic and complex environments. For each evader, a probability matrix is maintained which contains the probability of finding the evader at that position. These probabilities are used to predicting locations of evaders and subsequently solve the problem. Multiple pursuers coordinate among themselves to update and reduce probability matrix. The algorithm uses Wavefront algorithm for path finding and Hungarian algorithm for minimum cost assignment. The algorithm is used in various static and dynamic environments with a different number of pursuers and evaders to solve the problem.

Keywords Multiple pursuers · Multiple evaders · WaveFront Hungarian method · Static and dynamic obstacles

A. Mittal · A. Jain · A. Kumar (✉) · R. Tiwari
Robotics and Intelligent System Design Lab, Indian Institute of Information Technology and Management, Gwalior, India
e-mail: akumar15292@gmail.com

A. Mittal
e-mail: ayush2709@gmail.com

A. Jain
e-mail: akshay.jain2007@gmail.com

R. Tiwari
e-mail: tiwariritu2@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_46

Introduction

The search problem of moving evader is of utmost important in the field of robotics and computer games. In robotics, the moving evader search problem deals with the use of a pursuer to search for an evader which is moving over a particular area. Pursuer attempts to capture evader whereas evader attempts to evade from the pursuer [1].

Pursuit-Evasion is a group of problems in the field of computer science and mathematics, where one group (pursuers) attempts to track down another group (evaders) in an environment having static and dynamic obstacles. Pursuit-Evasion is relevant for many applications. For instance, the automated characters in games should show intelligence, for which intelligent evading behavior is essential. This can be used in a variety of applications, such as predator chasing a prey, cops chasing robbers, a missile chasing an aircraft, monster princess problem, hunter chasing a rabbit [2].

In Pursuit-Evasion problem, each pursuer tries to track down the evader by minimizing the distance between the pursuer and the evader, whereas the evader attempts to maximize the distance to avoid from being seized. In Pursuit-Evasion there can be different capturing states as shown in Fig. 46.1.

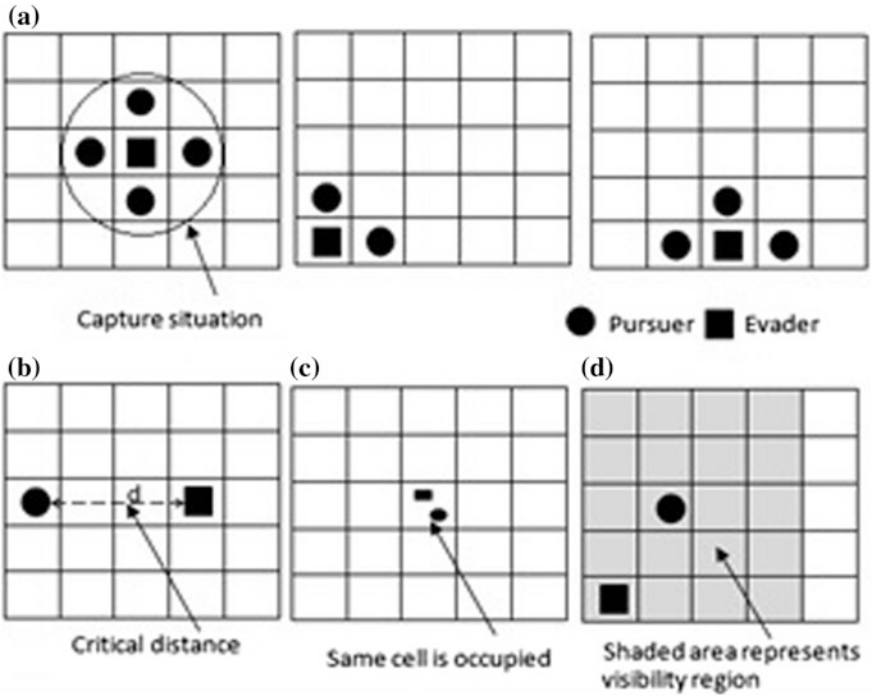


Fig. 46.1 Different capturing states

- Capturing by surrounding the evader, as shown in Fig. 46.1a.
- Capturing by approaching to the critical distance evader, as shown in Fig. 46.1b.
- Capturing by occupying the same position as an evader, as shown in Fig. 46.1c.
- Capturing by spotting evader in the visibility region, as shown in Fig. 46.1d.

This problem can be of four types:

- Single Pursuer Single Evader (SPSE)
- Single Pursuer Multiple Evader (SPME)
- Multiple Pursuer Single Evader (MPSE)
- Multiple Pursuer Multiple Evader (MPME)

Most of the research has been done for SPSE and this can be categorized into two types: single side search and both side search. In single side search, the evader is not aware of pursuer or it is busy in accomplishing its own task, whereas, in both sided search, both pursuer and evader make effort. Pursuer attempts to capture evader while evader attempts to move away from a pursuer. This is the simplest kind of pursuit-evasion. In this problem, no coordination or cooperation is needed among pursuers because pursuer is single [3].

In SPME, single pursuer attempts to capture multiple moving evaders. For example, single predator chasing multiple preys [4].

In MPSE multiple pursuers attempt to capture a single evader. A lot of work has been done on this problem. Most of the work makes various assumptions, for example, the position of evader is considered to be known all the time or the path along which evader is moving is considered known beforehand. Some of the algorithms assume that the target does not move. In our work, all these assumptions are relaxed as we calculate the positions of the evader on the basis of probability [5].

In MPME multiple moving pursuers attempt to capture multiple moving evaders [6]. Very little research work has been done on this problem. This has various real-life applications such as multiple cops chasing many robbers. We have proposed the algorithm for this case. In our work, all four cases are tested.

In MPME problem coordination among pursuers is necessary to reduce the overall time for capturing all the evaders. Pursuers should spread the information about evader's position to other pursuers. One pursuer should guide another pursuer so that they do not explore the same area while searching for moving evaders.

Related Work

The process of estimating a path from a starting point to an endpoint is called path planning. Path planning algorithms can be categorized into two types: online algorithms and offline algorithms. The algorithms that perform whole processing in advance before starting the movement of the robot are called offline algorithms. There are algorithms like Dijkstra's algorithm [7] which do not need the knowledge

of the environment. These are uninformed algorithms. Offline algorithms that use environment knowledge are called informed algorithm such as probabilistic roadmaps, random trees [8–10], A* [11, 12], genetic algorithms [13]. It is difficult to use these algorithms as they consume large time in dynamic environments. Offline algorithms can be used in dynamic environments by making them incremental; it is a technique which uses the previous searches information and finds the solution of the problem faster than solving the problem from starting, i.e., scratch. Some renowned algorithms available in the literature of such type are D* [14], Focused D* [15], D* Lite [16–18] and MT-Adaptive A* [19]. Because of efficiency problems in offline techniques, various online approaches are developed. At a time, online algorithms determine single step toward the goal. One of the oldest heuristic search algorithms available in the literature is Tangent-Bug [20]. In this algorithm, vision information is used to reach the target. Korf [21] introduced LRTA*—Learning Real Time A*, a generic heuristic search algorithm that is used for real-time path planning to fixed goals. However, most of such online search algorithms cannot be used against a moving evader since they are usually developed for fixed goals but in moving, evader search position of goal continuously changes.

MTS [22] is one of the algorithms capable of pursuing a moving evader. This algorithm is based on LRTA*. The algorithm calculates the heuristic values for each possible pair (x, y), where x is the starting point, i.e., the location of pursuer and y is the endpoint, i.e., the location of evader. It maintains a table of all these values. It is a slow algorithm because as evader moves (i.e., endpoint changes), the whole process starts again which take a huge amount of time and causes a performance bottleneck. Therefore, two new MTS are proposed which improve the solution quality of original MTS, these are called MTS-c and MTS-d.

Very recently, two real-time search algorithms, RTEF [23] and RTTES [24] were proposed for partially observable environments, on which our algorithm is built. These algorithms are developed for stationary evaders, but they can be used for moving evaders with some modifications.

The algorithms till now are applicable to only those problems which have single pursuer. The other problem comes in that is the case where there are multiple pursuers. Moving pursuers in coordination to capture a moving evader is a major challenge. Most of the work done till now for coordination is done in a simple environment that is free of static and dynamic obstacles.

In evader algorithms, usually hybrid techniques are suggested such as moving randomly in any possible direction [22, 25, 26]; move away from the pursuer in a circle or straight line [25, 27]; if evader sees a single pursuer then move in direction opposite of pursuer, if evader sees multiple pursuer then move in such a way that direction of evader forms the largest angle from the any two pursuers. Undeger [28] proposed a real-time moving target search in partially observable and dynamic environment called moving target evaluation search (MTES). MTES detects the

directions that are closed around the pursuer and estimates the best direction to capture a moving evader avoiding all the obstacles nearby. Various coordination strategies for coordinating among multiple agents are proposed. Such as keeping the robots apart from the centroid of robots.

Zheng [29] proposed the coordination strategy to coordinate multiple moving pursuers in tracking and seizing a moving evader, targeting the reduction in capture time. They assumed that the evader can be seen by any pursuer and can be captured by two or more pursuer's Mark. proposed an approach to capture single moving evader by multiple pursuers using BDI model [30].

Methodology

In this work, we have provided an algorithm using which multiple moving pursuers coordinate among themselves to capture multiple moving evaders in a dynamic environment. Pursuers use probability matrix to find the possible locations of the evaders.

Input: Map of size M X N, number of pursers and their location, number of evaders and their location, number of dynamic obstacles, and their location and locations of static obstacles.

Initially, we divide the available environment map into cells forming a square grid. All the analysis is done on this grid. These grid cells form a single entity and group of these cells will be treated as a set. After formation of cells, we repeat the steps shown in Fig. 46.2.

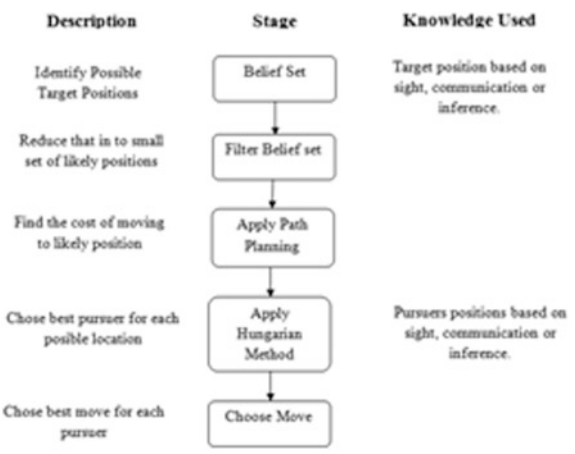


Fig. 46.2 Proposed methodology architecture

Belief Set

This considers the possible grid cells of the environment where evader can be found. These are formed on the basis of previous knowledge about the evaders' location and updated on the basis of probability. If an evader is found in a visible range by a pursuer, then its belief set contains only that grid cell and rest all cells are removed from the belief set, and the probability of that cell is made 1.

Filtered Belief Set

This considers the most likely position of the evader. After movement of evader, the probability of all grid cells is updated, the equal probability is propagated in all four directions. If there is no evader present in all the locations which are in the visible range of pursuers, then the probability of these cells is made zero. If any evader is seen by any pursuer then the probability of all cells for that evader is made zero except the location where it is seen. Locations with the highest probability are chosen from this set for allocation to pursuer for movement.

Path Planning

In this step, path planning algorithm (Wave Front) [31] is applied to know the cost of reaching the possible locations present in the filtered belief set from the current locations of the pursuers.

Hungarian Method

After getting the cost of each pursuer to reach each possible location of evader, Hungarian method [32] is used to choose the best target for each purser in terms of cost.

Choose Move

After choosing the pursuer and evader possible position best move is chosen for the pursuer to reach the evader. This move is based on the path planning algorithm. This process is repeated until all evaders are captured.

Design Details

Probability Matrix

We have used the concept of probability to find the positions of evaders. For each evader a probability matrix is made of dimensions M X N, if our map size is M X N. Initially, we know the positions of each evader so that location has probability 1 and all other locations have zero probability as shown in Fig. 46.3. In the left side, the map is shown in which X denotes evader, and in the right side, corresponding probability matrix is shown with all zeroes except one position where evader is present. These probability matrices are formed for each evader and as evader moves these are updated. A total probability matrix is formed of equal dimension which has some of all probabilities. Allocation of location for movements of pursuer is based on the total probability matrix.

Updating Probability

Every time an evader takes a step, its probability matrix is updated. We have considered equal probabilities for all four possible directions, i.e., 0.25 for one direction. At each step probability of a cell is propagated to its adjacent four cells. Summation of all incoming probability at a cell is a new probability. This is shown in Fig. 46.4.

Reduce Probability

After updating of probability matrix their reduction is done to minimize cells in belief set.

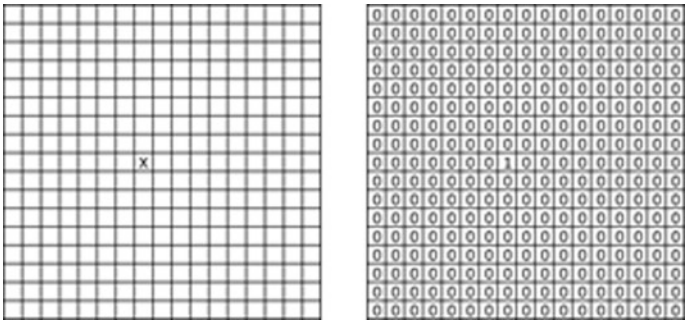


Fig. 46.3 Evader location probabilities

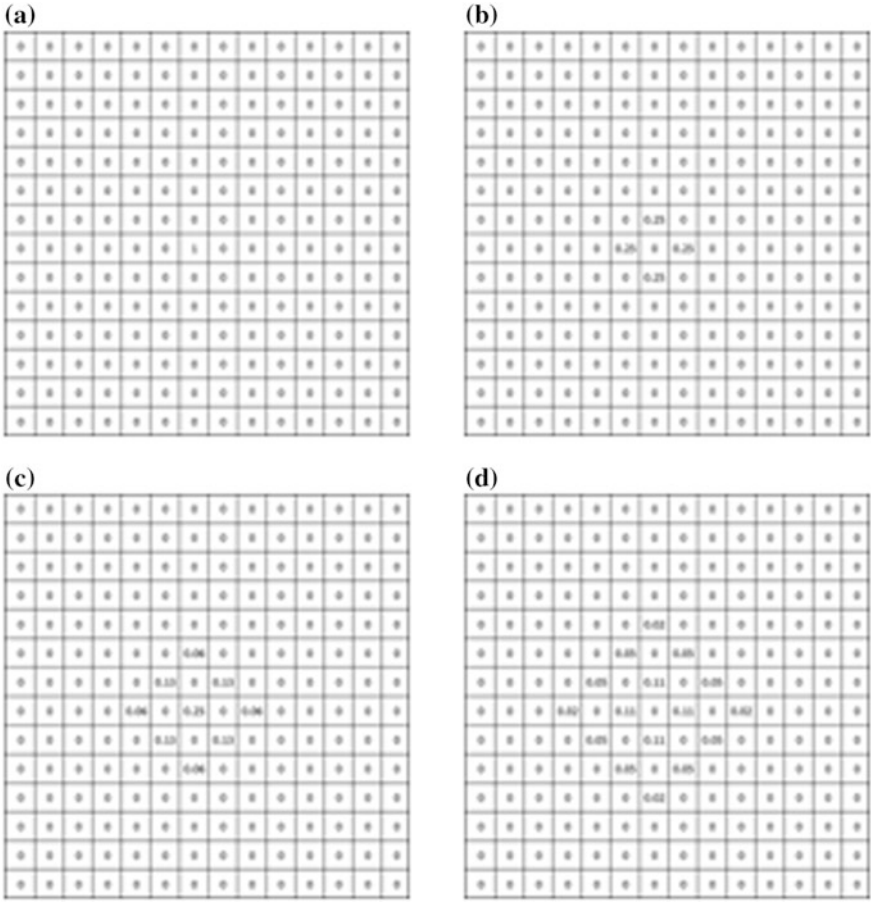


Fig. 46.4 **a** Probability matrix of evader, **b** probability matrix of evader after one step, **c** probability matrix of evader after two steps, **d** probability matrix of evader after three steps

This involves two cases: (1) when an evader is seen in a visible range of one of the pursuer then its probability matrix is made zero except that position as shown in Fig. 46.5. (2) If no evader is seen by the pursuer then the probability of all the cells which are in the visible range of the pursuer is made zero as shown in Fig. 46.6. This is done for probability matrices of all the evaders and total probability matrix is updated.

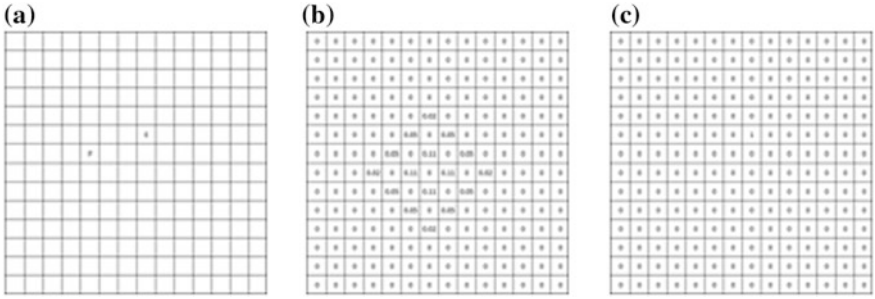


Fig. 46.5 **a** Positions of pursuer and evader, **b** probability matrix of evader, **c** probability matrix of evader after reduction

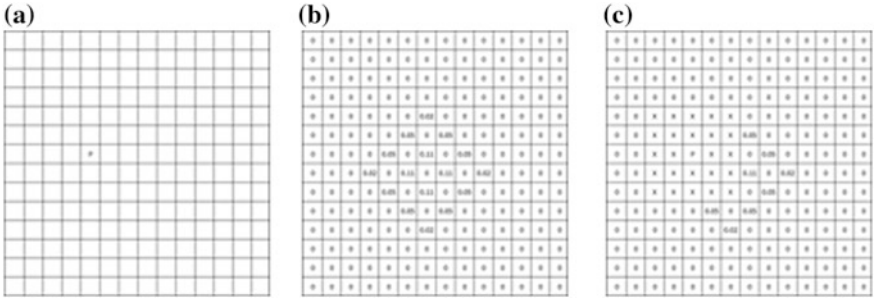


Fig. 46.6 Probability reduction **a** positions of pursue, **b** probability matrix of evader, **c** probability matrix of evader after reduction

Algorithm

Step 1: Initialize probability Matrix;
Step 2: If Evaders turn go to step 10;
Step 3: Find grid cells with the highest probability in total probability matrix and find minimum cost path from each pursuer's cell to these cells;
Step 4: Form a Square matrix in which minimum cost from each pursuer's location to location with the highest probability is stored;
Step 5: Apply Hungarian Method to find the minimum cost move for each pursuer;
Step 6: Move each pursuer;
Step 7: If all evaders captured, terminate; else go to step 2;
Step 8: Reduce probability matrices;
Step 9: Move all dynamic obstacles;
Step 10: Move all evaders and update all probability matrices and go to step 2;

Experimental Results

In this paper, we conducted experiments on Java Applet. Experiments are done for four types of map:

- Without Obstacles.
- Static Obstacles less Complex.
- Static Obstacles More Complex.
- Dynamic Obstacles.

In all four cases, evaders are fixed and numbers of pursuers are varied. It is concluded, that increasing the number of the pursuer, decreases the time elapsed in searching for an evader. That means the algorithm is working well and pursuers are coordinating. In all experiments pursuers are shown in green color, evaders are shown in red color, dynamic obstacles are shown in gray color and static obstacles are shown in black color.

Case 1: No Obstacles (Map 1)

Map with Pursuers and Evaders with no obstacles is shown in Fig. 46.7, in which each robot is supplied with range sensors.

The time taken by a number of pursuers in a no obstacle map is shown in Fig. 46.8.

Fig. 46.7 Map 1 with no obstacles

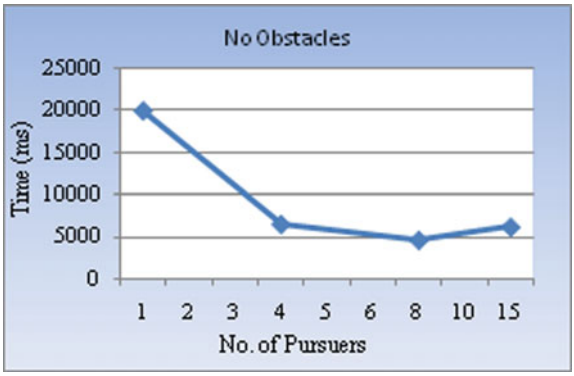


Fig. 46.8 Graph showing variation of time elapsed in capturing evaders in map 1

Case 2: Static Obstacles in Complex Environment (Map 2)

Map with Pursuers and Evaders with static obstacles is shown in Fig. 46.9, in which static obstacles are of different shapes at fixed positions. The Evaders attempts to hide behind the static obstacles thereby increasing the search time for Pursuers.

The time taken by a number of Pursuers in a static obstacle map is shown in Fig. 46.10.

Fig. 46.9 Map 2 with static obstacles

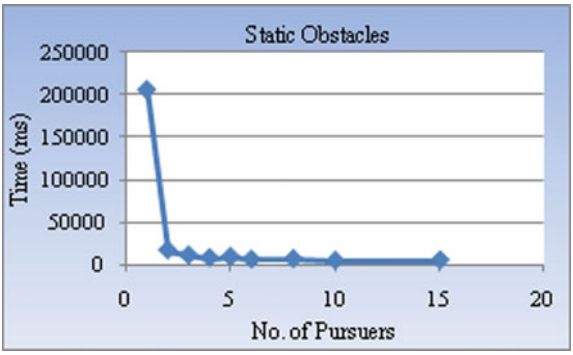


Fig. 46.10 Graph showing variation of time elapsed in capturing evaders in map 2

Case 3: Static Obstacles in Complex Environment (Map 3)

Complex Map with Pursuers and Evaders is shown in Fig. 46.11, in which complex environment with static obstacles reduces the speed of evaders' movement and thus enables it with less hidden positions to evade. This scenario decreases the search time for the pursuers.

The time taken by a number of Pursuers in a Complex map is shown in Fig. 46.12.

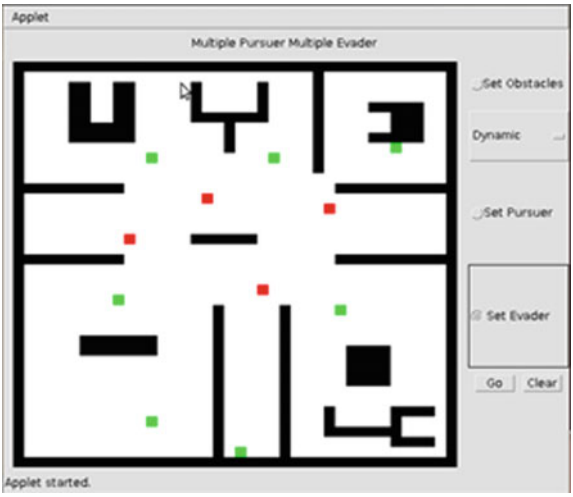


Fig. 46.11 Map 3 with static obstacles

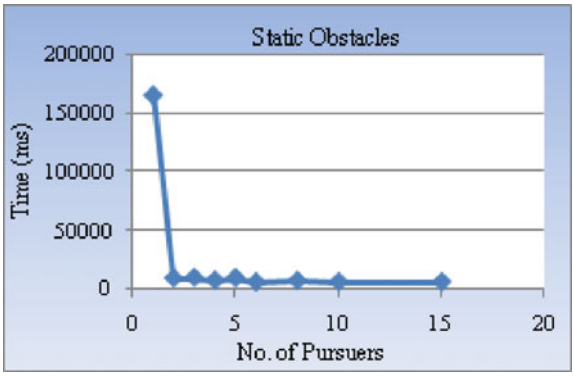


Fig. 46.12 Graph showing variation of time elapsed in capturing evaders in map 3

Case 4: Dynamic Obstacles in (Map 4)

Map with Pursuers and Evaders with static and dynamic obstacles is shown in Fig. 46.13, in which the static obstacles locations and the speed of dynamic obstacles create traffic for the Evaders and thus leading to decrease in evading speed itself. The Evaders gets trapped in obstacles, thereby, decreasing the search time.

The time taken by a number of Pursuers with Static and Dynamic Obstacles is shown in Fig. 46.14.

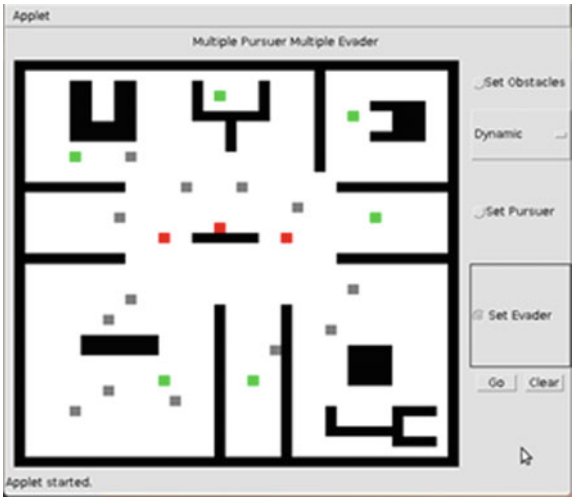


Fig. 46.13 Map 4 with dynamic obstacles

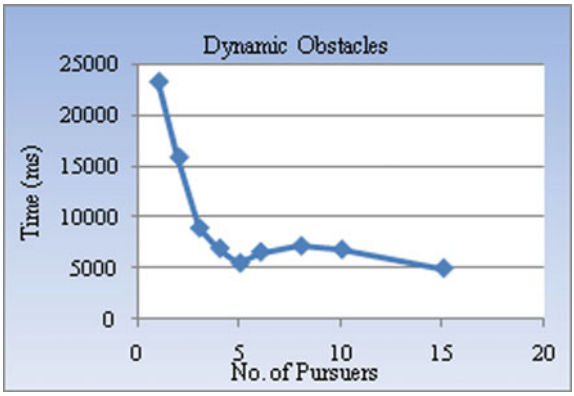


Fig. 46.14 Graph showing variation of time elapsed in capturing evaders in map 4

Conclusion and Future Scope

The problem of pursuit-evasion is of utmost important in the field of robotics and computer games and has many real-life applications. Most practical applications exhibit multiple pursuer multiple evader (MPME) type problems but not much research work has been done on this type. So in our work, we provided an algorithm based on the probability matrix using which multiple moving pursuers coordinate among themselves to capture multiple moving evaders in a dynamic environment. The algorithm is tested on Java Applets in various scenarios showing the capability of pursuers to be able to capture the moving evaders without knowing the positions of evaders beforehand. It is also concluded from the results that if numbers of pursuers are increased than time elapsed in search of moving target is reduced if the environment is kept same. The algorithm is giving positive results under varied environments and hence can be applied to solve any MPME type problem.

References

1. Oyler, D.W., Kabamba, P.T. and Girard, A.R.: Dominance in pursuit-evasion games with uncertainty. In Proceedings of IEEE Conference on Decision and Control, 5859–5864 (2015)
2. Amigoni, F. and Basilico, N.: A game theoretical approach to finding optimal strategies for pursuit evasion in grid environments. In Proceedings of IEEE International Conference on Robotics and Automation RiverCentre, 2155–2162 (2012)
3. Aleem, S.A., Nowzari, C. and Pappas, G.J.: Self-triggered pursuit of a single evader. In Proceedings of IEEE Conference on Decision and Control (CDC), 1433–1440 (2015)
4. Liu, S.Y., Zhou, Z., Tomlin, C. and Hedrick, J.K.: Evasion of a team of dubins vehicles from a hidden pursuer. In Proceedings of IEEE International Conference on Robotics and Automation, 6771–6776 (2014)
5. Festa, A. and Vinter, R.B.: A decomposition technique for pursuit evasion games with many pursuers. In Proceedings of IEEE Conference on Decision and Control, 5797–5802 (2013)

6. Petrov, N.N. and Shchelchkov, K.A.: On the "Equivalence" of Two Evasion Problems with Multiple Evaders. Journal of Computer and Systems Sciences International, Vol. 53, No. 6, 819–823, (2014)
7. Tanenbaum, A.: Computer networks. New Jersey: Prentice-Hall, (1996)
8. Cheng, P., & LaValle, S. M.: Resolution complete rapidly-exploring random trees. In Proceedings of IEEE international conference on robotics and automation, 267–272 (2002)
9. LaValle, S., & Kuffner, J.: Randomized kinodynamic planning. In Proceedings of the IEEE international conference on robotics and automation, 473–479 (1999)
10. LaValle, S. M., & Kuffner, J. J.: Rapidly-exploring random trees: Progress and prospects. Algorithmic and computational robotics: New directions, 293–308 (2001)
11. Russell, S., & Norving, P.: Artificial intelligence: A modern approach. New Jersey: Prentice Hall, (1995)
12. Gutmann, J., Fukuchi, M., & Fujita, M.: Real-time path planning for humanoid robot navigation. International joint conference on artificial intelligence IJCAI-05, 1232–1237 (2005)
13. Michalewicz, Z.: Genetic algorithms + data structure = evolution programs. New York: Springer, (1986)
14. Mudgal, A., Tovey, C., Greenberg, S., & Koenig, S., Bounds on the travel cost of a mars rover prototype search heuristic, SIAM Journal on Discrete Mathematics. 19(2), 431–447 (2005)
15. Stentz, A.: The focused D* algorithm for real-time replanning. In Proceedings of the International Joint Conference on Artificial Intelligence., 1652–1659 (1995)
16. Koenig, S., & Likhachev, M.: D* lite. In Proceedings of the national conference on artificial intelligence, 476–483 (2002)
17. Koenig, S., & Likhachev, M.: Improved fast replanning for robot navigation in unknown terrain. In Proceedings of the international conference on robotics and automation, 968–975 (2002)
18. Koenig, S. and Likhachev, M.: Fast replanning for navigation in unknown terrain. IEEE Transactions on Robotics and Automation. 21, no. 3, 354–363 (2005)
19. Koenig, S., Likhachev, M. and Sun, X.: Speeding up moving-target search. In Proceedings of sixth International Joint Conference on Autonomous Agents and Multi-Agents Systems, 188–188 (2007)
20. Kamon, I., Rivlin, E., & Rimon, E.: A new range-sensor based globally convergent navigation algorithm for mobile robots. In Proceedings of the IEEE international conference on robotics and automation. 1, 429–435 (1996)
21. Korf, R.: Real-time heuristic search. Artificial Intelligence. 42(2–3), 189–211 (1990)
22. Ishida, T. and Korf, R.E.: Moving target search: A real-time search for changing goals. IEEE Transactions on Pattern Analysis and Machine Intelligence. 17, no. 6, 97–109 (1995)
23. Undeger, C. and Polat, F.: Real-time edge follow: A real-time path search approach. IEEE Transaction on System, Man and Cybernatics, C. 37, no. 5, 860–872 (2007)
24. Undeger, C. and Polat, F.: RTTES: Real-time search in dynamic environments. Appl. I ntell. 27, 113–129 (2007)
25. Ishiwaka, Y.,Sato, T. and Kakazu, Y.: An approach to the pursuit problem on a heterogeneous multiagent system using reinforcement learning. Elsevier J. Robotics and Autonomous System. 43, no. 4, 245–256 (2003)
26. Goldenberg, M., Kovarsky, A.,Wu X. and Schaeffer, J.: Multiple agents moving target search. In Proceedings of International Joint Conference on Artificial Intelligence, 1536–1538 (2003)
27. Kota, R., Braynov, S. and Llinas, J.: Multi-agent moving target search in a hazy environmen. In Proceedings of IEEE International Conference on Integr. Knowledge on Intensive Multi-Agent System, 275–278 (2003)
28. Undeger, C. and Polat, F.: Real-Time Moving Target Evaluation Search. In Proceedings of IEEE Transactions On Systems, Man and Cybernetics—Part C: Applications and Reviews. 39, No. 3, 366–372 (2009)

29. Zheng, J., Yu, H., Zheng, M., Liang, W. and Zeng, P.: Coordination of multiple mobile robots with limited communication range in pursuit of single mobile target in cluttered environment. in J Control Theory Appl, 441–446 (2010)
30. Al, A. and Hoogendoorn, M.: Moving Target Search Using Theory of Mind. In IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 66–71 (2011)
31. Pal, A., Tiwari, R. and Shukla, A.: A focused wave front algorithm for mobile robot path planning. In International Conference on Hybrid Artificial Intelligence Systems, 191–197 (2011)
32. Date, K. and Nagi, R.: GPU-accelerated Hungarian algorithms for the Linear Assignment Problem. Elsevier J. Parallel Computing. 57, 52–72 (2016)

Chapter 47
An Efficient Multiscale Wavelet Local Binary Pattern for Biomedical Image Retrieval

Vijay Kumar Nath, Rakcinpha Hatibaruah and Deepika Hazarika

Abstract A method for biomedical image retrieval using multiscale wavelet local binary pattern (LBP) is presented in this paper. The method first decomposes a biomedical image into approximation and oriented detail subbands using discrete wavelet transform (DWT). Since the oriented detail subbands at each scale exhibit distinct directional features the proposed method employ a new 4-point LBP with selected non-diagonal neighbors in horizontal and vertical subbands, and a 4-point LBP with selected diagonal neighbors in diagonal subband to extract the LBP histogram. An 8-point LBP is employed in approximation subband to extract the LBP histogram. The biomedical image is finally represented by a single feature histogram that is formed by concatenation of all the LBP histograms. The proposed method provides significantly reduced feature vector size while maintaining same or most of the times better retrieval efficiency compared to original LBP and other relevant wavelet-based LBP schemes. The Euclidean distance measure is used for query matching and retrieval is performed based on the least matching distance. The method is tested using OSIRIX image data sets and experimental results validating the efficiency of the proposed method over other relevant schemes, are presented.

Keywords Discrete wavelet transform · Local binary pattern · Feature extraction · Similarity measurement · Biomedical image retrieval

V. K. Nath (✉) · R. Hatibaruah · D. Hazarika
Department of Electronics and Communication Engineering,
School of Engineering, Tezpur University, Assam, India
e-mail: vknath@tezu.ernet.in

R. Hatibaruah
e-mail: rakcinp@tezu.ernet.in

D. Hazarika
e-mail: deepika@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_47

Introduction

Biomedical image retrieval systems are gaining a lot of importance among the researchers for its widespread applicability and utility in patient care, diagnosis, and treatment. Biomedical images are considered as the best possible tool for in vivo analysis of human body organs. The efficient handling of thousands of images produced daily for diagnosis of patients is quite difficult. For treatment and clinical purpose, the doctors or the operating personal require them in organized and indexed form. Biomedical image retrieval systems are designed to fulfill this purpose, as on a particular query they present a set of similar type of images. The highly erroneous and scrupulous text-based retrieval systems are presently replaced by content-based image retrieval systems (CBIR) which are highly effective and swift. The efficiency of a CBIR system is specifically dependent on two aspects: (1) the feature extraction method and (2) the technique used for similarity measure.

Local binary pattern (LBP) operator was first introduced by Ojala et al. [4] as an efficient tool for texture description and is found to be highly effective in facial expression analysis, background modeling, image retrieval, and face recognition. Commercial CBIR systems like Flexible image database system (FIDS) and ImageScape are among the first ones to use LBP for retrieval purpose [9]. Takala et al. [7] proposed a block-based LBP scheme for image retrieval which gained a lot of attention among the researchers.

In the last few years, there has been good interest in multiresolution LBP with most of its application still limited to face recognition [1, 3, 5, 6, 8, 10, 11]. By varying the sampling radius the LBP's of various resolutions were obtained [5]. In [11], the LBP operator with a fixed radius is employed after the downsampling of the original image. In multiscale approaches, the image is first decomposed into various resolutions and then LBP is applied on each resolution to extract distinct features. Wang et al. [10] proposed a pyramid-based LBP method where they used a Gaussian filter for decomposition of the image into different resolutions. Liu et al. [3] used wavelet decomposition for the same purpose. However, the drawback in these methods lies in the fact that they do not consider the high frequency subbands which leads to loss of useful information. Tang et al. [8] eliminated this drawback by employing LBP on each subband of Haar wavelet decomposed image for face recognition. Yi Ding et al. [11] used the same concept for a hand vein recognition scheme, where they decomposed the image up to two levels and then employed LBP on each sub band excluding the diagonal subbands. The problem of information loss was minimized to some extent but at the cost of increased computational complexity as the feature vector size was increased due to the consideration of the high frequency sub bands.

Ojala et al. [5] observed that there are certain patterns in LBP that can be termed as uniform. By incorporating these uniform patterns (LBP_8^{u2}) they eliminated some of the redundant features and reduced the LBP feature vector size for a single image to 59 bins from 256 bins. The original 8-point LBP (LBP_8) is usually effective for spatial domain images, so employing the same method for both LL and non-LL sub band

will undoubtedly carry useful as well as redundant information. Although [5] and [6] suggested two different ways of reducing redundant features, we observe that the directional features in each oriented subband in these two methods are not exploited properly. Rashid et al. [6] reduced the feature vector size by applying uniform LBP (LBP_8^{u2}) for encoding the approximation subband and four neighbors LBP (LBP_4) for coding the detail subbands. Through the use of four neighbors LBP (LBP_4), each detail or high-frequency subband was represented with 32 bins instead of 59 bins thus reducing the feature vector size up to 46%.

In this paper for biomedical image retrieval, we propose an efficient feature extraction method using LBP and DWT which provides significantly reduced feature vector size and better retrieval performance than Rashid et al's method [6].

The organization of the rest of the paper is as follows. Section "Multiscale LBP and Discrete Wavelet Transform" presents a brief discussion about the use of LBP in multiscale approach. Section "Proposed Method" discusses the proposed method for image retrieval and section "Experimental Results and Discussions" presents the detailed analysis of experimental observations. At last conclusion of the work is presented in section "Conclusion".

Multiscale LBP and Discrete Wavelet Transform

An LBP operator [4] basically encodes a pixel element with an 8-bit binary pattern and replaces the element with its corresponding decimal value in the pattern image. LBP value for a given pixel is computed by comparing its gray value with all the neighbors in its neighborhood and is given by:

$$LBP_{P,R} = \sum_{i=1}^{P} 2^{P-1}(f(I(g_i)) - I(g_c)) \tag{47.1}$$

$$f(x) = \begin{cases} 1, \text{ if } x \geq 0 \\ 0, \text{ otherwise} \end{cases} \tag{47.2}$$

where,
$I(g_c)$ = gray value of the center pixel
$I(g_i)$ = gray value of the ith neighbor
P = number of neighbors

The whole image is represented by constructing a histogram, from the obtained LBP.

$$H_{LBP}(l) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} f_2(LBP(i,j), l) \tag{47.3}$$

$$f_2(x, y) = \begin{cases} 1, \text{ if } x = y \\ 0, \text{ otherwise} \end{cases} \tag{47.4}$$

Using this 8-point LBP operator each image can be represented with 256 (2^8) bins, which results in a large feature vector for the operation. As there may be many redundant features, Ojala et al. in [5] introduced the concept of uniform patterns (LBP_8^{u2}) which is only a subset of earlier 256 features. According to them, if a local binary pattern contains maximum two-bit wise transitions from 0 to 1 or vice versa then that pattern is called uniform binary pattern, e.g.,10000001 (2 transitions), 11111110 (1 transitions), 00000000 (0 transitions) are uniform, while 10000101 (4 transitions), 01111001 (3 transitions) are not. There are only 58 uniform patterns among 256 possible patterns in a 3 × 3 neighborhood and considering all other nonuniform patterns in a single bin, the feature vector size is reduced to 59 bins. This is a huge improvement in feature vector size optimization with almost 77% reduction in feature vector size without any significant degradation in efficiency.

In a multiscale approach, an image is first decomposed into various resolutions using a transformation method. The target pattern is then applied on each image separately to extract distinct features from them [2]. The final feature vector is constructed by concatenating all the features together.

DWT decomposes an image at different scales into low-and high-frequency subbands. The low-frequency approximation subband (LL) is a low-resolution approximation of the original image and the detail subbands contain features of the image in horizontal, vertical and diagonal directions. The LBP can be applied on each subbands to extract the features. It was demonstrated that the redundant features can be well reduced if approximation and detail subbands are treated with different LBP codes [6]. In [6], Rashid et al. applied the LBP_8^{u2} on approximation (LL) subband to extract the features for face recognition. For each detail, subbands they calculated two separate LBP codes, where first LBP code is computed based on the four non-diagonal main neighbors and second LBP code is based on the four diagonal neighbors. The final histogram of each detail subband was calculated based on the concatenation of histograms of above two LBP sets. It should be noted that that all the three detail subbands were treated with the same strategy for feature extraction [6].

Proposed Method

Although the work discussed in [6] has provided a good reduction in feature vector size, still the authors have not considered the directional features of each oriented subband properly. This paper proposes a method of extracting features more efficiently than [6] by considering the directional features of oriented subbands which results in further decrease of redundant features.

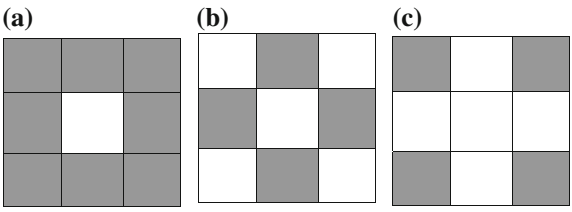


Fig. 47.1 **a** 8 neighbors **b** 4 non-diagonal neighbors **c** 4 diagonal neighbors

We first decompose the biomedical image into approximation and detail subbands, using Haar wavelet. Each oriented detail subband usually shows different anisotropic feature. We propose to use the non-diagonal neighbors $LBP_4^{h,v}$ in horizontal and vertical subbands (Fig. 47.1b) and diagonal neighbors LBP_4^d in diagonal subband (Fig. 47.1c) to extract the features. The use of $LBP_4^{h,v}$ in horizontal and vertical subbands, and LBP_4^d in diagonal subband captures the directional information very efficiently.

As a result only 16 bins is required to represent each of these three subbands instead of previously required 32 bins [6]. Approximation subband (LL) can be best encoded with $LBP8_u^2$. The single feature histogram is formed by computing the histogram of each subband image separately and are finally concatenated.

Experimental Results and Discussions

To analyze the retrieval performance, we have used a popular medical image database namely OSIRIX database, which contains various images of human body organs taken at various instants of interest. The database chosen in this experiment contains 271 computerized tomography (CT) thoracic aorta images of size 512 × 512 which were acquired on a 64 detector scanner. In the preprocessing part for measuring efficiency we classified them into four categories, each of which indicates the various status of a spiral aortic dissection after surgical repair of ascending aorta, containing 74, 67, 69, and 61 images, respectively. One image from each category is shown in Fig. 47.2. For query matching, we use Euclidian distance measure which is defined as

$$D(Q, DB_j) = \sum_{i=1}^{L} ((f_{DB_{ji}} - f_{Q_i})^2)^{\frac{1}{2}} \tag{47.5}$$

where,
f_{Q_i} = ith feature of the query image feature vector
$f_{DB_{ji}}$ = ith feature of the jth image in the database

$D(Q, DB_j)$ = Matching distance of the jth image in the database from the query image

L = Length of the feature vector

We have considered each image in the database as query image once and against each of them 'n' number of relevant images are retrieved from the database. Then, we have checked whether the retrieved images for each query, belongs to the same category (Fig. 47.2) or not. Two standard parameters namely average retrieval precision (ARP) and average retrieval recall (ARR) are used to measure the retrieval efficiency which are defined as follows.

$$ARP = \frac{1}{L}\sum_{i=1}^{L} P(I_i) \tag{47.6}$$

and

$$ARR = \frac{1}{L}\sum_{i=1}^{L} R(I_i) \tag{47.7}$$

where $P(I_i)$ = (Number of relevant images retrieved)/(Total number of images retrieved), $R(I_i)$ = (Number of relevant images retrieved)/(Total number of relevant images in the database) and L = Length of the feature vector.

We have implemented spatial domain LBP_8, wavelet domain LBP_8, wavelet domain LBP_8^{u2}, Rashid et al.'s [6] method and proposed method in Matlab environment and their retrieval performances are compared in terms of ARP and ARR using the selected database. The retrieval performance comparison of various methods and proposed method are shown in Figs. 47.3 and 47.4.

The feature vector size comparison of various methods including proposed method is provided in Table 47.1.

Figures 47.3 and 47.4 and Table 47.1 clearly reveals that the proposed method has smallest feature vector size and much less (about 30%) than that of [6] while maintaining similar or even better retrieval efficiency at many top matches. The retrieval efficiency degrades slightly when compared to wavelet LBP_8 method, however, the

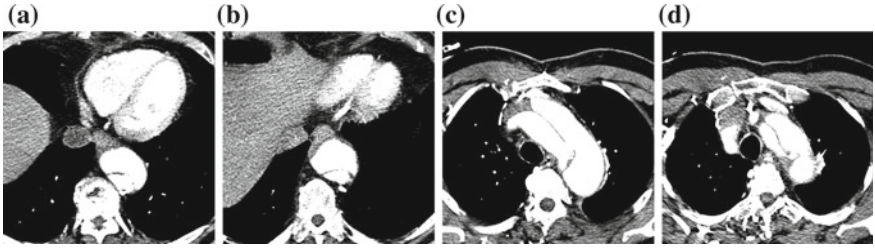


Fig. 47.2 One image from each category

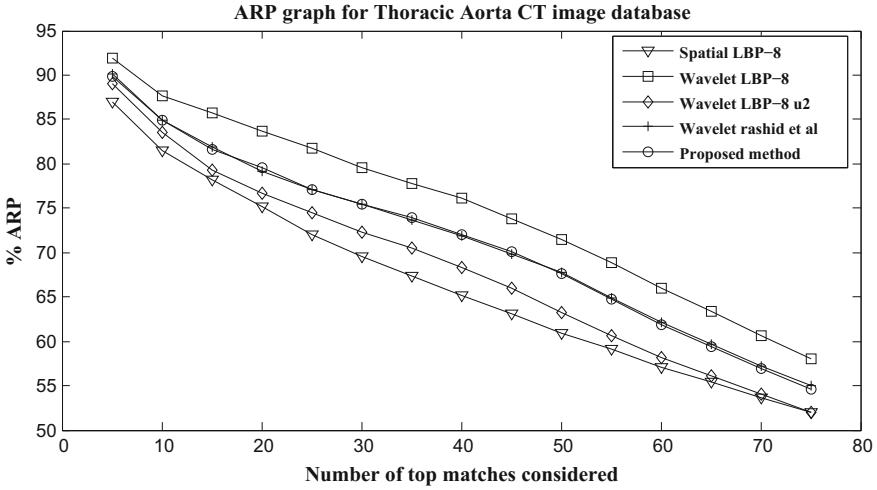


Fig. 47.3 ARP comparison graph for the different top matches. X axis shows number of top matches considered and Y axis shows % ARP for respective methods

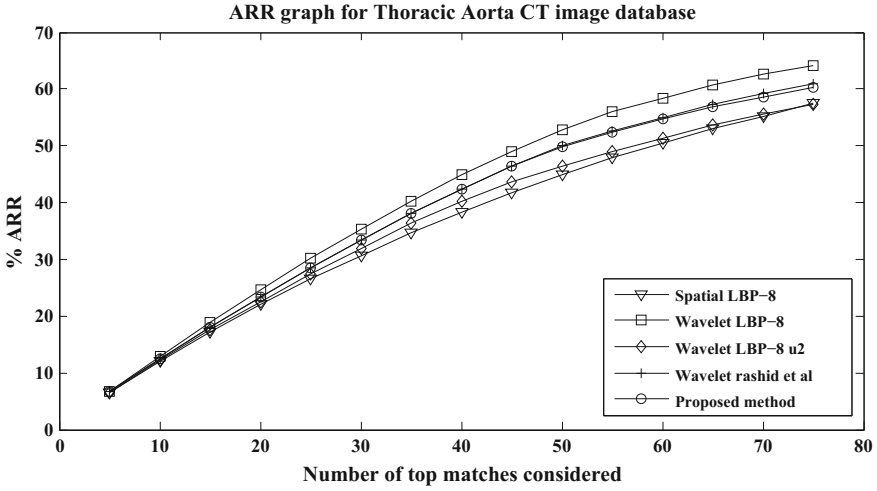


Fig. 47.4 ARR comparison graph for the different top matches. X axis shows number of top matches considered and Y axis shows % ARR for respective methods

Table 47.1 Comparison of feature vector sizes

Method	Feature vector size
Spatial LBP_8	256
Wavelet LBP_8	1024
Wavelet LBP_8^{u2}	236
Wavelet LBP [6]	155
Proposed method	107

reduction of feature vector size, in our case, is also extremely high (almost 90%). The proposed method, for all the top matches outperforms spatial domain LBP_8 and wavelet LBP_8^{u2}.

Conclusion

In this article, a wavelet LBP-based approach with significantly reduced feature vector size is presented for retrieval of CT class of biomedical images. The proposed method takes the full advantage of directional information present in each detail subband, for extracting the features. The approximation subband and each high-frequency detail subbands are treated differently for feature extraction. The use of LBP_8^{u2} in LL, $LBP_4^{h,v}$ in horizontal and vertical subbands and LBP_4^d in diagonal subband provides the best retrieval results with 30% reduction in feature vector size than one of the very recent related method.

References

1. Adelson, E.H., Anderson, C.H., Bergen, J.R., Burt, P.J., Ogden, J.M.: Pyramid methods in image processing. RCA engineer 29(6), 33–41 (1984)
2. González, E., Bianconi, F., Fernández, A.: An investigation on the use of local multi-resolution patterns for image classification. Information Sciences 361, 1–13 (2016)
3. Liu, X., Du, M., Jin, L.: Face features extraction based on multi-scale lbp. In: Signal Processing Systems (ICSPS), 2010 2nd International Conference on. vol. 2, pp. V2–438. IEEE (2010)
4. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern recognition 29(1), 51–59 (1996)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence 24(7), 971–987 (2002)
6. Rashid, R.D., Jassim, S.A., Sellahewa, H.: Lbp based on multi wavelet sub-bands feature extraction used for face recognition. In: 2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2013)
7. Takala, V., Ahonen, T., Pietikäinen, M.: Block-based methods for image retrieval using local binary patterns. In: Scandinavian Conference on Image Analysis. pp. 882–891. Springer (2005)

8. Tang, H., Sun, Y., Yin, B., Ge, Y.: Face recognition based on haar lbp histogram. In: 2010 3rd international conference on advanced computer theory and engineering (ICACTE). vol. 6, pp. V6–235. IEEE (2010)
9. Veltkamp, R.C., Tanase, M.: Content-based image retrieval systems: A survey (2001)
10. Wang, W., Chen, W., Xu, D.: Pyramid-based multi-scale lbp features for face recognition. In: Multimedia and Signal Processing (CMSP), 2011 International Conference on. vol. 1, pp. 151–155. IEEE (2011)
11. Wang, Y.D., Yan, Q.Y., Li, K.F.: Hand vein recognition based on multi-scale lbp and wavelet. In: 2011 International Conference on Wavelet Analysis and Pattern Recognition. pp. 214–218. IEEE (2011)

Chapter 48
Coupling Characteristic of Silicon-Based Optical Directional Coupler

Himanshu R. Das, Isac Daimary and Subhash C. Arya

Abstract This paper demonstrates an optical directional coupler for communication purpose. The light propagating in one waveguide couples with the other thereby transferring energy. The core has a refractive index of 3.5 and the cladding has a refractive index of 1.5. When the length of the waveguide is changed from 670 to 820 μm, their behavior is observed in terms of coupling length and electric field.

Keywords Optical directional coupler · Refractive index · Coupling length Electric field

Introduction

A directional coupler is a device which couples the energy traveling in a waveguide to the other in a particular direction. The concept behind this mechanism is the couple mode theory. It has a simple structure and is widely used in the field of photonic devices. Directional couplers are nowadays used in designing plasmonic-based optical modulators [1]. Different types of plasmonic materials are being used like graphene and indium tin oxide [2]. Also, it can be used in power dividers, switches, filters, optical interleavers, multiplexers, and demultiplexers [3]. However, conventional directional couplers can be III-IV semiconductor group-based and lithium

H. R. Das (✉) · I. Daimary · S. C. Arya
Lightwave Communication Systems Lab, Department of Electronics and Communication Engg, School of Technology, North-Eastern Hill University, Shillong, Meghalaya, India
e-mail: das.himanshu69@gmail.com

I. Daimary
e-mail: iscdaimary@gmail.com

S. C. Arya
e-mail: aryasubh@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_48

niobate-based depending upon their operation and function. An optical directional coupler made of these waveguides typically has a size of the order of few millimeters to centimeters because of the large separation between the waveguides and hence, have a long coupling length. This directional coupler has been designed and analyzed for electro-optical modulation and its compatibility with CMOS technology [4, 5]. The silicon-based directional coupler is attractive structure to use in optical interconnection in Si-chips. Silicon-on-insulator can be used as a platform for integration of such optoelectronic devices as it is of low cost and compatible with mature CMOS technology [6, 7]. In this paper, we demonstrate an optical directional coupler of the submicron level and observe its behavior in terms of electric field and frequency. We also discuss the structural parameters of the optical waveguide.

Structure of the Directional Coupler

In a directional coupler light traveling through one waveguide couples with the other thereby transferring energy. When both the waveguides are kept in a close proximity, coupling phenomena occurs. So, the gap between the waveguides is an important parameter. Here, 1 μm is taken as the gap between the waveguide cores. Width, length, and height of the waveguide also play a significant role in determining the effectiveness of a directional coupler. The 3D view of the optical directional coupler is shown in Fig. 48.1. The values of different parameters like the width, length, height, and gap between the waveguides are given in Table 48.1.

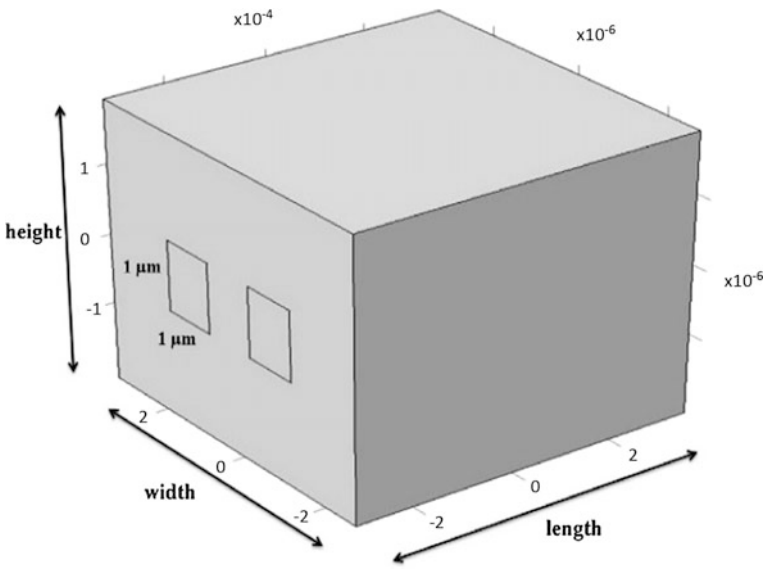


Fig. 48.1 Structure of the optical directional coupler

Table 48.1 Structural parameters of the optical directional coupler based on the silicon waveguides

Parameter	Value
Length of the waveguide	670, 820 μm
Width of the waveguide	6 μm
Height of the waveguide	4 μm
Gap between the waveguide cores	1 μm
Refractive index of core	3.5
Refractive index of cladding	1.5
Dimension of the waveguide core	1 × 1 μm

Silicon is used as the core material having a refractive index of 3.5 and the cladding is made up of silica having a refractive index of 1.5. Due to this difference in the refractive index of core and cladding, light is strongly confined inside the waveguide core. The dimension of both the waveguide core is 1 × 1 μm.

Results and Discussion

The characteristic behavior of the optical directional coupler is observed using a finite element method solver. Figure 48.2 shows the symmetric and asymmetric mode of the directional coupler when a light of 1.55 μm wavelength is launched from the lower port of the left side waveguide core. The effective refractive index for the symmetric and asymmetric mode is 2.9577 and 2.9502. The red spot indicates the presence of an electric field in the waveguide having length 670 μm. Also, Fig. 48.3 shows the coupling phenomena in the optical directional coupler when the waveguide length is 670 and 820 μm. It can be clearly seen that light

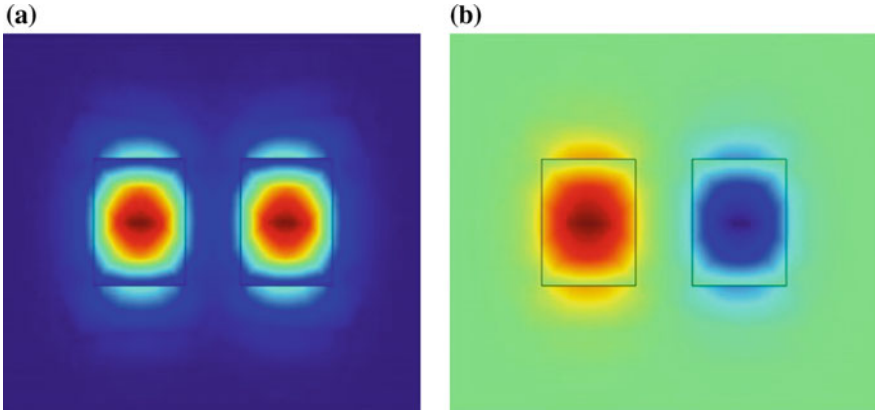


Fig. 48.2 Structure showing the **a** symmetric tangential boundary mode electric field, z component (V/m) **b** asymmetric tangential boundary mode electric field, z component (V/m)

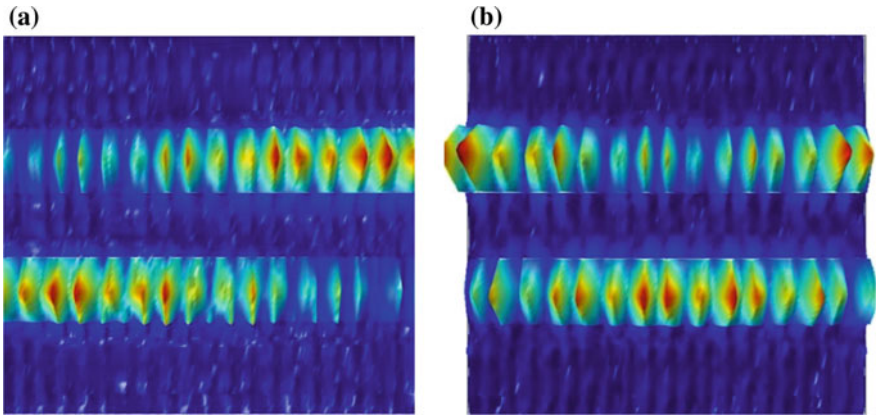


Fig. 48.3 Showing the coupling phenomena in the optical directional coupler **a** when the waveguide length is 670 μm **b** when the waveguide length is 820 μm

couples from one waveguide to the other in both the cases but in Fig. 48.3b the light couples onto the second waveguide and comes out from the first waveguide due to increase in waveguide length. The main concept behind this mechanism is coupling mode theory. The total electric field is calculated as the sum of electric fields of the two modes:-

$$E = E_1 \exp(-j\beta_1 x) + E_2 \exp(-j\beta_2 x) \tag{48.1}$$

$$\varepsilon_r = (n - ik)^2 \tag{48.2}$$

The electric field and the refractive index of the directional coupler can be easily found out from the given Eqs. 48.1 and 48.2. E_1 and E_2 are the electric fields in both the waveguides, β_1 and β_2 are the propagation constant, n and k are the refractive index of the real part and imaginary part. The graphs are shown in Fig. 48.4a and b give the details analysis of electric field distribution over the wavelength (1.33–1.55 μm) in both the waveguides having 670 and 820 μm waveguide length. Also, from Fig. 48.3a and b it can be seen that the energy couples from one waveguide to the other. Both the proposed models work well in the frequency range between 1.94×10^{14} Hz to 2.24×10^{14} Hz. The coupling length of the waveguide with 670 μm length was observed to be 640 μm and the coupling length for the waveguide with 820 μm was observed to be 400 μm. Also, the waveguide with length 670 μm shows a maximum electric field around 1.40 μm wavelength and waveguide with 820 μm shows a maximum electric field around 1.35 μm wavelength.

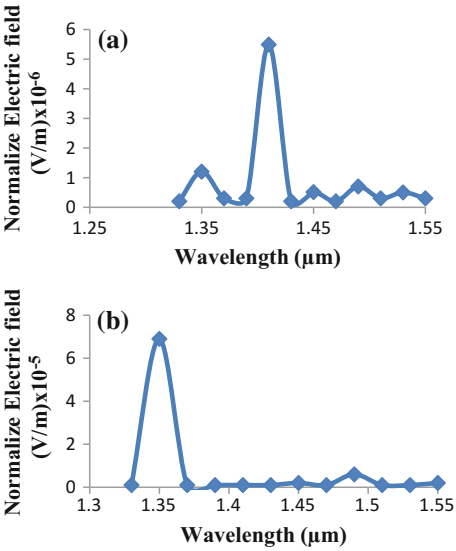


Fig. 48.4 The graph above shows the **a** electric field versus wavelength curve for waveguide having length 670 µm, **b** shows the electric field versus wavelength curve for 820 µm length waveguide

Conclusion

We have demonstrated an optical directional coupler based on the silicon waveguide. The symmetric and asymmetric mode of the directional coupler has been shown along with the coupling phenomena of both the waveguide with 670 and 820 µm waveguide lengths. The electric field distribution in both the waveguide along the wavelength is shown in Fig. 48.4a and b. Also, by varying the structural parameters the maximum electric field can be brought to a wavelength of 1.55 µm. This optical directional coupler may play an important role in the field of optical modulators as well as next-generation CMOS compatible photonic IC's.

Acknowledgements The authors would like to thank North-Eastern Hill University for providing UGC non-NET fellowship and department of electronics and communication engineering for providing all the required facilities to carry out the research.

References

1. Hosseininejad, S.E., Komjani, N. and Noghani, M.T., "A comparison of graphene and noble metals as conductors for plasmonic one-dimensional waveguides," IEEE Trans. Nanotechnol., vol. 14, 2015.
2. Kim, J. T., "Silicon Optical Modulators Based on Tunable Plasmonic Directional Couplers," IEEE journal of selected topics in quantum electronics, vol. 21, no. 4, 2015.
3. Shalaby, Hossam M. H., "Bi-Directional Coupler as a Mode-Division Multiplexer/Demultiplexer," J. Lightw. Technol., vol. 34, no. 15, pp. 3633–3640, 2016.

4. Kim, J. T. and Park, S., "The design and analysis of monolithic integration of CMOS-compatible plasmonic waveguides for on-chip electronic-photonic integrated circuits," J. Lightw. Technol., vol. 31, no. 18, pp. 2974–2981, 2013.
5. Vasudev, A. P, Kang, J.-H., Park, J., Liu, X., and Brongersma, M. L., "Electrooptical modulation of a silicon waveguide with an "epsilon-near-zero" material," Opt. Exp., vol. 21, no. 22, pp. 26387–26397, 2013.
6. Jun, Q. Y., de, H. P., jiang, R. Q., ping, Z. F., peng, G. L. and hua, Z. C., "A photonic wire-based directional coupler based on SOI," Optics Communications, vol. 281, no. 11, pp. 3105–3110, 2008.
7. Yamada, H., Chu, T., Ishida, S., and Arakawa, Y., "Optical directional coupler based on Si-wire waveguides," IEEE photonics Tech. lett., vol. 17, no. 3, 2005.

Chapter 49
Characteristics of Visible Light Communication Using Light-Emitting Diodes

Sujit Chatterjee, Rubi Baishya, Kubla Khan, Priya Sarma and Banty Tiru

Abstract In this paper, visible light communication using the white light-emitting diode based system for indoor communication is considered. The experimental setup for studying the characteristics of the system is presented. An analysis of the received signal strength as a function of the distance between the transmitter and receiver, number of receiving elements, and angle from the direct path is done. Results show that with a 1 W light-emitting diode, a transmission is possible to a distance of 6 m and message of 60 kHz. The arrangement can be used exclusively or as components of hybrid networking facilities.

Keywords Visible light communication · Light-emitting diode Power amplifier · Photodiode

Introduction

In recent years, the rapid development of semiconductor lighting devices such as light-emitting diodes (LED) provides an alternative for future lighting technologies [1]. The inherent advantages of LED over Compact Fluorescent Light (CFL) and halogen bulbs are energy efficiency, long lifetime, ruggedness, environmentally

S. Chatterjee (✉) · R. Baishya · K. Khan · P. Sarma · B. Tiru
Department of Physics, Gauhati University, Guwahati 781014, Assam, India
e-mail: sujitc43@gmail.com

R. Baishya
e-mail: rubibaishya89@gmail.com

K. Khan
e-mail: kublakhan21dec@gmail.com

P. Sarma
e-mail: priyasarma8@gmail.com

B. Tiru
e-mail: banty_tiru@rediffmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_49

friendly, and controllable. LED can also be used for visible light communication (VLC) by pulsing it at very high speed. This speed is very high for the human eye to detect [2, 3]. The inherent advantage of using VLC is that the portion of electromagnetic (EM) spectrum is license free and requires no regulations. There are no health issues and the transmitter and the receiver are very cheap compared to that of radio frequency (RF) [1]. Moreover, VLC does not interfere with the RF-based system and a good opportunity exists to incorporate in varied applications like airplanes, hospitals, and underwater [4]. In the recent years, a lot of research is going on in Asia, Europe, and the US for developing this mode of communication [5, 6]. Some of the research fields are spatial intensity on amplitude-frequency characteristics of VLC [7], modulation characteristics [8], and as broadband wireless home networking (WHN) [9]. A problem to be tackled is finding ways to increase the beam angle of LED so as to obtain large coverage area [10].

VLC can also be used as components of hybrid networking configuration that can supplement the shortcomings of the individual components. In the work, we aim to study the characteristics of VLC using a suitable transceiver system. The long-run plan of the work is to use the results for developing a hybrid communication network [11]. The VLC will be incorporated into Power Line Communication (PLC) for remote sensing of environmental parameters or for data communication between devices [12]. The block diagram of the proposed hybrid model is shown in Fig. 49.1. Sensor nodes will collect the required data and fed into the PLC modem [13, 14]. Via the available indoor power line of the building, the same will be made available in the same/different room via VLC. The paper gives the experimental setup and a study of the dependencies of the variability's for the VLC component.

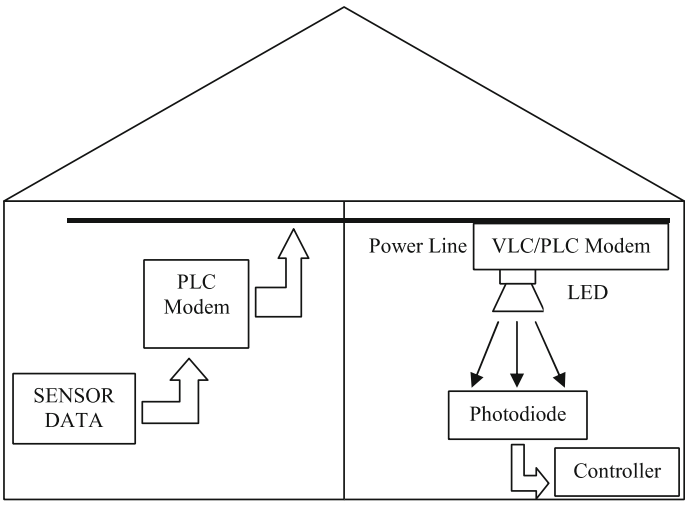


Fig. 49.1 Block diagram of the proposed hybrid system using power line communication and visible light communication

Experimental Setup and Measurement Technique

The block diagram for studying the characteristics of the VLC transceiver is shown in Fig. 49.2. The experimental setup consists of a transmitter and receiver module with the direct line of sight visible light channel. The schematic of the experimental setup is in Fig. 49.3.

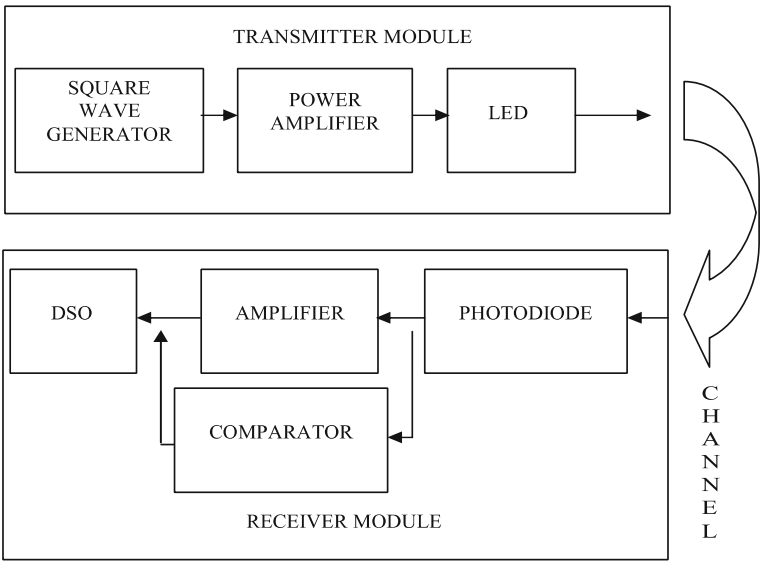


Fig. 49.2 Block diagram of VLC transceiver

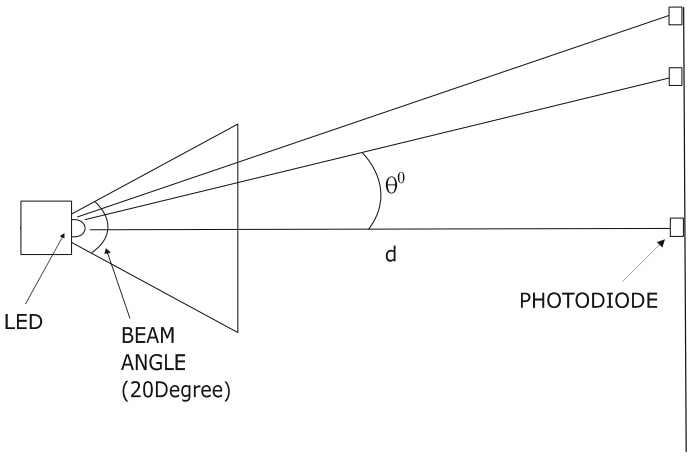


Fig. 49.3 Schematic of the transmitter- receiver system, the block diagram is given in Fig. 49.2

Transmitter Module

The transmitter module consists of a square wave generator, a power amplifier (PA) and LED. The square wave generator produces the square wave which acts as a message signal. Here, a square wave of 400 Hz and 1.1 V amplitude is fed into the PA that acts as the LED driver circuit. The PA is in class A biasing mode using IC TIP41C having a maximum collector to the base as well as a collector to emitter voltage of 100 V. The biasing voltage and current of 4.8 V and 0.5 A, respectively are supplied by a programmable power supply (PSD9001). A white LED (1 W) is placed at the load and acts as the transmitter of data transmission section. The beam angle of the LED is 120° and luminous flux is approximately of 90 lumens. The LED has a DC forward current of 250 mA and a forward voltage of 3.4 V. The maximum allowed forward current and voltages are 700 mA and 10.5 V, respectively. Due to the message signal, the LED blinks at the same frequency which is undetectable by eye. Using a suitable arrangement the beam angle of emitted light is nearly 20° to get a strong directivity.

Receiver Module

The receiver module consists of silicon PIN Photodiode (PD) BPW34, amplifier, or comparator circuit and a Digital Storage Oscilloscope GWINSTEK GDS-1052-U (50 MHz, 250 MS/s). The dimension of the PD is $5.4 \times 4.3 \times 3.2$ mm^3 with a radiant sensitive area of 7.5 mm^2. The typical dark current is about 2 nA and is sensitive to visible and near-infrared radiation. Further, the received signal from PD is allowed to pass through an amplifier or comparator circuit using IC LM318 having typical slew rate 70 V/μS. The amplifier designed has a gain of 10^5. Finally, the signal is stored in the DSO. The light from the LED falls directly on the PD and the results analyzed.

Channel

The channel between the LED and PD is the free space connected by visible light. The distance between the transceivers is changed and the results observed. The numbers of PD at the receiver module are also changed to check the sensitivity.

Results and Discussion

To study the performance of the VLC communication link, a number of experiments have been carried out by varying the (i) distance between the transceiver keeping the number of PD fixed and (ii) number of PD keeping the distance fixed and (iii) angle of the receiver keeping the number of PD and distance fixed. The results are shown in Table 49.1. In each case, the output of the amplifier is recorded. A snapshot of the received signal from the amplifier at a distance of 1 m and 5 m for 1 PD is shown in Fig. 49.4a and b, respectively. Figure 49.4c and d show the received signal when the frequency of the square wave is increased to 60 kHz and the output of the comparator, respectively. Analyzing the results, in

Table 49.1 Variation of the output voltage of the photodiode with varying dependencies

Experiment 1			Experiment 2			Experiment 3			
d (m)	N	O (mV)	N	d (m)	O (mV)	d (m)	N	A (°)	O (mV)
1	1	2180	1	1	2180	2	1	1	640
	2	4240		2	800			2	520
	3	6240		3	340			4	360
	4	8160		4	280			5	300
	5	9760		5	200			10	120
	…	…		7	Noisy			11	Noisy
2	1	800	2	1	4240		2	1	1260
	2	1520		2	1520			2	1010
	3	2120		3	590			4	680
	4	2700		4	450			5	600
	5	3320		5	340			7	400
	…	…		6	280			12	140
	…	…		9	Noisy			13	Noisy

Experiment 4							
d (m)	N	A (°)	O (mV)	d (m)	N	A (°)	O (mV)
2	1	0	172	2	2	0	220
		10	160			10	206
		20	144			20	190
		30	132			30	165
		40	120			40	150
		50	108			50	138
		60	98			60	128
		70	88			70	118
		80	75			80	115
		85	Noisy			85	Noisy

d: Distance between the transmitter and receiver, N: Number of photodiode at the receiver, O: Output (peak–peak) at the amplifier, A: Angle in degrees of the photodiode from the direct path
*Results shown in the table only up to 2 m

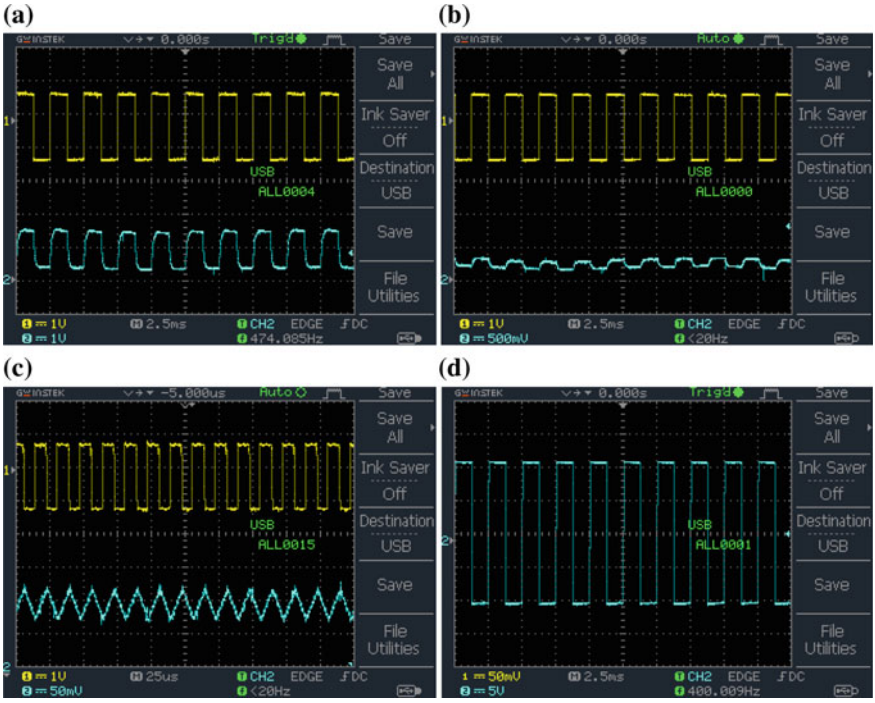


Fig. 49.4 **a** Snapshot of the input (top) and received (bottom) signal from the amplifier for 1 photodiode placed at a distance of 1 m from the LED and **b** at a distance 5 m from the LED **c** Snapshot of the input (top) and received (bottom) signal for 60 kHz signal for 1 photodiode placed at a distance of 1 m from the transmitter and **d** Output of the comparator

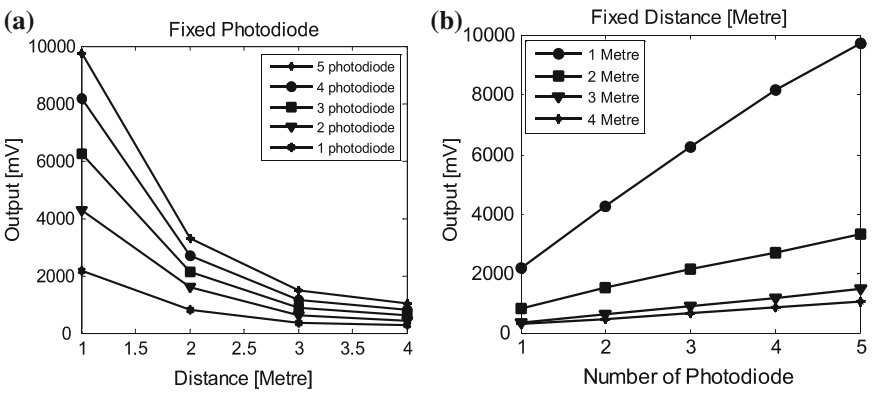


Fig. 49.5 **a** The output voltage as a function of the number of photodiodes and **b** distance

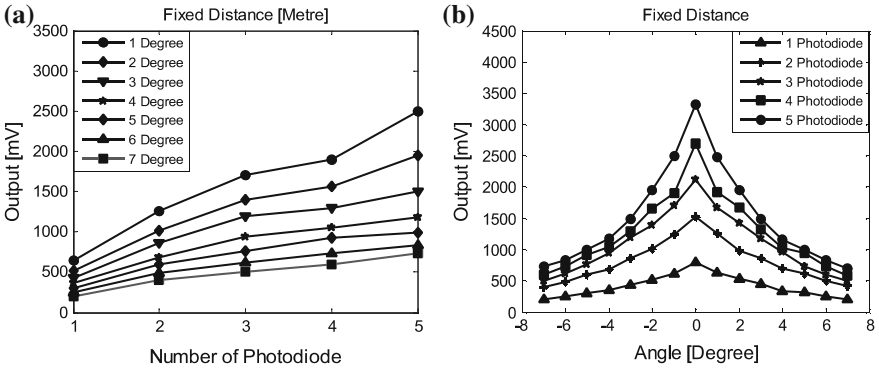


Fig. 49.6 **a** Variation of the output voltage of 1 photodiode at a distance of 2 m with different angle of the later and **b** The output voltage on both side of the axis

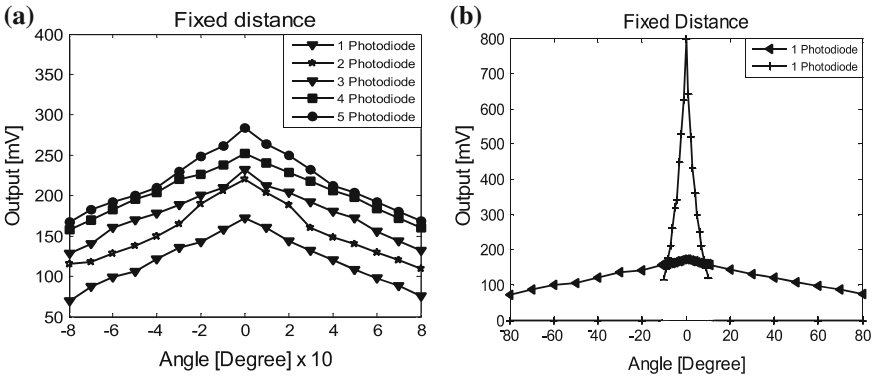


Fig. 49.7 **a** Variation of output voltage of both side of the axis at a distance of 2 m for 10 LED at the transmitter **b** Comparison of beam angle from the axis at a distance of 2 m for 1 and 10 LED at the transmitter

Fig. 49.5a it is found that when the number of PD is increased for a particular distance, the output voltage from the amplifier also increases. Also from Fig. 49.5b when the distance between the transceiver is increased, then the voltage at the receiver PD decreases as expected. However, it is to be noted that the increases of the output voltage with PD number are more significant at smaller distances. As the distance is increased, the received output power does not show improvement even when the number is increased to such as 5. From Fig. 49.6a and b, it is seen that as the angle of the receiver is changed, the received signal power decreases in both the angles from the direct path. The number of LED is also increased at the transmitter place in a horizontal spread to increase the coverage of the transmitter on both sides of the direct path. Figure 49.7b shows that for an increase of LED to 10 numbers (1 W each), the coverage increases by 700%.

Conclusion

In this work, we have designed an experimental setup that can be used for VLC using white LED and photodiode. It is found that the received signal strength decreases with distance. The same is also increased when the number of photodiodes is increased. However, the increase is significant for the small distance between the transceiver. The system works for the small angular width of the receiver and for a frequency of 60 kHz message signal. With 2 PDs the distance available is 8 m. The received signal at 60 kHz is highly distorted because of the PA. Further frequency can be transmitted by proper design of this. In the future, this arrangement will be used to acquire data from environment monitoring setups and other data transmitting devices. The increase in the number of LED at the transmitter in a horizontal spread increases the coverage of the receiver in both the directions of the axis. It is concluded that the coverage area is directly dependent on the beam angle of the LED. In the future, the effect of increasing the power of the LED will be studied. Hardware will be developed for remote data acquisition and control using VLC. This system will be incorporated as a part of a hybrid network comprising of wireless, power line, and VLC to be used as a robust and efficient communication system.

References

1. Mostafa Z. Afgani, Harald Haas, Hany Elgala and Dietmer Knipp.: Visible Light Communication Using OFDM. Testbeds & Research Infrastructures for the Development of NeTworks & Communities, Barcelona, Spain, March 1–3, (2006)
2. Irina Stefan, Hany Elgala and Harald Haas.: Study of Dimming and LED Nonlinearity for ACO-OFDM Based VLC Syetems. IEEE Wireless Communications and Networking Conference: PHY and Fundamentals, (2012)
3. Arnold Wilkins, Jennifer Veitch and Brad Lehman.: LED Lighting Flicker and Potential Health Concerns IEEE Standard PARI1789 Update. Energy Conversion Congress and Exposition (ECCE), (2010)
4. Wenbo Ding, et al.: A hybrid power line and visible light communication system for indoor hospital applications. Computers in industry, pp. 170–178, (2015)
5. H.L. Minh, et al.: 100-Mb/s NRZ Visible Light Communications Using a Post equalized White LED. in IEEE Photonics Technology Letters, Vol. 21, No. 15, (2009)
6. Z. W, J. Chau and T. Little.: Modeling and Designing of a New Indoor Free Space Visible Light Communication system. Networks and Optical Communications (NOC) in 16th European Conference, pp. 72–75, (2011)
7. Z. Xu, et al.: Effect of High-Power Led Spatial Light Intensity on Amplitude-Frequency Characteristics of VLC. Optical Communication Technology, Vol. 35, (2011)
8. J. Cao, Z. Liang and Z. Ma.: White LED Modulation Bandwidth and Modulation Characteristics of the Study. Hans Journal of Wireless Communications, (2012)
9. K. D. Langer et al.: Optical Wireless Communications for Broadband Access in Home Area Networks. in Proc. Of 10th Anniversary International Conference on Transparent Optical Networks (ICTON), Athens, Greece, vol. 4, pp. 149–154, Jun. 22-26, (2008)

10. Hany Elgala, Raed Mesleh, Harald Haas.: Indoor broadcasting via white LEDs and OFDM. IEEE Transactions on Consumer Electronics, Vol. 55, August (2009)
11. Banty Tiru.: Exploiting Power Line for Communication Purposes, Intelligent Applications for Heterogeneous System Modeling and Design, IGI Global, Chapter 14, pp 327–340
12. Banty Tiru, PK Boruah.: Design and Testing of a Suitable Transceiver for Full Duplex Communication in Indoor Power Line. IETE Journal of Research 56(5), https://doi.org/10.4103/0377-2063.72776, September (2010)
13. B. Tiru, P. K. Boruah, H. Medhi.: A DSP based Channel Selection Algorithm for Adaptive Transmission in Indoor Power Line, Indian Journal of Physics, vol 84(6), pp. 745–749, ISSN: 0973–1458 (2010)
14. B. Tiru, R. Baishya, U. Sarma.: An analysis of indoor power line network as a communication media using ABCD matrices, Lecture Notes in Electrical Engineering, Springer, Chapter 14, 2015

Chapter 50
C-Band Silicon Optical Modulator for High-Speed Optical Communication System

Silpeeka Medhi, Subhash C. Arya and Balbindar Kaur

Abstract The static performance of MOS-based silicon optical modulator has been depicted by analyzing the behavior of figure of merit (VπL) with respect to optical modulator height (h). The loss characteristics of modulator have also been shown. It has been observed that figure of merit is improved (from 3.5 to 1.5 V cm) by decreasing the modulator height (from 1600 to 800 nm) at the cost of an increase in loss coefficient (12–13.5 db/cm). These results are useful in designing high-speed silicon optical modulator, which is compatible with matured CMOS fabrication technology.

Keywords Optical modulator · MOS · Figure of merit · Effective index Loss coefficient

Introduction

Optical modulators are used for modulating any one of the parameters of the beam of light, phase, frequency, or amplitude using the electro-optic effect. There are around five major emerging technologies for broadband optical modulator design: (i) Electro-Optic Polymer Modulator, (ii) Gallium Arsenide Modulator technology, (iii) Complex Oxide Electro-Optic Modulator, (iv) Organic Electro-Optic Crystal

S. Medhi (✉) · S. C. Arya · B. Kaur
LWCS Laboratory Department of Electronics and Communication Engineering, North Eastern Hill University, Shillong 793022, Meghalaya, India
e-mail: silpeekamedhi@ymail.com

S. C. Arya
e-mail: aryasubh@gmail.com

B. Kaur
e-mail: balbindar27kaur@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_50

modulator, and (v) Silicon High-Speed Modulator [1]. Among these, the silicon-based optical modulator is thrust area at present which aims to exploit silica for optical modulator purpose. In general, silicon photonics is now innovative technology; however, researchers are refining the technology to circumvent the inherent performance limitations of silicon and to meet evolving application requirements. Metal-oxide-semiconductor (MOS) capacitor-based modulators works with carrier accumulation principle, where only majority carriers accumulate on either side of the insulating layer of silicon dioxide. The same layer after being placed in the wave-guiding region limits the bandwidth of the device till resistance-capacitance cut-off frequency.

The modulation of the optical signal can be attained by varying the refractive index of silicon through plasma dispersion effect, which is relatively less explored zone and hence results in a significant gap between the technology of electronic and photonic device fabrication. The proposed research work aims to minimize this technological gap between electronics and photonics for device fabrication using SOI state of art. The Silicon-on-Insulator (SOI) technology allows large integration of electronic and photonic devices for miniaturization of devices, used in photonics and fiber communication systems.

Various designs of optical modulator have been studied, namely: p-n, p-i-n, MOS, etc. silicon optical modulator. One of the initial experimental results of optical modulator using MOS has been reported in the year 2004. A study on various structures of SOI optical phase modulators was carried out by Mardiana et al. [2] based on the free-carrier dispersion effect at low-loss optical window. By selecting ideal position of doping regions in SOI optical phase modulator, its modulation efficiency of 0.015 V cm with a length of 155 μm was achieved using simulation.

Rao and Della Corte [3] have numerically simulated the free-carrier injection and depletion-based electro-optic modulator. Their experimental results attained birefringence free, low-loss, single-mode waveguide-integrated phase modulator. An upgraded phase modulation efficiency of 1.6 V cm had been achieved in noble compact configuration, requiring lower bias voltages for a better CMOS compatibility.

It is a challenge to attain high-speed optical intensity modulation in silicon (Si) because the material does not exhibit the appreciable electro-optic effect. Silicon being indirect band-gap material and due to its crystal symmetry, fast and effective modulation response is difficult to design when compared to direct band-gap materials such as III-V group materials. Modulation efficiency of 0.015 V cm with a length of 155 μm was found using simulation by selecting the ideal position of doping regions in Silicon-on Insulator optical phase modulator. Figure 50.1 shows the structural diagram of MOS-capacitor-based optical modulator designed using Atlas Silvaco software.

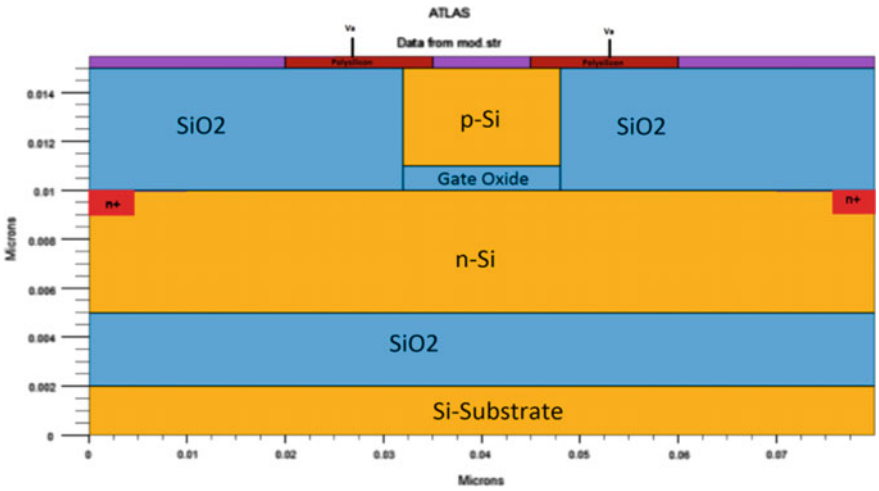


Fig. 50.1 MOS-capacitor-SOI optical modulator

The front, top, and side view of MOS capacitor-based optical modulator have been shown in the Fig. 50.2, which is derived from reference [1, 4] where only cross-section has been given. It comprises an n-type doped crystalline silicon slab and a p-type doped polysilicon rib with a gate oxide sandwiched between them. The

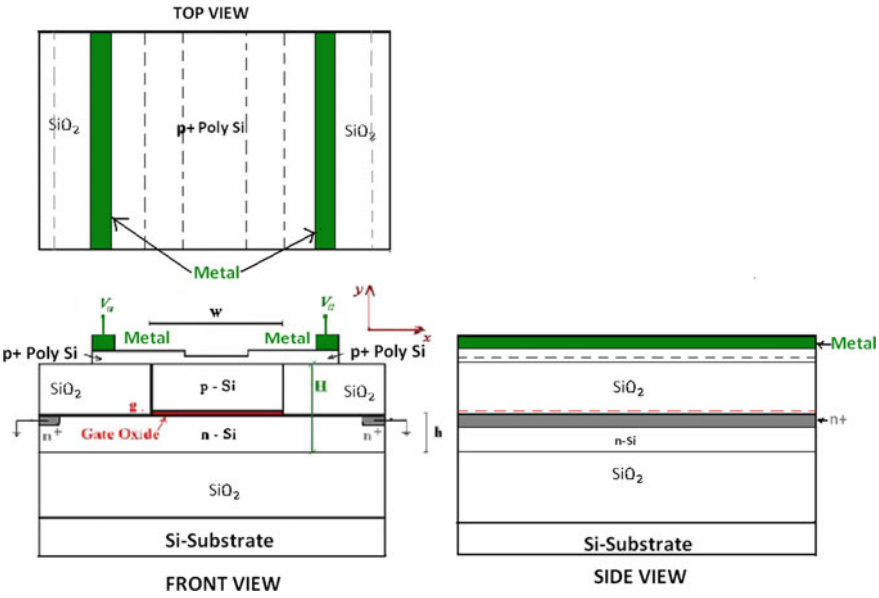


Fig. 50.2 Cross section of MOS-capacitor-based SOI optical modulator, as in [1]

optical absorption by metal contact has been minimized by using a wide polysilicon layer on the top of the oxide layers on both sides of the polysilicon rib. The oxide regions on either side of the rib maintain optical confinement and prevent optical filed penetrating into areas where the metal contacts are located [5].

Mathematical Modeling

The working principle of CMOS optical modulator is that the refractive index of silicon varies with the voltage applied to the modulator, which in return varies the speed of light passing through the modulator.

$$\text{Light Speed} \propto 1/(\text{Refractive Index of material}) \tag{50.1}$$

The refractive index of the material used in the modulator is due to the change in the carrier concentration, i.e., electrons and holes. The phenomenon of change of phase velocity with respect to the change in the refractive index of the modulator is called as the free-carrier dispersion effect [6].

The electrical and optical behavior of modulator is shown separately. The electrical characteristic can be modeled using 2-D drift diffusion equations, which are given as the three-coupled partial differential equations (PDEs) describing electrons and hole distribution [1–7]:

$$\in_0 \in_r (\nabla \Psi) = q(N_e - N_h - C), \tag{50.2}$$

$$\nabla.\left[\left(\mu_p v_t \nabla N_h\right) + \left(\mu_p N_h \nabla \Psi\right)\right] - \frac{\partial N_h}{\partial t} = R_{SRH}, \tag{50.3}$$

$$\nabla.(\mu_n v_t \nabla N_e - \mu_n N_e \nabla \Psi) - \frac{\partial N_h}{\partial t} = R_{SRH}, \tag{50.4}$$

where R_{SRH} is the recombination rate Shockely-Hall-Read, N_h is the hole distribution, N_e is the electron distribution, $\in_r$ is the relative permittivity, $\in_0$ is the vacuum permittivity, C is the doping profile, μ_n is the electron mobility, μ_p is the hole mobility, v_t is the thermal voltage, Ψ. the electric potential, and R_{SRH} can be defined as [4].

Next, the static figure of merit $V_\pi L$, (V_π the half-wave voltage and L modulator length) needs to be calculated using the formula [4]:

$$V_\pi L = \frac{\lambda}{2}\left(\frac{\Delta \eta_{eff}}{\Delta V_a}\right)^{-1}, \tag{50.5}$$

Or $$V_\pi L = \frac{\lambda}{2}\left(\frac{\Delta V_a}{\Delta \eta_{eff}}\right),$$

Where η_{eff} is the effective index.

The Effective index for Quasi-TE mode and Quasi-TM mode has a linear relation with applied voltage on the modulator [8].

Results and Discussion

After solving the 2-D drift diffusion equations (Eqs. 50.1–50.3), the values of the potential and the density of electrons and holes are calculated, which are used in $V_{\pi}L$ using Soref–Bennetd relationship [2].

Using Eq. 50.5, the variation of figure of merit (VπL) for varying wavelength can be seen in Fig. 50.3. From Fig. 50.3, it can be said that the value of Figure of merit for quasi-TM (Transverse Magnetic) mode is much higher than that of quasi-TE (Transverse Electric) mode. On the other hand, the graph also shows the dependence of optical modulator performance on polarized incoming light.

Figure 50.4 shows the relation between figure of merit (VπL) and loss coefficient (α_{loss}) with respect to height of the modulator for quasi-TE and quasi-TM Mode. The value for VπL and α_{loss} have been calculated for different value of gate oxide thickness of g = 4, 6, 10.5 nm for Transverse electric (TE), and Transverse Magnetic (TM) modes. It has been observed the trend lines of Vπ L and α_{loss} follow linear characteristic (y = mx + c). Table 50.1 shows such linear equations at g = 6 nm quasi-TE and quasi-TM modes. Minimum value of $V_{\pi}L$ is observed from 0.52 to 1.237 for gate oxide thickness (g) of 4 nm at the cost of loss coefficient α 13.45 for quasi TE mode. Minimum value of $V_{\pi}L$ is observed from 0.242 to 1.08 for gate oxide thickness (g) of 4 nm at the cost of loss coefficient 16.75 for quasi-TM mode. Since loss coefficient is 19.7% more in case of TM mode compare to TE, therefore, TE mode is more efficient for plasma dispersion silicon MOS optical modulator.

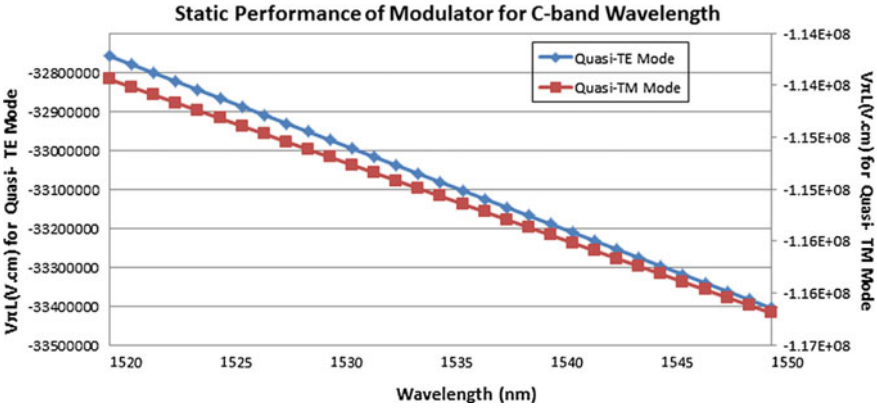


Fig. 50.3 Static performance of silicon optical modulator for C-band wavelength

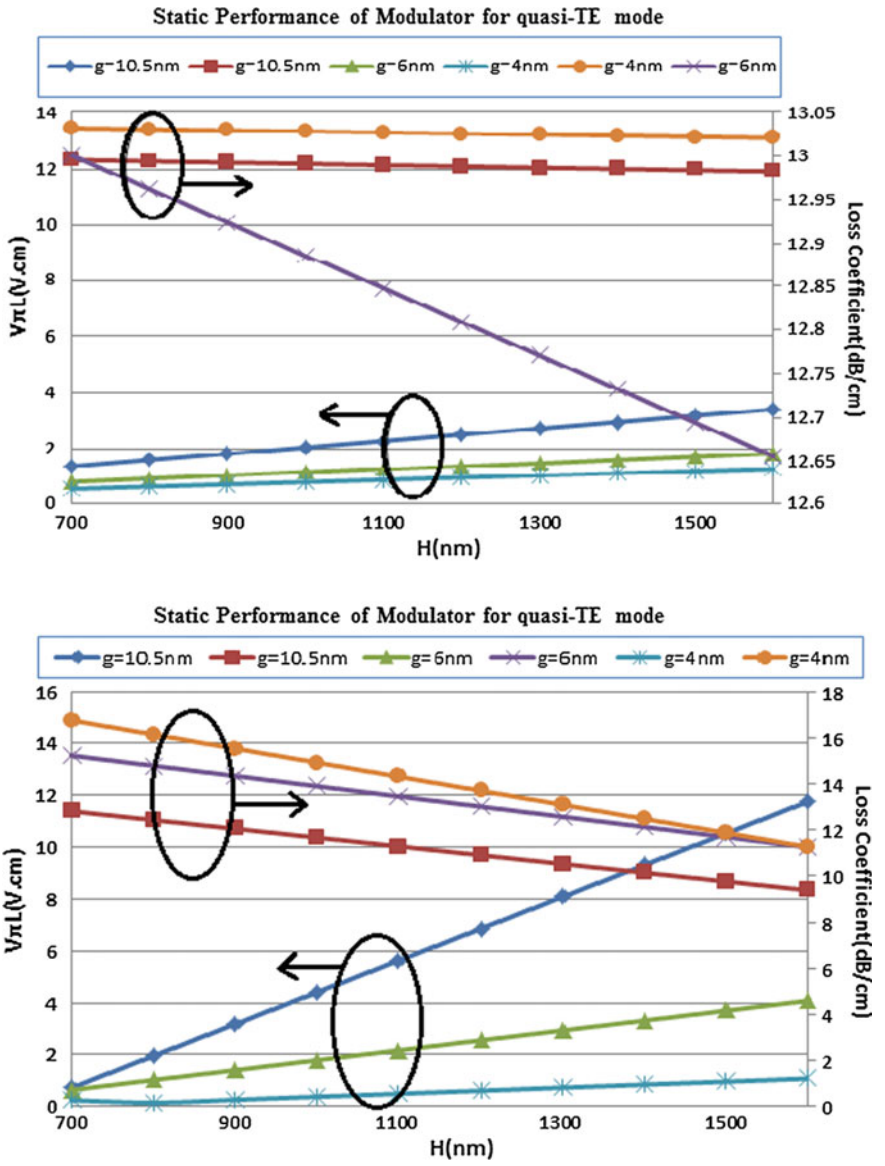


Fig. 50.4 Figure of merit (V_πL) and loss coefficient (α_{loss}) behavior with respect to height of the modulator for quasi-TE and quasi-TM. Mode

Table 50.1 Fitting parameters for the Fig. 50.4 at gate oxide thickness (g) = 6 nm

Modes	Figure of merit (V_πL)	Loss coefficient (α_{loss})
TE	0.012H−0.0432	−0.0004H + 13.27
TM	0.0038H−2.0315	−0.00393H + 16.28

Conclusion

In this paper, Figure of merit of the optical modular has been calculated and observed that the modulation of the information passing through C-band silicon optical modulator highly depends on the input polarized light. In addition, the efficiency of the C-band optical modulator can be improved by minimizing the figure of merit decreasing the height of optical modulator but at the cost of loss coefficient. Therefore, there is trade-off between figure of merit and loss coefficient about which comparative analysis has been done for quasi TE and TM modes. Our reported results are in confirmation with previous published results.

References

1. Chen, A., and Murphy, E. J., "Broadband Optical Modulators: Science, Technology and Application", CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suit 300 Boca Raton FL33487-2742, USA.
2. B. Mardiana, Sahbudin Shaari, P. Susthitha Menon, H. Hazura, A.R. Hanim, N. Arsad, H. Abdullah "Analyses for various doping structures of SOI-based optical phase modulator using free carrier dispersion effect", Journal of Optic, 125, pp. 1800–1803, (2013).
3. S. Rao, and F. G. Della Corte, "Numerical Analysis of Electro-Optical Modulators Based on the Amorphous Silicon Technology", Journal of Lightwave Technology, vol. 32, no. 13, pp. 2399–2407, (July 1, 2014).
4. V. M. N. Passaro, and Dell'Olio, F., "Scaling and Optimization of MOS Optical Modulators in Nanometer SOI Waveguides", IEEE Transactions on Nanotechnology, vol. 7, no. 4, pp. 401–408, (2008).
5. Ansheng Liu et al., "A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor", Letter to Nature, vol. no. 427, pp. 615–618, 12 Feb. 2004.
6. Graham T. Reed et al., "Silicon Optical Modulators", Materials Today, January 2005, pp. 4–50
7. Silpeeka Medhi et al., "Mathematical modeling of silicon based optical modulator for high speed fiber optic communication", Int. Conf. on Computing and Communication System (I3CS15) pp. 171–175, at NEHU, Shillong, India during 9–10th April, 2015.
8. D. Marris-Morini et al., "High speed all-silicon optical modulator", Journal of Luminescence 121 (2006), pp. 387–390.

Chapter 51
Power Spectral Study of EEG Signal from the Frontal Brain Area of Autistic Children

Bablu Lal Rajak, Meena Gupta, Dinesh Bhatia, Arun Mukherjee, Sudip Paul and Tapas Kumar Sinha

Abstract Autism or autism spectrum disorder (ASD) represents complex developmental disabilities characterized by deficits in social communications, interactions, and cognitive development. The prevalence of ASD shows a growing trend both in developed and developing countries. ASD occurs due to improper brain development in early life and individuals characterized as ASD possesses abnormal brain activity that is commonly studied using electroencephalography (EEG). Our present work analyzes the EEG of ASD children from the frontal lobe of the brain that is responsible for social, emotion, and cognitive functions, which was compared with the EEG signals of normal healthy children. The power spectra (PS) of EEG signal were obtained using fast Fourier transformation (FFT) algorithm in MATLAB. EEG recording was performed on all the ten selected children (five ASD and five normal) using two electrodes placed on F3 and F4. The artifact-free EEG signals of 10 min duration were extracted and used for obtaining PS. The PS revealed high-intensity power peak at frequency 50 Hz, for all healthy children; but in case of ASD participants, there existed two peaks at 100 and 50 Hz. The intensity of 50 Hz peak in ASD cases was not as intense as those of normal children

B. L. Rajak · M. Gupta · D. Bhatia (✉) · S. Paul
Department of Biomedical Engineering, North Eastern Hill University, Shillong, Meghalaya, India
e-mail: bhatiadinesh@rediffmail.com

B. L. Rajak
e-mail: bablurajak@gmail.com

M. Gupta
e-mail: meenaguptaphysio@gmail.com

S. Paul
e-mail: sudip.paul.bhu@gmail.com

A. Mukherjee
UDAAN-for the Differently Abled, New Delhi, India
e-mail: docarun@yahoo.com

T. K. Sinha
Computer Centre, North Eastern Hill University, Shillong, Meghalaya, India
e-mail: tksinha001@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_51

but the 100 Hz peak was highly intense. The existence of high-intensity peaks in ASD can be attributed to the imbalance in high-frequency EEG rhythm that is responsible for perpetual and cognitive processing in humans.

Keywords Autism spectrum disorder · Electroencephalogram
Fast Fourier transformation · Power spectrum

Introduction

Autism or autism spectrum disorder (ASD) is an umbrella term that includes child autism and developmental disorders. ASD represents complex developmental disabilities characterized by deficiency in social communications, interactions, and cognitive development. These characteristics appear in early childhood and tend to persist lifelong with a poor outcome in adulthood [1]. From public health perspective, ASD is an important cause of morbidity due to their early onset, lifelong persistence, associated impairments, and the absence of effective treatment. The prevalence of ASD is reported to be 1 in 88 children in USA [2], and in India, it is around 1–1.5% or 1 in 66 children [3]. The actual cause of autism largely remains unknown but some studies found that abnormal brain growth in early age cause changes in gray and white matter contents, leading to increased cerebral volume [4, 5]. Magnetic resonance imaging study on young children affected with ASD revealed bilateral enlargement of amygdalae and hippocampi, which was proportional to overall increase in volume of the cerebral cortex [6]. However, recently published literatures based on family studies and genetics established higher genetic and environmental contributions that disturb the normal biological pathways contributing to the disorder [7, 8]. Whatever could be the cause, it is known that abnormalities in the frontal lobe are responsible for generating severe impairment in social, emotional, and cognitive functions in autism [9].

Despite extensive research toward understanding the neural basis of transformed behavior in ASD, there still exists substantial debate about the functional and neurophysiological workings of an "autistic" brain [10, 11]. In view of this, several neuroimaging and neurophysiological methods have been used to comprehend the correlation between normal brain function and its autistic counterpart. Among these, clinical electroencephalography (EEG) studies are finding great interest generally due to high prevalence of epileptic abnormalities in autistic patients [12, 13] and for evaluating diverseness of behavioral disorders, any treatment responses, and consequences among other issues [14]. The simplicity of the EEG procedure and its determination of brain activity along with consistent analysis protocols provide an opportunity for intricate analysis of functions and dysfunctions of the brain.

EEG is a recording of spontaneous brain activity, recorded from the scalp surface of the brain employing metal electrodes pasted using conductive media [15]. EEG recording is widely used to distinguish brain's neurological and

neurophysiological disorders. The importance of EEG is its ability to be characterized by linear or nonlinear investigation that depicts distinct state of the brain. One of the most widely used feature extraction techniques is a computerized analysis of EEG using Fourier transformations that convert signal from time domain into frequency domain. Fast Fourier transformation (FFT) gives relative power spectra (PS), which represents power change with change in frequency, or in other words, the slope of the power curve as a function of frequency [16]. In this study, we analyzed the power spectrum of EEG signals of ASD children and compared it with those of normal children to detect change in periodicity or intensity of the signal.

Materials and Method

Participants

Ten children (all male) matched by age were recruited in this study after obtaining written consent from their parents/guardians. Five children (mean age: 9.55 ± SD 2.38) were ASD cases as diagnosed by consultant physician and another five (mean age: 8.26 ± SD 3.05) were normal healthy children without any neurological disorders. The recruited ASD children were from UDAAN for the differently abled Delhi, a nonprofit organization that has pioneered in therapy-based treatment for cerebral palsy and ASD using physical, occupational, and speech therapy since 1992. The normal children were also from the same organization that runs an outreach learning program for the poor children of nearby areas. This study was conducted after approval from the Institutional Ethics Committee for Human Samples/Participants (IECHSP) of the host institution.

EEG Recording

The EEG of all the recruited participants was recorded using Nexus Mark II, neurofeedback system (Mind Media B.V., Netherlands). Prior to EEG recording, the participants went through a preparatory procedure where the selected recording area on their scalp was cleaned using alcohol swabs, and the metal electrodes connected to the device was fixed using a conductive gel. This was a single-channel EEG recording; the placement of electrode on the scalp was according to Montage 10–20 system [17]. The selected recording area was F3–F4 with electrode at right mastoid as reference. After placement of the electrodes, the recording was taken for 15 min without using any anesthetics; that is, the patients were awake.

Data Analysis

Fifteen-minute EEG data was recorded at the rate of 256 samples per second using a single-channel bipolar electrode. The recorded data was exported in a text file for analysis in a user-designed FFT program in MATLAB. Preprocessing of the data was carried out to remove the 50 Hz artifact by using the device in battery mode. The EEG data of 10 min duration was used from different groups of children for this power spectral study. The analysis was performed with FFT–Welch method to obtain the power spectrum that was plotted power versus frequency. Welch method is a nonparametric analysis accomplished by dividing the signal into smaller segments. Multiplying these smaller segments with an appropriate window size, the periodogram was computed by calculating the squared magnitude on the result for obtaining a distinct Fourier transform. The periodogram obtained was then averaged resulting in power intensity with respect to frequency [18]. In this method, the data sequences $x_i\ (n)$ can be represented as

$x_i(n) = x(n + iD)$ here $n = 0, 1, 2, \ldots, M - 1$ and $I = 0, 1, 2, \ldots, L - 1$

where iD is the beginning of ith sequence. This will lead to formation of L data segments each of length $2M$. The modified periodogram is given as

$$\tilde{P}_{xx}^{(i)}(f) = \left| \frac{1}{MU} \sum_{n=0}^{M-1} x(n)\omega(n)e^{-j2\pi fn} \right|^2,$$

where U is the standardization factor for the power in the window function given as

$$U = \frac{1}{M} \sum_{n=0}^{M-1} \omega^2(n)$$

and $\omega(n)$ is the window function. Finally, the Welch power spectrum is obtained which is the mean of this periodogram and is given as

$$P_{xx}^{W}(f) = \frac{1}{L} \sum_{i=0}^{L-1} \tilde{P}_{xx}^{(i)}(f)$$

Result and Discussion

This study was performed to compare the power spectrum density of EEG signal from the frontal area of the brain of ASD and normal children using FFT. Previous studies using FFT demonstrated PS of different EEG sub-bands (alpha, beta, delta, gamma, and theta waves) but we could not find any literature closely related to employment of FFT used in raw EEG signals of ASD cases. Thus, the discussions presented here are based on different possibilities that could be viable reasons for

validating our results. In the study, we observed that power spectra of EEG signal from normal children revealed high-intensity peak at 50 Hz (Fig. 51.1a), but two peaks of comparatively high intensity at 100 and 50 Hz were observed in PS of ASD children (Fig. 51.1b). The presence of high-intensity power peak at 100 Hz in ASD EEG signal may be attributed to the existence of irregular electrophysiological brain activity as reported in [19] due to the existence of excess of high-frequency oscillations. High-frequency EEG rhythm, often referred to as high gamma oscillations, is known for cognitive processes such as memory and attention; and an increase in gamma frequency is associated with neuropsychiatric disorders such as schizophrenia, Alzheimer's epilepsy, and ASD [20]. Moreover, imbalance in the excitation–inhibition stability within the cerebral cortex also produces excessive EEG oscillations that contribute to abnormal electrophysiological brain activity in

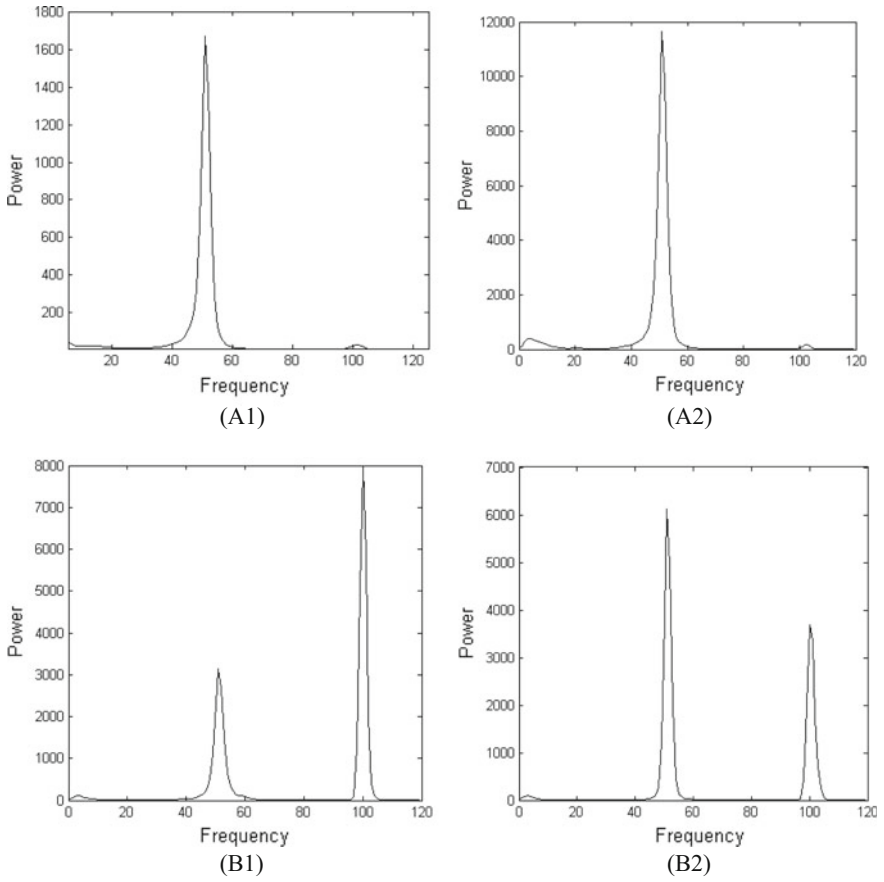


Fig. 51.1 Sample power spectrums of EEG signal from the frontal brain area (F3–F4) of normal (a1, a2) showing prominent 50 Hz peak and ASD (b1, b2) children showing two intensity peaks at 50 and 100 Hz

ASD [19]. Another factor responsible for unusual brain activity is the underconnectivity within and between the frontal hemispheres (prefrontal cortex) of the brain which is associated with cognitive dysfunction [21]. Additionally, power intensity of 50 Hz peak in ASD cases was lower in comparison to those of normal cases; this may be due to improper development of the autistic brain as reported in [5, 6].

Conclusion

Using our FFT algorithm, we observed high-intensity power spectral peaks at 50 Hz in all the EEG of normal children and complementary peak at 100 Hz in ASD cases. The presence of two peaks at 50 and 100 Hz in autistic children may originate due to imbalance in electrophysiological activity of "autistic brains", especially from the frontal lobe since this part of the brain plays important role in higher order social, cognitive, language, and emotional functions; each of which is extremely lacking in autism or it can be attributed to the improper development of their brain due to volumetric reduction in regions of corpus callosum, basal ganglia, frontal lobes, and cerebellum. These are regions of the prefrontal cortex responsible for attention, memory, and information processing. Moreover, if it is due to any other reason of abnormal activity of the brain in ASD, it is important to understand the actual cause that leads to this disorder and corrective approaches employed along with proper diagnostic tools to detect ASD in early developmental stages of the child. In this line of work, we intend to optimize the FFT algorithm that can be used as an early diagnostic tool for detecting ASD cases using simple EEG recordings.

Acknowledgements This work is supported by fund received from (Ref: SEED/TIDE/007/2013) from the Department of Science and Technology (DST), Government of India, New Delhi. A special thanks to all the participants and their parents for participating in this study.

References

1. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, (2013).
2. Baio, J.: Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. Morbidity and Mortality Weekly Report. Surveillance Summaries. vol. 61, no. 3. Centers for Disease Control and Prevention. (2012).
3. Secrets of the autistic mind, India Today, (2013). http://indiatoday.intoday.in/story/autism-autisitic-mind-western-syndrome-myths-about-autism-autistic-children/1/322242.html.
4. Aylward, E. H., Minshew, N. J., Field, K., Sparks, B. F., Singh, N.: Effects of age on brain volume and head circumference in autism. Neurology, vol. 59, no. 2, pp. 175–183, (2002).
5. Courchesne, E., Karns, C. M., Davis, H. R., Ziccardi, R., Carper, R. A., Tigue, Z. D., et. al.: Unusual brain growth patterns in early life in patients with autistic disorder an MRI study. Neurology, vol. 57, no. 2, pp. 245–254, (2001).

6. Sparks, B. F., Friedman, S. D., Shaw, D. W., Aylward, E. H., et al.: Brain structural abnormalities in young children with autism spectrum disorder. Neurology, vol. 59, no. 2, pp. 184–192, (2002).
7. Realmuto, George M. "Autism Spectrum Disorder." In The Medical Basis of Psychiatry, pp. 401–422.Springer New York, 2016.
8. Bolkan, S., Gordon, J. A.: Neuroscience: Untangling autism. Nature, vol. 532, no. 7597, pp. 45–46, (2016).
9. Courchesne, E., Karen, P.: Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. Cur. opinion in neurobio., vol. 15, no. 2, pp. 225–230, (2005).
10. Schipul, S. E., Keller, T. A., Just, M. A.: Inter-regional brain communication and its disturbance in autism. Front.Syst.Neurosci. vol. 5, pp. 10, (2011).
11. Narzisi, A., Muratori, F., Calderoni, S., Fabbro F., Urgesi, C.: Neuropsychological profile in high functioning autism spectrum disorders. J. Autism Dev. Disord. vol. 43, pp. 1895–1909, (2012).
12. Rossi, P. G., Parmeggiani, A., Bach, V., Santucci, M., et al.: EEG features and epilepsy in patients with autism. Brain Dev. vol. 17, pp. 169–74, (1995).
13. Chez, M. G., Chang, M., Krasne, V., Coughlan, C., et al.: Frequency of epileptiform EEG abnormalities in a sequential screening of autistic patients with no known clinical epilepsy from 1996 to 2005. Epilepsy Behav. vol. 8, pp. 267–71, (2006).
14. Sheikhani, A., Behnam, H., Noroozian, M., Mohammadi, M. R., et al.: Abnormalities of quantitative electroencephalography in children with Asperger disorder in various conditions. Res. Autism Spectr. Disord. vol. 3, pp. 538–546, (2009).
15. Niedermeyer, E., da Silva, F. L.: Electroencephalography: basic principles, clinical applications, and related fields, Lippincott Williams & Wilkins, (2005).
16. Dumermuth, G., Flühler, H.: Some modern aspects in numerical spectrum analysis of multichannel electroencephalographic data. Medical and biological engineering. vol. 5, no. 4, pp. 319–31, (1967).
17. Klem, G. H., Lüders, H. O., Jasper, H. H., Elger, C.: The ten-twenty electrode system of the International Federation. Electroencephalogr. Clin. Neurophysiol. vol. 52, no. 3, pp. 3–6, (1999).
18. Welch P. D.: The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics. vol. 15, no. 2, pp. 70–3, (1967).
19. Orekhova, E. V., Stroganova, T. A., Nygren, G., Tsetlin M. M., et al.: Excess of high frequency electroencephalogram oscillations in boys with autism. Biological psychiatry. vol. 62, no. 9, pp. 1022–1029, (2007).
20. Herrmann, C.S., Demiralp T.: Human EEG gamma oscillations in neuropsychiatric disorders. Clinical neurophysiology. vol.116, no. 12, pp. 2719–33, (2005).
21. Murias, M., Swanson, J.M., Srinivasan, R.: Functional Connectivity of Frontal Cortex in Healthy and ADHD Children Reflected in EEG Coherence. Cerebral Cortex. vol.17, pp. 1788–1799, (2007).

Chapter 52
An Euler Path Based Online Testing Technique to Detect Catastrophic Fault in Triangular DMFBs

Piyali Datta, Amartya Dutta, Riya Majumder, Arpan Chakraborty, Debasis Dhal and Rajat Kumar Pal

Abstract Defective microfluidic chip increases the assay completion and total turnaround time and it is the main reason of deviation of the actual result of assay operation. In online testing, the test droplets are moving out of step with each other. Again testing of DMFB is NP-hard in nature. Using Euler path based test technique, we can test the whole chip but we cannot apply Hamiltonian path based test method in case of equilateral triangular electrode array due to routing constraints in triangular DMFB. A graph-based test planning technique for online testing is proposed in this paper. Considering the active parts of the chip, like mixer and store cell, as obstacle in the triangular microfluidic array, we test the chip during the assay operation running in some portions of the array. Here, we focus on Euler path based test technique for catastrophic fault detection.

Keywords Lab-on-a-chip · Catastrophic and parametric failure
Online testing · Hamiltonian path · Euler circuit/path · Eulerisation
Graph matching · Testing

Introduction

Microfluidic biochips, also renowned as "*lab-on-a-chip*", have become very important biomedical analytic instruments [1, 2] as the experiments in a laboratory can be efficiently performed within a chip. It manipulates nanolitre or microlitre volumes of biological samples and reagents which lead to consumption of a very low cost and relatively a very less time. Moreover, higher sensitivity of DMFBs and less involvement of human manipulation make it less erroneous than laboratory experiments. In spite of all the facilities of a biochip system, physical defects may arise during microfluidic operations. Some manufacturing defects may be realised

P. Datta (✉) · A. Dutta · R. Majumder · A. Chakraborty · D. Dhal · R. K. Pal
Department of Computer Science and Engineering, University of Calcutta, JD-2, Sector—III Saltlake, Kolkata 700 098, West Bengal, India
e-mail: piyalidatta150888@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_52

during the assay operations and may lead to a crucial diagnosis process in vain. It is worthwhile to mention here that effective testing methodologies to test these devices after manufacture and during assay operations are very essential as biochips are generally used for vital biochemical and medical applications.

Generally, the testing procedure of the DMFBs is classified into structural testing and functional testing. Structural testing aims to detect the physical faults, while the functional testing focuses in identifying the faulty functional modules. One of the important testing techniques is droplet trace-based fault detection where a test stimulus droplet is moved through the testable cells, and depending on the presence of the droplet at its desired position at scheduled time, the defected cells are recognised. Such transportation of test droplets may be planned in terms of the Euler cycle or Euler path problems in an undirected or a directed graph, respectively. Sometimes, due to the unintended droplet, residual on the chip introduces particle contamination [1, 3] during assay operation that leads to physical defects afterwards. If one cell (Electrode) becomes faulty during the assay operation, even the entire assay operation may be fouled depending on the positional impact of the defective electrode(s). Thus, online and offline (e.g. Post-manufacturing) testing techniques are required to ensure system reliability and to augment the system performance [4, 5].

Now, in traditional square electrode array, we find a number of droplet trace-based testing mechanisms that are not beneficial for Triangular Electrode based Digital Microfluidic Biochip (*TEDMB*) [6] due to its routing constraints. In this paper, we have developed an Euler path based online testing algorithm for *TEDMB*. Moreover, a test planning procedure for online testing is introduced. Procedure for finding Euler circuit in a graph is presented here which runs in O $(n + e)$, where a number of vertices and edges in the graph are, respectively, denoted by n and e. In the next section, we first discuss the types of faults that may occur in biochip and the basis of online testing as well as some already existing testing algorithms for traditional DMFB. Section "Test Planning Issues for Online Testing" formulates the problem as an Euler circuit finding problem in graph theory and we discuss our algorithm in detail. In Sect. "Conclusion", a comparative study has been cited to explain the efficacy of our algorithm.

Classification of Faults in DMFB: Fault Modelling

Typically, the faults in DMFBs are of *catastrophic* and *parametric* [4, 5]. Catastrophic (i.e. hard) faults lead to a complete malfunction of the underlying system, whereas parametric (i.e. soft) faults may introduce a deviation in the system performance. Catastrophic faults have the highest priority for detection as they may cause complete breakdown of the system. Catastrophic faults can occur due to some physical imperfections such as (i) dielectric breakdown, (ii) short between a pair of adjacent electrodes, (iii) degradation of electrode performance, (iv) open circuit in the metal connections between the electrode and the control source and (v) defected

configuration of parallel plates, to name a few. Physical defects that lead to parametric faults are deviation in geometrical parameters, particle contamination due to residual effect and deviation in viscous force acting between a droplet and filler medium.

To avoid the digression of the overall performance of the system, testing of the electrodes is a crucial issue before the chip is subjected to market. Chip-level testing is an iterative procedure where a chip is tested before or after the bioassay operation (offline testing) or a parallel testing operation may be carried on simultaneously with an assay in a time-overlapped manner. Parallel scan like testing [7] (using both single and multiple droplets) and path-based testing (Hamiltonian and Euler circuit [8, 9]) belong to offline testing.

Test Planning Issues for Online Testing

Problem Formulation

One of the most important catastrophic faults is stuck-at fault where a test stimuli droplet may be stuck at a faulty cell during the droplet movement through the electrode array from an original or customised source to a sink. The detection of all test stimuli droplets by a sensing circuit placed at a droplet sink indicates that the electrodes on the path are fault free. An efficient test plan must ensure two things: (i) no conflict of the testing operation with the normal biomedical assay and (ii) guarantee of the coverage of test droplets over all the microfluidic chip cells that are available for testing.

Here, it is worth mentioning that we consider catastrophic faults detection and we assume that every catastrophic fault in the microfluidic device affects only a single cell on the chip array. Nevertheless, some faults, like electrode shorts, affect more than one cell in the microfluidic array [1]. The fundamental idea behind the graph-theoretic testing optimization approach is to formulate the 2D chip as a directed graph followed by partitioning into sub-graphs such that each partition requires one test stimuli droplet to cover the associated portion of the chip and it can be tested independent of the other parts of the chip.

Testing by traversal generally finds a path such that the test droplet can be routed through the array visiting every cell exactly once though a path connecting all available electrodes does not exist in a chip as shown in Fig. 52.1a. Although this method ensures the fault detection concerning a single electrode, it fails to identify the faults associated with electrode-short and fluidic-open faults that influence two neighbouring electrodes, e.g. as shown in Fig. 52.1b; the test droplet path $6 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 5 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 1$ is unable to detect an electrode-short present between electrodes 3 and 8. But the Hamiltonian path based tour visits each electrode just once. Thus, Hamiltonian path does not pledge to detect a fault-free microfluidic array. Therefore, a novel method for test planning is

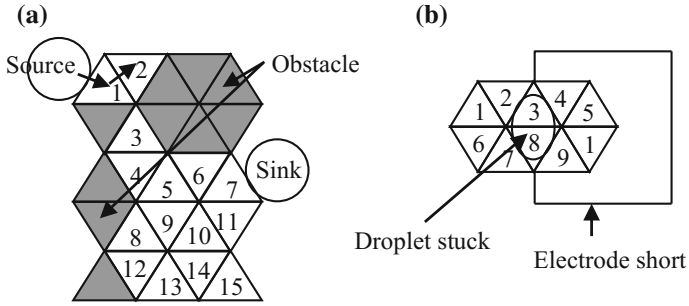


Fig. 52.1 **a** A test droplet from the source can traverse only two electrodes, but cannot reach the sink. **b** Test stimuli droplet from the source can reach the sink leading two electrodes not reachable

essential to solve this complication. Since this type of defect can happen in the DMFBs not only in the manufacturing (in the fabrication) process but also during the on-chip bioassay operations (e.g. owing to contamination of particles and relocation of electrode metal), the offline and online testing techniques are equally mandatory.

Here, we formulate the test planning problem into two well-known graph-theoretic problems, i.e. the 'Euler circuit' and 'Euler path' [8]. The basic idea behind this outlook is to construct an undirected graph representing the microfluidic chip array where each electrode acts as a vertex. Between any two vertices, i.e. two neighbouring electrodes, there exists an edge. Now, the problem of finding a test path is comparable to finding an Euler path of the graph. A flow-based test path for the test droplet can be acquired by using Euler's theorem. Such a path enables to detect the shorts between any two neighbouring electrodes. As the Euler path/circuit based test technique visits all edges exactly once, it guarantees any electrode-short fault within the chip array.

Initially, we model the chip array by employing an undirected graph $G = (V, E)$, where V is the set of microfluidic chip cells and E is the set of edge $\{u, v\}$ between any two vertices u and v if there exists a connectivity between two cells represented by u and v, respectively. Euler theorem gives us the technique to traverse every edge in the undirected graph only once. For an undirected graph, Euler path exists if the graph is in connected pattern and the connected graph has exactly two odd-degree vertices. Again, an undirected graph has Euler circuit if the graph is connected and the degree of every vertex in the connected graph must be even. Now, in general the graphical representation of the 2D microfluidic chip array has more than two vertices with odd degree in nature. So, we have to retrace some of the edges such that all the edges must be visited at least once. In G, retracing signifies an additional edge between two adjacent vertices. Therefore, our objective is to minimise the number of retracing throughout the testing, i.e. the requisite number of additional edges to Eulerise the graph.

As each electrode in *TEDMB* has three direct neighbours, in the dual graph of the array, each vertex other than the peripheral one has degree three [6]. At the

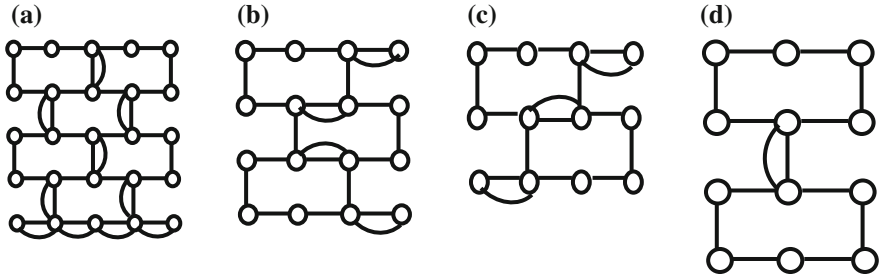


Fig. 52.2 a–d corresponding graph model and Eulerisation of different-sized 2D triangular microfluidic array

beginning of the algorithm, we identify all odd-degree vertices and transform them into even degree by adding minimum number of extra edges. For *type* 1 triangle, upward movement is restricted and for *type* 2 triangle downward movement is restricted [6]. So, there is an edge between two vertices, belonging to two adjacent rows, if and only if one vertex corresponds to a *type* 1 triangle and another vertex is associated with a *type* 2 triangle. To find an Euler circuit/path in a connected undirected graph, we use Fleury's algorithm [4, 8].

In Fig. 52.2a, b, c, d, associated graph models for different triangular 2D arrays are shown, i.e. in the 2D $m \times n$ array four cases may occur, (i) m = odd, n = odd, (ii) m = even, n = even, (iii) m = odd, n = even and (iv) m = even, n = odd. We next model the 2D microfluidic array using undirected graphs and then modify these graphs so that the condition for possessing an Euler circuit is satisfied. Applying Fleury's algorithm [4, 8], the complexity becomes $O\ (n + e)$, where n = the number of vertices and e = the number of edges. So it requires linear time.

An Efficient Eulerisation Technique for Online Testing

To ensure unpredictable faulty status, online testing is performed throughout the free cells during normal bioassay operation (say, on-chip mixing in some regions). Here, the active parts of the chip along with its guard band may be considered as obstacles to the test stimuli droplets. The graph model of a triangular biochip with obstacle in it (i.e. with mixing and other operations running within the array) is not regular structure as an $m \times n$ array. As a result, Eulerising it by adding extra dummy edges for optimising test cost may not be straightforward as eulerising the graph model of $m \times n$ triangular microfluidic array, as the odd degree nodes geometrically, may be random in nature. For more than one test droplet, we partition the graph model of the microfluidic array into sub-graphs and then manipulate them individually such that there exists an Euler circuit in each sub-graph. Now, multiple test droplets are dispensed in the chip for performing the concurrent

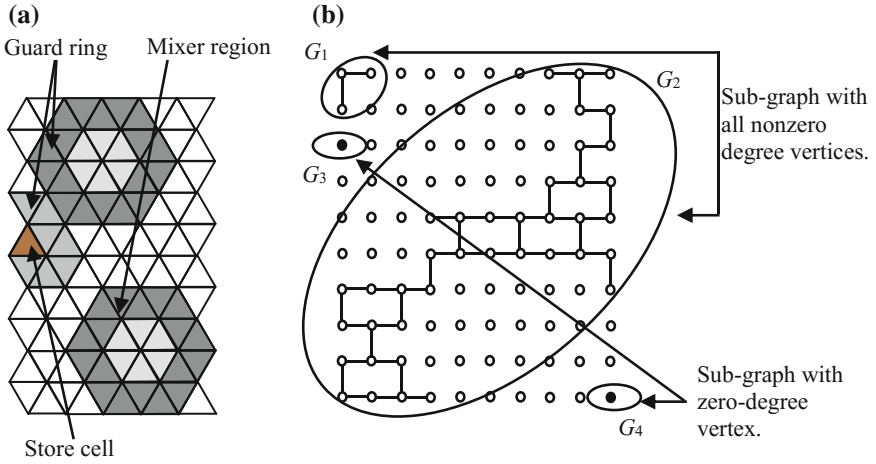


Fig. 52.3 **a** 10 × 10 chip array with obstacle, **b** Four sub-graphs G_1, G_2, G_3, G_4 constitute the graph representation of the 2D microfluidic array with mixer and storage cell along with guard bands as obstacle

edge-tour based testing in various sections of the chip array. Thus, entire testing application time is equal to the maximum testing time among all these sub-graphs.

Let us illustrate the procedure with an example. A 10 × 10 microfluidic chip array is shown in Fig. 52.3a where two mixing operations are going on and one droplet is stored in a storage cell; hence, the active cells as well as their guard band are not accessible during this period of time. In order to test the remaining free cells, we first check the connectivity of the associated dual graph. If it is a disconnected graph, we need to deploy more than one test droplet to perform the testing properly.

If there exist even number of odd degree vertices, we can pair them in such a way that all vertices in the graph have even degree [8]. A matching M in a graph $G = (V, E)$ comprises pair-wise non-adjacent edges, i.e. a common vertex is not shared by any two edges [8]. Appending additional edges between a pair of non-adjacent vertices is not as easy as adding extra edges between a pair of adjacent vertices. An extra edge in the graph model connecting two non-adjacent vertices i and j signifies an order of edges formulating a path with endpoints at node i and node j.

The Euler tour must be minimised to reduce the testing time. In a graph, this kind of problem can be modelled as the classical *Chinese Postman Problem* [10]. For doing this, instead of locating the arbitrary matching among the odd degree vertices in the graph representation of the 2D microfluidic array, we choose the closest pairs of odd degree vertices. There may be more than one connected sub-graph in the graphical representation of the microfluidic chip array with obstacle.

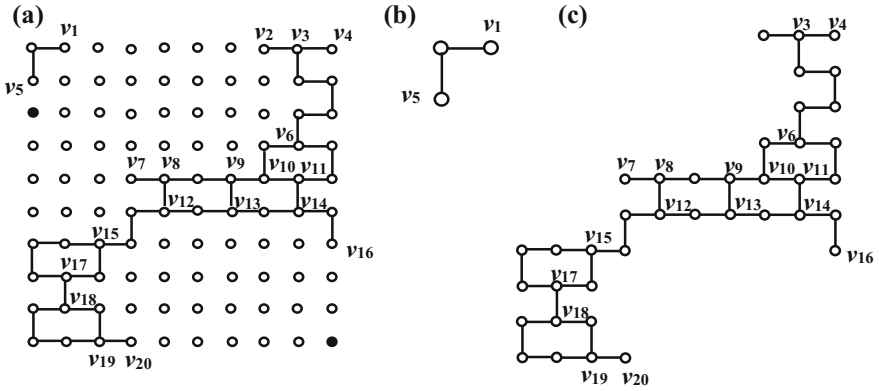


Fig. 52.4 **a** In sub-graphs G_1 and G_2, the vertices having odd degree are labelled, **b** Sub-graph G_1 with odd-degree vertices, **c** Another sub-graph G_2 with odd-degree vertices

We establish a complete, undirected, weighted graph $G = (V, E)$. Nodes of this graph represent the vertices with odd degree where microfluidic operations are not running. The least number of edges necessary to reach node j from node i, i.e. shortest Manhattan distance between node i and node j, is symbolised by $w(i, j)$. Then, a perfect matching with minimum weight is found in the graph G. Perfect matching defines the matching where all vertices of the graph are matched, i.e. every vertex of the graph is incident to exactly one edge of the matching. Minimum matching in a weighted graph usually denotes the perfect matching with minimum weight, i.e. a perfect matching with the weighted sum of all the edges in the matching minimised [9]. This kind of matching is subsequently used for the Eulerisation of the sub-graph model by adding extra edges that guarantee Eulerisation having least possible additional edges. All such sub-graphs of the graph model abstracted from the 2D microfluidic chip array with obstacle are Eulerised (Fig. 52.4).

Figure 52.3b shows that there are four sub-graphs with two sub-graphs having zero degree vertices. G_1 and G_2 are two sub-graphs where vertices have non-zero degree. But G_3 and G_4 cannot be Eulerised, since these sub-graphs have zero degree vertices. Next, as shown in Fig. 52.5a, we pair the odd-degree vertices which are close to each other in these two sub-graphs G_1 and G_2 and then we add extra edges to make these sub-graphs eulerize, as shown in Fig. 52.5b. Deploying the above-said algorithm, this is evidenced that required number of edges for different types of array configuration is different. Hence, it depends on the row–column combination of *TEDMB* as shown in Table 52.1. After performing a comparative study between the number of extra edges required to Eulerise a square electrode array and that for a *TEDMB*, we obtain Table 52.2.

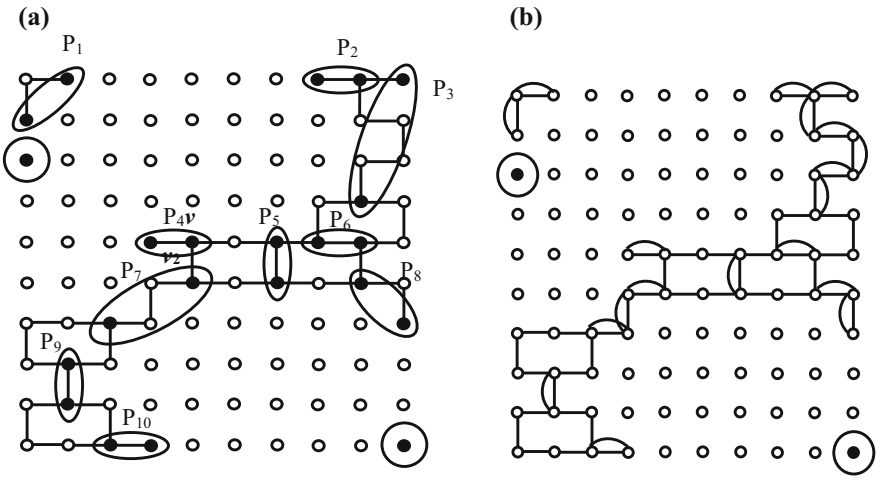


Fig. 52.5 **a** Pairing the odd-degree vertices in sub-graphs G_1 and G_2, **b** Eulerisation of the sub-graphs

Table 52.1 Number of additional edges for eulerising the graphs for different *TEDMB*s

Number of rows (m)	Number of columns (n)	Remaining columns $n_1 = (n - i)$, i is either 4 or 3 depending on whether n is even or odd, respectively	Number of repetitive clusters (d)	Number of additional edges N_a
Even	Even	$n_1 = n - 4$	$d = n_1/2$	$m + (d \times (m - 1))$
Odd	Odd	$n_1 = n - 3$	$d = n_1/2$	$(n - 1) + (d \times (m/2)) + ((d + 1) \times (m/2))$
Odd	Even	$n_1 = n - 4$	$d = n_1/2$	$m + (n - 4) + (d \times (m - 1))$
Even	Odd	$n_1 = n - 3$	$d = n_1/2$	$((d + 1) \times ((m - 2)/2)) + (d \times (m/2))$

Table 52.2 Comparative study of requisite number of additional edges between DMFB and TEDMB

Square *DMFB*	*TEDMB*
$m + n - 4$, if m = even, n = even, $m + n - 2$, otherwise	$m + (d \times (m - 1))$, for even–even
	$(n - 1) + (d \times (m/2)) + ((d + 1) \times (m/2))$, for odd–odd
	$m + (n - 4) + (d \times (m - 1))$, for odd–even
	$((d + 1) \times ((m - 2)/2)) + (d \times (m/2))$, for even–odd

Conclusion

In triangular electrode array, Hamiltonian path based testing is not possible due to the restricted movement of droplets along vertical axis (although Hamiltonian path based test is not sufficient testing, it cannot detect electrode short), but Euler circuit based testing works. This Euler circuit based test can be used both in offline and online testings. In this paper, a graph-based droplet traversal testing algorithm has been developed. Here, we have presented an Euler path based testing for online testing that is easy to implement. Some physical failures are thus far not well identified, like those faults related with power supply or deviation of microfluidic assay operation as a consequence of unknown thermal effect or environmental temperature variation [1]. Competent modelling of faults and generation of test stimuli methods are required for the testing of biochips. Although the detection of parametric faults is challenging and may lead in break down at a later stage, catastrophic faults result in the complete abnormality of the system structure, and hence, it should have the highest priority to be detected.

References

1. Chakrabarty, K., Xu, T.: Digital Microfluidic Biochips, Design Automation and Optimization. CRC Press, Taylor and Francis Group, Boca Raton, Fl, USA (2010)
2. Fair, R. B.: Digital microfluidics: is a true lab-on-a-chip possible? In: Microfluid Nanofluid, vol. 3, pp. 245–281, (2007)
3. Chakrabarty, K., Su, F.: Digital Microfluidic Biochips, Synthesis, Testing and Reconfiguration Techniques. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA (2007)
4. Xu, T., Chakrabarty, K.: Fault Modeling and Functional Test Methods for Digital Microfluidic Biochips. In: IEEE Transactions on Biomedical Circuits and Systems, vol. 3, no. 4, (2009)
5. Pasaniuc, B., Garfinkel, R., Mandoiu, I., Zelikovsky, A.: Optimal Testing of Digital Microfluidic Biochips. In: Informs Journal on Computing. *vol. 23, Issue 4, pp.* 518–529, (2010)
6. Datta, P., Dutta, A., Majumder, R., Chakraborty, A., Dhal, D., Pal, R. K.: A Technology Shift towards Triangular Electrodes from Square Electrodes in Design of Digital Microfluidic Biochip. In: IEEE Eighth International Conference on Electrical and Computer Engineering, pp. 1–4, Bangladesh (2014)
7. Xu, T., Chakrabarty, K.: Parallel Scan-Like Test and Multiple-Defect Diagnosis for Digital Microfluidic Biochips. In: IEEE Transactions on Biomedical Circuits and Systems, vol. 1, no. 2, pp. 148–158, (2007)
8. Cormen, TH., Leiserson, CE., Rivest, RL., Stein, C.: Introduction to algorithms. Prentice-Hall of India Pvt. Ltd., New Delhi (2004)
9. Sherwani, N. A.: Algorithms for VLSI Physical Design Automation. Third Edition, Academic Publishers, New York, (2002)
10. Edmonds, J., Jhonson, EL.: Matching, euler tour and the Chinese postman. Math Program, North-Holland Publishing Company, vol. 5, no. 1, pp. 88–124, (1973)

Chapter 53
Realization of Basic Gates Using Universal Gates Using Quantum-Dot Cellular Automata

Jayanta Pal, Paramartha Dutta and Apu Kumar Saha

Abstract Among the limitations of Complementary Metal Oxide Semiconductor (CMOS) technology, some serious observations are robustness and accuracy of the circuits. Nanotechnology, on the other hand, has played a vital role in finding the possible candidates as a replacement of CMOS. Quantum Cellular-dot Automata (QCA) plays a vital role as one of the most promising nanotechnology candidates to fix those limitations. This paper presented the implementation of basic gates like AND, OR, and NOT gates using universal gates such as NAND and NOR gates using QCA. To implement the basic gates, we have applied the implementations of NAND and NOR using QCA. The results of the designed circuit have been verified by using QCADesigner tool.

Keywords Quantum-dot cellular Automata · QCA · Universal gate Majority voter gate · QCADesigner

Introduction

The concept of Quantum-dot Cellular Automata (QCA) technology was proposed by Lent and Tougaw [1]. QCA is one of the emerging technologies in the field of nanotechnology which uses an array of quantum-dot cells for implementing the digital logics [2]. The ability of working in extremely small space, consuming

J. Pal (✉)
Department of Information Technology, Tripura University, Suryamaninagar, India
e-mail: jayantapal.tu@gmail.com

P. Dutta
Department of Computer and System Science, Visva-Bharati University, West Bengal, India
e-mail: paramartha.dutta@gmail.com

A. K. Saha
Department of Mathematics, NIT Agartala, Barjala, Tripura, India
e-mail: apusaha_nita@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_53

ultra-low power, makes it one of the emerging technologies concerning future computers [3]. The advantage of using QCA is discussed in [4]. The implementation of QCA circuits using majority gate and inverter gate is discussed in [5]. Along with the circuits, in [6], it is nicely described how to implement the Boolean operators and other logical functions using majority voter gate. This paper proposed a coplanar 4-dot 2-electron QCA design for universal gates and the implementations of basic gates using them.

This chapter is organized into different sections as follows: The section followed by introduction describes the concept of two-dimensional 4-dot 2-electron QCA. Next section deals with clocking in QCA. After that, the schematic design of basic gates and universal gates using QCA have been described. The proposed design and simulation result follows next two consecutive sections and final section comprises simulation setup of the proposed design.

Quantum-Dot Cellular Automata

The structure of QCA cell is shown in Fig. 53.1, which shows a QCA cell consists of four quantum dots placed at the four corners of position. The electrons reside at any two dots which can move to any neighboring dot within the cell through electron tunneling [7] maintaining the maximum possible distance. We can have only two distinct stable positions of electron pair due to the Columbic interactions between them. One position can have polarization (P) of −1 representing a logic 0 and polarization (P) of +1 representing a logic 1 in binary [8].

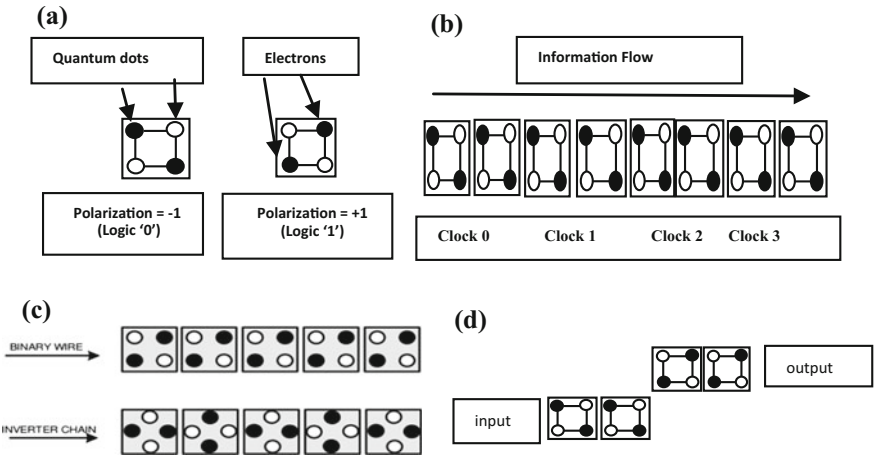


Fig. 53.1 **a** QCA cell with different polarizations, **b** QCA wire, **c** Wire with 90° and 45° cell, **d** QCA inverter

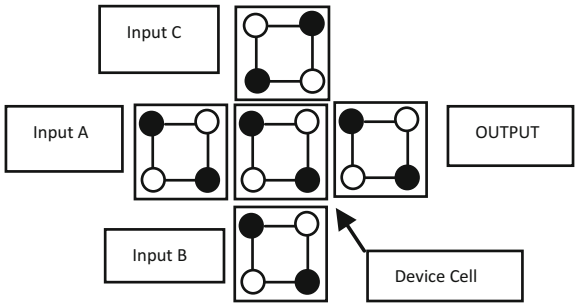


Fig. 53.2 Majority voter gate in QCA

The majority voter gate [7, 9] consists of three inputs and produces a single output which is described by the following function. The basic design is shown in Fig. 53.2.

$$\mathrm{MV}(\mathrm{A},\mathrm{B},\mathrm{C}) = \mathrm{BC} + \mathrm{CA} + \mathrm{AB} \tag{1}$$

The central cell also known as device cell senses the field imposed by the top, left, and bottom neighboring cells. It assumes the polarization equal to the polarization produced of the majority of the given three inputs. Therefore, the logical AND operation can be performed as M(a, 0, b) and the logical OR sum can be performed as M(a, 1, b) [10].

QCA Clocking

The QCA clock consists of four separate phases: 1. Switch (Clock 0), 2. Hold (Clock 1), 3. Release (Clock 2), and 4. Relax (Clock 4) as shown in Fig. 53.3. The clocking is described in detail in [1].

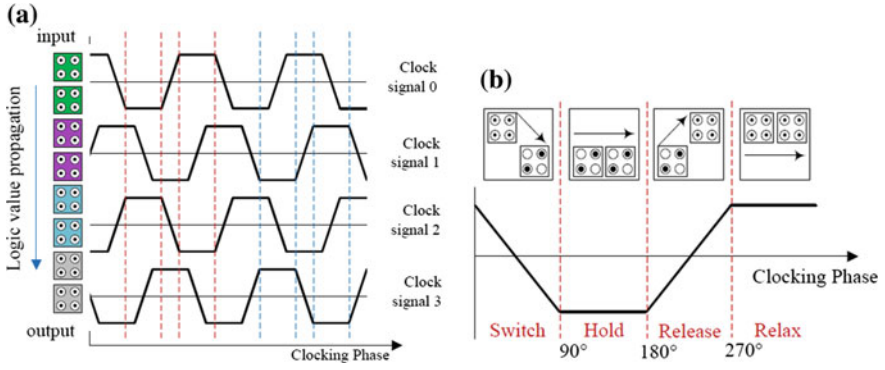


Fig. 53.3 **a** Four-phase clocking mechanism **b** QCA operations in each clock cycle

The concept of QCA clocking is discussed in [11, 12]; the inter-dot potential barrier was being modulated simultaneously for all cells in the array. The computational problem can be partitioned by subdividing an array into sub-arrays and the advantages of multiphase clocking and pipelining can be achieved. These sub-arrays (group of cells) are known as clock zones [11].

It is discussed in [1] that the probability of successful switching of all consecutive cells will decrease proportionally with the increase of wire length due to thermodynamic limitations.

QCA Gate Design

Basic Gates

Figure 53.4 shows the schematic diagram of implementation of AND, OR, NAND, and NOR gates.

Universal Gate

Figures 53.5 and 53.6 show the schematic diagram of implementation of basic gates using NAND and NOR gates, which means they can be used in place of basic gates such as AND, OR, and NOT gates.

They can be used to implement any circuit. They are often known as universal gates because we can use only one of the universal gates to implement any circuit.

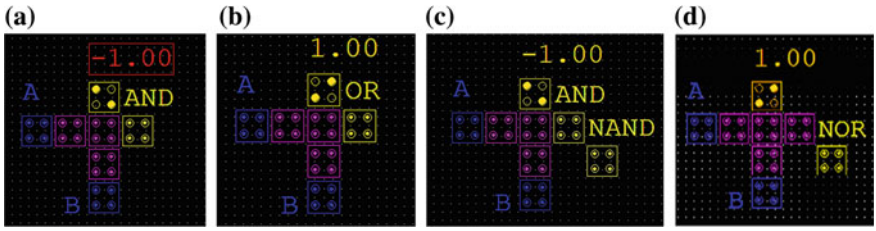


Fig. 53.4 QCA implementation of **a** AND gate, **b** OR gate, **c** NAND gate, **d** NOR gate

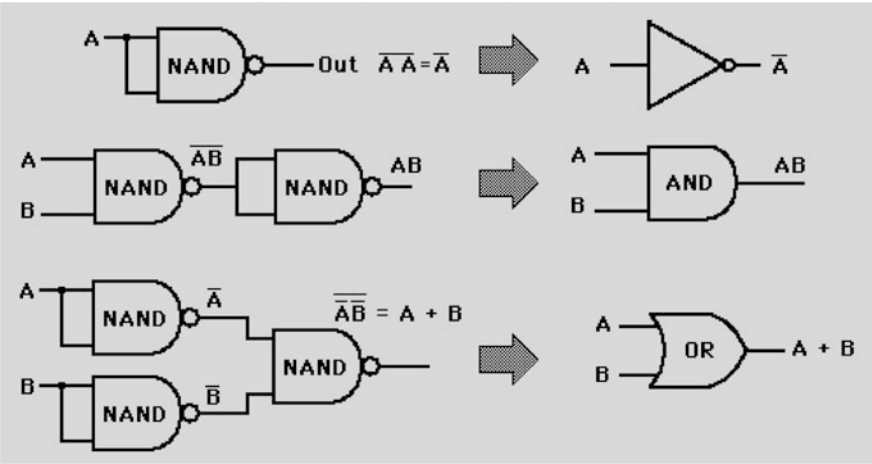


Fig. 53.5 Schematic diagram of implementation of basic gates using NAND gate

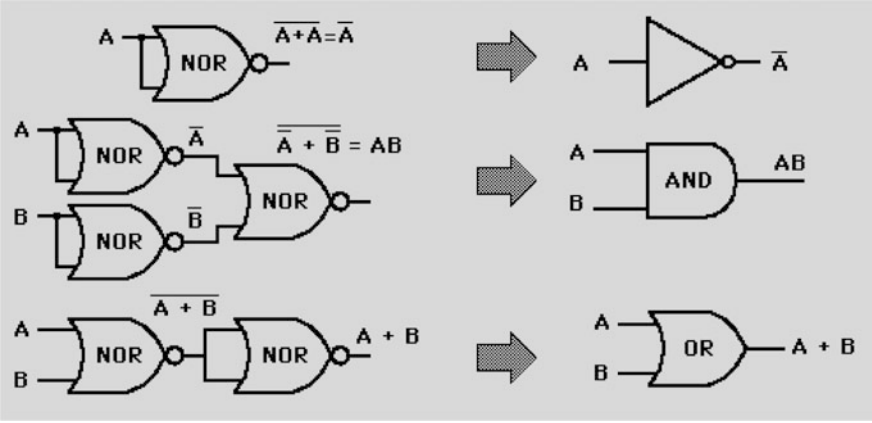


Fig. 53.6 Schematic diagram of implementation of basic gates using NOR gate

Proposed Designs and Simulation Result

Implementation of AND Gate

In Figs. 53.7a and 53.8a, the proposed design of implementation of AND gate using NAND and NOR gates, respectively, is shown.

To implement an AND gate using NAND gate, we have to use two NAND gates connected serially where the input bits are given as the input of the first gate and the generated output is shared among the inputs of the second gate, whereas, to implement the same using NOR gate, we have to use two NOR gates parallelly

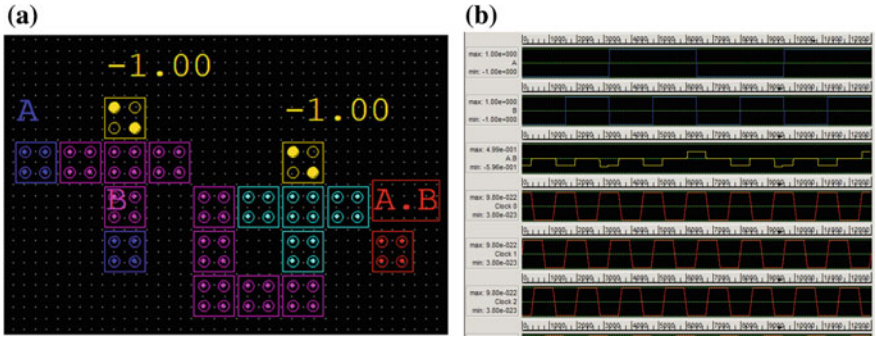


Fig. 53.7 **a** Design of AND gate using NAND gate, **b** Simulation result

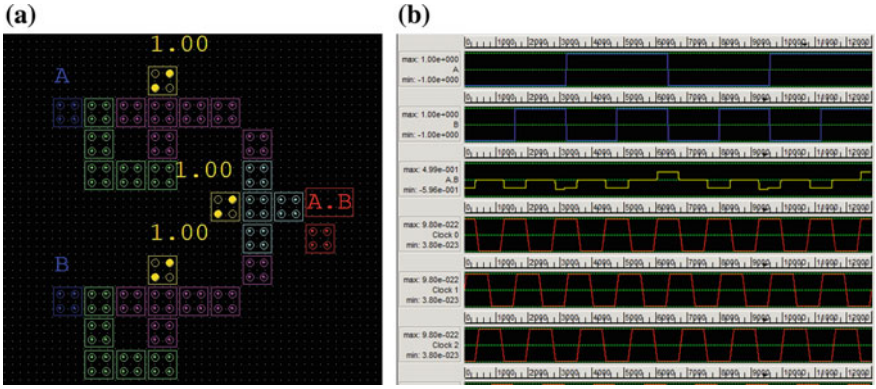


Fig. 53.8 **a** Design of AND gate using NOR gate, **b** Simulation result

followed by another NOR gate. The two separate input bits are shared among the inputs of the first NOR gate and outputs are given to the final NOR gate.

The simulation result can be observed in Figs. 53.7b and 53.8b, respectively.

Implementation of NOT Gate

The proposed design for implementation of NOT gate using NAND and NOR gates is shown in Figs. 53.9a and 53.10a, respectively.

The implementation of NOT gate is quite simple; if a NAND gate or NOR gate is used with common input, the output generated would be the inverted of the given input. The same is experienced with the proposed design.

The inverted property can be verified by the simulation result as shown in Figs. 53.9b and 53.10b, respectively.

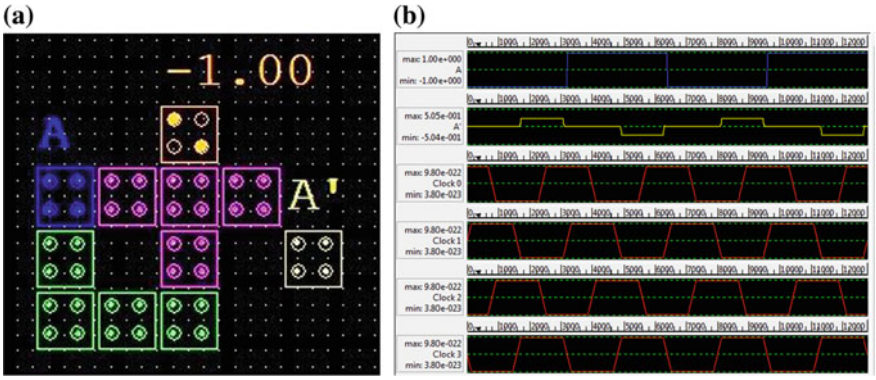


Fig. 53.9 **a** Design of NOT gate using NAND gate, **b** Simulation result

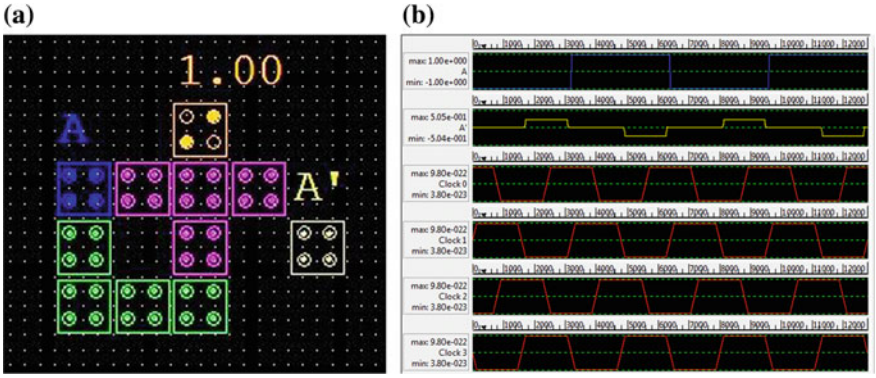


Fig. 53.10 **a** Design of NOT gate using NOR gate, **b** Simulation result

Implementation of OR Gate

Figures 53.11a and 53.12a, respectively, describe the proposed design for the implementation of OR gate using NAND and NOR gates.

To implement a OR gate using NOR gate, we have to use two NOR gates connected serially where the input bits are given as the input of the first gate and the generated output is shared among the inputs of the second gate, whereas, to implement the same using NAND gate, we have to use two NAND gates parallelly followed by another NAND gate. The two separate input bits are shared among the inputs of the first NOR gate and outputs are given to the final NAND gate.

The simulation result can be observed in Figs. 53.11b and 53.12b, respectively.

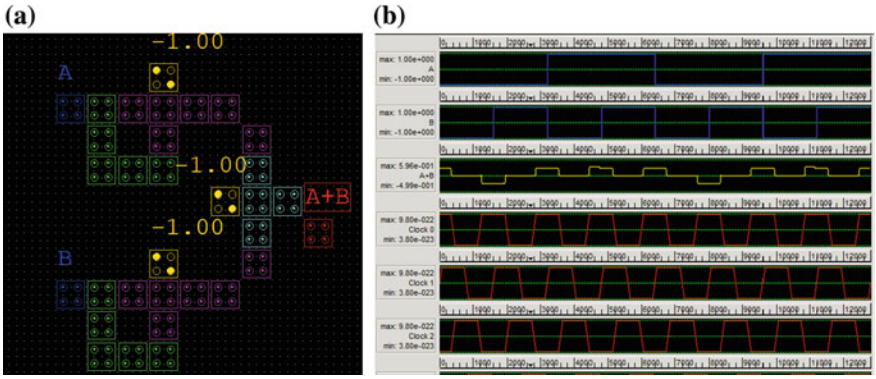


Fig. 53.11 Design of OR gate using NAND gate, **b** Simulation result

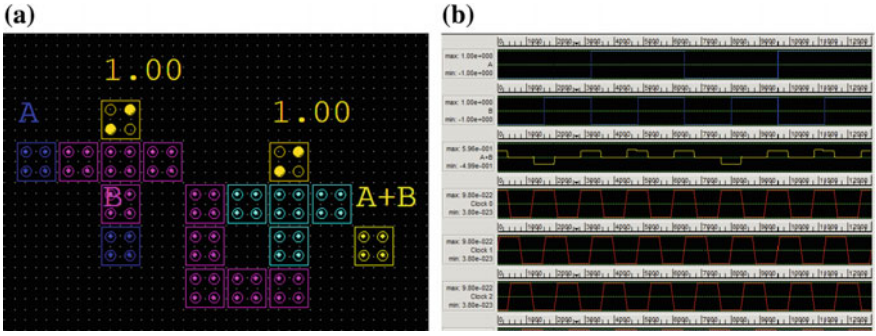


Fig. 53.12 **a** Design of OR gate using NOR gate, **b** Simulation result

Conclusions

In this paper, we have proposed single-layer QCA design for the implementation of basic gates using universal gates in QCA. The universal gates can be used to implement any circuits in the replacement of basic gates. The detail analysis of simulation-based results affirmed the reliability of the proposed design.

Simulation Setup

All the experiments have been tested and verified using QCADesigner version 2.0.3 [13].

References

1. C. Lent and P. Tougaw, "A device architecture for computing with quantum dots," Proceedings of the IEEE, vol. 85, pp. 541–557, 1997.
2. SureshRai, "Majority Gate Based Design for Combinational Quantum CellularAutomata (QCA) Circuits," 40th Southeastern Symposium on System Theory, pp. 222–224, March 16–18, 2008.
3. K. Datta, D. Mukhopadhyay and P. Dutta "Design of a 2-Dot I-Electron QCA Full Adder using Logically Reversible Half Adders", ISACC, 2015, IEEE.
4. P. Gupta et al, "Test generation for combinationalquantam.ce11ular automata (QCA) circuits," EDAA, 2006, pp. 311–316.
5. R. Farazkish, M. Azghadi, K. Navi, and M. Haghparast, "New Method for Decreasing the Number of Quantum-dot Cells in QCA Circuits," World Applied Sciences Journal, vol. 6, pp. 793– 02, 2008.
6. T. Oya, T. Asai, T. Fukui, and Y. Amemiya, "A Majority-Logic Device Using an Irreversible Single-Electron Box," IEEE Transactions on Nanotechnology, vol. 2, 2003.
7. P. D. Tougaw and C. S. Lent, "Logical devices implemented using quantum cellular automata," *J. Appl. Phys.*, vol. 75, no. 3, pp. 1818–1825, Feb. 1994.
8. Dariush Abedi, Ghassem Jaberipur, and Milad Sangsefidi, "Coplanar Full Adder in Quantum-Dot Cellular Automata via Clock-Zone-Based Crossover", IEEE TRANSACTIONS ON NANOTECHNOLOGY, VOL. 14, NO. 3, MAY 2015.
9. I. Amlani, A. O. Orlov, G. Toth, C. S. Lent, G. H. Bernstein and G.L. Snider, "Digital logic gate using quantum-dot cellularautomata," Applied Physics Letters, pol. 74, pp. 2875, 1999.
10. S. Ghosal and D. Biswas, "Study and Defect Characterization of a Universal QCA Gate", International Journal of Computer Applications (0975 – 8887) Volume 74– No.15, July 2013.
11. CaioAraujo T. Campos et al, "USE: A Universal, Scalable, and Efficient Clocking Scheme for QCA," IEEE Transactions on computer-aided design of integrated circuits and systems, vol. 35, no. 3, march 2016.
12. Mahalakshmi K S, Shiva Hajeri et al, "Performance Estimation of Conventional and Reversible Logic Circuits using QCA Implementation Platform", International Conference on Circuit, Power and Computing Technologies, 2016, IEEE.
13. K. Walus, T. Dysart, G. A. Jullien, and R. Budiman, "QCADesigner:a rapid design and simulation tool for quantum-dot cellular automata," Trans. Nanotechnology, vol. 3, no. 1, pp. 26–29, March 2004.

Chapter 54
Performance Optimization in Nonuniform Directive Arrays

A. V. L. Narayana Rao, N. Bala Ankaiah and Dharma Raj Cheruku

Abstract Antenna parameters can be optimized by considering various properties of the array. Antenna performance may be improved by reducing mutual coupling. Along with the reduction in mutual coupling between elements of the array, planar pattern and interelement spacing can be altered to optimize the performance. In this paper, arrays like circular and elliptical structures with nonuniform interelement spacing are considered. It is evident that when the variation of the interelement spacing between 0.5λ and λ is logarithmically varied in elliptical array, side lobe level has been decreased significantly though the gain is constant.

Keywords Mutual coupling · DOA · Uniform · Nonuniform spaced arrays Elliptical array

Introduction

Adaptive array performance is considerably affected by electromagnetic characteristics. Mutual coupling also affects greatly the beamwidth, the signal-to-noise ratio (SNR), resolution, and array gain [1]. This paper deals with side lobe reduction and mutual coupling effect on nonuniform array. Directional array with dipoles in elliptical and circular geometry is examined in this paper. Signal-to-noise ratio of adaptive antenna array accounts for mutual coupling among the elements and their normalized impedance. For improved directivity, elements used are dipoles in the array.

A. V. L. Narayana Rao (✉)
Department of ECE, SSNCET, Ongole, AP, India
e-mail: narayanaavlrao@gmail.com

N. Bala Ankaiah
CSIR, Bangalore, India
e-mail: bala.nunna8@gmail.com

D. R. Cheruku
Department of ECE, GIT, GITAM University, Visakhapatnam, AP, India
e-mail: dharmarajc@yahoo.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_54

Hence, compensation methods for mutual coupling and interelement spacing are required. Out of many types of compensation methods proposed in [2] for receiving antenna, receiving mutual impedance method (RMIM) [3] with nonuniform spacing with Gaussian distribution is preferred [4, 5].

Mutual Coupling Compensation

Considering an antenna array of N elements,

$$V_k = U_k + W_k, \tag{54.1}$$

where V_k = kth element terminal voltage while receiving,

Uk = Received voltage of kth antenna element, and W_k = Scattered field coupled voltage from the other antenna elements. The equation of W_k becomes [6, 7]

$$W_k = Z_t^{k1} I_1 + Z_t^{k2} I_2 + \cdots + Z_t^{k(K-1)} I_{k-1} + Z_t^{k(K-1)} I_{k+1}, \tag{54.2}$$

where Zk is the mutual impedance while receiving among elements k and i. I_i is the ith terminal current given by

$$I_i = V_i / Z_L, \quad where \;\; i = 1, 2, \ldots N$$

where Z_L = Load impedance at the terminal. Substituting Eqs. (54.2) into (54.1),

$$\begin{aligned} Vk = U_k + Z_t^{k1}\frac{V_1}{Z_L} + Z_t^{k2}\frac{V_2}{Z_L} + \cdots + Z_t^{k(k-1)}\frac{V_{k-1}}{Z_L} \\ + Z_t^{k(k-1)}\frac{V^{k+1}}{Z_L} + \cdots + Z_t^{KN}\frac{V^N}{Z_L} \end{aligned} \tag{54.3}$$

In Fig. 54.1, the received voltages at the terminal loads are given as V_1 and V_2. I_1 and I_2 are the currents through the two antennas. U_1 is the isolation voltage on antenna 1 when excited by plane wave source externally.

U_k (Uncoupled voltages) and V_k (received coupled voltage) can be given by matrix in the equation as (54.4)

$$\begin{bmatrix} U1 \\ U2 \\ \vdots \\ UN \end{bmatrix} = \begin{bmatrix} 1 & \frac{Z_t^{12}}{Z_L} & \cdots & \frac{Z_t^{1N}}{ZL} \\ \frac{Z_t^{21}}{Z_L} & 1 & \cdots & \frac{Z_t^{2N}}{Z_L} \\ \vdots & \vdots & \vdots & \\ \frac{Z_t^{N1}}{Z_L} & \frac{Z_t^{N2}}{Z_L} & \cdots & 1 \end{bmatrix} \begin{bmatrix} V1 \\ V2 \\ \vdots \\ VN \end{bmatrix} \tag{54.4}$$

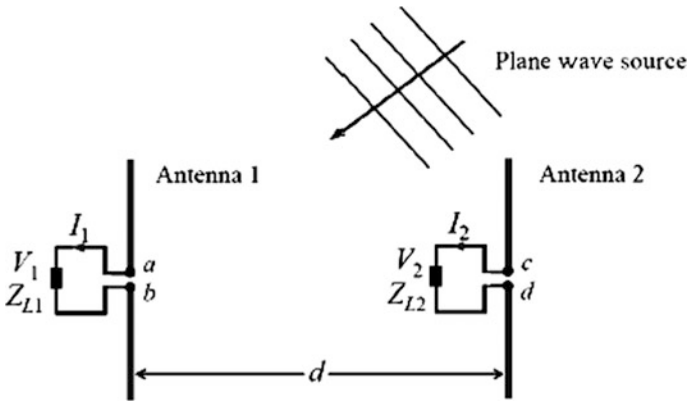


Fig. 54.1 Mutual impedance between antennas

The Receiving Mutual Impedance

Effect of coupling can be compensated by a modest approach called receiving mutual impedance method (RMIM). This mutual impedance caused by receiving current distribution in 1, 2 elements is given by Eq. (54.5):

$$Z_t^{12} = -(V_1 - U_1)/I_2 \tag{54.5}$$

In section "Optimal Size and Spacing of Elements in Array", design strategy of optimal size arrays is presented. In section "Results", simulation results in FORTRAN 90 for elliptical arrays and circular arrays with uniform and nonuniform spacing are shown. Section "Conclusion" deals with the conclusion.

Optimal Size and Spacing of Elements in Array

Elliptical Arrays

Consider an elliptical array with N elements for solution of optimal array problem. Let M be the number of interference signals along with mutually coupled signals from neighbor elements. Elevation angle (θ) is from 0 to π and azimuth angle (∅) is from 0 to 2π. Figure 54.2 shows the alignment of elliptical array.

The position of elements on x-y axis is given by

$$ri = x, \text{ y}, 0, \quad i = 1, 2, \ldots N \tag{54.6}$$

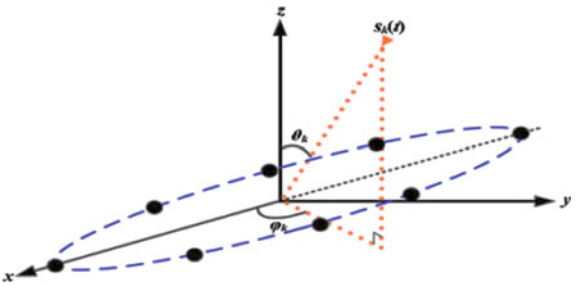


Fig. 54.2 Geometrical diagram of the ellipse

Input of array is given by

$$X(t) = \sum_{n=1}^{M} f(\Theta, \varnothing) s_n(t) v_n(k_n) \tag{54.7}$$

$f(\Theta, \varnothing)$ is the element pattern in array, and $s_n(t)$ and $v_n(t)$ are the signal and steering vectors, respectively.

Autocorrelation matrix becomes

$$R_{XX} = E\left[X(t)x^H(t)\right] = \mathrm{E}\left[\left(\sum_{n=1}^{M} s_n(t)v_n(k_n)\right)\left(\sum_{n=1}^{M} s_n(t)v_n(k_n)^H\right)\right] \tag{54.8}$$

or trigonometrically expressed as

$$[\mathrm{R_{xx}}]_{\mathrm{ab}} = \sum_{n-1}^{M} \sigma_{\mathrm{n}}^2 \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\lambda}\sin\Theta(\cos\varnothing_{\mathrm{n}}(\mathrm{x_a}-\mathrm{x_b})+\sin\varnothing_{\mathrm{n}}\left(\mathrm{y_a}-\mathrm{y_{b)}}\right)} \tag{54.9}$$

The expected value of above equation is given by

$$E[R_{xx}]_{ab} = \int_0^{2\pi} \sum_{n-1}^{M} \sigma_{\mathrm{n}}^2 \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\lambda}\sin\Theta(\cos\varnothing_{\mathrm{n}}(\mathrm{x_a}-\mathrm{x_b})+\sin\varnothing_{\mathrm{n}}\left(\mathrm{y_a}-\mathrm{y_{b)}}\right)} \frac{d\varnothing_n}{2\pi} \tag{54.10}$$

For evaluating integral substitutions, it can be considered as

$$(\mathrm{x_a}-\mathrm{x_b}) = R_{ab}cos\varnothing_{ab} \tag{54.11}$$

$$\left(\mathrm{y_a}-\mathrm{y_{b)}}\right) = R_{ab}sin\varnothing_{ab} \tag{54.12}$$

Using trigonometric formula

$$\mathrm{Cos}\,(\mathrm{A}-\mathrm{B}) = \mathrm{Cos}\,(\mathrm{A})\,\mathrm{Cos}\,(\mathrm{B}) + \mathrm{Sin}(\mathrm{A})\,\mathrm{Sin}(\mathrm{B}) \tag{54.13}$$

The integral (54.10) becomes

$$\int_0^{2\pi} \mathrm{e}^{-\mathrm{j}\frac{2\pi}{\lambda}R_{ab}\,\mathrm{sin}\Theta\cos(\varnothing_\mathrm{n}-\varnothing_\mathrm{ab})}\frac{d\varnothing_n}{2\pi} \tag{54.14}$$

Bessel function of the first-order n is

$$j_n(x) = \frac{j^{-n}}{2\pi}\int_0^{2\pi} e^{j(xcos\varnothing+n\varnothing)}d\varnothing \tag{54.15}$$

Integral of autocorrelation can be written using 54.14 and 54.15

$$[S_x]_{ab} = E[R_{xx}]_{ab} = \left[\sum_{n-1}^{M}\sigma_n^2\right]j_0\left(\frac{2\pi}{\lambda}R_{ab}sin\Theta\right) \tag{54.16}$$

The above matrix in Eq. (54.16) is dependent on the interferences and mutual coupling signals. To optimize the array from this problem, optimizing algorithms are helpful. In this paper, perturbation of element position is done by Gaussian distribution with *norm* μ, *variance* δ given by

$$\rho = f(x/\mu,\delta) = \frac{1}{\delta\sqrt{2\pi}}\,\mathrm{e}^{\frac{-(\mathrm{x}-\mu)^2}{2\delta^2}} \tag{54.17}$$

The spacing is chosen such that the perturbation of the average between elements is order of the 0.001λ. If the perturbation causes interelement spacing closer to 0.25λ or nearby λ, new perturbation is selected. This simulates the nonuniform spacing in array with different spacings.

The elements in circular or elliptical array are with an average distance of 0.25λ, but not more than 0.75λ. The tradeoff between this and largely spaced arrays is used for minimizing side lobe and mutual coupling.

Results

Simulation results using FORTRAN 90 for side lobe reduction by mutual coupling compensation and optimized spacing by Gaussian distribution are analyzed in this session.

Case1: Nonuniform array with nonsymmetrical (Gaussian) spacing:

This section describes the simulation results where array factor was used as algorithm for evaluating the pattern in dB with azimuth angle. Spacing among the antenna elements is perturbed nonuniformly from 0.5λ for 8-element antenna spacing variation of +0.03 and −0.01 respectively, in either direction from the center of the array. The results are shown in Figs. 54.3 and 54.4, respectively.

Consecutive element spacing is +0.01 and −0.01, respectively, for 16-element antenna from the center element towards the edges of the array. The results are shown in Figs. 54.5 and 54.6 for elliptical and circular arrays.

From the simulations, Figs. 54.3, 54.4, 54.5, and 54.6, elliptical array has lower side lobe level compared to circular array. The details of Gaussian spacing for circular and elliptical arrays of 8 and 16 elements are given in Table 54.1. The table also shows the comparison of results of elements which generate considerable side lobes and half-power bandwidth (HPBW).

Observations:

The array factor increases with spacing between the elements. Referring to the result obtained by the different number of elements, the gain is constant when the

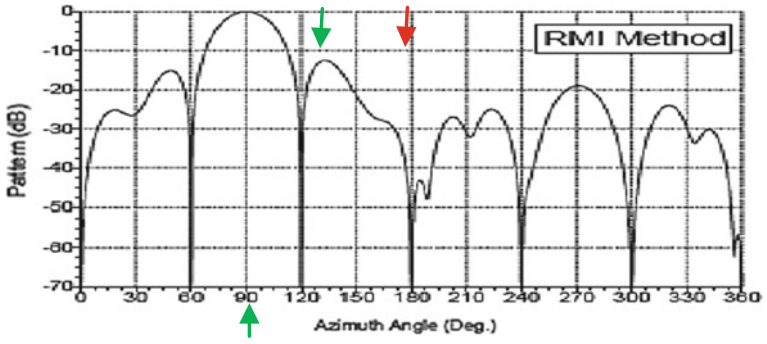


Fig. 54.3 8-element elliptical array with desired and probe signals 90° and 180°

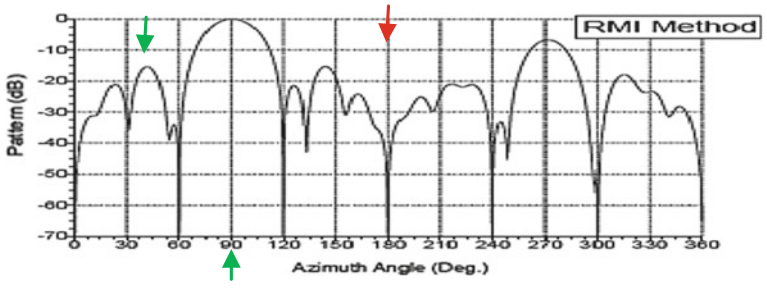


Fig. 54.4 8-element circular array with desired and probe signals 90° and 180°

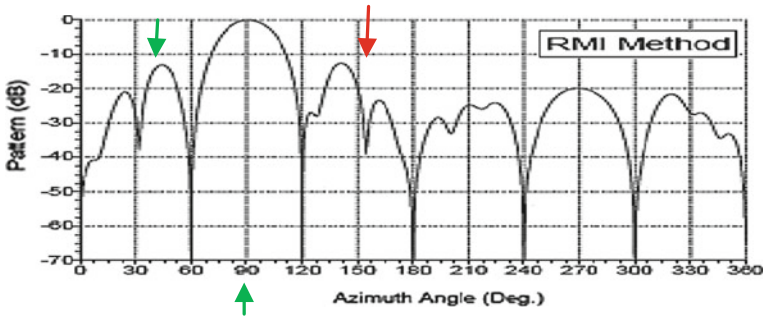


Fig. 54.5 16-element elliptical array with desired and probe signals 90° and 150°

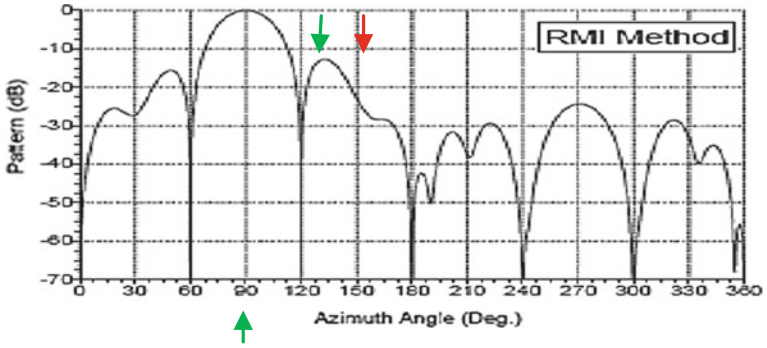


Fig. 54.6 16-element circular antenna with desired and probe signals 90° and 150°

Table 54.1 Comparison of nonuniform spaced arrays

Element (*n*th)	d = 0.43 to 0.58 (N=16)		d = 0.46 to 0.59 (N = 8)	
	HP BW	Max. side lobe level (dB)	HP BW	Max. side lobe level (dB)
2	42.1°	−25	53.4°	−25
4		−29		−24
8		−20		−25

spacing of element is increasing toward lambda. But half-power bandwidth reduces when spacing increases.

Comparing all the results from Figs. 54.3, 54.4, 54.5, and 54.6, the number of side lobes is strictly increased when the spacing between elements is increased. Figure 54.5 shows 19 dB gain for 0.1λ at one pair of elements. At the same time, main lobe width is decreased. Results show that spacing between elements has an effect on array pattern.

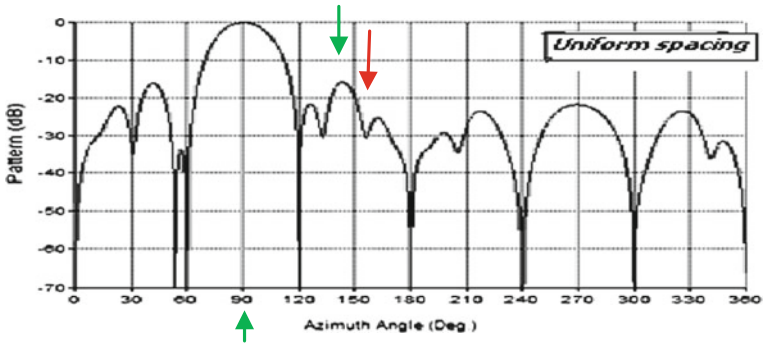


Fig. 54.7 16-element elliptical array with uniform spacing with desired and probe signals 90° and 150°

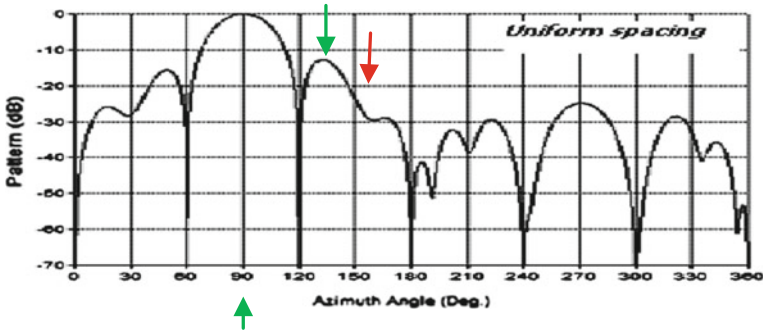


Fig. 54.8 16-element circular antenna of uniform spacing (desired and probe signals 90° and 150°)

Case2: Uniform spacing Arrays

Antenna arrays of geometrically elliptical and circle shaped with 16 elements with the uniform spacing among the elements are illustrated here. Simulation results are shown in Figs. 54.7 and 54.8. It is clearly elevated that elliptical array has low side lobe level when compared with its counterpart. The elliptical array has magnitude −19 dB for side lobe, whereas circular antenna has −13 dB.

Comparison of the suppressed probe signals at 30° and 150° in uniform and nonuniform elliptical arrays is shown in Fig. 54.9. Nonuniform spacing is computed logarithmically.

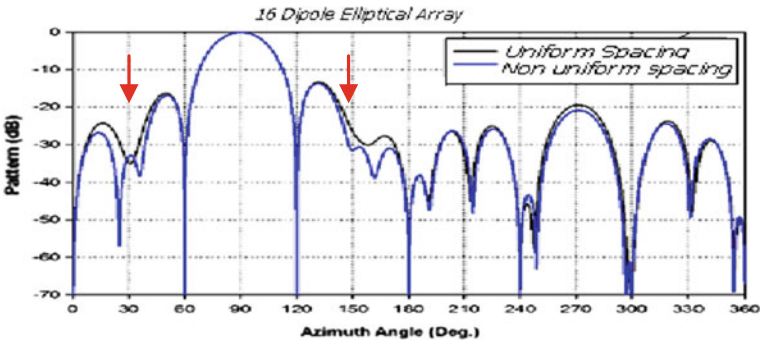


Fig. 54.9 Jammer suppression at 30° and 150° in uniform and nonuniform elliptical array

Conclusion

This paper presents reduction in side lobe with uniform and nonuniform spacing of elements. The numerical results show that the logarithmic (nonlinear) array geometry provides with best values for low side lobe levels.

Conclusion from the results is that the HPBW increases according to the number of the elements and their spacing. Also, the gain increases proportionally with the number of elements. The result obtained is evident for good performance in the antenna at the maximum gain of 19 dBattained when the element spacing is equal to 0.1λ in nonuniform spaced array. Furthermore, increasing the interspacing of elements is not suitable for the application to improve the system performance when the degradation is mainly caused by interference or jamming.

References

1. A.V.L. Narayana Rao, N. Balaankaiah, and D.R. Cheruku, 2016, "Reduction of Mutual Coupling Effect in Adaptive Arrays", © Springer India, Lecture Notes in Electrical Engineering vol. 372.
2. Inder J. Gupta and Aharon A. Ksienski, Fellow, IEEE, "Effect of Mutual Coupling on the Performance of Adaptive Arrays", IEEE Transactions on Antennas and Propagation VOL. Ap-31, NO. *5, September 1983785.*
3. P. Ioannidis and C. A. Balanis, 2005: "Uniform circular arrays for smart antennas," IEEE Antennas and Propagation Magazine, vol. 47, pp. 192–206, http://dx.doi.org/10.1109/MAP.2005.1589932.
4. T. T. Zhang, H. T. Hui, and Y. L. Lu, "Compensation for the mutual coupling effect in the ESPRIT direction finding algorithm by using a more effective method," *IEEE Transactions on Antennas and Propagation*, vol. 53, no. 4, pp. 1552–1555, 2005.
5. H. T. Hui, H. P. Low, T. T. Zhang, and Y. L. Lu, "Receiving mutual impedance between two normal-mode helical antennas (NMHAs)," *IEEE Antennas and Propagation Magazine*, vol. 48, no. 4, pp. 92–96, 2006.

6. H.-S. Lui, H. T. Hui, and M. S. Leong, "A new calculation method of the receiving mutual impedance for linear antenna array," in *Proceedings of Asia Pacific Microwave Conference (APMC'08)*, Macau, China, December 2008.
7. Ahmadi, N.; Neyestanak, A.; Dawes, R., 2008: "Elliptical array antenna design based on particle swarm method using fuzzy decision rules". In: 24th Biennial Symposium on Communications, Kingston, ON, pp. 352–355. http://dx.doi.org/10.1109/BSC.2008.4563274.

Chapter 55
Stepped Impedance Resonator Based U-Shaped Two-Element Monopole Microstrip Antenna Array

Rakita Shaw, Pritam Roy, Pankaj Sarkar and Goutam Saha

Abstract This paper presents stepped impedance resonator (SIR) based U-shaped two-element monopole microstrip antenna array. The stepped impedance resonator has been used to achieve a wideband response. The separation between the array elements has been optimized using fractal structures. Microstrip line feed has been used as the feeding network for the antenna array. The proposed design achieves the operating range of 800–1020 MHz (fractional bandwidth 24%) with a peak gain of 5.7 dBi at 1020 MHz. The antenna has been simulated using CST Microwave Studio (CST-MWS). The analytical characteristic of the proposed antenna agrees well with the simulated response.

Keywords Stepped impedance resonator (SIR) · Microstrip · Monopole Fractional bandwidth

Introduction

There is a tremendous demand for microstrip antennas in numerous fields ranging from mobile communications, satellite communications, and radio and television broadcasts to integrated antennas. Several advantages of microstrip antenna have led to this wide usage. Some of them are small size, easy fabrication, lightweight, and low cost. A key point of concern for microstrip antennas is its limited band-

R. Shaw (✉) · P. Roy · G. Saha
Department of IT, NEHU, Shillong, India
e-mail: rakita.shaw91@gmail.com

P. Roy
e-mail: pritam5791.roy@gmail.com

G. Saha
e-mail: dr.goutamsaha@gmail.com

P. Sarkar
Department of ECE, NEHU, Shillong, India
e-mail: pankajsarkar111@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_55

width keeping a reasonably high gain and affordable size especially in the MHz frequency range. Different works have been done on microstrip antenna arrays as presented in [1–8]. In [1], a quarter-wave U-shaped patch antenna has been designed with two unequal arms and shorted wall to obtain broadband response. The unequal arms were used to get two resonant modes at 900 and 1800 MHz which lead to a wide impedance bandwidth. A novel 2 × 2 microstrip antenna array has been presented in [2]. A tooth-like slot patch has been used as the array element. This antenna array design provides a gain of 14–16.3 dBi at 4.8–6.1 GHz. The design of planar slot array antenna for dual-band operation covering the bandwidth specification for LTE (2.5–2.7 GHz) and WiMAX (3.3–3.7 GHz) has been proposed in [3]. This design provides peak gains of 13.9 and 14.1 dBi across 2.5 and 3.5 GHz bands, respectively. In [4], various array configurations of proximity-fed gap-coupled ring microstrip antennas have been discussed. In a 3 × 3 array of ring patches, rectangular slot is cut on the edges of the patch which is gap-coupled along x-axis. Then, cuts are made inside the patches which are gap-coupled along x- and diagonal axes. This antenna structure shows a bandwidth of more than 500 MHz and a peak gain of around 10 dBi. Parasitically coupled feed reduces the size of the array as illustrated in [5], where a four-element microstrip array is directly fed through a feeding network consisting of T-junctions and quarter-wave transformers. A wideband antenna array has been designed in [6] having the operating range of 2.55–6.1 GHz using open-ended quarter-wavelength slot antenna elements. In [7], aperture-coupled, L-shaped feeding structure has been used to achieve a varying gain of 10.3–11.52 dBi over a bandwidth of 11.6–15 GHz. A microstrip antenna array design of frequency range 11.8–12.8 GHz having a maximum gain of 17 dB has been reported in [8].

This paper presents a novel configuration of two-element monopole microstrip antenna array. U-shaped SIR is used as the array element. The proposed array design achieves a wideband response and has been realized using FR4 substrate having dielectric constant of 4.4 and height 1 mm. The dimension of the designed antenna is 10 × 11.5 cm^2. The antenna design procedure is presented in section "Proposed Antenna Design". Experimental results are subsequently reported and discussed upon in section "Experimental Results".

Proposed Antenna Design

The layout of the proposed antenna array is depicted in Fig. 55.1. The antenna is designed on FR4 substrate of height 1.0 mm. The U-shaped array element is configured as a stepped impedance resonator (SIR) so that the first three resonant modes appear within the passband making it wide. The SIR is designed at the center frequency of 1 GHz. The SIR consists of two sections: a λ/2 section has low impedance, Z_1, in the center and two identical λ/4 sections having high impedance, Z_2, as depicted in Fig. 55.2. The impedance ratio $R_z = Z_2/Z_1$ can be adjusted to

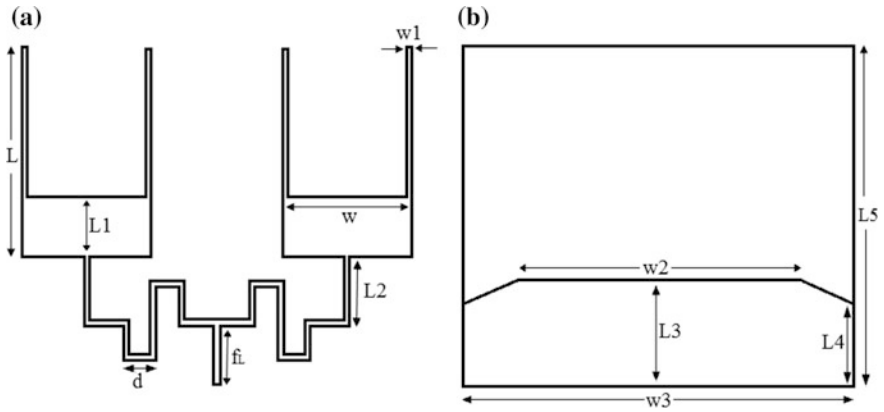


Fig. 55.1 Configuration of the proposed antenna array **a** Top layer, **b** Bottom layer

have the first three resonant modes in the desired frequency band using Eqs. (55.1) and (55.2) [9].

$$\frac{f_1}{f_0} = \frac{\pi}{2\tan^{-1}\sqrt{R_z}} \tag{55.1}$$

$$\frac{f_2}{f_0} = \frac{\pi}{\tan^{-1}\sqrt{R_z}} - 1 \tag{55.2}$$

For the three resonant modes to be $f_0 = 800$ MHz, $f_1 = 1000$ MHz, and $f_2 = 1200$ MHz, R_z is found to be 9.47 from Fig. 55.3.

A single U-shaped SIR structure gives a gain of 3 dBi, 3.9 dBi, 4.4 dBi, and 4.7 dBi at 800 MHz, 900 MHz, 980 MHz, and 1020 MHz, respectively. To improve the gain of this structure, an antenna array has been designed using two U-shaped SIR elements. The separation distance between the two array elements is in the order of λ/6. The array factor is found to be

$$AF = \frac{2\pi}{3}\cos\theta \tag{55.3}$$

At λ/6 element spacing, the nulls can be found in the direction

$$\theta = \pm\frac{\pi}{2} + 2n\pi \tag{55.4}$$

The feed length between the U-shaped SIR elements is in the order of λ at the center frequency of the SIR. To reduce the size of the antenna array as well as to maintain the symmetry of the design λ/2 separation, distance has been implemented in the form of S-shaped fractal structure as shown in Fig. 55.4. The unit element of the fractal has a length of d. Hence, the Euclidean distance reduces to 4d, although

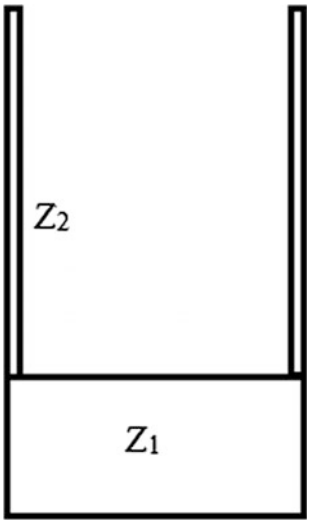


Fig. 55.2 U-shaped SIR

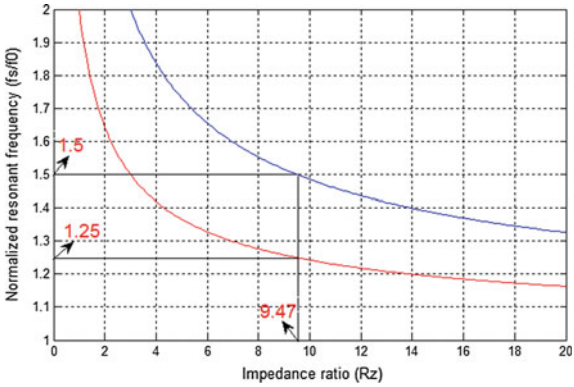


Fig. 55.3 Impedance ratio versus normalized resonant frequency

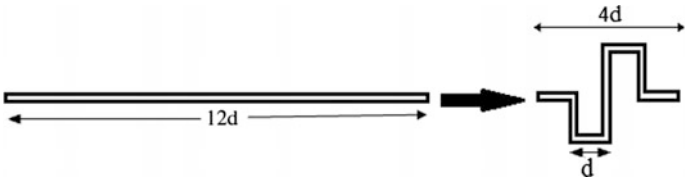


Fig. 55.4 S-shaped fractal

Table 55.1 Dimensions of the proposed antenna array (in millimeters)

Parameter	L	L1	w	w1	L2	D	f_L	w2	w3	L3	L4	L5
Value	70	15	36	1	25	7.1	15	60	100	37	34.5	115

the overall electrical length is 12d. The length of the ground plane is optimized to achieve better S-parameter characteristics. At the edges, the ground plane is cut for impedance matching purpose. The dimensions of the proposed antenna design are given in Table 55.1. The antenna array elements are fed through 50 Ω microstrip lines.

Experimental Results

The full-wave electromagnetic software CST-MWS is used for simulation. Figure 55.5 represents the simulated S_{11} response of the designed antenna array. It can be seen that the proposed antenna has the operating band from 0.8 to 1.02 GHz. The far-field radiation patterns of the designed antenna array at four different frequencies, 800 MHz, 900 MHz, 980 MHz, and 1020 MHz for $\varphi = 90°$ and $\varphi=0°$, are shown in Figs. 55.6 and 55.7, respectively. The far-field gain is found to be 3.8 dBi, 4.5 dBi, 5.3 dBi, and 5.7 dBi at 800 MHz, 900 MHz, 980 MHz, and 1020 MHz, respectively. Thus, proposed antenna array design shows approximately 22% improvement in gain over single element antenna structure.

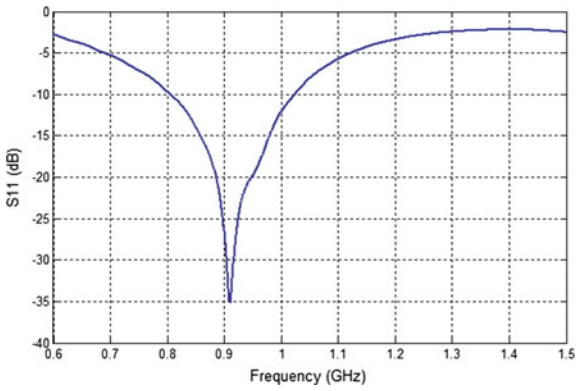


Fig. 55.5 Simulated S_{11} response of the proposed antenna design

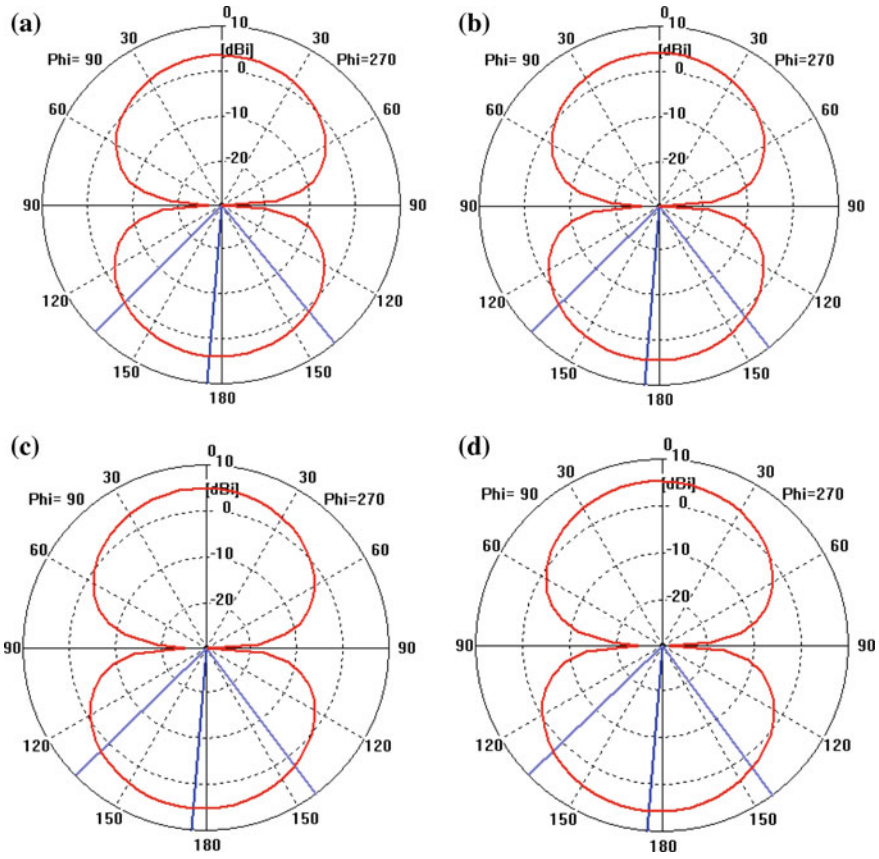


Fig. 55.6 Far-field radiation patterns for $\varphi = 90°$ of the proposed antenna array at **a** 800 MHz, **b** 900 MHz, **c** 980 MHz, and **d** 1020 MHz

Conclusion

This research paper described a U-shaped two-element SIR based monopole antenna array for broadband operation. The antenna array exhibits a fractional bandwidth of 24% (800–1020 MHz) with broadside far-field radiation patterns and a peak gain of 5.7 dBi at 1020 MHz.

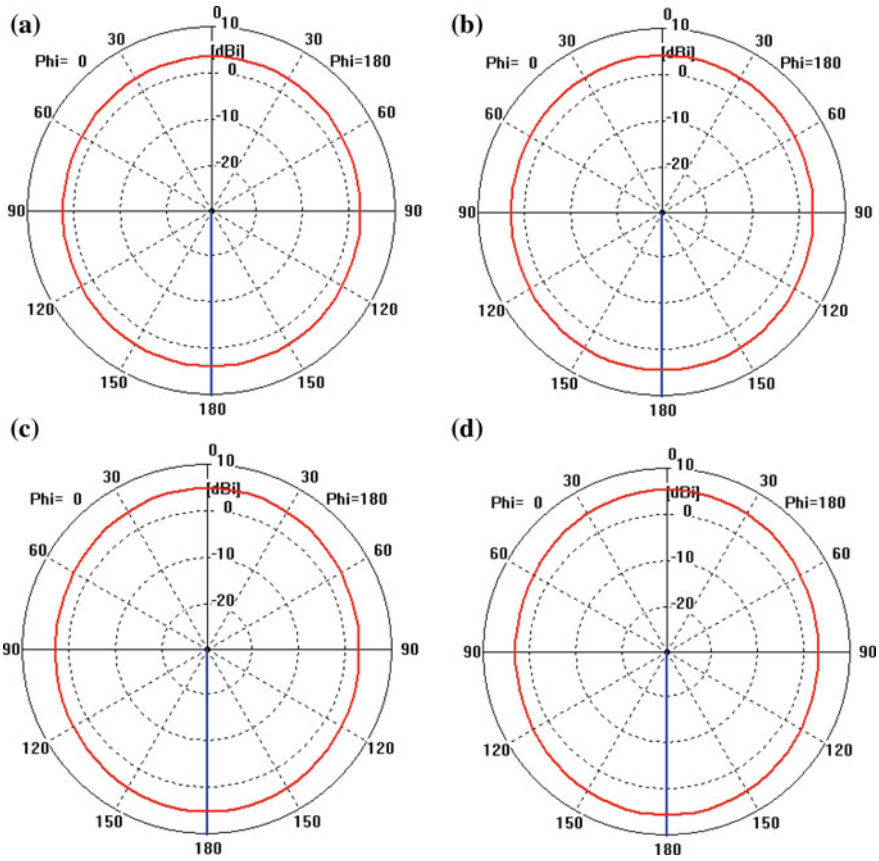


Fig. 55.7 Far-field radiation patterns for $\varphi = 0°$ of the proposed antenna array at **a** 800 MHz, **b** 900 MHz, **c** 980 MHz, and **d** 1020 MHz

References

1. Yong-Xin Guo, Kwai-Man Luk, Kai-Fong Lee and Ricky Chair, "A quarter-wave U-shaped patch antenna with two unequal arms for wideband and dual-frequency operation", IEEE Transactions on Antennas and Propagation, vol. 50, no. 8, pp. 1082–1087, August 2002.
2. H. Wang, X. B. Huang, D. G. Fang, and G. B. Han, "A Microstrip Antenna Array Formed by Microstrip Line Fed Tooth-Like-Slot Patches", IEEE Transactions on Antennas and Propagation, vol. 55, no. 4, pp. 1210–1214, April 2007.
3. Jui-Han Lu and Yen-Hung Liu, "Novel dual-band design of planar slot array antenna for 4G LTE/WiMAX access points", Microwave and Optical Technology Letters (Wiley), vol. 54, no. 5, pp. 1193–1196, May 2012.
4. Amit A. Deshmukh, "Broadband proximity fed ring microstrip antenna arrays", International Journal of Electronics and Communication (Elsevier), vol. 68, pp. 710–716, February 2014.
5. Sai Ho Yeung, Alejandro Garc´ıa-Lampérez, Tapan Kumar Sarkar and Magdalena Salazar-Palma, "Comparison of the Performance Between a Parasitically Coupled and a

Direct Coupled Feed for a Microstrip Antenna Array", IEEE Transactions on Antennas and Propagation, vol. 62, no. 5, pp. 2813–2818, May 2014.
6. Marko Sonkki, Douglas Pfeil, Veikko Hovinen and Kapil R. Dandekar, "Wideband Planar Four-Element Linear Antenna Array", IEEE Antennas and Wireless Propagation Letters, vol. 13, pp. 1663–1666, 2014.
7. Huayan Jin, Kuo-Sheng Chin, Wenquan Che, Chih-Chun Chang, Hua-Juan Li, and Quan Xue, "A Broadband Patch Antenna Array With Planar Differential L-Shaped Feeding Structures", IEEE Antennas and Wireless Propagation Letters, vol. 14, pp. 127–130, 2015.
8. Rahyanditya Ilham and Adit Kurniawan, "Design and Implementation of Microstrip Antenna Array on Ku-Band for Satellite TV Reception", 2015 International Telecommunication Networks and Applications Conference (ITNAC), IEEE, 2015.
9. M. Makimoto, S. Yamashita, "Microwave Resonators and Filters for Wireless Communication", Springer, 2001.

Chapter 56
Spatial Array of Microwave Sensors for IoT-Based Wireless Connectivity

Seyi Stephen Olokede and Babu Sena Paul

Abstract Spatial array of microwave sensors for IoT-based wireless connectivity is presented. The traditional challenges of poor input impedance matching associated with small antenna are analytically characterized using the many available formulae based on a novel 2 × 2 excitation network. Alternative microwave sensor solution designed at originally known low data throughput IEEE 802.11x standard was previously investigated to support multichannel bandwidth capacity, and now examined for robust link budget to provide complementary leverage for IoT-based applications.

Keywords Big data · Insertion loss · Integrated circuit · Latency
Parasitic effect · System-on-chip

Introduction

It is anticipated that internet of things (IoTs) would have grown considerably to become internet of all things in the next few decades. The growth is driven by the recent increase of smart homes, smart wearable devices, smart cars, smart cities, etc. Fundamentally, IoT-based technology requires an efficient information management system for low-cost data acquisition technique, effective data mining, process, collection, and transmission. In most cases, IoT ecosystem, usually consisting of thousands of interconnected sensors, laptops, and RFID tags, must be robust enough to generate, collect, communicate, and store these generated data securely without distortion or data loss. The transmission of these data to the local or cloud storage system must be through low-power wireless connectivity. Existing IEEE 802.11x wireless standard has been disadvantaged for highly densified IoT inter-

S. S. Olokede (✉) · B. S. Paul
Department of Electrical and Electronic Engineering Technology,
University of Johannesburg, Johannesburg, South Africa
e-mail: solokede@gmail.com

B. S. Paul
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_56

connection networks due to its power-hungry architecture, limited link budget, set up difficulty, and its cost ineffectiveness. Nonetheless, availability and adaption of its existing infrastructure as well as its user-friendliness could make it a complimentary alternative as heterogeneous interconnected wireless connectivity for IoT, in particular, when an efficient low-power alternative radio connectivity is provided. In [1], we presented a novel microwave sensor whose form factor does not depend on the radiation wavelength, but rather on the lumped elements. The sensor was reducible to smaller aperture size with respect to the resonance. Because of this capability, the structure becomes miniaturized to mitigate its power-hungry propensity. In this work, we intend to focus on the possibility of enhancing the link budget rather than estimating the power-hunger property of the proposed design or to assess to what extent the hunger has been mitigated. We intend to consolidate on many of the advantages of the structure as identified in [1, 2]. We have demonstrated that the structure when properly matched could deliver close to 17 bandwidth channels of 20 MHz each based on the IEEE 802.11 U-NII-3 standard, with link budget of more than 9 dBi. We assume that a better link budget could be achieved if an array is implemented, provided good impedance matching can be attained. We therefore out of necessity investigate the input impedance matching as a precursor to establishing our set goal. We subsequently examine the effect of efficiently coupled proximity coupled array to enhancing the link budget.

Problem Formulation

Figure 56.1 depicts a typical IoT technology. The data are collected by the interconnected sensors, which in turn are transmitted to low-power, high-speed, large-scale, and highly interconnected microcontroller (servers). Wireless networks such as BLE, Bluetooth, ZigBee, Wi-Fi, etc. are normally used to connect these thousands of sensor nodes to the gateway. However, the transmission of data from the transmitter to the cloud requires high-speed and long-range wireless internet connectivity depending on the systemic requirement. This is the focus of this work.

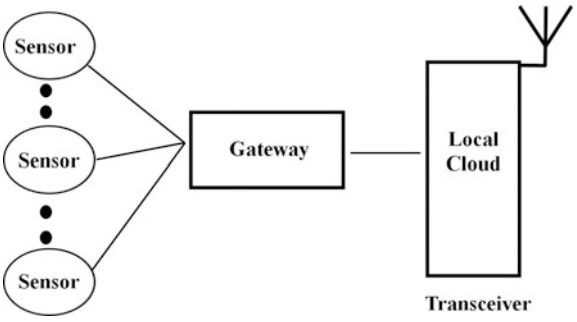


Fig. 56.1 Typical layout of IoT system

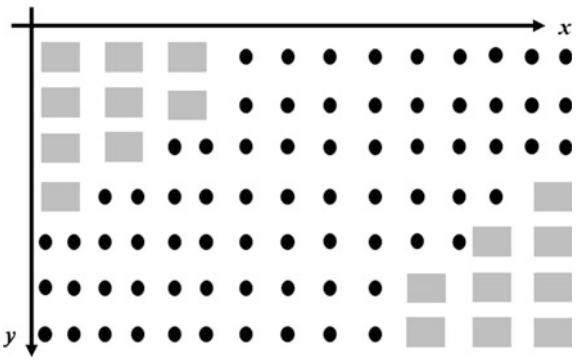


Fig. 56.2 Spatial array of microwave sensor

Input Impedance Matching

In Fig. 56.2, the spatial array configuration to enhance both the channel bandwidth and the link budget in order to provide a robust platform for high data throughput is depicted. We assume that the depicted $m \times n$ array elements are optimally power coupled by the excitation network. We provided $x \times y$ excitation network to ensure that optimal coupling is guaranteed. We ensure that each element is identical and is at equidistance position in both x- and y-directions. We further map the array into sizeable rectangular sub-array lattices, such that array pattern of $\lambda/2$ interelement spacing is maintained in both x- and y-planes. We exited each sub-array lattice with an excitation source to ensure that each element is optimally power coupled. On excitation, both magnetic and electric flux densities are created on the plane perpendicular to the plane where the array pattern is laid. The flux densities in turn excite the other array sensors via mutual and self-couplings. The equation to determine the mutual coupling is stated in matrices of Eq. (56.1).

$$
\begin{aligned}
Z_{Port1} &= \begin{bmatrix} Z_{11}+Z_L & Z_{12} & Z_{13}\ldots & Z_{15} \\ Z_{21} & Z_{22}+Z_L & Z_{23}\ldots & Z_{25} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{51} & Z_{52} & Z_{53}\ldots & Z_{55}+Z_L \end{bmatrix} \\
Z_{Port2} &= \begin{bmatrix} Z_{65}+Z_L & Z_{66} & Z_{67}\ldots & Z_{65} \\ Z_{75} & Z_{76}+Z_L & Z_{77}\ldots & Z_{79} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{105} & Z_{106} & Z_{107}\ldots & Z_{109}+Z_L \end{bmatrix} \\
Z_{Port3} &= \begin{bmatrix} Z_{61}+Z_L & Z_{62} & Z_{63}\ldots & Z_{65} \\ Z_{71} & Z_{72}+Z_L & Z_{73}\ldots & Z_{75} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{101} & Z_{102} & Z_{107}\ldots & Z_{105}+Z_L \end{bmatrix} \\
Z_{Port4} &= \begin{bmatrix} Z_{65}+Z_L & Z_{66} & Z_{67}\ldots & Z_{65} \\ Z_{75} & Z_{76}+Z_L & Z_{77}\ldots & Z_{79} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{105} & Z_{106} & Z_{107}\ldots & Z_{109}+Z_L \end{bmatrix} \\
&\vdots
\end{aligned}
\tag{56.1}
$$

The total input impedance (Z_{Total}) due to the mutual coupling is stated in Eq. (56.2). An array of 9 × 10 is configured such that the entire array is subdivided into four 5 × 5 sub-array lattices and excited by a 2 × 2 excitation source.

$$Z_{TOTAL} = \begin{bmatrix} Z_{Port(1,1)} & Z_{Port(1,2)} & \cdots & Z_{(1,x)} \\ Z_{Port(2,1)} & Z_{Port(2,2)} & \cdots & Z_{Port(2,x)} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{Port(y,1)} & Z_{Port(y,2)} & \cdots & Z_{Port(y,x)} \end{bmatrix} \tag{56.2}$$

The total input impedance as a result of the mutual coupling based on the four probes is as stated in Eq. (56.3). The input impedance is determined using the scattering parameter coefficients stated in Eq. (56.4) [3].

$$Z_{TOTAL} = \begin{bmatrix} Z_{Port(1,1)} & Z_{Port(1,2)} \\ Z_{Port(2,1)} & Z_{Port(2,2)} \end{bmatrix} \tag{56.3}$$

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad \begin{bmatrix} b_1 \\ b_3 \end{bmatrix} = \begin{bmatrix} S_{31} & S_{32} \\ S_{41} & S_{42} \end{bmatrix} \begin{bmatrix} a_1 \\ a_3 \end{bmatrix}$$
$$\begin{bmatrix} b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} S_{13} & S_{14} \\ S_{23} & S_{24} \end{bmatrix} \begin{bmatrix} a_3 \\ a_4 \end{bmatrix}, \quad \begin{bmatrix} b_2 \\ b_4 \end{bmatrix} = \begin{bmatrix} S_{33} & S_{34} \\ S_{43} & S_{44} \end{bmatrix} \begin{bmatrix} a_2 \\ a_4 \end{bmatrix} \tag{56.4}$$

The excitation voltage waves have been normalized using values a_1, a_2, … and b_1, b_2, … Therefore, the respective currents and voltages as a function of excitation sources are stated in Eqs. (56.5) and (56.6).

$$\begin{aligned} I_1 &= \frac{1}{(Z_{Port1})^{0.5}}(a_1 - b_1), \quad I_2 = \frac{1}{(Z_{Port2})^{0.5}}(a_2 - b_2) \\ I_3 &= \frac{1}{(Z_{Port3})^{0.5}}(a_3 - b_3), \quad I_4 = \frac{1}{(Z_{Port4})^{0.5}}(a_4 - b_4) \end{aligned} \tag{56.5}$$

$$\begin{aligned} v_1 &= (Z_{Port1})^{0.5}(a_1 + b_1), \quad v_2 = (Z_{Port2})^{0.5}(a_2 + b_2) \\ v_3 &= (Z_{Port3})^{0.5}(a_3 + b_3), \quad v_4 = (Z_{Port4})^{0.5}(a_4 + b_4) \end{aligned} \tag{56.6}$$

The total impedance (Z_{total}) is analyzed for a 2 × 2 excitation network using Matlab code. Figure 56.3 shows the obtained total impedance (Z_{total}) with respect to interelement spacing. Perfect matching was observed at Z = 0.5λ. Based on this, a proof of concept based on interelement spacing of 0.5λ was designed using EM microwave HFSS Ansys solver. Figure 56.4 shows the obtained results.

The input impedance is theoretically estimated to be 49.66 + j1.07 Ω, whereas the HFSS simulated result indicates a value of 49.98 + j.23 Ω. It is evident from the two results that the reactive component is very meager. As such, the impedance match is considerable.

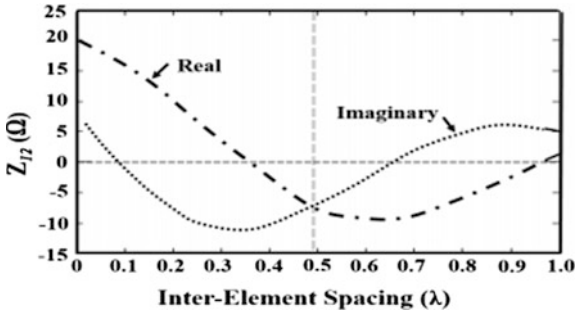


Fig. 56.3 Transfer impedance versus interelement spacing

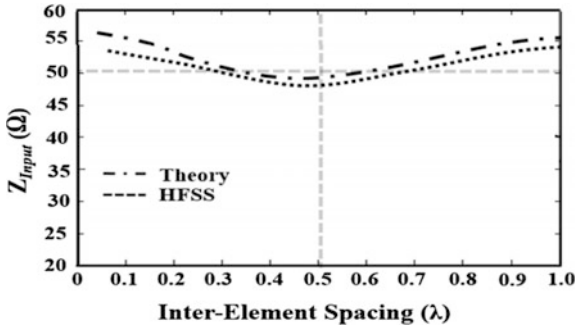


Fig. 56.4 Input impedance versus interelement spacing

Link Budget

The extracted values of the voltages and currents of Eqs. (56.5) and (56.6) are applied to the array configuration for the purpose of the proof of concept. We assume that element factor $f(\theta)$ is identical due to identical elements of the array, the uniform interelement spacing, and $f(\theta)$ is a function of the excitation sources and also indirectly depends on the impedance matching emphasized in section "Input Impedance Matching". Equation (56.7) indicates the current density for an excited microwave sensor, where $J_x(x, y)$ and $J_y(x, y)$ are defined by Eqs. (56.8) and (56.9), and N is equal to the number of interdigit fingers. The electric field density per a sensor is stated in Eq. (56.10) [4–7].

$$J_z(z) = \sum_{i=n}^{N} \left(J_x(x,y) + J_y(x,y)\right) \tag{56.7}$$

$$J_x\left(x^{'},y^{'}\right)=\begin{cases}\dfrac{\cos\left(\frac{2\pi}{w_c}x^{'}\right)}{\pi w_c\left(1-\left(\frac{2x^{'}}{w_c}\right)^2\right)^{0.5}} & n=0,2,4,6,\ldots,2n\\[2ex] \dfrac{\sin\left(\frac{2\pi}{w_c}x^{'}\right)}{\pi w_c\left(1-\left(\frac{2x^{'}}{w_c}\right)^2\right)^{0.5}} & n=1,5,7,9,\ldots,2(2n+1)\end{cases} \tag{56.8}$$

$$J_y\left(x^{'},y^{'}\right)=\begin{cases}\dfrac{\cos\left(\frac{2\pi}{w}y^{'}\right)}{\pi w\left(1-\left(\frac{2y^{'}}{w}\right)^2\right)^{0.5}} & n=0,2,4,6,\ldots,2n\\[2ex] \dfrac{\sin\left(\frac{2\pi}{w}y^{'}\right)}{\pi w\left(1-\left(\frac{2y^{'}}{w}\right)^2\right)^{0.5}} & n=1,5,7,9,\ldots,2(2n+1)\end{cases} \tag{56.9}$$

$$E_z(x,y,z)=\int_{-\infty}^{\infty}G(k_z)E(R,z=0)e^{-jkz}dk_z \tag{56.10}$$

$$G(k_z)=\int_{-\infty}^{\infty}J_z(z)e^{-jkz}dk_z,$$

where $G(k_z)$ is the Fourier transform of the current distribution function upon excitation. The far-field radiation is stated as shown in Eq. (56.11). Therefore, Eq. (56.11) can be decomposed to give Eqs. (56.12) and (56.13) [8].

$$E^{Far-field}=\frac{E_z^{Array}}{\sin\theta}$$

$$E^{Far-field}(r,\theta,\varphi)=\frac{jk_0G(k_z)e^{-jkr}}{2nr}f_{(x,y)}(\theta,\varphi)\times 2\sum_{p(x,y)}I_pG(k_z)e^{jkr(x\cos\theta+y\sin\varphi)\sin\theta} \tag{56.11}$$

$$EP=\frac{jk_0G(k_z)e^{-jkr}}{2nr}f_{(x,y)}(\theta,\varphi) \tag{56.12}$$

$$AF=2\sum_{p(x,y)}I_pG(k_z)e^{jkr(x\cos\theta+y\sin\varphi)\sin\theta}, \tag{56.13}$$

where AF is the array factor and EP is the product of the radiation intensities of the element. Equation 56.11 is the radiation pattern of the proposed $m \times n$ array.

Experimental Results

Equations (56.7) to (56.13) were coded using Matlab, to first extract the required values to characterize the link budget of the proposed microwave sensor array. A proof of concept was realized based on the extracted values using 3D EM numerical finite integration technique code. The design is photoetched on Roger printed circuit microwave laminate board.

Validation

To validate our proposition, an (9 × 10) array is implemented such that the array is subdivided into (5 × 5) sub-array lattice as shown in Fig. 56.2 of [2]. The (9 × 10) array is excited by (2 × 2) coaxial probe array coupling network, such that every (5 × 5) sub-array lattice is excited by at least one excitation source. Each microwave sensor is designed for IEEE 802.11 U-NII-3 standard, with resonance at 5.8 GHz. The interelement spacing of 0.5λ is determined to be 14 mm where $\lambda = 28$ mm.

Results and Discussions

Figure 56.5 depicts the obtained normalized array factor (AF) based on Eq. (56.13) using Matlab code. It is evident that the radiation pattern of the proposed array can be expressed as the product of the excited element factors $f(\theta)$ and the array factor (AF), i.e., AF × $f(\theta)$ as exemplified by Eq. (56.11) where $f(\theta)$ is expected to be identical for all the array elements. The obtained 3D AF plot is computed at $\theta = 90$ and $\varphi = 0$ as shown in the figure.

Table 56.1 summarizes the obtained performance profile of the proposed design. Findings indicate first that the input impedance matching is substantial as demonstrated by considerable reflection coefficient $|S_{11}|$ of well over 30 dB, with superior voltage standing wave ratio (VSWR) of 1.081:1. The radiation efficiency is 69.18% and aperture efficiency is 72.80%. The proposed microwave sensor demonstrated a robust area occupancy of (2.7 × 2.4 × 0.0315)λ and a link budget of about 20 dBi. The obtained link budget is very superior and can readily support long rage data transmission. The design is slim enough and is therefore adaptable for monolithic integrated microwave circuit.

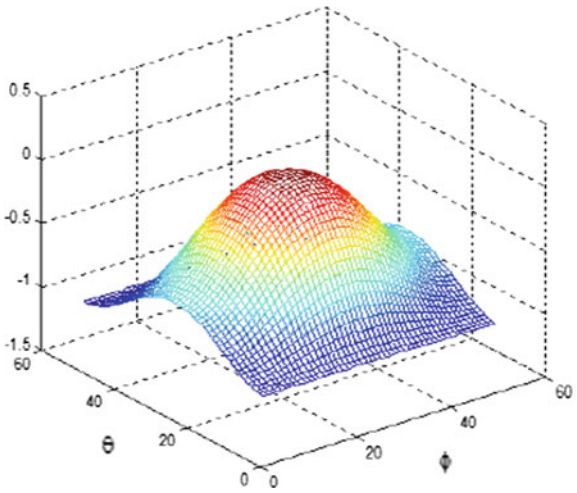


Fig. 56.5 Obtained normalized array factor (AF) of Eq. (56.13)

Table 56.1 Performance profile of the proposed sensor

|S11| (dB)	VSWR	Freq (GHz)	Gain (dBi)	Radiation efficiency (%)	Aperture efficiency (%)	Aperture size (mm)
34	1.081:1	5.80	19.79	69.18	72.80	2.7λ × 2.4λ

Conclusion

In this paper, we presented a spatial array of microwave sensors for IoT-based wireless connectivity for improved link budget. We opined that the IEEE 805.11x though have low data throughput could be used as complementary solution alongside the IEEE 805.15.4, in particular, when efficient alternative solutions are provided to leverage on the interconnections of heterogeneous wireless networks. We demonstrated in our previous work that alternative solutions could demonstrate substantial channel bandwidth enhancement to support big data, and more so if they are interconnected. We went further in this work to prove that such could also achieve significant link budget if an array is implemented. We first demonstrated that a miniaturized microwave sensor is achievable at lower frequency spectrum and, second, showed that the prevailing impedance matching challenges of small antennas at lower frequency band are surmountable.

References

1. Olokede S.S, Paul B., (2016). "Miniaturized microwave sensor for internet of things wireless connectivity," 16th Mediterranean Microwave Symposium, 14–16 November, Abu Dhabi, UAE. Accepted for publication.
2. Ain, M. F., S. S. Olokede, Y. M. Qasaymeh, A. Marzuki, J. J. Mohammed, S. Srimala, S. D. Hutagalung, Z. A. Ahmad, and M. Z. Abdulla, (2013). "A novel 5.8 GHz quasi-lumped element resonator antenna," *Int. J. Electron. Commun. (AEU)*, 67, 557–563.
3. Olokede S. S., and Adamariko C. A., (2015). "Analysis of the proximity coupling of a planar array quasi-lumped element resonator antenna based on four excitation sources," *Progress In Electromagnetics Research B*(63), 187–201.
4. Meixner, J. (1972). "The Behaviour of Electromagnetic Fields at Edges". IEEE Transactions on Antennas and Propagation, 20(4), 442–446.
5. Olokede S. S., (2015). "A Quasi-Lumped Element Series Array Resonator Antenna," Radioengineering, 24(3), 695–702.
6. Huang, F., Avenhaus, B., Lancaster, M. J. (1999). "Lumped Element Switchable Superconducting Filters". IEE Proceedings on Microwave, Antennas and Propagation, 146(3), 299–233.
7. Mittra, R., Lee, S. W. (1971). "Analytical Techniques in the Theory of Guided Waves". McMillan. Aspin: B0006c0e1m.
8. Simosvki, C. R. and S. He, (2001) "Antennas based on modified metallic photonic band gap structures consisting of capacitively loaded wires," *Microwave and Optical Tech. Letters*, 31(3), 214–221.

Chapter 57
Miniaturized Microwave Sensor for Internet of Things Wireless Connectivity

Seyi Stephen Olokede and Babu Sena Paul

Abstract A miniaturized microwave sensor for internet of things (IoTs) is presented. The proposed sensor though a periodic structure exhibits two intrinsic resonances, namely the spatial wavelength due to its periodic geometric structure and the radiation wavelength due to applied voltage source to the microwave sensor. The wavelength difference between the spatial and radiation wavelengths is employed for the sensing based on the electrical impedance tomography of the sensed material. The miniaturized capability of the proposed sensor is investigated based on some available formulae using Matlab code for parameter extraction, and also with finite integration technique electromagnetic (EM) codes. A proof-of-concept prototyped sensor is fabricated on a printed circuit microwave laminate board in order to validate the miniaturized capability of the proposed sensor. Findings indicate a superior impedance match, substantial impedance bandwidth, robust gain, and cost-effectiveness, compared to AoC/AiPs with associated losses due to the silicon substrate.

Keywords Antenna-on-chip/antenna-in-package · Big data · Channel bandwidth · IoTs · Integrated circuit · Link budget · Reflection coefficient System-on-chip

Introduction

The sustainability of big data technology may also depend on the efficiently implemented interconnections of highly integrated, high-speed, large-scale heterogeneous servers. Highly integrated IoTs heterogeneous sensor nodes and

S. S. Olokede (✉) · B. S. Paul
Department of Electrical and Electronic Engineering Technology,
University of Johannesburg, Johannesburg, South Africa
e-mail: solokede@gmail.com

B. S. Paul
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_57

gateways must be implemented to provide robust platforms to collect and transmit substantially massive and diverse data to the central controller through high channel bandwidth semi- to long-range wireless connectivity. For these massive data to be analyzed, stored, and mining, a real-time processing based on ultrahigh round-trip latency is mandatory. Interestingly, for a robust round-trip latency to be guaranteed, good link budget wireless connectivity, higher speed, low-power consumptions are additional but essential design requirements. Millimeter wave applications have indeed offered antidote for low channel bandwidth limitation, where possibilities of extremely-wide impedance bandwidth enhancement in the excess of 12 GHz and data throughput well above 1–20 Gbps have been demonstrated in some literature. Ironically, the next generation of wireless digital interface data rates of 1–20 Gbps is prerequisite for the mainstream deployment of big data technology. Unfortunately, inferior link budget in the order of negative dB is evident in most cases due to reduced radiation losses as a result of small aperture size of say 5 mm in particular for 60 GHz SoC radios. High radiation losses due to low resistivity silicon substrate, low-loss interconnect between components, propagation of surface wave at the interface between the sensor and the substrate, low integration levels, bulky and expensive waveguides and filters, conductor/dielectric losses, and lack of high-quality machined metal housings [1–5] are other debilitating challenges.

Though it has been determined that the existing IEEE 802.11x and ultra-wideband (UWB) standards [6] are incapable of supporting future multimedia applications, congestion of the band with too many competing applications are few of the many drawbacks of this standards. Nonetheless, for big data communication technology to be robust, highly densified large-scale heterogeneous nodes must be configured to provide the hardware platforms for its ecosystem. If the architecture must be heterogeneous, existing technologies such as IEEE 802.11x standards must be factored in order to leverage on channel bandwidth enhancement, need for more functionality, and also for IEEE 802.11x standards enable or adaptable applications. Moreover, current advances in semiconductor technology based on the bandgap reduction re-engineering could be complemented by these standards for low-cost alternatives, rather than to warrant millimeter wave wideband wireless applications' costs.

The basic requirement (among others) to achieve very high levels of integration and functionality is to consolidate on different existing sub-micron IC technologies, and also prospect for further new alternatives. In this work, therefore, we intend to explore IEEE 802.11x standards to determine to what extent it can be configured such that a miniaturized microwave sensor for wireless connectivity can be achieved. We will also investigate if the proposed sensor is robust enough in terms of directivity to leverage on long-range link budget. Finally, we will in addition estimate the channel bandwidth, the cost consequence, and the replicability.

Proposed Problem Formulation

The front-end design of the wireless connectivity device for big data transmitter/receiver via the gateway is constrained by electrical length. If this challenge is resolved, the problem of low radiative losses, poor directivity, and inferior link budget due to small form factor will be ameliorated to certain degree. Conventional sensor design procedure of transceivers at IEEE 802.11x and even at IEEE 802.15.3 has been inadequate to meet the system specifications. Hardware engineers are prospecting alternative robust solution in order to meet the systemic considerations. We proposed a microwave sensor depicted in Fig. 57.1. We explore alternative such that the resonance does not depend on the electrical length to avoid the dimensional limitation, and rather explore a geometry that depends on some lumped components. The structure consists of an assemblage of microstrip sections, bends,

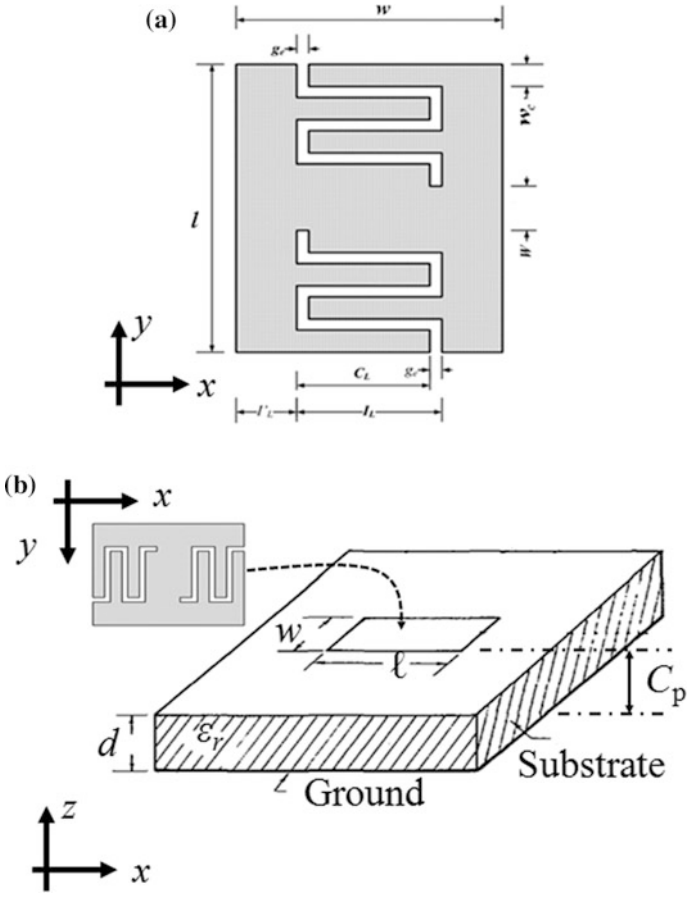


Fig. 57.1 The proposed sensor structure

gaps, and joint discontinuity as demonstrated in Fig. 57.1a of [7]. The performance of the structure is predicated on the intrinsic capacitive and inductive fringing fields, in particular the interdigit capacitance. Moreover, the size reduction is contingent on the long but folded interdigit meander line. The equation to determine the resonance is stated in Eq. (7) of [7], whereas this said resonance function is inversely proportional to square root of strip inductor (L), long but folded interdigit capacitance (C), and also on the pad capacitance (C_p). These parameters (i.e., L, C, C_p) are defined by Eqs. (1), (2), and (6) as stated in [7].

The values of these parameters are subsequently extracted based on these equations using Matlab code. Various parameter values with respect to resonance are determined exhaustively in order to investigate the miniaturized capability of the structure. Table 57.1 below depicts the obtained result. Findings indicate an aperture size of 528 mm^2 at a resonance of 1 GHz and 29 mm^2 at 6 GHz. The aperture size is reduced by 28.81% between 1 and 2 GHz resonance pattern, 86.55% between 1 and 3 GHz, 92.05% between 1 and 4 GHz, 94.13% between 1 and 5 GHz, and 94.51% between 1 and 6 GHz. Therefore, there is no limitation to the extent of size reduction as demonstrated by the table. However, difficulty of impedance matching, low radiative resistance, and etching limitation may constrain further reduction.

The proposed structure is further investigated using 3D finite integration technique code. We first understand the effect of printed microwave laminate board on the size reduction capability. The result is as tabulated in Table 57.2. The effect of

Table 57.1 Aperture size reduction capability of the proposed sensor

W_1 (mm)	g_e (mm)	Γ_L (mm)	N	C_L (mm)	I_L (mm)	w_C (mm)	Freq (GHz)	Length l (mm)	Width w (mm)	Aperture size (mm^2)
5.84	0.25	15.8	16	15.8	0.127	0.55	1.00	18.2	29.00	528
5.84	0.25	9.40	16	9.14	0.508	0.51	2.00	17.6	23.00	397
1.60	0.35	5.59	10	5.29	0.813	0.45	3.00	8.90	7.99	71
1.40	0.30	4.19	8	3.89	1.570	0.40	4.00	6.40	6.60	42
1.20	0.30	3.35	8	3.05	2.540	0.35	5.00	5.40	5.80	31
1.20	0.30	2.79	8	2.49	3.175	0.30	6.00	5.40	5.30	29

Table 57.2 Reduction capability based on permittivity function

Substrate	Substrate thickness (mm)	ε_r	Frequency (GHz)	Length (mm)	Width (mm)	Aperture size (mm^2)
R03003	0.127	3.00	5.784	7.000	5.680	39.760
RT6202PP	0.508	2.90	5.795	6.200	5.440	33.728
R04003C	0.813	3.38	5.800	5.800	5.600	32.480
RT5870	1.570	2.33	5.867	4.205	5.500	23.128
RT6010LM	2.540	10.2	5.882	4.205	4.335	18.229
RT5880	3.175	2.20	5.795	4.205	5.680	23.884

Table 57.3 Dimensions of the proof-of-concept prototype

W_1 (mm)	g_e (mm)	I_L (mm)	C_L (mm)	N	I'_L (mm)	w_C (mm)	Aperture size (mm)
3.35	0.30	3.35	3.05	8	1.125	0.35	5.8 × 5.6

two variables namely the dielectric thickness of the laminate board and the dielectric permittivity function—on the size reduction, they were examined using numerical EM code and available analytical formulae reported in [7]. It is evident from the table that the both parameters influence the reduction capability. A respective reduction ranging from 15, 18.19, 41.74, 54.1, and 39.84% is feasible when comparing Roger 3000 to other microwave laminates. It is evident that RT6010LM will offer the most promising reduction, directly followed by RT5870, and finally by RT5880. Interestingly, the increase of the thickness of the substrate may indeed enhance the reduction capability of the area occupancy, and also enhance impedance bandwidth; it may, however, be counterproductive in some practical applications where slimness could be a premium, as such prototypes could be bulky.

A proof-of-concept was designed for IEEE 802. 11g standard for WLAN ISM band for U-NII-3 indoor wireless applications. The dimensions are determined using the earlier stated Matlab code based on the said equations. The parameter values are stated in Table 57.3. The aperture size is 5.8 × 5.6 mm^2 at resonance of about 5.8 GHz. This aperture size is much smaller than the conventional rectangular metal patch of about 17 × 13 mm^2 at the same resonance. The proposed structure therefore demonstrates miniaturized capability of 85.30% over the metal patch at the same resonance, substrate permittivity, and thickness.

Experimental Results

Though a brilliant aperture size reduction of the proof-of-concept prototype is reported in Table 57.3 above, it is, however, pretty early to conclude a premium based on size occupancy only, in particular as regards small antennas. Feeding challenges, poor impedance matching, low-power handling, parasitic effects on the fingers, and poor radiative resistance could depreciate the otherwise gained advantage if not carefully fabricated. To understand these effects, a 3D numerical evaluation was done, and the prototype was measured. The numerical and the measured results are tabulated in Table 57.4. The numerical impedance bandwidth

Table 57.4 Performance results of the proposed sensor

	$\lvert S_{11}\rvert$ (dB)	VSWR	Freq (GHz)	BW (MHz)	Gain (dBi)	Input impedance (Ω)
Simulation	61.51	1.002:1	5.80	382	11.68	49.92−j1.32
Measurement	50.97	1.006:1	5.83	340	9.38	50.40−j2.16

is 382 MHz, whereas the measured is 340 MHz. For IEEE 802.11 for U-NII-3 standard, the proof-of-concept bandwidth can support 17 channel bandwidth of the transceiver with a standardized channel requirement of 20 MHz per channel. The link budget is in excess of 9 dBi, which is very substantial for a single element compared to what is obtainable when millimeter wave AoP or AiP is implemented. Instead, an inferior impedance bandwidth of 79.73% is demonstrated by a metallic patch even with such a large aperture size occupancy. It also demonstrates inferior link budget due to significant dielectric and conduction losses. The impedance matching is also reasonable both for the numerical and measured results as depicted in the table. The reactive components of both results are moderate as they are less than 3 Ω. Thus, the excitation power loss to the reactive component is negligible, with an antenna efficiency of 98.16%. A casual comparison of the efficiency of the proposed to that of metallic patch indicates an inferior patch efficiency of 64.1%.

Conclusion

In this paper, we investigated the feasibility of implementing a miniaturized microwave sensor for IoTs application. We posit that if the interconnections of densified heterogeneous IoT nodes are implemented, it could serve as platform for big data communication technology. We subsequently investigate the realization of IEEE 802.11x as a complimentary alternative for the heterogeneous interconnections of wireless connectivity. We introduce a sensor that does not depend on the electrical length and as such, it is not limited by its radiation wavelength. We established that other factors such as low radiative resistance, poor impedance matching, and etching limitation do not limit the performance of our proposed sensor. A proof-of-concept was implemented and measured to further validate our position. We opined that if interconnections of the sensor are done, an efficient cost-effective highly integrated, high-speed, large-scale sensor networks for big data communication can be implemented.

References

1. Oppermann M., (1998). “Multichip-modules for micro- and millimeter-wave applications—a challenge*?” International Conference on Multichip Modules and High Density Packaging*, 279–284.
2. Parker D., (2002). “Microwave industry outlook—defense applications”, *IEEE Trans. Microw. Theory and Tech*. 50(3), 1039–1041.
3. Olokede Seyi Stephen, Adamariko Clement Anowe, QasaymehYazeed Mohammed, (2015). “Equivalent circuit model of a coaxial excited microstrip-fed quasi-lumped element resonator antenna array”, IET Microwaves, Antennas and Propagation, 9(5), 446–453.
4. Jain N., (2000). “Designing commercially viable mm-wave modules”, in IEEE MTT-S Int. Microw. Symp. Dig., Boston, MA, 1, 565–568.

5. Olokede, S.S., Ain, M.F. and Ahmad, Z.A. (2016). "Dielectric loaded quasi-lumped element resonator antenna circuit model for U-NII/ISM band wireless applications," Ann. Telecommun. 71(9), 527–537.
6. Kitazawa K., Koriyama S., Minamiue H., and Fujii M., (2000). "77-GHz-band surface mountable ceramic package", *IEEE Trans. Microw. Theory and Tech*. 48(9), 1488–1491.
7. Ain M. F., Olokede S. S., Qasaymeh Y. M., Marzuki A., Mohammed J. J., Sreekantan S., Hutagalung D. S., Ahmad Z. A., Abdulla M. Z., (2013). A novel 5.8 GHz quasi-lumped element resonator antenna, AEU—International Journal of Electronics and Communications, 67(7), 557–563.

Part VII
Cryptography and Network Security

Chapter 58
Collaborative Intrusion Detection System in SDN Using Game Theory

V. N. Gowtham, R. N. Baratheraja, G. Jayabarathi and V. Vetriselvi

Abstract The SDN architecture decouples the data plane and control plane supporting virtualization in business organizations. As with any new technology, potential threats need to be addressed. The common vulnerability is the possibility of attack at various layers of SDN. The ability to implement security mechanisms on top of the central controller enables it to detect attacks in the network. In general, the SDN controller would steer the potentially offending traffic to an Intrusion Detection System (IDS) for analysis. There are various IDSs available. Each IDS is capable of detecting only a few kinds of attacks. Hence, more than one IDS has to be deployed in the controller to cover a wide range of attacks. However, deploying more than one IDS in the controller will lead to overhead by overloading it. This overhead can be handled by having each IDS in different controllers and making them communicate among themselves forming a collaborative system. The collaborative system is formalized in SDN using game theoretical framework that optimizes the behaviour of each IDS with respect to other IDS by achieving Nash equilibrium state. The proposed framework is evaluated by simulation to show the detection of wide range of attacks and reduce the overload of the controller.

Keywords Collaborative intrusion detection · Game theory · Nash equilibrium

V. N. Gowtham · R. N. Baratheraja · G. Jayabarathi · V. Vetriselvi (✉)
Department of Computer Science and Engineering, College of Engineering, Guindy, Chennai, India
e-mail: vetri@annauniv.edu; vetri@cs.annauniv.edu

V.N. Gowtham
e-mail: vngowthamceg@gmail.com

R. N. Baratheraja
e-mail: barathe94@gmail.com

G. Jayabarathi
e-mail: jayaguna66@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_58

Introduction

The dynamic environment enabled by virtualization has become a hindrance to traditional network. Hence, the new SDN technology has evolved to overcome this challenge by supporting virtualization, since it separates control plane from the data plane. OpenFlow protocol focuses specifically on these aspects. OpenFlow protocol is used for communication between switches and the controller. In the OpenFlow protocol, each switch will have a set of forwarding rules as instructed by the controller. When a new packet flow that does not match an entry in the forwarding table comes, the packet will be sent to the controller as PACKET-IN message. The controller then inspects the packet and adds the corresponding new flow in the forwarding table. Next time when a similar packet arrives, the switch need not consult the controller.

Though SDN has the advantage of virtualization, it is open for many kinds of attacks since the switches are not intelligent by themselves like in traditional network. The switches in traditional network use packet switching concept to receive and forward the data to the destination, whereas the switches in SDN work according to the rules specified by the controller. Hence, the problem of detecting security attacks on SDN network topology has to be addressed by the controller.

Many kinds of attacks like Denial of Service, worm and black hole attack are possible in SDN. An IDS mechanism is required to detect these attacks. It becomes impossible for the native IDS to detect other attacks which it is not designed to do so. In such a case, the controller can consult other controllers with a different IDS that is capable of detecting such an attack being deployed in it. From the response obtained from the neighbourhood controller, the controller can decide on its actions.

Trust management comes into play in such a case whether to trust the controller response or not. Based on the number of false positives or the probability of capture failure, trust can be defined between the controllers. Each controller is assumed to be the player of the game and the action of each controller as to whether pass or drop the packet will affect the other controller. A state in which each controller action would optimally help other controllers is called NASH EQUILIBRIUM. We propose a game theoretical model to attain the Nash equilibrium state between the controllers using Lagrangian relaxation to detect wide range of attacks without overloading the controller.

Related Work

Software-Defined Networking (SDN) is growing with the development of SDN-enabled devices and applications. The programmability and centralized control in SDN opens opportunity for many security attacks. The SDN scalability issue can be handled by using sFlow [1] protocol instead of OpenFlow. sFlow sends only a sample of packet to the controller when there is large traffic thereby avoiding flooding of

the controller in SDN. A combined approach of OpenFlow and sFlow may be used to detect the attacks in SDN like DDOS, port scan, etc.

An IDS is a device or software that monitors the network traffic for malicious activities or policy violation. There are network-based and host-based IDS [2]. NIDS is a network security system that focuses on the attacks that come from the inside of the network. The attacks caused by the traffic flows can be detected by using SPHINX IDS [3]. SPHINX detects the attacks in SDN with low-performance overhead and requires no changes to controllers for deployment. Gaikwad et al. [4] discussed signature-based IDS using multi-threading in SDN. Multi-threading enables parallel processing of the captured packets.

Snort IDS is a type of signature-based IDS that takes actions based on certain rules and generates an alert on detecting the attack in traditional network [5]. The snort rules are defined in such a way to alert, drop, or change packets when the incoming packet matches a particular rule. Snort rules can be defined and updated dynamically.

A particular IDS is capable of detecting only a few attacks. In order to cover wide range of attacks, a distributed and self-organized framework collaborating multiple heterogeneous IDS based on game theory may be used [6]. The harder problem of collaborating multiple IDS is done using game theory in traditional network. The architecture considers the game to be between the attackers and defenders. Here, each IDS reconfigures its mechanism based on the response from the other IDS. Fung and Boutaba [7] proposed the four main aspects of collaborative IDN (Intrusion Detection Network) system such as trust management, resource allocation, collaborative intrusion decision and collaborative selection in traditional network.

Among the four aspects, trust and proper resource allocation among the collaborating IDS are achieved by using a game theoretical model in traditional network [8]. Quanayan et al. [8] proposed the GUIDEX architecture which is based on reciprocal incentive-based resource allocation and trust management in which the amount of contribution from one IDS is proportional to the contribution of the neighbouring IDS to this IDS.

In a peer-to-peer network, some network nodes may not contribute their resources to other nodes but get benefited from other nodes. In order to resolve this issue, service differentiation can be used in a P2P network [9]. The mechanism is driven by a linear time distributed algorithm that works by assuming a game between the nodes in the traditional network. All the nodes in the network communicate with the information providing node to attain the Nash equilibrium state in a dynamic and effective way.

Yan et al. [10] proposed an approach for resolving the selfish behaviour of peers in the traditional network. Each peer in the network is assigned a ranking and those with bad rankings are considered to be free riders. The peer can increase its rank by contributing its services to the network and deteriorates its ranking by consuming resources from the network.

Optimal resource allocation among the collaborating nodes in traditional network may be achieved by following an economic model that uses cooperative game theory [11]. In this model, the resource allocation is done according to the trust values.

Here, the requests with lower trust value will be dropped. It is unfair because the entire resources would be consumed by the nodes with highest trust.

Traditional IDS finds it difficult to find the new threats. An incentive-based model using game theory [12] uses trust management for peers to collaborate truthfully in an IDN environment. An iterative algorithm that converges to Nash equilibrium is used based on the duality of the problem.

Dirichlet-based trust management [15] can be applied to measure trust among IDSs according to the mutual experience among the IDS nodes. Feldman et al. [13] proposed a system to study the phenomenon of free riding and free identities in peer-to-peer system. The nodes that leave the system and rejoin the system are termed as whitewashers.

The collaboration among multiple IDS using game theory is mostly dealt only in traditional networks. In this paper, an incentive-based game theoretical framework for the collaboration of multiple controllers in SDN deployed with heterogeneous IDS is proposed. This framework could detect wide range of attacks in SDN platform and can reduce the overload in the controller.

Collaborative Intrusion Detection Framework

The proposed architecture explains the collaboration of multiple IDSs to detect wide range of attacks with optimal resource allocation and trust management. The architecture of the proposed system is shown in Fig. 58.1. It consists of three main components, two controllers with different IDS being deployed in it and a common game theoretical component that takes care of resource management between the controllers. The number of controllers getting involved in collaboration can be extended. The two different IDSs deployed here includes entropy-based IDS and snort IDS.

The entropy-based IDS detects the attacks based on the behavioural change in the entropy of the packets. It uses the PACKET-IN messages for inspection and analysis. Snort IDS is a rule-based IDS that checks each new packet against the predefined rules and takes action according to the rules. Once the attack has been detected, proper mitigation steps have to be taken to avoid the occurrence of similar attacks in future.

A controller with single IDS is capable of detecting only a few attacks. If a controller A is not sure about the attack or if it finds a packet to be suspicious, it may send a consultation message to the another controller B asking for help. The controller B then may get the required information from controller A network by sharing the pcap file that would capture the traffic in binary format.

Based on the information obtained from the neighbourhood controller, the controller applies its own algorithm to detect the attack. If an attack is detected, corresponding mitigation process is also done accordingly for both entropy-based IDS or snort IDS.

To help nodes decide on the amount of resources it should allocate for a neighbouring IDS; a game theoretical incentive-based resource allocation mechanism is

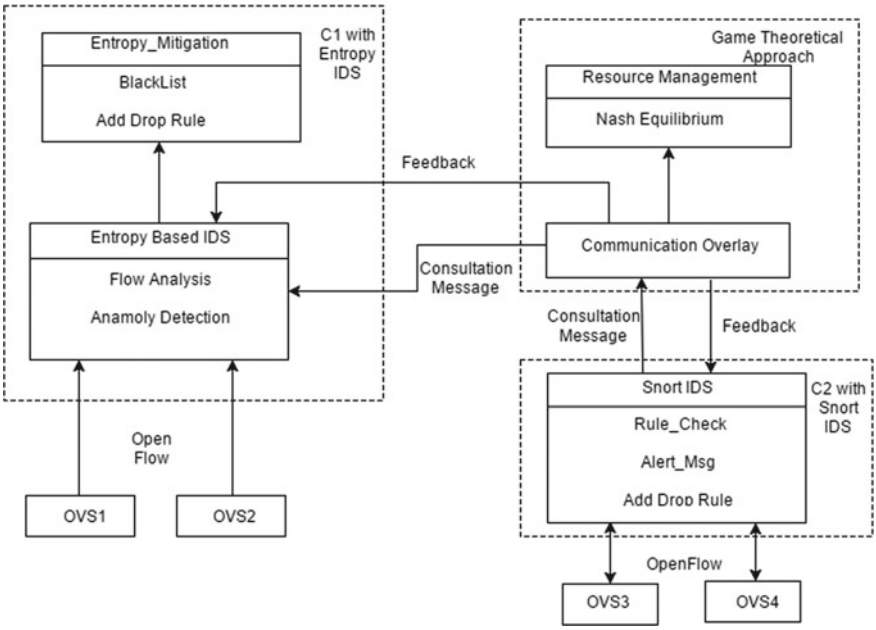


Fig. 58.1 Architecture of the collaborative IDS framework in SDN

used. The resource allocation is done based on their past assistance. The proposed framework aims at achieving Nash equilibrium state among the collaborating IDS nodes thereby avoiding free riders in the collaboration network.

Module Design

The design of the proposed architecture involves three major modules to be considered as depicted in Fig. 58.1.

Flow Statistics Collection

Using native OpenFlow approach The native OpenFlow protocol uses the forwarding logic in the controller and collects ordinary flow statistics from the switches along with their corresponding counters [1]. Those counters will be updated only when a look-up process matches the flow entry in the forwarding table. In the native OpenFlow approach, when the controller places the flow-stat request, the switches reply with large chunks of the flow table contents. Each chunk contains a portion of the

flow entry along with the total count of packets from the instantiation of the rule. For entropy-based detection, we require this count of packets during the last time window.

Similar to flow-stat request sent by the controller, a switch would send a PACKET-IN message to the controller when a packet that does not match any flow in the flow table arrives. Thus, when large traffic is encountered, a number of PACKET-IN messages would flood the controller. Hence, the controller might struggle to handle large traffic.

The flow entry of an OFF switch contains the 12-tuple flow definition. The action rule that specifies the forwarding path for a packet that is matched with the flow entry is used for forwarding the packets to the destination. The number of packets matched since the establishment of the flow, source IP and destination IP are used in the detection process of entropy-based IDS. Snort IDS uses source and destination IP and protocol in its detection process.

Entropy-Based IDS

Entropy is chosen as a measure to detect attacks. Entropy is a measure of randomness. By measuring the randomness of packets in the network traffic, certain attacks can be found. Controller maintains a window of packets to measure the randomness. Also, it is important to calculate the entropy independent of the number of distinct IP values in the window, so we calculate the normalized entropy by dividing it with the maximum entropy value of the window. The entropy can be calculated by using the formula

$$H_n(p) = \sum_i \frac{p_i log_b p_i}{log_b n}, \tag{58.1}$$

where $H_n(p)$ is the normalized entropy and $0 \leq H_n(p) \leq 1$ and n is the total number of occurrence of the entropy parameter under consideration and p_i is the probability of occurrence of the entropy parameter within a window. The entropy is divided by the $log_b n$ for normalization.

For the occurrence of DOS attack, the attacker might flood a single target host with large number of packets. The attacker may spoof the source IP of all the packets with different IPs so that the IDS will not be able to identify the attacker. Hping3 is a tool used to generate the DOS attack with spoofed source IP by flooding the controller. During this DOS attack, the controller will be flooded with PACKET-IN messages due to different source IPs. These PACKET-IN messages will be inspected by the IDS and the entropy value will be calculated for each PACKET-IN message.

The entropy value will be around 1 for normal traffic. But during the DOS attack, the entropy value of the destination IP decreases drastically. The controller is designed to identify such anomaly by using entropy values. Thus, when the entropy value decreases, the DOS attack will be detected.

Entropy IDS algorithm contains the sequence of steps followed by entropy-based IDS. The algorithm uses a queue to maintain the window of PACKET-IN messages. When a new PACKET-IN message occurs, the old packet is dequeued and the new packet is enqueued. Then, the probability of occurrence of the arrived PACKET-IN message in the current window is found. This probability is used to calculate the entropy. The DOS attack will be detected when there is a drastic decrease in the entropy value.

Snort IDS

Snort is a multi-mode packet analysis tool. Snort has certain rules that form the signatures to detect malicious behaviour. Any new traffic that does not match any flow entry in the switch will be checked against the snort rules. When a flow entry matches the snort rule, the corresponding traffic alert will be sent to the log file. The snort detection engine is programmed in such a way that it describes per packet tests and actions [14]. The components used by snort IDS in its detection process include sniffer, packet logger and data analysis tool.

In a software-defined network, the IDS is deployed in the controller of the network. When a new packet arrives, the snort IDS deployed in the controller would inspect the packet against the snort rules checking for intrusion. Based on the rule matched, the controller would take the corresponding action.

The packet decoder or sniffer would have decoding routines that are called in order to use the protocol stack. The main functionality of the decoder includes setting pointers into the packet for later analysis by the detector. The decoder would send details like inport, outport, source IP, destination IP and the protocol used to the detection engine. The detection engine maintains the detection rules in a two-dimensional linked list. The detection engine recursively checks this list of rules, and the first rule that matches a decoded packet in the detection engine would trigger the corresponding action.

The logging or alerting subsystem would send the alert to either syslog or store it in the text file or display it as a window pop-up message. Here, we would send the alert message to the other controllers in a collaborative network when asked for consultation.

The three basic actions carried out by the snort IDS are pass, log, or alert. Pass rules simply forward the packet, and log rules would write the full packet to the log file. Alert rule would create a notification as specified by the user at command line. Snort allows us to create new rules and update. Snort rules mainly focus on protocol, direction and port of interest in the packet.

The working of snort IDS is evaluated by generating port scan attack. Packets with specific destination IP with randomly changing destination ports will be used to generate port scan attack. Nmap is the tool used to generate port scan attack. The rules defined in snort IDS will help in detecting the occurrence of attack.

Resource Management

Resource Allocation Modelling

In our collaboration network, we consider N peers or controllers. The IDS nodes can communicate by sharing information using the REST APIs for sending consultation messages and feedbacks. We denote the number of controllers as $N = 1, 2, \ldots, N$ [8]. The neighbouring node of a controller u is denoted by N_u. We consider the network to be symmetric in such a way where when u is a neighbour of v, then v is also a neighbour of u. The topology of the collaboration network can be represented by a graph $G := (N, \xi)$, where ξ represents the set of (u, v) pairs in the network. We use r_{vu} to represent the units of resource that node u should allocate to v for the full satisfaction of v, whereas m_{vu} represents the minimum units of resource that node u should allocate to node v. These parameters are decided by v and sent to u during communication and establishment of connection. Let $p_{uv} \epsilon R_+$ over the interval $[m_{vu}, r_{vu}]$ be the units of resource that u allocates to v and it should lie within the specified interval for u to satisfy v. This is considered to be the deciding variable of controller u and is known only to u and v.

The response given by u to v is closely related to the trust value of controller v denoted by $T_v^u \in [0, 1]$. This represents the amount of trust that controller u has on v and is assessed by u. In our model, we assign various trust values for the controllers and learn the behaviour of the collaborative intrusion detection network.

Based on the capacity constraint C_u, which includes bandwidth, CPU and memory, each controller would try to maximize the help it offers for the neighbouring nodes. Thus, resource allocation should satisfy the following capacity constraint:

$$\sum_{v \in N_u} p_{uv} \leq C_u, \ \forall \ u \ \in N \tag{58.2}$$

We here use the non-cooperative approach in our game theory to model the altruistic behaviour among controllers. The non- cooperativeness is appropriate here because there is no centralized control among the controllers. The controllers introduced in the network intend to help other controllers and if one controller refuses to help the other controllers, the other controllers would correspondingly decline to help them in return, thus avoiding free riders.

The proposed framework is suitable for collaborative networks with reciprocal altruism. Also, the distinct features of intrusion detection system are incorporated. The insider attacks in IDS networks have been handled by imposing upper and lower bounds on p_{uv}, which can be used to prevent DOS (denial of service) attacks from the insiders. We have our players (controllers) aligned in a certain way to facilitate efficient sharing of knowledge with each other.

Incentive Properties

A collaborative network is said to be incentive if the helping resource p_{uv} from controller u to v increases as p_{vu} from controller v to u increases. Also, the helping resource depends on the trust value, i.e., the more the controller u trusts v, the more will be the helping resource p_{uv}.

The lower and upper bounds on the incentives will be decided by the controller. The bounds are decided because when a higher level request occurs, the controller would prefer to satisfy the lower level ones by allocating resources for them first. Hence, increased request rate will result in lowering the reply rates.

Primal/Dual Iterative Algorithm

This is the dynamic algorithm to compute the unique Nash equilibrium. Let $p_{uv}(t)$ be the resource allocated from controller u to v at step t. Consider the following algorithm:

$$\begin{cases} p_{uv}(t+1) = s_{uv} + t_{uv} p_{vu}(t) \\ p_{vu}(t+1) = s_{vu} + t_{vu} p_{uv}(t) \end{cases} \tag{58.3}$$

where $s_{uv} = \left(1 + \frac{1}{\alpha}\right) m_{vu} - \frac{1}{\alpha} r_{vu}$, $t_{uv} = \frac{T_v^u}{\xi_{uv}(ln(1+\alpha))}$, and s_{vu}, t_{vu} can be found in the same manner with initial conditions $p_{uv}(0) = min\left\{\frac{C_u}{N_u}, r_{uv}\right\}$, $\forall u, v \, \epsilon N$. This algorithm depends on the Lagrange multiplier λ_u.

Hence, the algorithm in (58.3) will be used to calculate the amount of resources shared between controllers at different instances when a consultation message is sent.

Experimental Evaluation

We evaluate the performance of the collaborative IDS in SDN by simulation. We focus on the detection accuracy of the collaborative framework and the range of attacks it could detect using OpenFlow mechanism in the Floodlight controller using Mininet simulation tool. Communication among the controller is achieved using the REST API. The entropy-based IDS and snort IDS are deployed in two separate controllers. The result from each IDS will be sent to the anomaly mitigation module for proper countermeasures. The proposed mechanism can be extended to a generic SDN topology with more number of OpenFlow-enabled switches and controllers.

Whenever a new packet arrives, the entropy will be calculated by sliding the window according to the window size. The entropy variation for the arrival of each packet is shown in Fig. 58.2. When the attacker floods the same destination with large number of packets, the entropy decreases drastically.

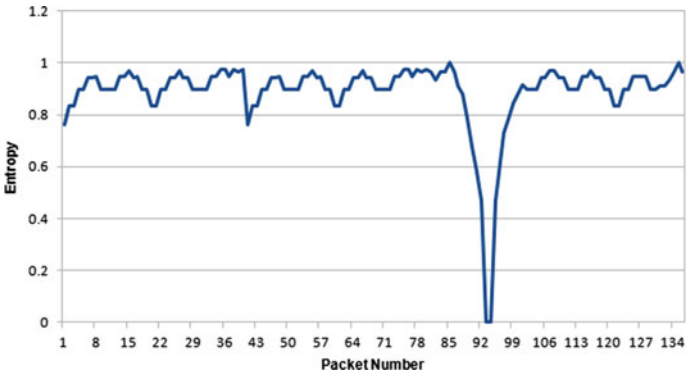


Fig. 58.2 Entropy variations of destination IP to detect DOS attack

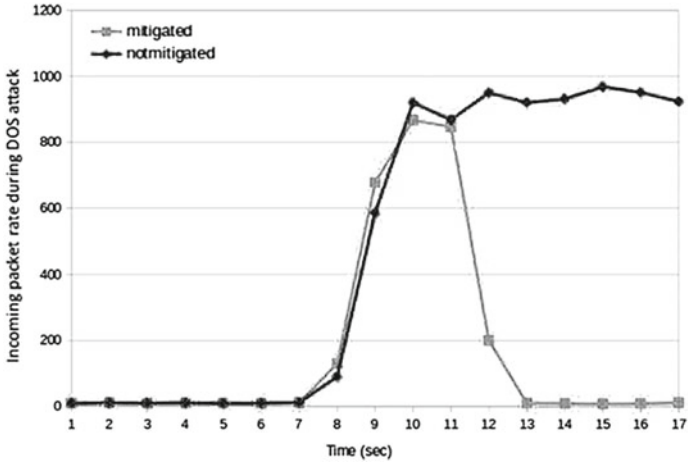


Fig. 58.3 Time versus Incoming packet rate

The incoming packet rate increases drastically during a DOS attack because of flooding and stays high if mitigation is not handled. If proper mitigation is done, the incoming packet rate would be restored back to original state since the flooding will be blocked by dropping the packets from the attacker. This is depicted in Fig. 58.3.

The graph in Fig. 58.4 shows the attainment of NASH equilibrium state by varying the trust among the controllers. The higher the trust, the sooner the attainment of Nash equilibrium. The response rate varies based on the maximum capacity of the controllers to help others. For experimentation, the controllers are initially assumed to respond to all requests with the maximum capacity of 30. The number of responses decreases over iterations and becomes stable at some particular point based on trust. It is evident that as trust decreases, the attainment of Nash equilibrium is delayed.

The term 'Capacity' here refers to the number of responses a controller could send to other controllers for helping them. The response rate increases with the capacity

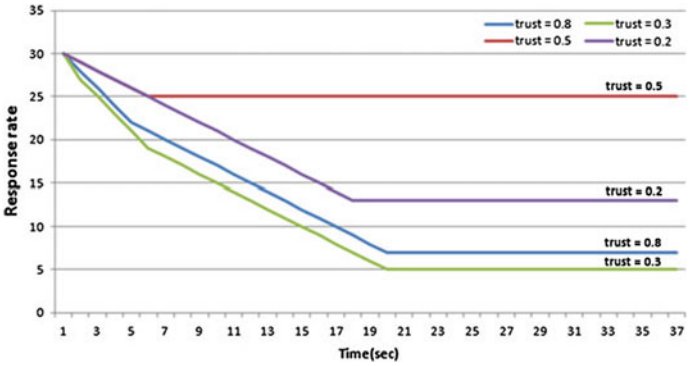


Fig. 58.4 Time versus response rate

of the controller to help other nodes. The response rate becomes stable at a particular point based on the trust and the maximum capacity to help. Each controller is assigned with a capacity value so that it could monitor the attacks in its own network and respond to the consultation messages from the neighbouring controllers. The stable response rate would be equivalent to capacity when trust value is 1 but for lesser values of trust, the response rate attains stability at a value less than the maximum capacity $\mathbf{R}_{vu}$. This is shown in Fig. 58.5.

The detection accuracy remains the same with or without collaboration when the IDS is handling the attacks that it is designed to identify. But for other new attacks, the detection accuracy is very low for the IDS without collaboration. Hence, it is observed that the detection accuracy of the proposed collaborative framework is found to be higher than that of the IDS without collaboration. Thus, it is evident that

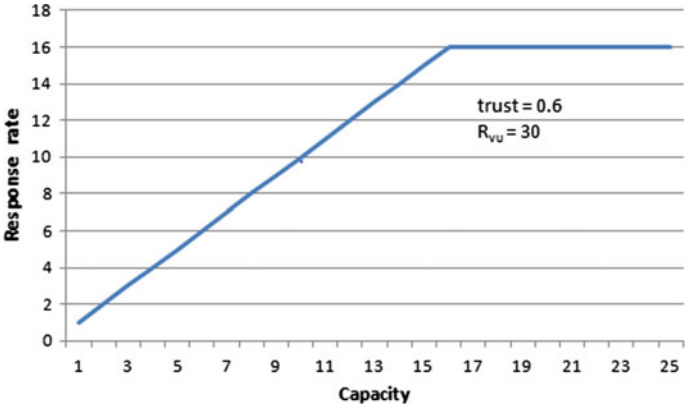


Fig. 58.5 Capacity versus Response rate

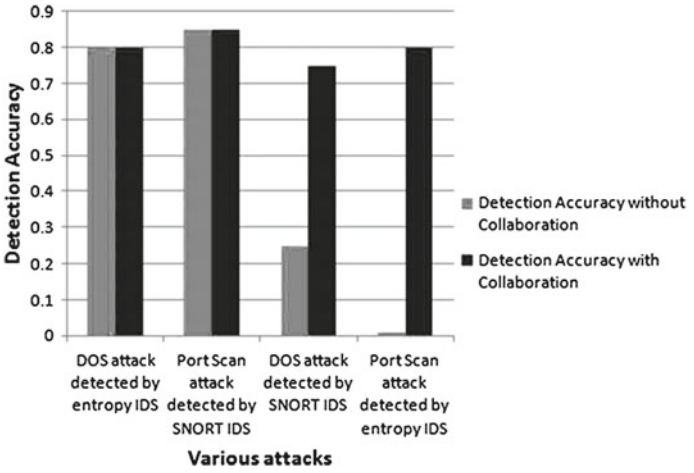


Fig. 58.6 Detection accuracy for various attacks

the proposed framework covers detection of wide range of attacks. This is shown in Fig. 58.6.

The observations from all the above graphs show that the efficiency of detection is increased in the proposed framework.

References

1. Giotis, K., Christos Argyropoulos, Georgios Androulidakis, Dimitrios Kalogeras, Vasilis Maglaris.: Combining OpenFlow and sFlow for an effective and scalable anomaly detection and mitigation mechanism on SDN environments. Computer Networks 62 (2014): pp. 122–136.
2. Scott-Hayward, Sandra, Gemma O'Callaghan, Sakir Sezer.: Sdn security: A survey. In Future Networks and Services (SDN4FNS), 2013 IEEE SDN For, pp. 1–7.
3. Dhawan, Mohan, Rishabh Poddar, Kshiteej Mahajan, Vijay Mann: SPHINX: Detecting security attacks in software-defined networks. InProceedings of the 2015 Network and Distributed System Security (NDSS) Symposium (2015).
4. Gaikwad, D. P., Pooja Pabshettiwar, Priyanka Musale, Pooja Paranjape, Ashwini S. Pawar.: A Proposal for reading Technique. International Journal of Computational Engineering Research 2, no. 7 (2012): pp. 59–65.
5. Xing, Tianyi, Dijiang Huang, Le Xu, Chun-Jen Chung, Pankaj Khatkar.: Snortflow: A openflow-based intrusion prevention system in cloud environment. In Research and Educational Experiment Workshop (GREE), 2013 IEEE Second GENI, pp. 89–92.
6. Bartos, Karel, Markus Rehak.: Self-organized mechanism for distributed setup of multiple heterogeneous intrusion detection systems. In Self-Adaptive and Self-Organizing Systems Workshops (SASOW), 2012 IEEE sixth international conference on, pp. 31–38.
7. Fung, Carol J., Raouf Boutaba. Design and management of collaborative intrusion detection networks. In Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on, pp. 955–961. IEEE, 2013.

8. Zhu, Quanyan, et al.: GUIDEX: A game-theoretic incentive-based mechanism for intrusion detection networks. Selected Areas in Communications, IEEE Journal on 30.11 (2012): pp. 2220–2230.
9. R. Ma, S. Lee, J. Lui, D. Yau: A game theoretic approach to provide incentive and service differentiation in P2P networks. In Sigmetrics/Performance, 2004.
10. Y. Yan, A. El-Atawy, E. Al-Shaer: Ranking-based optimal resource allocation in peer-to-peer networks In Proceedings of 26th annual IEEE conference on computer communications (INFOCOM 2007), May, 2007.
11. Grothoff, Christian.: Resource allocation in peer-to-peer networks: An excess-based economic model. Wirtschaftsinformatik 45, no. 3 (2003): pp. 285–292.
12. Zhu, Quanyan, Carol Fung, Raouf Boutaba, Tamer Baar.: A game-theoretical approach to incentive design in collaborative intrusion detection networks. In Game Theory for Networks, 2009. GameNets'09. International Conference on, IEEE, 2009. pp. 384–392.
13. Feldman, Michal, Christos Papadimitriou, John Chuang, Ion Stoica.: Free-riding and white-washing in peer-to-peer systems. Selected Areas in Communications, IEEE Journal on 24, no. 5 (2006): pp. 1010–1019.
14. Roesch, M.: Snort: Lightweight Intrusion Detection for Networks. In LISA (Vol. 99, No. 1, pp. 229–238).
15. C. Fung, J. Zhang, I. Aib, and R. Boutaba, "Dirichlet-based trust management for effective collaborative intrusion detection networks," *IEEE Trans. Network Service Management (TNSM)*, Vol. 8, No. 2 (2011): pp. 79–91.

Chapter 59
Intrusion Detection System for Software-Defined Networks Using Fuzzy System

Shalini S., Shafreen Nihara A., Sathiya Priya L. and Vetriselvi V.

Abstract In the field of computer networks, one among the latest technologies which is attracting researchers is software-defined networking (SDN). The network is centrally managed through software-based SDN controllers. SDN offers several advantages over conventional networks. Our area of interest is to enhance the security of present networks to a greater extent through centralized SDN controllers. Mechanisms to achieve this goal via SDN should be devised. An efficient intrusion detection system which is capable of monitoring real-time network traffic and reporting about intrusion if any to the controller is a good solution to this problem. Intrusion detection system (IDS) exists for traditional networks. Introducing IDS in SDN field results in achieving better efficiency compared to traditional networks, since it allows the controller to take immediate action on the attacker as soon as the attack is found. The aim of this project is to enhance security in networks through SDN by developing IDS using machine learning technique, namely fuzzy approach. The advantage of using this approach is that it provides high attack detection rate and less false alarm rate.

Keywords Intrusion detection system · Software-defined networks
Fuzzy system · Machine learning

Shalini S. · Shafreen Nihara A. · Sathiya Priya L. · Vetriselvi V. (✉)
Department of Computer Science and Engineering, College of Engineering Guindy, Chennai, India
e-mail: vetri@annauniv.edu

Shalini S.
e-mail: shalinishanmugamcse@gmail.com

Shafreen Nihara A.
e-mail: shafshafrin@gmail.com

Sathiya Priya L.
e-mail: sathiyapriyalcse@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_59

Introduction

In the networking world, security becomes a basic requirement as more sensitive data are available on Internet. When information gets transferred from one system to another in the network, it has to go through a number of intermediate nodes. This gives other users in the network an opportunity to gain access over the data which may result in compromising confidentiality, integrity, and availability of the system. In the field of computer networks, more efforts have been done till date to tackle security threats and attacks. The latest technology SDN provides a new paradigm of separating control plane from the underlying data plane so that the control of the network becomes directly programmable. Through SDN, the entire network can be viewed globally [1] and network configurations can be adjusted through centralized SDN controller which acts as a backbone of SDN network.

In conventional networks, the control resides in each and every networking device which takes forwarding decisions. But in SDN, separating the control plane from those devices and centralizing, it provides the following advantages. It facilitates to allow or block certain packets, set priorities to them, set path of packet flow over the network, etc. The control plane acts as a master by sending instructions to the data plane which serves as a slave. These networking devices simply act as forwarding devices by checking the flow entries installed and managed by the controller in flow tables. The messages between controller and data plane were exchanged securely via OpenFlow protocol.

There are two key interfaces in SDN network. They are Northbound and Southbound interface. A northbound interface is between external applications and the control plane. Through this interface, external applications can influence the network behavior. A southbound interface is between the controller and the physical data plane. Using this interface, SDN controller implements certain actions such as setting flow entries, dropping packets, disconnecting the host from network, etc. in forwarding plane.

SDN suffers from same security issues as of conventional networks. Apart from these, SDN has additional security issues also. As the controller is centralized in SDN, serious issue arises when the attacker gains control over the controller. To address the security problems that exist in both traditional and SDN networks, security defence mechanisms working in traditional networks can be deployed in SDN framework but some changes have to be applied to adapt to software-defined network. Some of the mechanisms which were developed to filter out and detect various attacks, threats, and malicious behaviors are firewalls, antivirus, and intrusion detection systems.

Firewall screens out viruses, worms or any malicious behavior that tries to reach the end host over the Internet. But there is no guarantee that every packet on the Internet passes through these firewalls. Hence, an intrusion detection system which

monitors and analyzes overall network traffic and reports malicious activities in the network is essential. Signature-based IDS is not efficient in detecting unknown attacks. Hence, anomaly-based IDS must be developed to address emerging new attacks. There were many techniques available for developing such anomaly-based IDS. Some of them include rule-based expert systems, data mining techniques, machine learning techniques, etc. Among those various techniques, machine learning proves to be the best technique to develop anomaly-based IDS. This technique uses supervised learning approach for this particular classification problem taken. This project deals with development of intrusion detection system using fuzzy logic which is a machine learning technique.

The paper is further organized as follows: Section "Related Works" deals with related works of the project. Section "Proposed System" discusses the proposed system. Sections "Feature Extraction and Controller Action", "Early Detection Algorithms", and "Anomaly-Based Fuzzy Intrusion Detection System" explain about each and every module in detail. Section "Experimental Framework" speaks about experimental setup and performance metrics. Finally, the paper is summarized and concluded in section "Conclusion".

Related Works

Software-defined networking is gaining the interest of business and academia. Recently, many researches have been done in the field of SDN. Since it is an emerging field, there is a need to analyze the key benefits and challenges of software-defined networks. These key concepts were addressed and SDN reference model was proposed [2].

Security proves to be a great challenge in software-defined networks. SDN has same security threats as that of traditional networks. In addition, it also suffers from new security threats such as flow entries being overwritten and fraudulent flow entries being inserted, etc. To address these security issues, a solid security architecture for software-defined network is essential [3]. In this architecture, various security policies were received from third-party applications by controller through northbound interface. The SDN controller follows those security policies in order to achieve security in the network.

The policy from application with high privilege level is given the first choice by the controller in terms of collision. Any malicious behavior in the network is reported to external security management application which sends information to the controller immediately for necessary action.

Software-defined networking was initially designed for wired networks. Nowadays, wireless environment is gaining much attention among people. Hence, carrying huge volumes of traffic and providing support for more sophisticated services such as VoIP, streaming media, and messaging has become inevitable for

mobile operators. Offering these services using software-defined networking has more advantages over traditional networking. SDN enables path management based on bandwidth demands which vary according to the type of content being sent. Also, deploying SDN in wireless is cost effective. Many works have been done in software-defined wireless networks and an architecture for SDWN has been proposed [4]. Also, a survey on security for mobile devices was held [5].

In order to overcome the security threats in the network, several efforts have been taken. One among them is intrusion detection system. Intrusion detection can be categorized into two types, namely misuse and anomaly detection. Combining both misuse detection and anomaly-based detection resulting in hybridized detection system yields much better results [6].

In the introduction of this paper, the benefits of using fuzzy logic for developing intrusion detection system are discussed. An intrusion detection system for traditional networks was developed using genetic fuzzy systems [7]. Fuzzy logic algorithms were devised and combined with genetic algorithms for development of efficient IDS [8].

Generating fuzzy rule base is essential to train the machine learning component to differentiate between normal and abnormal behavior. Fuzzy rule base can be manually created by recording entries given by experts. But logging numerous records covering wide variety of attacks is a challenging task. Thus, mechanisms to automatically generate fuzzy rules have gained more interest [9]. KDD Cup 99 dataset is used for training the machine learning component and for testing the system [10].

From the brief review, we can infer that efficient intrusion detection systems developed using fuzzy logic exist only for traditional networks. There is no such IDS to achieve better security in software-defined networks. Some early detection algorithms to detect attacks prior to connection establishment phase devised using fuzzy logic help in achieving some amount of security in SDN [11]. But these algorithms handle only two scenarios using connection success ratio and throttling connection algorithms.

Thus, our interest is to develop an intrusion detection system using fuzzy logic capable of detecting all kinds of attacks and informing the controller about intrusion. Our IDS achieves security to a greater extent in SDN, since it identifies all types of attacks.

Proposed System

Figure 59.1 depicts the overall architecture of the proposed system. The proposed system consists of centralized SDN controller, early detection algorithms, and anomaly-based fuzzy IDS which were embedded as modules within the controller and the forwarding plane, i.e., switches.

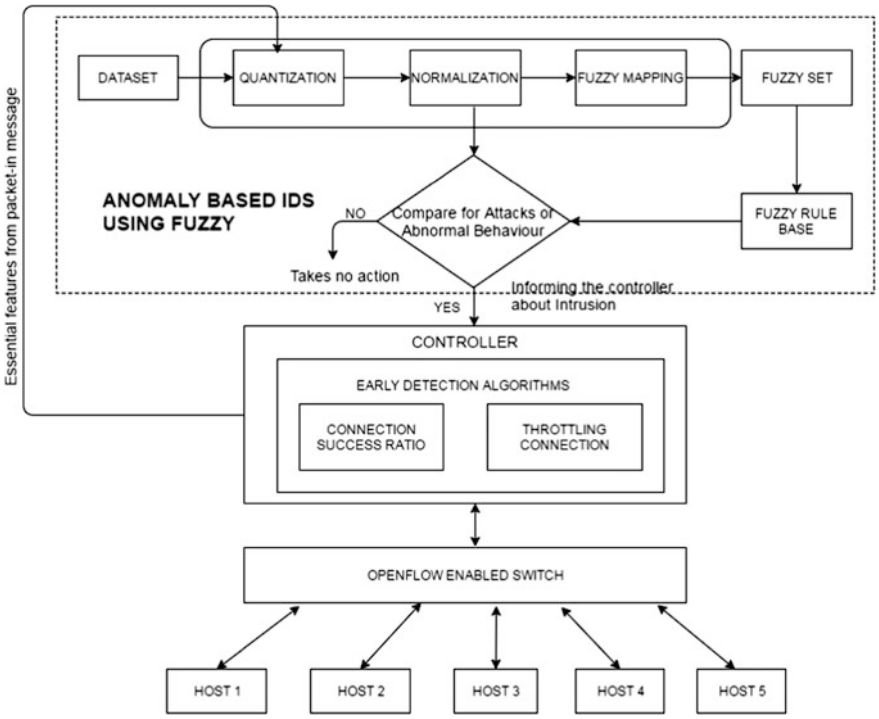


Fig. 59.1 Architecture for intrusion detection system in SDN

In this architecture, only one OpenFlow-enabled switch which is controlled by SDN controller is exposed for ease of exposition. Hosts in the network are connected to the OpenFlow-enabled switch. The number of switches and hosts can be extended.

The OpenFlow-enabled switch forwards the network traffic to the SDN controller and waits for the flow responses from it via secure OpenFlow channel. Whenever a switch receives a packet from a host, it first checks whether the packet is of request type or data. The switch forwards the packet to early detection algorithms module if it is of request type. Else, it forwards it to anomaly-based IDS module.

Early detection algorithms module consists of two detection algorithms, namely connection success ratio and throttling connection [12] which are capable of detecting attacks at an early stage, i.e., before the connection is established. These algorithms use the PACKET-IN message of connection request packets sent from switches, process them, and detect attackers if any. Then they inform the SDN

controller about the malicious host. Else, it sends the PACKET-IN message to anomaly-based IDS for further inspection.

Anomaly-based fuzzy IDS which is capable of detecting unseen attacks is developed using supervised machine learning approach. Dataset is collected from software-defined network and KDD Cup 99. The dataset consists of 500 training samples. Each sample consists of eight features (x1, x2, …, x8) which were extracted from PACKET-IN message and a class label (attack or normal) distinguishing normal host from malicious host. The collected dataset is quantized, normalized, and then mapped to fuzzy. Fuzzy membership functions were defined. Fuzzy rule base is created. The machine learning component is trained with the fuzzy rule base. This is a classification problem where the IDS classifies the incoming packet as either attack or normal based on the knowledge learned by training. IDS reports to the controller if it detects any abnormal behavior.

The controller receives information about intrusion either from anomaly-based fuzzy IDS or early detection algorithms in real time. The controller takes immediate action by disconnecting malicious host from the network. In this manner, the control is logically centralized.

Feature Extraction and Controller Action

The OpenFlow-enabled switches which act as just forwarding devices in SDN rely on the controller for each and every new incoming packet they encounter. Suppose let us say host A needs to communicate with host B and connection needs to be established between the two hosts. For requesting the connection, host A needs to send connection request packet to host B via switch to which it is connected. Once the connection is established, both the hosts can send or receive data between each other. Whenever a switch encounters a packet, it checks in the flow table for the corresponding source and destination host in order to route the packet. If entry is present, it simply routes the packet via the path present in flow table. Else, it sends PACKET-IN message of the received packet to the controller and waits for the flow response.

Algorithm 1. Controller Action

```
Input : PACKET-IN message from switch to controller
Output : Disconnects host from network if attack is found

if PACKET-IN msg arrives then
  if the packet is connection request then
    Extract src_ip, dest_ip, flag
    Pass PACKET-IN message and extracted features to Ear-
ly detection algorithms module for inspection
  else if packet is data
    Extract src_ip, flag, protocol, diff_serv count
    Pass PACKET-IN message and extracted features to
Anomaly based IDS module for inspection
  end if
if either Early detection algorithms or Anomaly based
fuzzy IDS module reports intrusion then
  Disconnects host which sent the packet from the switch
by dropping future packets from the particular host.
end if
end if
```

Features such as source address, destination address, and flag were extracted and sent to early detection algorithms since those features were the essential ones to detect attacks if any. For anomaly-based IDS, source address, protocol, flag, diff_serv count were extracted and sent since the machine learning component which resides in anomaly-based fuzzy IDS is trained with these features. Hence, it is essential to pass input packet's necessary features to the IDS so that the trained model compares and detects attacks.

Early Detection Algorithms

The early detection algorithms module receives PACKET-IN message and essential features from SDN controller. Based on the observed fact that most attacks require Internet connection, early detection algorithms [12] to detect attacks prior to connection establishment phase were used. The motive of using these algorithms in addition to anomaly-based fuzzy IDS is to enhance security by detecting certain attacks at early stage avoiding huge loss. These algorithms are as follows.

Connection Success Ratio

This algorithm works on the observed fact that if a good host attempts for a connection with any other host in the network, the probability that it succeeds in its connection should be much higher compared to any other malicious host. This algorithm is designed based on connection success ratio. To implement this, detection module maintains a pending request list. When the number of pending connections commenced from the specific host is high, the more likely the host has been infected. A threshold T1 is defined which represents that maximum number of pending requests a host can have. Whenever the count of pending connection requests for a host exceeds certain threshold value T1, the algorithm considers the host to be malicious and informs the controller.

Throttling Connection

This algorithm works on the fact that attackers normally try to propagate to different systems within very short duration. Connection rates will be higher for malicious hosts. The detection module maintains Recent Accessed Hosts (RAH) list for each host. It checks the list whenever it receives a new connection request from a host. If RAH list contains the destination host, connection request will be forwarded directly to destination, else the request will be forwarded after making an entry in waiting queue. R connection requests per N seconds will be continuously popped out from the waiting queue. Despite popping, if the length of queue exceeds certain threshold value T2, corresponding host will be considered suspicious and will be informed by the controller.

Till any attack is found, all the modules will be called one after another say connection success ratio, throttling connection, and anomaly-based fuzzy IDS, respectively, for inspection of particular packet that entered early detection algorithms module. A host that remains to be normal in all the modules is considered as normal host and is not disconnected from network. Any host which is detected as malicious by any of the modules will be considered as attacker and is disconnected immediately by the controller.

Anomaly-Based Fuzzy Intrusion Detection System

Anomaly-based fuzzy IDS is built using machine learning technique. Anomaly-based IDS is achieved by training the machine learning component with normal behavior of network without intrusion. Any deviation from this behavior will be treated abnormal by IDS and reported to the controller. Also, the machine learning component is trained with some kinds of attacks so that the IDS detects

particular attacks exactly when it encounters and informs the controller. Fuzzy approach is used for the development of anomaly-based IDS for the following reasons. One such reason is that intrusion detection problem involves many features that are quantitative which could be viewed as fuzzy variables. The second reason is that security itself includes fuzziness. Fuzzy logic allows concepts to be considered in more than one category, i.e., overlapping of categories. Rules are created using fuzzy logic and input network traffic is compared with these rules to find any abnormal behavior in the network. The training is given with the help of fuzzy rule base which covers normal behavior of the network without intrusion as well as attack types. Hence, the initial step in developing anomaly-based fuzzy IDS is fuzzy rule base creation.

Creation of Fuzzy Rule Base

Fuzzy rule base creation involves the following stages:

Dataset Collection: A precise dataset consisting of SDN adapted features of 500 training samples is collected. The dataset contains significant amount of records from KDD Cup 99 dataset covering normal network behavior and also two categories of attacks, namely dos and probe. Collected dataset also contains samples that are collected from SDN network for both normal and abnormal behavior. Every sample in the dataset contains below eight features as shown in Table 59.1 say (x1, x2, …, x8) in order terminated by class label (y) which is the output representing whether the sample with specified features is normal or attack. This is a supervised machine learning approach because the dataset tells the model the right answers, i.e., which is attack and which is normal.

Table 59.1 The features in the dataset are as follows

S.No	Selected features	Feature description
1.	Duration	Gives the length of the connection in number of seconds
2.	Protocol	Type of protocol used such as tcp, udp, etc.
3.	Flag	Gives the status information of connection such as normal or error
4.	Source bytes	Gives the number of data bytes sent from source to destination
5.	Destination bytes	Gives the number of data bytes from destination to source
6.	Urg	Gives the number of packets flagged urgent
7.	Packet count	Gives the total number of connections to the same host as the current connection in the past two seconds
8.	Diff_serv count	Gives the total number of connections to different services

The reason for training the model with such limited features arrives from the limitation of feature extraction in SDN environment. The controller receives only PACKET-IN message from switch and not the entire packet. PACKET-IN message contains only these limited features that could be extracted.

Preprocessing: *Quantization of non-numeric attributes:* The collected dataset contains features such as protocol, flag, and output label, namely class whose values are non-numeric. These non-numeric values need to be quantized to numeric values for ease of processing in upcoming stages.

Algorithm 2. Quanitize non-numeric attributes

```
Input : Collected dataset
Output : Quantized dataset

Let m be number of training samples (500)
Let x1,x2,..x8 represent features and x9 represent class in
dataset
for i from 1 to m do
  for j from 1 to 9 do
    if xj equals protocol then
      set xj to 0 for tcp
      set xj to 1 for icmp
      set xj to 2 for udp
    end if
    if xj equals flag then
      set xj to 2 for SYN
      set xj to 4 for RST
    end if
    if xj equals class then
      set xj to 0 for normal
      set xj to 1 for attack
    end if
   end for
end for
```

Normalization: The dataset contains only numeric values after quantization step. But the values in the dataset will not be in uniform range and vary widely. Hence, value of each and every feature (x1, x2, …, x8) in the dataset should be normalized to fit in uniform range (0, 1). Normalization is done using Eq. (59.1). The maximum and minimum values in each feature need to be calculated in order to proceed with normalization.

$$\text{NormAttrVal} = (\text{actualattrval} - \text{minval}(\text{attr}))/(\text{maxval}(\text{attr}) - \text{minval}(\text{attr})) \tag{59.1}$$

Algorithm 3. Normalize attribute values

```
Input : Quantized dataset
Output : Normalized dataset

Let m be number of training samples (500)
Let x1,x2,..x8 represent features in dataset
for i from 1 to 8 do
  for j from 1 to m do
    calculate max(xi) and min(xi)
   end for
end for
for i from 1 to 8 do
  for j from 1 to m do
    Normalized_Val =(actual_val - min(xi)) / (max(xi)-
min(xi))
  end for
end for
```

Defining membership function: Each and every feature in the normalized dataset is then mapped to a fuzzy variable. A fuzzy variable may take either discrete or continuous value. Fuzzy variables which take discrete values were defined using singleton membership function. For fuzzy variables which takes continuous values, fuzzy sets such as low, medium, and high were defined using triangular membership function. For example, triangular membership function for fuzzy variable source bytes is defined as follows:

```
FUZZIFY  source bytes
   TERM low1 :=  trian 0  0.0005  0.001;
   TERM medium1 := trian  0.001  0.013  0.025;
   TERM high1 := trian  0.02  0.51  1
ENDFUZZIFY
```

Membership value assignment is done by inference. With sufficient knowledge about the problem, the dataset is thoroughly analyzed and conclusions were inferred to assign values to membership function.

Deriving fuzzy rules: Once membership functions were defined, fuzzy rules are manually created with the knowledge of the problem. A typical fuzzy rule consists of complex fuzzy expression in the IF part followed by class label in THEN part. All possible combinations of fuzzy sets for all possible combinations of fuzzy variables constitute fuzzy rule base. A sample fuzzy rule is shown below:

```
if (PROTOCOL is TCP and FLAG is SH) and (COUNT is MEDIUM3 and
DIFF_SERV_COUNT is MEDIUM4) then CLASS is PROBE
```

Here, PROTOCOL and FLAG are discrete fuzzy variables, whereas COUNT and DIFF_SERV_COUNT are continuous fuzzy variables. CLASS label tells that the packet is of attack type and is of PROBE category. The machine learning component is trained with this fuzzy rule base.

Intrusion Detection Using Anomaly-Based Fuzzy IDS

The data packets from the controller were sent to anomaly-based fuzzy IDS to detect intrusion. Also, the attacks that are not dealt by early detection algorithms (i.e., packets which resulted normal from early detection algorithms) were forwarded to anomaly-based fuzzy IDS for further investigation. Hence, the input to this fuzzy IDS is PACKET-IN message and essential features from either controller in case of data packets or from early detection algorithms in case of request packets.

Since the machine learning model is trained with fuzzy rule base, it is essential to map the input to fuzzy before carrying out comparison. The same preprocessing steps such as quantization and normalization are applied to the input traffic. Then, the features are mapped to fuzzy variables. Then, the fuzzy inference system maps the values of input fuzzy variable to fuzzy sets and compares input with that of fuzzy rules and calculates support value for all rules. The rule whose support value is maximum is selected and the final outcome of IDS is the class of the selected rule. For example, consider the following fuzzy rules:

```
RULE 1: if (PROTOCOL is TCP and FLAG is SH) and (COUNT is
MEDIUM3 and DIFF_SERV_COUNT is MEDIUM4) then class is PROBE
```

```
RULE 2: if (PROTOCOL is TCP and FLAG is SH) and (COUNT is
LOW3 and DIFF_SERV_COUNT is MEDIUM4) then class is NORMAL
```

Then the support value for each fuzzy rule is calculated as

TV(*normal*) = TV(condition) * weight
TV(*abnormal*) = TV(condition) * weight

where condition is a complex expression involving fuzzy variables and weight is a real number which defines the strength of the rule [8]. There exist various techniques to discover the class label to which an object belongs to. Maximum technique is one such technique which classifies based on maximum support value.

Class is normal if TV(*normal*) > TV(*abnormal*)
Class is abnormal if TV(*normal*) < TV(*abnormal*).

The output fuzzy variable which is class label is defuzzified to crisp values which helps in plotting the outcome. If the class predicted by IDS is of attack type, then IDS sends information about attack to the controller which in turn takes immediate action. Thus, our work on developing intrusion detection system for

software-defined network results in enhancing security to a greater extent. All intrusion detection systems till date detect attacks and inform the administrator to take necessary action against the attack. This process of relying on the administrator may result in time delay to take appropriate action. Meanwhile, the attacker would have accomplished his task. But our idea of adapting IDS to SDN helps in taking immediate action by controller preventing the attacker to proceed further, thus avoiding huge damage.

Experimental Framework

In section "Experimental Setup", details about experimental setup and descriptions about the dataset collected are given. Section "Evaluation Metrics" discusses about various evaluation metrics and performances of our proposed system.

Experimental Setup

For training the machine learning component in IDS, dataset has to be collected. KDD Cup 99 dataset which is a standard dataset for network intrusion detection system is used. The original size of KDD Cup 99 dataset is huge. Hence, a small subset of records were selected from KDD Cup 99. Remaining records for the dataset were collected from the behavior of software-defined network for both normal and abnormal cases. The dataset comprises features for attacks such as dosnuke, pingofdeath, ipsweep, smurf, and tcpreset. Various attacks are grouped under the following four categories in KDD Cup 99 dataset. They are as follows:

Denial of Service (DOS): DOS attack makes a particular resource become exhausted or unavailable to users by sending numerous requests from various hosts in the network.

Remote to Local (R2L): R2L attacks are a type where an attacker gains local access to a machine on a network without having any account on that machine by sending packets to it over the network.

User to Root (U2R): User-to-root attacks are deployed by attackers who gain root access to particular host by cracking passwords, pins, etc.

Probing: Probe attack is a type where an attacker collects significant information which like machines, services, etc. could be accessed on a given network, and discovers well-known vulnerabilities by examining the network which is useful for attacking in future.

To setup software-defined network, Mininet emulator is used. By using Mininet, a network of virtual hosts, switches, controllers, and links could be created. Real-time network can be simulated using Mininet environment. Mininet can be

used to create custom topology and connect it to remote controller. Several researches have been carried out using Mininet by simulating the network and observing how it responds to various tests. We use Floodlight as the controller. Floodlight is the SDN controller that works with OpenFlow protocol to manage traffic in SDN environment. Floodlight installs and maintains flow table entries. Modules can be easily embedded in Floodlight controller. It offers developers the ability to easily adapt software and develop applications which are written in Java. Third-party applications can be created and can be integrated to Floodlight using REST APIs. Early detection algorithms and fuzzy-based intrusion detection system were developed as modules and embedded within the controller in Eclipse environment. Fuzzy logic is implemented using fuzzy control language. JFuzzyLogic is used to integrate fuzzy control language with Java.

Evaluation Metrics

The following evaluation metrics are used to measure the performance of the SDN controller in the presence of early detection algorithms:

Latency: The time interval between request packets being sent to the controller and the response being received from the controller is termed as Latency. This evaluation metric helps in measuring the performance of the SDN controller when the traffic is not much heavy. For achieving high performance, latency should be low.

Throughput: It represents number of packets processed by controller per unit time. This evaluation metric helps in measuring the performance of the controller when the traffic is much heavy.

The performance of the SDN controller is evaluated in the following different scenarios:

- SDN controller runs with only one early detection algorithm being enabled at a particular point of time.
  - That is, both the algorithms are enabled individually (one after another but not at the same time).
- SDN controller runs with both early detection algorithms being enabled.

In this paper, several metrics were selected for analyzing the performance of intrusion detection system.

TP = number of detected attacks and it is truly attacked (True positive).
TN = number of detected normal instances and it is truly normal (True Negative).
FP = number of detected attacks but it is truly normal (False positive).
FN = number of detected normal instances but it is truly attacked (False negative).

Accuracy: It is the proportion of total number of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{59.2}$$

Intrusion Detection Rate: It represents the accuracy rate for the attack classes. IDR is calculated as follows:

$$\text{IDR} = \frac{\sum_{i=0}^{1} TPi}{\sum_{i=0}^{1} TPi + FNi}, \tag{59.3}$$

where i is the class type. If i = 0, it denotes DOS class and if i = 1 it denotes probe class.

True Positive rate: TPR is when it is truly an attack and how often does it predict it as attack:

$$\text{TPR} = \frac{TP}{FN + TP} \tag{59.4}$$

False Positive rate: FPR is when it is truly not an attack and how often does it predict it as attack:

$$\text{FPR} = \frac{FP}{TN + FP} \tag{59.5}$$

False Alarm rate: FAR is the percentage of false negatives found. The false alarm rate can be calculated as follows:

$$\text{FAR} = \frac{FP}{TP + FP} \tag{59.6}$$

The performance of IDS by various metrics can be visualized from Fig. 59.2. This graph makes sense that the model detects attacks at a better rate with less number of false alarms.

The rate of correct detection and false alarms for each and every attack can be seen from Fig. 59.3. From the graph, it is concluded that IDS detects dosnuke and tcpreset attack types at a better rate with small number of false alarms. PingofDeath and ipsweep categories suffer considerable amount of false alarms. IDS sees equal detection and alarm rate in the case of smurf attack. This is because of limited features that are extracted from SDN packets.

ROC graph can be seen from Fig. 59.4 where all the test sets get plotted above threshold proving that the model is a good classifier.

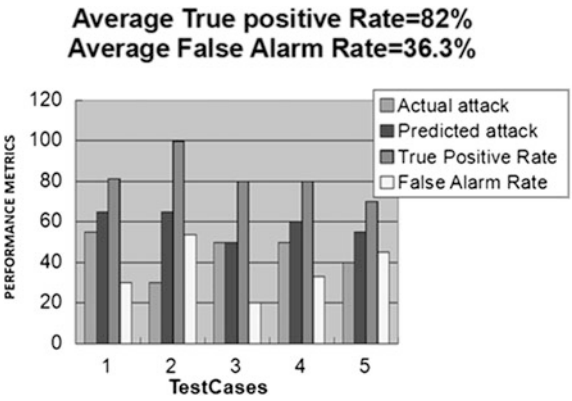


Fig. 59.2 IDS evaluation

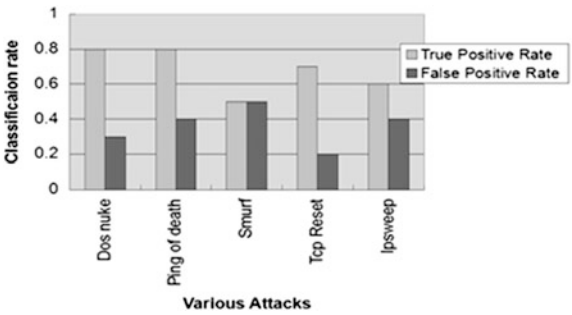


Fig. 59.3 TPR and FPR per attack

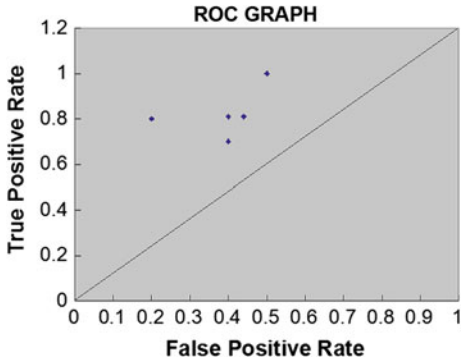


Fig. 59.4 Region operational characteristic

Conclusion

In this work, a better approach for the development of IDS to detect attacks in software-defined network compared to other techniques is proposed. Early detection algorithms employed in detecting attacks at early stage do not degrade the

performance of the system. Thus, the combination of both early detection algorithms and intrusion detection system has resulted in achieving better security in the field of SDN. Several testsets are employed in evaluating the performance of IDS. The results are quite good. The good behavior of this model is derived from the advantage of fuzzy logic. The machine learning component is trained with precise dataset collected from KDD Cup 99 and SDN containing both normal behavior and several attack types. Experimental results portray the good performance of our system in correctly identifying attacks with high accuracy which are already trained. This better performance paves way to explore several ways for future work. The efficiency of the system could be increased by training the machine learning component with large dataset containing diverse attack types. Genetic algorithms [13] could also be applied to obtain accurate and compact rule base which helps in attaining maximum accuracy.

References

1. Dabbagh, M., Hamdaoui, B., Guizani, M., Rayes, A.: Software Defined Networking Security—Pros and Cons, IEEE Communications Magazine 53(6) 73–79 (2015); Published by IEEE. ISSN: 0163-6804.
2. Xia, W., Wen, Y., Foh, C.H., Niyato, D., Xie, H.: A survey on software defined networking. IEEE Communication Surveys and Tutorials 17 (1) 27–50 (2015); Published by IEEE. ISSN 1553-877X.
3. Hu, Z., Wang M., Yan X., Yin Y., Luo Z.: Comprehensive Security Architecture for SDN. 18th International Conference on Intelligence in Next Generation Networks, February 17-February 19, pp 30–37. IEEE, Shanghai (2015).
4. Bernardos, C.J., Oliva, A., Serrano, P., Banches, A., Contreras, L.M., Jin, H., Zuniga, J.C.: An Architecture for Software Defined Wireless Networking. IEEE Wireless Communications 21(3) 52–61 (2014); Published by IEEE. ISSN 1536-1284.
5. Polla, M.L., Martinelli, F., Sgandurra, D.: A survey on security for mobile devices. IEEE Communications Surveys and Tutorials 15(1) 446–471(2012); Published by IEEE. ISSN: 1553-877X.
6. Bridges, S.M., Vaughn, R.B.: Fuzzy Data Mining and Genetic algorithms applied to intrusion detection. National Information Systems Security Conference, October 16–October 19. Baltimore MD (2000).
7. Elhang, S., Fernandez, A., Bawakid, A., Alshomrani, S., Herrera, F.: On combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems. Expert Systems with Applications 42(1) 193–202 (2015).
8. Cerli, A.A., Ramamoorthy, S.: Intrusion Detection System by Combining Fuzzy Logic with Genetic Algorithm. Global Journal of Pure and Applied Mathematics 11(1) 105–111 (2015); ISSN 0973-1768.
9. Mankad, K., Sajja, P.S., Akerkar, R.: Evolving Rules Using Genetic Fuzzy Approach-An Educational Case Study. International Journal on Soft Computing 2(1) 35–46 (2011).
10. Hassan, M.M.M.: Network Intrusion Detection System Using Genetic Algorithm and Fuzzy Logic. International Journal of Innovative Research in Computer and Communication Engineering 1(7) 1435–1445 (2013); ISSN 2320-9801 (Online), ISSN 2320-9798 (Print).

11. Dotcenko, S., Vladyko, A., Letenko, I.: A Fuzzy Logic-Based Information Security Management for Software-Defined Networks. 16th International Conference on Advanced Communication Technology, February 16–February 19, pp 167–171. IEEE (2014).
12. Jin R., Wang B.: Malware Detection for Mobile Devices Using Software-Defined Networking. Second GENI Research and Educational Experiment Workshop, March 20–March 22, pp 81–88. IEEE, Washington (2013).
13. Reddy, R., Ramadevi, Y., Sunitha, K.: Fuzzy logic and genetic based Intrusion Detection System. International Journal of Advanced Research in Computer Science and Software Engineering 5 (8) 792–797 (2015); ISSN 2277-128X.

Chapter 60
A (t, n) Multi-secret Sharing Scheme with Updated Secret Shadows

Shyamalendu Kandar and Bibhas Chandra Dhara

Abstract In multi-secret sharing, a number of secrets are shared among a group of participants in a single-secret sharing process. Each participant gets only one share from the sharing process and secrecy of one secret does not depend on the reconstruction of any other secrets. YCH method is a well-known multi-secret sharing scheme proposed by Yang, Chang, and Hwang. This scheme has used two-variable one-way function. Among these two variables, one is made public and another is a randomly distributed secret shadow for each participant. In the current work, YCH method is extended where the "public variable" is shared and each participant gets one of them. The shadows are updated periodically to prevent any adversary from collecting all the secret shadows within a stipulated time period.

Keywords Multi-secret sharing · Secret shadow · One-way function · Updated secret shadow

Introduction

Some real-life applications require a secret information to be shared among a group of persons. As an example, for missile launching, it requires a secret code. If it is kept under the possession of a single person, there is a threat of misuse of power. In this case, secret sharing can be chosen to play its role. Secret sharing is a part of information security evolved from the context of safeguarding cryptographic keys. In 1979, Shamir [1] and Blakley [2] individually proposed secret sharing scheme. Lagrange interpolation polynomial is the backbone of Shamir's scheme where hyperplane

S. Kandar (✉)
Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur 711103, West Bengal, India
e-mail: shyamalenduk@it.iiests.ac.in

B. C. Dhara
Department of Information Technology, Jadavpur Univesity, Salt Lake Campus, Kolkata 700098, West Bengal, India
e-mail: bibhas@it.jusl.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_60

geometry is used in Blakley's method. Using Chinese Remainder Theorem (*CRT*), a secret sharing scheme is proposed by Asmuth and Bloom [3]. Secret sharing is found useful in managing missile launching code, digitally operated secret vault key management, e-auction, etc.

In (t, n), threshold secret sharing scheme from secret n shares are generated (by dealer) and distributed to the participants so that secret can be obtained (by combiner) by using at least t shares. The success of (t, n) sharing scheme is based on the honesty of dealer and participants. Since dealer has no role after distribution of shares, the dealer entity is deleted.

In classical secret sharing scheme, single secret is shared in one secret sharing process and that particular is retrieved in the reconstruction process. Multi-secret sharing scheme (MSSS) is very useful in case of multiple secrets. In MSSS, a number of secrets can be shared in a single-sharing process without consuming extra data as compared to single secret sharing.

MSSS classified as either (i) single-time use scheme or (ii) multi-time use scheme. In the first category, dealer issues fresh shares to the participants after reconstruction of some or all secrets. In the second category, each participant needs to keep single share known as secret shadow and dealer does not require to resend new secret shadows after reconstruction of some or all secrets.

In [4], a multistage secret sharing is addressed based on one-way function and public shift technique to share multiple secrets. In the scheme secret, shadows are obtained using public shift technique and one-way function is applied in stage by stage reconstruction of the secrets. Sharing p secrets among n participants, total public value is $p.n$. Chien et al. [5] have used systematic block codes in their proposed multi-secret sharing scheme. The advantages of the scheme are dynamic determination of the distributed shares, multi-use policy, and parallel reconstruction of secrets. This method is with less public values but highly time constraint due to its iterative nature and in each iteration, several linear equations need to be solved. This has made the method inefficient to use for large number of secrets.

In 2004, Yang et al. [6] have proposed a two-variable one-way function based on multi-secret sharing scheme (YCH scheme). One variable is made public and another one is randomly chosen secret shadows for each participant. The method has two cases: (i) $v \leq t$ and (ii) $v > t$ for v secrets in (t, n) multi-secret sharing scheme. Case (i) requires $(n + 1)$ public values and case (ii) needs $(n + v - t + 1)$ public values. After the reconstruction phase, dealer is initiated to start sharing of new set of secrets. Dealer just takes a new one-way function but does not require to choose fresh secret shadows for the participants. This makes YCH method a multi-use scheme. A recent proposal [7] tries to reduce the requirement of public values in YCH scheme. Authors have used set-wise co-prime number to address the same. Threshold number of such numbers reveals a variable by performing GCD. A MSSS scheme is addressed in [8]. The scheme used linear recursion and linear feedback shift register based public key cryptosystem.

In dynamic multi-secret sharing scheme [9], each participant holds one master share. Using the master share along with the threshold value, different groups of secrets can be reconstructed. Step-by-step execution of one-way function with

XOR operation provides more security in the scheme. A MSSS scheme with minimal linear code is available in [10]. Tentu et al. [11] used Asmuth-Bloom sequence to address a multistage multi-secret sharing. Some recent proposals of multi-secret sharing schemes are available in [12–14].

Image is multimedia data that consists of correlated pixels. In recent years, using and transmitting image in web is gaining its popularity in social media, medical applications, military, etc. Image sharing is considered as a type of multi-secret sharing scheme. Refs. [15–18] are some image secret sharing schemes available in literature. In image secret sharing scheme, shares are generated by taking group of pixels one at a time as secret; thus, the size of the shares becomes a fraction than that of the original image. Some multi-secret sharing proposals are addressed in [19, 20] which perform secret sharing of multiple images in one secret sharing process.

For a long-lived secret, there is a possibility of collecting shares by corrupting one or more participant by an adversary. This can be prevented by proactive secret sharing scheme [21]. In [22], a proactive secret sharing scheme is discussed where the shares of the participants are updated periodically keeping the secret same. The duration of the shares is partitioned into periods, and in each period the shares hold by each participant are updated by mutual exchange of data among the participants. The share accumulated at time period T is no use of adversary at time period $T + i, i = 1, 2, \ldots$.

Right shares collected from participants will reveal the original secret. Thus, the backbone of secret sharing lies on the trust over the participants. It is believed that dealer and participants are trusted. But in real-life application, one or more participant/s may submit a wrong share resulting in retrieval of wrong secret of no use. Cheating is a vulnerability of secret sharing scheme. Verifiable secret sharing (VSS) scheme deals with cheating detection/cheater identification in secret sharing. Some of the verifiable secret sharing schemes are available in [23–26].

Though YCH is a popular MSSS scheme, it has a number of disadvantages. In this article, we have highlighted some of the drawbacks of YCH method. In the current work, we have presented a new (t, n) MSSS scheme based on YCH method. In this work, no variable is made public and the secret shadows are updated periodically.

The organization of the paper is as follows. A brief description of YCH scheme is given in section "YCH Method". Drawbacks of YCH method are pointed in section "Drawbacks in YCH Method". Proposed scheme is presented in section "Proposed Method". Section "Performance Analysis" contains the performance analysis of the proposed scheme and conclusions are drawn in section "Conclusion".

YCH Method

The term YCH scheme is abbreviated after a multi-secret sharing proposal of Yang et al. [6] in 2004. $f(a, b)$ is a two-variable one-way function which is the backbone of the method. This type of function is mainly a cryptographic primitive used in several

fields. $f(a, b)$ is computationally easy but finding $f^{-1}(a, b)$ is very hard. Multi-use property of YCH scheme is achieved due to the following properties of $f(a, b)$:

(a) $f(a, b)$ is easily computable when a and b are available.
(b) Having knowledge of both $f(a, b)$ and b, it is difficult to compute a.
(c) Without any knowledge of b, computation of $f(a, b)$ is difficult though a is known.
(d) For a known b, finding a_i and a_j is computationally intensive when $f(a_i, b) = f(a_j, b)$.
(e) Finding b is difficult, if a and $f(a, b)$ are known.
(f) If a_i and $f(a_i, b)$ are given, then for $a_j (\neq a_i)$ finding the value of $f(a_j, b)$ is very hard.

In YCH scheme, a fixed length bit string is generated from f using a public value w and a secret shadow sw. This string is used to generate the shares for n participants from a set of v secrets $S_1, S_2, \ldots, S_v \in \mathbb{F}_p$ (For a prime field p) using (t, n) secret sharing. n secret shadows $sw_1, sw_2, \ldots, sw_n$ are randomly chosen by the dealer and sw_i is sent to $Participant_i$ using a secret channel. Multi-secret sharing is performed depending on the following two conditions:

I. If $v \leq t$

(a) A $(t-1)$ degree polynomial

$$g(x) = S_1 + S_2 x^1 + \cdots + S_v x^{v-1} + c_1 x^v + c_2 x^{v+1} + \cdots + c_{t-v} x^{t-1} mod\, p$$

is constructed by the dealer with random coefficients $c_i \in F_p$.
(b) The shares are computed as $z_i = g(f(w, sw_i)) mod\, p$ for $i = 1, 2, \ldots, n$
(c) $w, z_1, z_2, \ldots, z_n$ are published publicly following the methods proposed in [27, 28].

Here, total $(n + 1)$ values are publicly published.

II. If $v > t$

(a) A $(v-1)$ degree polynomial

$$g(x) = S_1 + S_2 x^1 + \cdots + S_v x^{v-1} mod\, p$$

is constructed by the dealer.
(b) Participants' share is computed as $z_i = g(f(w, sw_i)) mod\, p$ for $i = 1, 2, \ldots, n$.
(c) Dealer computes $g(i) mod\, p$ for $i = 1, 2, \ldots, (v - t)$
(d) $w, g(1), g(2), \ldots, g(v-t), z_1, z_2, \ldots, z_n$ are published publicly following the methods proposed in [27, 28].

Here, total $(n + v - t + 1)$ values are public.

Drawbacks in YCH Method

YCH scheme is well-known multi-use MSSS scheme. But it has a number of disadvantages listed below.

(i) According to YCH scheme, sw_i are valid throughout the lifespan of the secret sharing process. This leads to a threat to YCH scheme. An adversary may collect sufficient number of sw_i from the participants by corrupting or managing them. For a long-lived secret, adversary gets enough time to make some deal with the participants.
(ii) $f(w, sw_i)$ is easily computable if any intruder is able to collect sw_i, as w is made public in YCH scheme.
(iii) How the dealer is initiated to choose a new w as public variable is not addressed clearly in YCH scheme.

Proposed Method

In the current section, a new multi-secret sharing is proposed. This method takes an integer $w \in \mathbb{F}_p$ (For a prime number p) as one of the two variables of the one-way function. The integer w is shared to $sh_i, i = 1, \ldots, n$ using Shamir secret sharing [1]. $w - (n - i)$ acts as the second variable of the one-way function where i is the participant number. v secrets are shared using YCH method. The pair (i, sh_i) is sent to $Participant_i$ using some secure channel. To protect w from any adversary, sh_i is updated periodically. The proposed method is described in the following section.

A. **Initialization** Let v secrets $(S_1, S_2, \ldots, S_v) \in \mathbb{F}_p$ are to be shared using (t, n) MSSS scheme. A random integer $w \in \mathbb{F}_p$ is chosen by the dealer and a $(t - 1)$ degree polynomial

$$Y(x) = w + C_1 x^1 + C_2 x^2 + \cdots + C_{t-1} x^{t-1}$$

is constructed with random coefficients $C_i \in p, i = 1, \ldots t - 1$ such that $P(0) = w$. The pair (i, sh_i) is sent to $Participant_i$ through a secure channel. A two-variable one-way function $f(a, b)$ is taken by the dealer and it is made publicly known.

B. **Share Generation**: Depending on the number of secrets, there are two cases in share generation process. (I) $v \leq t$ and (II) $v > t$.

Case I: If $v \leq t$:

(a) Dealer constructs

$$g(x) = S_1 + S_2 x^1 + \cdots + S_v x^{v-1} + c_1 x^v + c_2 x^{v+1} + \cdots + c_{t-v} x^{t-1} mod\ p$$

where $c_j, j = 1, 2, \ldots, (t - v)$ are random values taken as coefficients.

(b) $b_i = (w - (n - i))$ is computed and is taken as the second variable of f.
(c) $z_i = g(f(w, b_i))\ mod\ p$ is computed for $i = 1, 2, \ldots, n$.
(d) z_i, $i = 1, 2, \ldots, n$ are made public such that those obey [27, 28].

Total publicly published value is n.

Case II: If $v > t$:

(a) Dealer constructs

$$g(x) = S_1 + S_2 x^1 + \cdots + S_v x^{v-1} mod\ p$$

(b) $b_i = (w - (n - i))$ is computed and it is taken as the second variable of f.
(c) $z_i = g(f(w, b_i))\ mod\ p$ is computed for $i = 1, 2, \ldots, n$.
(d) $g(j)\ mod\ p$ is computed for $j = 1, 2, \ldots, (v - t)$.
(e) z_i, $i = 1, 2, \ldots, n$ and $g(1), g(2), \ldots, g(v - t)$ are made public such that those are in [27, 28].

Publicly published value is $(n + v - t)$.

C. **Secret Reconstruction**:
Case I: If $v \leq t$:

(a) Combiner chooses atleast t participants from the set of participants.
(b) $Participant_i$ $i = 1, \ldots, t$ sends the pair (i, sh_i) to the combiner.
(c) Combiner computes
$\sum_{i=1}^{t} sh_i \prod_{j=1, j\neq i}^{t} \frac{(-j)}{i-j} mod\ p$
to get w and finds $val_i = f(w, (w - (n - i)))$.
(d) A $(t - 1)$ degree polynomial

$$g(x) = \sum_{i=1}^{t} z_i \prod_{j=1, j\neq i}^{t} \frac{(x - val_i)}{val_i - val_j} mod\ p$$
$$= S_1 + S_2 x^1 + \cdots + S_v x^{v-1} + c_1 x^v + c_2 x^{v+1} + \cdots + c_{t-v} x^{t-1} mod\ p$$

is constructed from t pairs of (val_i, z_i) and the secrets S_i; $i = 1, \cdots, v$ are retrieved.

Case II: If $v > t$:

(a) Combiner chooses atleast t participants from the set of participants.
(b) $Participant_i$ $i = 1, \ldots, t$ sends the pair (i, sh_i) to the combiner.
(c) Combiner computes
$\sum_{i=1}^{t} sh_i \prod_{j=1, j\neq i}^{t} \frac{(-j)}{i-j} mod\ p$
to get w and finds $val_i = f(w, (w - (n - i)))$.
(d) The $(v - 1)$ degree polynomial

$$g(x) = \sum_{i=1}^{t} z_i \prod_{j=1, j\neq i}^{t} \frac{(x - val_j)}{(val_i - val_j)} + \sum_{i=1}^{v-t} g(i) \prod_{j=1, j\neq i}^{v-t} \frac{x-j}{i-j} mod\, p$$
$$= S_1 + S_2 x^1 + \cdots + S_v x^{v-1} mod\, p$$

is constructed from t pairs of (val_i, z_i) and $(v - t)$ pairs of $(i, g(i))$ and the secrets are retrieved.

D. **Updating the secret shadows**: In the proposed scheme, $f(w, b_i)$ is fully unknown to some participant as w is unknown to him and thus no question of knowing b_i arises. Having sh_i unchanged throughout the lifetime of the secret sharing raises the threat of getting sh_i from some corrupted $Participant_i$ by some adversary. Being able to collect sufficient (i, sh_i) pair, w is made disclosed. To remove the threat, sh_i is updated periodically in such a way that value of w remains intact in reconstruction process from the shares accumulated from a particular period. The updation process is done by message exchanging among participants as described in the following section.

(a) $Participant_i$, $i = 1, \ldots, n$ individually constructs $(t - 1)$ degree polynomial

$$V_i^q(x) = C_{1,i}^q x + C_{2,i}^q x^2 + \cdots + C_{(t-1),i}^q x^{(t-1)} mod\, p$$

with random coefficients $C_{j,i}^q$ for $j = 1, \ldots, (t - 1)$,
where q denotes the number of period (times) the secret shadows are updated. It is to be noted that $C_{0,i}^q$ is 0 for all q and i.

(b) $sh_{i,j}'^q$ denotes the jth share generated from the polynomial taken by $Participant_i$ in period q and computed as $sh_{i,j}'^q = V_i^q(j)$ for $j = 1 \cdots n$. $Participant_i$ sends $sh_{i,j}'^q$ to $Participant_j$ for $j \neq i$.

(c) $Participant_i$ updates its secret shadow by computing $sh_i^q = (sh_i^{q-1} + \sum_{j=1}^{n} sh_{j,i}'^q)\ mod\ p$ and destroys the old shadows.

Performance Analysis

In the proposed method, w is not public, and thus one less public value is required to made public than YCH method. Another variable of f is computed as $(w - (n - i))$. For each participant having no knowledge of w, it is impossible to guess the value from f. Thus, the property of two-variable one-way function is preserved. Periodical updation of the secret shadows removes the threat of collecting sufficient number of such from a particular period by corrupting the participants. Shares collected at period q are of no use in period $q + i$, $i > 0$. Collecting t shares belonging

to a particular period can retrieve w as in each polynomial taken by the participant, $C_{0,i}^{q}$ is zero [22]. The duration of the period is set in such a way that the adversary cannot get sufficient time to pollute significant number of participants.

Conclusion

An MSSS scheme based on YCH method is proposed in this paper. In the proposed method, none of the two variables of f is made public but one of the variables w is shared among the participants. Thus, no participant is able to calculate f as w is unknown to him. Even any adversary fails to get w if less than t participants are managed. For long-lived secret, there is always a threat of disclosure of shares from the participants by some unauthorized persons as it gets enough time to make sufficient number of participants corrupt. This is prevented by updating the shares generated from w in each time period. Shares gathered by an adversary at period q are of no use in period $q + i$. The updating is done in such a way that any t shares accumulated from a particular time period can retrieve w.

The major drawback of the proposed scheme is that it is a one-time use scheme. After retrieval of some or all the secrets, dealer has to take new w and new f to share another set of secrets. Taking the burden of updating the shares generated from w and sending information to each participant after a certain time interval are pointed as other disadvantages of the scheme.

Cheating is a rising concern in information security and thus in secret sharing also. During the updation process in some period, any participant acting as a cheater may provide arbitrary information to other participants. This spoils the value of w and the secrets retrieved are the fake ones. Development of a verifiability scheme with cheating detection over the proposed technique is left for future work.

References

1. Shamir, Adi.: How to share a secret. Communications of the ACM 22.11, 612–613 (1979)
2. Blakley, George Robert.: Safeguarding cryptographic keys. Proceedings of the national computer conference. Vol. 48, pp. 313–317 (1979)
3. Asmuth, C., Bloom, J.: A modular approach to key safeguarding. IEEE transactions on information theory, 30(2), 208–210 (1983)
4. He, Jingmin, and Edward Dawson.: Multistage secret sharing based on one-way function. Electronics Letters 30.19, 1591–1592 (1994)
5. Chien, Hung-Yu, J. A. N. Jinn-Ke, and Yuh-Min Tseng.: A practical (t, n) multi-secret sharing scheme. IEICE transactions on fundamentals of electronics, communications and computer sciences 83.12, 2762–2765 (2000)
6. Yang, Chou-Chen, Ting-Yi Chang, Min-Shiang Hwang: A (t, n) multi-secret sharing scheme. Applied Mathematics and Computation 151.2, 483–490. (2004)
7. Kandar, Shyamalendu, Bibhas Chandra Dhara: A (k, n) Multi Secret Sharing Scheme Using Two Variable One Way Function with Less Public Values. International Conference on Information Systems Security. Springer International Publishing, 532–541. (2015)

8. Mashhadi, Samaneh, Massoud Hadian Dehkordi.: Two verifiable multi secret sharing schemes based on nonhomogeneous linear recursion and LFSR public-key cryptosystem. Information Sciences 294, 31–40 (2015)
9. Lin, Han-Yu, Yi-Shiung Yeh.: Dynamic multi-secret sharing scheme. Int. J. Contemp. Math. Sciences 3.1, 37–42. (2008)
10. Song, Yun, Zhihui Li, Yongming Li, Jing Li.: A new multi-use multi-secret sharing scheme based on the duals of minimal linear codes., Security and Communication Networks 8, no. 2, 202–211, (2015)
11. Tentu, Appala Naidu, V. Kamakshi Prasad, V. Ch Venkaiah.: Multi-Stage Secret Sharing Schemes Based on Asmuths Bloom Sequence, Networking and Communication Engineering 8, no. 3, 75–79, (2016)
12. Rao Subha, Bhagvati, C.: CRT based threshold multi secret sharing scheme. International Journal of Network Security 16.4, 249–255 (2014)
13. Endurthi, A., Bidyapati Chanu, O., Naidu Tentu, A., Venkaiah, V. C.: Reusable Multi-Stage Multi-Secret Sharing Schemes Based on CRT. Journal of Communications Software and Systems, 11.1 (2015)
14. Dong, Xuedong.: A Multi-secret Sharing Scheme Based on the CRT and RSA. International Journal of Network Security 2.2, 69–72 (2015)
15. Thien, C. C., Lin, J. C.: Secret image sharing. Computers and Graphics, 26(5), 765–770, (2002)
16. Bai, Li.: A reliable (k, n) image secret sharing scheme. Dependable, Autonomic and Secure Computing, 2nd IEEE International Symposium. IEEE, 31–36 (2006)
17. Das, S. K., Dhara, B. C.: An Image Secret Sharing Technique with Block Based Image Coding. Communication Systems and Network Technologies (CSNT), 2015 Fifth International Conference, IEEE, 648–652, (2015)
18. Zarepour-Ahmadabadi, Jamal, MohammadEbrahim Shiri Ahmadabadi, AliMohammad Latif.: An adaptive secret image sharing with a new bitwise steganographic property., Information Sciences 369, 467–480, (2016)
19. Rajput, Mohit, Maroti Deshmukh.: Secure (n, n+ 1) Multi Secret Image Sharing Scheme Using Additive Modulo, Procedia Computer Science 89 677–683, (2016)
20. Deshmukh, M., Nain, N., Ahmed, M.: An (n, n) Multi Secret Image Sharing Scheme Using Boolean XOR and Modular Arithmetic, 30th International Conference on Advanced Information Networking and Applications(AINA), IEEE, 690–697, (2016)
21. Ostrovsky, R., Yung, M.: How to withstand mobile virus attacks. Proceedings of the tenth annual ACM symposium on Principles of distributed computing, ACM, 51–59, (1991)
22. Herzberg, Amir, Stanislaw Jarecki, Hugo Krawczyk, Moti Yung.: Proactive secret sharing or: How to cope with perpetual leakage. In Annual International Cryptology Conference, Springer Berlin Heidelberg, 339–352, (1995)
23. Feldman, Paul.: A practical scheme for non-interactive verifiable secret sharing, Foundations of Computer Science, 1987, 28th Annual Symposium on, IEEE, 427–438, (1987)
24. Pedersen, Torben Pryds.: Non-interactive and information-theoretic secure verifiable secret sharing. Annual International Cryptology Conference, Springer Berlin Heidelberg, 129–140, (1991)
25. Gan, Yuanju, Lihua Wang, Licheng Wang, Ping Pan, Yixian Yang.: Publicly Verifiable Secret Sharing Scheme with Provable Security against Chosen Secret Attacks, International Journal of Distributed Sensor Networks (2013)
26. Harn, Lein, Miao Fuyou, Chin-Chen Chang.: Verifiable secret sharing based on the Chinese remainder theorem., Security and Communication Networks 7, no. 6, 950–957, (2014)
27. Rivest, Ronald L., Adi Shamir, Len Adleman.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM 21.2, 120–126, (1978)
28. ElGamal, Taher: A public key cryptosystem and a signature scheme based on discrete logarithms. Advances in cryptology. 10–18 Springer Berlin Heidelberg, (1985)

Chapter 61
Design and Analysis of LFSR-Based Stream Cipher

Subhrajyoti Deb, Bubu Bhuyan and Navin Ch. Gupta

Abstract Stream ciphers are increasingly used for lightweight applications like RFID, Bluetooth, and Wi-Fi communications where the length of the plaintext is initially unknown. Generally, the stream ciphers are characterized by fast encryption and decryption speed. The LFSR-based stream cipher can generate pseudorandom binary strings with good cryptographic properties. Hardware implementation cost is also minimum for it. In this paper, we have discussed the architecture and properties of the LFSR. We have also discussed the properties of secured Boolean function as one of the important components of stream cipher. Here, we have implemented a generalized nonlinear combination of generator-based model comprising LFSR/NLFSR and Boolean function for designing a pRNG. The bitstream properties of pRNG are tabulated and compared with their best attainable parameters.

Keywords LFSR · Stream · Cryptography · Sagemath

Introduction

Cryptography is the study of mathematical techniques used for achieving secured communication over the unsecured channel. Cryptographic primitives are designed to deal with the basic security issues like confidentiality, integrity, authentication, and non-repudiation. It can be classified into two categories, namely symmetric key and asymmetric key primitives. Symmetric key cryptographic primitives have the advantage of higher throughput over asymmetric ones, and therefore symmetric key cryptographic primitives are widely used for bulk data encryption and decryption.

S. Deb (✉) · B. Bhuyan · N. C. Gupta
Department of Information Technology, North-Eastern Hill University, Shillong, India
e-mail: subhrajyotideb1@gmail.com

B. Bhuyan
e-mail: b.bhuyan@gmail.com

N. C. Gupta
e-mail: rainynavin@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_61

Symmetric key ciphers are again subdivided into two categories, namely block cipher and stream cipher. The stream ciphers have the advantages of higher throughput, low latency, and lesser error propagation effect than that of block cipher. The basic working principle of the stream cipher is to generate an arbitrarily long pseudorandom bit stream from a given random string and that pseudorandom bitstream is used to encrypt the message stream. Further, stream cipher based on LFSR (Linear Feedback Shift Register) is characterized by its lightweight property and ease of implementation in hardware. A few examples of popular stream ciphers are A5/1 used in GSM security, E0 used in Bluetooth, RC4 used in SSL, etc. A few prominent lightweight stream ciphers are Grain, WG, Trivium, SNOW, Salsa 20, Sprout, SOBER, etc. [1–3].

The outline of the paper is as follows: section "Literature Survey" provides the literature survey. Section "Background and Preliminaries" provides the background and preliminaries related to stream cipher. Section "Analysis of Feedback Shift Registers for Stream Cipher" provides the FSR analysis part. Section "Proposed Design" describes the architecture of implemented nonlinear generator model and its result analysis. Finally, section "Conclusion and Future Work" concludes the paper.

Literature Survey

Here, we have discussed the different types of LFSR-based stream ciphers. A renewed interest has grown among the research communities for analysis and design of stream ciphers due to the launch of eSTREAM project [1]. This research project was maintained by European Network of Excellence for Cryptology from 2004 to 2008. Only seven candidates are chosen from long-term research project in Europe known as ECRYPT [4]. In Table 61.1, we have listed a few LFSR-based stream ciphers and their building process [2, 3].

Background and Preliminaries

In this section, we have included mathematical preliminaries related to LFSR-based stream cipher design.

Boolean Function

A Boolean function is a mapping from $F_2^n \rightarrow F_2$, over the finite field with two elements$\{0, 1\}$. If the number of combination mapping consists of an equal number of $1's$ and $0's$, then the Boolean function is called as balanced.

Table 61.1 LFSR-based stream cipher list

Cipher name	Platform used	Building process
A5/1	Hardware	Combination of three LFSRs with irregular clocking
E0	Hardware	LFSR of length 4
Sosemanuk	Software	LFSR of length 10
HC-256	Software	Nonlinear Filter 32-bit-to-32-bit mapping Linear masking
Trivium	Hardware	NLFSR (Nonlinear Feedback Shift Register) of lengths 93, 84, and 111
Grain	Hardware	NLFSR of length 128 and LFSR of length 128
Sprout	Hardware	LFSR and NLFSR of the length 40
WG	Hardware and software	LFSR of length 11
Espresso	Hardware	256-bit NLFSR are in Fibonacci configuration
Decim v2	Hardware	LFSR of length 192
Decim-128	Hardware	LFSR of length 288

For example, let us consider $n = 3$ variable Boolean function $f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3$. The input sequences of (x_1, x_2, x_3) are (000, 100, 010, 110, 001, 101, 011, 111) and the final output of the Boolean function is depicted as (0, 0, 0, 1, 1, 1, 0, 1).

A cryptographically secured Boolean function should satisfy the following properties [5, 6]:

- Boolean function should be balanced.
- The nonlinearity and correlation immunity of the function should be high so that it can resist correlation attack.
- The algebraic degree and algebraic immunity of the function should be high so that it can resist algebraic attack.

Algebraic Normal Form

Usually, every Boolean function has a distinct representation as a multivariate over F_2 which is known as algebraic normal form (ANF). This function can be represented as

$$f(x_1, x_2, \ldots, x_n) = c_0 \oplus \sum_{1 \le i \le n} c_i x_i \oplus \sum_{1 \le i \le j \le n} c_{i,j} x_i x_j \oplus \ldots \oplus c_{(1,2,\ldots,n)} x_1 \ldots x_n$$

where the coefficients $c_0, c_i, c_{i,j}, \ldots, c_{(1,2,\ldots,n)} \in F_2$. In this function, the number of variables in the highest order product term (with coefficient non zero) is known as the algebraic degree. In general, when the degree of the function f is at most one, it can be described as an affine function. The affine functions with ($c_0 = 0$) are known as linear functions [6, 7].

Walsh Transform

This transformation function is an n variable Boolean function. In that case, $c = \{c_1 \ldots c_n\} \in F_2^n$ and a n variable linear function can be represented as $l_c(x) = c_1 x_1 \oplus \ldots \oplus c_n x_n$. So, this transformation function can be described as

$$W_f(c) = \sum_{x \in F_2^n} (-1)^{f(x) \oplus l_c(x)}$$

From the above definition of $W_f(c)$, it can be observed that, when $f(x) \oplus l_c(x)$ value is 0, then sum is increased by 1, and when this value is 1, sum is decreased by 1 [6, 7].

Nonlinearity

Nonlinearity of a Boolean function f of n variables can be described as the distance between the function and the set of all possible affine functions. Nonlinearity can be defined in terms of Walsh transform as given below [7, 8]:

$$\mathrm{nl(f)} = 2^{n-1} - \frac{1}{2} \, max \, |W_f(c)|$$

Correlation Immunity

A Boolean function f on F_2^n is said to be correlation of order m, where $1 \le m \le n$, if the output of f and any m input variables are statistically independent. A Boolean function f on F_2^n is correlation immune of order m iff $W_f(c) = 0$ for all vectors $c \in F_2^n$ such that $0 \le |c| \le m$ [6, 8].

Analysis of Feedback Shift Registers for Stream Cipher

The normal way of designing Feedback Shift Register (FSR) through binary sequences, $\{c_0, c_1, \ldots c_n\} \in GF(2)$, fulfills the recurrence relation of order n. FSR is created by *D Flip-flops* which are connected serially and each *D Flip-flop* constructed in such a way that each gate is simulating the Boolean logic for feedback function. Moreover, if FSR runs with linear recurrence, feedback function is known as LFSR and if it runs with nonlinear recurrence, then feedback function is known as NLFSR [9]. eStream selected Grain, Trivium, and MICKEY stream cipher are designed by NLFSR [1].

LFSR

Linear Feedback Shift Register (LFSR) is a shift register with a feedback path. Here, the output sequence of each *D Flip-flop* is joined to the input of the adjacent *D Flip-flops*. Feedback path is defined as the `tap` position of *D Flip-flop* which takes part in the XOR (modulo 2) operation and provides input to the last *D Flip-flop*. Initial value of LFSR is known as seed value of LFSR. The feedback path is also known as feedback function or connection polynomial [9, 10].

For example, let us consider a LFSR degree is $m = 3$. The LFSR structure and feedback path are shown in Fig. 61.1. Here, this feedback path can be represented in polynomial form as $(x^3 + x^2 + 1)$. The internal state bits are expressed as a_i and it has been shifted by one to the right at each clock. In that case, rightmost state bit is considered as present output bit and the leftmost state bit is calculated by feedback path. Let us consider the output bit is a_i and assuming the initial state bits are (a_0, a_1, a_2).

Now, the output sequence of the LFSR can be calculated as $a_{i+3} = a_{i+1} + a_i \bmod 2$, where $i = 0, 1, 2, 3, \ldots,$.

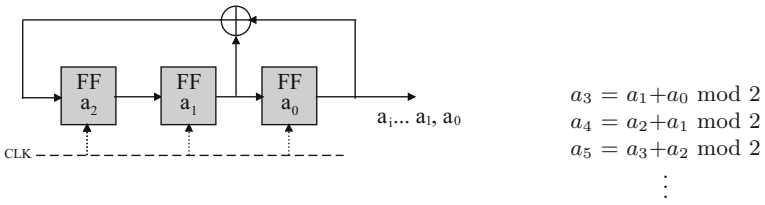


Fig. 61.1 Schematic diagram of LFSR

Properties of LFSR

- In l-stage LFSR, if l number of registers are available in the LFSR, then the number of states is equal to 2^{l-1}.
- However, every feedback path or connection polynomial will not give maximum length. The LFSR will yield maximum length if and only if the corresponding feedback path is primitive polynomial.

Klapper and **Goresky** developed similar type of LFSR design, known as Feedback with Carry Shift Register (FCSR). There are two different types of LFSR, namely Fibonacci and Galois [11]. In FCSR and LFSR, linear sequences are possible to employ through Berlekamp–Massey algorithm [10, 11].

LFSR-Based Stream Ciphers

The main use of LFSR in stream cipher is to produce pseudorandom sequence. We know, LFSR can generate an infinite bitstream. In the most common form, multiple LFSRs are used to build a stream cipher. But LFSR exhibits linear property. Thus the nonlinearity concept has been introduced to overcome the drawback of linear property [12] by irregularly changing the clock of the LFSR. LFSR-based stream cipher uses mainly three classes of pRNGs (pseudorandom number generators), namely *nonlinear combination, nonlinear filter, and clock-controlled generators* [13]. Almost every LFSR-based stream ciphers follow any of these nonlinear techniques or use a combination of these techniques with some extra efforts like adding counter or a combination of different LFSRs with NLFSR [1]. Usually, nonlinear combiner design employs n number of LFSRs of different lengths. In that case, all are initially assigned with nonzero seed values. During each clock pulse, n number of results from the LFSRs are taken and filled as n data inputs to an n variable Boolean function. In case of nonlinear filter generator, n numbers of outputs from different positions of the LFSR are filled as n inputs to an n variable Boolean function. Moreover, the Boolean function and memory collectively construct an FSM [7].

Proposed Design

Our proposed model follows a simple implementation of nonlinear combination generators shown in Fig. 61.2. The model comprises cryptographically secure Boolean function and number of LFSR can be added in a customized fashion. Here, for simplicity purpose, we use less number of the LFSR.

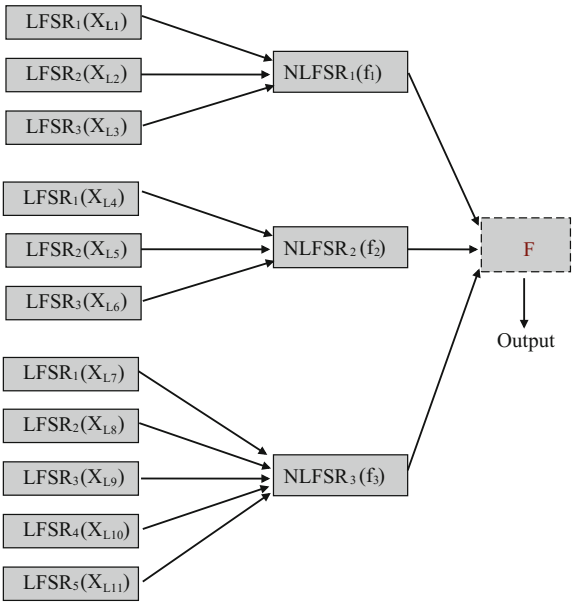


Fig. 61.2 Proposed design structure

Design Specifications

This design consists of three main blocks. In the first block, three LFSRs are initially loaded and passed through $NLFSR_1(f_1)$ function. The second block is also loaded with three LFSR and passed through $NLFSR_2(f_2)$ function and in the third block, five LFSRs are loaded and passed through $NLFSR_3(f_3)$ function. In LFSR, the feedback polynomial values are listed in the next subsection. Finally, three blocks are passed through one nonlinear function (F).

Initialization

Before the output sequence generation, the structure must be initialized with nonzero seed values. Usually, LFSR connection polynomial over *GF(2)* is the primitive polynomial or it can be called as the update function. Now, LFSR filled with a sequence of bits or it can be loaded like a fixed sized bit of hex values, or a string. Table 61.2 shows the LFSR `tap` polynomials. Specifically, results of the structure, i.e., binary sequences of the functions, are listed in the matrix order.[1]

[1]LF is represented as LFSR.

Table 61.2 Parameters of the proposed model

Tab Polynomials / Feedback Path	$NLFSR_1$	Output function:
$LFSR_1 = x_6 + x_5 + 1$ $LFSR_2 = x_5 + x_3 + 1$ $LFSR_3 = x_4 + x_3 + 1$	Combining function: $f_1(x_1, x_2, x_3) = x_2 + x_1x_3 + x_1x_2$	$F(x_1, x_2, x_3) = x_3 + x_1x_2 + x_3x_2$
Tab Polynomials/ Feedback Path $LFSR_4 = x_7 + x_6 + 1$ $LFSR_5 = x_5 + x_3 + 1$ $LFSR_6 = x_{11} + x_9 + 1$	$NLFSR_2$ Combining function: $f_2(x_1, x_2, x_3) = x_3x_2 + x_3x_1 + x_1$	
Tab Polynomials/ Feedback Path $LFSR_7 = x_5 + x_3 + 1$ $LFSR_8 = x_7 + x_6 + 1$ $LFSR_9 = x_8 + x_6 + x_5 + x_4 + 1$ $LFSR_{10} = x_9 + x_5 + 1$ $LFSR_{11} = x_{10} + x_7 + 1$	$NLFSR_3$ Combining function: $f_3(x_1, x_2, x_3, x_4, x_5) = x_1x_2 + x_2x_3 + x_3x_4 + x_4x_5 + x_5$	

$$
\begin{pmatrix}
LF_1 & 1\,0\,0\,1\,0\,0\,1\,0\,1\,1\,1\,0\,1\,1\,1\,0 \\
LF_2 & 1\,0\,1\,0\,1\,0\,0\,0\,0\,1\,0\,0\,1\,0\,1 \\
LF_3 & 0\,0\,1\,1\,0\,1\,0\,1\,1\,1\,1\,1\,0\,0\,0\,1 \\
LF_4 & 1\,0\,0\,1\,0\,0\,1\,1\,0\,1\,1\,0\,1\,0\,1 \\
LF_5 & 0\,1\,0\,1\,0\,0\,0\,0\,1\,0\,0\,1\,0\,1\,1 \\
LF_6 & 1\,0\,0\,0\,1\,0\,0\,1\,1\,0\,1\,1\,0\,1\,0 \\
LF_7 & 1\,1\,0\,0\,0\,1\,1\,0\,1\,1\,1\,0\,1\,0\,1 \\
LF_8 & 0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,1 \\
LF_9 & 1\,1\,1\,1\,0\,0\,0\,0\,1\,0\,1\,1\,1\,1\,0 \\
LF_{10} & 1\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\
LF_{11} & 1\,1\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1 \\
f_1 & 0\,0\,1\,1\,1\,0\,0\,0\,1\,1\,0\,0\,0\,0\,1 \\
f_2 & 0\,0\,0\,1\,0\,0\,1\,0\,1\,1\,0\,1\,1\,1\,1 \\
f_3 & 1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,1\,0 \\
F & 1\,1\,0\,1\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,1
\end{pmatrix}
$$

Results and Performance Analysis

We have done all our experiments in SageMath tool. Here, the final bits obtained from the nonlinear filter, i.e., F, will be considered as the final output bit which is used as a nonlinear sequence. We have taken only 15 bits of output; however, the number of bits can be increased or decreased as per requirement and also can be further converted to fixed bit hex values or string. In this paper, we have analyzed the nonlinearity property of the proposed schemes which are numerically depicted. Table 61.3 shows the typical values of parameters such as balancedness, nonlinearity,

Table 61.3 Cryptographic properties obtained after using the Boolean functions

Function	Balancedness	Nonlinearity	Maximum nonlinearity	Algebraic immunity	Correlation immunity	Walsh transform
f_1	YES	2	2.59	2	0	(0, 4, 4, 0, 0, 4, −4, 0)
f_2	YES	2	2.59	2	0	(0, 0, −4, 4, −4, −4, 0, 0)
f_3	YES	12	13.18	2	0	(0, 8, 0, −8, 8, 0, 8, 0, 0, −8, 0, 8, −8, 0, −8, 0, 8, 0, 8, 0, 0, 8, 0, −8, 8, 0, 8, 0, 0, 8, 0, −8)
F	YES	2	2.59	2	0	(0, 4, 0, −4, 4, 0, 4, 0)

maximum nonlinearity, algebraic immunity correlation immunity, and Walsh transform. More specifically, proposed model resultant bits are shown in Table 61.3. It shows the maximum possible nonlinearity and proposed design nonlinearity.

Conclusion and Future Work

In this paper, we have surveyed LFSR-/NLFSR-based stream ciphers. We have also implemented one nonlinear-based generator model to generate cryptographically secured bitstream. The various properties of randomness like algebraic immunity, correlation immunity, Walsh transformation, nonlinearity, etc. are listed in the tabulated form. The nonlinearity of the bitstream is compared with maximum nonlinearity achievable for a particular Boolean function. Research problems on NLFSR are still not well understood like patterns and its behaviors. Development of an intelligent algorithm for designing customized LFSR-based stream cipher using the generalized model shall be our future research work.

References

1. eSTREAM: the ECRYPT Stream Cipher Project. http://www.ecrypt.eu.org/stream/
2. Kocheta, M., Sujatha, N., Sivakanya, K., Srikanth, R., Shetty, S., Mohan, P.A.: A review of some recent stream ciphers. In: 2013 International conference on Circuits, Controls and Communications (CCUBE). (2013)
3. Eisenbarth, T., Kumar, S., Paar, C., Poschmann, A., Uhsadel, L.: A survey of lightweight-cryptography implementations. IEEE Design & Test of Computers **24**(6) (2007) 522–533
4. Videau, M. In: eSTREAM. Springer US, Boston, MA (2011) 426–427
5. Shehhi, M.A.A., Baek, J., Yeun, C.Y.: The use of boolean functions in stream ciphers. In: Internet Technology and Secured Transactions (ICITST), 2011 International Conference for. (Dec 2011) 29–33
6. Maitra, S., Sarkar, P.: Cryptographically significant boolean functions with five valued walsh spectra. Theoretical Computer Science **276**(1) (2002) 133–146
7. Maitra, S.: Autocorrelation properties of correlation immune boolean functions. In: International Conference on Cryptology in India, Springer (2001) 242–253
8. Nawaz, Y., Gong, G., Gupta, K.C. In: Upper Bounds on Algebraic Immunity of Boolean Power Functions. Springer Berlin Heidelberg, Berlin, Heidelberg (2006) 375–389
9. Deb, S., Biswas, B., Kar, N. In: Study of NLFSR and Reasonable Security Improvement on Trivium Cipher. Springer India, New Delhi (2015) 731–739
10. Klapper, A., Goresky, M. In: Cryptanalysis Based on 2-Adic Rational Approximation. Springer Berlin Heidelberg, Berlin, Heidelberg (1995) 262–273
11. Klapper, A., Goresky, M.: Feedback shift registers, 2-adic span, and combiners with memory. Journal of Cryptology **10**(2) (1997) 111–147
12. El Hennawy, H.M., Omar, A.E., Kholaif, S.M.: Lea: link encryption algorithm proposed stream cipher algorithm. Ain Shams Engineering Journal **6**(1) (2015) 57–65
13. Khan, M.A., Khan, A.A., Mirza, F.: Transform domain analysis of sequences. CoRR arXiv:1503.00943 (2015)

Chapter 62
A Modified RSA Cryptography Algorithm for Security Enhancement in Vehicular Ad Hoc Networks

Debasish Roy and Prodipto Das

Abstract Secure the communication over network is an important task. Security of Vehicular Ad hoc Networks (VANET) is prime concern due to the numbers of attacker's existence. We can protect information by using different cryptography methods. The Rivest–Shamir–Adleman (RSA) is one of the asymmetric key cryptography systems to protect information. In asymmetric key cryptography system from the generated key pair, one key is used for information to encrypt and the another key is used for information to decrypt. In RSA, the value of n is computed through the multiplication of two prime numbers p and q. If the factors of n can be found by the intruder using brute force attack, therefore, the security level will reduce. In this paper, we proposed a modified RSA algorithm called MRSA with three prime numbers j, k, and l for VANETs to increase the level of security in vehicular communication. RSA algorithm is with one more prime number, i.e., three prime numbers are used rather than two prime numbers, for increasing the brute force attack time. The lifetime of a message in VANETs may be less. Therefore, MRSA may provide us better result by enhancing the level of security with small key size. The experimental results and analyses are shown for both the RSA and MRSA cryptography algorithms.

Keywords VANETs · RSA · MRSA · Cryptography · RSUs
SKC · PKC

D. Roy (✉) · P. Das
Department of Computer Science, Assam University, Silchar 788011, India
e-mail: debasish.aucs@gmail.com

P. Das
e-mail: prodiptodas@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_62

Introduction

In VANETs [1], wireless communication between vehicles provides road safety and efficient transportation by spreading the road related information such as traffic condition, road condition, accident, etc. VANET is highly dynamic in nature due to its mobility and fast changing of network topology. In VANETs, each vehicle is allowed to transmit traffic-related information to its neighbors. Therefore, an existence of attackers may lead to security threats. Security threats may be avoided by giving the better security solution. One of the security solutions is cryptography [2]. Cryptography is the art of science to write the information in secret code. Therefore, to whom it is made for can read and process the information; no one else can read or process. Cryptography provides secure communication over the network. Cryptography has came out with the objectives of ensuring integrity, privacy or confidentiality, authentication, key exchange, and non-repudiation. There are basically three types of cryptography: secret key cryptography (SKC) [2], public key cryptography (PKC) [2], and hash functions [2]. In SKC for both encryption and decryption processess, a single key is used. In PKC, key pair is generated and from this key pair, one key is used to perform encryption and the other is used to perform decryption. Hash function is the one-way encryption. It does not use any key; instead, it uses mathematical transformation to encrypt the data. Encryption is the process of transforming plain text to the ciphertext and decryption is the process of converting the ciphertext to the original plain text. RSA [3] is one of the PKC techniques and we modified the RSA to enhance the security in vehicular communication over the network. The attackers in RSA may attack using brute force mechanism in which the attacker tries all the possible combination to identify the private key value. Therefore, we modified the RSA cryptography with three prime numbers to enhance the level of security. We have used the prime numbers, length of at least half of the key size. In VANET, MRSA with large key value may be used during the registration process of newly arrived vehicle into the network because here the encryption and decryption times are negligible with respect to the security

Fig. 62.1 Proposed registration process layout

level of communication. The MRSA algorithm consists of three steps mainly: Key generation, encryption, and decryption. Figure 62.1 shows the registration process of newly arrived vehicle into the network. The roadside units (RSUs) [4] are deployed near the side of the road and V represents the vehicle.

The remaining part of the paper is organized as follows: In section "Literature Review", existing works of VANET communication security framework are reviewed. Section "Algorithm" discusses about the RSA and proposed MRSA algorithms. Section "Experimental Results, Analysis and Discussion" discusses about the experimental results and analysis. Lastly, conclusion and future work are given in section "Conclusion and Future Work".

Literature Review

Security of VANET is a crucial issue due to the fact that all the transmission of information is done on the network environment which is open to all. The network is open to each vehicle to transmit information to its neighbors. Therefore, to solve the security issue in VANET is essential. There are significant and large numbers of research work that has been done regarding the security in VANET. In this section, we briefly reviewed some of them.

Nema et al. [5] proposed a mechanism to secure the intelligent traffic system of VANET by using RSA algorithm. Trust values are assigned to each vehicle to identify and remove the malicious nodes from the network. The RSA algorithm is used during the communication between the RSU and vehicle. The authors are mainly focused on maintaining the message integrity, providing confidentiality, identify, and remove the malicious as well as misbehaving nodes. Ad hoc On-demand Distance Vector routing protocol (AODV) is used. They have implemented the boundary for vehicle communication with RSU and done simulation using MATLAB to show the result. This scheme improves the system efficiency.

Barani et al. [6] proposed a mechanism to implement application for transforming safety-related information in VANET. They introduce self-organized VANET development. Each vehicle registration has done using RSA cryptography algorithm and the verification has done by using Zero-Knowledge Proof (ZKP) algorithm and also identify and remove the attackers from the network. NS2 is used for simulation. The public and private keys are generated through MATLAB. In future, plan to establish more secured encryption algorithm.

Choi et al. [7] proposed a security scheme with privacy and strong non-repudiation. They used ID-based cryptosystem. To verify the vehicle-trusted ID, third-party ID is used and developed RSA public key is used in terms of peers ID. This scheme is efficient than that of signature and verification.

Leu et al. [8] proposed a mechanism Ambulance Traffic Control System (ATCS) for ambulance (AMU) to reach the hospital as early as possible. The mechanism is to turn on the green light when an ambulance is near the intersection of the road. One ambulance can communicate with the other, and all the transferred information

are encrypted using RSA cryptography algorithm. When an accident occurs, the Roadside Transportation Authority (RTA) will find out the nearest and suitable ambulance, calculate the shortest path for the selected ambulance, and control all the traffic lights along this shortest path so that ambulance may reach the destination as fast as possible. In future, plan to establish formal behavior, reliability model, and also to give the control of the traffic lights to ambulance.

Sarkar et al. [9] proposed a mechanism, RSA threshold cryptography by using verifiable secret sharing (VSS) which is based on Chinese remainder theorem for MANET. Asmuth bloom secret sharing scheme is used. Assumptions taken are unique node identification number for each node in the network and a broadcast channel where the broadcast of data must be same for all nodes. The routes are established using AODV routing protocol. In future, plan to develop in test beds.

Isaac et al. [10] proposed a payment protocol in VANET where no direct communication is possible between user and issuer for authentication. Nontraditional digital signature scheme is used in this protocol and satisfies credit as well as debit card transactions; it also maintains the privacy. In user registration part, RSA cryptography technique is used. In future, plan to increase the functionalities of the proposed protocol.

Vijay et al. [11] proposed an intrusion detection system for detecting the misbehaving and malicious nodes. An Enhanced Adaptive ACKnowledgment (EAACK) system is developed. It has mainly three parts: end-to-end acknowledgement (ACK) scheme, Secure ACK (S-ACK) scheme, and Misbehavior Report Authentication (MRA) scheme. AODV is used as a routing protocol. An EAACK scheme consists of the mixture of both RSA and AES algorithms. In future, plan to establish a key exchange mechanism and test in real environment.

Yang et al. [12] proposed RSA with threshold-based multi-signature mechanism for Mobile Ad hoc Networks (MANET). A mechanism for maintaining the sequence is described for multi-signature. The proposed mechanism shows efficient result in computational and spatial complexities. In future, plan to establish a scheme for group partial signature validity.

Jagdale et al. [13] proposed a protocol using cryptography technique to protect the data packets and control packets. Group signature and ID-based encryption technique are used. Two different stages of the protocol are route discovery and key generation. The protocol is implemented using three encryption algorithms: RSA, Advanced Encryption Standard (AES), and Data Encryption Standard (DES) with AODV protocol, and the performance is compared with AODV protocol. The simulation is performed in NS2 simulator. In future, plan to scale down encryption time.

Caballero-Gil et al. [14] proposed an approach for secure authentication in VANET. In this algorithm, each vehicle selects the public key. Authentication of new vehicle is done using Zero-Knowledge Proof (ZKP). The simulation is performed using NS2 network simulator tool. Each vehicle is characterized by using unique ID, fixed public or private key pair of RSA algorithm, and key store.

Aswathy et al. [15] proposed a modified RSA cryptography algorithm. Field-Programmable Gate Array (FPGA) development of modular exponentiation

depends on two schemes: one is maximum sequential operation and another is parallel. It has two stages: first, mapping of two operations in parallel with two multipliers and second, in case of modular multiplication, the square and multiplication operations are done in parallel. The analysis of results is done based on area, speed, and power. The Xilinx ISE software is used for experiment.

Chhabra et al. [16] proposed a modified method for RSA cryptography algorithm for secure communication. They proposed a scheme where transformation of n is not the multiplication of two large prime numbers.

Moghaddam et al. [17] proposed a user authentication and encryption scheme to secure access in cloud servers based on clients. Trust is also ensured or identified to give access. Encryption is done by using the modification of Diffie–Hellman and RSA algorithm. Client-based system validates the user ID and controls the access. The algorithm is analyzed based on key value, time correctness, and security.

Mainanwal et al. [18] proposed an algorithm for authentication in web login process using zero-knowledge and RSA cryptography (Z-RSA) method. In the client part, Z-RSA is used and performed its task between client and server. Password hashes are not needed to save on the server. They have shown the registration and authentication process. It reduces the computation and cost of computation.

Wang et al. [19] proposed a mechanism to the personal information using RSA cryptography algorithm. Personal information is transferred original text to encrypted form. Privacy is maintained in communication between customer representative and clients.

Sone [20] proposed a mechanism to increase the security level and performance in wireless networks using efficient key management process. The author aims to decrease the execution time, increase throughput, etc. RSA cryptography, convolutional codes, and sub-band code are used to give security in wireless networks. Execution time is reduced by using the small key size value. In future, plan to embed the code in Field-Programmable Gate Array (FPGA) device.

Sahana et al. [21] proposed security protocol in wireless sensor networks (WSN) using RSA cryptography. They propose a cluster-based routing protocol. RSA cryptography is used during the communication between the cluster node and cluster head. Base station is broadcasting its own public key. Cluster head is also broadcasting the public key of its own to its cluster node. Cluster head is responsible for generating the public and private key pair.

Arazi et al. [22] proposed an RSA-based model to mitigate the Denial of Service attack (DOS) in Wireless Sensor Networks (WSN). Three mechanisms are used to generate the key. Elliptic Curve Digital Signature Algorithm (ECDSA) and self-certified DH fixed key generation are used for validation purpose,

Sulochana et al. [23] proposed a mechanism for secure data communication in the presence of false data injection attack. RSA is used to generate key pair. Keys are assigned to each node in the network. Each node is having unique network ID which is used during the communication with destination. They proposed a Polynomial-based Compromise Resilient En-route Filtering (PCREF) scheme to identify fake data. PFDI Fuzzy algorithm is used to prevent the fake information.

Zhao et al. [24] proposed an approach to secure communication using RSA cryptography in wireless sensor networks. They have performed the encryption of digital image and shown the result. They have shown two types of image file encryption: one is gray image encryption and another is color image encryption.

Algorithm

The RSA and MRSA cryptography algorithms are described as follows:

RSA Algorithm

The steps of RSA algorithm are as follows [3]:

Step 1: Generate two random prime numbers p and q of same size, where $p \neq q$.
Step 2: Calculate n, which is the multiplication of two generated prime numbers p and q.

$$n = p \times q$$

Step 3: Evaluate $\varphi = (p - 1) \times (q - 1)$.
Step 4: Choose an integer e, so that gcd $(e, \varphi) = 1$, where $1 < e < \varphi$.
Step 5: Evaluate d such that $d \equiv e^{-1} \pmod{\varphi}$ or $e \times d \equiv 1 \pmod{\varphi}$, where $1 < d < \varphi$.
Step 6: (e, n) represents public key and (d, n) represents private key. Keep the values of d, p, q, and φ as secret.
Step 7: The encryption is done by $C = M^e \bmod n$.
Step 8: The decryption of ciphered text is done by $M = C^d \bmod n$.

Proposed MRSA Algorithm

The steps of proposed MRSA algorithm are as follows:

Step 1: Generate three random prime numbers of same size j, k, and l where $j \neq k \neq l$.
Step 2: Compute the value of z that is the product of three prime numbers j, k, and l.

$$z = j \times k \times l$$

Step 3: Compute the value of ϕ,

$$\phi = (j-1) \times (k-1) \times (l-1)$$

Step 4: Choose an integer u, such that gcd (u, ϕ) = 1, where $1 < u < \phi$.
Step 5: Multiply u with the third prime number which is l.

$$w = u \times l$$

Step 6: Compute the value of v so that $v \equiv w^{-1} \pmod{\phi}$ or $w \times v \equiv 1 \pmod{\phi}$, where $1 < v < \phi$.

Step 7: The public key is (w, z) and the private key is (v, z). Keep the values of v, j, k, l, and ϕ as secret.
Step 8: The encryption of message is done by $E(M) = M^w \bmod z$.
Step 9: The decryption of encrypted message is done by $M = E(M)^v \bmod z$.

In Step 5, we multiply the value of u with third prime number l because if attackers find out the public key pair (w, z), still it will be difficult for them to find the value of u.

Experimental Results, Analysis, and Discussion

Both RSA and MRSA cryptography algorithms are developed using Java.

Figures 62.2, 62.3, 62.4, 62.5, 62.6, 62.7, 62.8, 62.9, 62.10, 62.11, 62.12, and 62.13 show the comparison between RSA and MRSA cryptography algorithms. The time for encryption and decryption increases in both the algorithms with the increase in key size and chunk size of data. We are getting better security with the increase in key size in both the algorithms but at the same time speed also decreases. Therefore, by increasing one more prime number in MRSA cryptography algorithm, we are getting better security results with small value of key size in comparison with RSA cryptography algorithm. MRSA takes more time to generate the key in comparison with RSA due to the additional multiplication of one more prime number. In the comparison of MRSA with key size 512, encryption time is 361 ms and decryption time is 734 ms with 1024 chunk size of data, whereas the RSA with key size 1024, encryption time is 554 ms and decryption time is 993 ms with 1024 chunk size of data; and the MRSA with key size 1024, encryption time is

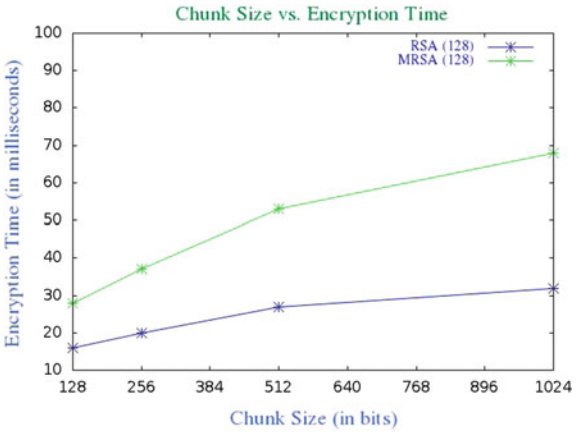


Fig. 62.2 Data size versus encryption time with 128-bit key

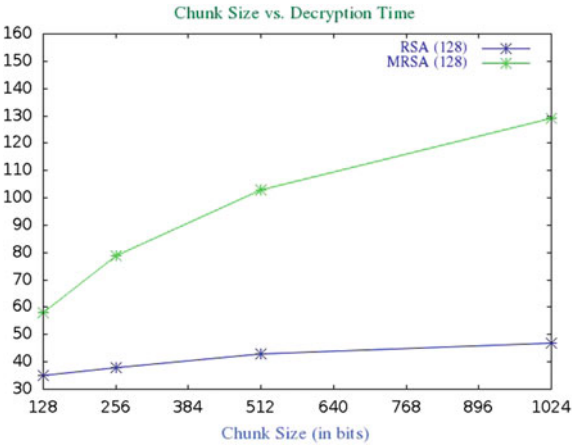


Fig. 62.3 Data size versus decryption time with 128-bit key

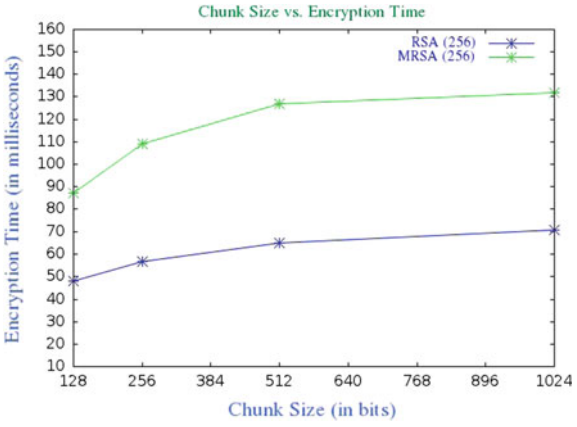


Fig. 62.4 Data size versus encryption time with 256-bit key

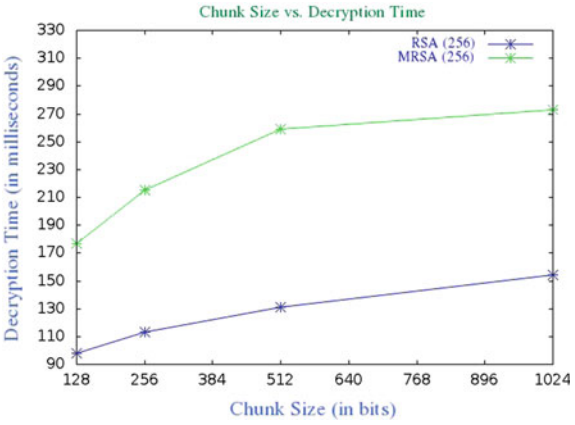


Fig. 62.5 Data size versus decryption time with 256-bit key

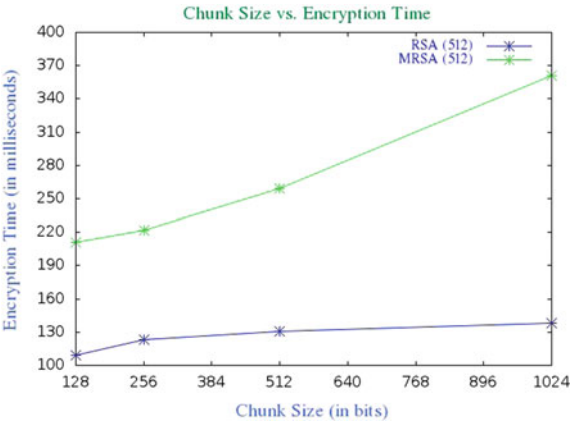


Fig. 62.6 Data size versus encryption time with 512-bit key

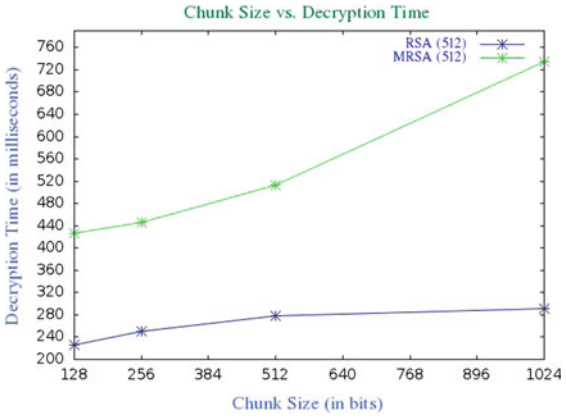


Fig. 62.7 Data size versus decryption time with 512-bit key

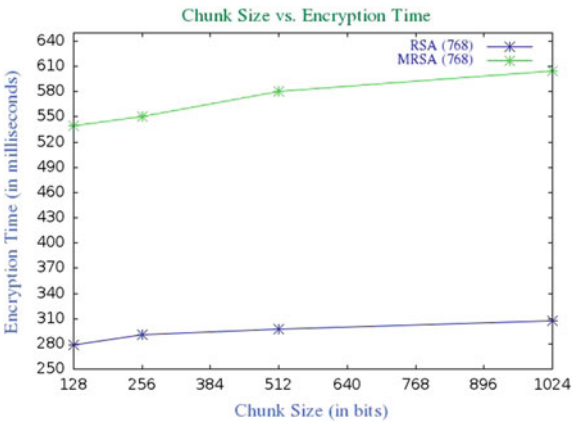


Fig. 62.8 Data size versus encryption time with 768-bit key

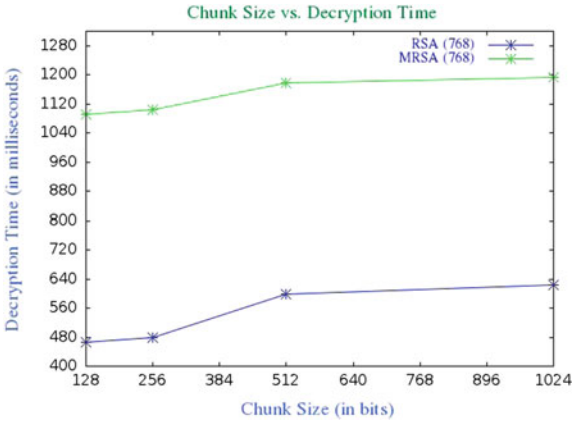


Fig. 62.9 Data size versus decryption time with 768-bit key

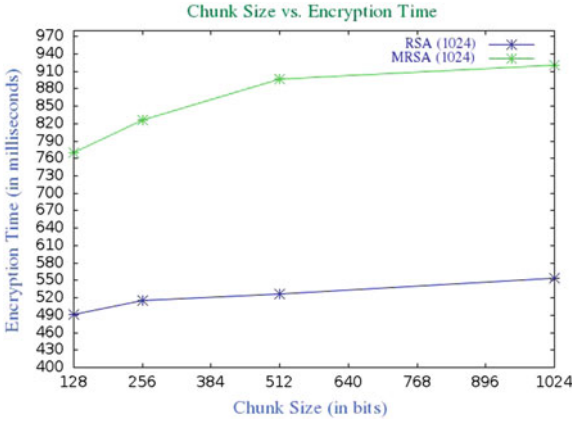


Fig. 62.10 Data size versus encryption time with 1024-bit key

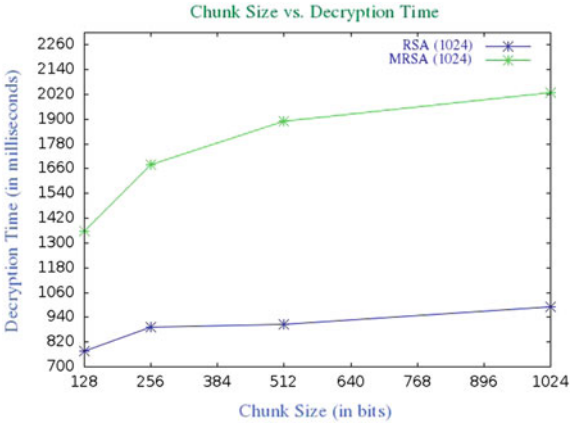


Fig. 62.11 Data size versus decryption time with 1024-bit key

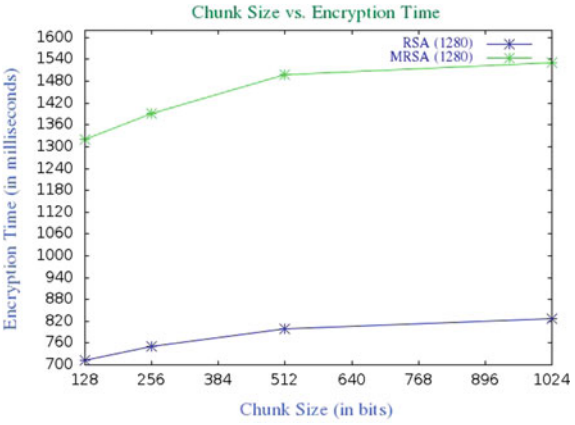


Fig. 62.12 Data size versus encryption time with 1280-bit key

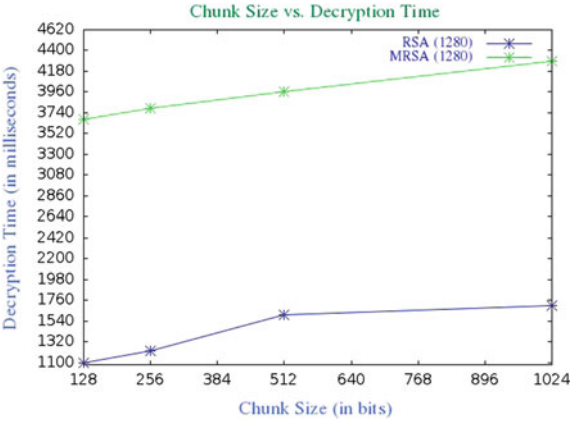


Fig. 62.13 Data size versus decryption time with 1280-bit key

920 ms and decryption time is 2028 ms with 1024 chunk size of data, whereas RSA with key size 1280, encryption time is 827 ms and decryption time is 1708 ms with 1024 chunk size of data. Therefore, MRSA is performing better than that of RSA from the viewpoint of encryption and decryption as well as MRSA enhancing the security.

Conclusion and Future Work

The paper presents a modified RSA cryptography algorithm called MRSA to enhance the security. The security is enhanced by increasing the number of factors of n, three prime numbers rather than two prime numbers. We are maximizing the time taken for brute force attack by increasing the numbers of factors of n in MRSA cryptography algorithm. Time complexity increases with the increase in factors of n but from the viewpoint of security it is negligible. Therefore, MRSA is more secured compared to RSA. In future, have a plan to reduce key generation time and also encryption, decryption time for large key value.

References

1. Shao, C., Leng, S., Zhang, Y., Vinel, A., Jonsson, M.: Performance Analysis of Connectivity Probability and Connectivity-Aware MAC Protocol Design for Platoon-Based VANETs. IEEE Transactions on Vehicular Technology, vol. 64, no. 12, pp. 5596–5609 (2015)
2. Stallings, W.: Cryptography and Network Security Principles and Practice. Pearson Education, Prentice Hall Press Upper Saddle River, vol. 5, pp. 1–719 (2010)
3. Rivest, R.L., Shamir, A., Adleman, L.: A Method for Obtaining Digital Signatures and Public Key Cryptosystems. Communications of the ACM, vol. 21, no. 2, pp. 120–126 (1978)
4. Song, J., He, C., Zhang, L., Tang, S., Zhang, H.: Toward an RSU-unavailable lightweight certificateless key agreement scheme for VANETs. IEEE China Communications, vol. 11, no. 9, pp. 93–103 (2014)
5. Nema, M., Stalin, S., Tiwari, R.: RSA algorithm based encryption on secure intelligent traffic system for VANET using Wi-Fi IEEE 802.11p. IEEE International Conference on Computer, Communication and Control (IC4), Indore, pp. 1–5 (2015)
6. Barani, P.S., Elizabeth, N.E.: Registration and verification of vehicles in VANET's. IEEE International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, pp. 1087–1092 (2015)
7. Choi, J., Jung, S.: A Security Framework with Strong Non-Repudiation and Privacy in VANETs. IEEE 6th Consumer Communications and Networking Conference, Las Vegas, NV, pp. 1–5 (2009)
8. Leu, F.-Y., Chen, M.-H., Huang, Y.-L., Lin, C.-C.: Controlling Traffic Lights for Ambulances. IEEE 7th International Conference on Broadband, Wireless Computing, Communication and Applications (BWCCA), Victoria, BC, pp. 462–467 (2012)
9. Sarkar, S., Kisku, B., Misra, S., Obaidat, M.S.: Chinese Remainder Theorem-Based RSA-Threshold Cryptography in MANET Using Verifiable Secret Sharing Scheme. IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, Marrakech, pp. 258–262 (2009)

10. Isaac, J.T., Camara, J.S., Zeadally, S., Marquez, J.T.: A secure vehicle-to-roadside communication payment protocol in vehicular ad-hoc networks. ELSEVIER computer communications, vol. 31, no. 10, pp. 2478–2484 (2008)
11. Vijay, M., Sujatha, R.: Intrusion detection system to detect malicious misbehaviour nodes in MANET. IEEE International Conference on Information Communication and Embedded Systems (ICICES), Chennai, pp. 1–6 (2014)
12. Yang, M., Hong, F., Li, J., Su, L.: Secure Threshold Order-Specified Multi-signature Scheme in MANET. IEEE International Conference on Computational Intelligence and Security, Guangzhou, vol. 2, pp. 1279–1282 (2006)
13. Jagdale, B.N., Patil, M.S.: Emulating cryptographic operations for secure routing in Ad-hoc Network. IEEE International Conference on Pervasive Computing (ICPC), Pune, pp. 1–4 (2015)
14. Caballero-Gil, C., Caballero-Gil, P., Molina-Gil, J.: Mutual authentication in self-organized VANETs. ELSEVIER Computer Standards & Interfaces, vol. 36, no. 4, pp. 704–710 (2014)
15. Aswathy, B.G., Resmi, R.: Modified RSA public key algorithm. IEEE 1st International Conference on Computational Systems and Communications (ICCSC), Trivandrum, pp. 252–255 (2014)
16. Chhabra, A., Mathur, S.: Modified RSA Algorithm: A Secure Approach. IEEE International Conference on Computational Intelligence and Communication Networks (CICN), Gwalior, pp. 545–548 (2011)
17. Moghaddam, F.F., Varnosfaderani, S.D., Ghavam, I., Mobedi, S.: A client-based user authentication and encryption algorithm for secure accessing to cloud servers based on modified Diffie-Hellman and RSA small-e. IEEE Student Conference on Research and Development (SCOReD), Putrajaya, pp. 175–180 (2013)
18. Mainanwal, V., Gupta, M., Upadhayay, S.K.: Zero Knowledge Protocol with RSA Cryptography Algorithm for Authentication in Web Browser Login System (Z-RSA). IEEE 5th International Conference on Communication Systems and Network Technologies (CSNT), Gwalior, pp. 776–780 (2015)
19. Wang, L., Zhang, Y.: A new personal information protection approach based on RSA cryptography. IEEE International Symposium on IT in Medicine and Education (ITME), Cuangzhou, vol. 1, pp. 591–593 (2011)
20. Sone, M.E.: Efficient key management scheme to enhance security-throughput trade-off performance in wireless networks. IEEE Science and Information Conference (SAI), London, pp. 1249–1256 (2015)
21. Sahana, A., Misra, I.S.: Implementation of RSA security protocol for sensor network security: Design and network lifetime Analysis. IEEE 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), Chennai, pp. 1–5 (2011)
22. Arazi, O., Qi, H., Rose, D.: A Public Key Cryptographic Method for Denial of Service Mitigation in Wireless Sensor Networks. IEEE 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, San Diego, CA, pp. 51–59 (2007)
23. Sulochana, S., Manjula, V.: Resilient system for secure sharing of information against false data injection attack. IEEE International Conference on Information Communication and Embedded Systems (ICICES), Chennai, Tamil Nadu, India, pp. 1–5 (2016)
24. Zhao, G., Yang, X., Zhou, B., Wei, W.: RSA-based digital image encryption algorithm in wireless sensor networks. IEEE 2nd International Conference on Signal Processing Systems (ICSPS), Dalian, vol. 2, pp. V2-640 –V2-643 (2010)

Chapter 63
HLR_DDoS: A Low-Rate and High-Rate DDoS Attack Detection Method Using α-Divergence

Nazrul Hoque and Dhruba K. Bhattacharyya

Abstract In this paper, an effective method called HLR_DDoS is proposed to detect both low- and high-rate flooding attacks using a statistical approach. The method detects both types of attacks in two steps: (i) normal traffic analysis using cross-correlation measure and (ii) identification of suspicious high- and low-rate attack traffic using α-divergence. The proposed method is evaluated on DDoS CAIDA 2007 and DARPA 2000 datasets.

Keywords Flooding attacks · Anomaly detection · Correlation · Accuracy

Introduction

DDoS attack also known as coordinated attack is performed on system(s) with the help of many compromised machines called zombies, and it prevents legitimate users to access the services provided by the system(s). The attack is catastrophic as it can crash a system in a short attacking period due to its cooperative and large-scale nature. The main purpose of DDoS attack is twofolds: resource exhaust and bandwidth consumption [1, 2]. According to Mirkovic et al. [3], a DDoS attack can be classified as continuous and variable rate attacks based on attack rate dynamics. Moreover, a variable rate attack can be either low- or high-rate attack. Xiang et al. [4] defined low- and high-rate attacks in terms of number of packets sent to the victim. Low-intensity flooding attack is very similar to normal traffic. Therefore, low-intensity attack can easily bypass the security mechanism of an anomaly-based defense system. Today, hackers create a large-scale low-rate DDoS attack combining various low-intensity attacks that are spread in a distributed manner. As a

N. Hoque (✉) · D. K. Bhattacharyya
Department of Computer Science and Engineering, Tezpur University, Sonitpur 784028, Assam, India
e-mail: tonazrul@gmail.com

D. K. Bhattacharyya
e-mail: dkb@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_63

consequence, traditional anomaly detection mechanism cannot detect low-rate DDoS attack. Hence, many researchers as well as security experts use statistical mechanism to overcome the problem because low-intensity DDoS attack can be identified from the statistical behavior of network traffic parameters, such as entropy of source IPs or destination IPs, packet arrival rate, etc.

Motivation: Anomaly-based DDoS detection mechanisms consider normal network traffic to identify anomalous traffic based on the divergence between normal and attack traffic patterns. In both the signature- and anomaly-based detection systems, network traffic is analyzed using different mechanisms by considering multiple features of network traffic. However, an intelligent attacker might generate attack traffic in such a way that a defense mechanism cannot identify the divergence between attack and normal network traffic. Moreover, performance of an anomaly-based DDoS detection mechanism depends on certain parameters or thresholds and it is very difficult to predict the correct range of values for those parameters a priori. Although many statistical methods have been developed to identify DDoS attacks as early as possible with low false alarm rate, there is no any method that can identify both low- and high-intensity traffic using only one threshold value. We propose a method using α-divergence and cross-correlation for attack detection using one threshold value and less number of network features.

Contributions:

1. We propose a DDoS attack detection mechanism called HLR_DDoS to protect a victim from DDoS attacks. In the first phase, the method identifies anomalous patterns coming to a system. In the next phase, it identifies both the low- and high-rate attack traffic.
2. HLR_DDoS considers only one threshold parameter called β to detect DDoS attacks.
3. HLR_DDoS identifies DDoS attack traffic using α-divergence.

Paper Organization: Rest of the paper is organized as follows. In section "Related Work", related work on various low- and high-rate DDoS attack detection methods is discussed. The proposed DDoS attack detection method is explained in section "Proposed Method". Results are analyzed and compared in section "Experimental Results". Conclusion and future work are discussed in section "Conclusion and Future Work".

Related Work

In literature, we found many DDoS attack detection methods for high-rate attacks [5–7]. Most DDoS attack detection methods use statistical and machine learning approach to identify high-rate DDoS flooding attacks. A few statistical methods such as entropy and chi-square are discussed by Feinstein et al. [8]. Yu et al. [9] proposed

Table 63.1 Comparison among detection methods

Author	Year	Method	Parameter used	LDoS/ LDDOS
Yang et al. [11]	2004	Randomized algorithm	minRTO, RTT	LDoS
Chen et al. [12]	2006	Signal analysis (DSP) using DFT	Packet arrival rate	LDDoS
Zhang et al. [13]	2011	Robust RED	RTO, burst rate	LDoS
Xiang et al. [4]	2011	Information metric	IP address and protocol	LDDoS
Zhang et al. [14]	2012	Congestion participation rate	Probability of packet drop	LDDoS
Zhi-jun et al. [15]	2013	Chaos-theory		LDoS
Zhijunt et al. [16]	2014	Cross-correlation	Attack periods, pulse peak	LDDoS

a collaborative method using entropy rate to identify DDoS attack traffic from legitimate traffic. The method identifies attack at an early stage calculating entropy of flows on routers and if the flow entropy is less than a particular threshold, then an alarm is generated. A TCP SYN flooding attack detection method is developed by Xiao et al. [10]. The method uses an active probing scheme to capture attack signature for detecting DDoS attack in an early stage.

Due to the similarity between normal traffic and low-intensity attack traffic, detection of low-rate attack is very difficult. Many researchers use behavioral analysis of network traffic patterns for low-rate attack detection. In literature, we found a few methods to detect LDDoS attacks [6]. A comparison among the detection methods is shown in Table 63.1.

Proposed Method

The proposed method consists of two phases, namely analysis of normal network traffic and analysis of captured traffic. The captured traffic analysis phase identifies anomalous traffic in the first step and discriminates low- and high-rate DDoS attack traffic in the next step. The method uses cross-correlation and α-divergence for anomaly detection. However, the method discriminates low- and high-rate attacks using α-divergence only. The intuition behind using two measures is that in many situations identification of low-intensity traffic is difficult due to its similar nature with normal traffic. Hence, if one measure is incapable of identifying their difference then another measure can discriminate them. So, during captured traffic analysis, both the measures, i.e., cross-correlation and α-divergence, are used. The cross-correlation between a normal network traffic and a similar low-rate attack traffic may be very high and hence, the method cannot find any difference between them. On the other

hand, α-divergence is very effective to identify the difference between low- and high-rate attack traffic. The proposed method splits network traffic into 1-s time window samples and from each 1-s time window, it extracts four features, viz., distinct source IPs, packet rate, entropy, and variation of source IPs [6].

Cross-correlation for DDoS attack: Correlation between any two samples or vectors can be computed using cross-correlation. Cross-correlation defines a linear relationship between two variables or objects. Cross-correlation between two variables is +1 if they rise in an identical rate and if one of the two variables does not move then cross-correlation is 0; whereas if the other variable falls at an identical rate, then cross-correlation is −1. Using this measure, a normal profile is generated from the computed cross-correlation values during normal traffic analysis. First, normal network traffic is sampled into 1-s time window and from each time window distinct source IPs, packet rate, entropy, and variation of source IPs are extracted. Now for each two consecutive time windows, say X and Y, we compute cross-correlation using the four extracted features as follows:

$$r_d = \frac{\sum[(x[i] - \bar{x}).(y[i-d] - \bar{y})]}{\sqrt{(\sum[(x[i] - \bar{x})])}\sqrt{(\sum[(y[i-d] - \bar{y})])}}, \tag{63.1}$$

where d is the time lag, and $\bar{x}$ and $\bar{y}$ are the mean values of X and Y. Cross-correlation between any two normal network traffic samples or between any two attack samples at a different time lag d will be close. But the correlation value between a normal sample and an attack sample certainly have diverged value. So, the measure can effectively discriminate DDoS attack samples from normal.

Normal traffic analysis: During normal traffic analysis, the proposed method considers normal network traffic and extracts features from each 1-s normal traffic sample. The method computes cross-correlation and divergence between any two normal network traffic samples using the extracted normal traffic features. Normal profile stores four parameters, viz., mean of normal samples, mean cross-correlation distance, mean α-divergence, and mean distance between individual cross-correlation and mean cross-correlation value computed from normal traffic samples.

Algorithm 1 describes the steps of normal profile generation.

Captured traffic analysis: In the second phase of the proposed method, raw network traffic is captured and sampled the traffic into 1-s time window samples. Like normal traffic analysis, the method extracts four distinct features from each 1-s traffic sample and computes cross-correlation and α-divergence between a captured test sample, say T_i and the normal mean sample, i.e., meanN. If the cross-correlation value is greater than a particular threshold value or the α-divergence is greater than 1, then the test sample is marked as an attack. Algorithm 2 describes the steps of the detection phase.

The proposed method detects anomalous traffic sample in the first phase of detection. In the next phase of detection, it discriminates low- and high-rate attack traffic using α-divergence. The method considers the traffic samples that are detected as

Data: Samples of normal network traffic
Result: mean of normal samples (meanN), mean alphaDiv (meanDiv), mean cross-correlation (meanCross) mean distance (meanD)
for *each normal sample* S_i, $i = 1, 2, \cdots, n$ **do**
  compute
  $O_i = \{E_{sip}, V_{sip}, D_{sip}, P_{rate}\}$
end
meanN= mean of all normal samples O_i
for *each normal sample* O_i, $i = 1, 2, \cdots, n$ **do**
  Xcorr(i)=cross-correlation(O_i, O_{i+1})
  alphaDiv(i)=D($O_i||O_{i+1}$)
end
Compute meanDiv and meanCorr using the following equations, respectively
$meanDiv = \frac{\sum_{i=1}^{n} alphaDiv(i)}{n}$
$meanCorr = \frac{1}{n}\sum_{i=1}^{n} Xcorr(i)$
for *each normal sample* O_i, $i = 1, 2, \cdots, n$ **do**
  Compute Distance, Dist(i)=distance(Xcorr(i), meanCorr)
end
Compute mean distance meanD
$meanD = \frac{1}{n}\sum_{i=1}^{n} Dist(i)$
Store meanDiv, meanD, meanCorr and meanN as profile parameters

Algorithm 1: Normal profile Generation

Data: meanN, threshold β, samples of captured network traffic
Result: normal and attack samples
for *each captured sample* T_i, $i = 1, 2, \cdots, n$ **do**
  compute
  $T_i = \{E_{sip}, V_{sip}, D_{sip}, P_{rate}\}$
end
for *each captured sample* T_i, $i = 1, 2, \cdots, n$ **do**
  Xcorr(i)=cross-correlation(T_i, *meanN*)
  alphaDiv(i)=D($O_i||meanN$)
  Compute distance Dist(i)=distance(Xcorr(i), meanCorr)
end
if $((Dist(i) - meanD) > \beta || (|meanDiv - alphaDiv(i)|) > 1)$ **then**
  mark T_i as attack
else
  T_i is normal sample
end

Algorithm 2: Captured Traffic Analysis

anomalous in the first phase and compares the α-divergence value with maximum (maxDiv), mean (meanDiv), and minimum (minDiv) α-divergence values of normal network traffic samples. If an anomalous traffic sample T_i satisfies the following condition, then the sample is marked as high-rate DDoS sample:
if((maxDiv-alpha(T_i))>(meanDiv-alpha(T_i)) >(minDiv-alpha(T_i))).
maxDiv, meanDiv, and minDiv are computed from normal traffic and alpha(T_i) represents alphaDiv of a test sample T_i. The above condition ensures that α-divergence is less than minDiv value but in case of high-intensity DDoS attack, attack samples

Table 63.2 Low- and high-rate classification accuracies on different threshold values

Threshold values					
	0.01	0.5	0.9	1	2
Detection rate	0.99	0.99	0.98	0.99	0.99
FPR	0	0	0	0	0
FNR	0.119	0.078	0.044	0.041	0.029

must have high packet rate, high variation index, and high distinct source IPs which are not close to normal network traffic sample. Therefore, the α-divergence value is greater than the normal range of α-divergence. But in case of low-intensity attack, attack samples have low packet rate, low entropy, and variation of source IPs. So, the α-divergence for low-intensity samples will be close to normal traffic samples and hence, the α-divergence value of a low-intensity samples is bounded by the meanDiv value of the normal profile.

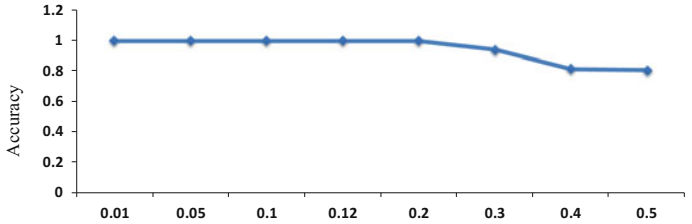
Complexity analysis of the proposed method: The complexity of the method depends only on the captured traffic analysis phase. Complexity of the captured traffic analysis phase is $O(m) + (O(m * k))$, where m is the total number of sample and k is the possible cross-correlation values between two samples at different lags.

Experimental Results

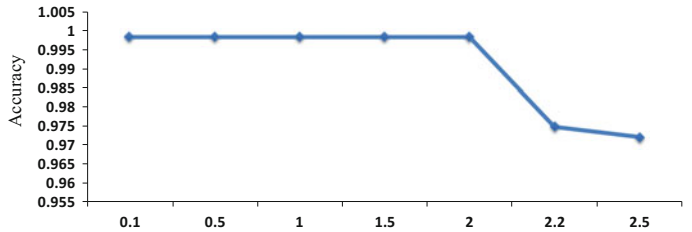
The experiment is done on a test bed that consists of 40 machines and five workstations. Both normal and attack traffic are generated from the test bed. The proposed method is implemented using MATLAB R2015 software. We use DDoS CAIDA 2007 and DARPA 2000 intrusion datasets to validate the method.

Performance analysis on CAIDA and DARPA dataset: In the first phase of detection, the method identifies the samples either as normal or attack but in the second phase of detection, the method identifies the type of the attacks, i.e., low- or high-rate attacks. On CAIDA dataset, HLR_DDoS yields high detection accuracy on certain threshold values. As shown in Fig. 63.1a, if the detection threshold value is less than 0.2, then the proposed method gives 99–100% detection accuracy. However, if the threshold value increases, performance of HLR_DDoS decreases. Similar to DDoS CAIDA 2007, the method gives 100% detection accuracy when the threshold value is less than 2. However, as shown in Fig. 63.1b, detection accuracy of the proposed method is degraded sharply as increased the threshold values beyond 2. So, for the DARPA dataset, the optimal range of threshold is any value smaller than 2.

The method gives high detection accuracy during classification of low-rate as well as high-rate attacks as shown in Table 63.2. The method is compared with three existing methods as shown in Table 63.3. From the comparison table, it can be observed



(a) Accuracy on CAIDA Dataset



(b) Accuracy on DARPA Dataset

Fig. 63.1 Performance analysis of the proposed method on CAIDA and DARPA datasets

Table 63.3 Comparison with other methods on CAIDA dataset

Method	DR (%)	FPR
NOX/OpenFlow [17]	99.11	0.46%
Traffic flow [18]	95	NA
Wavelet analysis	97.95	1.75%
Proposed method	99.8	0.4 %

that on certain threshold values HLR_DDoS efficiently discriminates attacks with low false alarm rate on CAIDA 2007 DDoS dataset.

Conclusion and Future Work

An effective method called HLR_DDoS is proposed to detect DDoS attacks using cross-correlation and α-divergence. Moreover, the method classifies low-intensity attack traffic from high-intensity attack traffic using α-divergence. The method gives high detection accuracy on CAIDA and DARPA datasets. The advantage of the proposed method is that it uses only one threshold value to detect DDoS attacks and identifies attack types either as low rate or high rate. Work is going on to develop a real-time DDoS traceback mechanism to prevent DDoS attacks.

References

1. Hoque, N., Bhuyan, M.H., Baishya, R.C., Bhattacharyya, D., Kalita, J.K.: Network attacks: Taxonomy, tools and systems. Journal of Network and Computer Applications **40** (2014) 307–324
2. Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: Botnet in ddos attacks: trends and challenges. IEEE Communications Surveys & Tutorials **17** (2015) 2242–2270
3. Mirkovic, J., Reiher, P.: A taxonomy of ddos attack and ddos defense mechanisms. ACM SIGCOMM Computer Communication Review **34** (2004) 39–53
4. Xiang, Y., Li, K., Zhou, W.: Low-rate ddos attacks detection and traceback by using new information metrics. Information Forensics and Security, IEEE Transactions on **6** (2011) 426–437
5. Ahmed, E., Mohay, G., Tickle, A., Bhatia, S.: Use of ip addresses for high rate flooding attack detection. In: Security and Privacy–Silver Linings in the Cloud. Springer (2010) 124–135
6. Hoque, N., Bhattacharyya, D.K., Kalita, J.K.: Ffsc: a novel measure for low-rate and high-rate ddos attack detection using multivariate data analysis. Security and Communication Networks **9** (2016) 2032–2041
7. Bhattacharyya D. K., Kalita, J.K.: Ddos attacks: Evolution, detection, prevention, reaction, and tolerance, CRC Press, 2014
8. Feinstein, L., Schnackenberg, D., Balupari, R., Kindred, D.: Statistical approaches to ddos attack detection and response. In: DARPA Information Survivability Conference and Exposition, 2003. Proceedings. Volume 1., IEEE (2003) 303–314
9. Yu, S., Zhou, W.: Entropy-based collaborative detection of ddos attacks on community networks. In: Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on, IEEE (2008) 566–571
10. Xiao, B., Chen, W., He, Y., Sha, E.H.: An active detecting method against syn flooding attack. In: Parallel and Distributed Systems, 2005. Proceedings. 11th International Conference on. Volume 1., IEEE (2005) 709–715
11. Yang, G., Gerla, M., Sanadidi, M.: Defense against low-rate tcp-targeted denial-of-service attacks. In: Computers and Communications, 2004. Proceedings. ISCC 2004. Ninth International Symposium on. Volume 1., IEEE (2004) 345–350
12. Chen, Y., Hwang, K.: Collaborative detection and filtering of shrew ddos attacks using spectral analysis. Journal of Parallel and Distributed Computing **66** (2006) 1137–1151
13. Zhang, C., Yin, J., Cai, Z., Chen, W.: Rred: robust red algorithm to counter low-rate denial-of-service attacks. Communications Letters, IEEE **14** (2010) 489–491
14. Zhang, C., Cai, Z., Chen, W., Luo, X., Yin, J.: Flow level detection and filtering of low-rate ddos. Computer Networks **56** (2012) 3417–3431
15. Wu, Z.j., Lei, J., Yao, D., Wang, M.h., Musa, S.M.: Chaos-based detection of ldos attacks. Journal of Systems and Software **86** (2013) 211–221
16. Zhijun, W., Yi, C., Meng, Y., Lan, M., Lu, W.: Cross-correlation based synchronization mechanism of lddos attacks. Journal of Networks **9** (2014) 604–611
17. Braga, R., Mota, E., Passito, A.: Lightweight ddos flooding attack detection using nox/openflow. In: Local Computer Networks (LCN), 2010 IEEE 35th Conference on, IEEE (2010) 408–415
18. Liu, H., Sun, Y., Valgenti, V.C., Kim, M.S.: Trustguard: A flow-level reputation-based ddos defense system. In: 2011 IEEE Consumer Communications and Networking Conference (CCNC), IEEE (2011) 287–291

Chapter 64
A Symmetric Key-Based Image Encryption Scheme

Jaydeb Bahumik and Supriyo De

Abstract In this information age, secure transmission and storage of digital image are an important requirement. Due to its better performance, symmetric key algorithms are popularly employed to provide confidentiality of digital information. In this paper, a new symmetric key-based image encryption technique has been introduced. In proposed scheme, firstly, a plain image is divided into blocks of 16 pixels and then each block is permutated by an invertible linear transformation. After that, permutated image pixel values are XOR-ed with expanded key bytes. The scheme is implemented and analyzed against statistical analysis and cryptanalysis. It is shown that proposed scheme is secure and easy for implementation.

Keywords Image security · Symmetric key · Linear transformation · Cipher image · Histogram · Correlation · Entropy · Differential attack

Introduction

Cryptography is a tool to transform readable information into a meaningless one. It plays a vital role when the secure information is transmitted through a non-secure channel. This scenario makes it more complex and significant for digital image. Encryption plays an important role which makes the image meaningless to the attacker. Although there exists several well-established encryption [1, 2] techniques such as Data Encryption Standard (DES), International Data Encryption Algorithm (IDEA), RSA, and Advanced Encryption Standard (AES) but with respect to robustness and complexity, they are not suitable for image encryption.

J. Bahumik (✉)
Department of ECE, Haldia Institute of Technology, Haldia, WB, India
e-mail: I.bahumik.jaydeb@gmail.com

S. De
Department of ECE, Saroj Mohan Institute of Technology, Guptipara, WB, India
e-mail: II.supriyo.tech@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_64

In last few years, researchers developed several image encryption schemes. But till date, it is becoming a challenge to develop an efficient image encryption scheme due to large size, redundancy, and strong correlation between pixel values of plain image. In [3], skew tent chaotic map and permutation–diffusion architecture have been proposed. But with respect to differential cryptanalysis, it does not provide satisfactory result in round one and the paper concluded that more than two times iteration of the algorithm can resist the differential attack. Bhowmik et al. [4] developed their work with the help of genetic algorithm (GA). Here, GA has been applied to generate encryption key with the combination of well known Blowfish encryption technique. Using GA, a new approach of image encryption has been developed in [5]. Beside GA-based scheme, several researchers focused their work on AES and modified AES for image encryption [6, 7]. In [8, 9], image encryption schemes based on AES in Electronic Code Book (ECB) mode have been presented. In these schemes, two different linear operations are performed before applying AES. Most of the existing image encryption schemes either do not satisfy all cryptographic parameters or they are costly for implementation in real-time application.

In this paper, a new image encryption scheme has been introduced. The scheme consists of three parts: image permutation, key expansion, and the final one is key XOR-ing with the permutated image. The correlation between adjacent pixels of cipher image significantly reduced by the proposed scheme, and it is observed that linear transformation with optimum number of iterations makes the algorithm lighter and faster.

The rest of the paper is organized as follows. In section "Proposed Scheme," proposed image encryption scheme is described. Section "Experimental Results" elaborately represents the experimental outcomes. In section "Security Analysis," security analysis of the proposed scheme is discussed and finally conclusions are drawn in section "Conclusion."

Proposed Scheme

The said scheme consists of three functional blocks: linear transformation for plain image, key expansion, and the last one is encryption by XOR operation with expanded key and outcome of the permutation block. The details of the each block are explained below.

Linear Transformation for Plain Image (LT_PI)

LT_PI is a linear transformation, and it is a 128×128 binary matrix. The matrix is a tri-diagonal matrix where upper and lower diagonal elements are all ones and main diagonal elements are as follows:
01001000100010000010111110111101011001110000001110011110010111 10

01011110010110001100000011100110101111011111010000010001000100010
In *LT_PI*, at a time, 16 bytes data are XOR-ed with 16 bytes output of *LT_PI*. Then, XOR-ed output is passed through the linear transformation.

Key Expansion

In key expansion algorithm, cipher key is expanded to produce key of size equal to image size. Key is expanded by employing linear transformation for key and nonlinear mixing function *Nmix*.

Linear Transformation for key (*LT_Key*) It takes 16 bytes input (initially 16 bytes cipher key) and produces a 16 bytes output by employing the matrix T_{key} shown in Eq. 64.1(a) and the 16 bytes previously permutated output. T matrix (Eq. 64.1(b)) is a (8×8) binary non-singular matrix and $T^L = I$, where I is a (8×8) identity matrix and $L = 255$. T matrix is as follows:

$$T_{key} = \begin{pmatrix} T^1 & T^2 & T^3 & .... & T^{15} & T^{16} \\ T^2 & T^4 & T^6 & & T^{30} & T^{32} \\ T^3 & T^6 & T^9 & & T^{45} & T^{48} \\ ... & & & & & \\ ... & & & & & \\ T^{15} & T^{30} & T^{45} & & T^{225} & T^{240} \\ T^{16} & T^{32} & T^{48} & & T^{240} & T^1 \end{pmatrix} ..(a) \; and \; T = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} ..(b) \quad (64.1)$$

The characteristic polynomial associated with T matrix is $(T + I.x)$, and it is also a primitive polynomial of degree 8. Input 128-bits cipher key is first passed through *LT_Key*. Then, output of *LT_Key* is mixed with input of *LT_Key* using *Nmix* function. For even/odd key generation, *Nmix* is operated from left to right/right to left, respectively. After 16th iteration, 16th 128-bits key is loaded as an input of *LT_Key* and it repeats until it expands the key size at par with size of plain image. Process has been presented in Fig. 64.1.

Nmix, the nonlinear key mixing function [10], is applied here for the key expansion from the permuted keys. Two different directions of operation of *Nmix* in key expansion algorithm are considered here to produce odd and even 128-bit keys. For an odd set of 128 bits output (obtained from *LT_Key*), the *Nmix* operation is done from the right to left (LSB to MSB). On the other hand, in an even set of key generation, the *Nmix* operation is done from the left to right (MSB to LSB) direction. The detailed equations for right to left *Nmix* operation are expressed in Eqs. 64.2 and 64.3. Similarly, left to right operation is performed starting from MSB.

$$y_i = x_i \oplus k_i \oplus c_{i-1} \quad (64.2)$$

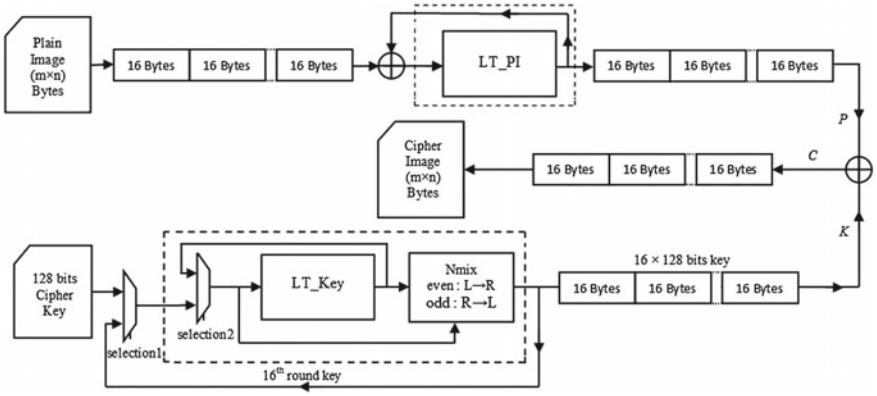


Fig. 64.1 Block diagram of proposed image encryption scheme

$$c_i = \oplus_{j=0}^{i} x_j k_j \oplus x_{i-1} x_i \oplus k_{i-1} k_i \tag{64.3}$$

where *X, K*, and *Y* are n-bit variables, c_i is the carry term propagating from the *i*th bit position to $(i + 1)$th bit position; $0 \leq i \leq n - 1$ and $0 \leq j \leq n - 1$.

Encryption

Image encryption part is done by bitwise XOR operation of permutated image with expanded keys. Here, the cipher image is obtained from the XOR operation between expanded keys(K) and permuted image(P).

$$C = P \oplus K \tag{64.4}$$

where P: permuted image, K: expanded key, C: cipher image

Decryption

Image decryption is done using following steps.
Step 1: Expand cipher key
Step 2: XOR between cipher image and expanded key
Step 3: Reverse image permutation

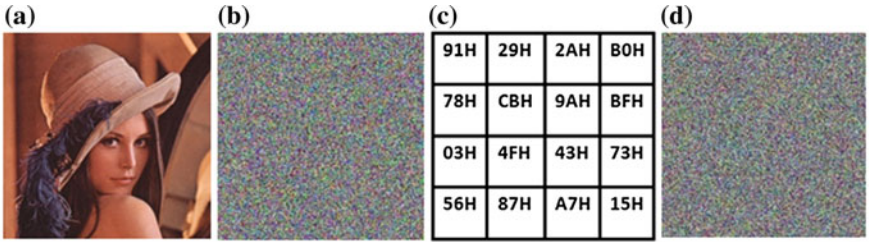


Fig. 64.2 **a** Lena image; **b** Permuted Lena; **c** 128 bits key; **d** Expanded key(256×256)

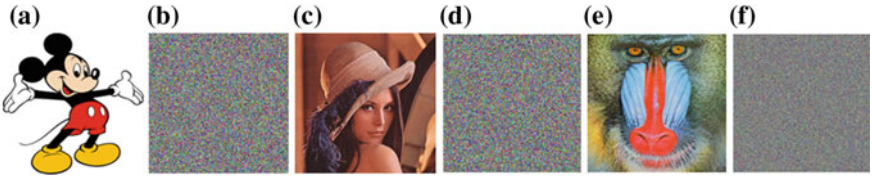


Fig. 64.3 **a** Micky image; **b** Encrypted Micky; **c** Lena image; **d** Encrypted Lena; **e** Baboon image; **f** Encrypted Baboon

Experimental Results

Proposed scheme is tested by using several plain images. The plain image "Lena" and result of the permutation block are shown in Figs. 64.2(a) and 64.2(b), respectively. On the other hand, initial 128 bits key and expanded keys for encryption are shown in Figs. 64.2(c) and 64.2(d), respectively. Figure 64.3 represents the plain images and corresponding cipher images obtained by applying the proposed scheme.

Security Analysis

In the following section, security analysis of proposed scheme is discussed.

Key Space Analysis

Key space indicates the all possible set of keys which can be used for encryption. Security level is proportional to the key space size. In proposed scheme, key length is 128 bits, i.e., the key space size is 2^{128}. As the key space size is larger than 2^{104} [11, 12], it can be stated that the proposed scheme can prevent the brute force attack.

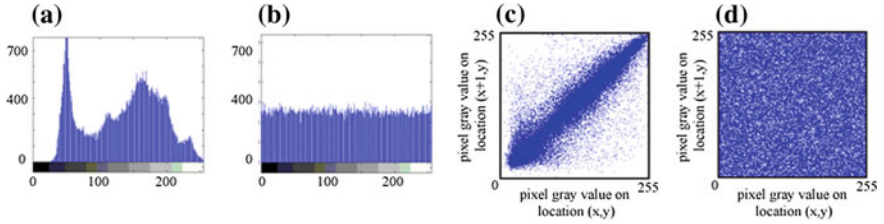


Fig. 64.4 **a** Histogram of Lena; **b** Histogram of encrypted Lena; **c** Correlations of two horizontally adjacent pixels of Lena; **d** Correlations of two horizontally adjacent pixels of encrypted Lena

Statistical Analysis

Statistical analysis is absolutely necessary to verify the encryption scheme. Robustness of the algorithm and the random distribution of the cipher images can be categorized by this analysis.

Histogram Analysis An image histogram illustrates how the pixel values are distributed throughout the image. The uniform distribution of histogram indicates to have equal probability of all color intensity level which is highly required for the cipher images. Plot of histograms for "Lena" image and corresponding cipher image are shown in Figs. 64.4(a) and 64.4(b), respectively. Here, the histogram analysis is justifying the proposed scheme.

Correlation Analysis Correlation analysis is used to find out the relation between two adjacent pixels of an image in all possible direction. High correlation is an intrinsic nature of general digital image. On the other hand, for an encryption scheme, it is the primary target to break the correlation between two adjacent pixels of the image. Correlation analysis has been done for plain image and cipher image produced by proposed scheme. The equation to compute correlation may be found in [3]. Comparison of correlation with other existing methods of image encryption schemes are provided in Table 64.1. Figure 64.4c, d shows the correlation of two horizontally adjacent pixels of "Lena" image and corresponding cipher image, respectively.

Entropy Analysis

Entropy is a mathematical tool which is basically used for checking the uncertainty of a signal or sequence. For an image encryption scheme having maximum values of entropy implies that it is more significant encryption. The column 5 of the Table 64.2 represents the entropy value of cipher images obtained by employing proposed scheme which is nearly closed to ideal 8-bits random sequences. Table 64.2 shows that the proposed scheme is competent enough with other existing image encryption schemes.

Table 64.1 Correlation coefficient of two adjacent pixels in the plain image and the cipher images

Image	Direction	Plain image	Cipher image [3]	Cipher image [7]	Cipher image-proposed
Micky	Horizontal	0.9350	–	−0.0112	−0.0024
	Diagonal	0.8891	–	0.0009	−0.0019
	Vertical	0.9244	–	−0.0813	0.0012
Lena	Horizontal	0.9327	−0.0008	–	0.0031
	Diagonal	0.9288	−0.0001	–	−0.0049
	Vertical	0.9468	0.0037	–	−0.0019
Baboon	Horizontal	0.9210	–	–	−0.0025
	Diagonal	0.8287	–	–	0.0033
	Vertical	0.8550	–	–	−0.0018

Table 64.2 Entropy value for the cipher images

Image	Cipher image [3]	Cipher image [7]	Cipher image [8]	Cipher image-proposed
Micky	–	7.9992	7.9973	7.9971
Lena	7.9974	–	–	7.9972
Baboon	7.9993	–	–	7.9993

Table 64.3 NPCR and UACI of cipher images with respect to key sensitivity

Image	NPCR		UACI	
	Reference [3]	Proposed	Reference [3]	Proposed
Micky	–	99.6332	–	33.5005
Lena	–	99.6332	–	33.5629
Baboon	99.6052	99.6187	33.4132	33.4645

Sensitivity Analysis

Block cipher-based cryptosystem can be analyzed by sensitivity analysis. It is a technique to check how the encryption scheme reacts with respect to one-bit change of key and/or plaintext. The number of pixels change rate (NPCR) and unified average changing intensity (UACI) [13, 14] are measured for this analysis. Mathematical expressions to compute NPCR and UACI may be found in [14].

Key Sensitivity In this test, one-bit difference between two 128 bits keys is employed for encrypting a plain image and then from the two different cipher images, NPCR and UACI values are computed and comparative study is shown in Table 64.3.

Plaintext Sensitivity The sensitivity is measured by finding out the NPCR and UACI from two cipher images(using same key) obtaining from two plain images with one-bit difference. The Table 64.4 shows the result of this study.

Table 64.4 NPCR and UACI of cipher images with respect to plaintext sensitivity

Image	NPCR				UACI			
	Reference [3] 1st round	Reference [3] 2nd round	Reference [7]	Proposed	Reference [3] 1st round	Reference [3] 2nd round	Reference [7]	Proposed
Micky	–	–	99.58	97.5631	–	–	29.63	32.6957
Lena	37.6389	99.6063	–	97.7071	12.7034	33.4758	–	32.8327
Baboon	84.1255	99.6048	–	98.7178	28.1799	33.4554	–	33.1385

From Table 64.3, it is observed that NPCR and UACI values with respect to key sensitivity are better compared to existing scheme. On the other hand, in Table 64.4 the NPCR values with respect to plaintext are slightly lesser than existing scheme in [3, 7], whereas in Table 64.4, UACI values of proposed scheme is better than existing scheme in [3, 7].

Conclusion

In this paper, a new image encryption scheme is introduced. The scheme consists of three functional blocks. In permutation step, the scheme generates same size of image from plain images using *LT_PI*. Next step, the key expansion module expands the key with same size of plain image using *LT_Key* and *Nmix* function. Third block performs XOR-ing of permutated image with the expanded same size key. The experimental results show that the scheme is able to break the correlation and provides the random outcome as a cipher image. Proposed scheme is able to prevent the brute force attack, statistical attack as well as differential attack.

References

1. Stallings, W.: Cryptography and Network Security: Principles and Practice. Prentice-Hall, Upper Saddle River, NJ. (1999)
2. Daemen, J., Rijmen, V.: The Design of Rijndael- AES, The Advanced Encryption Standard. Springer-Verlag. (2002)
3. Zhang, G., Liu, Q.: A Novel Image Encryption Method Based on Total Shuffling scheme. Journal of Optics Communications, Elsevier, vol. 284, Issue 12, 2775–2780 (2011)
4. Bhowmik, S., Acharyya, S.: Image Cryptography: The Genetic Algorithm Approach. IEEE International Conference on Computer Science and Automation Engineering, vol. 2, 223–227 (2011)
5. Kumar, J., Nirmala, S.: Encryption of Images Based on Genetic Algorithm A New Approach. Advances in Computer Science, Engineering & Applications, vol. 167 of the series Advances in Intelligent Systems and Computing, 783–791 (2012)
6. Huang, C. W., Tu, Y. H., Yeh, H. C., Liu, S. H., Chang, C. J.: Image observation on the modified ECB operations in Advanced Encryption Standard. Proc. of Int. Conf. on Information Society (i-Society), London, 264–269 (2011)
7. Shtewi, A. A., Hasan, B. E. M., Hegazy, A. E. F. A.: An Efficient Modified Advanced Encryption Standard (MAES) Adapted for Image Cryptosystems. International Journal of Computer Science and Network Security, vol. 10, 226–232 (2010)
8. De, S., Bhaumik, J.: TBLT-AES: A Robust Image Encryption Scheme. Journal of Discrete Mathematical Sciences & Cryptography, (Taylor & Francis, Co-published with TARU Publications) vol. 17, no. 3, 273–288 (2014)
9. De, S., Bhaumik, J.: An AES-based Robust Image Encryption Scheme. Int. Journal of Computer Applications, vol. 109, no. 12, 29–34 (2015)
10. Bhaumik, J., Chowdhury, D. R.: Nmix: An Ideal Candidate for Key Mixing. Proc. of Int. Conf. on Security and Cryptography (Secrypt), Milan, Italy, 285–288 (2009)
11. Alvarez, G., Li, S.: Some basic cryptographic requirements for chaos-based cryptosystems. Int. Journal of Bifurcation and Chaos, vol. 16, no. 8, 2129–2151 (2006)

12. Stinson, D.: Cryptography: Theory and Practice. Second ed. CRC/C&H
13. Wu, Y., Noonan, J. P., Agaian, S.: NPCR and UACI Randomness Tests for Image Encryption. Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT), 31–38 (2011)
14. El-Wahed, M. A., Mesbah, S.,Shoukry, A.: Efficiency and Security of Some Image Encryption Algorithms. Proc. of the World Congress on Engineering, vol. 1. (1982)

Chapter 65
Anonymous RFID Authentication for IoT in LTE-A

Hiten Choudhury

Abstract Due to advantages like global reach, cost-effective connectivity and easy maintenance, machine to machine (M2M) communication through long-term evolution advanced (LTE-A) is emerging as an infrastructure for the Internet of things (IoT). In LTE-A, base stations can be configured as evolved NodeBs (eNBs) in macrocells and home eNBs (HeNBs) in femtocells. Through these stations, connections among M2M devices can be extended in both indoor and outdoor conditions. Radio frequency identification device (RFID) has the potential to play a crucial role as an M2M device in the IoT. However, anonymous authentication of RFID in LTE-A is a challenge, as pseudonymization is used instead of anonymization in LTE-A. Moreover, a typical anonymization technique may not be suitable for RFID, because it is limited by power, storage, and computational capability. In this paper, a lightweight scheme for anonymous RFID authentication in LTE-A is proposed.

Keywords Anonymous authentication · Anonymization · IoT · RFID
Lightweight cryptography · Security · Privacy

Introduction

Machine to machine (M2M) communication allows machines or devices to communicate with each other without the intervention of any human beings. The Internet of things (IoT) may involve a huge number of interconnected devices and appliances that are used in our day-to-day life. M2M communications through mobile networks like long-term evolution advanced (LTE-A) are expected to play an important role in the deployment of the IoT [13]. LTA-A features several advantages, such as global reach, cost-effective connectivity, and easy maintenance, thereby providing a ready-to-use infrastructure for the IoT. In LTE-A, evolved NodeBs (eNBs) in macrocells

H. Choudhury (✉)
Department of Computer Science and Information Technology,
Cotton University, Guwahati 781001, Assam, India
e-mail: hiten.choudhury@cottonuniversity.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_65

and home eNBs (HeNBs) in femtocells can be configured as base stations [3, 9]. Through these base stations, connections among M2M devices can be extended in both outdoor and indoor conditions.

Radio frequency identification device (RFID) has the potential to play a crucial role as an M2M device in the IoT [12, 13]. However, due to limitation of RFID in terms of power, computation, and storage, anonymous authentication of RFID in LTE-A is a challenge.

In the authentication protocol used in LTE-A, pseudonymization is used instead of anonymization for protecting the identity privacy of the subscriber/device. Which means, for identity presentation during authentication and key agreement in LTE-A, pseudonyms (temporary identities) are used instead of the real identity of the endpoints [1].

Recently, 3GPP has released a technical specification [4]. The objective of this document is to analyze privacy- related key issues impacting 3GPP networks and to establish privacy- handling guidelines for future specifications. It presents privacy principles that should be followed in 3GPP when designing new systems, security architectures, and protocols. In this document, it is mentioned that 'pseudonymization' (use of temporary identities) is less efficient than 'anonymization.'

Considering 3GPP's observation about anonymization [4] and keeping the limitations of RFID in mind, in this paper a lightweight cryptographic scheme for anonymous RFID authentication in LTE-A is proposed.

The remaining portion of the paper is arranged as follows: In section "Authentication in LTE-A," the authentication and key agreement protocol of LTE-A is studied; in section "The Problem," the security problem that is addressed in this paper is discussed; in section "Related Work," similar works that have already been done in this area are discussed; in section "The Proposed Scheme," a proposal for anonymous authentication of RFID in LTE-A for IoT is presented; a computational cost analysis of the proposal is performed in section "Computational Overhead"; in section "Space Overhead," the space overhead analysis of the proposal is performed; in section "Security Analysis," a security analysis of the proposal is made; the paper is concluded in section "Conclusion."

Authentication in LTE-A

In LTE-A, each and every M2M device, like any normal human subscriber, has to be registered with its home network, where the subscription information of the device is stored in a server called the home subscriber server (HSS) [1] (Fig. 65.1). In a visited network, the device (for example RFID) is authenticated on behalf of the HSS by an entity called the mobility management entity (MME). The authentication protocol requires the following to be shared in advance between the device and the HSS [1].

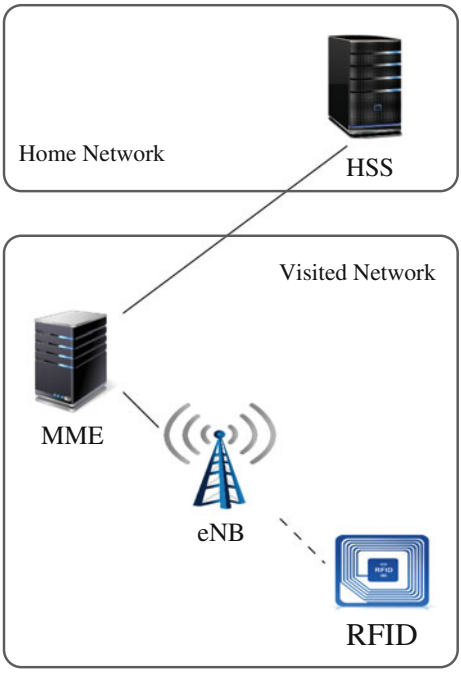


Fig. 65.1 Security architecture of LTE-A

- A secret key: k.
- The identity of the device called the international mobile subscriber identity: *IMSI* [2].
- A set of hash functions: $f_0 - f_5$.

The authentication protocol used for M2M communication in LTE-A works as follows (Fig. 65.2) [1].

1. **Device**: Transmits its *IMSI* to the MME (through the eNB).
2. **MME**: Forwards the *IMSI* to the HSS.
3. **HSS**: Locates the secret key k stored against *IMSI* in its database and computes a vector V, which it sends along with *IMSI* back to the MME. V is computed as follows.

$$R = f_0(Seed) \tag{65.1}$$

$$X = f_{2_k}(R) \tag{65.2}$$

$$C = f_{3_k}(R) \tag{65.3}$$

$$I = f_{4_k}(R) \tag{65.4}$$

$$A = f_{5_k}(R) \tag{65.5}$$

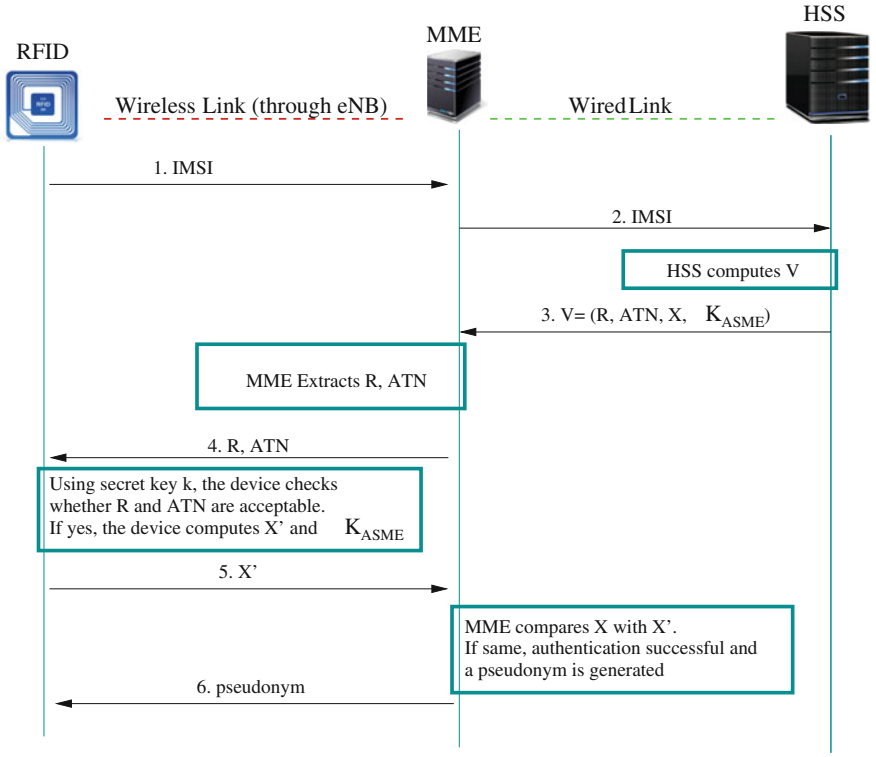


Fig. 65.2 Authentication protocol used in LTE-A

$$M = f_{1_k}(SQ \| R) \tag{65.6}$$

$$ATN = (SQ \oplus A \| M) \tag{65.7}$$

$$K_{ASME} = KDF(C, I) \tag{65.8}$$

$$V = (R, ATN, X, K_{ASME}) \tag{65.9}$$

where, R: random number, X: expected response, C: cipher key, I: integrity key, A: anonymity key, SQ: sequence number, M: message authentication code, ATN: authentication token, K_{ASME}: access security management entity, KDF: a key derivation function, '||': concatenation, and $\oplus$: XOR operation.

4. **MME**: *IMSI* and K_{ASME} are stored in the *IMSI*-to-K_{ASME} mapping table that is maintained in the MME database. After this, R and ATN are extracted from V and are transmitted to the device.
5. **Device**: Computes A (using Eq. 65.5) and M (using Eq. 65.6). SQ is then extracted from the ATN using A. If ATN is acceptable (which is verified based on the fresh-

ness of SQ and the comparison of the computed M with the received M), the device computes the following.

$$X' = f_{2k}(R) \tag{65.10}$$

$$C = f_{3_k}(R) \tag{65.11}$$

$$I = f_{4_k}(R) \tag{65.12}$$

$$K_{ASME} = KDF(C, I) \tag{65.13}$$

It then transmits X' back to the MME.

6. **MME**: Compares X' with X. If they are same, the authentication is successful, otherwise the MME sends an authentication failure message to the device.

Therefore, at the end of a successful authentication, a K_{ASME} is shared between the device and the MME [1]. The K_{ASME} is used for confidentiality and integrity protection of the communication between the device and the MME. It may also be used for future authentications that may be performed locally between the device and MME. The MME assigns a new pseudonym [2] to the device through the confidentiality-protected channel, so that for identity presentation during future authentications, it may be used by the device instead of the *IMSI* .

The Problem

The transmission of the *IMSI* of a device should as far as possible be avoided. If it is exposed, a person/device with malicious intent may track it and prepare an elaborate profile about the movement of the device. Therefore, for identity presentation, a pseudonym is transmitted in place of the *IMSI*. However, transmission of *IMSI* cannot always be avoided through pseudonymization, like in the following situations.

- During the first authentication in the lifetime of the device, when the device has not even acquired a pseudonym [1].
- When the MME cannot map a received pseudonym to its *IMSI*.
- When the MME of a newly visited network, cannot acquire the *IMSI* from the previous MME.

Therefore, it has become extremely important that effort be made to adopt anonymization instead of pseudonymization, so that anonymity of the device may be strengthened.

Related Work

Identity privacy in mobile networks has been an area of research for quite some time. In the early days of mobile communications, when researchers started recognizing privacy as an important security issue, several schemes and protocols were proposed to improve identity privacy. Some of these proposals are the ones proposed by Asokan [6], Lin et al. [19], Park et al. [20], Barbeau et al. [7]. However, during those days, emergence of IoT was not anticipated.

Keeping the computational and power constraints of mobile computing environment in mind, several schemes and protocols in general, that are not targeted at any specific mobile technology, were proposed recent times. In these proposals, computationally light techniques that use symmetric key-based cryptosystems, hash functions, XOR operations, temporary identities, alias names, etc., are used. Some of these protocols are the ones proposed by Tang et al. [21], Zhou et al. [23], Chen et al. [10], He et al. [14], Jiang et al. [15], etc.

Several schemes and protocols were proposed to strengthen privacy of a subscriber in LTE. Li et al. [18] proposed an enhanced authentication and key agreement protocol. This protocol is designed to overcome the existing vulnerabilities such as disclosure of identity. However, computationally intensive public key infrastructure is used in this proposal and are not suitable for RFID. In [11], Choudhury et al. proposed a protocol for enhanced identity privacy. The proposal uses symmetric key cryptography, XOR operations, and hash functions. However, this work is based on pseudonymization. In [17], Koien et al. have proposed to enhance mutual identity privacy in LTE where a dummy identity (pseudonym) is used. In [5], Angermeier et al. proposed a privacy-preserving scheme for LTE in vehicular communication. The scheme makes use of pseudonymous identifiers. In [12].

Recently, Fan et al. proposed an anonymous RFID authentication scheme for IOT in 5G. However, the scheme allows the shared secret keys, between the RFID and the database, to be transferred to the RFID readers. This may lead to serious compromise in privacy and overall security of the devices, as there may be several readers in a network and there could be possibility of the keys being leaked through any of them.

Thus, the need to design an anonymous authentication scheme for devices like RFID that are going to contribute in a big way to the IoT of things through networks like LTE-A continues to exist.

The Proposed Scheme

In this section, a lightweight anonymization scheme that may be adopted in LTE-A for authentication of M2M devices like RFID is proposed.

The scheme requires that a unique *IMSI* is assigned to every device through the identity module. For successful implementation of this scheme, the set of *IMSI*s

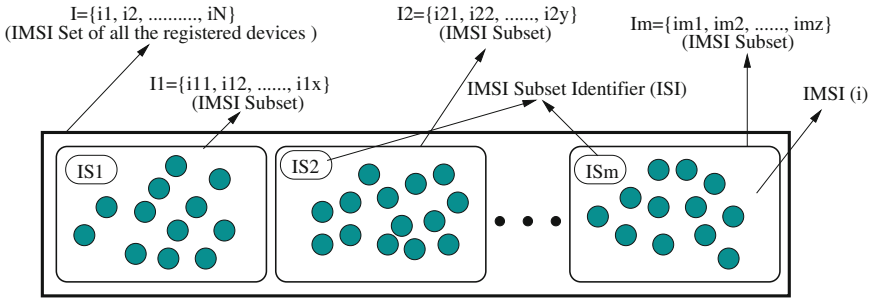


Fig. 65.3 Division of the IMSI Set *I*, of all the *N* registered subscribers at the HSS, into *m* IMSI subsets

(say *I*) stored at the HSS, containing the IMSIs, viz., i_1, i_2, ..., i_n, of all the (say *N*) registered devices, is divided into *m* subsets, viz., $I_1, I_2, ..., I_m$ (Fig. 65.3). Therefore,

$$I = \{i_1, i_2, ..., i_N\} \tag{65.14}$$

$$I = \{I_1, I_2, ..., I_m\} \tag{65.15}$$

The value of *m* is to be decided by the operator, depending on the pace at which it wishes to identify the *IMSI* of a device. The larger is the value of *m*, the faster it is for the operator. The *N IMSI*s in *I* are randomly distributed among the *m* subsets of *I*. Therefore,

$$I_1 = \{i_{11}, i_{12}, ..., i_{1x}\} \tag{65.16}$$

$$I_2 = \{i_{21}, i_{22}, ...i_{2y}\} \tag{65.17}$$

$$I_m = \{i_{m1}, i_{m2}, ..., i_{mz}\} \tag{65.18}$$

where $x + y + ... + z = N$.

Each of the *m* subsets of *I*, viz., $I_1, I_2, ..., I_m$, has a unique IMSI-Subset-Identifier (ISI), viz., $SI_1, SI_2, ..., SI_m$, respectively. The ISI of the subset to which the *IMSI* of a particular device belongs has to be stored in the device's identity module, as well as in the device's profile in HSS database. Therefore, in addition to *IMSI*, *k* and $f_0 - f_5$ (as explained in section "Authentication in LTE-A"), the *ISI* is also stored at the device and at the HSS. Also, two one-way hash functions, *H* and *h* (explained later in this section), are shared between the M2M device and the HSS.

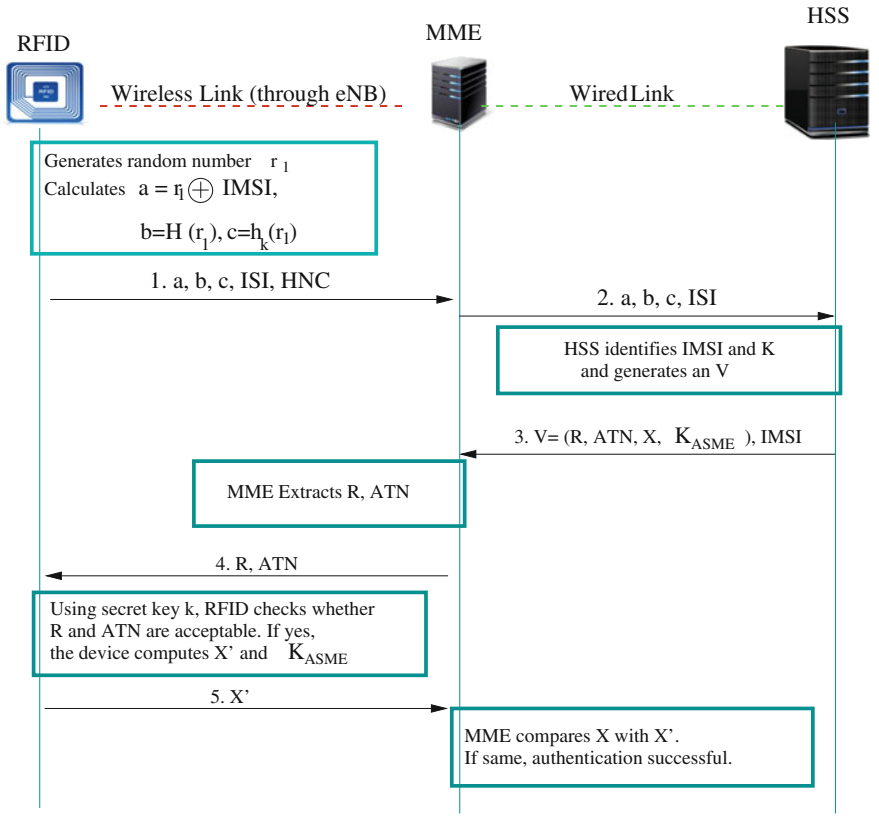


Fig. 65.4 Initial connection through a particular MME using the proposed scheme

The Initial Connection Through a Particular MME

For the initial connection through a particular MME that is located in a newly visited network, the device begins by generating a 128 bit random number r_1 (Fig. 65.4). It then calculates the following.

$$a = r_1 \oplus IMSI \tag{65.19}$$

$$b = H(r_1) \tag{65.20}$$

where H is a one-way hash function.

$$c = h_k(r_1) \tag{65.21}$$

where k is the secret key that is shared between the device and the HSS, and h is a keyed one-way hash function.

The rest of the scheme proceeds according to the following steps:

1. **Device**: Initiates the procedure by transmitting a, b, c, ISI, and HNC to the MME (through the eNB).
2. **MME**: Forwards a, b, c, and ISI to the HSS.
3. **HSS**: $\forall IMSI' \in I_{ISI}$ (where I_{ISI} is the IMSI subset whose subset identifier is ISI), calculates $r'_1 = a \oplus IMSI'$, and verifies if $H(r'_1) = b$ and if $h_{k'}(r'_1) = c$; where k' is the key that is stored against $IMSI'$ at the IMSI-to-key mapping in the HSS database. If for any $IMSI'$ (say $IMSI_d$) and k' (say k_d) the verifications are successful, $IMSI_d$ is established as the IMSI of the device and k_d is identified as the secret key of the device.
   After secret key k_d is identified, a vector V is computed using k_d in the same way as in section "Authentication in LTE-A." The HSS then sends V and $IMSI_d$ back to the MME.
4. **MME**: Takes out R and ATN from V and sends them to the device as a challenge. $IMSI_d$ and K_{ASME} are stored in the $IMSI$-to-K_{ASME} mapping table that is maintained in the MME database.
5. **Device**: Computes A (using Eq. 65.5) and M (using Eq. 65.6). SQ is then extracted from the ATN using A. If ATN is acceptable (which is verified based on the freshness of SQ and the comparison of the computed M with the received M), the device computes the following.

$$X' = f_{2k}(R) \tag{65.22}$$

$$C = f_{3_k}(R) \tag{65.23}$$

$$I = f_{4_k}(R) \tag{65.24}$$

$$K_{ASME} = KDF(C, I) \tag{65.25}$$

   It then transmits X' back to the MME.
6. **MME**: Compares X' with X. If they are same, the authentication is successful, otherwise the MME sends an authentication failure message to the device.

At the end, a shared secret K_{ASME} is established between the device and the MME.

Subsequent Connections Through the Same MME

For subsequent connections through the same MME, the AKA procedure may be performed locally between the device and the MME, without involving the HSS, thereby reducing communication latency. However, the MME may choose to fall back on the full authentication procedure (presented in section The Initial Connection Through

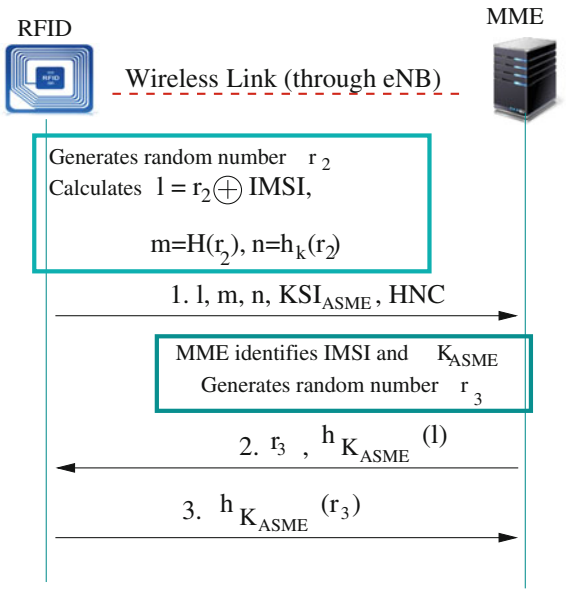


Fig. 65.5 Subsequent connections through the same MME using the proposed scheme

a Particular MME) beyond a certain threshold time (say T) that the device has spent with the same MME. The value of T depends on the operator's security policy.

The authentication procedure that can be executed locally between the device and the MME is as follows (Fig. 65.5).

The device begins by generating a 128 bit random number r_2. It then calculates the following.

$$l = r_2 \oplus IMSI \tag{65.26}$$

$$m = H(r_2) \tag{65.27}$$

$$n = h_{k_{ASME}}(r_2) \tag{65.28}$$

The rest of the scheme proceeds according to the following steps.

1. **Device**: Initiates the procedure by transmitting a request to the MME (through the eNB). The request should include l, m, n, the KSI_{ASME} (say KSI'_{ASME}), and the HNC (say HNC_d) of the device.
2. **MME**: Extracts the *IMSI*s of all the devices whose HNC is HNC_d and that were authenticated by it in the past time period T (where T is the threshold time). All these extracted *IMSI*s are then kept in a set (say I_T). The MME then, $\forall IMSI' \in I_T$, calculates $r'_2 = l \oplus IMSI'$, and verifies if $H(r'_2) = m$ and if $h_{k'_{ASME}}(r'_2) = n$; where k'_{ASME} is the key that is stored against $IMSI'$ and KSI'_{ASME} at the IMSI-to-key mapping in the MME database. If for any $IMSI'$ (say $IMSI_d$) and k'_{ASME} (say k'_{ASME_d}) the verifications are successful, the MME is convinced that k'_{ASME_d} is the

key for access security management of the device and $IMSI_d$ is the *IMSI* of the device. The MME then generates a random number r_3 and calculates an expected response X.

$$X = h_{k_{ASME}}(r3) \tag{65.29}$$

It then transmits r_3 to the device.

3. **Device**: Calculates a response X'.

$$X' = h_{k_{ASME}}(r3) \tag{65.30}$$

X' is then transmitted back to the MME.

4. **MME**: Verifies whether $X = X'$. If they match, the authentication is considered successful. Otherwise, the MME communicates an authentication failure message to the device.

Computational Overhead

In the proposed scheme, lightweight cryptographic functions like hash functions and XOR operations are used. While designing the scheme, the following factors are taken into account: (a) the limited battery, space, and computational capability of the device/RFID (b) the theoretically unlimited battery, space, and computational capability of the HSS. Considering the computation involved in one hash operation as h, one XOR operation as x and generation of a random number as g, the computational overhead introduced at the device, the MME and the HSS, over and above what is required by the authentication protocol used in LTE-A, can be derived as follows.

During Initial Authentication

- *Device*: The cryptographic operations that are performed at the device and the computation involved therein are as follows.
  - For generating a random number r_1, computation involved is g
  - For calculating a, the computation involved is one XOR operation: x
  - For calculating b and c, the computation involved is $2h$

  Therefore, the total computation involved in executing the scheme at the device is $g + x + 2h$.
- *MME*: No additional computation is introduced.

- *HSS*: For every request, the HSS, $\forall IMSI' \in I_{ISI}$, calculates $r'_1 = a \oplus IMSI'$, and verifies if $H(r'_1) = b$ and if $h_{k'}(r'_1) = c$. Considering the average number of IMSIs in a particular IMSI subset to be v, the computational overhead introduced at the HSS is $v * (x + 2h)$.

During Subsequent Authentication

- *Device*: The cryptographic operations that are performed at the device and the computation involved therein are as follows.
  - For generating a random number r_2, computation involved is g
  - For calculating l, the computation involved is one XOR operation: x
  - For calculating m and n, the computation involved is $2h$

  Therefore, the total computation involved in executing the scheme at the device is $g + x + 2h$.
- *MME*: For every request, the MME, $\forall IMSI' \in I_T$, calculates $r'_2 = a \oplus IMSI'$, and verifies if $H(r'_2) = b$ and if $h_{k'_{ASME}}(r'_2) = c$. Considering the number of *IMSI*s that belongs to the same HNC as the received *IMSI* and that are with the MME for the past T seconds, to be u, the computational overhead introduced at the HSS is $u * (x + 2h)$.
- *HSS*: There is no computation at the HSS.

Therefore, as evident from the above derivations, most of the computation is introduced at the HSS, a minimal amount of computation is introduced at the device, and no additional burden is introduced at the MME during the initial authentication. However, during subsequent connections within time T, some amount of computation is involved at the MME, with no additional computation at the HSS.

Space Overhead

The proposed scheme needs additional memory space at the device, MME and the HSS for its successful implementation. The amount of space required in the random access memory and in the read-only memory of the device, MME and the HSS, may be computed as follows.

- *Device:*
  - 128 bits, to temporarily store r_1/r_2.
  - 128 bits to temporarily store a/l.
  - 128 bits to temporarily store b/m.
  - 128 bits to temporarily store c/m.
  - 128 bits to permanently store KSI_{HN}.

Therefore, a total of 56 bytes in the random access memory and 16 bytes in the read-only memory will be required additionally at the device.

- *MME:*
  - 128 bits to temporarily store r_2.
  - 128 bits to temporarily store a/l.
  - 128 bits to temporarily store b/m.
  - 128 bits to temporarily store c/n.
  - 128 bits to temporarily store KSI_{HN}.

  Therefore, a total of 62 bytes in the random access memory will be additionally required at the MME.
- *HSS:*
  - 128 bits to temporarily store a.
  - 128 bits to temporarily store b.
  - 128 bits to permanently store KSI_{HN} against each of the device's profile.

  Therefore, a total of $N \times 32$ bytes in the random access memory and $N \times 16$ bytes of space in the read-only memory will be additionally required at the HSS where N is the total number of registered devices/subscribers at the HSS.

Security Analysis

In this section, a security analysis of the proposed scheme is performed with respect to some common vulnerabilities that a device is susceptible to.

- *Non-eavesdropping:* Eavesdropping is the act of listening to the private conversation of others without their consent or knowledge [22]. In LTE-A, the MME and the device are connected through a wireless link, which is vulnerable to eavesdroppers due to its open nature. However, neither the *IMSI* nor any other private information is transmitted in clear text. Therefore, eavesdroppers cannot obtain any secret information regarding the device.
- *Non-tracking:* The *IMSI* is never transmitted through the radio interface. Moreover, the values of a/l, b/m, and c/n are different for every session because r_1/r_2 is randomly chosen. Therefore, there is no fixed information that can be used by an adversary to link two subsequent connections.
- *Non-denial-of-service:* A Denial-of-service attack (DoS attack) is an attempt to make a service unavailable to its intended users. Such an attack involves flooding the target device with fake communications requests, such that the target device cannot respond to legitimate request or responds so slowly as to be considered effectively unavailable [8]. With no knowledge of the *IMSI*, the secret key k and k_{ASME}, the attacker cannot disguise a device to affect the authentication procedure. Therefore, the proposal can prevent the denial-of-service (DoS) attack.

- *Replay attack:* A replay attack is a form of network attack in which a valid data transmission is maliciously or fraudulently repeated at a later instance, masquerading it as a regular transmission. This is carried out either by the originator or by an adversary who intercepts the data and retransmits it [16]. In our scheme, even if an adversary retransmits a, b, c, ISI, and HNC or X', the involvement of a sequence number SQ, as in the original authentication protocol, and a freshly generated random number r_3, during subsequent authentication, ensures that a replay attack cannot succeed.
- *Forward security:* Even if the attacker obtains a/l, b/m, c/n, and ISI, she/he cannot compromise any precious information regarding the device, because the value of a/l, b/m, and c/n (that depends on random number $r1/r2$) changes with every authentication. Therefore, the proposal ensures forward security.

Conclusion

M2M communication in LTE-A is emerging as an infrastructure for the IoT, and RFID has the potential to play a crucial role as an M2M device in the IoT. However, anonymous authentication of RFID in LTE-A is a challenge, as pseudonymization is used instead of anonymization in LTE-A. Moreover, a typical anonymization technique may not be suitable for RFID, because it is limited by power, storage, and computational capability. There are several proposals for anonymous authentication in mobile networks and for RFID, in general. However, none of them are specifically targeted at authentication of RFID over LTE-A. In this paper, a lightweight scheme that may be adopted in LTE-A for anonymous authentication of RFID is proposed.

References

1. 3GPP: 3GPP System Architecture Evolution (SAE); Security architecture. TS 33.401, 3rd Generation Partnership Project (3GPP) (2011), http://www.3gpp.org/ftp/Specs/html-info/33401.htm
2. 3GPP: Numbering, addressing and identification. TS 23.003, 3rd Generation Partnership Project (3GPP) (2011), http://www.3gpp.org/ftp/Specs/html-info/23003.htm
3. 3GPP: Evolved universal terrestrial radio access (E-UTRA) and evolved universal terrestrial radio access network (EUTRAN), Overall Description. TS 36.300, 3rd Generation Partnership Project (3GPP) (2012)
4. 3GPP: Study on subscriber privacy impact in 3GPP. TS 33.849, 3rd Generation Partnership Project (3GPP) (2016)
5. Angermeier, D., Kiening, A., Stumpf, F.: Pal-privacy augmented lte: A privacy-preserving scheme for vehicular lte communication. In: Proceeding of the tenth ACM international workshop on Vehicular inter-networking, systems, and applications. pp. 1–10. ACM (2013)
6. Asokan, N.: Anonymity in a mobile computing environment,! In: IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 1994), Santa Cruz, CA. pp. 200–204 (1994)

7. Barbeau, M., Robert, J.M.: Perfect identity concealment in UMTS over radio access links, In: IEEE International Conference on Wireless And Mobile Computing, Networking And Communications (WiMob 2005), Montreal, Canada. pp. 72–77 (2005)
8. Carl, G., Kesidis, G., Brooks, R., Rai, S., A, B., C, D., D, E.: Denial-of-service attack-detection techniques, IEEE Internet Computing 10(1)(1), 82–89 (2006)
9. Chandrasekhar, V., Andrews, J.G., Gatherer, A.: Femtocell networks: a survey. IEEE Communications magazine 46(9), 59–67 (2008)
10. Chen, C., He, D., Chan, S., Bu, J., Gao, Y., Fan, R.: Lightweight and provably secure user authentication with anonymity for the global mobility network, International Journal of Communication Systems 24(3)(3), 347–362 (2011)
11. Choudhury, H., Roychoudhury, B., Saikia, D.K.: Enhancing user identity privacy in lte. In: 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. pp. 949–957. IEEE (2012)
12. Fan, K., Gong, Y., Liang, C., Li, H., Yang, Y.: Lightweight and ultralightweight rfid mutual authentication protocol with cache in the reader for iot in 5g. Security and Communication Networks (2015)
13. Ghavimi, F., Chen, H.H.: M2m communications in 3gpp lte/lte-a networks: architectures, service requirements, challenges, and applications. IEEE Communications Surveys & Tutorials 17(2), 525–549 (2015)
14. He, D., Ma, M., Zhang, Y., Chen, C., Bu, J., A, B., B, C.: A strong user authentication scheme with smart cards for wireless communications, Computer Communications 34(3)(3), 367–374 (2011)
15. Jiang, Q., Ma, J., Li, G., Yang, L., A, B., B, C., C, D.: An enhanced authentication scheme with privacy preservation for roaming service in global mobility networks, Wireless Personal Communications 68(4)(4), 1–15 (2012)
16. Kavitha, T., Sridharan, D.: Security vulnerabilities in wireless sensor networks: A survey, Journal of information Assurance and Security 5(1)(1), 31–44 (2010)
17. Koien, G.M.: Privacy enhanced mutual authentication in lte. In: Wireless and Mobile Computing, Networking and Communications (WiMob), 2013 IEEE 9th International Conference on. pp. 614–621. IEEE (2013)
18. Li, X., Wang, Y.: Security enhanced authentication and key agreement protocol for LTE/SAE network, In: IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM' 11), Wuhan, China. pp. 1–4 (2011)
19. Lin, H.Y., Harn, L.: Authentication protocols for personal communication systems, ACM SIGCOMM Computer Communication Review 25(4)(4), 256–261 (1995)
20. Park, J., Go, J., Kim, K., A, B., B, C., C, D.: Wireless authentication protocol preserving user anonymity, In: Symposium on Cryptography and Information Security (SCIS 2001), Oiso, Japan. pp. 159–164 (2001)
21. Tang, C., Wu, D.O.: Mobile privacy in wireless networks-revisited, IEEE Transactions on Wireless Communications 7(3)(3), 1035–1042 (2008)
22. Zhang, Y., Zheng, J., Ma, M., A, B., B, C., C, D.: Handbook of Research on Wireless Security,. Information Science Reference-Imprint of: IGI Publishing, Hershey, PA (2008)
23. Zhou, T., Xu, J.: Provable secure authentication protocol with anonymity for roaming service in global mobility networks, Computer Networks 55(1)(1), 205–213 (2011)

Chapter 66
An Indirect Addressing Image Steganographic Scheme Using 9 × 9 Sudoku Matrix

Ambar Bandhyopadhyay, Debanjali Dey, Rajat Kumar Pal and Arnab Kumar Maji

Abstract In recent world, everybody wants to transmit their valuable information securely without the knowledge of the intruder. One of the way-out is steganography. Using steganography, information can be transferred using cover media. Cover media can be image, video, sound, or text. Most popularly used cover media is a digital image because the slight modification of a cover image is difficult to distinguish by human eye. İn our paper, the proposed method uses Sudoku puzzle to hide secret message within a cover image. The benefit of using Sudoku is if any intruder modifies the secret message then receiver will be able to detect this. In this proposed method, the length of the secret message is directly proportional to the resolution of the cover image. The proposed method has been developed to transmit the secret message more securely. Here, a single 9 × 9 Sudoku puzzle is used, and this Sudoku puzzle is embedded within the image by modifying the RGB value of pixels. Within each cell of the Sudoku puzzle we indirectly store one character of the secret data; hence, the intruder cannot fetch the secret message directly, to detect if intruder modifies the secret message or not, and associated information is stored which makes the proposed method more reliable than other existing methods. This proposed method is better than the existing methods because it uses RGB value, whereas the most of the methods used grayscale value.

A. Bandhyopadhyay (✉) · D. Dey · R. K. Pal
Department of Computer Science and Engineering, University of Calcutta, JD-2, Sector-III, Saltlake 700098, Kolkata, India
e-mail: ambarbandanyopadhyay@gmail.com

D. Dey
e-mail: de.debanjali@gmail.com

R. K. Pal
e-mail: pal.rajatk@gmail.com

A. K. Maji
Department of Information Technology, North Eastern Hill University, Shillong 793022, India
e-mail: arnab.maji@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_66

Keywords Steganography · Sudoku puzzle · Secret message
Cover image

Introduction

In today's world, the most important issue of information and communication technology is the security of information. Cryptography is a methodology for preserving the secrecy of message. To keep the message secret, different methodologies are developed for encryption and decryption of data. The popularity of multimedia data creates a new opportunity for communication using digital steganography.

The word "steganography" came from Greek word "Steganos." The meaning of it is "covered" or "reticent." The following formula provides a standard description of the steganographic process:

$$cover\_medium + hidden\_data + stego\_key = stego\_medium$$

The cover_medium is the file in which we hide the hidden_data, which can be encrypted using the stego_key. The consequential file is the stego_medium (which is, of course be the similar category of file as the cover_medium). Thus, the cover_medium (and hence, the stego_medium) is naturally image or audio/video file. In this paper, we focus on image files and, therefore, refer them as *cover_image* and *stego_image*.

The primary goal of our proposed method is to transfer the secret message embedded securely so that the intruder will not get any clue about the message in the image and if somehow intruder modifies the secret message knowingly or unknowingly, it will be detected.

Some of the key terminologies that are used in this paper are discussed in the next subsection.

Key Terminologies

- Secret message: Data which sender wants to send to receiver securely.
- Cover image: Taking the cover object (within which we embed the secret message) as image in steganography is known as cover image.
- Intruder: A person who intrudes, especially who fetch and modify the secret data
- Sudoku puzzle: A puzzle in which players insert the numbers into a grid consisting of blocks into a further smaller subblocks in such a way that every number appears once in each horizontal line, vertical line, and blocks.

- Encryption and decryption: The process of converting the data into cipher text is called encryption. So that the intruder cannot get any secret information from that cipher text. The process of converting the cipher text into data is called decryption.

The paper is organized into six sections. In section "Introduction to Sudoku," Sudoku is introduced where the logic behind Sudoku is discussed. In section "Existing Literature," existing literatures are discussed. In section "Proposed Technique," the proposed technique is introduced and the algorithm of the proposed technique is also written. In section "Experimental Results," the experimental result of the proposed technique is given. In section "Conclusion," conclusions are given where a judgment about the proposed technique is written.

Introduction to Sudoku

Sudoku is a puzzle in which some predefined clues are given. The most common size of Sudoku is 9 × 9, i.e., 9 rows and 9 columns. Nine rows and columns, formulates 3 × 3 smaller subgrid, which is popularly known as minigrid. The aim of this game is to fill up all 9 rows, columns, and minigrids with the numbers (generally 1–9) in such a way that, all the numbers occurs exactly once.

In Fig. 66.1, a sample Sudoku puzzle is shown and its solution (we called it as Sudoku matrix) is shown in Fig. 66.2.

		3	7					
2					1	6	9	
	8			9			5	
					9			
6				3			2	
		4		1		5	7	
3					5	9	6	
	5					1		
8						3		

Fig. 66.1 An instance of a Sudoku puzzle

Fig. 66.2 A solved Sudoku puzzle

4	9	3	7	5	6	2	1	8
2	7	5	8	4	1	6	9	3
1	8	6	3	9	2	7	5	4
5	2	8	6	7	9	4	3	1
6	1	7	5	3	4	8	2	9
9	3	4	2	1	8	5	7	6
3	4	2	1	8	5	9	6	7
7	5	9	4	6	3	1	8	2
8	6	1	9	2	7	3	4	5

Existing Literature

There are several algorithmic techniques for image steganography; in this section we are going to discuss some of them.

Gupta et al. [1] have proposed a data concealing plan utilizing the minimum huge piece steganography alongside cryptographic strategy. In this proposed methodology, the raw information was scrambled before inserting it into the picture. This framework makes utilization of Rivest, Shamir, Adleman (RSA) calculation, and the Diffie Hellman calculation to scramble the secret data. To give higher security, the secret information is encoded and after that it was changed over to double frame. In the mean time, the picture pixels were additionally changed over to parallel shape and afterward the scrambled mystery data was installed into the picture by a LSB encoder.

Khare et al. [2] have proposed a framework that enables client to exchange instant messages safely by concealing them into a computerized picture document. This framework was a blend of both steganography and encryption calculations and gives a solid spine to its security. This framework additionally proposed a structure for concealing the extensive volume of information in pictures while acquiring negligible perceptual debasement. In this proposed strategy, entropy thresholding (ET) and selectively embedding in coefficients (SEC) were utilized to choose whether or not to install the mystery information into the picture. This framework was mostly utilized for applications that require high-volume implanting with vigor.

Saradha et al. [3] have proposed a plan for enhancing information concealing limit utilizing Sudoku confound in shading pictures. The primary thought of the plan was to utilize a Sudoku astound, in which each esteem relates to a pixel match (red, blue) of the picture mapped with the mystery by supplanting a couple of one pixel of two hues. This plan was proposed to enhance the visual nature of the stego image and to enhance the normal concealing limit of the picture to 4 bpp (bits per pixel). This plan makes utilization of 24 bit of a picture and adjusts 16 bit of every pixel. The Sudoku arrangement was utilized as a kind of perspective lattice for both inserting and extricating the secret information into and from the picture.

Tawade et al. [4] have proposed a proficient information concealing plan utilizing mystery reference networks. The information was covered up in 8-bit grayscale image utilizing 256 × 256 grid which was built by utilizing 4 × 4 table with unreported digits from 0 to 15. The proposed technique was to enhance the holding limit of cover picture and increment the many-sided quality to split the secret reference matrix (SRM). They additionally proposed another spatial area information concealing plan by utilizing a secret reference grid (SRM) for information embedding and extraction.

Tyagi et al. [5] have proposed a steganographic technique utilizing least significant bit (LSB) alongside a cryptographic calculation. The symmetric cryptographic calculation was utilized for encryption of the mystery message. This calculation utilizes arbitrary size of the key. In the wake of changing over the data into mystery code or scrambled shape, it was fixed into the picture. For fixing the mystery information, the slightest noteworthy piece of the picture was utilized.

Khalaf et al. [6] have proposed a strong steganography procedure in view of LSB matching. In this plan, the mystery information was disguised in the picture in view of LSB addition and RSA encryption method. The key of the proposed system was to encode the mystery information. At that point the scrambled information was changed over into a bit stream and partitioned it into various sections. The cover picture is likewise separated into a similar number of portions. Each portion of information was contrasted and each section of the picture to locate the best match fragment, so as to make another irregular arrangement of portions to be embedded into the cover picture.

Chithra et al. [7] have proposed a steganography system by utilizing Sudoku Puzzle and ECC algorithm. It changes over the first information into mystery code. In the proposed technique, the picture steganography is finished utilizing Sudoku bewilder. Along these lines, this technique shrouds the information, as well as proselytes the first information into mystery code. In this way, by consolidating these two strategies the information will be all the more safely exchanged by this approach.

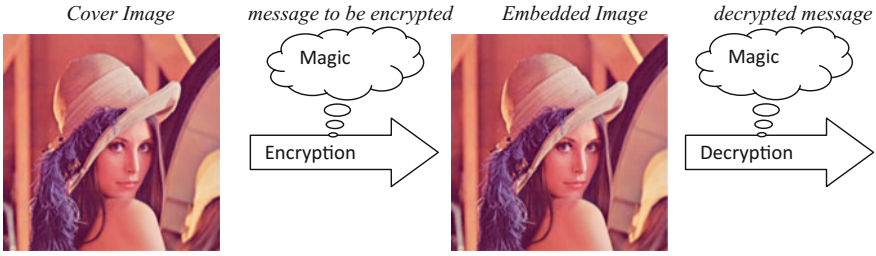


Fig. 66.3 Basic structure of the proposed method

Proposed Technique

All the existing methods which specify in the subpart "Existing Literature" uses "Tile Matrix." In order to make the tile matrix, more storage space is required. In some existing methodology, 256 × 256 tile matrix is also used which used too many storage space but here only 9 × 9 Sudoku matrix is used. And also in the existing methods, the length of the message is restricted, where as in this proposed method the length of the message directly proportional to the resolution of the image, for example, if the resolution of the cover image (in which image the secret message is hidden) is 256 × 256 then the number of characters it can hide will be greater than the data embedding capacity of the cover image whose the resolution is 128 × 128. Also in the existing methods, first RGB value is converted to gray scale then the secret message is embedded then again it is converted to RGB so extra computation time is needed for these conversions, whereas in this proposed method RGB value is directly used so extra computation is not needed here. One of the major drawbacks of the existing methods is that if any intruder changes the pixel values then it is no way possible to report about the corrupted image which is the most hazardous drawback among all. Now in this section, a new method has been proposed to omit the above disadvantages (Fig. 66.3).

Here, a 9 × 9 Sudoku matrix is used as shown in Fig. 66.4. Sudoku is used to encrypt and decrypt the secret message. The proposal method is discussed with an example. Let the secret message is "MBA" so first of all every character is converted into its ASCII value. The ASCII value of the characters "M", "B", "A" are 77, 66, 65, respectively. Now 32 is subtracted from each of the ASCII value because in the ASCII table the corresponding character of 1–31 is not been used which helps to make each value within two digit because the highest ASCII value that can be the part of the secret message is 127 and subtracting 32 from it will give value less than 100. So after subtracting the values will become 45, 34, 33. In each cell of the Sudoku, one character is stored so "M" will be stored at (1, 1)th position in the Sudoku, "B" will be stored at (1, 2)th position in the Sudoku, "A" will be

Fig. 66.4 A solver Sudoku puzzle used as an example

6	3	4	9	1	5	7	2	8
2	7	8	6	4	3	5	1	9
5	9	1	2	7	8	6	4	3
4	5	7	3	6	9	2	8	1
9	8	6	4	2	1	3	5	7
3	1	2	8	5	7	4	9	6
1	2	5	7	8	6	9	3	4
8	6	3	5	9	4	1	7	2
7	4	9	1	3	2	8	6	5

stored at (1, 3)th position in the Sudoku. In our proposed method, row and column start from 1.

Now the element of (1, 1)th position in the Sudoku is 6. So the mechanism is that six steps will be moved in forward side (counting the starting position) then the new position will be the pivot. Here, the pivot will be (1, 6). Now from that position the value (in this case it is 45) is searched along column wise taking one digit at a time. So first 4 will be searched and it is in (8, 6) so the displacement will be $(8 - 1 + 1)$ or 8 and also for 5 again it will be searched and it is in (1, 6) position so here the displacement is $(1 - 1 + 1)$ or 1. Now to check at the receiver side whether the intruder modifies the value or not another information also be sent and that is the element situated at (8, 1)th position in the Sudoku matrix (here 8 and 1 are two displacement), here it is 8.

Now, for "B" the elements of (1, 2)th position in the Sudoku is 3. So here 3 steps will be moved in forward direction then the new position will be the pivot. Here the pivot will be (1, 4). Now from that position first, we search 3 in column wise it is at (4, 4)th position so the displacement will be $(4 - 1 + 1)$ or 4 and also for 4 again it will searched and it is at (5, 4)th position so the displacement will be $(5 - 1 + 1)$ or 5. Now to check at the receiver side whether the intruder modifies the value or not another information also be sent and that is the element situated at (4, 5)th position in the Sudoku matrix (here 4 and 5 are two displacement), here it is 6.

Now, for "A" the elements of (1, 3)th position in the Sudoku is 4. So here 4 steps will be moved in forward side then the new position will be the pivot. Here the pivot will be (1, 6). Now from that position first, we search 3 in column wise it is at (2, 6)th position so the displacement will be $(2 - 1 + 1)$ or 2 and also for 3 (second one) again it will searched and it is at (2, 6)th position so the displacement will be

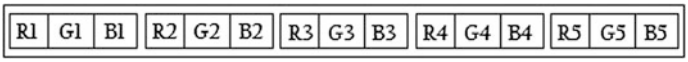


Fig. 66.5 Structure of cell of Sudoku matrix

(2 − 1 + 1) or 2. Now to check at the receiver side whether the intruder modifies the value or not another information also be sent and that is the element situated at (2, 2)th position in the Sudoku matrix (here 2 and 2 are two displacement), here it is 7.

So in this method one set we will transmit one Sudoku elements, two displacements digit, and one check digit. Each cell will have 5 pixels as shown in Fig. 66.5. So to store a Sudoku matrix along with the secret message we need total (9 × 9 × 5) or 405 pixels. Now each pixel contain three parts R(Red), G(Green), and B(blue). The value of Sudoku element must be within 1–9, so we need at most 4 binary bits; so for this 4 pixels are used, and it is stored in the LSB of the G part of the pixel, so that for this change there will be minimal change in the cover image, to store this we use G1, G2, G3, G4. Now the two displacements are stored in B part, the first displacement is stored in second last bit and the last displacement is stored in LSB now to store this value we also need 4 binary bits but from pivot we can go upward or downward for this an extra bit is needed for each displacement so for this 5 B parts are taken (for example, if the pivot is (4, 3) now to search 8 we need to move upward whereas to search 2 we need to move downward). The value of the check digit must also be within 1–9 so we need at most 4 binary bits; so for this 4 pixels are used, and it is stored in the LSB of the R part of the pixel, so that for this change there will be minimal change in the cover image, to store this we use R1, R2, R3, R4.

Now at the receiver end first take the Sudoku matrix, then start fetching the character by starting at the position (1, 1) that is 6 (if not modified by intruder) then again we move 6 steps from (1, 1) so now we are in (1, 6)th position now fetch two displacement from the received image it will be 8 and 1. Now take out the check digit from the received image (here it is 8) and check the element at the (8, 1)th position in the Sudoku matrix is 8 or not, if it is also 8 then we can proceed otherwise we report that intruder modifies the message. If the image is not corrupted then calculate the digit at 8 position displacement from (1, 6) and here it is (8, 6) and the element of the Sudoku matrix at that position is 4 now the digit at 1 position displacement from (1, 6) is (1, 6) and the element of the Sudoku matrix at this position is 5. Now concatenate 4 and 5 so the final value will be 45 now add 32 with it to get the ASCII value of the corresponding character here the final ASCII value will be 77 and the corresponding character will be “M” proceeding in this way decryption is done for all characters. Following this way all the encrypted characters can be retrieved successfully at the receiver end.

Algorithm for Encryption:

Input: Cover image, Sudoku matrix and Secret message.
Output: Cover image, where the secret message is embedded

Step 1: Start.
Step 2: Read the cover image.
Step 3: Calculate the length of the line and store the length into a variable 'msglength'.
Step 4: At the end of the secret message insert a symbol '$' to know the end of the secret message then convert each character of the secret message into the corresponding ASCII value and store it in the array.
Step 5: If the size of the message is greater than 81 then divide the message into several parts where each part except last part will contain 81 characters and the last part will contain less than or equal to 81 character then for each part do steps 6 to 31
Step 6: for i=1 to 9 by 1 do //Store the Sudoku matrix into the array
    Step 6.1: for j=1 to 9 by 1 do
        Step 6.1.1: content=Sudoku(i,j)
        Step 6.1.2: Convert the content into 4 digit binary number
        Step 6.1.3: fetch 5 pixels at a time.
        Step 6.1.4: Store this 4 digit binary number into the LSB of the Green part of the first 4 pixel.
    Step 6.2: end
Step 7: end
Step 8: i=1
Step 9: j=1
Step 10: for k=1 to msglength
    Step 10.1: content=Sudoku(i,j)
    Step 10.2: secretdata=asciimessage(k)
    Step 10.3: secretdata=data-32
    Step 10.4: firstbit=secretdata/10
    Step 10.5: secondbit=secretdata%10
    Step 10.6: From the point (i, j) go forward 'content' number of steps then this new position will be the pivot let it be (p_i, p_j) here.
    Step 10.7: From that position search first bit column wise and then find the displacement from that point let it be d1 here.
    Step 10.8: From that position search second bit column wise and then find the displacement from that point let it be d2 here.
    Step 10.9: checkbit = Sudoku(d1, d2)
    Step 10.10: Fetch 5 pixel at a time.
    Step 10.11: Store the check bit into LSB of the Red part of the first 4 pixel.
    Step 10.12: Store d1 into second Last bit of the Blue part of the last 4 pixels among 5 pixels and now if from pivot the d1 displacement comes going upward then in the second last bit of the Blue part of the first pixel store 1 else 0.
    Step 10.13: Store d2 into LSB the Blue part of the last 4 pixels among 5 pixels and now if from pivot the d1 displacement comes going upward then in the LSB of the Blue part of the first pixel store 1 else 0.
    Step 10.14: increment j by 1 and then if j exceeds 9 then initialize j by 1 and increment i by 1.
Step 11: end.
Step 12: Stop.

Algorithm for Decryption:

Input: Cover image in which secret message is embedded
Output: Secret message

Step 1: Start
Step 2: read the cover image
Step 3: imgrow=1
Step 4: imgcolumn=1
Step 5: for i=1 to 9 by 1 do
    Step 5.1: for j=1 to 9 by 1 do
        Step 5.1.1: Take out the 5 pixel set image.
        Step 5.1.2: Now from the set take the LSB of Green part of the first 4 pixels and after getting 4 binary digits concatenate it and convert it into decimal and store it in a variable 'content'.
        Step 5.1.3: Sudoku(imgrow, imgcolumn)=content
        Step 5.1.4: Increment imgcolumn by 1 after that if it exceed 9 then initialize imgcolumn by 1 and increment imgrow by 1
    Step 5.2: end
Step 6: end
Step 7: imgrow=1
Step 8: imgcolumn=1
Step 9: for i=1 to 9 by 1 do
    Step 9.1: for j=1 to 9 by 1 do
        Step 9.1.1: Take out the 5 pixels set from the cover image.
        Step 9.1.2: Now from the set take the LSB of Red part of the first 4 pixels and after getting 4 binary digits concatenate it and convert it into decimal and store it in a variable 'checkbit'.
        Step 9.1.3: Now from the set take the second last bit of Blue part of last 4 pixels and after getting 4 binary digits concatenate it and convert it into decimal and then store it in a variable 'd1' now if in the first pixel of the Blue part the second last bit is 1 then multiply d1 by -1 else don't do anything.
        Step 9.1.4: Now from the set take the second last bit of Blue part of last 4 pixels and after getting 4 binary digits concatenate it and convert it into decimal and then store it in a variable 'd2' now if in the first pixel of the Blue part the LSB is 1 then multiply d2 by -1 else don't do anything.

        Step 9.1.5: if Sudoku(absolute_value(d1),absolute_value(d2)) != checkbit then
            Intruder modifies the message , give error message to the receiver and exit.
        Step 9.1.6: end
        Step 9.1.7: Now from pivot go to displacement d1 and store the element stored in Sudoku in variable 'firstbit'.
        Step 9.1.8:Now from pivot go to displacement d2 and store the element stored in Sudoku in variable 'secondbit'.
        Step 9.1.9: Concatenate secondbit at the end of firstbit and then increment the resultant by 32 now this will be the ASCII value of one of the character of the secret message and store it in a variable 'asciino'.
        Step 9.1.10: After converting this asciino into the corresponding character using ASCII standard if the character equals to '$' then
            Step 9.1.10.1: this will be the end of the message so exit
        Step 9.1.11: else
            Step 9.1.11.1: show the character to the receiver.
        Step 9.1.12: end
    Step 9.2: end
Step 10: end
Step 11: Till now '$' is not recognized so go to Step 3.
Step 12: Stop

Advantages of this Proposed Technique:

(a) This method ensures that if the intruder modifies the data then it can be caught.
(b) All the existing methods first convert the color image into gray-level image and then encryption is done therefore it is time consuming, whereas the method directly encrypt in color image so conversion from color image to gray image not needed for this reason this method will need less computation.
(c) Here the size of the message totally depends on the resolution of the image so it is not constant.
(d) Here any kind of tile matrix is not used which generally uses more space.
(e) Here it is not needed to send Sudoku matrix explicitly, it is embedded into the cover image with the secret message.

Experimental Results

The distortion in cover image depends upon the change in value of pixels and number of pixels of cover image used for embedding which in turn depends upon the number of components of the pixel used and amount of input data. PSNR value is used to evaluate the quality of an image. The PSNR is defined as follows:

$$\text{PSNR} = 10\log_{10}255^2 / \text{MSE dB}$$

where MSE is the mean square error between the original image and the stego image. The MSE is defined as follows:

$$\overset{V-1W-1}{\text{MSE}} = 1/3 * V * W \sum_{i=0}^{} \sum_{j=0} \sum \left(R_{ij} - R'_{ij}\right)^2 + \sum \left(G_{ij} - G'_{ij}\right)^2 + \sum \left(B_{ij} - B'_{ij}\right)^2$$

V * W gives number of all pixels display in a picture. A bigger PSNR shows that the nature of the stego picture is nearer to the first one. Ordinarily, human's eyes think that it is difficult to recognize the mutilations on a stego picture contrasted with unique picture when its PSNR esteem is more noteworthy than 30 dB.

The width and height of the following image is 512 × 512 and throughout the discussion this image is used (Fig. 66.6 and Table 66.1).

Fig. 66.6 Cover image

Table 66.1 Performance of the proposed technique shown below

Message	PSNR (db)		
	Red	Green	Blue
SO	91.5583	81.4866	86.7871
ME	90.8889	81.4866	86.7871
MY	90.8889	81.4866	87.8786
US	91.5583	81.4866	87.1650
WE	91.2107	81.4866	86.6681
BYE	90.8889	81.4866	87.5789
YOU	91.2107	81.4866	87.5789
SIX	91.5583	81.4866	88.0365
OUT	90.3089	81.4866	88.0365
LAW	91.9362	81.4866	86.6681
FIND	90.0457	81.4866	85.7226
HANG	89.7974	81.4866	85.3604
LOCK	90.5892	81.4866	86.4395
OKAY	89.7974	81.4866	85.1074
GOAL	90.8889	81.4866	85.7226
HELLO	90.0457	81.4866	84.7914
ARRAY	89.3398	81.4866	87.4365
BASIC	89.7974	81.4866	86.0155
POINT	90.5892	81.4866	85.8180
TURBO	89.3398	81.4866	86.6681
%56#	90.0457	81.4866	87.0354
^&*123	87.2986	81.4866	87.7262
#58974*	86.6681	81.4866	77.4929
CALCUTTA UNIVERSITY	83.0052	81.4866	74.3293

Comparison of Experimental Results

See Table 66.2.

Analysis of the Experimental Result

Here Table 66.1 shows the experimental result of the new method, where as in Table 66.2 it is compared with some existing methods. By looking at Table 66.2, it can be seen that this proposed method is better than the existing methods. It can be seen that existing methods use gray level, whereas proposed method uses RGB

Table 66.2 Average PSNR in dB for n characters

Number of character	Existing algorithm		Proposed algorithm		
	Existing algorithm1 Gray level	Existing algorithm2 Gray level	Red	Green	Blue
2	85.81132	88.63408	91.22102	81.4866	87.05718
3	84.17294	86.58102	91.27355	81.4866	87.58
4	83.1824	85.2717	90.19405	81.4866	85.91465
5	81.62497	83.594757	89.82238	81.4866	86.1459
6	81.0947	82.9763	87.2986	81.4866	87.7262
7	80.0971	81.2331	86.6681	81.4866	77.4929

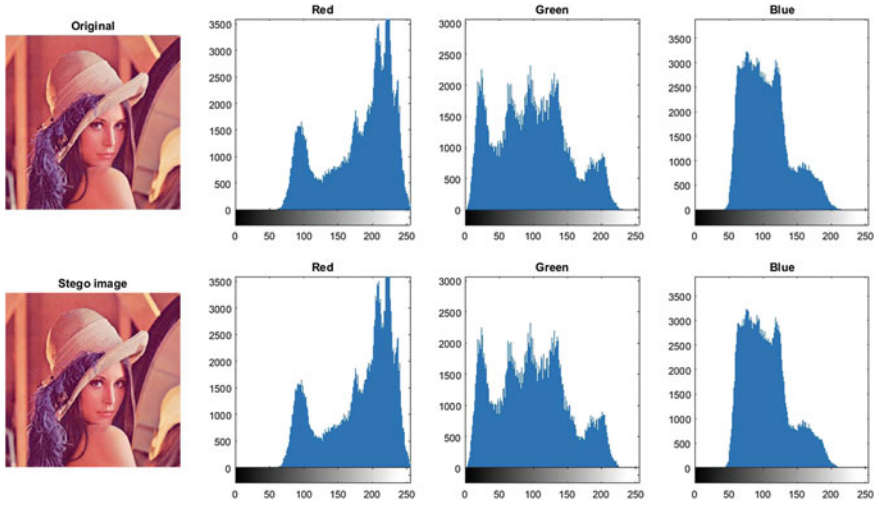


Fig. 66.7 Histograms of original and stego image

value and also for comparison PSNR value is used where the proposed method shows its quality. For these comparisons, same 21 text messages and three numerical values are used for comparison with the existing methods and the proposed methods.

Now for experimental purpose, a 65 characters secret message is embedded within an image using proposed method and after that the original and stego image and there histogram of red, green, blue component is shown in Fig. 66.7. It is clear from this figure that human cannot distinguish these two images easily by just seeing.

By using the proposed method steganalysis of the stego image is also performed, and we find that the proposed scheme in this paper is safe against geometrical attack, noise based attack, and so on.

Conclusion

Data transmission is an important issue, and everyone wants to transmit their data more securely. In this paper, a new method has been proposed to transmit data more securely. Here, we have used 9*9 Sudoku to secure the message. The best part of our algorithm is that if someone (intruder) modifies the data then this algorithm can detect this and show the appropriate error. Here, we have treated the elements of the Sudoku as index which is a new concept, and before this paper no one has ever used this concept, to the best of the author's knowledge. This paper presents a secure technique based on a steganographic method using Sudoku matrix. For embedding the cipher into the cover image, the Sudoku matrix is used. Thus, the data will be more securely transferred by embedding the cipher into an image. Hence, this technique can be used where we want to transmit data securely. Here, 9×9 Sudoku matrix has been used to transmit the data more securely. All the previous methods use either 256×256 or 27×27 or 18×18, whereas in this work 9×9 Sudoku matrix has been used; thus, it saves the computation time and space. Therefore, it can be concluded that the proposed method is more robust with less computation.

References

1. V.S. Gupta, A. Goyal and B. Bhushan, "Information Hiding Using Least Significant Bit Steganography and Cryptography". In: *International Journal of Modern Education and Computer Science*, Vol. 6, pp. 27–34 (2012).
2. A. Khare, M. Kumari and P. Khare, "Efficient Algorithm for Digital Image Steganography". In*: Journal of Information, Knowledge and Research in Computer Science and Applications*, Vol. 1, No. 1, pp. 1–5 (2010).

3. P. Saradha and B. Swamy, "Improving Image Data Hiding Capacity Scheme using Sudoku Puzzle in Color Images". In: *International Journal of Engineering Research and Applications*, Vol. 2, No. 3, pp. 2741–2744 (2012).
4. L. Tawade, R. Mahajan and C. Kulthe, "Efficient & Secure Data Hiding Using Secret Reference Matrix". In: *International Journal of Network Security & Its Applications*, Vol. 4, No. 1, pp. 43–50 (2012).
5. V. Tyagi, A. Kumar, R. Patel, S. Tyagi and S.S. Gangwar, "Image Steganography using Least Significant Bit with Cryptograph". In*: Journal of Global Research in Computer Science*, Vol. 3, No. 3, pp. 53–55 (2012).
6. E.T. Khalaf and N. Sulaiman, "A Robust Data Hiding Technique based on LSBMatching". In: *World Academy of Science, Engineering and Technology*, Vol. 58, pp. 117–121 (2011).
7. B. Chithra, D. Harinath, M.V.R. Murthy and K.R. Babu, "Enhancing Security through Steganography by using Sudoku Puzzle and ECC Algorithm". In: *International Journal for Research in Science Engineering and Technology*, Vol. 2, Issue 6, pp. 29–43 (2015).

Chapter 67
Video Steganography Techniques in Spatial Domain—A Survey

Mukesh Dalal and Mamta Juneja

Abstract Steganography is considered as a useful approach for hiding data in a carrier file; data could be any multimedia file such as image, text, audio, video. Steganography is referred to as the science of invisible communication. Sometimes steganography is used with encryption schemes of cryptography to provide additional security for the secret data. A good steganographic technique must fulfill three basic requirements of steganography, i.e., capacity, imperceptibility, and robustness. Videos have become a very popular choice for steganography due to its significant growth on Internet and are able to fulfill all these requirements due to its large size and statistical complexity. This paper presents the survey of video steganography schemes based on spatial domain techniques.

Keywords Cryptography · Video steganography · Spatial domain
Imperceptibility · Payload

Introduction

Steganography is defined as a technique which hides message in a way that the intruder not even realizes that the secret message is being exchanged. The word came from Greek words steganos meaning covered (secret) and graphy meaning drawing (writing). Steganography has been used since ancient times, where to send a message, a wood with wax tablets base was used keeping the top of the wax blank [1]. Almost every web page on Internet contains multimedia files of different types so to hide the secret message, any of these multimedia files can be utilized. Nowadays, due to the growth of videos over Internet, they are becoming powerful carrier for hiding secret data as compared to other files. They have high embedding

M. Dalal (✉) · M. Juneja
UIET, Panjab University, Chandigarh, India
e-mail: mukeshdalal05@gmail.com

M. Juneja
e-mail: mamtajuneja@pu.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_67

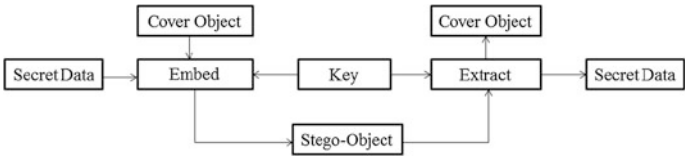


Fig. 67.1 Basic steganography model

capacity and are robust against attacks due to their complex structure. The embedding of secret data in videos can be done by using spatial domain techniques such as least significant bit (LSB), RGB (red, green, blue) components, YUV components, histograms. The goal of this paper is to do survey of various spatial domain techniques presented in the literature. The basic steganography model is shown below in Fig. 67.1.

Literature Survey

Most of the researchers utilized LSB technique of spatial domain for embedding secret data, and there also exist other spatial domain techniques such as YUV components, RGB components which are utilized by very few of the researchers. The literature survey related to these spatial domain techniques is discussed in this section.

Hanafy et al. [2] utilized two bits of the red, blue, and green channels for embedding. The secret data was first segmented into blocks, and its randomization was done before embedding. Pseudorandom locations were searched for embedding the secret data with the help of secret key. This approach was experimented upon different multimedia data types such as text, image, audio, and video with different size, and the obtained results demonstrated that PSNR was above 50 dB for embedding all type of data with high embedding capacity.

Bhattacharya et al. [3] proposed an approach for high-resolution AVI video steganography based on directed graph pattern. This approach converted the frames of cover video to bitmap format and utilized graph direction to insert data bits in cover video with LSB replacement. The method used two graph patterns for embedding process, and results demonstrated that they have high embedding capacity. This scheme was implemented in spatial domain; it can further improve robustness and security by applying the frequency domain methods.

Sherly and Amritha [4] proposed a video steganography algorithm called tri-way pixel-value differencing with pseudorandom dithering (TPVDD) to embed text in MPEG videos. Secret data was embedded in the I-frame micro-blocks with maximum scene change and P and B frames macro-blocks based on motion vectors with large magnitude. This scheme provides high capacity and imperceptibility. The results show that it can be applied to compress videos without any noticeable degradation in visual quality.

Balaji and Naveen [5] proposed a highly secure method by utilizing LSB technique for hiding data in a video. A frame of cover video was chosen to create an index for the secret data, and this index frame was used for secret data location while extraction. This was very simple method with high security and was also efficient in terms of calculation time.

Hu [6] presented a data hiding scheme to hide secret video in an uncompressed video of approximately same size. Embedding was done by partitioning the frames into non-uniform rectangular codes with 4 bits LSB. In non-uniform rectangular partition, three main factors were used; they are the initial partition, the bit-variate polynomial, and the suitable error control. Obtained results showed that the proposed algorithm was highly secure with high embedding capacity and without distortion with average PSNR above 28 dB.

Paul et al. [7] utilized a scene change detection algorithm for embedding data in a video. This algorithm detected abrupt changes in sequence of frames with histogram difference by checking each frame of the video, and embedding was done with 3-3-2 LSB substitution. They utilized a random sequence generator known as indexed-based chaotic sequence which helped in improving the security. The experimental results were evaluated through histogram with no perceptual difference between the cover video and the stego video.

Moon and Raut [8] proposed a steganography algorithm for an uncompressed AVI video file and improved the payload by utilizing 4LSB method for hiding secret text and image file. The security of this method was improved by using computer forensics as an authentication tool, and the security analysis was done with the help of histogram. This algorithm compared results with 1LSB and 2LSB methods and demonstrated that the proposed algorithm has four times more capacity (50%) than 1LSB. This algorithm provided three layers of security using encryption and computer forensics.

Gupta and Chaturvedi [9] proposed a method for embedding secret data in AVI video format by using 1, 2, or 3 LSB replacement technique. Before embedding, the video was converted to 20 grayscale images and after that, embedding was done using LSB replacement technique. They calculated correlation coefficient for each 1 bit, 2 bits, and 3 bits LSB embedding and the results demonstrated that correlation factor decreased with the increase in replacement of bits.

Yadav et al. [10] presented a novel video data hiding scheme with high capacity and security. This scheme divided cover videos frame into its individual components, and those components were converted into binary patterns of 8-bit. Encryption of each frame of the secret video was done with a secret key using XOR function, and encrypted frames were hidden in the cover video with the help of LSB sequential encoding using a pattern BGRRGBGR. The proposed scheme possesses high embedding capacity, and average PSNR was above 34 dB with no visual distortion.

Chaudhary [11] proposed a multiphase model with three steps where frame selection was done in first step using entropy method based on content level analysis. Using frequency-based analysis, adaptive mask identification was done in second step and in final step, LSB embedding was done over the mask. The results

of this scheme demonstrated that average PSNR was above 50 dB which resulted in less distortion.

Richa [12] proposed a method to embed text into AVI format videos using back propagation neural network for video steganography. Neural network was trained to perform XOR operation and select the symmetric key and bit pattern. In this method, predefined bit position was used for hiding the text and evaluation of the result was done using histogram.

Kelash et al. [13] utilized color histograms of AVI video formats to propose video steganography algorithm. This algorithm hides secret data in frames randomly selected using a threshold with the help of histogram constant value. The video frame pixels were divided into two parts where the embedding of bits was in the right part and number of bits was counted in the left part of the pixels. Results of the proposed scheme demonstrated that the algorithm was simple, secure with very good embedding capacity and PSNR above 47 dB.

Mstafa and Elleithy [14] presented a secure video steganography scheme using hamming code (7, 4) based on linear block code principle. The secret image was hidden by changing the positions of Y, U, and V component pixels. This method used three keys: one for the reposition of Y, U, V component pixels, and other two keys were used for embedding. To provide more security, the secret image was encoded with hamming code (7, 4) and was XORed with random values generated by using a key. This scheme was having high embedding efficiency with payload 16 Kbits that can be extended up to 90 Kbits in a single frame without noticeable distortion and with average PSNR above 51 dB.

Athira et al. [15] proposed a scheme to hide secret data in audio and image part of a video file. Digital signature was used for encryption of the data to be hidden. For image steganography, the proposed method used 4LSB technique, and for audio steganography, phase coding algorithm was used.

Kapoor and Mirza [16] proposed a method based on pixel extraction of RGB model, and LSB insertion was used to insert the data into video frames. In this secret, data was embedded after applying some checks in terms of capacity and imperceptibility. The secret text was embedded using LSB insertion to the MPEG video formats as a cover file. The results of this scheme demonstrated that it achieved better imperceptibility in terms of MSE and PSNR with average PSNR above 52 dB. This scheme could be used for other video formats also such as AVI with little modification.

Singh and Singh [17] utilized grayscale videos for data embedding using LSB substitution technique. They used background subtraction technique to detect the pixels for foreground, and embedding was done only in the foreground pixels to provide it a natural randomness. A major drawback of this method was that after extraction of the data cover, video was not recovered, and this method is applicable only to the grayscale videos.

Ramalingam and Isa [18] presented a video steganography technique for hiding text and images in the RGB components. This method used random key for encryption and decryption to provide more security, and embedding was done in RGBBGRRG order of the video frames. The extraction of the hidden data was done

with no error with PSNR above 62 dB. This method possessed high embedding capacity as it can hide different size data in the video without any change in the size of cover video.

Mstafa and Elleithy [19] utilized KLT tracking algorithm and hamming codes to purpose a video steganography algorithm. This scheme used two pre-processing techniques, hamming code for secret data and other was for face detection and tracking in cover videos. They utilized spatial domain LSB technique for embedding secret data in the covert video. This method was tested against different noise attacks and achieved average value of PSNR = 36.86 dB, PSNRH = 50.29 dB, and PSNRM = 55.56 dB which could be improved further.

Table 67.1 Comparison of spatial domain techniques

Authors	Steganography method	Carrier format	Additional key	Message supported	PSNR
Hanafy et al. [2]	LSB	AVI	Yes	Audio, image, and text	50
Bhattacharya et al. [3]	LSB replacement	AVI	No	Characters	–
Sherly and Amritha [4]	TPVD	MPEG	Yes	Text	40
Balaji and Naveen [5]	LSB replacement	AVI	No	Binary text	–
Hu [6]	4 LSB	AVI	No	Video	28
Paul et al. [7]	3-3-2 LSB	WMV, MPG, AVI	No	Text	–
Moon and Raut [8]	4 LSB	AVI	Yes	Text, image	–
Gupta and Chaturvedi [9]	LSB replacement	AVI (grayscale)	No	Text and binary images.	64
Yadav et al. [10]	LSB sequential	AVI	Yes	Video	34
Chaudhary [11]	LSB	AVI	No	Image	50
Richa [12]	LSB	AVI	Yes	Text	64
Kelash et al. [13]	Histogram	AVI	No	Text	47
Mstafa and Elleithy [14]	Y, U, V components	YUV	Yes	Binary logo image	51
Athira et al. [15]	4LSB	AVI	Yes	Text	–
Kapoor and Mirza [16]	LSB insertion	MPEG	–	Text	52
Singh and Singh [17]	LSB substitution	Grayscale	No	Image	54
Ramalingam and Isa [18]	RGB components	BMP (Images)	Yes	Text and image	62
Mstafa and Elleithy [19]	LSB	AVI	No	Text	36
Sudeepa et al. [20]	LSB insertion	AVI	Yes	Text	–

Sudeepa et al. [20] utilized randomization and parallelization for efficient video steganography algorithm using LSB insertion. The encryption of secret text was done using symmetric key, and the frames were chosen randomly with feedback shift register (FSR) for embedding the text. They used four threads for parallelization process, and the encryption/embedding and decryption/extraction processes were executed in parallel. The throughput of the proposed scheme increased as the size of input increased.

The comparative analysis of the above literature is done in Table 67.1 as shown below. This analysis is according to the literature presented by the corresponding researchers, and the PSNR value used here is the average of PSNR values in dB given in the literature.

Conclusion

The paper tried to present the literature survey of video steganography based on spatial domain techniques. It is seen that the spatial domain methods are very simple and easy to implement with high embedding capacity. The methods used for steganography are advancing day by day due to the increased use of the Internet over the past. Numerous techniques exist, and more updated and developed ways of detecting hidden data also advance quickly. So to avoid detection of secret message, methods must be secure enough and need to be combined with other security systems such as encryption techniques of cryptography.

References

1. T. Kellen, "Hiding in plain view: Could steganography be a terrorist tool," SANS Inst. InfoSec Read. Room, (2001).
2. Hanafy, Amr A., Gouda I. Salama, and Yahya Z. Mohasseb. "A secure covert communication model based on video steganography." MILCOM 2008-2008 IEEE Military Communications Conference. IEEE, pp. 1–6, (2008).
3. Bhattacharyya, Debnath, et al. "Directed graph pattern synthesis in LSB technique on video steganography." Advances in Computer Science and Information Technology. 61–69. Springer Berlin Heidelberg, (2010).
4. Sherly, A. P., and P. P. Amritha. "A compressed video steganography using TPVD." International Journal of Database Management Systems (IJDMS) Vol 2 (2010): 764–766.
5. Balaji, R., & Naveen, G. (2011, May). Secure data transmission using video Steganography. In Electro/Information Technology (EIT), IEEE International Conference on (pp. 1–5). IEEE. (2011).
6. Hu, Sheng Dun. "A novel video steganography based on non-uniform rectangular partition." In Computational Science and Engineering (CSE), 2011 IEEE 14th International Conference on, pp. 57–61. IEEE, (2011).
7. Paul, Rahul, Anuja Kumar Acharya, Virendra Kumar Yadav, and Saumya Batham. "Hiding large amount of data using a new approach of video steganography." In Confluence 2013:

The Next Generation Information Technology Summit (4th International Conference), pp. 337–343. IET, (2013).
8. Moon, Sunil K., and Rajeshree D. Raut. "Analysis of secured video steganography using computer forensics technique for enhance data security." In Image Information Processing (ICIIP), 2013 IEEE Second International Conference on, pp. 660–665. IEEE, (2013).
9. Gupta, Hemant, and Dr Setu Chaturvedi. "Video Data Hiding Through LSB Substitution Technique." Research Inventry: International Journal Of Engineering And Science 2, no. 10, pp. 32–39, (2013).
10. Yadav, Pooja, Nishchol Mishra, and Sanjeev Sharma. "A secure video steganography with encryption based on LSB technique." In Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, pp. 1–5. IEEE, (2013).
11. Chaudhary, Jyoti. "A Multi Phase Model to Improve Video Steganography." In Computational Intelligence and Communication Networks (CICN), 2014 International Conference on, pp. 725–729. IEEE, (2014).
12. Richa, K. "Video Steganography by LSB Technique using Neural Network." In Sixth International Conference. (2014).
13. Kelash, Hamdy M., Osama F. Abdel Wahab, Osama A. Elshakankiry, and Hala S. El-Sayed. "Utilization of steganographic techniques in video sequences." International Journal of Computing and Network Technology, vol 24, no 1, pp. 17–24, (2014).
14. Mstafa, Ramadhan J., and Khaled M. Elleithy. "A highly secure video steganography using Hamming code (7, 4)." In Systems, Applications and Technology Conference (LISAT), 2014 IEEE Long Island, pp. 1–6. IEEE, (2014).
15. Athira Mohanan, Reshma Remanan, Sasidhar Babu Suvanam, and N. V. Kalyankar. "AUDIO-VIDEO STEGANOGRAPHY USING FORENSIC TECHNIQUEFOR DATA SECURITY." Journal Impact Factor 5, no. 12, pp. 154–157, (2014).
16. Kapoor, Vivek, and Akbar Mirza. "An Enhanced LSB based Video Steganographic System for Secure and Efficient Data Transmission." International Journal of Computer Applications 121.10 (2015).
17. Singh, Dalwinder, and Birmohan Singh. "Data hiding in videos using Background Subtraction." In 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS), pp. 1–5. IEEE, (2015).
18. Ramalingam, Mritha, and Nor Ashidi Mat Isa. "A steganography approach over video images to improve security." Indian Journal of Science and Technology 8.1, pp. 79–86, (2015).
19. Mstafa, Ramadhan J., and Khaled M. Elleithy. "A video steganography algorithm based on Kanade-Lucas-Tomasi tracking algorithm and error correcting codes." Multimedia Tools and Applications, pp. 1–23 (2015).
20. Sudeepa, K. B., et al. "A New Approach for Video Steganography Based on Randomization and Parallelization." Procedia Computer Science 78, pp. 483–490, (2016).

Chapter 68
Trust-Based Routing for Mitigating Grayhole Attack in MANET

Kefayat Ullah and Prodipto Das

Abstract MANET is a temporary network without having any pre-existing network setup. In this network, each node participating in the network act both as host and router. Due to the dynamicity and mobility of the network, connections in the network change frequently. Since each participating node in the MANETs is mobile and distributive and also the nodes are resource restricted, traditional networks security systems cannot be applied directly. Also, due to MANET various features like mobility, dynamicity, distributive, cooperative nature, MANETs are vulnerable to various attacks like blackhole, wormhole, grayhole, flooding. This paper explores trust and security challenges to improve security level for MANET. In this study, a new dynamic source routing protocol for MANET based on trust model is developed to mitigate grayhole attack. Trust is calculated based on trust function. Nodes are selected based on the values of trust function and threshold value. The modified DSR protocol can effectively detect the grayhole nodes and isolates them from taking part on routing.

Keywords MANET · DSR · Trust · Grayhole · RREQ · RREP

Introduction

MANET [1] is a temporary network created and operated by the nodes themselves without having any centralized infrastructure. Nodes cooperate with each other by passing data and packets from one hop to another. Every nodes act as a router itself for communication. MANET is a multihop wireless network with dynamic topology, low battery power, and limited bandwidth. Due to dynamic nature of nodes, topology changes frequently at any time. So it is very difficult to find an optimal

K. Ullah (✉) · P. Das
Department of Computer Science, Assam University, Silchar, Assam, India
e-mail: kefayat86@gmail.com

P. Das
e-mail: prodiptodas@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_68

route for communication. The routing algorithm must act quickly as the topology changes. Different protocols [2] have been developed for such type of network. Since the network is dynamic in nature, the topology changes frequently and the network is open to attack and unreliability. Nodes misbehavior due to malicious intention could significantly degrade the performance of MANET. In MANET, the nodes act both as host and router and act as a communicating device in a wireless environment, so trust is a major factor in MANET because ad hoc network communicates depending upon the cooperative and trusting nature of nodes. Since the nodes change frequently, so the number of nodes participating in the route selection is always changing; thus, the degree of trust also changes. Trust is a relationship between two nodes. Trust value can be stated as the degree that one node expects another node to offer certain services. Trust can be calculated directly or by recommendation by other nodes. Grayhole attack [3], is unlike blackhole attack [4], is an active attack type which drops packets. In grayhole attack, the attacker node accepts the packets for forwarding, but without doing so, it just drops it. In grayhole attack, the malicious node initially behaves correctly and replies true RREP messages to that forwards RREQ messages. This way it takes the packets and later just drops the packets. The sender nodes thus loose the connection and again try to establish a new route, broadcasting RREQ messages. Attacker node again do the same things thus consumes battery power and other network resources. This study reveals a new dynamic source routing protocol for MANET based on trust model to mitigate grayhole attack. Trust is calculated based on trust function. Nodes are selected based on the values of trust function and threshold value. The modified DSR protocol can effectively detect the grayhole nodes and isolates them from taking part on routing. This paper tries to mitigate the grayhole attack using trust-based routing. In this paper, the original DSR protocol has been modified along with trust estimation model. The communication between the nodes in the MANET is based on the trust value of a node to its neighbors. Trust value is calculated based on the experiences of node. Grayhole attack, a type of blackhole attack [5], is an active attack type which drops packets. In grayhole attack, the attacker node accepts the packets for forwarding, but without doing so, it just drops it. In such attack, the malicious node initially behaves correctly and replies true RREP messages to nodes that forward RREQ messages. This way it takes the packets and later just drops the packets. The sender nodes thus loose the connection and again try to establish a new route, broadcasting RREQ messages. Attacker node again does the same things thus consumes battery power and other network resources. Trust is a critical factor based on uncertainty conditions and is used for decision-making on cooperating with unknown participants [6]. It includes establishment and updating of trust. In [7] the authors state that trust may not be transitive, symmetric. That means if node A trust node B, and node B trust node C that does not necessarily imply A trust C. Also, symmetric means if node A trust node B that does not guarantee that B will trust A. In [8] authors elaborate logic and reputation based trust. Logical policy-based mechanism takes binary decision depending upon which a node is certified as trusted or not. In [9], the authors proposed an aggregate signature algorithm for grayhole nodes. In that mechanism,

aggregate signature algorithm, DSR protocol, and network model were incorporated. Aggregate signature algorithm was used to trace grayhole nodes. In this paper, we present a trust-based DSR (TBDSR) to mitigate packet dropping that is grayhole attack. In the process, first, the nodes trust value is calculated using the trust calculating function and then, the DSR RREQ header and RREP header were updated. Based on trust value and threshold value, the optimal route has been selected for routing. Finally, the performance was measured based on various performance parameters. The remainder of this paper is structured like this. In section "Proposed Trust-Based Model (Trust-Based DSR)," the proposed trust model is discussed including the algorithm. In section "Simulation Environment," the experimental setup was presented. In section "Results and Discussions," the result and discussion were discussed. Finally, the paper is wind up with conclusion and future work in section "Conclusion and Future Work".

Proposed Trust-Based Model (Trust-Based DSR)

The proposed trust-based DSR (TBDSR) improves the DSR protocol [10] fortified with trust-based route selection. In this model, the trust values will be adjusted based on the experiences that the node has with its neighbor nodes. In the proposed work, the DSR routing protocol is embedded along with the trust estimation function. The communication among the nodes in the MANET depends on the coordination and the trust value with its neighbors (Fig. 68.1).

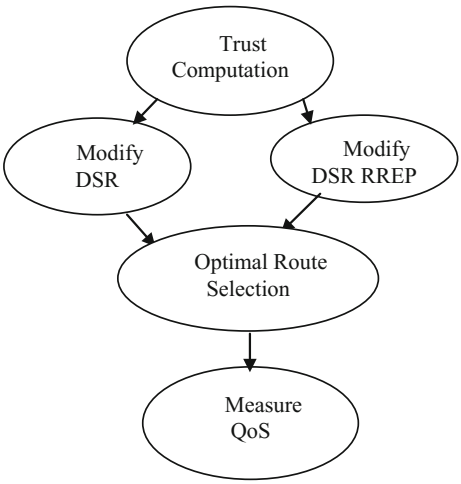


Fig. 68.1 Proposed trust model

Table 68.1 Modified DRS RREQ header for grayhole

IP header	DSR header	DSR RREQ header	Intermediate addresses Address 1, Address 2, …, Address n	Trust values

I. Trust Computation

We propose a very simple equation for the calculation of trust value.

$$V = tanh\,(r1 + r2 + A) \tag{68.1}$$

where V = trust value

$$r1 = \frac{\text{number of packets sent}}{\text{number of packets to be sent}}$$

$$r2 = \frac{\text{total no. of packets received from a node but not originated from that node}}{\text{total number of packets received from it}}$$

A = acknowledgement bit (0 or 1)

Here, we use hyperbolic tangent function [11] to limit the computational trust value within the range −1 and +1.

II. Maintaining Trust Information

In this scheme, the trust of each node is calculated locally. To store the trust information, the DSR's route request (RREQ) and route reply (RREP) headers are modified to attach the trust values of nodes. The modified RREQ and RREP header of DSR routing protocol are shown in the Table 68.1 and Table 68.2, respectively.

III. Decision Making using Threshold Value

To make a security decision with the computed trust value, we need to approximate the level of risk that can be affordable by each task. In other words, a threshold trust value has to be fixed for each task. Such threshold value may be varied depending on the security requirement of each ongoing task, for example, a very important message may have threshold value as 0.8,

Table 68.2 Modified DRS RREP header for grayhole

IP header reply	DSR header	DSR RREP header	Addresses source Address 1, Address 2, …, Address destination
Reply trust score values between source node and destination node	DSR source node route header	DSR source route Address 1 …, Address N	DSR source route trust values

whereas a less important one may have threshold value as 0.2. By comparing the computed trust value and the threshold trust value, it is easy to see whether the trustor node can trust the trustee. A simple equation for making decision is defined as follows:

$$D = V - \textit{threshold value} \tag{68.2}$$

If $D \geq \alpha$, it means the computed trust value satisfies the trust requirement of the ongoing task and is trusted as per the degree of trust. If $D < \alpha$, it means that the trust requirement is not satisfied.

IV. **Algorithm for Routing**

Notations:
SN: Source Node
IN: Intermediate Node
DN: Destination Node
Step 1: SN broadcasts RREQ
Step 2: SN receives RREP {
IF (RREP is from trusted node){
Route data packets (Secure Route) }
ELSE RREP is from Unknown node
Insecure Route
IN can be a grayhole node
} While (IN is NOT a reliable node)

Simulation Environment

The proposed model is implemented using network simulator 2. Here, the original DSR protocol is modified with calculated trust value. The DSR's route request (RREQ) and route reply (RREP) headers are modified to attach the trust values of nodes. Here, calculated trust is used to identify and handle grayhole attack in the routing process which drops the packets. For simulation, random waypoint model is used, that is, in random waypoint, any node can start at random position, waits for pause time, and then moves any other random position. The simulation parameters used are tabulated (Table 68.3).

Table 68.3 Simulation parameters

Parameter	Value
Simulator	NS2.34
Simulation duration	800 s
Number of nodes	100
Transmission range	250 m
Topology	1000 m × 1000 m
Routing protocol	DSR
Network traffic	CBR
No. of malicious node	9
Packet size	512

Results and Discussions

In this simulation experiment, we have used various parameters for performance analysis of the trust-based DSR protocol. Here, we have calculated packet delivery ratio, throughput of the network, routing overhead, and dropped data packets. The experimental values and their respective graphical representation are given below. In Fig. 68.2, we have calculated the packet delivery ratio of both DSR and TBDSR for measuring the performance of trust-based DSR. It is been observed that when there is no grayhole malicious node, both DSR's and TBDSR's packet delivery ratio is more than 90%, but as the number of malicious node increases the PDR decreases. It is estimated that with approximately 20% of grayhole node, the PDR for TBDSR is approximately 70% which is a 10% improvement over DSR. In Fig. 68.3, packet dropped by grayhole node is calculated. Here also, when there is no grayhole node, packets dropped were very few, but as the malicious node increases, percentage of packet drop is very high in DSR as compared to TBDSR. The simulation result shows that the proposed TBDSR can prevent packet dropping at a great since the routing takes place via secured and trusted node. In Fig. 68.4,

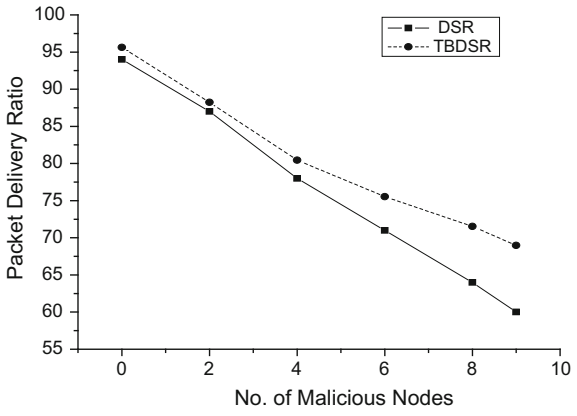


Fig. 68.2 Packet delivery ratio

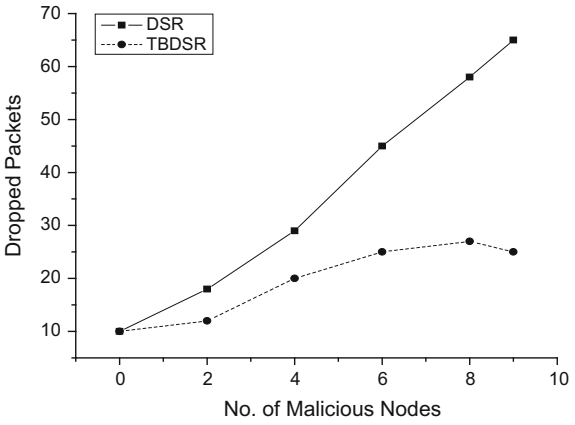


Fig. 68.3 Packet dropped

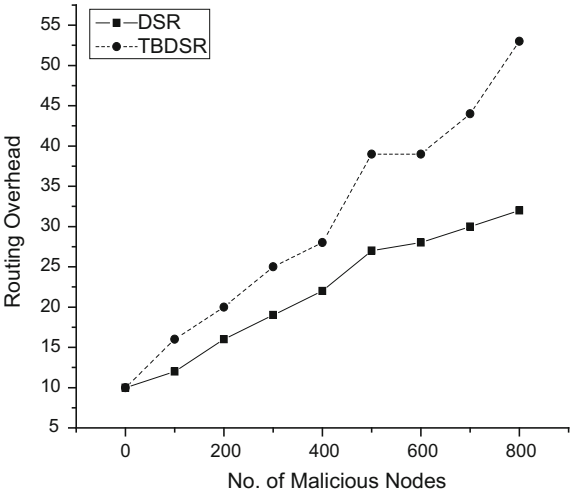


Fig. 68.4 Routing overhead

routing overhead is measured for both DSR and TBDSR. When grayhole node is present in the network, all the attacking nodes participate in the routing process and forward some packets. Since in TBDSR, routing is done through trusted node, routing overhead is higher than DSR. Finally, in Fig. 68.5, the simulation result shows the throughput of the DSR and TBDSR protocol in the presence of the grayhole malicious node. The simulation result shows that when there is no malicious node, the throughput was more than 90% both for DSR and TBDSR. But at 20% of grayhole node, the throughput for TBDSR is 80% while for DSR it is 60%. When grayhole node is 60%, the throughput for DSR is only 25% and for TBDSR is 62%, which is 37% improvement over DSR.

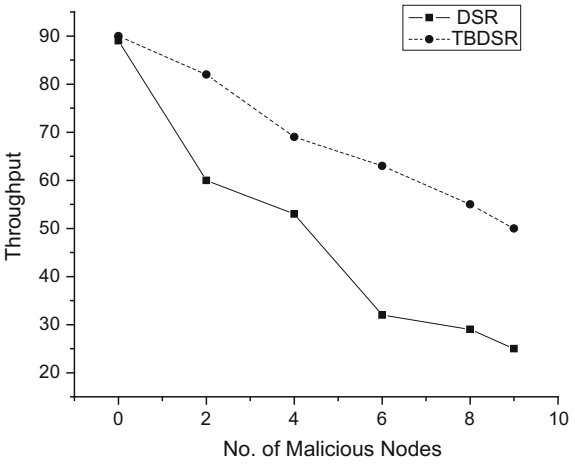


Fig. 68.5 Throughput

Conclusion and Future Work

This paper discussed about a security solution of grayhole attack in mobile environment. The performance of the proposed technique is evaluated with respect to routing overhead, PDR, dropping nodes, and throughput. The performance matrix shows a reasonable outcome. In future, more improved trust-based security protocol may be designed to secure the network from various kind of attack and same can be implemented for other protocol also. Also, the work can be tested using test bed.

References

1. Mishra and Nadkarni: Security in Wireless Ad-hoc Network, The Hand book of Ad Hoc Wireless Networks, CRC press LLC (2003)
2. Elizabeth M. Royer: A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks, University of California, Santa Barbara Chai-Keong Toh, Georgia Institute of Technology, IEEE personal communication, April (1999)
3. Xiao, Y., Shen, X. and D. -Z. Du(Eds.): A Survey on Attacks and Countermeasures in Mobile Ad Hoc Networks, Wireless/Mobile Network Security, Springer (2006)
4. Mohammad llyas and Richard Dorf., C.: The handbook of ad hoc wireless networks (2003)
5. Bhalaji, N, Shanmugam, A: Association between nodes to Combat Blackhole attack in DSR based MANET, in Proceedings of the Sixth IFIP-IEEE WOCN, April 28–30, Cairo, 403–407, ISBN: 978-1-4244-4704-, https://doi.org/10.1109/wocn.2009.5010579 (2009)
6. Solhaug, Elgesem, D., and Stolen, K.: Why Trust is not proportional to Risk?, Proc. 2nd Int'l Conf. on Availability, Reliability, and Security (ARES'07), 10–13 April, Vienna, Austria, 11–18 (2007)
7. Golbeck, J: Computing with Trust: Definition, Properties, and Algorithms, Securecomm and Workshops-Security and Privacy for Emerging Areas Communications Networks, Baltimore, MD, 28 Aug.–1 Sep., 1–7 (2006)

8. Yunfang, F: Adaptive Trust Management in MANETs, Proc. International conference on Computational Intelligence and Security, Harbin, China, Dec., 804–808 (2006)
9. Xiaopeng, G and Chen Wei: A Novel Gray Hole Attack Detection Scheme for Mobile Ad-Hoc Networks, IFIP International Conference on Network and Parallel Computing—Workshops, IEEE (2007)
10. Johnson, D. and Maltz, D.: Dynamic Source Routing in ad hoc wireless Networks, in: I. Tomasz, K. Hank (Eds.), Mobile Computing, first ed., Kluwer Academic Press, 153–181 (1996)
11. Kannan and Mohapatra: Trust Computations and Trust Dynamics in Mobile Adhoc Networks: A Survey, IEEE 14(2), 279–298 (2012)

Part VIII
Data Mining and it's Applications

Chapter 69
MaDGM: Multi-aspect Dynamic Graph Miner

Hardeo Kumar Thakur, Anand Gupta, Rahul Khanna and Sakshi

Abstract Dynamic graphs model time-varying interactions between related entities in a network. Extensive studies have been carried out on the mining of frequent, regular, and periodic behavior of such interactions. Some of the previous research focused on providing users the platform to mine periodic patterns on a single aspect (structure, weight, or direction) at a time. However, the designed tool needs to be run multiple times, to mine significant information that is gained by the study of multiple aspects simultaneously in a network. Moreover, it lacks capability of mining the regular patterns in a network, and the applicability of which lies in wide-ranging domains. In the present research, a tool, the Multi-aspect Dynamic Graph Miner, is proposed that fills up these gaps by providing users an integrated platform for mining regular and periodic patterns on multiple aspects. Besides the primary features, it facilitates easy visualization of tool output and provides a converter for the users of previous works that some portion of our software is based on. We further discuss its applicability by testing on real-world and synthetic datasets.

Keywords Multi-aspect · Data mining · Periodic pattern · Regular pattern
Dynamic graph · Weighted directed network · Repetition rule
Pattern miner

H. K. Thakur (✉) · A. Gupta · R. Khanna · Sakshi
Division of Computer Engineering, Netaji Subhas Institute of Technology, Delhi, India
e-mail: hardeokumar@gmail.com

A. Gupta
e-mail: omaranand@gmail.com

R. Khanna
e-mail: khannar1995@gmail.com

Sakshi
e-mail: sakshi_nsit@yahoo.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_69

Introduction

A dynamic graph is a means to represent time-varying interactions between related entities in a network. A vertex in such a graph represents an entity, the corresponding edge between any two entities denotes a relationship and the weight associated with each edge gives the strength of the corresponding relationship. The interaction between these entities being dynamic varies with time. The mining of these dynamic interaction patterns in such a given network is widely known as dynamic graph mining.

These interaction patterns have been divided previously into three major categories, which are—*frequent patterns* that occur frequently in graph data, *periodic patterns* which occur after a fixed interval of time with temporal periodicity, and *regular patterns* occurring repeatedly in a sequence, representing certain characteristic or behavior. Frequent pattern mining first put forth by Agrawal et al. [1] has been followed by significant contributions of many other researchers [2, 3]. Noteworthy research has been carried out in the fields of periodic [4, 5] and regular pattern mining [6, 7]. All types of pattern mining have been based broadly, on three major aspects of interactions in a time-varying dynamic network: (i) topology or structure, (ii) weight, and (iii) direction of the interactions.

Although a lot of work has been done in the field of data mining, none of it has taken into account the relation between various categories of interactions. A set of regular patterns is itself a subset of a set of periodic patterns. This can be illustrated as follows: if a pattern *abcdabcdabcd* occurs in an interaction, *abcd* can be seen as a regular pattern occurring in three consecutive intervals or as a periodic pattern with periodicity 4; but it also exhibits periodicity with respect to each of *a, b, c, d* (period 4). Hence, we might find in the real-world situations where a user requires to simultaneously study the periodicity and regularity in a network as illustrated with a real-world example.

A Web site owner wishes to analyze the response of customers to a set of Webpages that he creates. If a network is built out of this dataset, any customer and Webpage would form a node each and the presence of an edge between a customer and Webpage means that the customer has accessed that Webpage. The owner now wishes to analyze the results of mining these interaction patterns. Not only does he need the periodicity of occurrence of edges but also its regularity. There is no one platform for such a user to study both of these patterns simultaneously so as to save time, effort, and resources.

There are several tools in the field of data mining, like *sequential pattern mining using bitmaps* [8], *periodic subgraph embedding miner* [4] to name a few. To the best of our knowledge, there exists no tool that integrates mining of periodic and regular patterns. To fill up this gap, a tool, *Multi-aspect Dynamic Graph Miner (MaDGM)*, has been proposed in the present research to facilitate the user with an interactive platform for mining regular and periodic patterns in a network.

Algorithmic Background

Periodic Pattern Miner

Any network modeled from the real world contains patterns that recur with a natural periodicity [4]. Such patterns exhibit predictable behavior because of their property of repeating periodically. The study of these patterns called periodic pattern mining is a significant part of any network's analysis.

Gupta et al. [5] have proposed the *periodic pattern on structure and weight framework (PPSWF)* based on the PSEMiner [4] to efficiently mine periodic pattern on structure (PPS) and periodic pattern on weight (PPW) together in a weighted dynamic network. It can be generalized to mine any two related aspects at a time—*primary and secondary aspects* [9]. We use the existing PPSW framework as an instance that structure being the primary aspect and weight being the secondary aspect. In the *PPSWF*, snapshots of a weighted dynamic network are taken and the edges having occurrence less than threshold are removed as infrequent edges. Rest of the edges and edge-weight combinations are assigned a unique number each and stored in a structural input table and a unique numbered edge map, respectively. *Edited PSEMiner (EPSEM)* is applied to the structural input table to identify *structural inverse periodic patterns (SIPP)*. The unique numbers that occur in *SIPP* are mapped to their corresponding vertices to build a *periodic pattern on structure (PPS)*. Data regarding the edges present in structural edge lists are selected from the unique numbered edge map and are input to the *EPSEM* to find *weighted inverse periodic patterns (WIPP)*. Mapping unique numbers to edge-weight combinations and edges to their vertices results in *periodic patterns on weight (PPW)*.

We leverage this algorithm for *periodic pattern miner (PPM)* to mine periodic patterns on any two aspects at a time. The primary and secondary aspects for a directed network are structure *(PPS)* and direction *(periodic pattern on direction, PPD)* and that for a weighted directed network are weight *(PPW)* and direction *(PPD)* [9].

Regular Pattern Miner

Apart from periodic patterns, there may exist repeated patterns in a dynamic network that represents certain characteristic or behavior called regular patterns. Regular pattern mining has found a host of potential applications like weather forecasting, earthquake prediction.

We leverage the novel method put forth by Gupta et al. [7] focusing on mining regular patterns on weight and direction in a directed weighted dynamic network, as the second component of MaDGM called the *regular pattern miner (RPM)*. In this method, snapshots of the dynamic weighted directed network are taken, and from it, a sequence is obtained over time corresponding to occurrence, weight, direction,

and weight-direction. From this sequence, infrequent edges are removed and substring of maximum length that repeats consecutively at least threshold number of times along with its starting position, results in a repetition rule. Edges with same rule are grouped to form evolution patterns [7].

Descriptive Analysis of MaDGM

Interaction patterns can be mined for all three aspects using RPM and any two using PPMs, thus giving rise to the name '*Multi-aspect Dynamic Graph Miner.*' Following subsections describe how MaDGM integrates its components for the exploratory analysis of real-world dynamic graphs. Both the miner and user interface have been implemented in JAVA.

Periodic Pattern Miner

This component (b in Fig. 69.1) is used for mining periodic patterns. *Tree datastructure* has been used in its implementation, in accordance with the leveraged algorithm. The input to PPM is a text file having the number of timesteps, '*n,*' on which snapshots of the dynamic network are taken, in its first line. The lines that follow contain entries of the form $< v_1\, v_2\, w_1\, w_2 \ldots w_n >$, where v_1 and v_2 represent vertices and w_1 through w_n represent the weight on corresponding edge at n timesteps. A weight '0' at any timestep signifies that that edge does not exist on that particular timestep.

The output of the component is also available in a text file. Each line contains an entry of the form $< isp\, num[(v_1, v_2, w_1), (v_1, v_3, w_2), (), \ldots] >$ that represents a subgraph having *num* number of edges ($v_1 - v_2$, $v_1 - v_3$) with weights w_1 and w_2, containing a pattern that occurs for s number of timesteps starting from t_i and repeats after every p timesteps. This forms the *weight-rule file* of the output.

Regular Pattern Miner

This component (a in Fig. 69.1) is used for mining regular patterns. The UI contains multiple columns to be filled by the user—(i) *threshold*, minimum number of times a pattern should recur consecutively to be a regular pattern, (ii) *jitter*, number of miss occurrences allowed(to accommodate imperfect regularity), (iii) *input text file*. Each line of the input file contains entries of the form $<e_k\ v_i\ v_j$ *occurrence-string weight-string direction-string*> where e_k is edge number of the edge between vertices v_i and v_j. Of the strings that follow, *occurrence-string* is mandatory. The

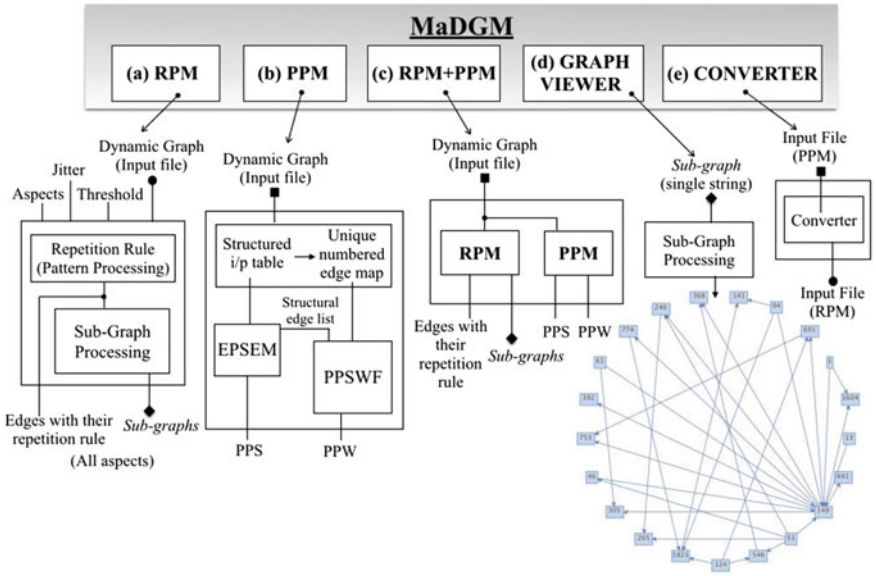


Fig. 69.1 MaDGM structure

length of the strings is total number of timesteps. A '0' on a timestep indicates that the edge does not occur at that timestep.

The output of this component is also available as text files with the following formats. First line of the first file with the repetition rule contains an edge $e_k(v_i, v_j)$, and the next line contains an entry of the form $< v\ i\ s\ c >$ where the repeated pattern 'v' starts at t_i has a length 's' (t_i to t_{i+s}) and occurs c number of times consecutively (c must be greater than threshold). There may exist more than one repetition rules for one edge, for instance, *'abcabcabcabdabdabd'* has two quadruplets <*abc 1 3 3*> and <*abd 10 3 3*>. As a result, many different <*v i s c*> entries exist for each edge. Also, multiple edges may be associated with each <*v i s c*> combination. Corresponding to every unique <*v i s c*>, a list of all edges containing it is constructed, within which some might end up having common vertices. All the edges in the same list sharing common vertices form a subgraph. Any edge that does not share a vertex with any other edge in its <*v i s c*> list is rejected, as a subgraph must have minimum two edges. Many subgraphs are mined from each list and are stored in the second output file, each line of which is in the format <*i s c num [*e_1 (v_1, v_2) e_2 (v_2, v_3), …*]*> where *i, s, c* are same as defined above; *num* is the number of edges; e_1, e_2 are two such edges between vertices $v_1 - v_2$ and $v_2 - v_3$, respectively. It makes use of linked list for the formation of these subgraphs.

RPM and PPM in MaDGM can be run simultaneously, using (c) in Fig. 69.1.

Converter

Since the input formats for the required and generated files of *PPM* and *RPM* are not the same, *MaDGM* enables the user to convert from one input format to another conveniently as per the requirement. This component of the tool aids the user of already existing *PSEMiner* to use the same input file which they would have used in PSEMiner, thereby saving their time and resources. This component is shown as (e) in Fig. 69.1.

Graph Viewer

This component makes MaDGM output easy to read and visualize. The input to this component is a file containing subgraphs with edges (that follow a common repetition rule) in each line. The format is of the type $< e_1(v_1, v_2)\, e_2(v_2, v_3)\, e_3(v_1, v_4) \ldots >$ with e_1, e_2, e_3 representing edges between $v_1 - v_2$, $v_2 - v_3$, and $v_1 - v_4$, respectively. The component along with a sample graphical output is shown as (d) in Fig. 69.1.

System Features and Comparison with PSEMiner

In 2007, Lahiri et al. proposed a new mining problem of finding a minimal set of periodically recurring subgraphs to capture all periodic behavior in a dynamic network. [4] The tool implementing their algorithm came into existence as the *PSEMiner*. However, considering only one of structure or weight of interactions, as proposed by the author, misses some important information. Therefore, it needs to be run twice, to identify both aspects of periodic interactions resulting in a loss of efficiency of the tool. Consequently, Gupta et al. [5] proposed *PPSWF* to efficiently mine on both aspects simultaneously. In this algorithm, only subgraphs having more than one edge are considered, thereby removing redundant PSEs. Besides this, it also provides a better user experience as user does not have to himself assign unique number to edges unlike the PSEMiner. *Clearly, PSEMiner is a subset of PPSWF.*

PPM component of MaDGM leverages *PPSWF* and exploits all of its advantages over *PSEMiner*. Moreover, it comprises of *RPM*, using which users can mine regular patterns too. Thus, *MaDGM* clearly supersedes *PSEMiner*.

We illustrate the applicability of MaDGM with a real-world example—consider a social network that facilitates interaction among its users. This interaction may not be limited to immediate as one might interact with people who are not known to him directly but share similarities with him due to a common bond. Therefore, one

of the key features of a social networking system is the ability to recommend to its users who they can interact with and then let them choose.

We take the case of a hypothetical social network *SN*. It aims to provide recommendation to its users to avoid them from leaving the network, because of lack of communication with mutual connections having similar interests. Communication among users in *SN* occurs by the exchange of messages, the length and frequency of which may vary.

Let users be represented by nodes and the corresponding edge between any two nodes represent the interaction between them. The weight of this edge is a function of the length of the message exchanged, and its direction originates from the sender of the message and terminates at its recipient. For instance, the existence of two nodes *X, Y* and an edge directed *XY* with a weight '*a*' at a particular instance indicates the presence of two users *X* and *Y* with messages of length '*a*' sent from *X* to *Y* at that instance.

Now suppose, another user *Z* signs up on *SN*, is a connection of *X* and *X, Z* also communicate with each other. Should *Z* be recommended as a connection to *Y* and vice versa due to common connection *X*? The answer would depend on the similarity in their communication patterns with *X*. Thus arises the need to study the network and give connection recommendations. In a large network, it becomes a tedious and time-consuming job. In this scenario, MaDGM studies the network and helps to generate necessary recommendations for the users.

First of all, the communication data among users is stored each day for say 70 days and kept in a file in a desired format which is input to MaDGM. Suppose now, in the pattern *'a0aba0aba0ab,'* regular pattern *'a0ab'* with count 4 or periodic pattern *'a'* with period 2 is mined in the same sequence for both edges *XY* and *XZ*. From this, it can be concluded that a great deal of similarity occurs in the communication between users *X, Y* and *X, Z*. This data is output as a subgraph consisting of nodes *X, Y, Z* and directed edges *XY* and *YZ*.

Based on this output, connection recommendations can be sent to all nodes (users) that belong to the same output subgraph. In this way, the owners reap savings in time and resources, which would have to be put in, had they not made use of MaDGM.

Performance

MaDGM execution has been done on Ubuntu 14.04 64 bit virtual machine, having 2 Intel core i7 CPU @ 2.2 GHz processor with 8 GB RAM. *RPM* can be run on any OS(JVM 1.7 or higher) while *PPM* requires a Linux environment to run. Also, RAM requirements for *PPM* are higher than that of *RPM*.

To test the efficacy of the proposed tool, we run it on various real and synthetic datasets. A real-world stock market dataset consisting of opening and closing prices of S&P 500 stocks has been used. A stock represents a node, and an edge denotes that both nodes are profitable. Weight represents the combined profit of the

Table 69.1 Experimental results

S No.	Name	Edges	Timesteps	Time (ms)
1.	Stock market	249220	74	9748
2.	Artificial	122037	100	8884
3.	Random_testcase1	79800	200	28290
4.	Random_testcase2	124750	200	43186
5.	Random_testcase3	124750	300	110323

corresponding nodes. Random test case datasets have been generated by random function for testing. The datasets collected are processed in RPM, and the results in terms of time of successful execution have been recorded in Table 69.1.

For big data, PPSWF is more efficient than PSEMiner in terms of time and space complexity. For detailed performance, comparison between PSEMiner and PPSWF (used for PPM) refers [5].

Conclusions and Future Work

In this paper, we have introduced a tool called the *Multi-aspect Dynamic Graph Miner (MaDGM)*, for mining of regular and periodic patterns. Importance of mining regular and periodic patterns simultaneously, for meaningful deductions from real-world networks, has been justified with the help of examples. *MaDGM* provides the platform to its users for this simultaneous mining, efficiently. For this reason, it would have a host of potential application domains including medicine, social networks etc. has been tested on synthetic and real-world datasets.

While regular pattern mining using *MaDGM* can be done on all aspects of structure, weight, and direction of interactions simultaneously; periodic patterns can be mined only on two of these aspects at a time. Also, the algorithmic background of *PPM (PPSWF)* does not take into account imperfect periodicity. We look forward to versioning the tool with these improvements in the future.

References

1. Agrawal R, Imielinski T, Swami A (1993) 'Mining association rules between sets of items in large databases.' In: Proceedings of the 1993 ACM-SIGMOD international conference on management of data (SIGMOD '93), Washington, DC, pp. 207–216.
2. Han, J., Pei, J., Yin, Y., and Mao, R. (2004). 'Mining frequent patterns without candidate generation: A frequent-pattern tree approach'. Journal of Data Mining and Knowledge Discovery, Vol. 8, pp. 53–87.
3. K.M. Borgwardt, H.P. Kriegel, and P. Wackersreuther (2006) 'Pattern Mining in Frequent Dynamic subgraph.' In Proceedings of the 6th International Conference on Data Mining (ICDM '06), Hong Kong, China, pp. 818–822.

4. M. Lahiri, Y. Tanya and B. Wolf (2008) 'Mining Periodic Behavior in Dynamic Social Networks.' In Proceedings of the 8th IEEE International Conference on Data Mining(ICDM '08), Pisa, Italy, pp. 373–382.
5. A. Gupta, H. K Thakur and A. Garg.' Periodic Pattern Mining in Weighted Dynamic Networks', International Journal of Intelligent Systems Design and Computing (Inderscience) (Accepted).
6. A. Gupta and H.K. Thakur (2013) 'A Novel Method to Determine Regular Pattern in Edge Labelled Dynamic Graph.' In Proceedings of 7th IEEE International Conference on Data mining and warehousing (ICDMW '2013), Bangalore, pp. 39–47.
7. Gupta, A., Thakur, H., and Kishore, P. (2014). 'Mining regular patterns in weighted directed networks', In Proceedings of Thirteenth International Conference Of Information Technology. Bhubaneswar, India, pp. 215–220.
8. J. Ayres, J. Gehrke, T. Yiu, J. Flannick. (2002). 'Sequential pattern mining using a bitmap representation', In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429–35.
9. H. K. Thakur (2016). 'Study of Evolving relation between patterns in Dynamic Graph', PhD thesis Submitted in University of Delhi, India.

Chapter 70
Efficient Prior-Art Retrieval of Patent Documents Using MapReduce Paradigm

K. Girthana and S. Swamynathan

Abstract A patent is a legal right given to novel, non-obvious and useful inventions. The prior-art search involves retrieving prior works related to it to avoid duplication of the invention and granting of the patent. Moreover, it analyzes a variety of documents like newspaper articles, proceedings, and journals. The amount of patent document and the volume of filings keep on increasing at an unprecedented rate every year. Processing on this enormous volume of data sequentially is time-consuming. Hence, the proposed Prior-Art Retrieval System (PARS) retrieves only the patent documents through Google patent API, and K-Means clustering was employed in a parallel mode to cluster the documents. Through Relevance Mapping prominent document clusters were identified. The documents within the relevant clusters are ranked based on the citations. The top ranked documents were displayed to the patent analyst.The results show that the processing time with map reduce has reduced significantly and accuracy of clusters was around 50%.

Keywords Prior-art search · Clustering · MapReduce

Introduction

A patent document is a license, conferring a right or title for a limited period. The patent provides rights to exclude others from making, using and selling it. When there is an innovation, prior-art search becomes much more important. It is undertaken to ascertain whether the invention is novel, non-obvious and useful or not. It uncovers any knowledge that exists before the filing of the current invention. Clustering, an unsupervised machine learning technique, is employed for this prior-art

K. Girthana (✉) · S. Swamynathan
Department of Information Science and Technology,
Anna University, Chennai 600025, India
e-mail: k.girthana@auist.net

S. Swamynathan
e-mail: swamyns@annauniv.edu

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_70

search to group relevant patent documents. It also describes the relationship between the patent documents and finds the intrinsic structure of the patent documents. The emerging increase in the size of patent filings and non-patent articles greatly influence the processing capability of the retrieval system. The patent documents are structured with specific field information while the non-patent literature is unstructured in nature.The non-patent literature includes newspaper articles, conference proceedings, videos, web documents and so on. Each varies in structure and analyzing mechanism. Analysis on huge volume and variety becomes tedious and may not be sufficient if processing is done sequentially; hence, the big data paradigm and related data processing strategies came into existence.

Hadoop, the core platform addresses the challenges of Big data by allowing multiple types of analytical workloads to run on the same data parallelly. The two main components of Hadoop are Hadoop Distributed File System (HDFS) and MapReduce paradigm. The computations are carried out with Map and Reduce functions, which operate as Key-Value pairs.

The remainder sections of the paper are organized as follows: Section "Literature Survey" discusses the prior works in the field of clustering. Sections "Architecture of Prior Art Retrieval System" and "Implementation" briefs the architecture of the proposed model and Implementation details. Section "Results and Discussion" discusses the results obtained through this system and section "Conclusion" concludes the paper with guidelines for future enhancement.

Literature Survey

The prior art is an art of searching the previous works which are related to the current work, which may not exist but might be cited by others. Gaff and Rubinger [1] discussed in detail about the prior-art search process. The various approaches involved in standard Information Retrieval process are explained with different combinations of patent document fields by Wanagiri and Adriani [2]. Their experimental results shows title, description, and claim field combination retrieves better. Xue and Croft [3] proposed a method for automatic generation of the search query from the full patent application. They considered different type of search features along with learning to rank method.

Document Clustering groups similar documents and has been applied widely in various fields. Patent document clustering includes clustering the data taken from various fields of the document like title, abstract and description of the patent document [4, 5]. Huang et al. [6] examine the Chinese patent documents based on the implicit and explicit structure using Self-Organizing Maps. Balabantaray et al. [7] compared K-Means clustering with the K-Medoids and found that K-Means suits well for summarization of documents. The efficiency of K-Means clustering depends on the initial set of seed points [8] and its sensitive to outliers.

Subspace clustering identifies clusters of similar vectors on a subspace of the data space [9]. It is suitable for clustering on high dimensional data, and they

usually follow either density-based or K-Means clustering approach. Though there are many clustering methods [10], currently, no subspace clustering algorithm can handle vast datasets in reduced time. For very high dimensional data such as documents where the features are the words in the document, parallelism will be effective and efficient. Ngazimbi [11] presented a case study of different clustering algorithms like K-means, Greedy Agglomerative, and Expectation Maximization algorithms in Map-reduce framework on Netflix Movie data. Sun et al. [12] use hierarchical clustering technique for the internet users by mining an enormous volume of web access logs. They improved the efficiency of the algorithm by co-occurrence based feature selection and by the batch update process. A PARABLE algorithm [13] for the MapReduce framework, solves the problem of using divide and conquer strategy.

This work relies on Map-Reduce based K-Means clustering [14] because the majority of the clustering algorithms require the dataset to reside in main memory, and the linkage clustering algorithms cannot be applied to mapreduce framework. This work differs from the rest by having an initial number of clusters formed by manually assigned International Patent classification (IPC) code [15]. The clustered documents are ranked to show the priority among the relevant documents. Aleman-Meza et al. [16] measured relevance by considering the semantics of relationships. The major drawback of this approach is ontology derived may negatively affect the semantic annotation and retrieval steps. Here, the authoritative importance or popularity along with similarity score is used as a measure for ranking the documents within the cluster.

Architecture of Prior Art Retrieval System

The prior art retrieval system (PARS) of the patent documents helps to retrieve the relevant patent documents by grouping the documents into a set of clusters. The system flow was outlined in Fig. 70.1. The patent analysts or patentee submits the innovation details in the form of new patent document to the system. The keyword extractor extracts the prominent keywords from the title, abstract portions of the patent document. Using this information, the query formulator formulates the query. This query, in the document searcher phase, helps to retrieve the top K documents from the search system through Google Patent API and stores their contents in the patent database. It also stores citation information retrieved from national patent databases ESPACENET. The patent database is pre-processed and used for further processing.

Hadoop-based patent document processor includes Patent document Indexer and Patent Document Clusteror. The Indexer allows fast and full-text searches at the cost of increased processing. The Patent Document Clusteror makes use of K-Means clustering in the hadoop environment to analyze the relationship between the document clusters. The inputs to the Map-reduce K-Means algorithm are the set of the documents with TF-IDF (Term Frequency and Inverse Document Frequency) scores for each word in the title and abstract sections of the document after preprocessing.

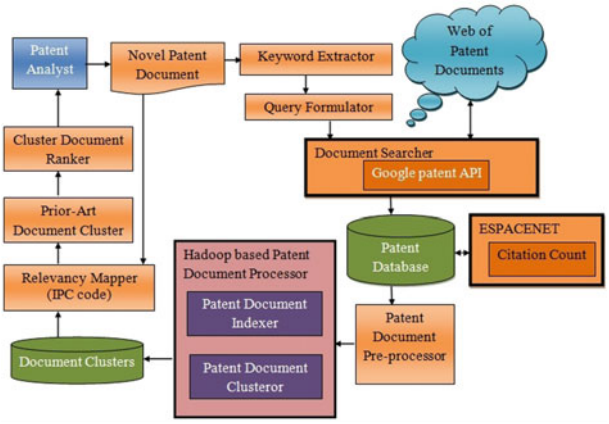


Fig. 70.1 Architecture of prior-art retrieval mechanism

The retrieval system then checks for relevance between the novel patent document and the clusters. Relevant clusters are retained and the documents within the relevant clusters are ranked based on the citation count of the patent document and the similarity score. Preference is given to the most cited patent documents.

Implementation

This section describes the execution environment and in detail about the various phases involved in the prior-art retrieval system.

Environmental Setup

The experiment was conducted by configuring Hadoop in Pseudo-Distributed Mode. The node is configured with 4 GB memory, 64 GB hard disk along with JDK 1.7.0 and Hadoop 2.6.1 version and Ubuntu 14.04 operating system.

Patent Retrieval from Google API

Google APIs are a set of Application Programming Interfaces (API) allow communication with Google Services and their integration with other services. When the patent analysts or patentee submits the query (Metaheuristic Algorithm) to the search system, the system retrieves the first 500 patent documents links through the

Fig. 70.2 Sample patent database

Patent No	Title	Abstract	IPC code	Citation
WO200300(	Robotic systen	In an automated processing syste	B25J9/18, G05	4
WO200807(	Method and a	A methodology is presented to ad	G06N3/00, G0	0
WO200707(	Method and a	The invention pertains to an algor	G06N3/12	1
US9015145	Method and a	A method and system are present	G06F17/30, G(	23
US6263325	System, meth	A system for executing a learning	G06N3/00, G0	37

Google Patent API. Their contents were parsed with an HTML parser (Jsoup). Since the title and the abstract portions of the patent document convey the major portion of the invention details, only those portions of the document were parsed along with the document ID and the IPC code. A snapshot of the patent database was shown in Fig. 70.2. The extracted citation details are also stored in the patent database.

Patent Document Pre-processor

The patent documents are checked for consistency and duplication. The redundant documents are removed and and inconsistent documents are updated with a single latest version of the patent document. These Pre-processing reduces the total number of documents to 350. In addition to them, other pre-processing steps like stemming, and stopword removal are carried out. The documents are represented using a bag of words model.

Patent Document Indexer

The Hadoop patent document Indexer helps to scan the documents quickly through map and reduce phases. The inputs to the indexer map phase are pre-processed patent documents. For each token in the pre-processed document, it computes the term frequency. The indexer map phase splits the document into multiple parts and for each word it encounters, the map phase emits (word, 1). The reducer combines the output of multiple map phases and sums up the count for each word (word, term frequency).

MapReduce K-Means Clusteror

The MapReduce framework splits the data into M splits and processes them in parallel. The clustering algorithm takes input as the set of the documents with TF-IDF values for each word and the set of K initial centroids chosen randomly.The mapper and reducer pseudo code for K-Means are given in Figs. 70.3 and 70.4 respectively. The number of clusters are determined based on the IPC code information of the

Fig. 70.3 Pseudo code of K-Means map function

```
Input : D = Set of documents (d1,d2...dn),
        C=set of centroids (c1, c2,..ck)
Output: (cluster id, Document id)
for all di ϵ D where 1≤i ≤ n do
  MinDis ←∞ ; best_cluster←1
  for all cj ϵ C where 1 ≤ j ≤ k do
      dist ←sim(di, cj)
      if dist < MinDis then
          MinDis ←dist
          best_cluster← cj
      end if
  end for
  display(best_cluster, di )
  i+=1
end for
```

patent document. The map function calculates the cosine similarity score between the input set of documents and each K centroids and produces output in the form of <centroid, document id>. The reducer function takes the output of map function and calculates the new k-centroids.

Results and Discussion

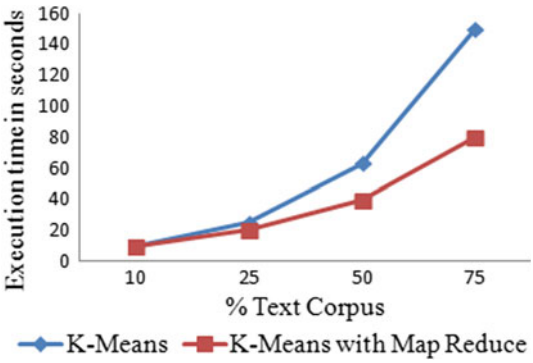
The results analysis was done for the query Meta-heuristic algorithm. Since the majority of the pre-processed patent documents belong to the IPC codes G06N3, G06F17 and G06Q10, the number of clusters was assigned as 3. Patent document clusteror iterates until the cluster centroids remain constant, and the system does a comparative analysis on K-Means clustering with and without Map Reduce. Figure 70.5 depicts the results of K-Means clustering with Map-reduce. Its clear that for small sized data corpus, there is only marginal difference between their execution times while for large corpus K Means clustering with Map-Reduce executes faster than traditional K-Means clustering. Table 70.1a outlines the cluster details along

Fig. 70.4 Pseudo code of K-Means reducer function

- **Input:**
  C_i = Cluster id
  D =List of documents belonging to cluster C_i -<$d_1, d_2, \ldots, d_x$>
- **Output:** < *Cluster id, newcentroid*>
  DocValue ← 0
  no _ docs ← 0

for all d_i ϵ D where $1 \leq i \leq x$ do
  DocValue+ ← DocValue
  no _docs ++
  newcentroid ← DocValue / no _ docs
  display(C_i, newcentroid)
End for

Fig. 70.5 Comparison of K-Means clustering with and without MapReduce



with clustering accuracy. It's known that the precision and recall for the clusters are around 50%. This is because the documents lexicon are unique and the semantics and relations between the lexicons was not considered. Relevance between the document clusters and the novel patent was determined based on both the word similarity and the IPC code description information. The most relevant cluster was chosen for further ranking and processing. Since two-third of the documents belongs to the cluster2 and also because of its closeness to the novel patent, cluster2 is chosen for providing the retrieval results to the user. Another search query "Smart Devices" was also executed on the prior art retrieval system that executes K-Means clustering in the Hadoop framework and the results obtained were similar to the one discussed. This query was also executed under the same environmental setup and 800 documents were retrieved. After pre-processing there were 715 documents and these documents were clustered. The clustering quality and the cluster details were briefed in Table 70.1b.

Table 70.1 Patent document clusteror

Query1: Metaheuristic algorithm			
Clusters	# of Documents	Cluster topic area	Precision
C1 (G0F17)	121	Digital computing and data processing techniques	0.57
C2 (G06N3)	167	Genetic algorithm, stochastic approaches, fuzzy logic	0.58
C3(G06Q10)	62	Optimisation and scheduling Approaches	0.62
Query2: Smart devices			
Clusters	# of Documents	Cluster topic area	Precision
C1 (H04W4)	285	Mobile application and substation equipment	0.57
C2 (G06Q20)	111	Payment architectures and protocols	0.52
C3(G06F3)	232	Image acquisition and digital information	0.45
C4(H04W84)	47	Network topologies	0.623
C5(G06Q30)	40	E-commerce	0.49

The documents within the cluster were ranked based on document citations. The citation information was retrieved from online patent databases like USPTO (United States Patent and Trademark Office) and ESPACENET. If the document citations are same, the retrieval ordering was used for ranking. Since the document citations were given preference, the most popular and widely cited documents are given more weight and they appear at the top of the result section. For performance evaluation, the PARS was compared with Google patent search. The top five relevant documents were retrieved for the query from both PARS and Google patent search along with their ranking and the same was depicted in Table 70.2. The retrieved top 5 documents were totally different because PARS clusters the documents and gives preferences to the citations. The top results of the Google patent search obtained lower rank in PARS because the cohesion between the novel patent and the patent document was less or the relevant terminologies that match the query were meagre or the citations are less.

Table 70.2 Prior-art retrieval system versus google patent search

Google top 5 documents	Google ranking	PARS top 5 documents	Google ranking
US9015145	1	US6360191	43
WO2008079097A1	2	US6055523	22
US9251460	3	US6327582	8
US9305257	4	US6263325	11
US20160140441	5	US20080077544	3

Conclusion

Patents are technical documents that provide exclusive rights to the invention. Prior-art search on the patent is an essential task to uncover the knowledge that exists in a particular field. Since the data for Prior-art retrieval keeps on increasing exponentially, finding insights into the data and invalidating a patent application becomes essential. Clustering on retrieved patent helps to group the patent documents based on the domain. Among several clustering algorithms, K-Means algorithm is straightforward and familiar. This work makes uses of K-means clustering algorithm in MapReduce framework to retrieve efficient and quality clusters in reduced time. It uses patent classification code information for deciding the number of clusters and deciding the relevant cluster. The patent documents within the cluster makes use of citation information for ranking.The results show that the processing time with mapreduce has reduced by almost 25% for larger data sizes and for smaller sized the variation has reduced and also the precision and recall for the clusters was around 55%. This work can be further enhanced with multi-node cluster setup and also by assigning importance to the words in the Patent document abstract and also through semantic analysis [15].

References

1. Gaff, Brian M., and Bruce Rubinger.: The significance of prior art. Computer. 8, pp. 9–11 (2014)
2. Wanagiri, M. Z., Adriani, M.: Prior Art Retrieval Using Various Patent Document Fields Contents. CLEF (Notebook Papers/LABs/Workshops), pp. 1–6, UK (2010)
3. Xue, X., Croft, W. B.: Automatic query generation for patent search. In: 18th ACM conference on Information and knowledge management, pp. 2037–2040, Germany (2009)
4. Jun, S., Park, S. S., Jang, D. S. : Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications. 41, 7, 3204–3212 (2014)
5. Andrews, N. O., and Fox, E. A.: Recent developments in document clustering. Technical Report TR-07-35 (2007)
6. Huang, S. H., Ke, H. R., Yang, W. P.: Structure clustering for Chinese patent documents. Expert Systems with Applications. 34, 4, 2290–2297 (2008)

7. Balabantaray, R. C., Sarma, C., Jha, M.: Document Clustering using K-Means and K-Medoids. International Journal of Knowledge Based Computer Systems. 1, 1 (2015)
8. Bradley, P. S., Fayyad, U. M., Reina, C.: Scaling Clustering algorithms to large databases. In: 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), pp. 9–15. (1998)
9. Kriegel, H. P., Kroger, P., Renz, M., Wurst, S.: A generic framework for efficient subspace clustering of high-dimensional data. In: Proceedings of the 5th IEEE International conference on data mining (ICDM), pp 250–257 (2005)
10. Han, J. and Kamber M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, Elsevier (2011)
11. Ngazimbi, M.: Data Clustering Using MapReduce. In Masters Thesis, Boise State University (2009)
12. Sun, T., Shu, C., Li, F., Yu, H., Ma, L., Fang, Y.: An Efficient Hierarchical Clustering Method for Large Datasets with Map-Reduce. In Proceedings of the International Conference on Parallel and Distributed Computing, Applications and Technologies, 12, 2, pp. 494–499 (2009)
13. Wang, S., Dutta, H.: PARABLE: A PArallel RAndom-partition Based HierarchicaL ClustEring Algorithm for the MapReduce Framework. In: 6th Annual Machine Learning Symposium at the New York Academy of Science. (2011)
14. Zhao, W., Ma, H.,He, Q.: Parallel K-means Clustering Based on MapReduce. In: IEEE International Conference on Cloud Computing, pp. 674–679 (2009)
15. Kang, I. S., Na, S. H., Kim, J., Lee, J. H.: Cluster-based patent retrieval. Information processing & management, 43, 5, 1173–1182 (2007)
16. Aleman-Meza, B., Arpinar, I. B., Nural, M. V., Sheth, A. P.: Ranking documents semantically using ontological relationships. In: 4th International Conference on semantic computing, pp. 299–304, US (2010)

Chapter 71
FCA Based Ontology Revision

Zubin Bhuyan and Shyamanta M. Hazarika

Abstract *Ontology* is a formal description of a conceptualization for a domain of discourse. Revision of an ontology is primarily required because of changes in the domain, its conceptualization, or in the defined specification. This paper proposes an approach for revision of $\mathcal{SHOIN}^D$ ontology, wherein changes are incorporated into the ontology without human intervention. Techniques from Formal Concept Analysis (FCA) form the backbone of the proposed method. The Description Logic (DL) knowledge base is emulated on an extended FCA framework. Thereafter, revisions are incorporated, following which the DL ontology is reconstructed. Throughout the process of revision, the system endeavors to minimize loss of information.

Keywords Ontology · Ontology revision · Formal concept analysis

Introduction

Ontology is a *data engineering* artifact which enables formal conceptualization of any domain of discourse [1]. An assortment of formalisms has been developed to represent ontologies, starting from the frame-based and object-based models, to the more recent Description Logics (DL). DL has emerged as the prime formalisms for ontology representation as it allows one to formally define terms and relations which constitute the vocabulary of a particular domain. The term *ontology change* is used to mean any kind of ontology modification, including the problem of deciding the modifications required, followed by implementation and their management. These modifications may be due to the requirement to improve the conceptualization of a

Z. Bhuyan (✉) · S. M. Hazarika
Biomimetic and Cognitive Robotics Lab, Department of Computer Science and Engineering, Tezpur University, Tezpur 784028, India
e-mail: zubin@tezu.ernet.in

S. M. Hazarika
e-mail: smh@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_71

domain, a requirement change from the users, a dynamic alteration in the domain, or due to perception of information which was previously unknown or unavailable. In this paper we make use of FCA techniques for revision of $\mathcal{SHOIN}^D$ ontology.

Background and Related Work

Description Logic

DL has a clear formal semantics, and expressive diversity to cover a variety of representation needs, and this has enabled it to become the preferred choice for ontology representation. The building blocks in DLs to represent knowledge are *concepts*,[1] *roles* and *individuals*. The *namespace* of an ontology defined using a DL is the set of allowable names and terms for concepts, roles and individuals. The namespace, usually referred to as N_{DL}, comprises of three disjoint sets, viz., N^C_{DL}, N^R_{DL} and N^I_{DL}, which respectively denote the sets of concepts, roles and individuals in the DL namespace. Primitive elements, i.e., concepts, roles and individuals may be combined using *operators* and *connectives* to form complex definitions. For example, the idea that concept *mother* is equivalent to a female human with at least one human child is expressed in (71.1).

$$\text{Mother} \equiv \text{Human} \sqcap \text{Female} \sqcap \exists \text{HasChild.Human} \tag{71.1}$$

The part of DL Knowledge Base (DLKB) involving concepts and roles is called the *TBox*, and the other part involving individuals is called the *ABox*.

Web Ontology language, or OWL, is a family of ontology languages which was primarily aimed at the semantic web community, but have been used in diverse area as medicine, bioinformatics, geography, astronomy and robotics. The relation between DLs and OWL was established in [2]; it was shown that OWL-DL and OWL-Lite were equivalent to the DLs $\mathcal{SHOIN}^D$ and $\mathcal{SHIF}^D$ respectively. Readers are referred to [3] for a detailed understanding of description logics.

Formal Concept Analysis

Formal Concept Analysis (FCA) is an order-theoretic technique of abstracting a *concept hierarchy* from a set of individuals and their corresponding properties [4]. Formally, a *context* is defined as $K = (O, A, I)$. Here O and A are sets of objects

[1] Note that the idea of *concepts* in description logic is different from that of Formal Concept Analysis.

and attributes respectively, and I is an incidence relation between O and A, i.e., $I \subseteq O \times A$.

For every *formal context* (O, A, I), the two concept forming operators $^{\uparrow} : 2^O \rightarrow 2^I$ and $^{\downarrow} : 2^I \rightarrow 2^O$ are defined for every $X \in O$ and $Y \in A$ as:

$$X^{\uparrow} = \{y \in A | \text{for each } x \in X : (x, y) \in I\}$$
$$Y^{\downarrow} = \{x \in O | \text{for each } y \in Y : (x, y) \in I\}$$

A *formal concept* is a pair (X, Y) where $X \in O$ and $Y \in A$ such that $X^{\uparrow} = Y$ and $Y^{\downarrow} = X$. X is referred to as the concept's *extent*, and Y is called it's *intent*. The *subconcept* relation between two concepts (X_1, Y_1) and (X_2, Y_2) is denoted by $\leq$, and is defined as:

$$(X_1, Y_1) \leq (X_2, Y_2) \Leftrightarrow X_1 \subseteq X_2 \Leftrightarrow Y_2 \subseteq Y_1$$

The set comprised of the concepts of a context (O, A, I) is denoted by $\mathcal{L}(O, A, I)$. This set of concepts is partially ordered by the *subconcept* relation, and is called the formal concept lattice. This lattice is represented as the tuple $(\mathcal{L}, \leq)$. For a review of the theoretical foundation and current trends in FCA, readers may refer to [5].

Existing Work on FCA and Ontology Dynamics

Albeit FCA and DL follow different methodologies, several attempts have been made to combine the two for expedient representation and reasoning. The prime difference between the two is that in DL, the concept description is given independent of the objects of the domain, whereas in FCA it is a fixed pair of object and attributes sets in a formal context.

Sertkaya's Ph.D. dissertation [6] presents a technique for assisting construction of DL KB, along with an extended form of FCA for completion of DL KBs. In [7], the author employs FCA to identify structure in theories which are used to construct ontology via conjunctive concepts and relation among them. Hance et al. in [8], proposes a platform for ontology management using FCA based *Relational Concept Analysis*.

Another FCA based approach is presented in [9] for designing of sensor ontology which allows knowledge sharing and reasoning on sensor equipments' characteristics. In [10], an FCA based method of development of ontology from ambiguous and implicit data is proposed. A concise state-of-the art survey of FCA based construction and merging of ontology can be found in [11].

Ontology Revision Using FCA Techniques

In this section we discuss the proposed ontology revision method. We espouse FCA based representation and revision for a $\mathcal{SHOIN}^D$ knowledge base. Such an approach will be especially be useful in scenarios where the *TBox* axioms and rules are generated based on the features and behaviour of entities of the domain being considered. It is to be noted here that, unlike standard ontology environments, we make the *Unique Name Assumption*, i.e., all elements of the DL ontology are assumed to have unique names.

Modeling DL Information in FCA

The data is organized as a triple $\mathcal{Z} = (\mathcal{K}, \mathcal{R}, \mathcal{M})$, where $\mathcal{K}$ is a set of contexts whose extents are composed of sets of objects, $\mathcal{R}$ is a context of roles whose lattice represents role hierarchy, and $\mathcal{M}$ is a set of functions, $m_R : O_i \rightarrow 2^{O_j}$, where O_i and O_j are the object sets from $K_i, K_j \in \mathcal{K}$, and R is a *role* between two DL *concepts*. Each m can also be represented as a relation $m \subseteq O_i \times C(\mathcal{R}) \times O_j$. Here $C(\mathcal{R})$ is the concept set of $\mathcal{R}$. We shall discuss about $\mathcal{K}$ and $\mathcal{R}$ in sections "Construction of $\mathcal{R}$" and "Construction of $\mathcal{K}$" respectively.

For ease of representation and understanding, three supplementary functions are defined as follows:

- **Domain function**: $\delta : C(\mathcal{R}) \rightarrow O$ where $\delta(m_R : O_i \rightarrow 2^{O_j}) = O_i$, where $C(\mathcal{R})$ is the concept set of $\mathcal{R}$, i.e., this function maps roles to its respective domain.
- **Range function**: $\rho : C(\mathcal{R}) \rightarrow O$ where $\rho(m_R : O_i \rightarrow 2^{O_j}) = O_j$. The ρ function maps the set of roles to their respective ranges.
- **Function to map a context to a set of roles**: $\mu : \mathcal{K} \rightarrow 2^{C(\mathcal{R})}$, i.e., $\mu(K_i) = \{R | \delta(r) = O_i\}$. The function μ maps contexts from $\mathcal{K}$ to a set of roles, such that their domain is the given context.

Construction of $\mathcal{R}$

DL *role hierarchies* are defined by axioms having the form $R \sqsubseteq S$ where R and S are both atomic.

In our model a formal context (R, R, I) is constructed where R is the set of all role names and I is a binary relation $\subseteq R \times R$. If rIr' we say that r' subsumes R. Hence, we get a concept lattice of role hierarchies.[2]

[2] All concepts in this lattice which directly subsume supremum are independent roles with no sub-roles or super-roles.

Construction of $\mathcal{K}$

Every context K_i in $\mathcal{K}$ has a set of objects forming its *extent*. Construction of every formal context begins with one primitive concept (class) and its immediate sub-concepts as its intent. The corresponding individuals of the *ABox* forms its *extent*. That is, for each *primitive* DL concept there exists a $K \in \mathcal{K}$. Thereafter, the intent of the formal context is iteratively augmented by concept descriptions like $\neg C$, $\sqcup C$, $\sqcap C$, $\forall r.C$ and $\exists r.C$, where C is the name of a DL concept.

Adding Conjunction to Attribute Set ($\sqcap C$) With respect to a context K_i, let C be a formal concept already defined in some other context, K_j. If a definition for an object is encountered with $\sqcap$, it is added to the intent of the context K_i. It is important to note here that C will be a concept of another lattice, i.e. $C \in \mathcal{L}_j$.

Mathematically it can be defined as a conjunction operator, $op_{\sqcap}^{\mathcal{L}_j} : \mathcal{K} \rightarrow \mathcal{K}$. For a given context $K_i = (O_i, A_i, I_i)$ the conjunction operator, with respect to the lattice, extends the A_i and I_i, but keeps the O_i unaltered.

$$
\begin{aligned}
op_{\sqcap}^{\mathcal{L}_j}(K_i) =& (O_i^{\mathcal{L}_j}, A_i^{\mathcal{L}_j}, I_i^{\mathcal{L}_j}) \\
where: \quad & O_i^{\mathcal{L}_j} = O_i, \\
& A_i^{\mathcal{L}_j} = A_i \cup \{\sqcap C | C \in \mathcal{L}_j\}, \\
& I_i^{\mathcal{L}_j} = I_i \cup \{(o, \sqcap C) | o \in O_i, C \in \mathcal{L}_j\}
\end{aligned}
$$

Adding Disjunction to Attribute Set ($\sqcup C$) If a definition for an object is encountered with $\sqcup$, it can be directly added to the intent of the context under consideration. For a given context $K_i = (O_i, A_i, I_i)$ the disjunction operator results in an extension of the A_i and I_i, keeping the O_i unaltered.

$$
\begin{aligned}
op_{\sqcup}^{\mathcal{L}_j}(K_i) =& (O_i^{\mathcal{L}_j}, A_i^{\mathcal{L}_j}, I_i^{\mathcal{L}_j}) \\
where: \quad & O_i^{\mathcal{L}_j} = O_i, \\
& A_i^{\mathcal{L}_j} = A_i \cup \{\sqcup C | C \in \mathcal{L}_j\}, \\
& I_i^{\mathcal{L}_j} = I_i \cup \{(o, \sqcup C) | o \in O_i, C \in \mathcal{L}_j\}
\end{aligned}
$$

Adding Negation to Attribute Set ($\neg C$) Attribute set extension can be defined via a *negation operator*, $op_{\neg}^{\mathcal{L}_j} : \mathcal{K} \rightarrow \mathcal{K}$.

$$
\begin{aligned}
op_{\neg}^{\mathcal{L}_j}(K_i) =& (O_i^{\mathcal{L}_j}, A_i^{\mathcal{L}_j}, I_i^{\mathcal{L}_j}) \\
where: \quad & O_i^{\mathcal{L}_j} = O_i, A_i^{\mathcal{L}_j} = A_i \cup \{\neg C | C \in \mathcal{L}_j\}, \\
& I_i^{\mathcal{L}_j} = I_i \cup \{(o, \neg C) | o \in O_i, C \in \mathcal{L}_j\}
\end{aligned}
$$

Scaling of Relations ($\forall r.C$ **and** $\exists r.C$) Rouane et al. in [12] have presented an adaptation of FCA which analyzes objects and their relational attributes. It proposes a method called *relational scaling* to incorporate relations in FCA. They define two encoding schemes for relational scaling, namely *narrow encoding* for $\forall R.C$ and *wide encoding* for $\exists R.C$.

We employ a similar technique to include the relations as binary attributes in the formal context. However, our technique also preserves the defined *role hierarchy*. The newly incorporated attribute name will be of the form $R : C$. (Attribute name should be based on the concept name for unique identification.)

Let us consider a DL *role* R with $\delta(R) = O_i$ of K_i and $\rho(R) = O_j$ of K_j.

The operator for *narrow scaling*, $op_{\times}^{\mathcal{L}_j} : \mathcal{K} \rightarrow \mathcal{K}$ is expressed as:

$$\begin{aligned} op_{\times}^{\mathcal{L}_j} =& (\mathcal{K}_i) = (O_i^{R,\mathcal{L}_j}, A_i^{R,\mathcal{L}_j}, I_i^{R,\mathcal{L}_j}) \\ where: \quad & O_i^{R,\mathcal{L}_j} = O_i,\ A_i^{R,\mathcal{L}_j} = A_i \cup \{\forall R : C \mid C \in \mathcal{L}_j\}, \\ & I_i^{R,\mathcal{L}_j} = I_i \cup \{(o, \forall R : C) | o \in O_i, C \in \mathcal{L}_j, m_R(o) \neq \phi, \\ & \qquad m_R(o) \subseteq extent(C)\} \end{aligned}$$

For $\exists R.C$, the operator for *wide scaling*, $op_{+}^{\mathcal{L}_j}$, is defined as:

$$\begin{aligned} op_{+}^{\mathcal{L}_j} =& (\mathcal{K}_i) = (O_i^{R,\mathcal{L}_j}, A_i^{R,\mathcal{L}_j}, I_i^{R,\mathcal{L}_j}) \\ where: \quad & O_i^{R,\mathcal{L}_j} = O_i,\ A_i^{R,\mathcal{L}_j} = A_i \cup \{\exists R : C \mid C \in \mathcal{L}_j\}, \\ & I_i^{R,\mathcal{L}_j} = I_i \cup \{(o, \exists R : C) | o \in O_i, C \in \mathcal{L}_j, m_R(o) \cap extent(C) \neq \phi\} \end{aligned}$$

Operators defined in the above subsections can be used in a composite manner. Considering the set of roles given by $\mu(K_i)$, whose domain is a context K_i, composite scaling can be defined as:

$$K_i = op_{\odot}(op_{\odot}(\ldots op_{\odot}(K_i))) \tag{71.2}$$

In (71.2), $op_{\odot}$ can be any operator defined above.

Transitive Roles and Cardinality Restrictions Transitive roles are handled by explicitly incorporating new mappings for $M_{R_t} \in \mathcal{M}$, where R_t is a transitive role. For example, let o_1 and o_2, be related via a known transitive role R_t. Similarily, o_2 and o_3 are also related via R_t, i.e., $m_{R_t}(o_1) = o_2$ and $m_{R_t}(o_2) = o_3$.

In such cases, an explicit mapping is introduced in m_{R_t} for $m_{R_t}(o_1) = o_3$. Accordingly attribute set and incidence relation in the corresponding context K_i has to be augmented as shown in sections "Ontology Revision Using FCA Techniques" and "Conclusion and Future Work". The t in the attribute name indicates that it is a transitive role.

$$A_i^{R,\mathcal{L}_j} = A_i \cup \{t : R_t : C | C \in \mathcal{L}_j\} \tag{71.3}$$

$$I_i^{R,\mathcal{L}_j} = I_i \cup \{(o_l, t : R_t : C) | o_l \in O_1, C \in \mathcal{L}_j, m_R(o_l) = o_3\} \tag{71.4}$$

Restrictions on cardinality can be checked or enforced by maintaining the cardinality of the set $m_R(o)$ where o is an object.

Inverse Relation (R^-) In DL, inverse of a role R is denoted by R^-, and its semantics is given by $\{(x, y)|(y, x) \in R^I\}$. For such a role $R \equiv S^-$, a set of functions in $\mathcal{M}$ have to be defined. Following our convention, the corresponding m_R function is defined as:

$$\{m_{R_k} : O_{j_k} \rightarrow 2^{O_i} | \text{k is the index representing elements of } 2^{O_i} \text{ in } m_R : O_i \rightarrow 2^{O_j}\}$$

Accordingly attribute scaling in $K_j$has to be done.

Nominals ($\mathcal{O}$) *Nominals* define a class in terms of a specified number of values. Nominal classes will have individual contexts in $\mathcal{K}$ just like any other primitive class. For a nominal DL concept C, with possible values $\{o_1, o_2, \ldots, o_n\}$, its corresponding context's intent will have attributes $\{o_1 : C, o_2 : C, \ldots, o_n : C\}$.

Datatype properties (^{D}) In many ontology languages, *object properties* specify relationships between pairs of objects, and *datatype properties* specify relationships between a resource and a datatype value. The sets of object and datatype properties are mutually disjoint.

An elegant approach on Typed FCA has been made by Peake et al. in [13]. Based on this work, we introduce datatype values as an additional context in $\mathcal{K}$ by defining an auxiliary mapping s which maps type labels, T, to the values it can take.

$$s : T \rightarrow 2^U, U \text{ is a countable set of all possible values.}$$

Let $\top_s = \bigcap s(t)$ and $\bot_s = \bigcup s(t)$, where $t \in T$, and $L_s = \{s(t)|t \in T\} \cup \{\top_s, \bot_s\}$. Hence, we get a complete lattice $(L_s, \subseteq)$ with $\top_s$ as the supremum and $\bot_s$ as the infimum. An example of the type lattice is shown in Fig. 71.1. This lattice can be treated as any another member of $\mathcal{K}$, and the *datatype properties* can be represented as *object type* property. Augmented attributes should be appropriately named so that datatype propertied can be explicitly identified while mapping back the contexts to DL.

Construction of Lattice from $\mathcal{Z}$ The construction of lattices for $\mathcal{Z}$ is done in an iterative fashion which alternates between lattice construction and expansion of attribute sets of the contexts of $\mathcal{K}$. We will adopt the notations used in [12].

The process initiates with the construction of lattices for each of the contexts considering only the concept definitions, i.e., excluding all relational information. Given a context K_i, its attribute set and incidence pairs gradually grow along the iterative process. Its evolution can be represent as a series, K_i^p, where p is the current number of iteration. The starting element, $K_i^0 = (O_i, A_i^0, I_i^0)$, is the original context K_i. The

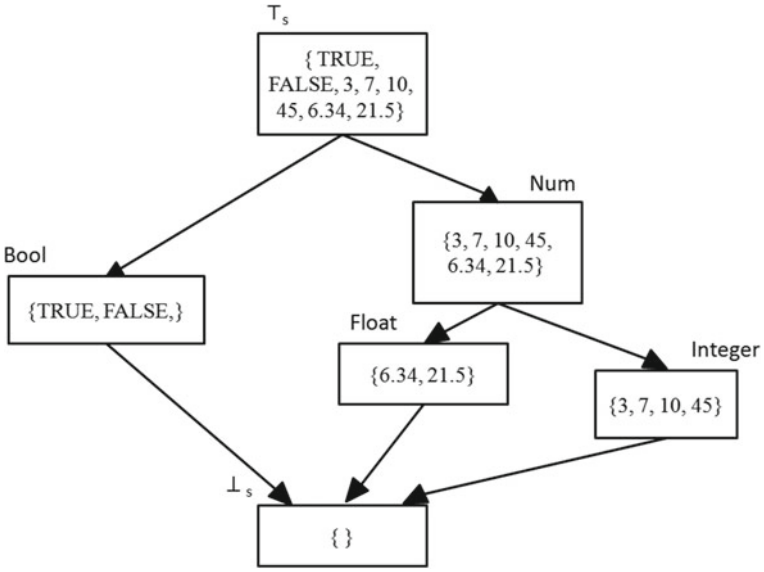


Fig. 71.1 A finite lattice of types

subsequent context state of the series is obtained from the previous state by application of any of the intent expansion operators defined above. It can be expressed as $K_i^{p+1} = (K_i^p)^{rel_p}$, where $.^{rel_p}$ indicates that an operator was applied at step p of the procedure. Augmenting a context with new attributes results in expansion of the corresponding lattice with new concepts; however it does not alter or remove any previously attained concepts [14]. One point to note here is that for any O_i there can be $2^{|O_i|}$ possible concepts, hence the concept set is bound from above by $2^{|O_i|}$.This guarantees termination of the context analysis and lattice construction process.

Whenever for two successive steps all pairs of lattices, say $\mathcal{L}_i^p$ and $\mathcal{L}_i^{p+1}$, remain isomorphic, termination of the procedure is affirmed. The final lattice formed is represented as $\mathcal{L}_i^\infty$. Another vital matter to be kept in mind is that the lattices thus formed will have several new concepts which were not originally defined in the DL *TBox*. Hence, for mapping back the FCA concepts to DL statements, we shall consider only those concepts whose definitions and subsumptions were used to augment the attribute sets of the contexts in $\mathcal{K}$.

Updating the Ontology

Update statements are meant to infuse new information into the ontology.
The types of update statements are as follows:

1. **Inclusion/Modification/Deletion of individuals of ABox and their properties**: If the statement involves addition, deletion or changing properties of individuals of *ABox*, it will only result in the change in extent of a context, migration of an item to a different concept or creation of a new concept. This form of revision is handled in the initial stage of lattice formation. For removal or inclusion of new individuals, the respective extents of the contexts are changed. If o is the new element to be removed or included, it can be expressed as $O_{i+1} = O_i \cup \{o\}$, or $O_{i+1} = O_i - \{o\}$, where O_i is the extent of a context K_i. Modifications to the property values of an object are reflected directly in the incidence mapping of the contexts where the modifications are made.
2. **Inclusion/Modification/Deletion of roles**: For updates of this kind, the relevant role(s) in $C(\mathcal{R})$ has to be added or shifted to the appropriate subsumption hierarchy level. If a role R is required to be deleted, it is removed from $C(\mathcal{R})$ along with all its sub-roles. Thereafter, all references to R are removed in the following process where $\mathcal{K}$ is constructed. However, if the sub-roles of R require preserving, these have to be first moved up in the lattice hierarchy and the role R has to be deleted.
3. **Inclusion of new concepts and Modification/Deletion of concepts**: Update statement for inclusion of a new primitive DL concept will result in creation of a new context in $\mathcal{K}$ whose extent will be the individuals of the new concept.

Constructing the Ontology from $\mathcal{Z}$

The process of building the Description Logic KB from $\mathcal{Z}$ begins with the mapping of the initial set of attributes of all contexts to relevant DL concepts. Concepts from the lattice of $\mathcal{R}$, i.e. $\mathcal{L}_{\mathcal{R}}$, are mapped to roles, objects to individuals and formal concepts in $\mathcal{L}_i^{\infty}$ are mapped to defined concepts. Let λ be a bijection from $\mathcal{Z}$ to the DL KB.

Denoting a DL knowledge base as $((T_C, T_R), ABox)$, λ is defined as:

$$\begin{gathered}\forall K_i \in \mathcal{K}, \forall a_i \in A_i^0 : \lambda(a_i) \in T_C \\ \forall R_i \in C(\mathcal{R}) : \lambda(R) \in T_R \\ \forall o \in O_i : \lambda(o) \in ABox \\ \forall c_i \in \mathcal{L}_i^{\infty} \text{ which are defined using DL constructs } : \lambda(c_i) \in T_C\end{gathered}$$

An operator used to augment the particular attribute set of a context implies a DL construct. The inverse of these operators are used to obtain the complete description of the defined concepts of $\mathcal{L}_i^{\infty}$.

For construction of the *ABox*, the object sets of formal contexts are associated with their respective concepts and the basic relations are deduced.

$$\forall a \in A_i^0, \forall o_i \in \{a\}^{\downarrow} : [\lambda(a)](\lambda(o_i)) \in ABox$$
$$\forall R_i \in C(\mathcal{R}), \forall o_1, o_2 \text{ in } m_R(o_1, o_2) : [\lambda(R_i)](\lambda(o_1), \lambda(o_2)) \in ABox$$
$$\forall c_i \in \mathcal{L}_i^{\infty}, \forall o_i \in extent(c_i) : [\lambda(c_i)](\lambda(o_i)) \in ABox$$

Value restrictions are translated into DL using the attributes which were introduced during relational scaling. For any concept $c_i \in \mathcal{L}_i^{\infty}$, with an attribute $R : C$, the DL mapping through λ will be $\lambda(c_i) = [\lambda(R)](\lambda(C)) \in T_C$).

Conclusion and Future Work

In this paper we propose an adaptation of FCA to represent and alter a $\mathcal{SHOIN}^D$ KB. Such FCA based manipulations holds promise in scenarios where ontology revisions are generated by the monitoring systems which report changes in relations and properties of object. Algorithmic and reasoning issues, such as decidability and complexity, for this proposed framework are topics of future research. Extending the current framework for other DLs like OWL-FULL, and its practical implementation appear to be a challenging and exciting path to head.

References

1. Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.
2. Ian Horrocks and Peter Patel-Schneider. Reducing OWL entailment to description logic satisfiability. *J. of Web Semantics*, 1(4):345–357, 2004.
3. Markus Krötzsch, Frantisek Simancik, and Ian Horrocks. A Description Logic Primer. *CoRR*, abs/1201.4089, 2012.
4. Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1997.
5. Achyanta Kumar Sarmah, Shyamanta M. Hazarika, and Smriti Kumar Sinha. FCA: Current trends and directions. *Artif. Intell. Rev*., 44(1):47–86, 2015.
6. Barış Sertkaya. *Formal Concept Analysis Methods for Description Logics*. PhD thesis, Institute of Theoretical Computer Science, TU Dresden, Germany, 2007.
7. Michael Bain. *Inductive Construction of Ontologies from Formal Concept Analysis*, pages 88–99. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.
8. Mohamed Rouane-Hacene, Petko Valtchev, and Roger Nkambou. Supporting ontology design through large-scale fca-based ontology restructuring. In *Proceedings of the 19th International Conference on Conceptual Structures for Discovering Knowledge*, ICCS'11, pages 257–269, Berlin, Heidelberg, 2011. Springer-Verlag.
9. P. Wang, W. Lu, Z. Meng, J. Wei, and F. Fogelman-Soulié. Towards Knowledge Structuring of Sensor Data Based on FCA and Ontology. In *Proc. of the 2nd Int. Conf. on FCA4AI*, pages 57–57, Aachen, Germany, 2013. CEUR-WS.
10. Gaihua Fu. Fca based ontology development for data integration. *Inf. Process. Manage*., 52(5):765–782, 2016.
11. M. Priya and Ch. Ashwini Kumar. A Survey of State of the Art of Ontology Construction and Merging using Formal Concept Analysis. *Indian Journal of Science and Technology*, 8(24), Sept. 2015.

12. Mohamed Hacene Rouane, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. A proposal for combining formal concept analysis and description logics for mining relational data. In *Proceedings of the 5th International Conference on FCA*, ICFCA'07, pages 51–65, Berlin, Heidelberg, 2007. Springer-Verlag.
13. Ian Peake, Ian Thomas, and Heinrich Schmidt. Typed formal concept analysis. In *Contributions to ICFCA 2009 (Supplementary Proceedings), 7th International Conference on FCA*, ICFCA'09, pages 35–51. Springer-Verlag, 2009.
14. Petko Valtchev, Rokia Missaoui, Robert Godin, and Mohamed Meridji. Generating frequent itemsets incrementally: Two novel approaches based on Galois lattice theory. *J. Exp. Theor. Artif. Intell*., 14(2-3):115–142, 2002.

Chapter 72
A Density-Based Clustering for Gene Expression Data Using Gene Ontology

Koyel Mandal and Rosy Sarmah

Abstract Gene expression clustering is built on the premise that similarly expressed genes are included in the same kind of biological process. Recent research has focused on the fact that incorporation of biological knowledge such as gene ontology (GO) improves the result of clustering. This paper demonstrates a *Semi-supervised Density-based Clustering* (SDC) which uses GO to detect positive and negative co-regulated patterns from the noisy gene expression data. SDC improves a previous algorithm *DenGeneClus* (DGC) which could handle only positive co-regulation and did not include GO in the clustering process. Experimental results on four real-life data show that SDC outperforms DGC based on z-score and gene ontology enrichment analysis.

Keywords Gene expression data · Semi-supervised clustering · Gene ontology Positive and negative co-regulated gene

Introduction

High-throughput experiments such as DNA microarray technology have generated huge amount of data, analysis of which require high performing computational methods such as clustering. DNA microarrays have helped researchers to observe the expressions of enormous amount of genes at different conditions such as time, different stages of diseases, or drug applications [1]. The output of DNA microarrays after several preprocessing steps is finally obtained as a gene expression numerical matrix which is mathematically written as, $GE_{G\times C} = \{ev_{i,t} | i \in G, t \in C\}$, where

K. Mandal (✉) · R. Sarmah
Department of Computer Science and Engineering, Tezpur University, Sonitpur 784028, Assam, India
e-mail: koyel@tezu.ernet.in

R. Sarmah
e-mail: rosy8@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_72

$ev_{i,t}$ denotes the expression value of ith gene at condition t, G the total gene set, and C the condition set, respectively [1].

Clustering helps to identify co-expressed, coherent, and co-regulated genes from the given expression data [2]. Co-regulated genes can be of two types: positively co-regulated and negatively co-regulated genes [3]. Genes g_1 and g_2 are said to be positively co-regulated if the expression value of g_1 increases (or decreases) from condition t_i to t_j then the expression level of g_2 also increases (or decreases) from t_i to t_j. Two genes g_1 and g_2 are said to be negatively co-regulated if the expression value of g_1 increases (or decreases) from condition t_i to t_j then the expression level of g_2 also decreases (or increases) from t_i to t_j.

Unsupervised clustering algorithms are built on the presumption that co-expressed genes are likely to have common biological functions. However it is seen that most of the algorithms miss the gene functional prediction at the time of clustering. This has motivated us to shift from unsupervised to semi-supervised clustering by incorporating gene ontology (GO) knowledge in the clustering process. GO is the fundamental database of bioinformatics that specifically gives the annotations for gene products with consistent and structured vocabularies [4].

Algorithms based on the density information give quality clusters, and when GO knowledge is incorporated into the clustering process, we get more biologically significant clusters. Therefore, in this paper, we have combined both density and GO information to get the benefits of both in our proposed algorithm.

A Semi-supervised Density-based Clustering (SDC) is being proposed, using the density information of a gene and external knowledge from GO to discover more biologically relevant clusters from noisy data. This work overcomes the drawbacks of a density-based clustering algorithm (DenGeneClus, DGC) [5] by discovering both the positively and negatively co-regulated genes.

Background

Clustering of gene expression data is widely classified into five types, viz. hierarchical, partitional, density-based, graph-theoretical, and model-based [1, 6, 7]. From the various approaches surveyed, we find that density-based algorithm is not dependent on number of clusters. DGC [5], DHC [8], OverDBC [9], and Bayesian-OverDBC [10] are the examples of density-based gene expression data clustering. Conventional clustering algorithms find sets of genes depending on their proximity (similarity or dissimilarity) measure. Most commonly used proximity measure is Euclidean distance which gives the dissimilarity between gene g_i, g_j as $Dis_{Euc}(g_i, g_j) = \sqrt{(\sum_{t=1}^{|C|}(g_{i,t}, g_{j,t})^2)}$ [1]. Expression based measures may not find the potential relationships among the genes. Therefore, it is necessary to guide the clustering process with external domain knowledge. Semantic similarity measure is the key technique to incorporate the knowledge of known genes from gene ontology and gene annotation file. Some of the well-known semantic similarities are

Resnik's, Jiang and Conraths's and Lin's [11]. These semantic similarities are built on the information theory which means how much information they commonly share. Information content of a term t represented as IC in a specific corpus is described by $IC = -log(P(t))$, where $P(t)$ represents probability of occurrence of t. In our proposed method, we have used Lin similarity between two terms say t_i and t_j and given by $Sim_{Lin}(t_i, t_j) = \frac{2 \times IC(LCA)}{IC(t_i)+IC(t_j)}$. Sim_{Lin} gives the IC between two terms by considering the IC of each individual term and the IC of lowest common ancestors (LCA). The value of Sim_{Lin} lies between 0 and 1. We combined Sim_{Lin} and Dis_{Euc} to improve the clustering result. We first convert Dis_{Euc} into a similarity measure as $Sim_{Euc} = \frac{1}{1+Dis_{Euc}}$. Then we find the combined similarity (Com_sim) given next.

$$Com\_sim = w1 * Sim_{Euc} + w2 * Sim_{Lin} \tag{72.1}$$

where, $w1 + w2 = 1, 0 \leq w1 \leq 1$ and $0 \leq w2 \leq 1$ [4]. $w1$ and $w2$ control the weights to two similarity measures. Hang et al. [12] proposed an algorithm using two information such as gene density function and biological knowledge, and the proposed one gave better result than standard algorithm. Zhou et al. [13] also proposed an algorithm incorporating density of data and gene ontology in distance-based clustering algorithm. Both the algorithms do not address the issue of identifying the positive and negative co-regulated genes. An algorithm which finds clusters comprised with co-regulated genes is being proposed by Ji and Tan [3]. To identify interesting partial negative positive co-regulated gene cluster, Koch et al. [14] proposed an algorithm which also discovers overlapping clusters.

Proposed Method

Our SDC is a density-based clustering algorithm which works in two phases (i) preprocessing and (ii) clustering phase.

Preprocessing step is initiated by normalizing (standard deviation 1 and mean 0) the gene expression data. Then, a discretization process discretizes the gene expression data, and the discretized data (GE_{disct}) is fed as input to the clustering algorithm.

In **Discretization** step, each cell $ev_{i,t}$, (where $t = 1$) of the gene expression data (GE) for the first condition is discretized by using Eq. 72.2, and for the other conditions ($C - t_1$), each cell $ev_{i,t}$ (where $t = 2, 3..., |C|$) is computed using Eq. 72.3. Each gene in GE_{disct} will now have a pattern of regulation values 0^s, 1^s, and 2^s across condition known as regulation pattern.

$$GE_{disct}(i, 1) = \begin{cases} 2 & \text{if } ev_{i,1} < 0 \\ 0 & \text{if } ev_{i,1} = 0 \\ 1 & \text{if } ev_{i,1} > 0 \end{cases} \tag{72.2}$$

$$GE_{disct}(i,t) = \begin{cases} 2 & \text{if } ev_{i,t} < ev_{i,t-1} \\ 0 & \text{if } ev_{i,t} = ev_{i,t-1} \\ 1 & \text{if } ev_{i,t} > ev_{i,t-1} \end{cases} \tag{72.3}$$

After the computation of each gene's regulation pattern, next job is to calculate the match (M) between genes g_i and g_j stated in Eq. 72.4.

Definition 1 Match: Match (M) gives the number of common regulation value according to the conditions except the first one, which signifies how similar two patterns are with respect to their expression values.

If $M = |C| - 1$, it can be said that two patterns are almost similar. The match between g_i and g_j is calculated as below.

$$Pat_t^{i,j} = \begin{cases} 1 & \text{if } GE_{disct}(ev_{i,t}) = GE_{disct}(ev_{j,t}) \text{where } t = 2, ..., |C| \\ 0 & \text{otherwise} \end{cases} \tag{72.4}$$

$$M(g_i, g_j) = number\ of\ 1^s\ in\ Pat_t^{i,j} \tag{72.5}$$

Definition 2 Maximal Match: If match between g_i and g_j is equal or greater to the minimum threshold value δ, ($M(g_i, g_j) >= \delta$) and no other gene exists whose match (M) with respect to g_i is greater than g_j, then g_i has a maximal match (MM) with another g_j ($g_i \neq g_j$).

Definition 3 Maximally Matched Regulation Pattern: For genes g_i and g_j, let g_i be maximally matched with g_j, then the Maximally Matched Regulation Pattern ($MMRP$) is computed using Eq. 72.6 by considering the subset (two gene profiles may not match throughout $|C| - 1$ conditions) of conditions where they maximally matched based on δ.

$$MMRP(g_{i,t}) = MMRP(g_{j,t}) = \begin{cases} 2 & \text{if } GE_{disct}(g_{i,t}) = 2 = GE_{disct}(g_{j,t}) \\ 0 & \text{if } GE_{disct}(g_{i,t}) = 0 = GE_{disct}(g_{j,t}) \\ 1 & \text{if } GE_{disct}(g_{i,t}) = 1 = GE_{disct}(g_{j,t}) \\ x & \text{otherwise} \end{cases} \tag{72.6}$$

where, $t = 2, 3, ..., |C|$. Therefore, for the whole set of t conditions, we obtain an $MMRP$ pattern of 0^s, 1^s, 2^s and x^s.

Definition 4 Negative Maximally Matched Regulation Pattern: The Negative Maximally Matched Regulation Pattern ($NMMRP$) of g_j is determined by comparing the $MMRP$ of g_i as stated in Eq. 72.7.

$$NMMRP(g_{j,t}) = \begin{cases} 2 & \text{if } MMRP(g_{i,t}) = 1 \\ 1 & \text{if } MMRP(g_{i,t}) = 2 \\ 0 & \text{if } MMRP(g_{i,t}) = 0 \\ x & \text{if } MMRP(g_{i,t}) = x \end{cases} \tag{72.7}$$

Therefore, we obtain a *NMMRP* pattern for t conditions ($t = 2, 3, ..., |C|$).

Definition 5 Rank: *Rank* gives the ascending order of expression levels of a gene across conditions.

Rank is measured by giving a ranked value starting from 1 to all the expression values in the MMRP pattern except for those conditions having a x value. The working example of the computation of *M*, *MM*, *MMRP*, *NMMRP*, and *Rank* is available in http://agnigarh.tezu.ernet.in/~rosy8/workingexampleSDC.pdf.

The second phase, **Clustering** of SDC is based on some of the fundamental concepts of density-based clustering. The following definitions are trivial to the clustering process.

Definition 6 ϵ-neighbor: ϵ-neighbors with respect to $g_i \in G$ are those genes $g_k \in G$, which have more similarity than the user defined threshold (ϵ). Here, we have used combined similarity which is mentioned in Eq. 72.1.

$$\epsilon - neighbors(g_i) = \{g_k | where\ g_k \in G\ and\ com\_sim(g_i, g_k) >= \epsilon\} \tag{72.8}$$

Definition 7 Core-neighbors: Core-neighbors of a gene $g_i \in G$ is described by a set of genes $G^* \in G$ and should satisfy the following four criteria. A gene, say g_i is considered as core gene if.

1. $\forall g_y \in G^*, g_y \in \epsilon - neighbors(g_i)$.
2. $MMRP(g_y) \approx MMRP(g_i)$.
3. $Rank(g_y) \approx Rank(g_i)$.
4. $|G^*| >= min\_points$ (a user defined threshold).

To compute the core-neighbors of a particular gene g_i, we check the above-mentioned four criteria for all the $C - 1$ dimensions (except condition 1). If we do not get the core-neighbors, we will go on checking the criteria by reducing the search space one condition at a time. At first we reduce the condition set by $C - \{t_1, t_l\}$, where t_1 and t_l are the first and the last condition respectively, i.e., $|C| - 1 - 1 = |C| - 2$. If we still do not find the core-neighbors of g_i, we further reduce the search space by the second last condition i.e., $C - \{t_1, t_{l-1}, t_l\}$, where t_1, t_{l-1} and t_l are the first, second last and the last condition respectively. In other words the condition set is reduced by $|C| - 2, |C| - 3, |C| - 4 \ldots$ and so on.

Definition 8 Direct density reachable: g_i is direct density reachable with respect to g_j if it fulfills three basic principles.

1. g_j must be a core gene or g_j must have core-neighbors.
2. $g_i \in \epsilon - neighbors(g_j)$.
3. $MMRP(g_i) \approx MMRP(g_j)$.

In case of pairs of core genes, direct density reachable relation holds symmetric relation.

Definition 9 Density reachable: Gene g_q is density reachable from g_p provided there is a chain of genes $g_1, g_2, g_3, ..., g_n$ such that $g_1 = g_p$ and $g_n = g_q$ and every g_{i+1} gene is directly density reachable from g_i^{th} gene.

Definition 10 Connected: Gene g_i is connected to g_j with respect to ϵ, provided g_i and g_j are reachable from a common gene say g_k.

This relation holds symmetric property.

Definition 11 Cluster: A cluster CL ($|CL| >= min\_points$) is a collection of reachable and connected genes. Say, a gene $g_i \in CL$ and the gene g_j is found to be reachable from g_i, then g_j must be in cluster CL. Similarly, if a gene $g_i \in CL$ and g_j is connected to g_i then g_j will be in the same CL cluster.

Definition 12 Noise: A noise gene is a gene which does not belong to any cluster.

The steps of SDC is given next. At first, all genes are not clustered.

Step 1 Start with an random unclustered gene say g_i.
Step 2 Find the $MMRP(g_i)$ and $Rank(g_i)$.
Step 3 Find the core-neighbors of g_i using Definition 7.
Step 4 For each core-neighbors of g_i.

Step 4.1 Identify all connected and reachable genes with respect to each core-neighbors.
Step 4.2 Give the same *cluster_id* for all these genes.

Step 5 End of step 4.
Step 6 Find the *NMMRP* from *MMRP* of the newly formed *cluster_id* .
Step 7 Find the unclustered genes which matches the *NMMRP*.
Step 8 For each gene g_j with matched *NMMRP*.
Step 9 Find the core-neighbors of gene g_j and all reachable and connected genes from it.
Step 10 Assign another cluster_id to all the reachable and connected genes of g_j.
Step 11 Repeat step 1 to 10 with the next unclustered gene.
Step 12 All the unclustered genes are marked as noise.
Step 13 End

Table 72.1 A brief description about the datasets

Serial no.	Name of dataset	Genes/samples	Source
D1	Yeast cell cycle	384/17	http://anirbanmukhopadhyay.50webs.com/data.html
D2	Yeast sporulation	474/7	http://cmgm.stanford.edu/pbrown/sporulation/
D3	Yeast diauxic shift	614/7	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28
D4	Prostate cancer	339/102	http://archive.broadinstitute.org/mpr/publications/projects/Cancer_Susceptibility/references_and_URLS_of_datasets.html

Experimental Result

Implementation of SDC and DGC was done in MATLAB 2015 platform and experimented over four publicly available real-life gene expression datasets. Table 72.1 gives a description about the used datasets in the experiment.
To compute Sim_{Lin}, we have downloaded the most recent gene ontology file (released on 2016-09-10) and annotation files (*Saccharomyces Genome Database* and *Homo Sapience*) from www.geneontology.org. To compare both the DGC and SDC clustering results, we use Dis_{Euc} for DGC and *Com_sim* for SDC. The parameter settings highly influence the clustering results. We keep the value of δ as minimum ($\delta = 3$) as possible. To determine the value of ϵ and *min_points* ($= 4$), we follow the method mentioned in [15]. As we want to give the more weightage on proximity measure than semantic similarity measure, we kept the value of $w1 = 0.6$ and $w2 = 0.4$. The ϵ for DGC and SDC changes from one dataset to another. The ϵ of DGC is 3 and 1.2 for D1 and D2; and for D3 and D4, it is 1 and 10, respectively. The ϵ of SDC is 0.3 for D1, 0.4 for D2, 0.3 for D3 and 0.3 for D4, respectively.

To assess the biological significance of clusters, we eventually investigated the clusters generated by DGC and SDC by functional enrichment analysis. A cluster is called enriched, if at least one of the GO term of a particular cluster from the Biological Process is below the level of significance. P value is being computed using FuncAssociate 3.0 with 5% level of significance [16]. We then analyze the functional category Biological Process (BP) using web (http://www.ebi.ac.uk/QuickGO/) based on GO annotation database. It can be observed from Fig. 72.1 that SDC finds more enriched clusters than DGC.

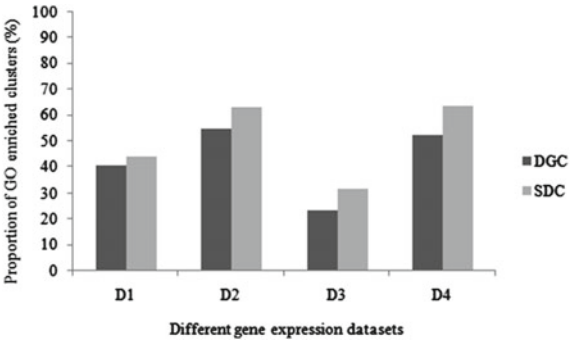


Fig. 72.1 Proportion of enriched clusters of DGC and SDC for different datasets

Table 72.2 Comparison of z-score between DGC and SDC on various datasets

	DGC		SDC	
Datasets	No. clusters	z-score	No. clusters	z-score
D1	27	2.494	25	5.042
D2	31	9.621	27	12.59
D3	47	4.121	41	4.718

To judge the quality of clusters, we have used the Web-based tool cluster judge [17]. The comparison of DGC and SDC based on z-score for different yeast datasets is shown in Table 72.2. Table 72.2 suggests that the clusters generated by SDC have higher z-score value than DGC which proves that the cluster quality is better for SDC.

Conclusion

We have proposed an algorithm incorporating gene ontology in a density-based clustering algorithm. It is being observed that external domain knowledge gives reliable clusters. The drawback of this algorithm is that it finds disjoint clusters and cannot find overlapping clusters. Biologically it is proven that one gene may participate in many biological pathways, and this allows it to belong to multiple clusters. Detecting overlapping clusters is a crucial task and will be incorporated in our future work.

References

1. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. IEEE Transactions on knowledge and data engineering **16**(11) (2004) 1370–1386

2. Jiang, D., Pei, J., Zhang, A.: Gpx: interactive mining of gene expression data. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, VLDB Endowment (2004) 1249–1252
3. Ji, L., Tan, K.L.: Mining gene expression data for positive and negative co-regulated gene clusters. Bioinformatics **20**(16) (2004) 2711–2718
4. Lee, W.P., Lin, C.H.: Combining expression data and knowledge ontology for gene clustering and network reconstruction. Cognitive Computation **8**(2) (2016) 217–227
5. Das, R., Bhattacharyya, D., Kalita, J.: Clustering gene expression data using an effective dissimilarity measure. International Journal of Computational BioScience (Special Issue) **1**(1) (2010) 55–68
6. Kerr, G., Ruskin, H.J., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in biology and medicine **38**(3) (2008) 283–293
7. Pirim, H., Ekşioğlu, B., Perkins, A.D., Yüceer, Ç.: Clustering of high throughput gene expression data. Computers & operations research **39**(12) (2012) 3046–3061
8. Jiang, D., Pei, J., Zhang, A.: Dhc: a density-based hierarchical clustering method for time series gene expression data. In: Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, IEEE (2003) 393–400
9. Mirzaie, M., Barani, A., Nematbakkhsh, N., Beigi, M.: Overdbc: A new density-based clustering method with the ability of detecting overlapped clusters from gene expression data. Intelligent Data Analysis **19**(6) (2015) 1311–1321
10. Mirzaie, M., Barani, A., Nematbakkhsh, N., Mohammad-Beigi, M.: Bayesian-overdbc: A bayesian density-based approach for modeling overlapping clusters. Mathematical Problems in Engineering **2015** (2015)
11. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS comput biol **5**(7) (2009) e1000443
12. Hang, S., You, Z., Chun, L.Y.: Incorporating biological knowledge into density-based clustering analysis of gene expression data. In: Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. Volume 5., IEEE (2009) 52–56
13. Zhou, X., Sun, H., Wang, D.P., Zhang, Y., Zhou, Y.: Analysis of gene expression data based on density and biological knowledge. In: 2010 Fifth International Conference on Frontier of Computer Science and Technology, IEEE (2010) 448–453
14. Xu, X., Lu, Y., Tung, A.K., Wang, W.: Mining shifting-and-scaling co-regulation patterns on gene expression profiles. In: 22nd International Conference on Data Engineering (ICDE'06), IEEE (2006) 89–89
15. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. Volume 96. (1996) 226–231
16. Hochberg, Y., Benjamini, Y.: More powerful procedures for multiple significance testing. Statistics in medicine **9**(7) (1990) 811–818
17. Gibbons, F.D., Roth, F.P.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome research **12**(10) (2002) 1574–1581

Chapter 73
Statistical Analysis of a Data Centre Resource Usage Patterns: A Case Study

Somnath Mazumdar and Anoop S. Kumar

Abstract Performance evaluation is necessary to understand the runtime behaviour of a computing system. A better understanding of resource usage leads to better utilisation and less energy cost. To optimise the server provisioning and also the energy cost of a data centre (DC), we should explore the underlying resource usage patterns to extract meaningful information. In this paper, our primary goal is to obtain correlation or cross-correlation among CPU, RAM, and Network at different timescales of a DC. To perform this analysis, we have collected Wikimedia grid traces and conducted an experimental campaign using rationally selected multiple statistical methods. They are: a univariate method (Hurst exponent), multivariate explanatory methods (such as wavelets, cross-recurrence quantification analysis (CRQA)), multivariate predictive methods (such as vector auto-regression (VAR), multivariate adaptive regression splines (MARS)). It is worth to note that, we analyse the data without any prior knowledge about running applications. We present the results together with a comprehensive analysis. In our case study, we found in long time scale CPU, and RAM is more correlated than Network. We also have shown that wavelet-based methods are superior to detect long-run relationship among these resource variables.

Keywords Time series · Pattern · Analysis · Hurst exponent · Wavelet · CRQA VAR · MARS

S. Mazumdar (✉)
University of Siena, Via Roma 56, Siena, Italy
e-mail: mazumdar@dii.unisi.it

A. S. Kumar (✉)
BITS Pilani–K.K. Birla Goa Campus, Goa 403726, India
e-mail: anoopk@goa.bits-pilani.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_73

Introduction

Energy efficiency is the primary concern to any Data centre (DC) infrastructure irrespective of the type of the running applications. DC consumes an enormous amount of power for running the whole infrastructure and a big chunk of energy cost, directly related to running servers. DCs are mixed of heterogeneous (e.g., micro-architecture, capacity) systems. Maintaining the availability and serviceability of any DC is a complex task. Compute are: oversubscribed to achieve higher utilisation using *virtualisation*, but the fundamental issue of resource *under-utilisation* has not yet been fully solved. A typical DC pays 40–50% of their total expenses as power bills [8], but the average server utilisation is around 10–50% [2]. Apart from that, an idle server also can consume up to 60% of its peak power [11]. Main computing hardware components such as CPU and RAM can consume up to 32% and 14.5% of the system's total peak power [5]. So, to optimise the power consumption, server allocation also has to be optimised. The server utilisation can be optimised by reducing the number of idle servers and also by allocating them efficiently.

The workload is a collective behaviour of various resource components localised in time and frequency. Hence, usage patterns can give valuable information related to running applications (like memory-I/O stalls, components behaviours during failures). Analysing resource usage patterns at proper granularity can increase the overall system performance. In this paper, we advocate for analysing the trend which is embedded into the workload traces. We can learn the necessary (time-dependent) relationships from the workload traces and then allocate the resources using extracted knowledge to optimise the overall server allocation. DCs manage multiple applications at the same time, where some of them are CPU hungry, and some of them are the memory and/or network intensive. In a real distributed environment, the stable resource usage pattern occurred more often, or the trace related data are more correlated [19]. Thus, there lies a space for opportunity to optimise the system further by understanding the data.

In this work, we want to understand how these computing resources behave at runtime and how they interact or influences each other at various timescales. Correlating performances of CPU, RAM, and Network are our main contribution in this paper. To achieve this, we have carefully collected relevant data sets and statistical methods to draw the conclusion. The summary of our contribution is: first, we check the persistence of data using Hurst exponent. Next, we extract the correlation and dependence structure between CPU, RAM, and Network using different multivariate methods (such as wavelets, cross-recurrence quantification analysis (CRQA), vector auto-regression (VAR), and multivariate adaptive regression splines (MARS)). The presented insights can help to improve resource utilisation, response time and also to make a better workload model for the system that could assist in optimising resource allocation.

Related Work

Here, we are analysing the workload traces to tune the resource allocation further, and in turn, this knowledge can optimise the energy cost while keeping the prediction out of the scope of this paper. Analysing network traffic has been very popular research topic to detect self-similarity, long-range dependence (LRD), and heavy-tailed distributions. However, analysing the correlation of Network to others (like CPU and RAM) has not received much attention. Hurst exponent has been used to estimate the web server request inter-arrival times in a DC [4] and also used to test self-similarity in the Grid's job arrival process [17]. In [14], Hurst exponent was used to detect the self-similarity in Ethernet LAN traffic. Wavelet-based Hurst exponent to analyse Internet traffic analysis has proven to be very efficient (such as in [1], where wavelet analysis and Hurst exponent detect LRD in Ethernet traces). This work is quite similar to our work regarding approach but does not consider CPU and RAM. This work also shows that wavelet is a very efficient method to expose LRD in Ethernet time series data.

Wavelet transformation is used to analyse various properties of network traffic [6, 18, 21]. In [21], an independent wavelet model is used to detect the relationship between the "smoothness/burstiness" of the network traffic and the "short-range/long-range" dependence in the data. Feldmann et al. [6] proposed a discrete wavelet transform based model (that multiplicatively generate a class of multifractals) to infer the scaling behaviour of the WAN network traffic from the traces. In another work [18], a multiplicative wavelet model (MWM)-based framework was proposed to characterise and synthesise the positive LRD data. Wavelet was also employed to analyse the correlation and the scaling behaviour of CPU within multiple Grids [15]. In this work, the application was running on a single core, and the pseudo-periodicity, LRD, and multifractals were identified by using a discrete wavelet transform.

In this work, we present a statistically in-depth analysis of collected traces to expose patterns and correlations among CPU, RAM, and Network. We also evaluate the interrelations among the resources at different time intervals. The novelty of our methodology is that it does not require a priori information about the type of the applications running on the DC.

Background

For ease of understanding, we broadly classify the employed methodologies into three groups. They are: (i) univariate methods (Hurst exponent), (ii) multivariate explanatory methods (wavelets and CRQA), and (iii) multivariate predictive methods (VAR and MARS). We start with univariate methods to understand how CPU, RAM, and Network are correlated or dependent on each other. For our analysis, all the selected methods are statistically significant and connected. First, we have

selected Hurst exponent due to its efficiency to detect the persistence of the data. After Hurst, wavelet helps us to identify the long-run relationship in the data. Further, CRQA is used to show how all three resources are affecting each other. Moreover, VAR is used to identify the functional relationship among the variables. Finally, to determine the best functional relationship among the variables, we have used MARS.

Hurst Exponent

Hurst exponent (denoted as H) [12] is a statistical method to measure the level of persistence (self-similarity) of a time series. H value (ranging from 0 to 1) can classify a time series behaviour into three classes. They are geometrical Brownian motion, persistent and anti-persistent time series. In this work, we employ a discrete wavelet transform (DWT) [13]-based method for estimating H. As the name suggests, in DWT, high-pass (details) and low-pass filters (approximation) are applied to the input signal at each level. If a series is decomposed to level J, it could be reconstructed by summing the individual coefficients. Consider a random process x_t such that $(1-L)^d * x_t = \varepsilon_t$, where L is the lag operator, ε_t is the error term (also independently and identically distributed normal) with zero mean, σ^2 as variance and d is the differencing parameter. For a fractionally integrated process $I(d)$ with $d < 0.5$, the autocovariance function shows that the detail coefficients w_j are distributed as $N(0, \sigma^2 * 2^{-2(J-j)d})$, where j is the corresponding scaling parameter. H is calculated as $H = d + 0.5$. The fractional integration parameter d is estimated from the slope of the following equation:

$$Var(j) = log\sigma^2 + d * log2^{-2(J-j)d} \tag{73.1}$$

Here, $Var(j)$ is the logarithmic transformation of variance of the wavelet detail coefficient estimated across different dyadic scales.

Wavelets: Small Waves

Wavelet transform decomposes time series into its high- and low-frequency components [3]. The attractiveness of wavelet transform lies in its decomposition properties for timescale localisation and not requiring the assumption of stationarity in the time series. We have used *Daubechies wavelets* as mother function with a window length of eight. Unlike Haar wavelet, Daubechies wavelet does not suffer from slow convergence issue. To analyse the relationship between variables, we employed wavelet multiple correlation (WMC) [7]. WMC calculates correlation among multivariate time series using *maximal overlap discrete wavelets transformation* (MODWT). A brief explanation of WMC is given below.

Let $\{X_t\}$ be the multivariate time series, and we obtain $\{W_{jt}\}$ as the respective jth level wavelet coefficients by the application of MODWT. The WMC $\phi_x(\lambda_j)$ can be then defined as the set of multiscale correlations estimated from X_t as follows. At each wavelet scale λ_j, we calculate the square root of the regression coefficient of determination in the linear combination of variables $w_{ijt}(i = 1, 2, \ldots n)$, for which the coefficient of determination is a maximum. However, these auxiliary regressions could be avoided as the coefficient of determination with respect to the regression of a variable z_i on a set of regressors $\{z_k, k \neq i\}$ could be calculated as $R_i^2 = 1 - 1/\rho^{ii}$ (where ρ^{ii} is the ith diagonal element of the inverse of the complete correlation matrix P). Hence, $\phi_X(\lambda_j)$ could be obtained as:

$$\phi_X(\lambda_j) = \sqrt{1 - \frac{1}{\max diag P_j^{-1}}} \tag{73.2}$$

where P is the $n \times n$ correlation matrix of W_{ji}, and the max diag (.) operator picks up the largest element in the diagonal of the argument. Since the R_i^2 coefficient in the regression of a z_i on the rest of variables is equal to the correlation between the observed values of z_i and the fitted values $\hat{z}_i$, obtained from the regression, we have $\phi_X(\lambda_j)$ that can be expressed in another way as:

$$\phi_X(\lambda_j) = Corr(w_{ijt}, \widehat{w_{ijt}}) = \frac{Cov(w_{ijt}, \widehat{w_{ijt}})}{\sqrt{Var(w_{ijt})}\sqrt{Var(\widehat{w_{ijt}})}} \tag{73.3}$$

where w_{ijt} is chosen to maximise $\phi_X(\lambda_j)$, $\widehat{w_{ij}}$ are the fitted values in the regression of w_{ij} on the rest of wavelet coefficients at scale λ_j.

Cross-Recurrence Quantification Analysis (CRQA)

Cross-recurrence quantification analysis (CRQA) [16] is a method to measure the manner in which two nonlinear systems exhibits similar activity at the same time by employing the process of delaying and embedding. It also sets a threshold to see if the proximity between the time series qualifies as recurrence or not.

$$\begin{aligned} RR &= \frac{1}{N^2} \Sigma_{i,j=1}^{N} R_i^{m,\varepsilon} \\ DET &= \frac{\Sigma_{l=l_{min}}^{N} lP^{\varepsilon}(l)}{\Sigma_{i,j=1}^{N} R_i^{m,\varepsilon}} \\ LAM &= \frac{\Sigma_{v=v_{min}}^{N} vP(v)}{\Sigma_{v=1}^{N} vP(v)} \end{aligned} \tag{73.4}$$

We have used CRQA method for examining cross-recurrences at different level and quantification of the recurrence plots (RPs) to extract similar information to cross-correlation. These patterns are captured in cross-recurrence plots (CRPs), which is the graphical representation of the cross-recurrence matrix (CRM) of distance elements within a cut-off limit. We have computed three complexity measures from the cross-recurrence plots. They are recurrence rate (RR), percentage determinism (DET), and laminarity (LAM) (see Eq. 73.4). RR is the ratio of all recurrent states to all possible states and is the probability of recurrence of a special state. RR is used to compute the correlation dimension of data. DET is the ratio of recurrence points forming diagonal structures to all recurrence points. DET measures the percentage of recurrent points forming line segments that are parallel to the main diagonal. LAM is related to a number of laminar phases in the system (intermittency). It is estimated from some recurrence points which form vertical lines.

Vector Auto-Regression (VAR)

VAR [20] estimation is a recursive process to analyse the behaviour of multivariate time series. The efficiency of the model is dependent on a number of variables (K) and lags (p) used in the equations, where lag length is typically determined using *Akaike information criteria* (AIC). The model is estimated by using the principle of least squares (PLS). We have chosen reduced form of VAR model where a single variable is expressed as a linear function of own values and as well as others. In the reduced form, every time series variable is composed of two components: one is deterministic (zero, constant or linear trend) and another is stochastic (zero mean and constant variance). To analyse possible causal relationships in the system, we have used *Granger causality* tests [10] and *impulse response functions* (IRFs) [20]. If a variable (x) can be used to predict another variable (y), then x is said to Granger-cause y. IRFs are to analyse how the change in one variable can influence the other. In VAR, changes in one variable are induced by nonzero residuals. In the reduced form of VAR, shocks enter through the residual vector. A VAR(p) model, where p is the lag, could be represented as follows: Let $Y_t = (y_{1t}, y_{2t}, \ldots y_{nt})'$ be an $n \times 1$ multivariate time series. A VAR(p) model could be written as:

$$Y_t = C + \prod_1 Y_{t-1} + \prod_2 Y_{t-2} + \ldots \prod_p Y_{t-p} + \varepsilon_t \tag{73.5}$$

where $\prod_i$ are $n \times n$ coefficient matrices and ε_t is an $n \times 1$ unobservable white noise process with zero mean and time invariant variance–covariance matrix $\sum_{n \times n}$.

Multivariate Adaptive Regression Splines (MARS)

MARS [9] is an additive, nonparametric, regression model with a combination of piecewise linear basis functions which can select or transform and detect the variables relationship or degree of interaction in the model with low overheads. It can also detect nonlinearity among variables. MARS does not need any pre-assumption about the functional relationship between the input and the outputs. MARS determine the basis functions from regression data using two-sided truncated functions. MARS implement generalised cross-validation (GCV) process to remove redundant basis functions. Both the basis functions and generic MARS equation (f(X)) are defined in Eq. 73.6.

$$
\begin{aligned}
[-(x-t)]_+^q &= \begin{cases} (t-x)^q & if\ x < t \\ 0 & otherwise \end{cases} \\
[(x-t)]_+^q &= \begin{cases} (x-t)^q & if\ x \geq t \\ 0 & otherwise \end{cases} \\
f(X) &= \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(x)
\end{aligned}
\tag{73.6}
$$

where q is the order of the spline, t is the knot, α_0 is an intercept parameter, α_m is coefficient, B_m is a mth spline basis function, x is the input, and M are the numbers of spline function.

Results and Discussion

In DCs, it is always difficult to collect the historical workload traces over an extended period. For our experiments, we have collected (not readily available as trace dump) a week-long (last week of July 2016) data with six-minute granularity from the Wikimedia Grid.[1] We have used R statistical language[2] for all experiments.

Hurst Exponent Analysis: To analyse the time-varying persistent behaviour and antipersistent behaviour, we estimate Hurst exponent with an overlapping window of length 512 on seven days data point with six-minute granularity. The results are shown in Fig. 73.1. It is evident from the figure that CPU is showing persistent behaviour. There is fluctuation in Hurst coefficient values. However, it remains greater than 0.5 ($H > 0.5$) throughout the period of analysis. From the graph, it could be seen that for RAM, there is an oscillation between persistent behaviour and antipersistent behaviour with increasing frequency. Compared to CPU, RAM has

[1] https://www.wikimedia.org/.

[2] https://www.r-project.org/.

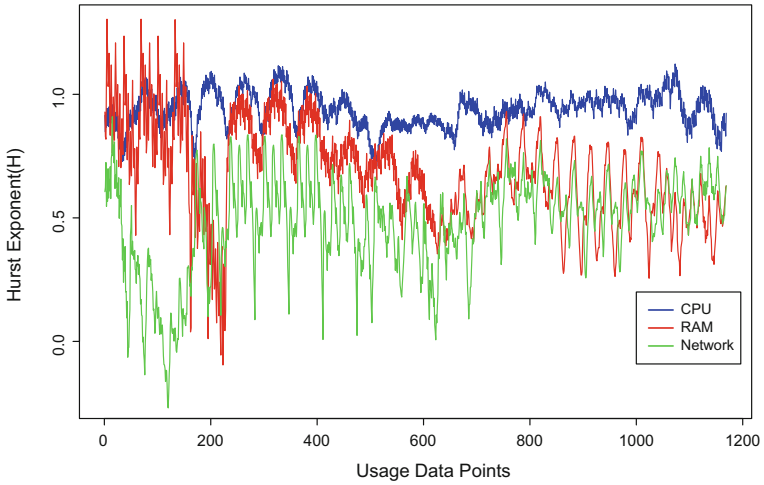


Fig. 73.1 Hurst estimation of resources over the time

shown a higher persistent behaviour ($H > 0.5$) for the initial period and followed by a drop around point 200 and demonstrate evidence of antipersistence ($H < 0.5$). Here, RAM usage is switched between two states, a state of relatively high usage, followed by a relatively low usage. It is followed by persistent behaviour up to the point 580, where antipersistent behaviour is observed. Network displays a persistent behaviour at the beginning. Later it switches to antipersistent state and also shows an oscillation between the states. We can conclude that Network is less persistent compared to CPU and exhibits antipersistence compared to RAM.

Wavelet Analysis: We have used WMC to see how CPU, RAM, and Network are related to each other across different timescales. From the result of WMC shown in Table 73.1, it is evident that the variables are not highly correlated with each other up to scale five (22–129 h), then registers a sudden increase across scales 7 and 8. It shows that the variables are moderately correlated at low scales (high frequency), and highly correlated at high scales (daily frequency and above). It could be said that the three resource components move together, in the long run, implying that the usage of one resource is dependent upon the usage of others. We establish evidence of persistence and co-movement among the data sets. Logically, the next step would be to see if there are any functional relationships between the DC resources and to do that we employed methods such as CRQA and VAR (including Granger causality).

CRQA Analysis: We use CRQA to analyse how the fluctuations in one series affect other. We estimated three complexity measures (namely RR, DET, and LAM) and presented the results in Table 73.2. RR is a measure for regularity obtained by estimating the probability of occurrence of similar states in two-time series from the cross-RP. We extracted RR measures across the diagonals, and for doing so, we set embedding dimension to one and the threshold to zero. Similarly, DET gives

Table 73.1 Experimental results of WMC

Scale	WMC	Lower CI	Upper CI
1	0.1147423	0.0474679	0.1809804
2	0.0828632	−0.0129255	0.1771449
3	0.0863950	−0.0495753	0.2192212
4	0.1284954	−0.0647649	0.3124653
5	0.0790983	−0.1980777	0.3445610
6	0.1126099	−0.2872728	0.4790652
7	0.8389531	0.5355057	0.9505472
8	0.9997307	0.9974145	0.9999720

[Note: CI means confidence interval]

Table 73.2 Experimental results of CRQA

	RAM→Netw	CPU→RAM	CPU→Netw
RR	0.4832	0.046	0.042
DET	0.974	0.2401	0.2457
LAM	10.3848	2.1851	2.1881

a measure of predictable or deterministic structure that may exist between the two series. From Table 73.2, it can be seen that the DET value is highest between RAM and Network, followed by CPU and Network. Finally, LAM shows the time interval during which both series are stable compared to the sudden activity. LAM value and RR value also show a similar trend indicating that the fluctuations in RAM usage and Network activity are interconnected. Therefore, we have shown how CRQA analyses CPU, RAM, and Network, and also their influence on each other. Now, to detect any functional relationships between these variables, we attempt VAR and Granger causality analysis.

VAR Analysis: We estimate the VAR model with the lag length 15. First, we analyse the IRF for each variable over a horizon of 30 periods and Fig. 73.2 presents the IRF result. From IRF analysis, it is visible that the system is stable for each of the three variables because the shocks die out after certain time horizons. For CPU, a shock affects both CPU and Network. Similarly, for RAM, the shock affects RAM and Network as well. However, a unit shock in Network only affects Network and the impact on other variables is negligible.

To identify the causal relationship between the variables, we employ Granger causality analysis. However, as the WMC indicated the presence of high correlation among values, in the long run, we employed a *wavelet augmented causality testing* procedure to see the causal behaviour over multiple time horizons. In this process first, we decompose the three series using an MODWT with a least asymmetric (LA) wavelet filter of length eight. Next, we apply a nonparametric version of Granger causality pairwise for each decomposed variable at each level. The results are shown

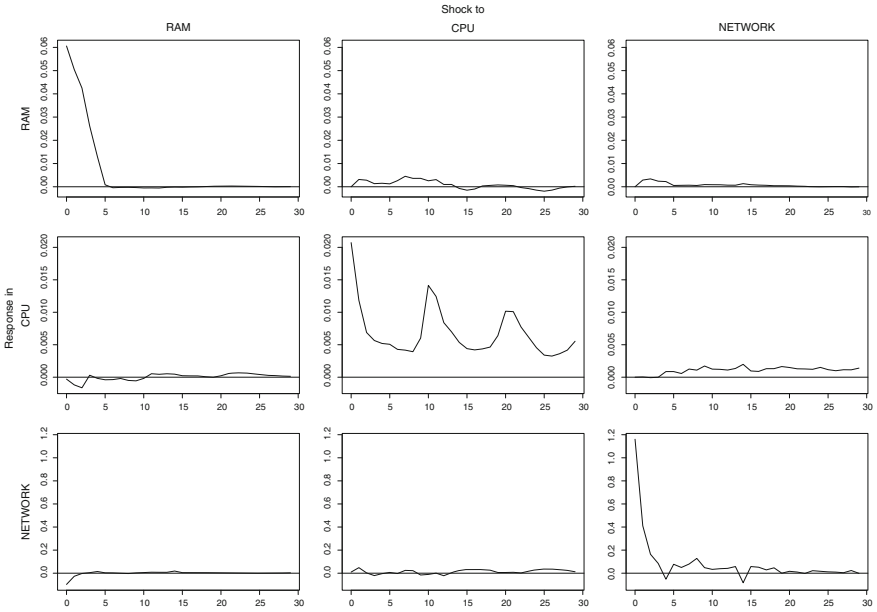


Fig. 73.2 Impulse response function analysis

Table 73.3 Wavelet-based granger causality result

Level	D1	D2	D3	D4	D5	D6	D7	D8
CPU-NW	−0.181 (0.571)	0.609 (0.728)	−0.397 (0.654)	−0.027 (0.510)	−0.21 (0.583)	0.056 (0.477)	1.72 (0.042)	0.434 (0.668)
NW-CPU	−2.305 (0.989)	−1.49 (0.931)	−0.992 (0.839)	0.702 (0.241)	0.048 (0.480)	1.192 (0.116)	3.595 (0.000)	2.702 (0.003)
CPU-RAM	1.042 (0.148)	−1.004 (0.842)	−0.258 (0.601)	−0.95 (0.828)	0.875 (0.190)	−1.102 (0.864)	−0.784 (0.783)	1.11 (0.133)
RAM-CPU	1.067 (0.143)	1.329 (0.091)	0.902 (0.183)	0.18 (0.428)	−0.891 (0.813)	−3.857 (0.999)	0.179 (0.429)	1.573 (0.057)
NW-RAM	−0.868 (0.807)	−1.005 (0.842)	−1.01 (0.843)	−1.01 (0.843)	−3.754 (0.999)	−1.119 (0.868)	−0.205 (0.581)	−1.787 (0.962)
RAM-NW	−1.8 (0.964)	−1.965 (0.975)	−2.316 (0.989)	−2.795 (0.997)	−1.41 (0.920)	−4.522 (1.000)	−4.561 1.000)	−4.757 (1.000)

in Table 73.3. From the results, it is evident that there exists *bidirectional causality* between CPU and Network at level seven (D7). Further, Network influences CPU at level eight (D8). In all other cases, the null of non-Granger causality is accepted for all the variables, as evident from the p value. From the preceding analysis, it is also apparent that while the variables influence each other, a shock in any of the constituent parts (CPU, RAM, Network) will mostly be carried by the variables. So, here we found the existence of the causal relationship between the variables. Finally,

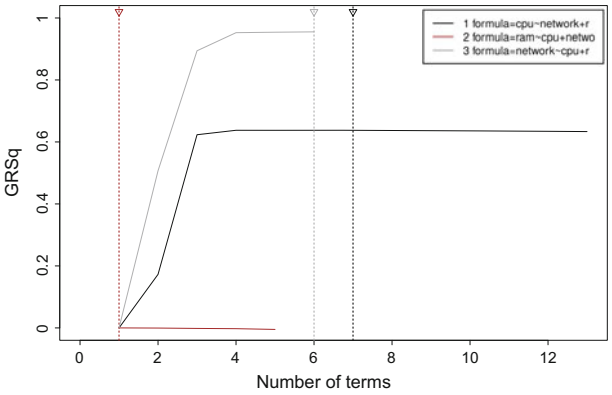


Fig. 73.3 Model comparison for all three MARS models

Table 73.4 MARS results: relationship measurement between the variables

Netw = f(CPU, RAM)		CPU = f(RAM, Netw)	
Parameters	Coefficient	Parameters	Coefficient
Intercept	16.6	Intercept	1.98
h(0.918018-cpu)	−1.4	h(netw-9.96054)	0.27
h(cpu-0.918018)	−15	h(netw-10.2148)	−0.48
h(13.4-ram)	−3.4	h(16.1488-netw)	−0.198
h(ram-13.4)	−548.1	h(netw-16.1488)	0.328
h(ram-13.4116)	549.9	h(netw-16.4498)	−0.15
–	–	h(ram-13.4221)	1.789
R^2	0.956	R^2	0.643

our final step is to determine the best functional relationship that could represent all the variables under study.

MARS Analysis: For MARS, we have estimated three models (`{x = CPU, y = (RAM, Netw)}`, `{x = RAM, y = (CPU, Netw)}`, and `{x = Netw, y = (CPU, RAM)}`) (see Fig. 73.3), and the results are shown in Table 73.4. From the estimated R^2 values, it is evident that third model `{x = Netw, y = (CPU, RAM)}` represents the best (up to 95.6%) relationship between the variables. It shows that the fluctuations in Network could be attributed largely to the fluctuations in CPU and RAM usage. Third model is followed by first model `{x = CPU, y = (RAM, Netw)}`, which gives a fit of about 64.3%. However, second model completely fails (R^2 is equal to 0). The GRSq is a measure of the predictive ability of the model, which is calculated from the GCV of the model. The GRSq value approaches 0.9 at sixth term of model three (see Fig. 73.3). For model two, the GRSq value approaches a maximum of 0.6 with seventh term in the model which is

also a good result. Hence, from Fig. 73.3 it can be seen that model three shows the best functional relationship between the resource components.

Applicability of this Analysis: Optimisation of power consumption in DC is highly important, and it can be achieved with the help of proper knowledge of usage pattern. Efficient resource allocation techniques enable us to optimise the resource allocation based on incoming task's requirement and the available resources. Knowing the insight of temporal relationship, we can scale up or down the computing resources for better performance. Our statistical analysis can be added on top of the resource allocation strategy which will help to tune the allocated resources further by learning the real-time behaviour and may also support the over-subscription. Over-subscription is an interesting approach to increase the system utilisation [22]. Analysing the statistical properties of different resource components as well as confirming the functional relationship between the variables can be beneficial to build an effective forecast model for the workload.

Conclusion

We have attempted statistically to analyse the relationship between computing resources at various time intervals without knowing the characteristics of running applications. As a first step, we analyse the persistence of resources captured in the traces by Hurst exponent. Further, we employed a wavelet-based measure and found that the series are weakly correlated in short-run, but they are strongly correlated in the long-run. To analyse further, we employed IRF analysis and found how a change in one variable influences others. Furthermore, CRQA is used to investigate how the fluctuations in one series affect other. Finally, in the last stage, we employed MARS to find the best relationship that could explain these variables. To summarise all the results, we found that change in Network is dependent upon the change in CPU and RAM. We also found that wavelet-based methods outperform others to extract long-term relations from the traces.

References

1. Abry, P., Veitch, D.: Wavelet analysis of long-range-dependent traffic. IEEE transactions on information theory 44(1), 2–15 (1998)
2. Barroso, L.A., Hölzle, U.: The case for energy-proportional computing (2007)
3. Chui, C.K.: An introduction to wavelets, vol. 1. Academic press (2014)
4. Ersoz, D., Yousif, M.S., Das, C.R.: Characterizing network traffic in a cluster-based, multi-tier data center. In: 27th International Conference on Distributed Computing Systems (ICDCS'07). pp. 59–59. IEEE (2007)
5. Fan, X., Weber, W.D., Barroso, L.A.: Power provisioning for a warehouse-sized computer. In: ACM SIGARCH Computer Architecture News. vol. 35, pp. 13–23. ACM (2007)

6. Feldmann, A., Gilbert, A.C., Willinger, W.: Data networks as cascades: Investigating the multifractal nature of internet wan traffic. In: ACM SIGCOMM Computer Communication Review. vol. 28, pp. 42–55. ACM (1998)
7. Fernández-Macho, J.: Wavelet multiple correlation and cross-correlation: a multiscale analysis of eurozone stock markets. Physica A: Statistical Mechanics and its Applications 391(4), 1097–1104 (2012)
8. Filani, D., He, J., Gao, S., Rajappa, M., Kumar, A., Shah, P., Nagappan, R.: Dynamic data center power management: Trends, issues, and solutions. Intel Technology Journal 12(1) (2008)
9. Friedman, J.H.: Multivariate adaptive regression splines. The annals of statistics pp. 1–67 (1991)
10. Granger, C.W.: Some recent development in a concept of causality. Journal of econometrics 39(1), 199–211 (1988)
11. Greenberg, A., Hamilton, J., Maltz, D.A., Patel, P.: The cost of a cloud: research problems in data center networks. ACM SIGCOMM computer communication review 39(1), 68–73 (2008)
12. Hurst, H.E., Black, R.P., Simaika, Y.: Long-term storage: an experimental study. Constable (1965)
13. Jensen, M.J.: Using wavelets to obtain a consistent ordinary least squares estimator of the long-memory parameter (1999)
14. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of ethernet traffic. In: ACM SIGCOMM Computer Communication Review. vol. 23, pp. 183–193. ACM (1993)
15. Li, H.: Workload dynamics on clusters and grids. The Journal of Supercomputing 47(1), 1–20 (2009)
16. Marwan, N., Kurths, J.: Nonlinear analysis of bivariate data with cross recurrence plots. Physics Letters A 302(5), 299–307 (2002)
17. Oikonomakos, M., Christodoulopoulos, K., Varvarigos, E.: Profiling computation jobs in grid systems. In: Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid'07). pp. 197–204. IEEE (2007)
18. Riedi, R.H., Crouse, M.S., Ribeiro, V.J., Baraniuk, R.G.: A multifractal wavelet model with application to network traffic. IEEE Transactions on Information Theory 45(3), 992–1018 (1999)
19. Schroeder, M.R.: Fractals, chaos, power laws: Minutes from an infinite paradise. Courier Corporation (2012)
20. Sims, C.A.: Macroeconomics and reality. Econometrica: Journal of the Econometric Society pp. 1–48 (1980)
21. Tian, X., Wu, H., Ji, C.: A unified framework for understanding network traffic using independent wavelet models. In: INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE. vol. 1, pp. 446–454. IEEE (2002)
22. Urgaonkar, B., Shenoy, P., Roscoe, T.: Resource overbooking and application profiling in shared hosting platforms. ACM SIGOPS Operating Systems Review 36(SI), 239–254 (2002)

Chapter 74
Classification of dsDNA Virus Through Sequence Patterns

Uddalak Mitra and Balaram Bhattacharyya

Abstract We attempt an in silico similarity analysis of virus genomes using feature mining based on interval distribution of k-words over the sequence. Patterns extracted offer an important measure for their classification of genomes. A set of highly diverge viruses from the family of Phycodnaviridae with large genome size (160–560 KB) and icosahedral shaped has been used to test the proficiency of the proposed method. We also observe that the viruses are segregated based on their respective host organism and their ecosystem. The finding indicates that respective host environment delivers unique selective pressure in genome organization.

Keywords Similarity mining · Giant DNA virus segregation · Evolution and dsDNA virus · K-word pattern analysis

Introduction

Virus enters living cell and takeovers protein-synthesis mechanisms of the cell by contaminating the DNA sequence of the cell using its genetic materials to replicate it into many folds, destroying the cell itself in the course. Identification with characteristic features of genome sequence of infecting virus is a necessity for its control and medical treatment. Subtyping virus through the process of classification is one of the fundamental steps for its detection. Conventionally, viruses are classified by phenotype characteristics like morphology, host types, cause of disease. Such classification often yields ambiguous results as different viruses cause the same disease or look alike morphologically. Sequence-based classification paves to be a better alternative.

Baltimore classification [1] based on genome properties like DNA or RNA types, single or double strands, sense or antisense and type of replication. Although

U. Mitra · B. Bhattacharyya (✉)
Department of Computer and System Sciences, Visva-Bharati University, Santiniketan 731235, India
e-mail: balaramb@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_74

it generates a type of classification based on overall features of genetic materials of the viruses, the study of in-depth genome topographies under a class would have been achieved through similarity studies on the basis of internal features, or more specifically, arrangements of nucleic acids among respective sequences.

The family Phycodnaviridae [2] is an assembly of double-stranded DNA viruses that infect marine or fresh water algae. The taxonomy of this family is based on host range. The genus chloroviruses infect chlorella-like green algae in terrestrial water and other genera infect marine green and brown algae [3]. Interestingly, more than a half of the predicted protein product of these viruses resembles prokaryotic and eukaryotic genes with known functions [4, 5]. In addition, virus-encoded proteins are either the smallest or smaller proteins of their class. Accumulating evidence also indicates that these viruses have a long evolutionary history perhaps dating back to the time when eukaryotic diverge from prokaryotic, making these viruses an important stains to study the ecology and evolution in marine and terrestrial environment [6].

The viruses in the family of Phycodnaviridae are genetically diverse but morphologically similar having large genome sizes ranging from 160 to 560 kb. It is reported that among over 1000 genes only 14 homologous genes are common among the genera [6], making the study of genome similarity difficult to trace. The information theoretic approach using the concept of super information reported in [7] is a computational attempt to trace genomic similarity eligible for segregation of the chloroviruses. It emphasizes local variation of entropy over the sequence. However, the studies on global similarity have profound impact in identifying characteristics among the large variety of target objects. In this study, we track down global similarity over the sequences of dsDNA viruses through interval distribution of all possible k-words of genome sequences.

Contribution of the Present Work

We conduct an in-depth analysis of arrangements of nucleotide bases comprising the virus genomes of the family Phycodnaviridae. The inherent information extracted from interval distribution of all possible k-words of a genomic sequence, at optimal k value is used to construct representative vector of the genome. Finally, we develop a dissimilarity matrix using Bhattacharyya distance [8] to study the degree of dissimilarity between pairs of sequences using feature vector representations, followed by construction of similarity relationship among the candidate genomes using N-J [9] hierarchical clustering algorithm.

Formalizing the Method

Let $\sum = \{A, T, G, C\}$ be an alphabet and S is a genome sequence formed by series of 'n' symbols from $\sum$. A subsequence of 'k' ($\leq n$) of consecutive symbols in S is designated as a k-word.

Let denote the jth k-word by w_j that occurs at locations l_i^j, i = 1(1)m, m being the maximum number of occurrence. We define the set of intervals between each successive locations of that jth k-word as:

$$\begin{aligned} d_i^j &= l_{i+1}^j - l_i^j, \text{ for } i < m \\ &= |S| - l_m^j + l_1^j \text{ for } i = m \end{aligned} \tag{74.1}$$

An obvious feature of the set is that sum of its elements equals the length of the sequence

$$\sum_{i=1}^{m} di = |S| \tag{74.2}$$

If $p_i = d_i / \sum_{i=1}^{m} d_i$, i = 1(1)m, be the probability of occurring the jth k-word, w_j at intervals d_i^j, then the statistical parameter h_j, summarized the interval distribution of the jth k-word, defined according to Shannon's entropy as

$$h_j = \sum_{i=1}^{|D|} p_i * \log(1/p_i) \tag{74.3}$$

Entropies of all possible k-words form a numeric vector of size 4^k. Upon normalization, we obtain the representation vector of the given sequence S for word size k

$$H = (h_1, h_2, \ldots, h_r)./ \sum_{j=1}^{r} h_j \tag{74.4}$$

Where./ denotes element wise division.

H is a numeric vector of size $r = 4^k$, each element represents normalized entropy of the interval distribution of a k-word of the sequence at a given k value.

Distance Measure Between Genomes Based on Parameter of Interval Distributions

To measure the distance between two genome sequences (S_i and S_j), we use Bhattacharyya distance [8] between given pair of genome sequences, as follows

$$d_{ij} = -\log\left(\sum\left(H_{irh} * H_{jrh}\right)^{1/2}\right) \quad (74.5)$$

'r' stands for interval distribution of a particular k-word. The terms H_{irh} and H_{jrh} mean normalized entropy of rth interval distribution in ith genome and jth genome, respectively, where r = 1, …, 4^k. In case of absence of a k-word or single occurrence, we assign zero to the parameter for that k-word.

Workflow of the Method

Input: sequences $\{S_1, S_2, \ldots, S_{N,}\ k\}$

Output: Distance matrix d(N × N) : each element holds pair-wise similarity

Steps:

1. For each sequence, search and locate each k-word of word size k.
1.1 For each k-word, use Eq. (74.1) to calculate intervals.
1.2 For each k-word, use Eqs. (74.2) and (74.3) to calculate entropy of the interval distribution.
2. For each sequence, construct 4^k component numeric vector using Eq. (74.4).
3. For any two sequences, use Eq. (74.5) to calculate distance between the two sequences.
4. Return {d}.

Time Complexity of the Algorithm

The proposed algorithm makes feature vector of a DNA sequence in $O(n + 4^k \times 4m)$ time, where n is the length of the DNA sequence, k is the optimal word length, and m is the average number occurrence of a word in the sequence. If each word is equal-probable to occur, the average word count can be given by $m = n/4^k$. Thus, making feature vector requires time $O(n + 4^k \times 4m)$, that is, $O(n + 4^k \times 4n/4^k) = O(5n)$. Thus, the algorithm possesses linear time complexity.

Results and Discussion

The virus genomes belonging to the family of Phycodnaviridae that we include in our study are summarized in Table 74.1. The selected viruses are suitable for analysis because the genomes are highly diverging in respect to gene content and homology but having morphologic similarity. We apply the proposed method to devise dissimilarity among the genome sequences based on their interval

distribution of k-words. We attempt to trace patterns of similarity or dissimilarity prevalent in the test sequences that would be evident enough for their classification.

Viruses in the family Phycodnaviridae replicate in host-specific way [6]. Virus EsV1 has a specific host Ectocarpus siliculosus, a type of marine brown algae. The viruses PBCV1, NY2A, AR158, KSB1, NEJV4, and NYs1 exclusively replicate in Chlorella Species NC64A. The hosts for the other Chlorella viruses are related species of the algae. The annotated 415 KB genome of PBCV1 shows 366 protein-coding genes, 11 t-RNA genes, 90% coding regions, and 39% (C + G) content. In addition the genome consists of introns, specifically self-splicing, spliceosomal-processed introns and small t-RNA introns. Virus EsV1 consists of 421 KB genomes of which 70% are coding regions with 244 predicted CDs and 51% (C + G) content. They are also expected to have common ancestor with nucleocytoplasmic large DNA viruses (NCLDVs).

Initially, we attempt to trace intervals for nucleotide base thymine (T) for the virus genomes PBCV1 and EsV1, summarized in the frequency plot of Fig. 74.1. The measured variation in term of interval distribution of 'T' of these viruses, having hosts in different environments (PBCV1 replicate in terrestrial water green algae and EsV1 replicate in marine water brown algae), is reflected in these plots. For example, the frequency of interval value '1' of nucleotide base 'T' in case of virus PBCV1 is 33791 and for EsV1 it is 21903. Moreover, all frequency distributions are positively skewed and show their local variability.

In addition, the plot also reflects local dependency property of nucleotide sequences and lower interval values occurred more frequently compared to higher interval values. Another interesting fact is that the interval distribution of the nucleotide thymine (T) in PBCV1 is more disperse compared to the virus EsV1.

Table 74.1 List of dsDNA virus from family Phycodnaviridae

Name of the virus	Host species	Size (in KB)a
AR-158	Chlorella NC6A	432
PBCV1	Chlorella NC6A	415
NY-2A	Chlorella NC6A	463
NYs1	Chlorella NC6A	437
KSB1	Chlorella NC6A	361
NE-JV-4	Chlorella NC6A	412
CVM1	Chlorella Pbi	410
FR483	Chlorella Pbi	403
MT325	ChlorellaPbi	394
ATCV1	Chlorella SAG 3.83	361
TN603	Chlorella SAG 3.83	403
EsV1	Ectocarpus siliculosus	421

a Denote windows file size to store the genomes

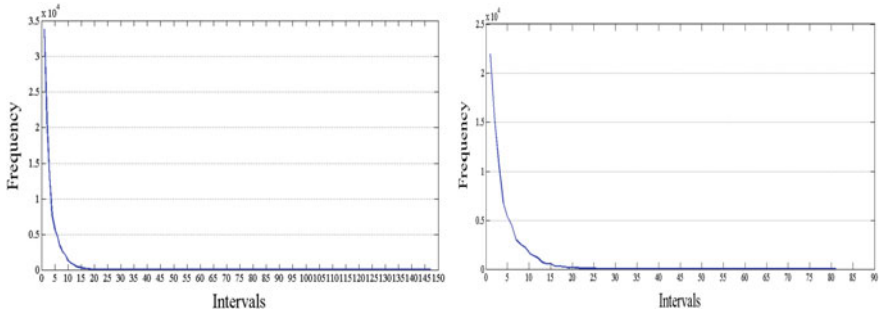


Fig. 74.1 Interval distribution of nucleotide base thymine (T), i.e., k = 1

We also inspect differences in variation of interval distributions for other nucleotide bases (C, G, T) in these viruses (not shown in the figure). Thus, we apprehend that the interval distribution of four nucleotide bases, at k value 1 together can summarized the genomic content with inherent ability to recognize homologous patterns for segregation of these viruses.

Although frequency plot can be used as a tool to trace pattern of dissimilarity among virus genomes, different sequence length generates unequal number of observation in terms of interval values. Thus, it is essential to summarized interval distributions with statistical parameters and we use Shannon entropy [10] to reflect disorder of the interval distributions.

Since a nucleotide sequence is composed of four bases, viz., adenine (A), guanine (G), thymine (T), and cytosine (C), theoretically 4^k number of k-words are possible. Thus, 4^k interval distributions corresponding to each k-word for a sequence can be computed. Thus, we conduct analysis for the virus genomes for k values ranging from 1 to 4. For each k value, we construct cladogram trees and observe that at k = 2 the virus genomes are completely segregated in host-specific way. The numeric vector representations of all the virus genomes are listed in supplementary Table S1. The Bhattacharyya distance measures between each possible virus are listed in Table 74.2.

Viruses with common host show less dissimilarity compared to other viruses with different host. For example, viruses CVM1, FR483, and MT325 possess common host Chlorella Pbi, having less dissimilarity among them, and have been grouped together (Fig. 74.2). All the viruses that replicated in Chlorella NC64A are also grouped together in a separate clade. Viruses, ATCV1 and TN603, replicated in marine water green algae SAG 3.83 place in a common clade and have been segregated from marine water brown algae virus EsV1. Moreover, viruses with host (Chlorella NC64A) in terrestrial water environment have been completely segregated from the viruses having host (Chlorella Pbi, SAG 3.83, and E. siliculosus) in marine water ecosystem.

The result suggests that the respective host environment gives rise to unique selective force on the genomic structure and may provide valuable clue to study

Table 74.2 Pair-wise Bhattacharyya distance between the virus genomes

Virus	1	2	3	4	5	6	7	8	9	10	11	12
AR-158	0											
PBCV1	0.0223	0										
NY2A	0.0002	0.0249	0									
NYs1	0.0002	0.0208	0.0003	0								
KSB1	0.0229	0.0003	0.0257	0.0215	0							
NEJV4	0.0216	0.0002	0.0240	0.0201	0.0005	0						
CVM1	0.0921	0.0629	0.0966	0.0931	0.0578	0.0641	0					
FR483	0.0999	0.0693	0.1044	0.1010	0.0641	0.0705	0.0003	0				
MT325	0.1055	0.0769	0.1101	0.1071	0.0711	0.0781	0.0012	0.0010	0			
ATCV1	0.1493	0.1487	0.1539	0.1544	0.1398	0.1490	0.0492	0.0501	0.0380	0		
TN603	0.1249	0.1208	0.1293	0.1292	0.1126	0.1215	0.0398	0.0415	0.0310	0.0022	0	
EsV1	0.2098	0.2268	0.2132	0.2164	0.2163	0.2273	0.1058	0.1068	0.0886	0.0151	0.0210	0

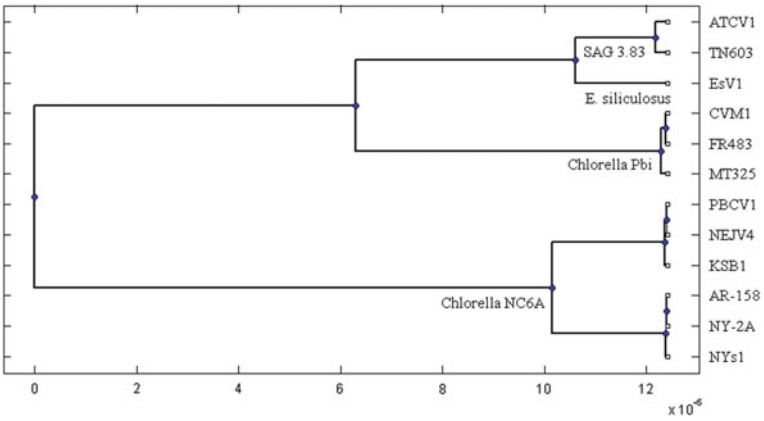


Fig. 74.2 Clustering of virus genomes annotated with their respective hosts

dynamics of ongoing changes in ecosystem and evolution history in terrestrial and marine water environment.

Conclusion

We present a similarity mining method based on interval distribution of all possible k-words for segregation of giant dsDNA virus genomes. Clustering of the genome sequences on the basis of Bhattacharyya distance fetches host specificity of these viruses, i.e., the viruses having common hosts are closer to each other in terms of arrangements of nucleic acids in respective genomes. The proposed k-words interval-based segregation is also able to completely separate viruses with hosts in different environments. Our findings are also consistent with the biological evidences existent in terrestrial and marine water ecosystems, thus extending the scope of studying the dynamics therein.

Supplementary File

See Table S1

Table S1 Numeric vector representation at k = 2 of the virus genomes used in the experiment

Seq1	0.0640	0.0635	0.0649	0.0628	0.0628	0.0583	0.0586	0.0651	0.0609	0.0630	0.0578	0.0633	0.0666	0.0611	0.0627	0.0643
Seq2	0.0648	0.0636	0.0635	0.0639	0.0632	0.0589	0.0586	0.0636	0.0612	0.0614	0.0586	0.0634	0.0659	0.0614	0.0630	0.0649
Seq3	0.0641	0.0636	0.0651	0.0627	0.0628	0.0582	0.0585	0.0652	0.0610	0.0631	0.0577	0.0633	0.0667	0.0610	0.0628	0.0642
Seq4	0.0641	0.0635	0.0650	0.0629	0.0628	0.0582	0.0585	0.0651	0.0611	0.0629	0.0578	0.0633	0.0667	0.0610	0.0627	0.0644
Seq5	0.0647	0.0635	0.0635	0.0639	0.0631	0.0590	0.0587	0.0636	0.0614	0.0614	0.0588	0.0633	0.0658	0.0615	0.0630	0.0648
Seq6	0.0647	0.0636	0.0636	0.0639	0.0632	0.0590	0.0585	0.0636	0.0612	0.0614	0.0585	0.0634	0.0659	0.0613	0.0631	0.0649
Seq7	0.0634	0.0648	0.0632	0.0613	0.0627	0.0600	0.0612	0.0635	0.0621	0.0598	0.0602	0.0646	0.0646	0.0624	0.0627	0.0634
Seq8	0.0634	0.0648	0.0632	0.0612	0.0626	0.0601	0.0613	0.0634	0.0622	0.0596	0.0603	0.0648	0.0647	0.0623	0.0626	0.0633
seq9	0.0632	0.0648	0.0632	0.0610	0.0628	0.0603	0.0615	0.0634	0.0622	0.0600	0.0605	0.0646	0.0644	0.0624	0.0626	0.0632
Seq10	0.0617	0.0639	0.0634	0.0596	0.0627	0.0613	0.0625	0.0637	0.0619	0.0630	0.0610	0.0643	0.0630	0.0623	0.0631	0.0626
Seq11	0.0622	0.0638	0.0634	0.0602	0.0628	0.0609	0.0622	0.0636	0.0618	0.0630	0.0610	0.0640	0.0631	0.0620	0.0632	0.0627
Seq12	0.0617	0.0639	0.0641	0.0584	0.0633	0.0613	0.0632	0.0639	0.0623	0.0643	0.0613	0.0638	0.0620	0.0620	0.0633	0.0614

References

1. Baltimore D (1971): Expression of animal virus genomes, Bacteriol Rev. 35 (3)(1971) 235–41
2. Van Etten, JL.: Unusual life style of giant chlorella viruses. Annu Rev Genet. **37**(2003) 153–95
3. Van Etten, JL.: Meints, RH.Giant viruses infecting algae, Annu Rev Microbiol. **53**(1999) 447–94
4. Winson V.H., Van Etten J. L., Allen M.J.: The Phycodnaviridae: the story of how tiny giants rule the world, Curr Top Microbiol Immunol,328 (2010), 1–42
5. Van Etten J. L., Lane C. L., Dunigan D. D.: DNA viruses: The really big one(giruses) Curr Top Microbiol Immunol,64 (2010), 83–99
6. Yamada T., Onimatsu H., Van Etten J. L.: Chlorella Viruses, Adv Virus Res., 66, (2006), 293–336
7. Bose R., Thiel G., Hamacher K.: Clustering of giant virus dna based on variation in local entropy, Viruses, 6, (2014), 2259–2267
8. Bhattacharyya A.: On a measure of divergence between two statistical populations defined by their probability distributions. Calcutta Mathematical Society, 35(1943) 99–109
9. Saitou N, Nei M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, (1987), 406–425
10. Shannon, Claude E.: A Mathematical Theory of Communication. Bell System Technical Journal. **27** (3): (1948). 379–423

Chapter 75
Analysis of the Firing Behavior of STN-GPe Network in Parkinson Disease

Jyotsna Singh, Phool Singh and Vikas Malik

Abstract Movement disorders are caused by the malfunctions of the basal ganglia. These disorders are dystonia and Parkinson disease. Various models have been used to display the pathophysiology of these disorders. We have chosen a conductance-based model to display the firing patterns of Parkinson disease with respect to calcium, potassium, and sodium currents. These patterns are generated in hyper direct and indirect pathway coupling neuron, subthalamic neuron. Subthalamic nucleus is one of the main nuclei involved in the origin of motor dysfunction in Parkinson disease. The cause to the generation of such activity pattern is not yet completely understood. In this paper, we are simulating the firing patterns of subthalamic nucleus in healthy and Parkinson primate and also to provide insight into the importance of different currents to improve the correlation among these two patterns.

Keywords Movement disorder · Parkinson disease · Basal ganglia · Subthalamic nucleus · Globus pallidus · Dysfunction

Introduction

Parkinson disease(PD) is one of many diseases that are collectively known as movement disorder. This disorder occurs due to the malfunction and death of neurons. Due to the malfunction of neuron, the amount of neurotransmitter known as dopamine produced in brain is decreased. Dopamine is the neurotransmitter which is involved in the control movement and coordination, hence its depletion leave a person unable

J. Singh (✉) · P. Singh
The NorthCap University, Gurgaon, Haryana, India
e-mail: singhjyotsna1@gmail.com

P. Singh
e-mail: phool.singh24@gmail.com

V. Malik
Department of Physics, JIIT Noida, Uttar Pradesh, India
e-mail: vikasm76@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_75

to control movements. The cause to the depletion of dopamine is not well known, and hence there is presently no diagnosis and cure at early stage [1].

Target region involved in the dysfunction of brain is basal ganglia (BG). BG has four main nuclei, substantia nigra (SNc), striatum, subthalamic nucleus (STN), and globus pallidus (GP). The striatum receives input directly from cortex and delivers neurotransmitter which is inhibitory in nature to the GP through different pathways (1) direct and (2) indirect. STN directly receives inputs from cortex through (3) hyper direct pathway [2]. Direct and indirect affect the basal ganglia network in opposite ways and are simultaneously involved with the control of voluntary movements [1]. It has been acknowledged by researchers in [2–5] that there is central origin to PD. Nowadays, STN is said to be mainly involved to display the Parkinson symptoms [2], but the localization to the origin is still not clear.

There are various models which depict the basal ganglia dysfunction on account of the development of Parkinson motor symptoms. The Rate Model [6–8] poised that the Parkinson motor symptoms were the result of change in mean discharge rate in basal ganglia circuit. Experimental studies [6–8] are although supported by human and non-human primates [9], but a recent study [10] failed to show in non-human primate the increase in rate of discharge in internal globus pallidus (GPi) in Parkinson disease. In addition to this, in Parkinson patients, the pallidotomy is said to be reduced rather than to increase [9–11], while there is an increase in rate of discharge in GPi in subthalamic nucleus(STN) during deep brain stimulation (DBS) with improvement in motor signs [12]. Above observations put doubt on the validity of Rate model and it led various other theories like bursting and oscillation in a particular frequency range [13]. In DBS, oscillation synchrony is observed in STN and GPi in the beta band, in a PD patient [14]. Other studies have shown that pathological STN drive in Parkinson disease modifies the electrophysiological activities in GP, which in turn disrupts the normal function of basal ganglia [15–17]. Thus, the role of neuronal activity within basal ganglia in the development of Parkinson disease remains unclear.

Above studies suggests that STN-GPe could account for the Parkinson symptoms in hyper direct and indirect pathway. Therefore, in this paper, we have analyzed the signals in STN-GPe network for exploring the dynamics of the subthalamo-globus loop subjected to various neuronal parameters. Work has been distributed in various sections. Section "Methods and Models" explains underlying Mathematical Model and Methods. Section "Simulation of Firing Patterns" explains firing behavior of model with respect to calcium, sodium and potassium currents. In section "Conclusion", we are concluding our work which describes the importance of calcium currents over sodium and potassium currents.

Methods and Models

A conductance-based model for STN and GPe has been demonstrated to display the firing behavior in PD. Each module in the pictorial representation of STN-GPe

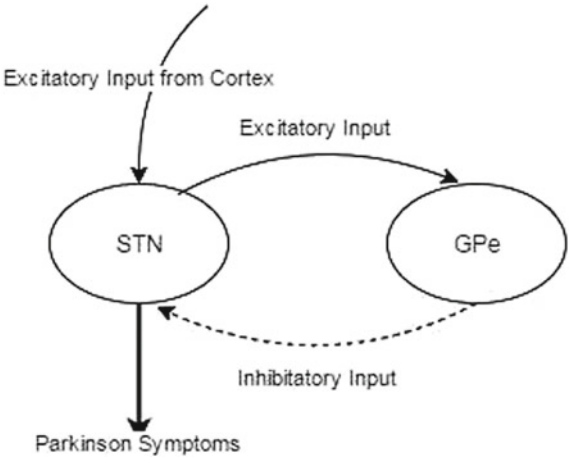


Fig. 75.1 Pictorial representation of STN-GPe coupling architecture

(Fig. 75.1) coupling architecture is represented by a single neuron. Each neuron in this model is a single compartment conductance-based model. Equations (75.1) and (75.8) form a system of differential equation. The system of differential equations along with initial condition specified in [18] has been numerically simulated in MATLAB 7.14 (i7 Intel processor, 16GB RAM machine) using ODE45. The model [18] includes a leak current (I_l), fast spike-producing potassium (I_k) and sodium currents (I_{Na}), low-threshold T-type (I_T) and high-threshold calcium currents (I_{Ca}), a calcium activated voltage independent after-hyperpolarization K^+ current (I_{AHP}), synaptic current from GPe (I_{syn}), and applied current ($I_{applied}$) [18], so that the equation governing the membrane potential V takes the form

$$C\frac{dv}{dt} = I_{applied} + (-I_l - I_k - I_{Na} - I_T - I_{Ca} - I_{AHP} - I_{syn}) \tag{75.1}$$

Different membrane currents are given by following equations:

$$I_l = g_l \cdot [V - V_l] \tag{75.2}$$

$$I_k = g_k \cdot n^4 \cdot [V - V_k] \tag{75.3}$$

$$I_{Na} = g_{Na} \cdot m_\infty^3(V) \cdot h \cdot [V - V_{Na}] \tag{75.4}$$

$$I_T = g_T \cdot a_\infty^3(V) \cdot r \cdot [V - V_{Ca}] \tag{75.5}$$

$$I_{Ca} = g_{Ca} \cdot s_\infty^2 \cdot (V) \cdot [V - V_{Ca}] \tag{75.6}$$

$$I_{AHP} = g_{AHP} \cdot \frac{[Ca]}{[Ca] + k_1} \cdot [V - V_k] \tag{75.7}$$

where k_1 is the dissociation constant of calcium-dependent AHP current. Membrane resting potential is adjusted with the help of applied current, $I_{applied}$ with the experimental data [19, 20]. Calcium current is described by the equation:

$$d\frac{[Ca]}{dt} = \epsilon[-I_{Ca} - I_T - kCa[Ca]] \tag{75.8}$$

Calcium influx is denoted by ϵ, and calcium pump rate is denoted by kCa. Gating variable and other parameters have been referenced from [21–24].

Simulation of Firing Patterns

In this experimental study, we have used MATLAB 7.14 to realize the system of differential equation (75.1) for STN-GPe network. Using this model, we have generated two spiking patterns for STN-GPe coupling network. These patterns represent the behavior of model in Non-PD and PD state. All the currents have been considered for STN neuron, and the synaptic input from GPe is considered. Behavior of spiking patterns have been analyzed by using the cross-correlation (CC) between two. CC is a measure of similarity between two series as a function of the lag of one relative to the other. We have considered the value of CC at lag = 0. The CC between A_t and B_{t+k} is called the k_{th} order cross-correlation of A and B. The sample estimate of this cross-correlation, called r_i, is calculated using the formula:

$$r_i = \frac{\sum_{j=1}^{n-i}(A_j - \overline{A})(B_{j+i} - \overline{B})}{\sqrt{\sum_1^n (A_j - \overline{A})^2 \sum_1^n (B_j - \overline{B})^2}} \tag{75.9}$$

The firing behavior of model has been studied for calcium V_{Ca}, sodium V_{Na}, and potassium V_k membrane potential.

Firing Patterns for Calcium Membrane Potential

To investigate the model assumptions about physiological properties and connectivity patterns that lead to the best explanation of (i.e., closest fit to) our in vivo electrophysiological data, we studied different spiking patterns generated by the mathematical descriptions presented above. This section describes the firing behavior of model with respect to the calcium membrane potential V_{Ca}. All the other values are kept same as mentioned in [18]. Initial value of V_{Ca} was 140 for both the cases, i.e., Non-PD and PD. In order to check the sensitivity of PD patterns with respect to V_{Ca}, calcium membrane potential has been decreased and increased gradually and different graphs have been generated. Here we are showing the graphs which are

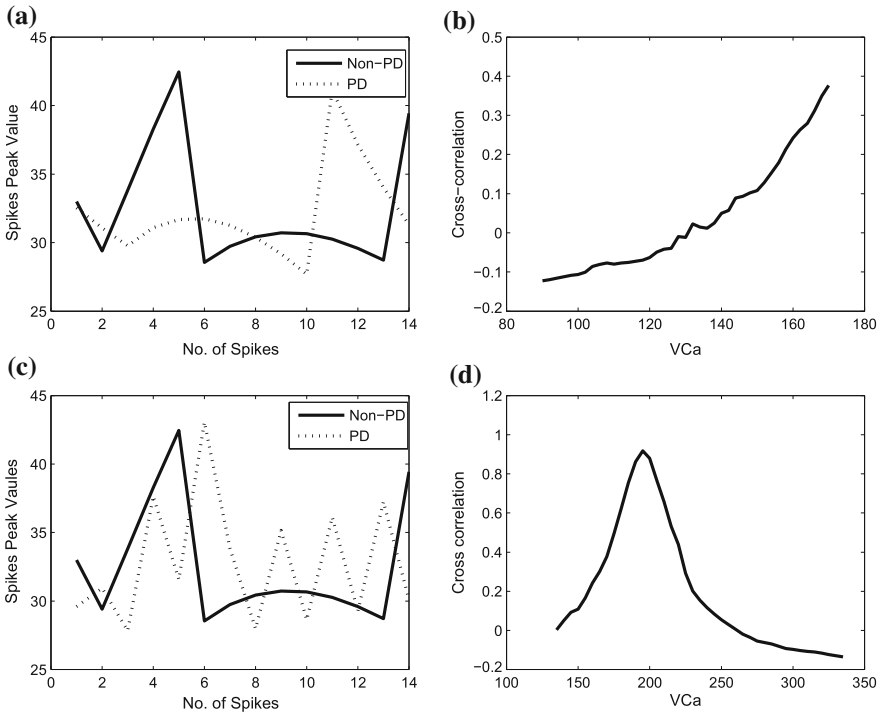


Fig. 75.2 Firing behavior for calcium membrane potential, **a** Spiking trend for $V_{Ca} = 140$, **b** Cross-correlation between Non-PD and PD patterns $V_{Ca} = 90$–170, **c** Spiking trend for $V_{Ca} = 256$, and **d** Cross-correlation between Non-PD and PD patterns $V_{Ca} = 135$–350

giving some promising outputs and results. Figure 75.2a show the spiking trend of Non-PD and PD patterns for $V_{Ca} = 140$. Figure 75.2b show the correlation among Non-PD and PD patterns for $V_{Ca} = 90$–170. Cross-correlation graph shows that the correlation is improving with the increase in calcium membrane potential. From 2a, we can see easily compare the spiking behavior of Non-PD and PD patterns. The spiking range is from 28 to 42 for both the cases but Non-PD patterns are irregular, whereas PD patterns are regular comparatively. Hence to further check the behavior of patterns, we have increased calcium membrane potential from 180 to 350 and plotted the graph. Figure 75.2c shows the spiking trend for the maximum correlation value, which is becoming smooth for PD patterns, and spikes are occurring in a limited range that is also varying gradually. Figure 75.2d clearly shows that the maximum correlation CC = 0.9 has been attained at $V_{Ca} = 256$ (approx.), and then, it starts decreasing gradually.

Firing Patterns for Sodium Membrane Potential

In order to further study the behavior of model, we have fixed other parameters [18] and checked the sensitivity with respect to sodium membrane potential V_{Na}. This is another important parameter after the calcium potential which needs to be observed in Parkinson disease. Initial value of V_{Na} was 55 for both the cases, i.e., Non-PD and PD. In order to check the sensitivity of PD patterns with respect to V_{Na}, we have increased it in the range from 50 to 95 and seen the trend of spiking. Figure 75.3a shows the spiking trend of PD patterns for $V_{Na} = 95$. Figure 75.3b shows the correlation among Non-PD and PD patterns for this range. From 3a, we can see easily compare the spiking behavior of Non-PD and PD patterns. The spiking range is from –80 to 80 for PD patterns and from 28 to 42 for Non-PD patterns. Non-PD patterns are irregular, whereas PD patterns are regular comparatively but spiking occurs in a wide range. The correlation is also varying between a wide range of sodium value. The maximum value attained in the range 55 to 95 is 0.07 (approx.)at. Further, the behavior of patterns has been analyzed in V_{Na} range from 55 to 275. Figure 75.3c shows the spiking trend, which has become slow for PD and peak are

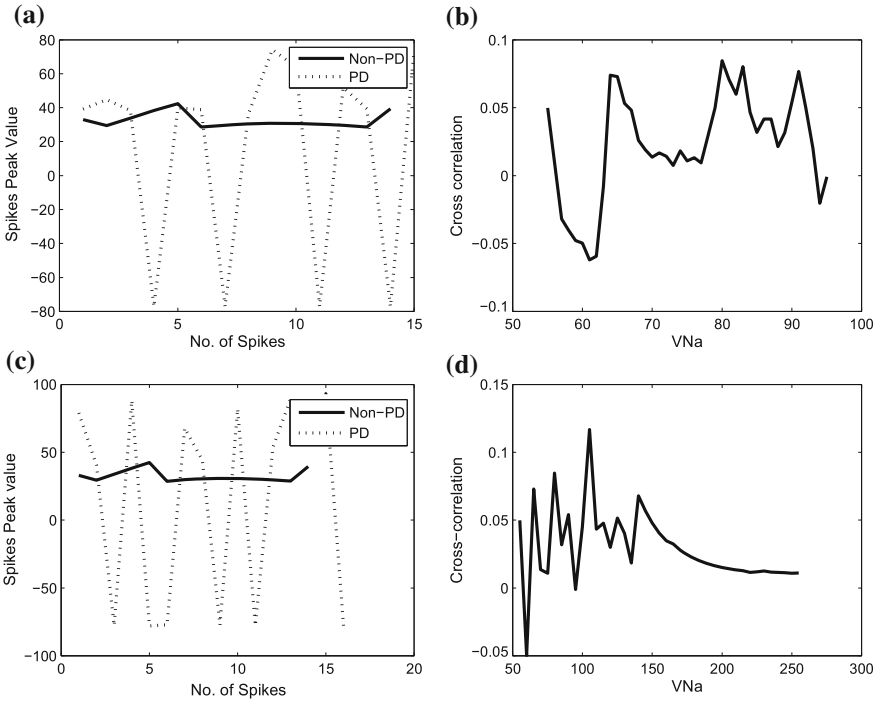


Fig. 75.3 Firing behavior for sodium membrane potential, **a** Spiking trend for $V_{Na} = 95$, **b** Cross-correlation between Non-PD and PD patterns $V_{Na} = 50$–95, **c** Spiking trend for $V_{Na} = 120$, and **d** Cross-correlation between Non-PD and PD patterns $V_{Na} = 50$–300

also occurring at the same value, i.e., spikes are occurring at a fixed peak value, i.e., 32. Figure 75.3d clearly shows that the maximum correlation CC = 0.13 has been attained at $V_{Na} = 120$ (approx.), and then, it fluctuates between the range 0–0.08.

Firing Patterns for Potassium Membrane Potential

Firing behavior of the model has been next analyzed for potassium membrane potential V_k. All the other values are kept same as mentioned in [18]. Initial value of V_k was –80 for both the cases, i.e., Non-PD and PD. In order to check the sensitivity of PD patterns with respect to V_k, we have increased it in the range from –80 to –40 and plotted the graphs. Figure 75.4a shows the spiking trend of PD patterns within this range, and Fig. 75.4b shows the correlation among Non-PD and PD patterns. From 4a, we can see easily compare the spiking behavior of Non-PD and PD patterns. Initially, the spiking is occurring at the same point, and then, the spikes are spread at a wide range but the peak value is almost same, i.e., –5. The maximum correlation

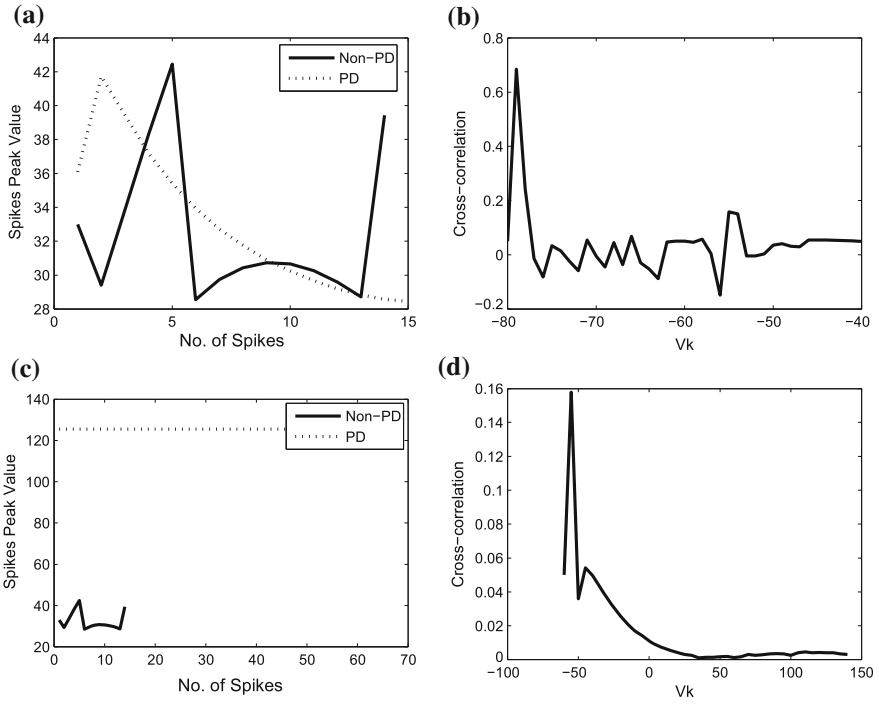


Fig. 75.4 Firing behavior for potassium membrane potential, **a** Spiking trend for $V_k = -77$, **b** Cross-correlation between Non-PD and PD patterns $V_k = -80$ to –40, **c** Spiking trend for $V_k = 150$, and **d** Cross-correlation between Non-PD and PD patterns $V_k = -100$ to 150

attained in this range of potassium potential is .7 as shown in Fig. 75.4b. Further the behavior of patterns has been analyzed in V_k range from –80 to 120. Figure 75.4c shows the spiking trend, which is now shifted to the positive region, i.e., 120. All the spikes are occurring at the same value, i.e., 120. Figure 75.4d clearly shows that the correlation value has been decreased instead of increasing with the increase in value of potassium membrane potential.

All the above simulation has been performed for time span 0–500 ms.

Conclusion

The results in the above study show that STN-GPe network is highly sensitive to calcium, potassium, and sodium currents. These two series, i.e., Non-PD and PD are not at all correlated initially. By changing one parameter and keeping others fixed, correlation patterns are also improving to some extent between Non-PD and PD patterns. The limitation of this study is that the firing behavior is studied for one parameter at a time. It does not account if we change two or more parameters simultaneously. Hence, it needs to be observed how the correlation patterns vary with variation in all the reported parameters.

Acknowledgements We thank Department of Science and Technology, Government of India for financial support vide Reference No SR/CSRI/166/2014(G) under Cognitive Science Research Initiative (CSRI) to carry out this work.

References

1. Kang G and M. Lowery M. Conductance-Based Model of the Basal Ganglia in Parkinsons Disease. *ISSC*. June 10–11 (2009), University College Dublin.
2. Santaniello S, Fiengo G, Glielmo L and M. Grill W. Basal Ganglia Modeling in Healthy and Parkinsons Disease State. I. Isolated Neurons Activity. *Proceedings of the 2007 American Control Conference*. July 11–13 (2007), USA.
3. Moran A, Bergman H, Israel Z, Bar-Gad I. Subthalamic nucleus functional organization revealed by parkinsonian neuronal oscillations and synchrony. *Brain.* 2008; 131: 33953409.
4. T. Wichmann, M.A. Kliem and J. Soares. Slow Oscillatory Discharge in the Primate Basal Ganglia, *J. Neurophysiol.* 2002; 87:1145–1148.
5. Weinberger M, Hutchison WD, Lozano AM, Hodaie M, Dostrovsky JO. Increased gamma oscillatory activity in the subthalamic nucleus during tremor in Parkinsons disease patients. *J Neurophysiol.* 2009; 101: 789802.
6. Albin R.L., Young A.B., Penney J.B., The functional anatomy of basal ganglia disorders, Trends Neurosci., 12(1989) 366–375. https://doi.org/10.1016/0166-2236(95)80020-3.
7. DeLong M.R., Primate models of movement disorders of basal ganglia origin, Trends Neurosci., 13(1990) 281–285. https://doi.org/10.1016/0166-2236(90)90110-V.
8. Miller W.C., DeLong M.R., Parkinsonian symptomatology. An anatomical and physiological analysis, Ann. N. Y. Acad. Sci., 515 (1988) 287–302. https://doi.org/10.1111/j.1749-6632.
9. Vitek J. L., Bakay Roy A. E., Freeman A., Evatt M., Green J., McDonald W., Haber M., Barnhart H., Wahlay N., Triche S., Mewes K., Chockkan V., Zhang J. and DeLong M. R., Ran-

domized trial of pallidotomy versus medical therapy for Parkinson's disease, Ann. Neurol., 53 (2003) 558–569. https://doi.org/10.1002/ana.10517.
10. Raz A., Vaadia E., Bergman H., Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl- 1,2,3,6-tetrahydropyridine vervet model of parkinsonism, J. Neurosci., 20(2000) 8559–8571. https://doi.org/10.1016/j.baga.2014.11.001.
11. Baron M.S., Vitek J.L., Bakay R.A., Green J., Kaneoke Y., Hashimoto T., Turner R.S., Woodard J.L., Cole S.A., McDonald W.M., DeLong M.R., Treatment of advanced Parkinson's disease by posterior GPi pallidotomy: 1-year results of a pilot study, Ann. Neurol., 40(1996) 355–366. https://doi.org/10.1002/ana.410400305.
12. Hashimoto T., Elder C.M., Okun M.S., Patrick S.K., Vitek J.L., Stimulation of the subthalamic nucleus changes the firing pattern of pallidal neurons, J. Neurosci., 23(2003) 1916–1923. https://doi.org/10.3389/fnsys.2011.00016.
13. Brown P., Abnormal oscillatory synchronisation in the motor system leads to impaired movement, Curr. Opin. Neurobiol., 17(2007) 656–664. https://doi.org/10.1016/j.conb.2007.12.001.
14. Devergnas A., Pittard D., Bliwise D.,Wichmann T., Relationship between oscillatory activity in the cortico-basal ganglia network and parkinsonism in MPTP-treated monkeys, Neurobiol. Dis., 68(2014) 156–166. https://doi.org/10.1016/j.nbd.2014.04.004.
15. Levy R., Hutchison W.D., Lozano A.M., Dostrovsky J.O., High-frequency synchronization of neuronal activity in the subthalamic nucleus of parkinsonian patients with limb tremor, J. Neurosci., 20(2000) 7766–7775. https://doi.org/10.1016/j.neulet.2008.06.087.
16. Brown P.: Oscillatory nature of human basal ganglia activity: relationship to the pathopysiology of Parkinsons disease, Mov. Disorders, 18(2003) 357–363. https://doi.org/10.1002/mds.10358.
17. Brown P., Abnormal oscillatory synchronizations in the motor system leads to impaired movement, Curr. Opin. NeuroBiol., 17(2007) 656–664. https://doi.org/10.1016/j.conb.2007.12.001.
18. Dovzhenok A., Rubchinsky L. L., On the Origin of Tremor in Parkinsons Disease, Plos One, 7(2012) e41598. https://doi.org/10.1371/journal.pone.0041598.
19. Terman D, Rubin JE, Yew AC, Wilson CJ (2002) Activity patterns in a model for subthalamopallidal network of basal ganglia. J Neurosci 22: 29632976.
20. Rubin JE, Terman D (2004) High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model,J Comput Neuroscience, May–Jun; 16(3):211–235.
21. Bergman H, Feingold A, Nini A, Raz A, Slovin H, Abeles M & Vaadia E. Physiological aspects of information processing in the basal ganglia of normal and parkinsonian primates. *Trends Neuroscience.* 1998; 21, 3238.
22. Hammond C, Bergman H & Brown P. Pathological synchronization in Parkinsons disease: networks, models and treatments. *Trends Neuroscience.* 2007; 30, 357364.
23. Jankovic, J. Parkinsons disease: clinical features and diagnosis. *J. Neurol. Neurosurg. Psychiatry.* 2008; 79, 368376.
24. Gelb DJ, Oliver E, Gilman S. Diagnostic Criteria for Parkinson Disease. *Arch Neurol.* 1999; 56: 3339.

Chapter 76
DCRS: A Multi-objective Protein Complex Finding Method

Pooja Sharma and Dhruba Bhattacharyya

Abstract Identifying quality protein complexes from the protein–protein interaction network is a burning issue. Many researchers have started using optimization techniques to parallely process the multiple objectives used during complex finding. We have proposed a method called DCRS, which is based on parallel optimization in order to identify quality complexes. The performance obtained using DCRS has been compared in terms of *sensitivity, positive predictive value*, and *accuracy*, and it has been found to perform better than the existing methods over the human PPI dataset.

Keywords Protein complex · Multi-objective · Optimization

Introduction

Advancement in the domain of biological information has led to the necessity of effective techniques in analyzing these data. A large number of studies are being carried out in various spheres of biological domain to understand the cellular and structural functioning of living body. Proteomics is one such study which deals with the characteristics associated with proteins and their organization at the molecular level. Proteins are polymers formed by chains of amino acids linked together. Permutation and combination of amino acids give rise to different proteins. These proteins work in coordination with each other to form groups. These groups are responsible for various functional, metabolic, and enzymatic activities in the living body. Examples

P. Sharma (✉) · D. Bhattacharyya (✉)
Department of Computer Science and Engineering, Tezpur University,
Sonitpur 784028, Assam, India
e-mail: poojasharma0303@gmail.com

D. Bhattacharyya
e-mail: dkb@tezu.ernet.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_76

may include the anaphase-promoting complex, RNA splicing complex [1], and many others. Identification of protein complexes from the molecular network of proteins is one of the prime focus of biologists. This is because identifying protein complexes can aid in understanding their structural and functional properties; they can also help in deriving the evolutionary orthology signals. Moreover, properties associated with protein complexes can also be used in drug designing these days [2].

A number of methods have been developed till date for identifying protein complexes, but they have been mostly limited to yeast dataset. However, with the increase in human PPI information, the trend has moved toward analyzing the protein complexes in humans. However, it has been found that the contemporary methods suffer from accuracy point of view when it comes to analyzing them in case of human PPI dataset. In this paper, we have introduced a complex finding method based on multi-objective optimization. The performance of our method has been compared with some state-of-the-art methods, and we found our results to be better than the existing methods in case of human PPI dataset.

Literature Review

Proteins along with their interactions are represented in the form of a graph called the protein–protein interaction network (PPIN), where the vertices represent proteins and edges between them represent interactions. The problem of complex finding is basically a clustering problem which involves identifying groups of similar proteins such that these subgroups lead to a dense structure as compared with the rest of the network. Complex finding methods rely on at least two objectives in order to find quality complexes. In the yesteryears, the trend was evaluating the objectives used during complex finding one at a time, but in the recent days, few researchers have worked toward parallel evaluation of objectives in order to get complexes. A list of some of the existing methods is reported in Table 76.1.

Proposed Method

We have modeled protein complex finding as a multi-objective optimization problem. We have used three topological measures and a biological measure. The optimization technique used for our purpose is NSGA II [13]. This is because it offers the best performance in less time due to its elitist approach. The various steps involved in our method is described here.

***A. Population initialization*:** In order to use NSGA II, there has to be a starting set of population. Each individual in the population set is referred to as a chromosome. A protein complex, $C_i = \{v_1, v_2, ...v_n\}$, where $v_i \in V$ is represented as a chromosome in our problem. The initial population can be any set of complexes obtained using

Table 76.1 Protein complex finding methods

Method	Mode of evaluation	Characteristic feature
MCODE [3]	Sequential	Cluster coefficients of highest k-cores determine the weight of each node in PPIN and thus starts the complex finding process
IPCA [4]	Sequential	Looks for subgraph structures that have small diameters, so as to form compact clusters
FAG-EC [5]	Sequential	Calculates edge weights before hand to start the cluster finding process
QCUT [6]	Sequential	Subdivides the network and chooses the one with the highest modularity for future processing
ClusterONE [7]	Sequential	Expands seed nodes based on a cohesive measure given by the weights of interactions in the graph to find complexes
CORE [8]	Sequential	It works by computing the probabilities whether two proteins can be in the same core or not. Once cores are found, they are merged to obtain the final set of cores. Then only attachments are added to the core
Neural-network-based [9]	Sequential	Uses a combination of topological and biological features to assign weights to subgraphs. It then iteratively trains a neural network for better protein complex prediction
GMFTP [10]	Sequential	It uses a probabilistic approach for generating model for complex identification using both functional and topological properties
PROCOMOSS [11]	Parallel	It uses density, bridge nodes, and semantic similarity to find protein complexes
Bandyopadhya et al. [12]	Parallel	It uses density, contribution of nodes, closeness centrality, and semantic similarity to find protein complexes

any clustering method on the PPI dataset. We have used the initial population set obtained from CNCM [14], which is our own protein complex finding method. This method uses the clustering coefficient and connectivity property of PPIN to obtain complexes.

***B. Choice of objective functions*:** Two classes of objective functions have been chosen: First one analyzes the topological properties of PPIN and the other emphasizes the biological property. We have used *density, contribution of a node into a cluster*, and *reachability contribution* for the topological features and *semantic similarity* for the biological feature. We used the framework given in [12] to set up

the choice of objective functions. However, we did not use the closeness centrality measure as used in their method [12]. This is because when we carried out an empirical study on the performance of different centrality measures over PPI dataset, we could find that none of these measures performed consistently well. Apart from this, the fact that so many centrality measures exist itself rules out each other's advantages. We have therefore used another objective function called reachability contribution along with the two topological objectives. This is because reachability index is known to capture the topological and biological characteristics of PPI networks very well. Reachability index uses the degree of the adjacent nodes, which has been used effectively in identifying the essentiality of proteins in yeast [15]. Therefore, using this as one of the objective functions helps to improve the overall complex quality. We now discuss the objective functions used for our method.

1. ***Objective functions based on topology:*** Topological characteristics of a network can very well define the subgroups present in it. Protein complexes being dense in nature [3], the first objective that we have used is density of the complex. Maximizing density function results in compact and functionally coherent complexes. For a complex C_i, density is given as

$$den_{C_i} = \frac{2 \times E_i}{|V_i| \times (|V_i| - 1)} \tag{76.1}$$

where E_i represents the number of edges among V_i vertices of complex C_i.
The next objective function is chosen as the contribution of a cluster which is the sum of the individual contribution of its member elements $cont_{v_i}$. Maximizing this function results in well-separated clusters.
The overall contribution of a cluster, *Clcont*, is given as the sum of the individual contribution, $cont_{v_i}$ of its nodes, i.e.,

$$Clcont_{C_i} = \sum_{i=1}^{k} cont_{v_i} \tag{76.2}$$

where $C_i = \{v_1, v_2 ... v_k\}$ are the cluster members and the individual node contribution is

$$cont_{v_i} = \frac{DN_{c_{v_i}}}{deg_{v_i}} \tag{76.3}$$

where $DN_{c_{v_i}}$ represents the direct neighbors of v_i within cluster C_i and deg_{v_i} is the total degree of node v_i.
The above two objective functions have been used in [12]. Another objective function used in our method depends on the reachability of the direct neighbors of a node in a cluster. Maximizing this objective function leads to formation of topologically coherent complexes. The reachability contribution of a cluster, C_i, is given as the sum of the reachability of its member nodes, Rby_{v_i}, i.e.,

$$RbyCont(C_i) = \sum_{i=1}^{k} Rby_{v_i} \tag{76.4}$$

where Rby_{v_i} is the reachability of a node v_i in a cluster C_i and is given as the ratio of the total number of links its direct neighbors have within C_i to the total number of edges in the cluster. Mathematically,

$$Rby_{v_i} = \sum_{dN} \frac{l_{wC}}{tedges_C} \tag{76.5}$$

where dN is the set of direct neighbors of node v_i within cluster C_i, l_{wC} is the number of links each node $v_x \in dN$ has within the cluster, and $tedges_C$ is the total number of edges in the cluster, C_i.

2. ***Objective function based on biological characteristics:*** Apart from these three topological functions, we have used semantic similarity as the biological measure. We have found the Wang's semantic similarity for each protein pair within a cluster. The overall semantic similarity of a cluster, C_i, is given as the sum of the semantic similarity of the combination of all its member nodes, $C_i = \{v_1, v_2 ... v_k\}$. Mathematically,

$$SS_{C_i} = \sum_{i=1, j \neq i}^{k} SS(v_i, v_j) \tag{76.6}$$

In order to get functionally enriched complexes, we need to maximize this criteria too along with the set of three topological criteria. We have named our method as *DCRS* as it uses **D**ensity, **C**ontribution, **R**eachability contribution, and **S**emantic similarity as its objective functions.

***C. Intermediate population generation*:** The next step involves intermediate population generation for a user-specified number of iterations. This is done using the non-domination sorting approach of NSGA II. Here, the objective functions are utilized to generate a set of optimal complexes. Then the genetic operators namely selection, crossover, and mutation are used. However, we do not use crossover here, as it would result in disconnected components. The selection operation is based on the traditional crowding distance method, which utilizes the objective function space to prioritize the solutions. In order to carry out mutation operation, the perturbed node is chosen only if it has higher probability, i.e., $p = 0.9$. Mutation can be done on the parent node in either of the two ways—insertion or deletion. In any case, we choose a set of random nodes, then add the direct neighbors of these nodes to get a new chromosome or delete the randomly selected nodes from the existing set to get the new chromosome.

***D. Final population generation*:** Once the number of iterations is over, it is time to produce the final set of population to be returned as solution to the user. The result set is also arranged on the basis of non-domination sorting method.

Experimental Results

We implemented our method in MATLAB running on an HP xw6600 workstation. We have performed analysis of our proposed method on HPRD dataset [16] consisting of 39,237 interactions and 9088 proteins. The semantic similarity used in our objective function is obtained using R's GOSemSim package [17]. The initial population is taken from CNCM [14], and the population to be returned in 5 iterations is given as 50. In order to analyze the performance of our proposed method with the existing ones, we have taken the help of three performance indices, viz., sensitivity (Sn), positive predictive value (PPV), and accuracy (Acc), derived from a matrix, P of the order $s \times t$, where s is the number of benchmark complexes and t number of predicted clusters [18]. Let C_{ij} be the number of common proteins between the ith benchmark complex and jth predicted cluster, P_i is the total number of proteins in the ith benchmark complex, and $C_j = \sum_{i=1}^{s} C_j$, then the three measures are given as

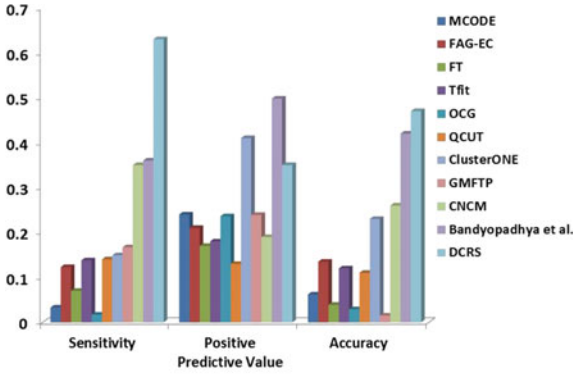
$$Sn = \frac{\sum_{i=1}^{s} max_j\{C_{ij}\}}{\sum_{i=1}^{s} P_i} \tag{76.7}$$

$$PPV = \frac{\sum_{j=1}^{t} max_i\{C_{ij}\}}{\sum_{j=1}^{t} C_j} \tag{76.8}$$

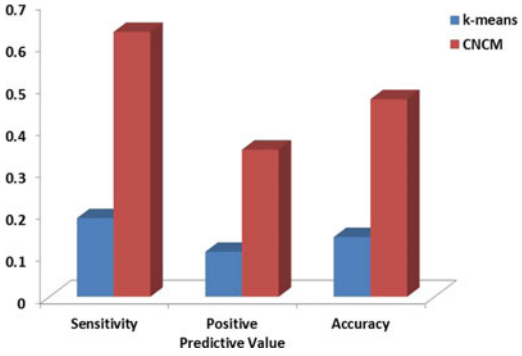
$$Acc = \sqrt{Sn \times PPV} \tag{76.9}$$

In recent years, the trend has moved toward parallel optimization of a set of objective functions to detect the set of complexes. We have compared the performance of our method with few of the well-known methods such as MCODE [3], FAG-EC [5], FT [19], TFit [19], OCG [20], QCUT [6], ClusterONE [7], GMFTP [10], CNCM [14], and Bandyopadhya et al. [12] and reported their performance in Fig. 76.1a. From Fig. 76.1a, it can be seen that DCRS is the clear-cut winner in case of sensitivity, which is around 63% which is much higher than the rest of the methods. In case of positive predictive value, Bandyopadhya et al.'s method emerges as the winner followed by ClusterONE. DCRS occupies the third rank here. However, in case of accuracy, DCRS is again at the top position beating all its competitors.

We have also shown the effect of using a well-formed cluster set during population initialization of our method as compared to any random clustering used for population initialization. An example performance comparison on HPRD dataset using k-means and CNCM method has been shown in Fig. 76.1b. It has been observed that using cluster set generated from CNCM results in better performance. A major



(a) Performance measures in terms of Sn, PPV and Acc of DCRS compared with other methods over HPRD dataset.



(b) Performance measures in terms of Sn, PPV and Acc of DCRS with k-means and CNCM as initial population.

Fig. 76.1 Performance measures of DCRS compared with other methods over HPRD dataset and variations in results using two different set of initial population

drawback of k-means method is that it requires number of clusters as input, which is usually difficult to provide. However, CNCM does not require the number of clusters as input. The performance improvement of our method using CNCM as initial population set can also be due to the fact that k-means is a partitioning-based algorithm, whereas CNCM is a graph-based method, and graph-based methods such as MCODE and IPCA [2] have already been established to be effective in PPI analysis.

Conclusion and Future Work

A multi-objective PPI complex finding method called DCRS has been proposed. The proposed method performs better than ten other existing methods over HPRD

dataset. DCRS is implemented as a tool to support the biologists in analyzing PPI data; however, the detailed discussion on the tool is out of scope of this paper. Work is also undergoing to rank the complexes based on some disease query.

References

1. Spirin Victor and Mirny Leonid A (2003) Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, National Academy of Sciences.
2. Sharma Pooja, Ahmed Hasin A, Roy Swarup, and Bhattacharyya Dhruba K. Unsupervised methods for finding protein complexes from PPI networks. (2015) *NetMAHIB*, 4(1): 1–15, Springer.
3. Bader Gary D and Hogue Christopher WV (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):1, BioMed Central.
4. Li Min, Chen Jian-er, Wang Jian-xin, Hu Bin, and Chen Gang (2008) Modifying the DPClus algorithm for identifying protein complexes based on new topological structures *BMC Bioinformatics*, 9(1):398, BioMed Central.
5. Li Min, Wang Jianxin and Chen Jian Er (2008) A fast agglomerate algorithm for mining functional modules in protein interaction networks. In *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*, volume 1, pages 3–7. IEEE
6. Ruan Jianhua and Zhang Weixiong (2008) Identifying network communities with a high resolution. *Physical Review E*, 77(1):016104, APS.
7. Nepusz Tamás, Yu Haiyuan and Paccanaro Alberto (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9(5): 471–472, Nature Publishing Group.
8. Leung CM Henry, Xiang Qian, Yiu Siu-Ming, and Chin Francis YL (2009) Predicting protein complexes from PPI data: A Core-Attachment approach. *Journal of Computational Biology*, 16(2):133–144, Mary Ann Liebert, Inc.
9. Shi Lei, Lei Xiujuan, and Zhang Aidong (2011) Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Science*, 9(1): 1–9, BioMed Central.
10. Zhang Xiao-Fei, Dai Dao-Qing, Yang Le Ou, and Yan Hong (2014) Detecting overlapping protein complexes based on a generative model with functional and topological properties. *BMC Bioinformatics*, 15(1):186, BioMed Central.
11. Mukhopadhyay Anirban, Ray Sumanta, and De Moumita (2012) Detecting protein complexes in a PPI network: A Gene Ontology based multi-objective evolutionary approach. *Molecular BioSystems*, 8(11): 3036–3048, Royal Society of Chemistry.
12. Bandyopadhyay Sanghamitra, Ray Sumanta, Mukhopadhyay Anirban, and Maulik Ujjwal (2015) A multiobjective approach for identifying protein complexes and studying their association in multiple disorders. *Algorithms for Molecular Biology*, 10(1):1, BioMed Central.
13. Deb Kalyanmoy, Pratap Amrit, Agarwal Sameer, and Meyarivan Tamt (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2): 182–197, IEEE.
14. Sharma Pooja, Ahmed Hasin A, Roy Swarup, and Bhattacharyya Dhruba K. (2015) Detecting protein complexes using connectivity among nodes in a PPI Network. *NetMAHIB*, 4(1): 1–17, Springer.
15. Hahn, Matthew W and Kern, Andrew D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks *Molecular Biology and Evolution*, 22(4):803–806, SMBE.
16. Prasad TS Keshava, Goel Renu, Kandasamy Kumaran, Keerthikumar Shivakumar, Kumar Sameer, Mathivanan Suresh, Telikicherla Deepthi, Raju Rajesh, Shafreen Beema, Venugopal

Abhilash (2009) Human Protein Reference Database-2009 update. *Nucleic Acids Research*, 37(suppl 1): D767–D772, Oxford University Press.
17. Yu Guangchuang, Li Fei, Qin Yide, Bo Xiaochen, Wu Yibo, and Wang Shengqi (2010) GOSemSim: An Rpackage for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7): 976–978, Oxford University Press.
18. Brohee Sylvain and Helden Van Jacques (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, 7(1):1, BioMed Central.
19. Gambette Philippe and Guénoche Alain (2011) Bootstrap clustering for graph partitioning *RAIRO-Operations Research*, 45(4) pages 339–352, EDP Sciences.
20. Becker Emmanuelle, Robisson Benoît and Chapple Charles E Guénoche Alain and Brun Christine (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network, *Bioinformatics*, 28(1): 84–90, Oxford University Press.

Chapter 77
Phenotype Interweaved Network of Genes using Rough Set

Arnab Sadhu and Balaram Bhattacharyya

Abstract Knowledge on inter-genetic activities under different cellular states can reveal the effect of phenotypic changes on core groups of interacting genes with a view to identify groups against disease. We attempt to present the genetic interaction under phenotypic condition as differential networks using rough set with the help of microarray data from normal and diseased cell samples. Weight of an edge between a pair of nodes is determined from associativity of the nodes with the cell state by using indiscernibility relation. A weight ranking formula is used to build the corresponding directed network. The method is tested on lung cancer data. The hubs of network modules thus obtained corroborate with previous studies and biological findings.

Keywords Phenotype interweaved network · Rough set-based network · Gene interaction network · Lung cancer

Introduction

Genes express at appropriate levels in incessant fashion for proper functioning and maintenance of cellular activities. Slightest imprecision in expressional network of a set of genes due to external perturbation, stimulation, or genetic mutation causes functional imbalance in cell coordination resulting in threats to the living organism. The early is the detection of such external disturbances, the more is the chance for recovery. This is particularly true for life-threatening diseases, particularly cancer. Microarray experiment offers an excellent opportunity to study co-expressional activity of large number of genes at an instant. Simultaneous experiments on normal and infected cell samples make it possible to analyze effect of certain disease on expressional behavior of specific genes. Identifying such genes through computational methods has become a fundamental task. Isolating genes from thousands

A. Sadhu · B. Bhattacharyya (✉)
Department of Computer and System Sciences, Visva-Bharati University, Santiniketan 731235, India
e-mail: balaramb@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_77

in microarray data becomes possible with classification and clustering techniques [1, 2]. Identifying genes showing distinctly different expressional patterns while in infected cells has been the motto. In studies based on classification, although accuracy is achieved with a small subset of differentially expressed genes, there remains differences from biological observations [3]. Affecting a small number of genes directly by the influence of disease results in instability in coordination of activities of a larger subset. Studying simultaneous interaction of a large number of genes will be of better alternative to study effect of disease on collective behavior of genes along with the phenotype state [4].

Gene–gene interaction network is extensively studied over the last decades to explore inter-regulatory mechanism of large set of genes [5–7]. Typically, correlation between the expression levels of genes under a certain phenotype is examined to assign the weights due to the hypothesis that genes which are co-regulated are controlled by similar subset of transcription factors. Independent networks formed under different phenotypes are finally compared to find changes in node connections. Usually, Pearson correlation coefficient is taken as the method for similarity measurement [6–8]. Roy et al. [5] proposed shift- and scale-based local pattern similarity measurement technique as genes are believed to be biologically associated when they have a pattern similarity.

In the current study, we aim to identify the nonlinear interactions of the genes interweaved with the diseased phenotype, which cannot be observed under correlation network. Rough set is an approach where integrity among attributes can be studied with respect to decision attribute. Here we present a differential network formation method which assigns weights of edges by a novel rough set-based approach. Further, a directed gene differential network is formed. The method is applied on a lung adenocarcinoma expression dataset. Genes which have high degree of connection with its neighborhood while scrutinized through the lens of differential network are claimed to be affected by the mutation of the cell.

Method

Analysis of gene interaction network under changing phenotype is a prospective scheme for identification of biomarker and to have knowledge on perturbation of translational mechanism of genes in the presence of disease. In an undirected co-expression network, the weights of the edges indicate expressional similarity among connecting genes, and changes in differential network can be interpreted as influence of disease under study. A directed network, however, will be of more informative.

Association of a subset of genes with the disease can be ascertained after analyzing differential network. Appreciable change in connection under differential condition for certain node is the clue. The process, however, cannot detect nonlinear relations within a gene set. Rough set [9]-based approach for differential network discovery can reveal uneven distribution in interaction patterns in the presence of disease over the normal status through the weights on edges in gene correlation network.

Rough set deals with uncertainty to find dependency among subsets of attributes. For the study of disease association disease phenotype is taken into account as the decision attribute. In the following, we present the formal structure of rough set.

Rough Set

Rough set is defined by means of topological connections with operations interior and closure, called approximations (Fig. 77.1). Let $I = (U, A)$ be the information system, where U is the non-empty finite set of all objects and A the is non-empty finite set of attributes such that for every $a \in A$, $a : U \to V_a$, V_a being the value set of a. $A = \{g_1, g_2, \ldots, g_m\}$ be the set of genes and $U = \{P_1, P_2, \ldots, P_p, P_{p+1}, \ldots, P_{p+q}\}$, the set of samples containing p numbers of normal samples and q numbers of diseased samples. Given such a dataset having discretized attribute values, rough set offers a scope to nd a subset of the original attributes having maximized information gain.

Indiscernibility Relation

Two or more objects are indiscernible in respect to an attribute subset if they are indistinguishable when only respective values for that subset of attributes are considered. Indiscernibility is mathematically defined as,

$$IND_I(C) = \{(t, t') \in U^2 | \forall a \in C, a(t) = a(t')\} \tag{77.1}$$

where $C \subset A$ (the set of all attributes).

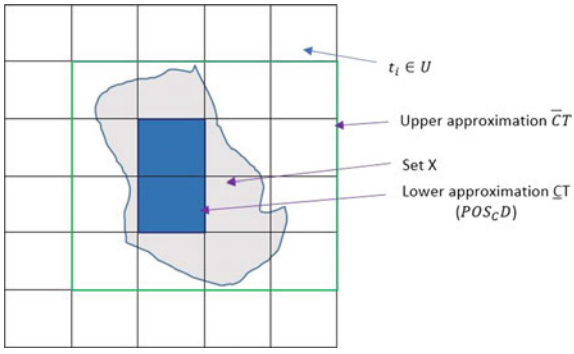


Fig. 77.1 Rough set approximation

C Positive Region of D

Let $C \subset A$ and $T \subset U$. T can be approximated by using the information contained in B by constructing the C-lower and C-upper approximations of T, denoted by $\underline{C}T$ and $\overline{C}T$, respectively, where

$$\underline{C}T = \{t|[t]_C \subset T\} \; and \; \overline{C}T = \{t|[t]_C \cap T \neq \phi\} \tag{77.2}$$

C positive region of D, $POS_C D$ is defined as,

$$POS_C D = \cup_{T \in U/D} \underline{C}T \tag{77.3}$$

Further, a set of *reducts* R are formed, where $R \subset C$ and $POS_R D = POS_C D$ and $S' = (U, R \cup D)$ is independent, i.e., all attributes in R are indispensable in S'. Intersection of all such *reducts* contains the most informative attributes.

With rough set, discretization of expression values and appropriate screening for the candidate genes are challenging.

Gene Selection upon Stable Expressional Pattern

Genes are selected by applying common subcluster mining [10] with all the normal samples. Number of cluster seeds are selected at equidistant points, and each samples are clustered individually. A gene G_i is a member of a common subcluster C_j if $G_i \in j^{\text{th}}$ cluster, for all the samples (Fig 77.2). Membership of a gene is controlled by two other thresholds, namely membership value threshold θ_1, and sample count threshold, θ_2. Furthermore, cluster indices are assigned as integral values of the genes' expression levels.

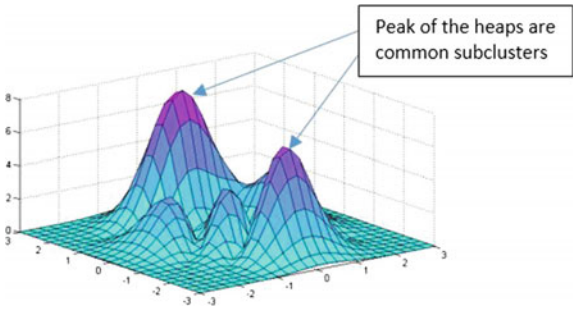


Fig. 77.2 Gene selection with CSM

Building the Weighted Undirected Network from Selected Genes

Weights between two attributes are assessed by examining similarities of objects along with the columns of selected attributes. Genes are treated as attributes. A high weight connection between two genes represents their impact on the disease state. In case of rough set, similarities of objects are evaluated by indiscernibility relation.

Let w_{ij} be the weight of connection between genes g_i and g_j, and C is the set $(\{g_i, g_j\})$. Let D be the set of phenotype state of samples. Thus,

$$w_{ij} = \gamma(C, D) = \sum X \in U/D|\underline{C}(X)|/|U| = |POS_C D|/|U| \tag{77.4}$$

$0 < w_{ij} < 1$ indicates partial dependency of D on C with degree $\gamma(C, D)$.

Retrieving the Un-weighted Directed Network

While an undirected weighted network of genes depicts the resultant pattern of a series of interaction among the participating genes, a directed networks among those genes can reveal successive states of their interactions and hence are potential source for discovery of biological phenomena at gene level. We thus proceed to generate directed networks of genes from the undirected networks with weights using rank-based method [11].

For each gene, weights of edges are ordered in descending manner, and top κ ordered edges are sustained with an updated weight which is universal, while all the remaining edges are removed from the network.

$$ranked\_network(i, j) = \begin{cases} 1, & \text{if } G_j \in \ top_i, \forall j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{77.5}$$

where top_i = top κ genes ordered by their respective edge weight with G_i.

Because of the ranked nature, the average degree of the network is between κ and 2κ. Out-degree of every node is κ, while the in-degree can be up to maximum $|G| - 1$. So the in-degree of the vertex will decide the impact of the gene on the whole network.

Algorithm

Input: Dataset $\mathbb{D} = (p + q) \times m$, normal and diseased samples together
Output: Directed network of genes

Procedure Build-network(N,D)
$C = (\{g_i, g_j\}) \forall i \in 1, 2, \ldots, p+q, j = i, i+1, \ldots, p+q;$
$\mathbb{P}_1 = \{N, N, N, \ldots p \; times\}, \mathbb{P}_2 = \{D, D, D, \ldots q \; times\};$
$\mathbb{P} = (\mathbb{P}_1, \mathbb{P}_2);$
$U = (\{\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_(p+q)\}), \; A = (\{g_1, g_2, \ldots, g_m\})$
$I = (U, A);$
Determine $\overline{C}\mathbb{P}_1$ and $\underline{C}\mathbb{P}_2$;
$\gamma(C, \mathbb{P}) = \frac{|POS_C(\mathbb{P})|}{|U|};$
$undirectedNet(g_i, g_j) = undirectedNet(g_j, g_i) = \gamma(C, \mathbb{P});$

$$directedNet(g_i, g_j) = \begin{cases} 1, & \text{if } g_j \in top(g_i), \forall j \neq i \\ 0, & \text{otherwise} \end{cases}$$

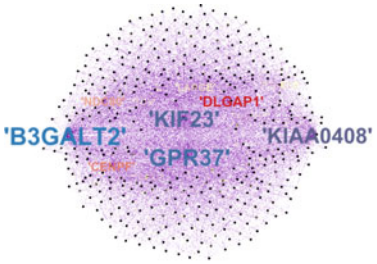
Results and Discussion

Lung Adenocarcinoma

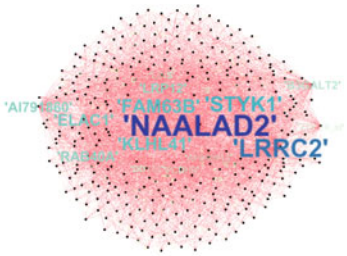
Adenocarcinoma is a common type of non-small cell lung cancer perceived by mostly persons possessing smoking habit. Landi et al. [12] made a study over 107 patient samples to observe the role of smoking in developing lung adenocarcinoma. We select the dataset to evaluate applicability of rough set on similar studies and assess scope of discovery of new knowledge. Missing values are dropped, and expression data of 21805 genes from a total of 36 samples (among them, 33 are pairs of normal and tumor samples) is taken into account. Expression levels are floating values which are replaced by cluster indices after applying CSM [10] to make them discrete.

A total of 3118 genes are selected by applying CSM with k = 10, and 10 clusters are treated separately while forming the undirected networks. Additionally, first and tenth clusters, due to their large sizes, are randomly and equally divided into two parts each. Each pair of genes is selected as attribute set C, and cell condition is categorized accordingly and designated as set D. Weight of an edge between two genes is an indication of phenotypic effect therein. Directed networks are formed with each of these undirected ones with $\kappa = 10$. Hence each gene has an out-degree connectivity of ten. However, the in-degree connectivity offers a scope of examining the impact of the corresponding gene on the differential network. Figure 77.3 shows all the networks. Labels of the nodes, i.e., genes names are shown in increasing size in order of their corresponding number of in-degree edges.

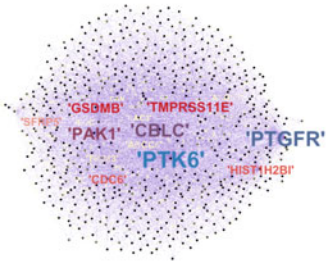
Emphasis is given in selecting genes having high degree of connectivity with the notion of its importance in inter-genetic activities. Such genes are listed in



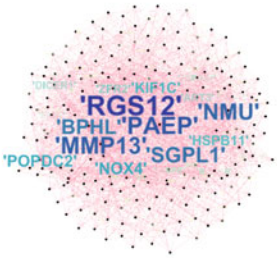
(a) First cluster



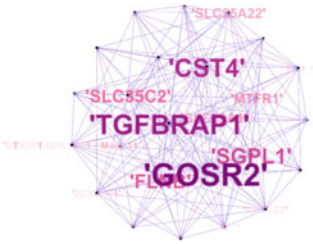
(b) First cluster



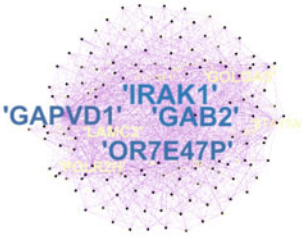
(c) Second cluster



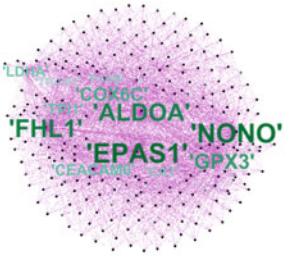
(d) Third cluster



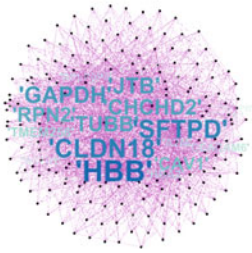
(e) Sixth Cluster



(f) Eighth cluster



(g) Tenth Cluster



(h) Tenth cluster

Fig. 77.3 Network modules

Table 77.1 List of genes with corresponding fold change and p-value

Gene	Connectivity ratio	Fold change	p-value
CLDN18	0.89	−11.12	$5.36 \times e^{-22}$
SFTPD	0.87	−6.43	$4.63 \times e^{-15}$
FHL1	0.87	−5.56	$9.63 \times e^{-22}$
EPAS1	0.96	−4.25	$8.95 \times e^{-25}$
NONO	0.93	+1.24	$4.69 \times e^{-06}$
KIF23	0.89	+1.45	$4.32 \times e^{-07}$
PTK6	0.67	+1.68	$5.90 \times e^{-08}$
TFAP2C	0.94	+1.70	$9.19 \times e^{-11}$
MMP13	0.61	+1.83	$3.74 \times e^{-07}$
EIF2AK1	0.97	+1.89	$1.51 \times e^{-19}$
IRAK1	0.99	+1.89	$1.36 \times e^{-14}$
GPR37	0.93	+2.21	$2.69 \times e^{-07}$

Table 77.1. Relevance of those molecules is further validated by their corresponding fold change and p-values.

KIF15 and KIF2C found differentially expressed between current from never smoker and former from never smoker by Landi et. al. [12]. We found KIF23 as a highly connected component. CLDN18 (familiar as Claudin-18) and FHL1 which are found down-regulated in current study are reported to be suppressed in lung cancer [10, 13]. Variant of SFTPD is also discovered for marker in lung adenocarcinoma [14]. Matrix metalloproteinases genes are found severely up-regulated lung carcinoma [15]. We found MMP13 as an oncogene. Several other genes found highly connected and severely differentially expressed in our result demand biological validation.

Conclusion

Cellular disorder upon infection of disease is a manifestation of a series of interactions among several genes, rather than a couple or a few of them. Transformation in genetic behavior in the presence of specific disease can be studied through changes in gene interaction network. Purpose of this study is to build differential network of gene interaction such that genetic variations become apparent in the presence of phenotypic agent. We present a method of interweaving gene expression values under different phenotypic condition of cells using rough set. Interaction network is constructed by assessing the changes in co-expressional activity. Identifying core groups and the effect of disease on them in case of lung cancer is ascertained through tests of validity and biological observations. The method leaves scope for mining knowledge on sequential changes in gene interaction network with availability of time series data.

References

1. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. 286(5439):531–537.
2. Hung-Chung Huang, Siyuan Zheng, Vincent VanBuren, and Zhongming Zhao. Discovering disease-specific biomarker genes for cancer diagnosis and prognosis. *Technology in cancer research & treatment*, 9(3):219–229, 2010.
3. Lesley T MacNeil and Albertha JM Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research*, 21(5):645–657, 2011.
4. Sonam Dolma, Stephen L Lessnick, William C Hahn, and Brent R Stockwell. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer cell*, 3(3):285–296, 2003.
5. Swarup Roy, Dhruba K Bhattacharyya, and Jugal K Kalita. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC bioinformatics*, 15(7):1, 2014.
6. Guanming Wu and Lincoln Stein. A network module-based method for identifying cancer prognostic signatures. *Genome biology*, 13(12):1, 2012.
7. Jie Zhang, Kewei Lu, Yang Xiang, Muhtadi Islam, Shweta Kotian, Zeina Kais, Cindy Lee, Mansi Arora, Hui-wen Liu, Jeffrey D Parvin, et al. Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput Biol*, 8(8):e1002656, 2012.
8. Denise M Wolf, Marc E Lenburg, Christina Yau, Aaron Boudreau, and Laura J vant Veer. Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. *PloS one*, 9(2):e88309, 2014.
9. Zdzisław Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 11(5):341–356, 1982.
10. A. Sadhu and B. Bhattacharyya. Discovery of cancer linked biomarker genes through common subcluster mining. *International Conference on Bioinformatics and Systems Biology (BSB),2016 IEEE*, pages 1–5, March 2016.
11. Jianhua Ruan, Angela K Dean, and Weixiong Zhang. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC systems biology*, 4(1):1, 2010.
12. Maria Teresa Landi, Tatiana Dracheva, Melissa Rotunno, Jonine D Figueroa, Huaitian Liu, Abhijit Dasgupta, Felecia E Mann, Junya Fukuoka, Megan Hames, Andrew W Bergen, Sharon E Murphy, Ping Yang, Angela C Pesatori, Dario Consonni, Pier Alberto Bertazzi, Sholom Wacholder, Joanna H Shih, Neil E Caporaso, and Jin Jen. Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival. *PLoS ONE*, 3(2):e1651, 2008.
13. Chang Niu, Chaoyang Liang, Juntang Guo, Long Cheng, Hao Zhang, Xi Qin, Qunwei Zhang, Lihua Ding, Bin Yuan, Xiaojie Xu, et al. Downregulation and growth inhibitory role of fhl1 in lung cancer. *International Journal of Cancer*, 130(11):2549–2556, 2012.
14. Ruiyun Li, Nevins W. Todd, Qi Qiu, Tao Fan, Richard Y. Zhao, William H. Rodgers, Hong-Bin Fang, Ruth L. Katz, Sanford A. Stass, and Feng Jiang. Genetic deletions in sputum as diagnostic markers for early detection of stage i non small cell lung cancer. *Clinical Cancer Research*, 13(2):482–487, 2007.
15. Wiebke Sauter, Albert Rosenberger, Lars Beckmann, Silke Kropp, Kirstin Mittelstrass, Maria Timofeeva, Gabi Wlke, Angelika Steinwachs, Daniela Scheiner, Eckart Meese, Gerhard Sybrecht, Florian Kronenberg, Hendrik Dienemann, Jenny Chang-Claude, Thomas Illig, Heinz-Erich Wichmann, Heike Bickebller, and Angela Risch. Matrix metalloproteinase 1 (mmp1) is associated with early-onset lung cancer. *Cancer Epidemiology Biomarkers and Prevention*, 17(5):1127–1135, 2008.

Chapter 78
River Length Calculation Using Map Data

Deepshikha Shrivastava, Amitabha Nath and Goutam Saha

Abstract Researchers over the years are trying to model different natural phenomenons for its better understanding and decision making, and river morphology is one of them. There are several features such as length, depth, flora and fauna, health of the river. River length has become an interesting topic for researchers over the years. Calculating exact river length is the prerequisite for various calculations in river geomorphology. Traditionally, river length has been determined by field survey or by interpreting vector-based topographic data. The topographic data represents a river using either a single line (narrow) or double line (wide) depending on the scale and size of the river. Often getting topographic data for a region of interest becomes the biggest hurdle in carrying out research in this domain. Fortunately, there is one source in the form of digital maps, e.g., Google Earth, ARC GIS map database, which are freely and readily available and can be brought into use for such calculation. This paper discusses an approach of determining length of a particular river section from map data by calculating centerline of the river using Delaunay triangle.

Keywords Centerline · Double line river · Delaunay triangle
River length

Introduction

In this paper, the main objective is to calculate the exact river length from map or Digital Elevation Model (DEM) data. In DEM, often shallow rivers are represented using single line, whereas wide rivers are represented using double line [1, 2].

D. Shrivastava · A. Nath (✉) · G. Saha
Department of Information Technology, North Eastern Hill University, Shillong, India
e-mail: amitabha.me@gmail.com

D. Shrivastava
e-mail: deepshikha.rke12@gmail.com

G. Saha
e-mail: dr_goutamsaha@yahoo.com

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_78

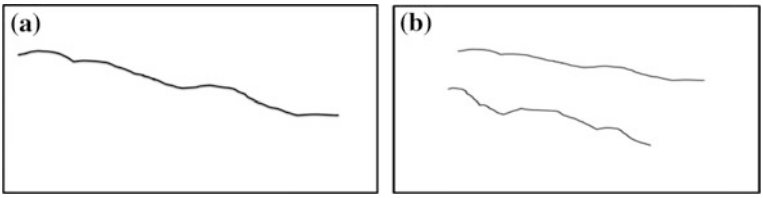


Fig. 78.1 Diagram illustrating **a** single line river, **b** double line river

Situations where rivers are represented as a single line things which are little easy as because calculating length of that single line can be interpreted as the actual river length, whereas in case of double line this kind of calculation does not hold true. As in double line river, it is not necessary that both the lines representing the two sides of the river are identical, so we cannot calculate the length of river directly. This can be seen from the following (Fig. 78.1).

As for calculating the river length, we are depending on the field survey or vector-based topographical sheets and that often fails to give the proper output [1]. This method is time-consuming, and many times it does not give good result. So our objective is to calculate the length of double line (wide) rivers as accurate as possible. As in double line river, it is not necessary that both the lines are identical so it is not easy to calculate the length directly just by calculating the length of any one line.

Input

For input data we can use Digital Elevation Model (DEM) or other source of map data like Google map images. Although images acquired using Google map requires some extra preprocessing to make it geo-referenced [3, 4]. Once geo-referenced images are available, Delaunay triangle can be formed and required calculation can be carried out.

Delaunay Triangle

For a set of points P, Delaunay triangle is with vertices P so that there is no point P in the each triangle's circumcircle [5]. Angle maximization or minimization performed by Delaunay triangle. It is formulated by Boris Delaunay for his work on this topic from 1934 (Fig. 78.2).

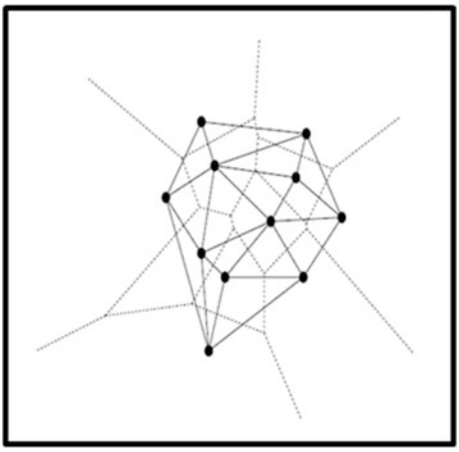


Fig. 78.2 Diagram illustrating Delaunay triangle [5]

Materials and Methods

The following Fig. 78.3 shows how we can calculate the river length of double line (wide) river after calculating the centerline of river with help of Delaunay triangle in MATLAB.

The input of the system is the double line river. The output is the centerline of the river.

The system performs the summarization in four main steps:

(1) It draws the Delaunay triangle
(2) calculation of midpoint
(3) calculating centerline
(4) calculation of river length.

For this, we have to calculate the midpoint of the Delaunay triangle which is further used for calculation of the centerline of river [1, 2]. The midpoints of adjacent and inside marked Delaunay triangle ABC with coordinates $d = (A + B + 2C)/4$ as A, B, and C are the vertices of the triangle, and d is called the midpoint of triangle (Figs. 78.4 and 78.5).

After calculating the medial axes point of Delaunay triangle, we joined these medial points and then calculated the centerline for the river. This centerline is helpful for the double line river as it provides the centerline in single format. Finally we calculate the length of the centerline and we get the length of the river.

After joining these midpoint obtained by Delaunay triangle, we used the straight line equation to calculate the length of that centerline. This length of centerline is the length of river. The equation which is used to calculate the length is given below

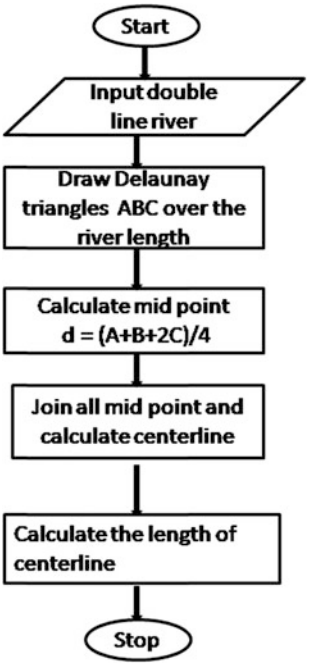


Fig. 78.3 Diagram illustrating flow chart for proposed system architecture

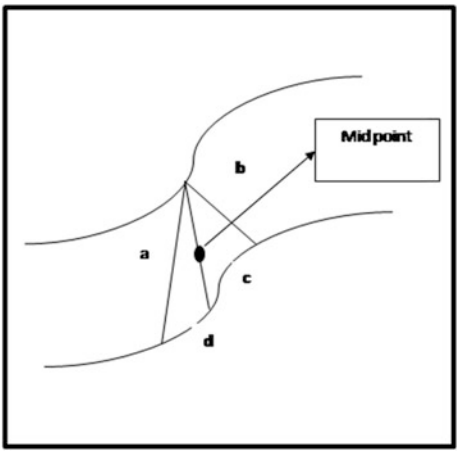


Fig. 78.4 Diagram illustrating medial axes point of triangle

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

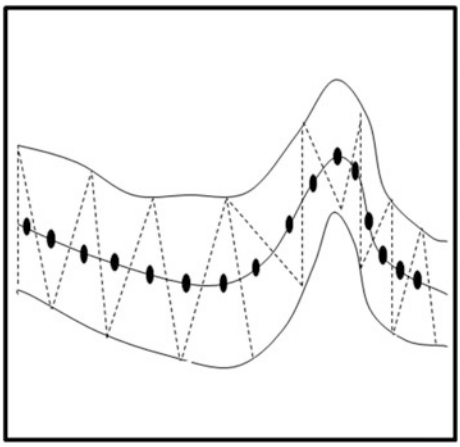


Fig. 78.5 Diagram illustrating middle axis for centerline

With help of above equation, we calculated the length of river in segments; after calculating each straight line segment, all line segments are combined and calculated the length of centerline.

Result

Fig. 78.6 shows the first phase of our approach. In this step, we loaded our lines for which the Delaunay triangle is calculated. In MATLAB, we make a GUI in which the lines are loaded. In this step, the loaded double lines are used for which the Delaunay triangle is calculated.

This is the GUI application form which we used to calculate the centerline with help of Delaunay triangle. With help of load line button, we can load the desired double line for grid of Delaunay triangulation (Fig. 78.7).

After this, Delaunay triangle is calculated for the input lines of the double line river with help of GUI which is developed in MATLAB. The grid of Delaunay triangle is made for entire line which is further used for calculating the centerline of the river (Figs. 78.8 and 78.9).

We have implemented the Delaunay triangle for calculating the centerline of the river which can be represented by a pair of double lines and then applied the straight line equation to find the length of the centerline. The result obtained by this approach is good enough as it provides the length near to the actual value

For testing purpose we have applied our technique on river Barak in south Assam, India. For a section of the river spanning from Silchar to Badarpur, its actual length is near about 30 kms. When same cross section is put into our system

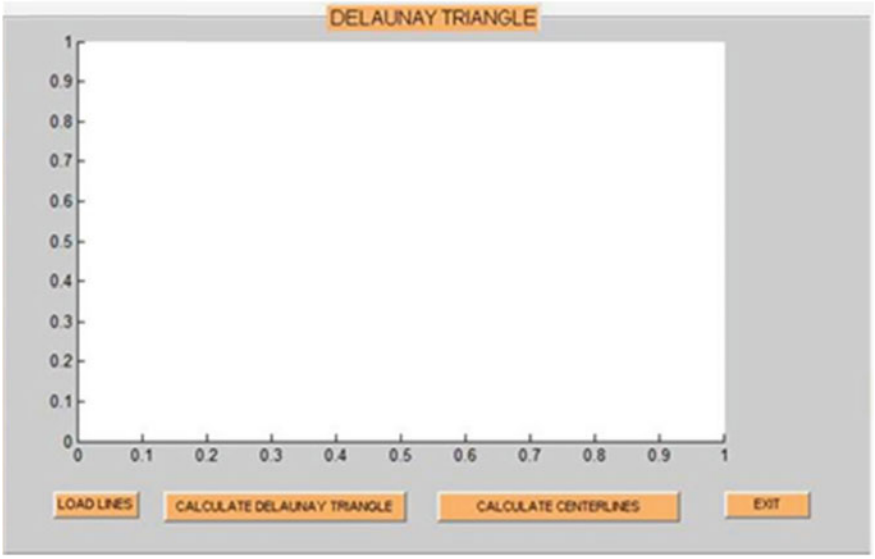


Fig. 78.6 Diagram illustrating GUI for calculating river length

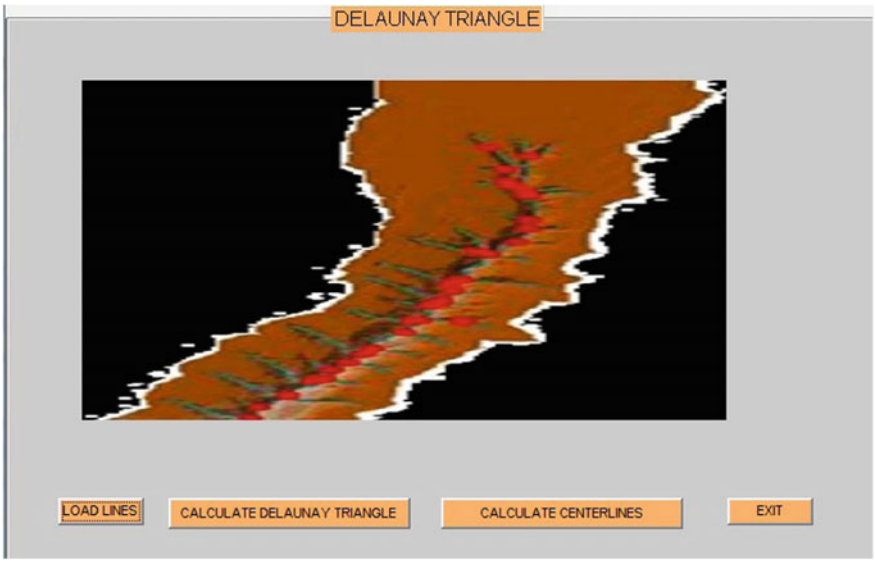


Fig. 78.7 Figure showing loaded image of a river

we got the length to be 28.73 kms. An accuracy of nearly 96% is quite good and acceptable given the scenario that substantial amount of ground work and physical surveying is reduced.

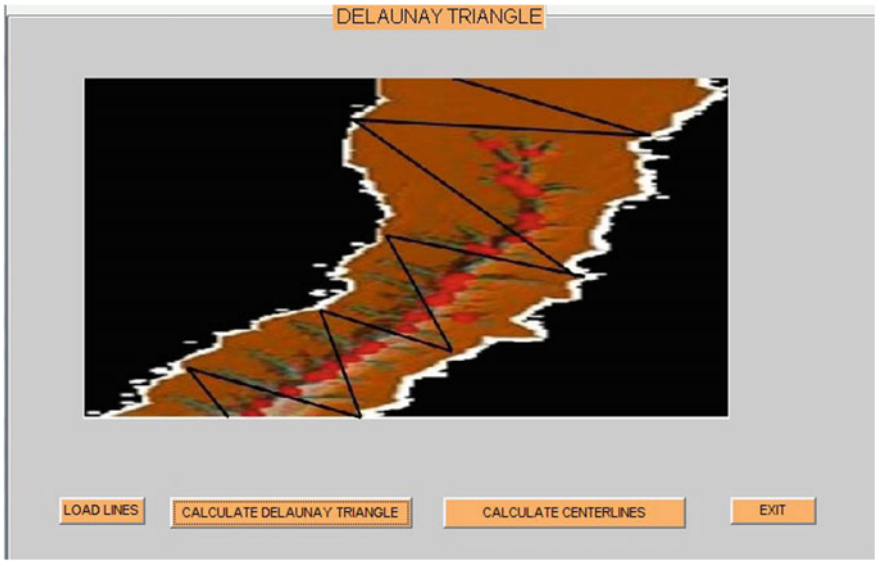


Fig. 78.8 Diagram illustrating form for calculating Delaunay triangle

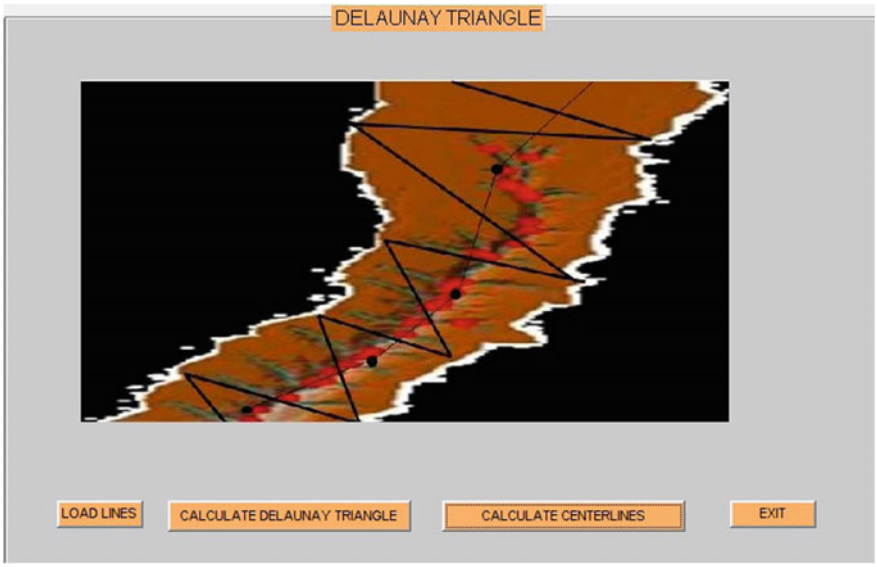


Fig. 78.9 Diagram illustrating form for calculating centerline for Delaunay triangle

Experiment and Conclusion

Finding exact river length is a difficult task. The objective of this work has been to calculate river length as accurately as possible. We have used map from ARC GIS map archive as input data and MATLAB program to segregate the river cross section from the map and marked Delaunay triangles to find the centerline for calculation of length. The method shows a great deal of accuracy for the instances where it has been applied. As we compare this with existed model, which have the manual calculation, then it shows the better result.

Large rivers have many tributaries which make calculation of length even more difficult. Manual intervention is required to cut the branches and isolates mainstream from rest of the branches. Beside that identification of origin (headwater) and end point (embouchure) of a river is also important in identifying the exact length. Further, work can be devoted in this direction to automate this whole manual process.

References

1. ZHANG Lijuan HUANG We JIANG Jie Calculating River Length Based on Topographic Data, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Vol. XXXVII. Part B4. Beijing 2008 pp-181–186
2. Hiep-Thuan Do, S´ebastienLimet, Emmanuel Melin Parallel Computing Flow Accumulation in Large Digital Elevation Model ICCS 2011 2277–2286
3. S.P.Wechsler Uncertainties associated with digital elevation models for hydrologic applications: a review, Hydrol. Earth Syst. Sci., 11, 1481–1500, 2012
4. Johann Muller, Isabelle Gartner, Accuracy assesment of airborne photogrammetrically derived high-resolution digital elevation models in high mountain environment, ISPRS journal of photogrammetry and remote sensing 98 (2014) 58–69
5. Quach Dong Thang Spatial Analysis in PostGIS based on Voronoi diagram/ Delaunay triangulation, Asia Geospatial Forum, 17–19 September 2012, Hanoi, Vietnam 3, I pp-1–11
6. CharrierYingkui Li, Assessing resolution and source effects of digital elevation models on automated floodplain delineation: A case study from the Camp Creek Watershed, Missouri Applied Geography 34 (2012) 38–46

Chapter 79
Cancer Detection with Prostate Zonal Segmentation—A Review

Gaurav Garg and Mamta Juneja

Abstract Urologist are treating approximately six out of ten prostate cancer (CaP) patients. Early CaP detection is major concerns so that immortality rates can be improved globally. Accurate and timely detection of CaP plays key role in proper diagnosis of cancer. This paper reviews various cancer localization techniques by considering prostate zonal segmentation, and it also addresses various challenges which are still open problem for biomedical engineers.

Keywords Prostate cancer · Segmentation · MRI

Introduction

CaP is a most frequently diagnosed cancer in which cancer cells formed in the tissue of the prostate gland, and it is one of the most common disease found among men after 50 years of age. Statistics of American Cancer Society of 2016 predict about 1.80 lakh new cases and approximately 26 thousands of death cases due to CaP in the USA [1]. Statistics of India states that where life span raised from 61.97 to 65.48 since 2001 to 2011, CaP occurrences nurtured by one percent each year. Early stage cancer initiates to grow slowly in the prostate gland which further results into highly malignant tumor and invades to surrounding organs. The main indications of CaP are recurrent urination, irregular or weak flow of urine, painful urination, blood in urine, pain in pelvis, hips, back, and other adjoining bony areas. Studies revealed that foremost reasons of CaP are due to hereditary, age [2], high BP [3], dietary routines [4],

G. Garg (✉) · M. Juneja
Department of Computer Science and Engineering,
UIET, Panjab University, Chandigarh, India
e-mail: ergaurav.garg@yahoo.com

M. Juneja
e-mail: mamtajuneja@pu.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_79

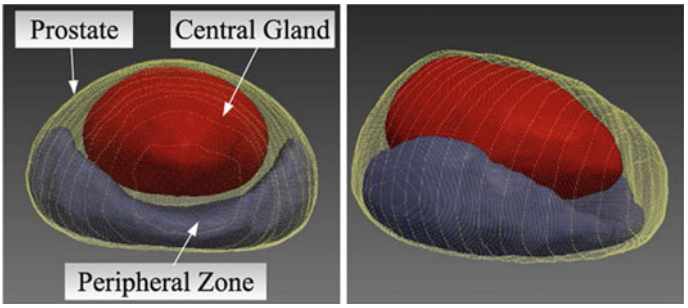


Fig. 79.1 Image courtesy of [6]

intake of processed, and red meat [5]. Consistent physical activity, consumption of green veggies, fish, vitamin E, and mineral supplements [2] can be counted as some preventive measures of CaP. CaP can be effectively treated by active surveillance, cryosurgery, vaccine treatment, hormone therapy, radiation therapy, and chemotherapy [2]. Diagnostic range of CaP involves digital rectal examination (DRE), prostate specific antigen (PSA) blood test, biopsy tissue examination, and imaging test.

Detection and segmentation of prostate gland or CaP are a perplexing task, and these issues considerably vary from one modality to another. Various modalities like TRUS, MRI, and CT efficiently assist in assessment of location and size of prostate gland or CaP. The usage of a specific modality is determined by the clinical intention as the information delivered by each modality considerably vary from each other. Prostate zonal segmentation is very fast growing research area because it is concluded in [6] that in utmost cases, tumor is present in the peripheral zone (PZ). Tumor can be located in the central gland (CG) also, which have significantly dissimilar aspect than PZ tumor. Whole prostate gland is combination of CG and PZ as shown with different colors (3D view) in Fig. 79.1.

Prostate Zonal Segmentation

Prostate gland is surrounded with bladder, seminal vesicles, and rectum. It is a challenging task to delineate the cancer accurately in different zones of prostate gland without delineating each zone separately. Following studies emphasizes on CaP detection w.r.t prostate zonal segmentation.

Derraz et al. [7] presented a novel framework for two-staged contours segmentation algorithm centered on directional active contours (DAC) by integrating preceding prostate shape knowledge. Author firstly addressed the prostate contour segmentation issue precisely by employing globally DAC that introduce both parametric and statistical shape prior model. This enables the exploitation of global characteristics of contour segmentation issue by integrating a user response in segmentation pro-

cedure to resolve the confusing states raised by DAC. After successful delineation of prostate contour, a cost function was developed which introduces both parametric shape prior model and user feedback communication. Author used the openly available MR datasets to prove the efficacy and proficiency of the proposed methodology. The shortcomings of this method was that it shows poor performance for some slices due to texture non-uniformity, unclear boundary along the prostate apex and base, and high variation of prostate visual characteristics among different patients. The major advantage of the proposed methodology was that it produces a fast, perfect, and reproducible delineation solution with least user interaction.

Qiu et al. [8] recommended a multi-region delineation algorithm to concurrently delineate the prostate capsule and its two main subregions, i.e., CG and PZ on a single 3D T2w image. Proposed delineation algorithm used the prior spatial region uniformity and a modified prostate appearance model. The expressed challenging combinatorial optimization issue was resolved by convex relaxation in which a new spatially continuous max-flow model was hosted as dual optimization formulation with region uniformity restrictions. The anticipated model developed an effective duality-based procedure that has numerical benefits and easily implementable on GPUs. The recommended methodology was authenticated with eighteen three-dimensional images with a bodycoil and twenty-five images with an endorectal coil. Author marked the incorporation of shape and texture information for improvement of delineation accurateness as future work. The limitation of work was that it was a semiautomatic algorithm requiring a user initialization which could lead to the observer variability. This study only measured the transversal volume of single image but volume measurement in both sagittal and coronal views will add more value to clinical decisions which was marked as future work by the author.

Litjens et al. [9] proposed a pattern recognition-based algorithm for prostate zonal delineation. Author considered the three different features, i.e., anatomical, texture, and intensity features which could distinguish the both zones. Evaluation was done with multi-parametric (mp) multi-atlas-based methodology considering forty-eight y mp-MRI cases. Inter-observer variability was assessed by three observers and comparison was done with existing studies. The main disadvantage of this study was that it was difficult to compare the outcomes from other research work because of data variability. Author concluded that proposed voxel classification methodology could be improved by incorporation of other texture features (like Gabor or Gaussian-based texture features) or by inclusion of global information like volume of prostate.

Toth et al. [10] proposed a domain constrained deformable (DoCD) biomechanical model to study the simulating effects of radiation on the internal subregions of the prostate. Author applied DoCD on thirty patients registered images who were subjected to external beam radiation therapy (EBRT) for CaP. Efficiency of the proposed methodology was validated by both quantitative and qualitative results.

Lemaitre et al. [11] recommended an automated algorithm to localize CaP in a per voxel strategy by means of 3 T mp-MRI and a gradient boosting classifier. By means of all mp and zonal information, author concluded that notable results were obtained in terms of three parameters, i.e., sensitivity, specificity, and an area under curve (AUC): 94.7%, 93.0%, and 0.968, respectively.

Trigui et al. [12] proposed a classification schema considering random forest and SVM classifier in three classes, i.e., healthy, benign, and malignant for 34 patients. Author contributed with improved automated classification outcomes as compared to parfait et al. [13] because of improved spectra quality. Author presented the enhanced global detection by employing mp-MRI-based features. Feature selection was done by evaluation of several MRI modalities combinations. Analysis was carried over whole prostate gland considering CG and PZ, and it was concluded that SVM classify the healthy and malignant voxels with sensitivity, specificity, and error rate as 99.1%, 98.4%, and 1%, respectively, and random forest with 72%, 88%, and 18.2%, respectively.

Allen et al. [14] recommended an automatic technique to segment the CG and prostate periphery. However, proposed technique delineated the prostate middle region only and it did not consider the gland base and apex.

Yin et al. [15] anticipated an automatic CG delineation procedure for T2w MR images on the basis of layered optimal graph image segmentation of multiple objects and surfaces (LOGISMOS). Author calculated the mean DSC of 80% for CG. Demerit of above studies [14, 15] is that it only considered the CG segmentation only and neglected the PZ which restricts its application for clinical usage.

Critical Inferences

On the basis of study of the above literature, Table 79.1 presents the quantitative comparison of different methodologies considered by various biomedical engineers. In order to avoid under or over segmentation issues, all adjoining organs of prostate, i.e., bladder, bones, seminal vesicule, rectum, and all different zones of prostate, i.e., CG, transition zone (TZ), and PZ must be delineated accurately so that tumor penetration in other zones or regions can be located or differentiated properly by urologist. The major drawback of all discussed studies [7–15] is that no one has considered all adjoining parts of prostate in segmentation algorithm which may produce erroneous outcomes in clinical exercise. Moreover, there is need of evaluation of different segmentation algorithms on both 1.5 T and 3.0 T mp-MRI to prove the robustness of the methodology, above studies are lacking in the same context.

Table 79.1 Quantitative comparison of different segmentation techniques

Study	Segmentation Criteria	Efficiency		Validation
		Measure	Value	
Derraz et al. [7]	Directional active contours (DAC) with prior shape knowledge	Dice similarity coefficient (DSC), Absolute relative volume difference (ARVD)	1.29 ± 0.23, 1.02 ± 2.30	50 Patients, MICCAI grand challenge 2012 dataset
Qiu et al. [8]	Multi-region delineation with convex relaxation	Mean DSC	89.3 ± 3.2% for bodycoil images, 89.5 ± 3.2% for endocoil images	18 body coil images, 25 endocoil images
Litjens et al. [9]	Multi-atlas delineation	Mean DSC	0.89 ± 0.03 for CG, 0.75 ± 0.07 for PZ	48 mp-MRI
Toth et al. [10]	DoCD	Mean DSC, Mean absolute distance (MAD), Root mean square error	0.896 for prostate and 0.902 for CG, 0.665 mm for prostate and 0.397 mm for CG, 2.994 mm	30 Patients
Lemaitre et al. [11]	Voxel-based and 3D texton-based feature extraction strategy	Sensitivity, Specificity, Area Under Curve (AUC)	94.7%, 93.0%, 0.968.	20 Patients
Trigui et al. [12]	mp-MRI based features classification schema	Error rate, Sensitivity, Specificity	1, 99.1, 98.4% for SVM; 18.2, 72, 88% for random forest	34 Patients
Yin et al. [15]	LOGISMOS	Mean DSC, Execution time	0.81, 15 s	261 images

Challenges in Diagnosis of Prostate Cancer

CaP delineation is still an unsatisfactorily defended problem. It is high requirement that proposed algorithms must work in real scenario in automated way to produce results with improved efficiency and accuracy. It is possible by fusion of at least two different modalities so that valuable evidence can be extracted. Second, an increase in 3D prostate delineation approaches will be the tendency in upcoming years because of the increased usage of three-dimensional imaging modalities, where proficient and perfect algorithms are essential. Third, endorectal magnetic resonance (MR) imaging contributes significant value to PSA, DRE, and biopsy findings by localizing cancer and by assessing its size and extension. However, MRI is insensitive to whether the cancer has metastasized to the lymph nodes and is only somewhat accurate in predict-

ing if the cancer has penetrated the prostate capsule. Fourth, a number of structures and conditions—including BPH nodules, prostatitis, and hemorrhage—show similar enhancement patterns to CaP on dynamic contrast-enhanced (DCE) MR images and therefore compromise the specificity of image analysis. Fifth, with respect to different modalities (CT and TRUS) used for diagnosis of CaP, non-availability of benchmarked dataset with ground truth of different prostate zones and CaP is of major concern and only one dataset is available w.r.t to multi-parametric MRI [16] of 30 patients only and it is under development phase. Last, the majority of prostate CADx systems employed a relatively small data set with no more than 50 patients. Validation on a large-scale data set with several hundred patients is required to make the systems usable in clinical settings.

Conclusion

CaP diagnosis is an active area of research among biomedical engineers since last two decades because of its unbearable impact on life expectancy. CaP localization is still an open problem, and it is high need to address the above discussed challenges unambiguously to invent a methodology which is capable to detect early stages of CaP so that mortality rates can be suppressed globally.

References

1. American Cancer Society, "Prostate cancer" 2016, http://www.cancer.org/Cancer/ProstateCancer/
2. G. Steinberg, B. Carterm, T. Beaty, B. Childs, and P. Walsh.: Family history and the risk of prostate cancer. Prostate 17(4), 337–47 (1990).
3. R. Martin, L. Vatten, D. Gunnell, and P. Romundstad.: Blood pressure and risk of prostate cancer: Cohort Norway (CONOR). Cancer Causes Control 21(3), 463–72 (2010).
4. V. Venkateswaran and L. H. Klotz.: Diet and prostate cancer: Mechanisms of action and implications for chemoprevention. Nature reviews Urology 7(8), 442–453 (2010).
5. D. Alexander, P. Mink, C. Cushing, and B. Sceurman.: A review and meta-analysis of prospective studies of red and processed meat intake and prostate cancer. Nutrition journal 9(50), (2010)
6. Toth R., Ribault J., Gentile J., Sperling D., and Madabhushi A.: Simultaneous segmentation of prostatic zones using active appearance models with multiple coupled levelsets. Computer Vision and Image Understanding 117(9), 1051–1060 (2013).
7. Derraz, F., Forzy, G., Delebarre, A., TalebAhmed, A., Oussalah, M., Peyrodie, L., and Verclytte, S.: Prostate contours delineation using interactive directional active contours model and parametric shape prior model. International journal for numerical methods in biomedical engineering 31(11), (2015)
8. Qiu, W., Yuan, J., Ukwatta, E., Sun, Y., Rajchl, M., and Fenster, A.: Dual optimization based prostate zonal segmentation in 3D MR images. Medical image analysis 18(4), 660–673 (2014).
9. Litjens, G., Debats, O., van de Ven, W., Karssemeijer, N., Huisman, H.: A pattern recognition approach to zonal segmentation of the prostate on MRI. In: International Conference on

Medical Image Computing and Computer-Assisted Intervention, pp. 413–420. Springer Berlin Heidelberg (2012)
10. Toth, R., Traughber, B., Ellis, R., Kurhanewicz, J., and Madabhushi, A.: A domain constrained deformable (DoCD) model for co-registration of pre-and post-radiated prostate MRI. Neurocomputing 144, 3–12 (2014).
11. Lemaitre, G., Massich, J., Martí, R., Freixenet, J., Vilanova, J. C., Walker, P. M., ... and Mriaudeau, F. : A boosting approach for prostate cancer detection using multi-parametric MRI. In: International Conference on Quality Control by Artificial Vision, pp. 95340A–95340A. (2015)
12. Trigui, R., Mitran, J., Walker, P. M., Sellami, L., and Hamida, A. B.: Automatic classification and localization of prostate cancer using multi-parametric MRI/MRS. Biomedical Signal Processing and Control 31, 189–198 (2017).
13. S. Parfait, P.M. Walker, G. Crhange, X. Tizon, J. Miteran.: Classification of prostate magnetic resonance spectra using support vector machine. Biomedical Signal Processing and Control 7(5), 499–508 (2012)
14. Allen, P., Graham, J., Williamson, D., Hutchinson, C.: Differential segmentation of the prostate in MR images using combined 3d shape modelling and voxel classification. In: 3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, pp. 410–413. (2006)
15. Yin, Y., Fotin, S., Periaswamy, S., Kunz, J., Haldankar, H., Muradyan, N., Turkbey, B., Choyke, P.: Fully automated 3d prostate central gland segmentation in mr images: a logismos based approach. In: SPIE Medical Imaging, International Society for Optics and Photonics, pp. 83143B–83143B. (2012)
16. G. Lemaitre, R. Marti, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau: Computer-Aided Detection and Diagnosis for prostate cancer based on mono and multi-parametric MRI: A Review. Computer in Biology and Medicine 60, 8–31 (2015).

Chapter 80
Clustering Based on Ant Colony Optimization and Relative Neighborhood (C-ACORN)

Parika Jhanji, Ankit Vij and Padmavati Khandnor

Abstract Wireless sensor network has emerged as a powerful technology and is growing day by day. Ease of availability and low maintenance of small, inexpensive, fault-tolerant, self-configured, self-reliant, easily deployable sensor nodes has made them useful in several critical areas like military, healthcare, industrial process control, security and surveillance, smart homes. But wireless sensor network face challenges of energy conservation, increasing the network lifetime. Clustering is the best-known solution to this problem. In this paper, ant colony optimization and relative neighborhood-based clustering algorithm have been proposed which uses graph-based techniques to form neighbors for the ants. The algorithm is evaluated for seven datasets using the cluster validity indices like Dunn's index (DI), modified Dunn's index (MDI), and Rand index (RI). The results are compared with the existing clustering techniques like density-based spatial clustering applications with noise (DBSCAN) and complete linkage clustering. The comparison is done on the parameters of the quality of solution. The proposed algorithm generates good clustering results and is able to detect all the target clusters efficiently.

Keywords Ant colony optimization · Clustering · Cluster validity indices Relative neighborhood

P. Jhanji (✉) · A. Vij · P. Khandnor
Department of Computer Science, PEC University of Technology, Chandigarh, India
e-mail: jhanjipari@gmail.com

A. Vij
e-mail: ankit.vij1@gmail.com

P. Khandnor
e-mail: padmavati@pec.ac.in

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_80

Introduction

The advancements of technology in the field of low-power design and miniaturization have led to the development of cheap, small-sized devices which have built-in computing, storage, and communication facilities with a power source [1]. This tiny smart device is known as a sensor node, and several such nodes combine to form wireless sensor network (WSN). This invention has emerged as one of the most powerful and innovative inventions of the twenty-first century. It has made possible to access the harsh, inaccessible terrains. The ability of these tiny, smart, battery-powered, fault-tolerant sensor nodes to sense various characteristics of the surroundings (e.g., humidity, temperature, pressure, light, sound, air or water quality, movement) and operate in self-configured and self-reliant mode and form network in ad hoc manner has made them useful in many critical fields like military, healthcare, industrial process control, security surveillance, and smart homes [2]. These sensors can be easily deployed in harsh environments (like forests, battlefields) by dropping them from helicopters. The sensor nodes can monitor the presence of vehicles, opposing forces and track their movement, protect the sensitive objects like oil and gas pipelines, atomic plants, and military headquarters, guide unmanned robotic vehicles, submarines, tanks, fighter planes against obstacles. Due to such important applications of sensor nodes, it becomes necessary for them to work efficiently. There are several challenges in WSN and energy conservation, and increasing the network lifetime is among the most important challenges in WSN. The researchers are working upon finding the best solution for these problems and till date, clustering is the best possible known solution [3]. Besides energy conservation and increasing the network lifetime, clustering also helps in scalability, collision avoidance, fault tolerance, and reduced routing delay [3].

Clustering is NP-hard problem of partitioning the given heterogeneous datasets into groups in such a way that the data points in the particular group are similar while they are different from points which are present in the other group. In WSN, we form clusters of sensor nodes that sense data and a cluster head (CH) [4] is formed which collects data from all other nodes in that cluster, aggregates the data, and then sends aggregated data to the base station (BS).

Figure 80.1 depicts clustering in WSN. There are three clusters which are sending data to BS. The red-colored node is CH. All the other nodes are sending the data to CH. The data aggregated by CH is sent to BS.

The clustering algorithms are broadly classified into the following five categories: partition based, hierarchical, density based, graph based, and meta-heuristics (simulated annealing, evolutionary algorithms, swarm intelligence, etc.) [5]. Swarm intelligence (SI) [6] is gaining a lot of popularity these days as it gives quite promising results for the optimization problems. The field of SI is based on the biological phenomenon of swarming, flocking, and herding known in the vertebrates. Among several SI techniques, ant colony optimization (ACO) is the most popular as a lot of research has been done in the field and also it is the oldest known SI technique. ACO derives its inspiration from the clustering and sorting behavior of ants [7].

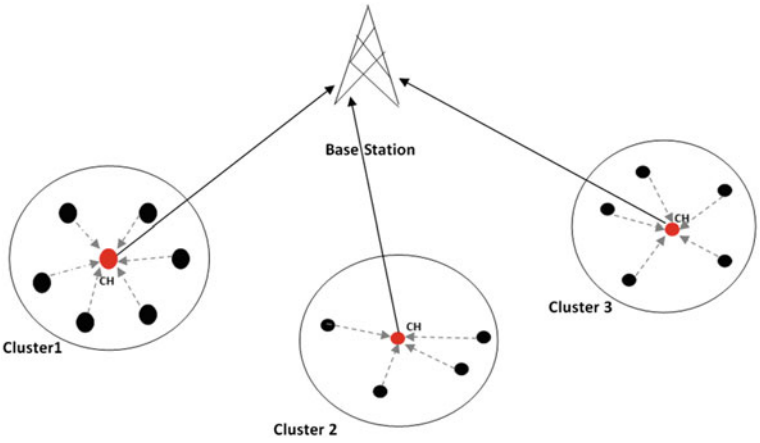


Fig. 80.1 Clustering in WSN

In this paper, we propose ACO-based clustering methodology which uses the relative neighborhood technique to form path for the ants. The results obtained from the proposed algorithm has been compared with other existing clustering algorithms like density-based spatial clustering of applications with noise (DBSCAN) and complete linkage clustering (one of the agglomerative hierarchical clusterings). The comparison has been done on the basis of constructed solution evaluation.

The related work has been discussed in section "Related Work". Section "The Proposed Algorithm: C-ACORN" provides the details of the proposed algorithm. Simulation results are discussed in section "Simulation Results," and section "Conclusion" concludes the paper.

Related Work

In this section, many existing ant colony-based clustering algorithms and clustering solution evaluation using cluster validity indices are presented.

Ant colony optimization was introduced by Marco Dorigo [8]. It is based on the foraging behavior of ants. Ants perform random search for food. While searching food, they deposit pheromones on the ground to mark their path trail from nest to the food source. The path with the highest accumulated pheromone is considered the best path. The other ants follow the path with more pheromones. This type of ants' behavior encouraged researchers to use the same technique for optimization problems. ACO has its applications in many engineering domains [9].

Many ant species cluster corpses to form cemetery and sort their larvae into several piles [10]. This behavior inspires to design clustering methodologies based on ACO. Deneubourg et al. [11] proposed two closely related models on this basis.

The general idea behind the models is to pick the isolated point and group it with similar type of data point. Lumer and Faieta [5] have provided a generalized approach for Deneubourg et al.'s model and have applied it to exploratory data analysis. Urszula Boryezka 2009 [7] has proposed ACA algorithm which is an improved version of Lumer and Faieta's clustering methodology. He also proposed modified ACA algorithm known as ACAM in [10].

Selvakennedy et al. 2007 proposed T-ANT protocol in [12]. The protocol forms optimal number of clusters and provides good CH distribution for load balancing and energy conservation of the network. The clustering operation is divided into several rounds and each round has two phases: cluster setup and steady-state phase. In setup phase, CHs are chosen (node with ants is chosen as CH) and all the other nodes become the part of the best cluster in their range. During steady phase, data is transferred, aggregated, and sent to the sink node at regular intervals of time.

Azzag et al. 2005 [13] proposed AntTree algorithm which is based on self-assembly behavior observed in few species of ants. The ants form complex structures by connecting to each other. The algorithm has better results as compared to K-means, ANTCLASS, and ascending hierarchical clustering.

A hybrid evolutionary algorithm has been proposed by Niknam et al. 2009 [14] which is the combination of fuzzy adaptive particle swarm optimization (FAPSO), ACO, and K-means algorithm. The proposed algorithm gives good results as compared to several other algorithms like ACO, PSO, simulated annealing (SA), PSO and SA (PSO–SA), ACO and SA (ACO–SA), PSO and ACO (PSO–ACO), genetic algorithm (GA), Tabu search (TS), honeybee mating optimization (HBMO), and K-means for partitional clustering problem.

Bandyopadhyay et al. [15] evaluated the performance of K-means, simulated annealing, and single linkage clustering algorithms by using the cluster validity indices. Davies–Bouldin index, Dunn index, Calinski–Harabasz index, and index I were used for comparison. These four indices provide better results for spherical dataset, and index I is more reliable for finding optimal number of clusters.

Ghosh et al. 2008 [16] APC algorithm is based on the concept of aggregation pheromones in ants. The results of the proposed algorithm were compared with agglomerative average linkage and K-means and ant-based method called adaptive time-dependent transporter ants for clustering (ATTA-C) based on the clustering solution and time of execution. APC gives quite promising results.

The Proposed Algorithm: C-ACORN

The proposed algorithm C-ACORN works on the basic principle of ACO. In proposed scheme, the artificial ants are deployed in the network and they work exactly like the biological ants. The neighborhood of the given data points is constructed based on relative neighborhood graph, and the ants follow the path formed by the graph. The pheromones are deposited by the ants while traversing. Greater pheromone density means better chances of finding the similar type of data

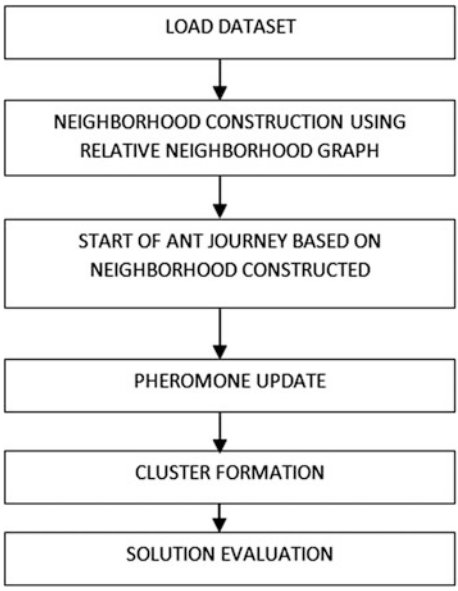


Fig. 80.2 Flowchart for C-ACORN algorithm

points and hence better rate of clustering. The flowchart of the proposed clustering algorithm is shown in Fig. 80.2.

Parameter settings for C-ACORN algorithm:

- no. of iterations: 2000
- no. of ants: 20
- initial pheromone value: 0.1

The detail of each step in C-ACORN is explained below:

Step 0: *Load dataset*:
The algorithm is tested on seven datasets (face, flag, half-ring, moon, ring, spiral, wave).
Step 1: *Neighborhood construction*
The neighbors of the data points are constructed using relative neighborhood graph (RNG). RNG is an undirected graph in which two points: p and q are connected by an edge if there does not exist a point closer to both p and q as they (p, q) are to each other.
Step 2: *Start of ant's journey*
The ants start their journey by traversing the path formed by RNG. The ants will deposit the pheromone on the path as they move. The greater pheromone density indicates the similar type of data points. Therefore, greater pheromone trails mean better clustering solution.
Step 3: *Pheromone update*
Pheromone's value is updated after each round.

Step 4: *Cluster formation*
The cluster is formed on the basis of the pheromone values. After each round, the pheromone values are checked and the data points are picked and dropped by ants in their appropriate cluster.
Step 5: *Solution evaluation*
The measure of the quality of clustering solution has been taken with the help of cluster validity indices (CVIs) [7]. Dunn index [7], modified Dunn index [17], and Rand index [18] are used to evaluate the clustering results.

Simulation Results

The C-ACORN algorithm is tested on seven datasets. The information regarding datasets is given in Table 80.1. The results are compared with the existing clustering algorithms like DBSCAN and complete linkage clustering.

Table 80.2 shows the cases where the C-ACORN, DBSCAN and complete linkage clustering algorithms detect the target clusters. The no. of target clusters to be detected for a particular dataset is provided in Table 1.

Table 80.1 No. of samples for datasets

S. no.	Dataset	Samples	No. of target clusters
1.	Face	320	4
2.	Flag	640	3
3.	Half-ring	400	5
4.	Moon	514	4
5.	Ring	800	2
6.	Spiral	200	2
7.	Wave	287	2

Table 80.2 Analysis of target cluster detection

Algorithm	Detection of target clusters						
	Face	Flag	Half-ring	Moon	Ring	Spiral	Wave
C-ACORN	Yes	Yes	Yes	Yes	Yes	Yes	Yes
DBSCAN	No	No	No	No	Yes	Yes	No
Complete linkage	No	No	No	No	No	No	No

Results for Face Dataset

Figure 80.3a shows the ants forming the cluster of points for face dataset. The red points are the ants, while the green points are the data points of the face dataset. Figure 80.3b shows the target clusters that need to be formed for face dataset.

Figure 80.4a shows the clusters being formed for face dataset by using C-ACORN. The results are shown for several iterations. Figure 80.4b shows the CVI values for C-ACORN algorithm for face dataset. The values of optimal number of target clusters are three for Dunn's index (DI) and four for modified Dunn's index, and Rand index. C-ACORN is able to detect the target clusters.

Figure 80.5a shows the clustering results for DBSCAN algorithm for face dataset. DBSCAN is not able to detect the target clusters. Figure 80.5b shows the

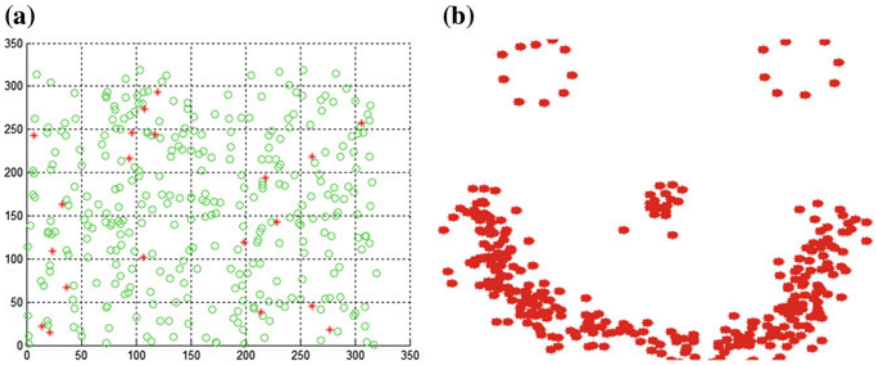


Fig. 80.3 **a** Ants forming the clusters. **b** Target clusters for face dataset

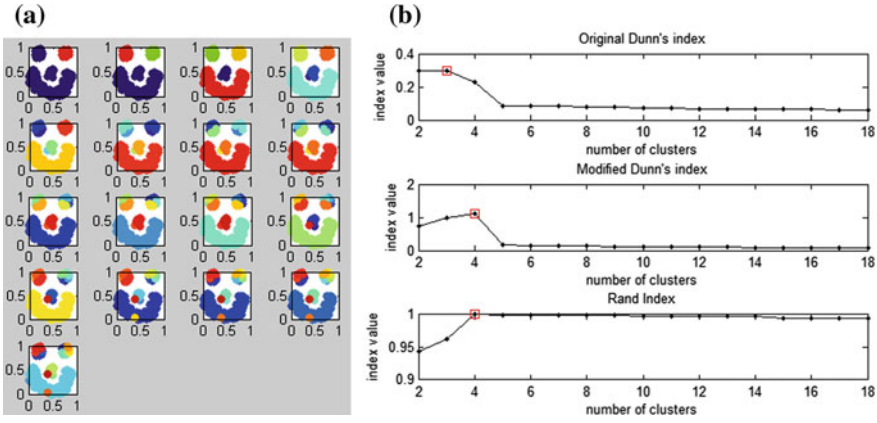


Fig. 80.4 **a** Clustering results for C-ACORN for face dataset. **b** CVI values for C-ACORN for face dataset

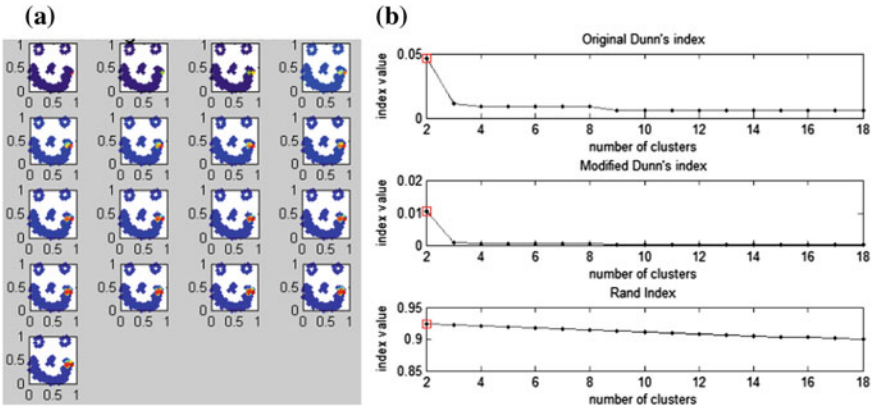


Fig. 80.5 **a** Clustering results for DBSCAN for face dataset. **b** CVI values for DBSCAN for face dataset

CVI values for DBSCAN for face dataset. The value of optimal number of target clusters is two for all the three indices. The optimal number of clusters should be four.

Figure 80.6a shows the clustering results for complete linkage algorithm. Figure 80.6b shows the CVI values for complete linkage clustering algorithm for face dataset. The optimal number of target clusters is two for all the three validity indices which are an incorrect value. Complete linkage clustering is unable to detect the target clusters properly.

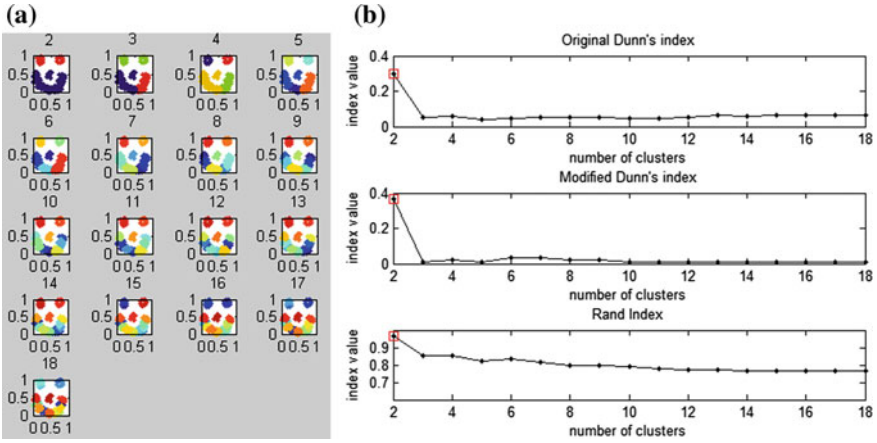


Fig. 80.6 **a** Clustering results for complete linkage clustering. **b** CVI values for complete linkage for face dataset

Conclusion

To address the challenges of energy conservation and increasing the network lifetime in WSN, clustering is the best solution proposed by the researchers. A lot of work is being done in this domain to enhance the existing clustering algorithms in order to improve the performance of sensor networks. In this work, a clustering approach has been proposed based on ACO and relative neighborhood construction. The proposed algorithm C-ACORN is tested for seven datasets, and the obtained results are compared with DBSCAN and complete linkage clustering algorithm. C-ACORN is able to form the target clusters for all the seven datasets, while DBSCAN is able to form target clusters only for ring and spiral dataset. Complete linkage clustering is not able to form target clusters efficiently for any dataset. C-ACORN can be further enhanced in terms of the execution time.

References

1. Mahale, Rashmi A., and S. D. Chavan. "Throughput aware ACO based routing protocol for wireless sensor network." Wireless Computing and Networking (GCWCN), 2014 IEEE Global Conference on. IEEE, 2014.
2. Tyagi, Sudhanshu, and Neeraj Kumar. "A systematic review on clustering and routing techniques based upon LEACH protocol for wireless sensor networks." Journal of Network and Computer Applications 36.2 (2013): 623–645.
3. Afsar, M. Mehdi, and Mohammad-H. Tayarani-N. "Clustering in sensor networks: A literature survey." Journal of Network and Computer Applications 46 (2014): 198–226.
4. Abbasi, Ameer Ahmed, and Mohamed Younis. "A survey on clustering algorithms for wireless sensor networks." Computer communications 30.14 (2007): 2826–2841.
5. Lumer, Erik D., and Baldo Faieta. "Diversity and adaptation in populations of clustering ants." Proceedings of the third international conference on Simulation of adaptive behavior: from animals to animats 3: from animals to animats 3. MIT Press, 1994.
6. Dressler, Falko, and Ozgur B. Akan. "A survey on bio-inspired networking." Computer Networks 54.6 (2010): 881–900.
7. Boryczka, Urszula. "Finding groups in data: Cluster analysis with ants." Applied Soft Computing 9.1 (2009): 61–70.
8. Dorigoa, Marco, and Christian Blumb. "Ant colony optimization theory: A survey." Theoretical Computer Science 344.2–3 (2005): 11.
9. Mohan, B. Chandra, and R. Baskaran. "A survey: Ant Colony Optimization based recent research and implementation on several engineering domain." Expert Systems with Applications 39.4 (2012): 4618–4627.
10. Cheng, Chi-Tsun, Chi K. Tse, and Francis Lau. "A clustering algorithm for wireless sensor networks based on social insect colonies." Sensors Journal, IEEE 11.3 (2011): 711–721.
11. Deneubourg, Jean-Louis, et al. "The dynamics of collective sorting robot-like ants and ant-like robots." Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats. 1991.
12. Selvakennedy, Selvadurai, Sukunesan Sinnappan, and Yi Shang. "A biologically-inspired clustering protocol for wireless sensor networks." Computer Communications 30.14 (2007): 2786–2801.

13. Azzag, Hanene, et al. “A hierarchical ant based clustering algorithm and its use in three real-world applications.” European Journal of Operational Research 179.3 (2007): 906–922.
14. Niknam, Taher, and Babak Amiri. “An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis.” Applied Soft Computing 10.1 (2010): 183–197.
15. Maulik, Ujjwal, and Sanghamitra Bandyopadhyay. “Performance evaluation of some clustering algorithms and validity indices.” Pattern Analysis and Machine Intelligence, IEEE Transactions on 24.12 (2002): 1650–1654.
16. Ghosh, Ashish, et al. “Aggregation pheromone density based data clustering.” Information Sciences 178.13 (2008): 2816–2831.
17. Ilc, Nejc. “Modified Dunn’s cluster validity index based on graph theory.” Przegląd Elektrotechniczny 88 (2012): 126–131.
18. Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. “On clustering validation techniques.” Journal of intelligent information systems 17.2 (2001): 107–145.

Chapter 81
Parasitic Effect on Reduced Latency of SoC-Based Big Data

Seyi Stephen Olokede and Babu Sena Paul

Abstract Big data technology sustainability is contingent on the availability of interconnections of large scale, ultra-high speed, densely integrated big data heterogeneous server platforms. For highly densified servers to be attainable, semiconductors technologies upon which these servers are predicated must further be miniaturized. It is recently not uncommon to implement band gap reduction engineering of SiGe HBT in a bid to attain highly densified integrated circuit for large-scale servers. Unfortunately, the parasitic effects become significant, in particular as these integrated circuits are targeted for high frequency of operations due to the interconnections links between the chip and the transceivers. Insertion loss $|S_{21}|$ becomes considerable, and both the signal level as well as noise figure depreciate substantially as a result. In this work, therefore, we investigate the parasitic effect of interconnections on the roundtrip latency of system-on-chip (SoC).

Keywords Big data · Insertion loss · Integrated circuit · Latency
Parasitic effect · System-on-chip

Introduction

The system requirements for big data communication in particular with regard to hardware demands are very stringent. The interconnections of highly densified chips to meet the resources requirements of massive data communications create more rooms for densified fine wire lines with serious challenges between the chip/transceivers, signal traces, and, interconnections. Unfortunately, the side effect of

S. S. Olokede (✉) · B. S. Paul
Department of Electrical and Electronic Engineering Technology,
University of Johannesburg, Johannesburg, South Africa
e-mail: solokede@gmail.com

B. S. Paul
e-mail: bspaul@uj.ac.za

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4_81

compromising these requirements could result in substantial performance degradation, signal distortion, low speed, and, inability to recover clock signal components correctly. Transmission losses along the signal traces as well as interconnections are considerable due to discontinuities as a result of bends and junctions along the transmission lines. The parasitic effects of these interconnections cause significant propagation delay as the length and frequency increases with attendant impedance bandwidth degradation. The increase effect of mutual coupling, initiate and propagate surface waves, and thus, limit the system efficiency in particular the low speed, and signal distortion.

This challenge is becoming exacerbated due to increasing need for more highly integrated SoC. The reduced chip size to meet systemic integration capability requirements put more stringent constraints on other ancillary associated circuits. For instance, the bond wire or flip chip needed to interconnect the chip to AoC/AiP is expected to be longer to be effective. Interestingly, longer length of interconnect —in order to accommodate more increasingly densified heterogeneous circuits in a chip—leads to more interconnect delay [1–4]. Unfortunately, the present deep submicron is dominated by increasing die size with decreasing wire pitch. The pressure for highly integrated single chip becomes more severe as the need for interconnections of heterogeneous integration of various technologies into one single chip increases. Research efforts have been ongoing in order to mitigate the SoC delay latency. In 1990, copper–low dielectric interconnects were propounded. However, the overtly longer interconnect delay with attendant systemic performance degradation of the SoC disqualifies the technology. The interconnects scaling is subsequently proposed. The intent of this work, however, is in first instance to empirically confirm that delay do exists in SoC due to interconnections, and the delay could cause signal distortion owing to parasitic effect. Secondly is to assess the extent of this delay, and specifically, we intend to characterize a typical SoC in order to proffer an approximate estimate for the delay latency.

Proposed Interconnect Configuration

In Fig. 81.1, a typical schematic diagram of interconnect of two wires is depicted. Fig. 81.1a is the cross-sectional schematic diagram, and Fig. 81.1b is the equivalent circuit diagram. The interconnects are on ungrounded silicon substrate with dielectric constant of 11.7. Though interconnect wires could be assessed based on being local, global, or semi-local. In this work, we intend to focus our work on the intermediate to global interconnects interconnections.

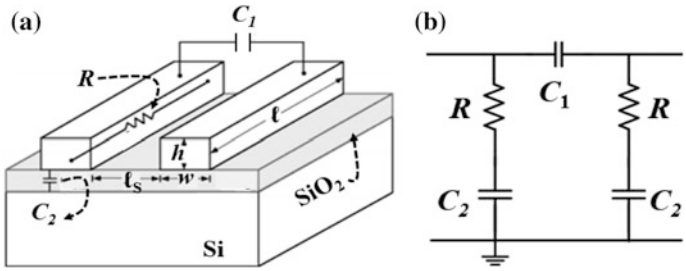


Fig. 81.1 Interconnect of two wires. **a** Cross section schematic, **b** equivalent circuit

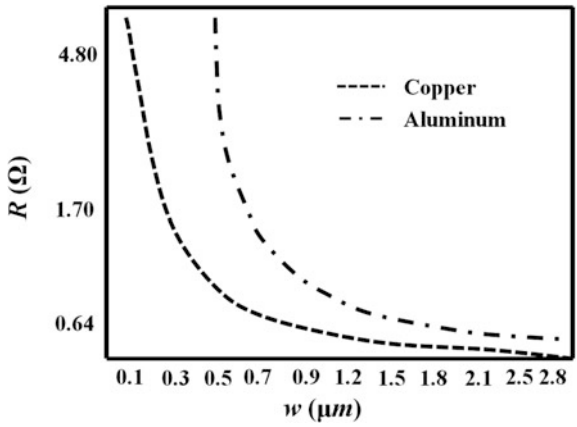


Fig. 81.2 Effect of line width on the resistance of interconnect wire

Equivalent Circuit Representation and Analysis

Figure 81.1b is the equivalent circuit representation of Fig. 81.1a. Equations (81.1) through to (81.4) are the formulae to determine the various equivalent circuit components of the proposed equivalent circuit, where R is the line resistance of one of the interconnect wires, C_{ox} is the substrate to line capacitance, C_1 is the capacitance between the two interconnect wires, C is the total wire capacitance, k_1, k_2 are the fringing fields to the substrate, and to the air, respectively, ε_r is the dielectric constant, and ℓ_s is the feature size, or lithographic resolutions (in $k\lambda$), X_{ox} = oxide thickness.

$$R = \frac{\ell}{\sigma w h} \tag{81.1}$$

$$C_1 = \varepsilon_0 \varepsilon_r \frac{\ell h}{\ell_s} \tag{81.2}$$

$$C_2 = \varepsilon_0 \varepsilon_r \frac{w\ell}{X_{0x}} \tag{81.3}$$

$$C = k_1 (C_1 + C_2) \tag{81.4}$$

The total line capacitance is determined by using Eq. (81.4). The delay function is as stated in Eqs. (81.5) and (81.6) below [5, 6].

$$\tau_L = 2\rho\varepsilon_0\varepsilon_r \frac{w}{wh}\left(\frac{w}{X_{0x}} + \frac{h}{\ell_s}\right) \tag{81.5}$$

$$\tau_L = 2R(C_1 + C_2) \tag{81.6}$$

Experimental Results

The circuit parameters of Fig. 81.1b are extracted using Matlab code. Wire interconnections made of copper and aluminum with resistivities (ρ) of 1.7×10^{-8} and 2.82×10^{-8} Ω m were both examined, with a constant aspect ratio (h/w). The effect of line-to-substrate capacitance with respect to delay time (τ_L), the effect of the line width on the resistance of the interconnect wire, and finally, the effect of lithographic resolution on interconnect wireline were examined. Figure 81.1 depicts the dependency of interconnection resistance on the line width. It is evident that the interconnection resistance depreciated with the line width. When this occurs, the feature size (ℓ_s) becomes wider. The capacitance between the two interconnected wires appreciates until there comes a time when the spacing between the two lines becomes too ($\ell_s \gg$) big with little charges. The capacitance then nose-dived. Consequently, the interconnect line thus affect the delay response, as supported by Eqs. (81.1) and (81.6). Interestingly, aluminum interconnection wires respond more abrupt to line width compared with the copper wire. In both, the responses become very steep at an average line width of about 0.9–1 μm. Alternatively, the interconnection resistance becomes very high at low line width. The effect of this also impacted on the delay time. In Fig. 81.3, the graph of interconnection length against the lithographic resolution is demonstrated. Findings indicate that the longer the interconnection length, the better to use bigger cross-sectional area with lower resistivity materials in order to achieve a reduced resistance. Conversely, it may be expedient and beneficial to place smaller length of interconnections in a layer of smaller cross-sectional area of higher resistivity materials (Fig. 81.2).

In Fig. 81.4, we demonstrate the effect of line-to-substrate with respect to delay time. Ironically, delay time is directly proportional to line-to-substrate capacitance. The higher the line-to-substrate capacitance, the more the delay time. However, abnormality was observed somewhere between 0–4 pF where the delay time depreciate before it appreciate again and then steadily. While we do not specifically

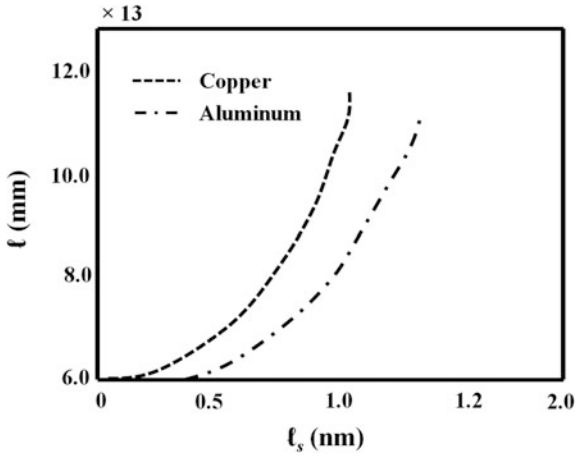


Fig. 81.3 Interconnect length of versus lithographic resolution

investigate what could be responsible to this due to exigency, we envisage that this may not be unconnected to the effect of line resistance as the line width becomes too large.

Though Fig. 81.5 depicts the effect of chip area with respect to delay time, we specifically draw our observation on both Figs. 81.4 and 81.5 to underscore our objective of this work. Obviously, the delay time responds to chip area [7, 8]. Alternatively, the interconnect delay increases with increase of interconnection lengths. By implication, the delay through interconnection length is a larger percentage of the entire model delay time. Therefore, our assertion is that if the interconnect line constitute very significant part of the entire circuit delay as demonstrated by this model which only focus on two wire interconnect lines, it won't be wrong therefore to infer the magnitude of delay that will be experienced in

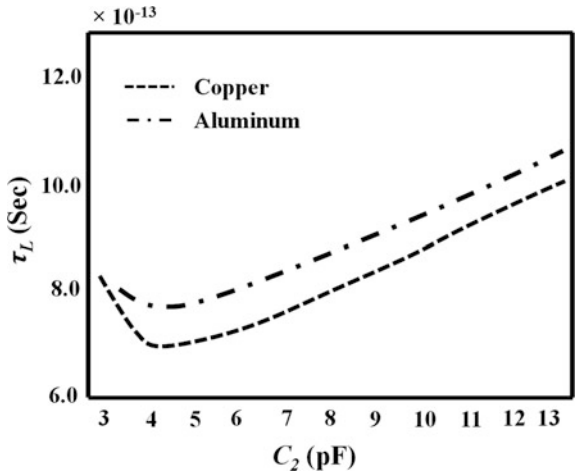


Fig. 81.4 Delay time versus line-to-substrate capacitance

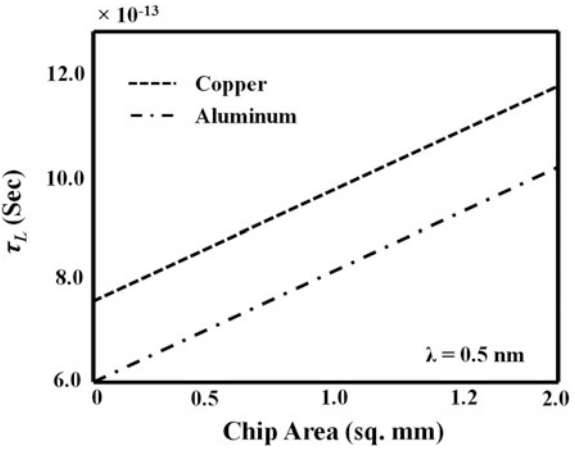


Fig. 81.5 Delay time versus chip area

a highly densified integrated circuit for large-scale servers to provide robust platform for big data [9]. Thus, if the delay time is that substantial, signal distortion, transmission losses, surface wave propagation, will depreciate the circuit performance. In turn, there would be no certainty that the transmitted data will be correctly recovered due to signal degradation and distortion.

Conclusion

In this paper, we have theoretically investigated the parasitic effect of reduced latency of SoC-based big data. We investigated interconnect of two wires on silicon substrate analytically by drawing the equivalent circuit representation. We extracted the equivalent circuit parasitics and examined its influence on the delay time. The effect of interconnect resistance, line-to-substrate capacitance and that of capacitance between the two interconnect lines on the delay time were all examined. We conclude that the interconnect wires contribute more than 50% of the delay. By implication ,therefore, is that, if this is to be robust for big data communication, tremendous effort must be put in place to reduce the delay due to interconnection lines. The obvious reason is that big data will require interconnections of highly densified networks of heterogeneous large-scale integrated circuits. It then means, this delay will multiply exponentially and the result is that the efficiency will depreciate. The roundtrip latency will severely be affected, and data communication will not be able to be recovered due to signal distortion, the effect of which will worsen as diverse of massive of data is communicated by big data technology.

References

1. Eduardo Abreu, Michael Lightstone, Sanjit K Mitra, and Kaoru Arakawa. (1996). "A new efficient approach for the removal of impulse noise from highly corrupted images," *IEEE Transactions on Image Processing*, 5(6):1012–1025.
2. Oistamo K., Urban J., and Heintz C., (1995). "A highly integrated processor for improved quality television," IEEE Trans. Consumer Electron., 41, 421–428.
3. Hoeneisen B. and Mead C. A., (1972). "Fundamental limitations in microelctronics -I. MOS technology," Solid State Electron., 15, 819–829.
4. Chung J. E., Jeng M. C., Moon J. E., KO P. K., and Hu C., (1990). "Low-voltage hot-electron currents and degradation in deep submicron MOSFET's," IEEE Trans. Electron Devices, 37, 1651–1657.
5. Davari B., Dennard R. H., and Shahidi G. G., (1995). "CMOS scaling for high performance and low power-The next ten years," *Proc. of the IEEE*, 83(4), 595–606.
6. Sai-Halasz G. A., (1995). "Performance trends in high-end processors," *Proc. of the IEEE*, 83 (1), 20–36.
7. Saraswat K. C., and Mohammadi F., (1982) "Effect of interconnection scaling on time delay of VLSI circuits," *IEEE Trans. Electron Devices*, ED-29, 645–650.
8. Rahman A., and Reif R. (2000). "System-level performance evaluation of three-dimensional integrated circuits," IEEE Transactions on Very Large Scale Integration (VLSI) Systems 8(6), 671–678.
9. Saraswat K. C., and Mohammadi F. (2005) "Effect of scaling of interconnections on the time delay of VLSI circuits," IEEE Transactions on Electron Devices 29(4), 645–650.

Author Index

© Springer Nature Singapore Pte Ltd. 2018

J. K. Mandal et al. (eds.), *Proceedings of the International Conference on Computing and Communication Systems*, Lecture Notes in Networks and Systems 24, https://doi.org/10.1007/978-981-10-6890-4