

Bibudhendu Pati

Chhabi Rani Panigrahi · Sudip Misra

Arun K. Pujari · Sambit Bakshi *Editors*

# Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2017, Volume 1

# **Advances in Intelligent Systems and Computing**

Volume 713

## **Series editor**

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland  
e-mail: [kacprzyk@ibspan.waw.pl](mailto:kacprzyk@ibspan.waw.pl)

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

### *Advisory Board*

#### Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India  
e-mail: [nikhil@isical.ac.in](mailto:nikhil@isical.ac.in)

#### Members

Rafael Bello Perez, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba  
e-mail: [rbellop@uclv.edu.cu](mailto:rbellop@uclv.edu.cu)

Emilio S. Corchado, University of Salamanca, Salamanca, Spain  
e-mail: [escorchado@usal.es](mailto:escorchado@usal.es)

Hani Hagrais, University of Essex, Colchester, UK  
e-mail: [hani@essex.ac.uk](mailto:hani@essex.ac.uk)

László T. Kóczy, Széchenyi István University, Győr, Hungary  
e-mail: [koczy@sze.hu](mailto:koczy@sze.hu)

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan  
e-mail: [ctlin@mail.nctu.edu.tw](mailto:ctlin@mail.nctu.edu.tw)

Jie Lu, University of Technology, Sydney, Australia  
e-mail: [Jie.Lu@uts.edu.au](mailto:Jie.Lu@uts.edu.au)

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico  
e-mail: [epmelin@hafsamx.org](mailto:epmelin@hafsamx.org)

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil  
e-mail: [nadia@eng.uerj.br](mailto:nadia@eng.uerj.br)

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland  
e-mail: [Ngoc-Thanh.Nguyen@pwr.edu.pl](mailto:Ngoc-Thanh.Nguyen@pwr.edu.pl)

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong  
e-mail: [jwang@mae.cuhk.edu.hk](mailto:jwang@mae.cuhk.edu.hk)

More information about this series at <http://www.springer.com/series/11156>

Bibudhendu Pati · Chhabi Rani Panigrahi  
Sudip Misra · Arun K. Pujari  
Sambit Bakshi  
Editors

# Progress in Advanced Computing and Intelligent Engineering

Proceedings of ICACIE 2017, Volume 1

 Springer



*Editors*

Bibudhendu Pati  
Department of Computer Science  
Rama Devi Women's University  
Bhubaneswar, Odisha, India

Arun K. Pujari  
Department of Computer Science  
Central University of Rajasthan  
Jaipur, Rajasthan, India

Chhabi Rani Panigrahi  
Department of Computer Science  
Rama Devi Women's University  
Bhubaneswar, Odisha, India

Sambit Bakshi  
Department of Computer Science  
and Engineering  
National Institute of Technology, Rourkela  
Rourkela, Odisha, India

Sudip Misra  
Department of Computer Science  
and Engineering  
Indian Institute of Technology Kharagpur  
Kharagpur, West Bengal, India

ISSN 2194-5357 ISSN 2194-5365 (electronic)  
Advances in Intelligent Systems and Computing  
ISBN 978-981-13-1707-1 ISBN 978-981-13-1708-8 (eBook)  
<https://doi.org/10.1007/978-981-13-1708-8>

Library of Congress Control Number: 2018948829

© Springer Nature Singapore Pte Ltd. 2019

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.

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

# Preface

This volume contains the papers presented at 2nd International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2017: The 2nd International Conference on Advanced Computing and Intelligent Engineering ([www.icacie.com](http://www.icacie.com)) held during 23–25 November 2017 at the Central University of Rajasthan, India. There were 618 submissions and each qualified submission was reviewed by a minimum of two Technical Program Committee members using the criteria of relevance, originality, technical quality, and presentation. The committee accepted 109 full papers for oral presentation at the conference and the overall acceptance rate is 18%.

ICACIE 2017 was an initiative taken by the organizers which focuses on research and applications on topics of advanced computing and intelligent engineering. The focus was also to present state-of-the-art scientific results, to disseminate modern technologies, and to promote collaborative research in the field of advanced computing and intelligent engineering.

Researchers presented their work in the conference and had an excellent opportunity to interact with eminent professors, scientists and scholars in their area of research. All participants were benefitted from discussions that facilitated the emergence of innovative ideas and approaches. Many distinguished professors, well-known scholars, industry leaders, and young researchers were participated in making ICACIE 2017 an immense success.

We had also industry and academia panel discussion and we invited people from software industries like TCS, Infosys, and DRDO.

We thank all the Technical Program Committee members and all reviewers/sub-reviewers for their timely and thorough participation during the review process.

We express our sincere gratitude to Honourable Vice Chancellor and General Chair, Prof. Arun K. Pujari, Central University of Rajasthan to allow us to organize ICACIE 2017 on the campus and for his valuable moral and timely support. We also thank Prof. A. K. Gupta, Dean Research for his valuable guidance. We appreciate the time and efforts put in by the members of the local organizing team at Central University of Rajasthan, especially the faculty members of different departments, student volunteers, administrative, account section, guest house

management and hostel management staff, who dedicated their time and efforts to make ICACIE 2017 successful. We thank Mr. Subhashis Das Mohapatra, System Analyst, C.V. Raman College of Engineering, Bhubaneswar for designing and maintaining ICACIE 2017 Website.

We are very grateful to all our sponsors, especially DRDO for its generous support towards ICACIE 2017.

Bhubaneswar, India  
Bhubaneswar, India  
Kharagpur, India  
Jaipur, India  
Rourkela, India

Bibudhendu Pati  
Chhabi Rani Panigrahi  
Sudip Misra  
Arun K. Pujari  
Sambit Bakshi

# About This Book

The book focuses on theory, practice and applications in the broad areas of advanced computing techniques and intelligent engineering. This two volumes book includes 109 scholarly articles, which have been accepted for presentation from over 618 submissions in the 2nd International Conference on Advanced Computing and Intelligent Engineering held at Central University of Rajasthan, India during 23–25 November, 2017. The first volume of this book consists of 55 numbers of papers and volume 2 contains 54 papers with a total of 109 papers. This book brings together academic scientists, professors, research scholars and students to share and disseminate their knowledge and scientific research works related to advanced computing and intelligent engineering. It helps to provide a platform to the young researchers to find the practical challenges encountered in these areas of research and the solutions adopted. The book helps to disseminate the knowledge about some innovative and active research directions in the field of advanced computing techniques and intelligent engineering, along with some current issues and applications of related topics.

# Contents

## Part I Advanced Image Processing

<b>Patch-Based Feature Extraction Algorithm for Mammographic Cancer Images</b> . . . . .	3
P. M. Rajasree and Anand Jatti	
<b>Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Techniques</b> . . . . .	13
Preeti Joon, Shalini Bhaskar Bajaj and Aman Jatain	
<b>A Correlative Study of Contrary Image Segmentation Methods Appending Dental Panoramic X-ray Images to Detect Jawbone Disorders</b> . . . . .	25
Krishnappa Veena Divya, Anand Jatti, P. Revan Joshi and S. Deepu Krishna	
<b>Image Quilting for Texture Synthesis of Grayscale Images Using Gray-Level Co-occurrence Matrix and Restricted Cross-Correlation</b> . . . . .	37
Mudassir Rafi and Susanta Mukhopadhyay	
<b>Tongue Recognition and Detection</b> . . . . .	49
Ravi Saharan and Divya Meena	
<b>Sample Entropy Based Selection of Wavelet Decomposition Level for Finger Movement Recognition Using EMG</b> . . . . .	61
Nabasmita Phukan and Nayan M. Kakoty	
<b>Skin Detection Using Hybrid Colour Space of RGB-H-CMYK</b> . . . . .	75
Ashish Kumar and P. Shanmugavadivu	
<b>GPU-Based Approach for Human Action Recognition in Video</b> . . . . .	85
Ishita Dutta, Vikas Tripathi, Vaishali Dabral and Pooja Sharma	

## **Part II Machine Learning and Data Mining**

<b>Protein Sequence in Classifying Dengue Serotypes . . . . .</b>	<b>97</b>
Pandiselvam Pandiyarajan and Kathirvalavakumar Thangairulappan	
<b>An Assistive Bot for Healthcare Using Deep Learning: Conversation-as-a-Service . . . . .</b>	<b>109</b>
Dhvani Shah and Thekkekara Joel Philip	
<b>A Comprehensive Recommender System for Fresher and Employer . . . . .</b>	<b>119</b>
Bhavna Gupta, Sarthak Kanodia, Nikita Khanna and Saksham	
<b>A New Approach of Learning Based on Episodic Memory Model . . . . .</b>	<b>129</b>
Rahul Shrivastava and Sudhakar Tripathi	
<b>A Hybrid Model for Mining and Classification of Gene Expression Pattern for Detecting Neurodegenerative Disorder . . . . .</b>	<b>139</b>
S. Geeitha and M. Thangamani	
<b>A New Deterministic Method of Initializing Spherical K-means for Document Clustering . . . . .</b>	<b>149</b>
Fatima Gulnashin, Iti Sharma and Harish Sharma	
<b>Learners' Player Model for Designing an Effective Game-Based Learning . . . . .</b>	<b>157</b>
Lamyae Bennis, Said Benhlima and M. Ali Bekri	
<b>Reducing Time Delay Problem in Asynchronous Learning Mode Using Metadata . . . . .</b>	<b>167</b>
Barsha Abhisheka and Rajeev Chatterjee	
<b>Improved Forecasting of CO<sub>2</sub> Emissions Based on an ANN and Multiresolution Decomposition . . . . .</b>	<b>177</b>
Lida Barba and Nivaldo Rodríguez	
<b>Clustered Support Vector Machine for ATM Cash Repository Prediction . . . . .</b>	<b>189</b>
Pankaj Kumar Jadwal, Sonal Jain, Umesh Gupta and Prashant Khanna	
<b>An Effective Intrusion Detection System Using Flawless Feature Selection, Outlier Detection and Classification . . . . .</b>	<b>203</b>
Rajesh Kambattan Kovarasan and Manimegalai Rajkumar	
<b>A Novel LtR and RtL Framework for Subset Feature Selection (Reduction) for Improving the Classification Accuracy . . . . .</b>	<b>215</b>
Sai Prasad Potharaju and M. Sreedevi	

**Gradient-Based Swarm Optimization for ICA** ..... 225  
 Rasmikanta Pati, Vikas Kumar and Arun K. Pujari

**Empirical Evaluation of Inference Technique for Topic Models** ..... 237  
 Pooja Kherwa and Poonam Bansal

**Action Recognition Framework Based on Normalized Local Binary Pattern** ..... 247  
 Shivam Singhal and Vikas Tripathi

**Enhancements to Randomized Web Proxy Caching Algorithms Using Data Mining Classifier Model**..... 257  
 P. Julian Benadit, F. Sagayaraj Fancis and A. M. James Raj

**Extraction and Classification of Liver Abnormality Based on Neutrosophic and SVM Classifier** ..... 269  
 Jayanthi Muthuswamy

**Improving Accuracy of Short Text Categorization Using Contextual Information** ..... 281  
 V. Vasantha Kumar, S. Sendhilkumar and G. S. Mahalakshmi

**Efficient Classification Technique on Healthcare Data** ..... 293  
 Rella Usha Rani and Jagadeesh Kakarla

**Part III Cryptography and Information Security**

**Two-Phase Validation Scheme for Detection and Prevention of ARP Cache Poisoning** ..... 303  
 Sweta Singh, Dayashankar Singh and Aanjey Mani Tripathi

**Software-Defined Networks and Methods to Mitigate Attacks on the Network** ..... 317  
 Shubham Kumar, Sumit Kumar and Valluri Sarimela

**A Fast Image Encryption Technique Using Henon Chaotic Map** ..... 329  
 Kapil Mishra and Ravi Saharan

**A New Approach to Provide Authentication Using Acknowledgment** ..... 341  
 Vijay Paul Singh, Naveen Aggarwal, Muzzammil Hussain and Charanjeet Kour Raina

**Prevention of Replay Attack Using Intrusion Detection System Framework** ..... 349  
 Mamata Rath and Binod Kumar Pattanayak

**Appending Photoplethysmograph as a Security Key for Encryption of Medical Images Using Watermarking** ..... 359  
 M. J. Vidya and K. V. Padmaja

<b>Hierarchical Autoconfiguration Scheme for IPv6-Based MANETs</b> .....	371
T. R. Reshmi	
<b>Improved (k, n) Visual Secret Sharing Based on Random Grids</b> .....	381
Pritam Kumari and Rajneesh Rani	
<b>Efficient Motion Encoding Technique for Activity Analysis at ATM Premises</b> .....	393
Prateek Bajaj, Monika Pandey, Vikas Tripathi and Vishal Sanserwal	
<b>EKRV: Ensemble of kNN and Random Committee Using Voting for Efficient Classification of Phishing</b> .....	403
A. Niranjan, D. K. Haripriya, R. Pooja, S. Sarah, P. Deepa Shenoy and K. R. Venugopal	
<b>Enhanced Digital Video Watermarking Technique Using 2-Level DWT</b> .....	415
Rashmi Jakhmola and Rajneesh Rani	
<b>Cryptanalysis on Digital Image Watermarking Based on Feature Extraction and Visual Cryptography</b> .....	425
Neha Shashni, Ranvijay and Mainekar Yadav	
<b>A Spoofing Security Approach for Facial Biometric Data Authentication in Unconstraint Environment</b> .....	437
Naresh Kumar and Aditi Sharma	
<b>Part IV Optical and Wireless Networks</b>	
<b>Design and Implementation of OFDM Transceiver Using Different Modulation Technique over CDMA</b> .....	451
Shikha Bharti, Hemant Rathore, Arun Kumar and Manish Kumar Singh	
<b>Energy Harvesting-Based Two-Hop Clustering for Wireless Mesh Network</b> .....	463
Sudeep Tanwar, Shivangi Verma and Sudhanshu Tyagi	
<b>A Compact and High Selective Microstrip Dual-Band Bandpass Filter</b> .....	475
Dwipjoy Sarkar and Tamasi Moyra	
<b>A Reliable Routing Protocol for EH-WSAN</b> .....	483
Jagadeesh Kakarla	
<b>A Simulation Study: LMI Based Sliding Mode Control with Attractive Ellipsoids for Sensorless Induction Motor</b> .....	493
Deepika, Shiv Narayan and Sandeep Kaur	



**A Study of Environmental Impact Assessment on the Performance of Solar Photovoltaic Module** . . . . . 505  
 Sanhita Mishra, S. C. Swain, P. C. Panda and Ritesh Dash

**Design of a Compact Ultra-wideband Bandpass Filter Employing Defected Ground Structure and Short-Circuited Stubs** . . . . . 513  
 Sarbani Sen and Tamasi Moyra

**Performance Evaluation of Wireless Sensor Network in the Presence of Wormhole Attack** . . . . . 523  
 Manish Patel, Akshai Aggarwal and Nirbhay Chaubey

**Part V Social Networks and Sentiment Analysis**

**Fused Sentiments from Social Media and Its Relationship with Consumer Demand** . . . . . 531  
 Pushkal Agarwal, Shubham Upadhyaya, Aditya Kesharwani and Kannan Balaji

**A Prototype for Semantic Knowledge Retrieval from Educational Ontology Using RDF and SPARQL** . . . . . 541  
 S. Mahaboob Hussain, Prathyusha Kanakam and D. Suryanarayana

**Social Trust Analysis: How Your Behavior on the Web Determines Reliability of the Information You Generate?** . . . . . 551  
 Rhea Sanjay Sukthanker and K. Saravanakumar

**Automatic Emotion Classifier** . . . . . 565  
 Hakak Nida, Kirmani Mahira, Mohd. Mudasir, Muttoo Mudasir Ahmed and Mohd. Mohsin

**Sentiment Analysis of Tweets Through Data Mining Technique** . . . . . 573  
 Taranpreet Singh Ruprah and Nitin Trivedi

**UPLBSN: User Profiling in Location-Based Social Networking** . . . . . 581  
 G. U. Vasanthakumar, G. R. Ashwini, K. N. Srilekha, S. Swathi, Ankita Acharya, P. Deepa Shenoy and K. R. Venugopal

**Performing Interest Mining on Tweets of Twitter Users for Recommending Other Users with Similar Interests** . . . . . 593  
 Richa Sharma, Shashank Uniyal and Vaishali Gera

**Author Index** . . . . . 605

## About the Editors

**Dr. Bibudhendu Pati** is Associate Professor Department of Computer Science at Rama Devi Women's University, Bhubaneswar, India. He has around 21 years of experience in teaching and research. His areas of research interests include Wireless Sensor Networks, Cloud Computing, Big Data, Internet of Things, and Advanced Network Technologies. He completed his Ph.D. from IIT Kharagpur. He is a Life Member of Indian Society for Technical Education, Computer Society of India and Senior Member of IEEE. He has got several papers published in reputed journals, conference proceedings, and books of international repute. He also served as Guest Editor of IJCND and IJCSE journals. He was the General Chair of ICACIE 2016, ICACIE 2018 and IEEE ANTS 2017 international conference.

**Dr. Chhabi Rani Panigrahi** is Assistant Professor in the Department of Computer Science at Rama Devi Women's University, Bhubaneswar, India. She completed her Ph.D. in the Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, India. Her areas of research interests include Software Testing and Mobile Cloud Computing. She holds 17 years of teaching and research experience. She has published several international journals and conference papers. She is a Life Member of Indian Society for Technical Education (ISTE) and Member of IEEE and Computer Society of India (CSI). She also served as Guest Editor of IJCND and IJCSE journals. She was the Organizing Chair of ICACIE 2016, ICACIE 2017, ICACIE 2018, and WiE Co-chair of IEEE ANTS 2017 and IEEE ANTS 2018.

**Dr. Sudip Misra** is Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology Kharagpur. Prior to this he was associated with Cornell University (USA), Yale University (USA), Nortel Networks (Canada) and the Government of Ontario (Canada). He received his Ph. D. degree in Computer Science from Carleton University, in Ottawa, Canada. He has several years of experience working in the academia, government, and the private sectors. His current research interests include Wireless Sensor Networks, Internet of Things (IoT), Software Defined Networks, Cloud Computing, Big Data

Networking, Computer Networks. Dr. Misra is the author of over 260 scholarly research papers, including 150+ reputed journal papers. Dr. Misra has published 9 books in the areas of Advanced Computer Networks.

**Prof. Arun K. Pujari** faculty and Dean of the School of Computer and Information Sciences at the University of Hyderabad (UoH) and has been appointed as the Vice-Chancellor of the Central University of Rajasthan. Professor Pujari has earlier served as the Vice-Chancellor of Sambalpur University, Odisha in 2008. Professor Pujari got his Ph.D. from IIT-Kanpur in 1980. He has more than 15 years' experience as Dean and Head of Department. He has served as a member of high-level bodies such as UGC, DST, DRDO, ISRO and AICTE. Professor Pujari has over 100 publications to his credit and has wide exposure in national and international arena. His two books published are *Data Mining Techniques* and *Database Management System*.

**Dr. Sambit Bakshi** is Assistant Professor in the Department of Computer Science and Engineering at NIT Rourkela, Odisha. He has received his Ph.D. degree from NIT Rourkela. His research interests include Biometric Security, Visual Surveillance, Medical Signal Processing and Social Security Analytics. He has published several journal and conference papers of international repute. He is a Life Member of CSI, IEEE and other technical societies. He is also the editor and associate editor of several reputed international journals.

**Part I**  
**Advanced Image Processing**

# Patch-Based Feature Extraction Algorithm for Mammographic Cancer Images



P. M. Rajasree and Anand Jatti

**Abstract** The study of mammography aims at identifying the presence of cancerous or non-cancerous tissue by using signs of bilateral asymmetry, masses, calcification and architectural distortion. The most vigilant one among them is the architectural distortion owing to speculated or random patterns. In this paper, a novel method for pectoral muscle removal and annotation removal is explained. A patch-based algorithm is implemented to extract textural features, and according to the features, a neural classifier has been classified into benign or malignant. The method was experimented on 88 images from MIAS database, and the proposed method has a total efficiency of 92.04% with respect to pectoral muscle and annotation removal.

**Keywords** Benign · Malignant · Pectoral muscle · Patch-based feature extraction

## 1 Introduction

Among all the developed and non-developed countries, breast cancer is the most common cancer in women, and though, around 58% of deaths were observed in developed countries [1]. Incident rate is below 40% for every 100,000 people in developing regions. The survival rate in low-income nations is below 40% [2]. Low survival rate indicates a lack of early detection facility or adequate facilities in terms of both instruments and personnel, causing the cancer detected at a very late stage. Breast cancer is fully treatable when detected at an early stage, and still mammography is a very effective modality to detect the cancer at an early stage for those who do not have symptom. Though every year the mortality rate count keeps increasing, some studies reported around 25% breast cancer remains undetected at an early stage [3]. The variability in detection occurs because mammographic examination is difficult to study, especially when positive signs are hidden by superimposition of soft

---

P. M. Rajasree (✉) · A. Jatti  
Department of Electronics & Instrumentation Engineering, R.V. College  
of Engineering, Bengaluru 560059, India  
e-mail: rajasreepm@rvce.edu.in

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_1](https://doi.org/10.1007/978-981-13-1708-8_1)

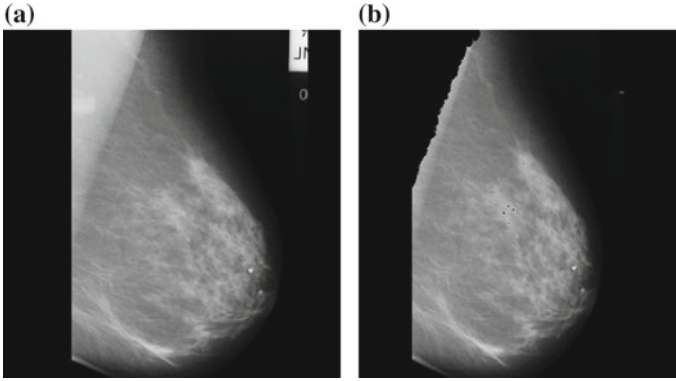
tissues. Second opinion is one of the very commonly followed procedures, and it has been very efficient [4] in curbing the miss prediction occurring in the diagnosis. Though second opinion could also have issues of its own [5] including human error, a diagnosis supported by an objective analysis could help reducing the mortality rate to a greater extent. Though computer-aided systems (CAD) cannot replace a doctor's diagnosis, it always aid in increasing the diagnosis capability. Development of CAD in mammogram has proven very effective in determining masses and calcification [6], though certain cases such as architectural distortion and ill-defined masses have reduced the accuracy and precision of identifying the malignancy [7]. Many researchers have proposed methods to categorize benign and malignant. Radovic et al. [8] proposed a method based on mass segmentation and feature extraction which provided a sensitivity of 77%. Swapnil et al. [9] proposed a method based on grid-based textural extraction which provided an efficiency of 91% for identifying the malignancy in the region of interest. Malkov et al. [10] proposed a method which used fractal dimension and statistical features corresponding to second order. With the aim of developing a novel method, a patch-based method is used. And gray-level co-occurrence matrix (GLCM) is used to extract the features.

## 2 Methodology

In the first section, the process of pectoral muscle and annotation is explained. In the second section, a patch-based method is implemented. The main concept of the patch-based method is to extract all the patches which are very small compared with the original image size with overlaps from the given image. Then, the interrelations between those patches are found out. In the patch-based method, there is an expectation that every patch taken from the image may find similar ones elsewhere in the image. Three texture features are extracted using GLCM method. Images are taken from MIAS database. Neural classifier is used.

### 2.1 *A Novel Method for Pectoral Muscle Removal and Annotation Removal*

The initial step in the dictionary-based learning is removal of annotation and pectoral muscle from mammogram images. Existing method of patch learning increases the CPU memory utilization and compilation time. Another method of connected component labeling requires a threshold value to be used which is ineffective in most cases. Hence, a different method is proposed in Algorithm 1 for removal of annotation and pectoral muscle. The first step involves the identification of right mediolateral oblique (MLO) or left MLO by the method of clustering to find the placement of the pectoral muscle. The second step involves removal of the annotation if it exists. Then, the normalization of the image is done to improve the range of pixel intensity values of poor contrast regions. Depending on the normalization of images, clustering is carried out at fourth step of the algorithm starting with two clusters. Every cluster



**Fig. 1** Mammogram image **a** before preprocessing (left), **b** after preprocessing (right)

has some dense components of pectoral muscle restored. Therefore, if the density is above a threshold, only then the cluster is considered and otherwise neglected. The mean value for the clusters is evaluated, and only the highest mean value in the cluster is retained. Then, the morphological operation is performed at each cluster stage with a disk as structuring element. The small blobs in the image are removed and remaining are retained. All the clusters crossing the threshold are combined using a mathematical operator. But because breast tissue is present, there could be a chance of other regions being affected in the process; as we observe morphological character, pectoral muscle shows an appearance of triangle, and as seen on the top half, this allows us to consider only half of the image in horizontal axis. The detailed algorithm is given in the following section, and the result is shown in Fig. 1.

#### Algorithm 1: Pectoral muscle and annotation removal

- 
1. Identify if Right(R) MLO or Left(L) MLO
    - Initialize  $I = 1 : M$   
 $J = 1 : N/2$  increment of +1
    - where  $m$  &  $n$  are the rows and column
 
$$\begin{cases} I_K = (I+1)_K \\ J_K = (J+1)_K \end{cases}$$
    - check if  $P(I, J)_k = 0$  for image P  
 if  $P(I, J) = 0$  at  $k$  then
    - check if  $P(I_K, J_K) = 0$   
 for  $I_K = I_k + 1$  and  $k = 1 : M$
    - if condition is true then it is a RMLO  
 else follow the same step with
    - Initializing  $I = 1 : M$   
 $J = N : -1 : N/2$  increment of -1
    - if condition is true then it is a LMLO

---

2. Once RMLO ( $P_R$ ) or LMLO ( $P_L$ ) is identified then annotation removal is done by

For RMLO,

$$P_R = \sum_{i=1}^{M/3} \sum_{j=j_K}^N P_R(i,j)=0 \quad (1)$$

And for LMLO,

$$P_L = \sum_{i=1}^{M/3} \sum_{j=j_K}^1 P_L(i,j)=0 \quad (2)$$

where  $j=1: (-1):J_k$

As it is observed annotation is always present on the Right top side for RMLO and on the left top side for LMLO image

---

3. P is the image which could be either  $P_R$  or  $P_L$  with dimension  $m \times n$

$$O_{\max} = \sum_{i=1}^M \sum_{j=1}^N \max P(I,J) \quad (3)$$

$$O_{\min} = \sum_{i=1}^M \sum_{j=1}^N \min P(I,J) \quad (4)$$

$$S_{(foreachi,j)} = Value(i,j) - O_{\min} / O_{\max} - O_{\min}$$

$$S_{(new)} = (newrange \times S_{(foreachi,j)} + S_{(newmin)})$$

$$G_{(i,j)} = (S_{(\max)} - S_{(\min)}) \times P(i,j) - I_{(\min)} / O_{\max} - O_{\min} \quad (5)$$

$O_{\max}$  and  $O_{\min}$  is from intensity levels of the original image,  $S_{\max}$  and  $S_{\min}$  from normalized image.

---

$$4. T_i = (T_{i\text{white}} / T_{\text{total}}) \times 100 \text{ for } i=2,3,4,5 \quad (6)$$

$T_{i\text{white}}$  is the total pixels in the upper half section which is a part of cluster and  $T_{\text{total}}$  is the image area in total and  $I$  is the cluster numbers

If  $T_i > 50\%$

$\forall i = 2, 3, 4, 5$  then

$$G = (G_T)^{\forall i=2,3,4,5} \otimes (G_T)^{\forall i=3,4,5} \otimes (G_T)^{\forall i=4,5} \otimes (G_T)^{\forall i=5}$$

If  $T_i < 50\%$

$$(G_T) = 0 \quad \forall_{i=2,3,4,5} \quad (7)$$

- Clear T

The region G is a binary mask, a 3<sup>rd</sup> order polynomial equation gives the curve estimation of pectoral muscle.

---



## 2.2 Patch Generation

Consider an grayscale image  $G$  with size  $m \times n$  and  $G(x, y)$  as coordinate. Define  $R$  as an operator for extracting patch that returns the image as patches with size  $M_p \times N_p$  as derived by Eq. (8). Total  $G_X$  patches are stacked together as vector defined by  $X(t)$ , where  $t$  is the length of the stack.

$$R \rightarrow M_p = N_p = \begin{cases} if \\ T \bmod B = 0 \forall B=2,3,5 \end{cases} \quad (8)$$

If  $T \bmod B = 0 \forall B = 2$   
 $M_p = N_p = 2; t = (m \times n)/B^2$

As the window of  $M_p \times N_p$  derived from Eq. (8) is moved over the image, patches ( $T_p$ ) are generated and simultaneously feature values are extracted with labeling which is explained in Sect. 2.3.

## 2.3 Patch-Based Dictionary: Feature Extraction

The patch is generated using  $R$  operator which consists of “ $t$ ” number of patches. Some patches “ $t_{NON}$ ” would not be relevant as it could belong to the area which is the background, marked by black in the image, and some patches would include the border region “ $t_{BR}$ ” of the RMLO or the LMLO where it is highly unlikely to get a benign or malignant tissue. To identify  $T_{NON}$ ,  $T_{BR}$  black pixel density needs to extract which is calculated in Eq. (9).

$$\forall T_z; z = 1, 2, 3 \dots t; \text{ where } t = (m \times n)/B^2, \\ \text{and } T_{NON}, T_{BR} \in T$$

$$D_z = (T_z(BLACK\ AREA)/(T_{total\ AREA})) \times 100,$$

$$\begin{aligned} T = T_{NOND} &\geq 90 \\ T_{BRD} &\geq 40\% . \\ T &\text{ otherwise} \end{aligned} \quad (9)$$

Only  $T$  with  $(x1, y1, x2, y2)$  is considered and other patches are neglected where  $x1, y1, x2, y2$  are coordinates of the patches extracted from  $G$  during the process of patch creation.

$$T_{CORD} = [x1, y1, x2, y2]. \quad (10)$$

To make a dictionary-based system, the texture features [11] are extracted using gray-level co-occurrence matrix (GLCM). The neighborhood pixel distance is taken as 1, and directions  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  are considered as only one direction might not be reliable information. Homogeneity, contrast and energy are chosen as descriptors. These textural descriptors are chosen because homogeneity shows the closeness in pixel intensity distribution, while local variation is derived from contrast and correlation provides information of highly correlated neighboring pixels. Energy determines the pixel repetitions occurring in pairs and hence detects disorders in texture. The values extracted are a single patch stored as an array of dimension  $1 \times 4$ , and for a set of 100 patches, the feature dimension would be  $100 \times 4$ . To convert the patches' features into dictionary-based system, we use the prior information for training the system using coordinates given in the MIAS database where  $x$  and  $y$  are the coordinates provided by database and  $r$  is the radius of tumor.  $I_{CORD}$  is defined by Eq. (11).

$$\begin{aligned} p1 &= x - r; z1 = 1000 - y; p2 = z2 = 2 \times r \\ I_{CORD} &= [p1, z1, p2, z2]. \end{aligned} \quad (11)$$

The patch  $T$  is distributed into three sections  $T_N$  for normal,  $T_M$  for malignant and  $T_B$  for benign which are labeled as 0, 1 and 2, respectively.

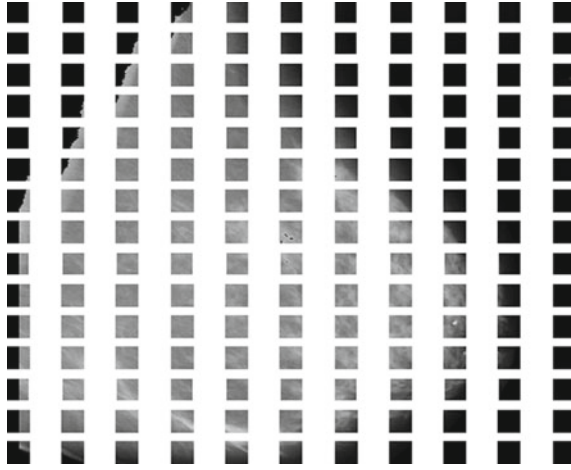
Equation (11) and prior label instruction from MIAS database are used to label the learning phase

$$T = \begin{cases} T_M & \text{if } I_{CORD} \equiv T_{CORD} \text{ and } label = M \\ T_B & \text{if } I_{CORD} \equiv T_{CORD} \text{ and } label = B \\ T_N & \text{otherwise} \end{cases} . \quad (12)$$

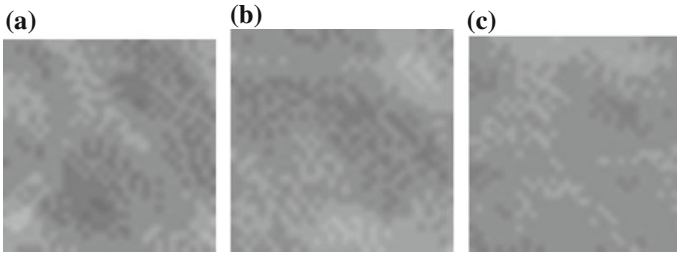
The values extracted forms the dictionary which are trained to a neural classifier with five hidden layers. Any image is considered with the same procedure as in Algorithm 1, and then,  $T_{NON}$ ,  $T_{BR}$  from the region are filtered. The filtered patches are prepared for feature extraction, and then, the patches are tested for a potential malignant or benign site in the image.

### 3 Results

To check the performance of the proposed method, experiments on MIAS database images were carried out [12]. The initial evaluation has been done to remove pectoral muscle and annotation if it exists. The technique used demonstrates its effectiveness in eliminating the unwanted region as shown in Fig. 1. The experiments were carried out to test 88 images, and it is observed that for 81 images, pectoral muscle and annotation were removed effectively. Among them, two images could not produce a satisfactory result in removing pectoral muscle; three images left behind some portion of annotation due to its placement of annotation on the center rather than on



**Fig. 2** Patch T which contains  $T_{NON}$   $T_{BR}$

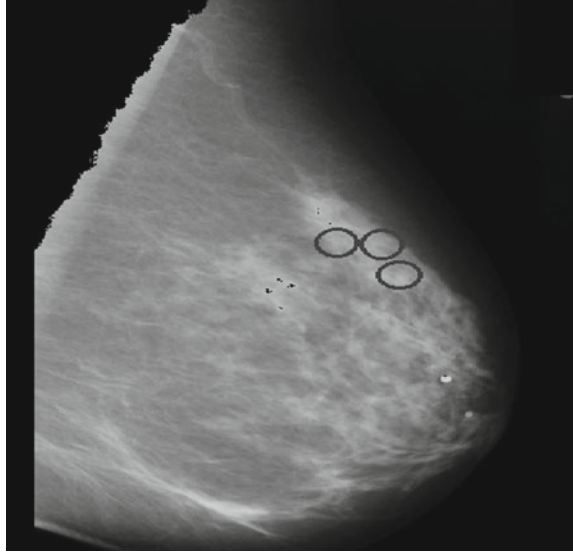


**Fig. 3** Patches extracted to be potential benign site and the patch (a–c)

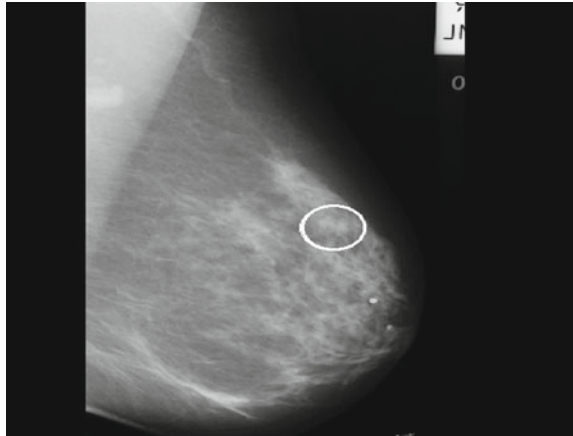
the top; and for two images, there was a hindrance due to breast tissue segmented along with pectoral muscle as those tissues had a dense structure which resembles the pectoral muscle; hence, the cluster algorithm showed a slighter ineffectiveness in the case where breast tissue is denser and looks alike to pectoral muscle. A total efficiency of 92.04% was observed with respect to pectoral muscle and annotation removal.

The algorithm for patch is evaluated, and the results for the same are shown in Fig. 2 and it was observed that one patch creation took about 0.003337 s with a size of 1.85 KB per patch; hence, it shows that the pectoral muscle removal and the annotation removal play an important role in faster execution and using memory effectively. It was observed that a total of 45 patches were classified as  $T_{NON}$  and 17 patches were classified as  $T_{BR}$ . The images generated from patches were trained using a neural classifier where a set of 14,025 patches were trained from 81 images which had normal, benign and malignant patches and validation showed an accuracy of 91.36%.

**Fig. 4** Location of tissue with respect to patch on the pectoral, annotation segmented image



**Fig. 5** Region of interest on the original image with respect to the MIAS database annotation as given by database



A new set of 21 data from MIAS database were used for evaluation, and the result for one of the images is shown in Fig. 3 where three patches were found to have signs of benign tissue, and Fig. 4 shows the mapping of the patch in the original image as per Eqs. (10) and (11). Overall accuracy of 85.7% for 21 test data was observed when tumor region was cross-checked against the annotations provided. Hence, the method shows its effectiveness in detecting tumor region being benign or malignant. Figure 4 shows the comparison with default annotation as per MIAS database and performance of this algorithm (Fig. 5).

It can be observed that this algorithm is able to analyze and detect the suspected region without the need of any tissue segmentation, which is a huge challenge, considering the texture of mammogram image.

## References

1. Breast cancer: prevention and control (n.d.). <http://www.who.int/cancer/detection/breastcancer/en/index1.html>. Accessed 1 Feb 2017
2. Coleman, M.P., Quaresma, M., Berrino, F., Lutz, J.M., De Angelis, R., Capocaccia, R., Baili, P., Rachet, B., Gatta, G., Hakulinen, T., Micheli, A., Sant, M., Weir, H.K., Elwood, J.M., Tsukuma, H., Koifman, S.E., Silva, G.A., Francisci, S., Antaquilani, M., Verdecchia, A., Storm, H.H., Young, J.L., CONCORD Working Group: Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol.* **9**(8), 730–756 (2008)
3. Breast Cancer Overview: Risk Factors, Screening, Genetic Testing, and Prevention (1 June 2015). <http://www.cancernetwork.com/cancermanagement/breast-erview/article/10165/1802560>. Accessed 3 Feb 2017
4. Lorenzen, J., Finck-Wedel, A.K., Lisboa, B., Adam, G.: Second opinion assessment in diagnostic mammography at a breast cancer centre. *Geburtshilfe und Frauenheilkunde* **72**(8) (2012)
5. Okamoto, S., Kawahara, K., Okawa, A., Tanaka, Y.: Values and risks of second opinion in Japan's universal health-care system. *Health Expect. Int. J. Public Particip. Health Care Health Policy* **18**(5), 826–838 (2015)
6. Baker, J.A., Rosen, E.L., Lo, J.Y., Gimenez, E.I., Walsh, R., Soo, M.S.: Computer-aided detection (CAD) in screening mammography: sensitivity of commercial CAD systems for detecting architectural distortion. *Am. J. Roentgenol.* **181**(4), 1083–1088 (2003)
7. Morton, M.J., Whaley, D.H.B., Kathleen, R., Amrami, K.K.: Screening mammograms: interpretation with computer aided detection prospective evaluation. *Radiology* **239**(2), 375–383. PMID: 16569779.2006
8. Radovic, M., Milosevic, M., Ninkovic, S., Filipovic, N., Peulic, A.: Parameter optimization of a computer-aided diagnosis system for detection of masses on digitized mammograms. *Technol. Health Care-Off. J. Eur. Soc. Eng. Med.* **27**; **23**(6), 757–774 (2015)
9. Swapnil, P., Pandey, E., Yathav, J.R., Baig, A., Bailur, A.: Region marking and grid based textural analysis for early identification of breast cancer in digital mammography. In: *IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, pp. 426–429 (2016)
10. Malkov, S., Shepherd, J.A., Scott, C.G., et al.: Mammographic texture and risk of breast cancer by tumor type and estrogen receptor status. *Breast Cancer Res. BCR* **18**, 122 (2016)
11. Kashou, N.H., Smith, M.A., Roberts, C.J.: Ameliorating slice gaps in multislice magnetic resonance images: an interpolation scheme. *Int. J. Comput. Assist. Radiol. Surg.* (2014)
12. MIAS database. <http://pejpa.essex.ac.uk/info/mias.html>

# Segmentation and Detection of Lung Cancer Using Image Processing and Clustering Techniques



Preeti Joon, Shalini Bhaskar Bajaj and Aman Jatain

**Abstract** Lung cancer is a most common disease nowadays, so to get rid of it we have made a detection system. In this paper, an active spline model is used to segment the X-ray images of lung cancer. The system formed acquired medical images of lung X-ray. First, in preprocessing median filter is used for noise detection. Then, segmentation is applied and further K-mean and fuzzy C-mean clustering is applied for feature extraction. This paper is an extension of techniques of image processing of lung cancer detection and produces the final results of feature extraction after X-ray image segmentation. Here, the proposed model is developed using SVM algorithm used for classification. Using MATLAB, simulation results are obtained for cancer detection system. This paper focuses thus on segmentation and detection of lung cancer by finding normality and abnormality of the images.

**Keywords** Median filtration · Segmentation · Active spline model · Clustering Feature extraction · Support vector machine · X-ray images

## 1 Introduction

Cancer is a common disease, which is formed by different divisions of abnormal cells in a region of human body. Cancer can occur almost anywhere in the human body. The number is in centuries nowadays of cancer. Lung cancer is today the most common cancer which originates in different cells of lung. Another name of lung cancer is lung carcinoma [1]. Lung cancer is also a malignant tumor, which means it can spread to other parts of body easily. The initial stage of lung cancer starts from

---

P. Joon (✉) · S. B. Bajaj · A. Jatain  
Department of Computer Science, Amity University, Gurugram, Gurugram, India  
e-mail: Joon.preeti6@gmail.com

S. B. Bajaj  
e-mail: shalinivimal@gmail.com

A. Jatain  
e-mail: amanjatainsingh@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_2](https://doi.org/10.1007/978-981-13-1708-8_2)

lung. Lung cancer has been further classified into two parts: small cell lung cancer and non-small cell lung cancer. Lung cancer is increasing day by day all over world in both males and females. Thus, lung cancer results in abnormality of the cell, where cell is a basic unit of life [1]. The early detection of lung cancer leads to a higher chance of successful treatment. Mostly, lung cancer affects males and females all over the world due to smoking, alcohol, etc.

In this paper, we have collected X-ray images of lungs from one of the hospitals. Thus, the main aim of this paper is to detect lung cancer and find normality and abnormality of X-ray images by different techniques such as filtration, segmentation process, clustering algorithms and SVM techniques [2].

## 2 Paper Preparation

Previously, many lung cancer detection techniques depend on human experience and image of CT scan observed by them. But we find hardly any paper of lung detection on X-ray images. So we have worked on lung X-ray images. Using image processing and clustering techniques, we can quickly and accurately detect lung cancer of X-ray images, and segmentation is done by active spline model. Explanation of some surveyed papers is as follows:

“Bhagyashri G. Patil and Prof. Sanjeev N. Jain” [3]: The paper uses two methods thresholding and watershed for segmentation used to detect CT images of lung region. The aim of this paper is to detect cancer as early as possible.

“Mr. Vijay A. Gajdhane and Prof. Deshpande L.M” [4]: The main aim of this paper is to find different levels of lung cancer. Many CT-scanned images are used and detected by image processing techniques.

“K. Kaviarasu, V. Sakthivel” [5]: This paper uses CT-scanned images. First image is segmented by using clustering techniques such as K-means and fuzzy C-means. Further, for cancer detection different image processing techniques are used like thresholding, etc.

“Santhosh T, Narasimha Prasad L V” [6]: This paper uses PET images of lung and it also uses fuzzy C-means clustering technique to find the cancerous part in image.

“P. Thangaraju, N. Mala” [7]: This paper proposes the study of lung cancer tumor using accurate image segmentation techniques. The proposed model compares lung tumor using three algorithms, namely K-harmonic means, expectation maximization and hierarchical clustering, using images.

“Joel George R, Anitha Jeba Kumari D” [8]: This paper uses optimization for segmentation of lung image suffering from cancer. After image processing techniques are applied on lung cancerous image like thresholding, then images are sorted according to their clusters, and further, for segmentation fuzzy C-means and particle swarm optimization are used.

“Preeti Joon, Aman Jatain, Shalini Bhaskar Baja” [9]: The main intention of this paper is early lung cancer detection as it increases the chance of survival among people. This paper first discusses the preprocessing techniques, and image segmen-

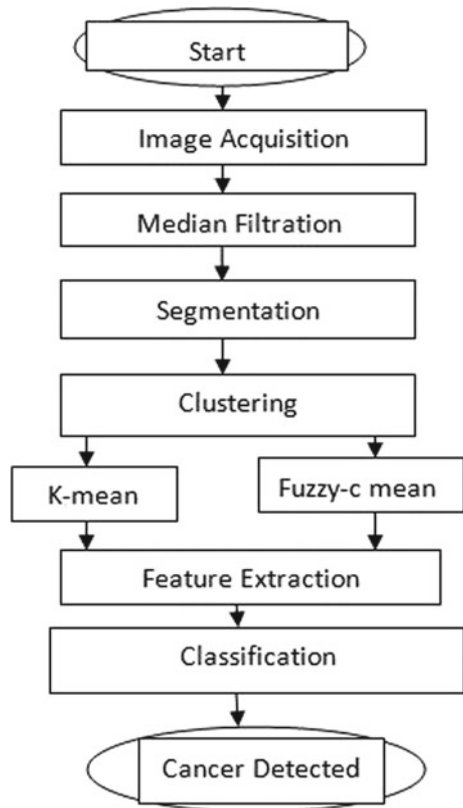
tation techniques have been used. This paper also finds the feature extraction and classification of data sets used in papers from year 2011 to 2017.

### 3 Figures

In our proposed methodology, we have taken X-ray lung images. Methodology is composed basically of two phases.

- 1 First phase: The X-ray image is preprocessed to remove noise by median filtration. Segmented image gives better accurate result. So, image is cleaned and segmented by using clustering algorithms. And then textural and features are taken or extracted from segmented image by the application of feature extraction techniques [10].
- 2 Second phase: Second phase uses the SVM classifier, which conveys whether lung is normal or abnormal as shown in Fig. 1.

**Fig. 1** Flowchart of lung cancer detection system





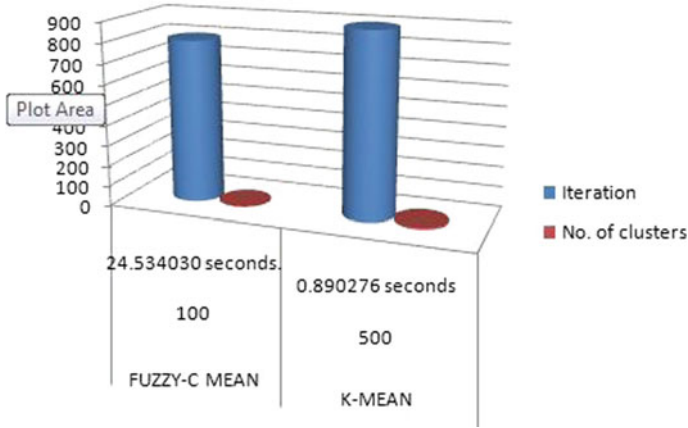


Fig. 2 Performance parameter of two clustering algorithms

### 3.1 Image Acquisition

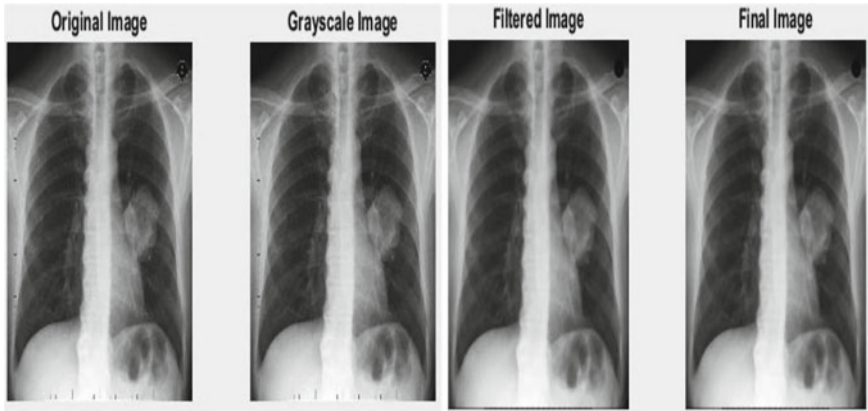
The first step in the proposed methodology is to capture the many normal and abnormal samples from the digital camera. The sample is captured from the digital camera, and the features are then stored in the database (Fig. 2).

### 3.2 Preprocessing

When images are captured by a digital camera, they are in RGB color. MATLAB does not support any image in RGB format. So, first step is to change these RGB color images to grayscale images. Then, second step is to remove noise which is a very common problem found in images like white noise and salt-and-pepper noise which is removed by median filtration as shown in Fig. 3 [4].

### 3.3 Segmentation

Segmentation is a process which means to segment or divide a image into different parts as it adds more meaning to image and is easy to analyze and also make image more simpler [4]. Segmentation is done by clustering techniques and also by image processing techniques, but in this paper we have applied segmentation on X-ray image of lung and focused on the point distribution model which is an active spline model.

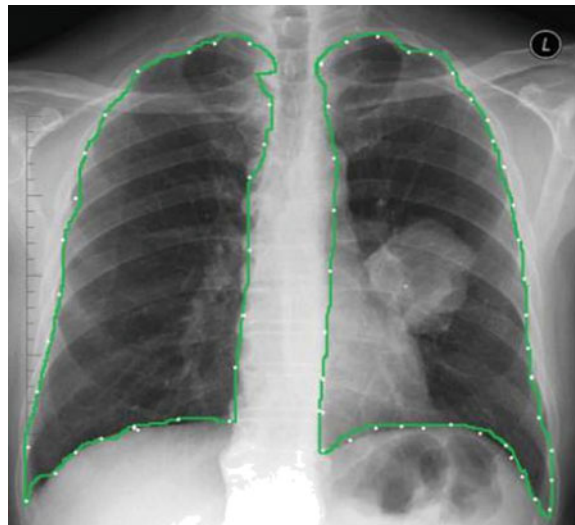


**Fig. 3** Grayscale and filtered image

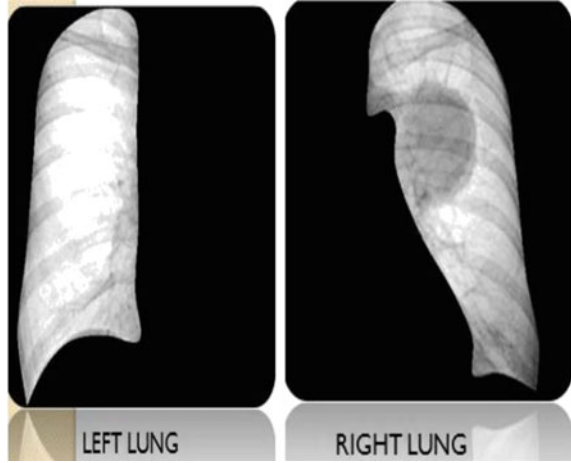
### ***3.4 Active Spline Model***

The segmentation also forms a method which is a combination distribution of point model and centripetal parameterized Catmull–Rom spline model. This method is termed as active spline. Active spline model works only on simple and small mouse operations [11]. For automated segmentation, it points in circle form almost the needed region as shown in Fig. 4. And finally, we get segmented X-ray image as left and right lung image as shown in Fig. 5.

**Fig. 4** Automated segmentation



**Fig. 5** Segmented image



### 3.5 Clustering

Cluster includes data of many objects which can be of same types which are “similar” or can be of different kinds which are “dissimilar.” Clustering is system of putting different objects into different groups based on their similar and dissimilar clusters [12]. These are classified as follows:

#### Fuzzy C-means Clustering

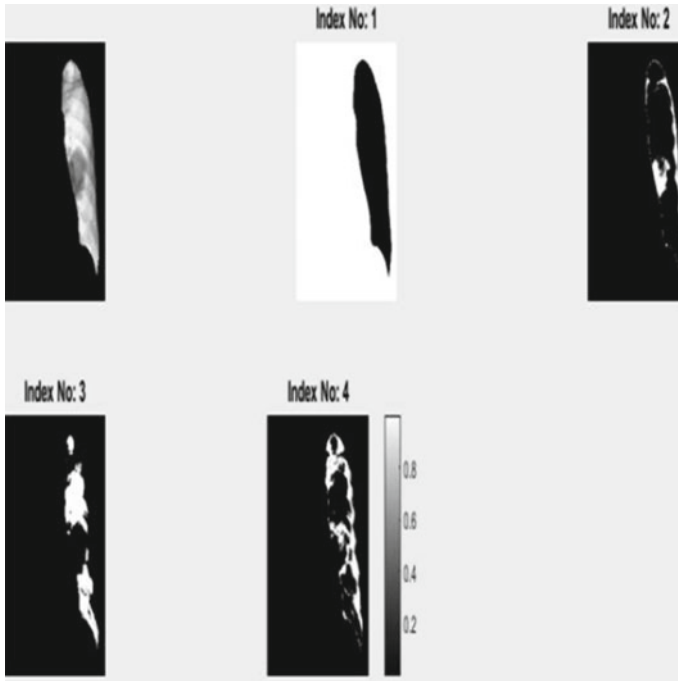
The fuzzy C-mean clustering is a type of algorithm. Fuzzy C-mean is an unsupervised clustering algorithm. This algorithm can be used as a classifier and for clustering designing. FCM also helps to find clusters in data. It is different from K-means clustering as it takes more time in process. In 1973, FCM was developed by Dunn and further proceeded in 1981 by Bezdek [13]. The advantages of fuzzy C-mean are as follows:

#### Advantages

1. Fuzzy C-means give better result than K-means clustering.
2. The data point completely belongs to single center of cluster, whereas in FCM the data point gives membership to each center of cluster.

Fuzzy C-means only allow one data to more than two clusters. It includes the following function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty.$$



**Fig. 6** Fuzzy C-means result

In above equation,  $m$  is known as real number which is more than 1 and also known as  $\|*\|$  which is any form denoting the similarity between any measured data and the center. And we have used four clusters for four images of right lung. And also elapsed time calculated in fuzzy C-mean is 24.534030 s as shown in Fig. 6.

As fuzzy C-mean is very complicated and takes much time to give result, we also apply K-mean algorithm.

### **K-means Clustering**

Simplest algorithm is only K-means which is used for collecting clusters. It is unsupervised learning. K-mean allows only one data for one cluster. This clustering does not give good result as fuzzy C-mean but gives faster result than fuzzy C-mean [8].

The K-means algorithm has the following properties:

1. K-means work better in processing many data sets.
2. It mainly completes at most favorable point.
3. It also works different on numerical values and expressions.
4. And the clusters are in shape like sphere or convex lens [14].

K-mean clustering is the simplest unsupervised learning algorithm. The objective function is:

$$J = \sum_{j=1}^k \sum_{i=1}^x \left\| x_i^{(j)} - c_j \right\|^2.$$

In above equation,  $\left\| x_i^{(j)} - c_j \right\|^2$  which is used to choose distance between data and center of cluster.

And we have used five classes for right lung image. And also elapsed time calculated in K-mean is 0.890276 s. Thus, it shows K-mean gives better and fast result as shown in Fig. 7.

## 4 Feature Extractions

It is very important stage in image processing. Cancer nodule always carries a large number of features. It is essential to extract interesting features from it to define shape of nodule uniquely. There are basically two types of descriptors as: textural and structural features [15]. Extracted features which are basically from affected region are classified on the following basis:

- (a) Area: Area is the number of the pixels present in the tumor region. Area has only magnitude but no direction.
- (b) Perimeter: Perimeter is the number of pixels which are linked or connected to each other on edge of tumor.
- (c) Eccentricity: Eccentricity is the roundness or circularity which is less than one for the circular or other shape [2].
- (d) Time: Time is taken by both clustering and iterations which are compared as shown in Fig. 2.

## 5 Classification

After feature extraction, we have classified the features into normal and abnormal lung, i.e., cancerous and non-cancerous types. Hence, for classification we have SVM classifier.

**Fig. 7** K-mean result

### 5.1 SVM Classifier

SVM is a support vector machine. It is supervised machine learning algorithm. To analyze the different data and to recognize different patterns for the classification work, supervised learning model with learning algorithms is used which is also known as SVM (support vector machine). Here, in this thesis main work of SVM is to find whether the lung images are normal or abnormal. SVM shows that particular lung image is normal, which means image does not have cancer and it gives the negative result, whereas SVM shows that lung image is abnormal, which means image has cancer and it gives the positive result.

SVM is also defined by separating hyperplane. Hyperplane separates the space into two half spaces as shown in Fig. 8.



**Fig. 8** Classification by SVM (abnormal image)

**Table 1** Performance parameter of two algorithms

Factors	Fuzzy C-mean	K-mean
Number of instances	100	500
Time (s)	24.534030	0.890276
Iteration	800	900
Number of clusters	4	10

## 6 Results and Discussion

The proposed detection of lung cancer is trained by taking three abnormal images and 80 normal lungs X-ray images in JPEG format. The database is taken from RJ Hospital. The trained system is tested by the experts of hospital. First, median filtration is applied. Then, segmentation is applied through active spline model, and then, by clustering techniques cells are separated. Thus, features are extracted and then classification is applied through SVM classifier on the images in GUI (Graphical user interface) in MATLAB, which shows the normality or abnormality of lung (Table 1).

## 7 Conclusions

An attempt is made to expose lung cancer using image processing algorithms and clustering algorithms. Initially, the X-ray image is captured and processed, and their cancer or tumor region is identified correctly from the original image. Then, in preprocessing stage median filtration is used to avoid salt-and-pepper noise of lung

image. After preprocessing, image segmentation is done through active spline model. Then, after segment of images clustering is applied for separating cells and different types of region so K-mean and fuzzy C-mean are applied for finding cancerous cell for features extraction using parameters such as area, shape and size of nodule. They help to identify different dimensions and mark boundaries of cancerous cell. Then, classifiers are proposed. In classification, support vector machine classifier is applied. Support vector machine shows which lung image is normal and which is abnormal. Hence, we conclude that from all above techniques and algorithms we can find normal and abnormal lung and can extract their features.

**Acknowledgements** Initially, we would like to thank our almighty in the success of completing this work. We want to thank the RJ Superspecialities Hospital & Heart Center for supporting and providing large collection of lung X-ray images, which have been valuable for this research.

## References

1. Thangaraju, P., Mala, N.: Segmentation of lung tumor using clustering techniques. *IJSART* **1**(8) (2015). ISSN: 2395-1052
2. Deshpande, A.S., Lokhande, D.D., Mundhe, R.P., Ghatole, J.M.: Lung cancer detection with fusion of CT and MRI images using image processing. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **4**(3) (2015)
3. Patil, B.G., Jain, S.N.: Cancer cells detection using digital image processing methods. *Int. J. Latest Trends Eng. Technol. (IJLTET)* **3** (2014). ISSN: 2278-621X
4. Gajdhane, V.A., Deshpande, L.M.: Detection of lung cancer stages on CT scan images by using various image processing techniques. *IOSR J. Comput. Eng. (IOSR-JCE)* **16**(5), 2278–8727 (2014). ISSN: e-2278-0061
5. Kaviarasu, K., Sakthivel, V.: K-Means clustering using fuzzy C-Means based image segmentation for lung cancer. *South Asian J. Eng. Technol.* **2**(17) (2016). ISSN No: 2454-9614
6. Santhosh, T., Narasimha Prasad, L.V.: Segmentation of lung cancer PET scan images using fuzzy C-means. *Int. J. Comput. Sci. Eng.* **6**(9) (2014). ISSN: 0975-3397
7. Thangaraju, P., Mala, N.: Segmentation of lung tumor using clustering techniques. *Online J. Sci. Res. Technol. (IJSART)* **1** (2015). ISSN: 2395-1052
8. George, R.J., Kumari, D.A.J.: Segmentation and analysis of lung cancer images using optimization technique. *Int. J. Eng. Innov. Technol. (IJEIT)* **3**(10) (2014)
9. Joon, P., Jatani, A., Bajaj, S.B.: Lung cancer detection using image processing techniques: review. *Int. J. Eng. Sci. Comput.* **7**(4) (2017). ISSN: 2321-3361
10. Malik, B., Singh, J.P., Singh, V.B.P., Naresh, P.: Lung cancer detection at initial stage by using image processing and classification techniques. *Int. Res. J. Eng. Technol. (IRJET)* **3** (2016). ISSN: e-2395-0056, p-2395-0072
11. Tan, J.H., Acharya, U.R.: Active spline model: a shape based model—interactive segmentation. *Digit. Signal Process.* **35**, 64–74
12. Lalitha, M., Kiruthiga, M., Loganathan, C.: A survey on image segmentation through clustering algorithm. *Int. J. Sci. Res. (IJSR)* **2** (2013). ISSN: 2319-7064
13. Sharma, P., Suji, J.: A review on image segmentation with its clustering techniques. *Int. J. Signal Process. Image Process. Pattern Recognit.* **9**(5), 209–218 (2016)
14. Pardhi, S., Wanjale, K.H.: Survey on techniques involved in image segmentation. *Int. J. Comput. Sci. Trends Technol. (IJCST)* **4**(3) (2016)
15. Singh, N., Asuntha, A.: Lung cancer detection using medical images through image processing. *J. Chem. Pharm. Sci. (JCPS)* **9**(3) (2016). ISSN: 0974-2115



# A Correlative Study of Contrary Image Segmentation Methods Appending Dental Panoramic X-ray Images to Detect Jawbone Disorders



Krishnappa Veena Divya, Anand Jatti, P. Revan Joshi and S. Deepu Krishna

**Abstract** Dental radiographs have been widely used by dentists to detect any bony pathology which is difficult to diagnose solely by clinical examination. The usage of dental X-rays images has brought about a great improvement in clinical diagnosis due to its immediate availability and relatively lesser radiation dose. Orthopantomograms (OPG) or panoramic imaging is one of the imaging modality frequently used in dentistry to detect any dental anomaly. But the dental panoramic images suffer from varying superimposition of lot of anatomical structures and have an inherent technical issue which leads to lot of ambiguity when applied as an aid to diagnosis. This paper presents the systematic review of image segmentation algorithms applied on dental X-ray images and its results with the supervision of radiologists. A generic comparison of segmentation algorithms has been discussed for the cysts and lesion segmentation in distinction to panoramic image. Thresholding watershed and level sets methods were chosen for segmenting the desired region of cysts and to study the various characteristics of cystic region. Level sets segmentation produces the better results in segmenting the cyst/tumors. The shape descriptors obtained for the region of cysts could conceivably be used as feature vectors in image classification where the classifiers can automatically detect the abnormal tissues or tumors from OPG images helping in its diagnosis and treatment.

**Keywords** OPG · Image preprocessing · Image segmentation · Watershed Thresholding · Level sets

---

K. Veena Divya (✉) · A. Jatti  
Department of Electronics and Instrumentation Engineering, R.V. College  
of Engineering, Bengaluru 560059, India  
e-mail: veenadivya@gmail.com

P. Revan Joshi  
Department of Oral Medicine and Radiology, D.A. Pandu Memorial  
R.V. Dental College and Hospital, Bengaluru, India

S. Deepu Krishna  
Apollo Hospitals, Bengaluru, India

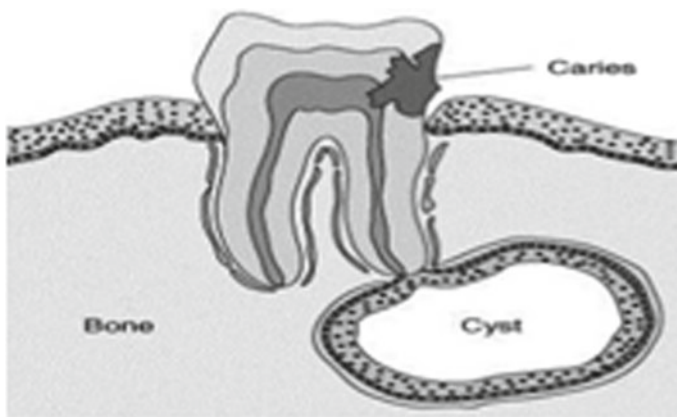
## 1 Introduction

Diagnosing cysts and tumors of jaws is one of the most perplexing challenges in maxillofacial radiology. A cyst is a fluid-filled pathological cavity lined by an epithelium. Cysts are more common in jaws than any other bone since most cysts originates from numerous cell rests of epithelium after the tooth formation. Basically, cysts are classified based on developmental and inflammatory origin. The most frequently occurring cysts are radicular cysts of inflammatory origin and dentigerous cysts of developmental origin. Usually, cysts are round or oval shape with corticated or scalloped boundary. A complete conclusion can be made by age, sexual orientation and predominance of the cyst. In the event that the cysts are not distinguished at a beginning period, it might prompt tumors [1].

Tumors and cysts have some common things. They occupy space and replace or displace the normal tissues. They may push the teeth out of their alignment and also resorb adjacent tissues. An usually painless expansion causing disfigurement of face is commonest presenting symptom. They may also compress the local nerves leading to altered sensation in the region supplied by the affected nerve [2] (Fig. 1).

### 1.1 Causes of Cysts Origin

Exact pathogenesis of a cyst is not clearly known. The origin of the cyst is usually attributed to the growth of the cell rests of Malassez (epithelium responsible for formation of teeth within the jaws). The trigger that induces this growth could be some inflammation or infection found in the dento-alveolar region. The epithelium thus formed behaves as a semipermeable membrane and is composed of connective



**Fig. 1** A growth creating from tooth bud [11]

fibrous tissue. Eventually, this permeability nature; steadily the cyst expands either by secreting fluid from the surrounding tissues or cells lining within its own cavity leading to the cyst formation. Cystic fluid contains proteins which exerts osmotic pressure. The breakdown of cellular debris within the fluid increases the protein concentration which increases the osmotic pressure. The total effect is that pressure developed by osmotic tension within the cavity of cyst and due to presence of the bone resorbing cytokines and chemical mediators, there is an increase in osteoclastic activity leading to bone resorption as well as expansion [3].

## ***1.2 Orthopantomogram***

Orthopantomograph is a mode of dental and maxillofacial imaging and consists of a panoramic image of all the dental anatomical and their supporting structures, both the jaws and the temporo mandibular joint along with a few contiguous structures. It is extremely useful in all dental specialties such as it is useful in diagnosing carious lesions, periodontology, prosthodontics, maxillo-facial surgery, implantology, pediatric dentistry and orthodontics [1]. These radiographs play a crucial role in diagnosing important pathologies and more often lead to diagnosis of asymptomatic pathologies, when used as routine screening method of imaging maxillofacial area [2].

In the image preprocessing stage, picture enhancement is done to upgrade the interpretability or impression of information in pictures for human view and to give “better” commitment for other mechanized image processing techniques. The info image is preprocessed by expanding the dim levels in the picture utilizing linear contrast extending, and some noise expulsion steps are incorporated into this stage.

## **2 Generic Segmentation Algorithms**

Image segmentation is the most essential part in medical image processing as it serves a better input for the automated models in segmentation of any objects from the whole image requiring less operator involvement. Medical image segmentation basically subdivides an image into objects so that interpretation and acquisition of information from an image becomes easier.

Jan Mikulka et al. presented a procedure for automatized evaluation of parameters in orthopantomographic images catching neurotic tissues made in human jaw bones. The consequences of quick automatized segmentation acknowledged through the live-wire strategy and contrast the got information and the outcomes gave by other division methods. In this unique circumstance, an examination of different classifiers is performed, including the decision tree, naive Bayes, neural network, k-NN, SVM and LDA characterization devices. Inside this correlation, the most elevated level

of exactness around 85% on the normal can be credited to the decision tree, naive Bayes and neural network classifiers [4].

P. L. Lin, P. Y. Huang et al. have proposed an effective arrangement to part every tooth in dental periapical radiographs in light of neighborhood peculiarity examination. The neighborhood singularities measured by Holder illustration are figured to secure a structure picture in which the structures of teeth are much smoother than the structures of gums. Otsu's thresholding is associated with divide teeth from gums; finally, related portion examination and morphological operations are associated with detach each tooth. Exploratory results show that out of 18 teeth in six attempted periapical pictures, all teeth are viably segmented with 17 expelled tooth shapes absolutely fitting in with human visual acumen [5].

Most of the segmentation methods depend on one of the essential properties of pixel intensities such as similarity and discontinuity. In the first classification, the methodology is used to divide an image into the regions that comparative as indicated by some predefined criteria and in the second method partitioning an image based on the huge changes in intensities across its edges. Existing non-specific segmentation techniques can be generally grouped into five classifications.

1. Thresholding
2. Region-Based Segmentation
3. Edge-Based Segmentation
4. Graph-Based Segmentation
5. Classification-Based Segmentation

These strategies are portrayed and their advantages and disadvantages are examined as takes after.

### 1. *Thresholding*

For the OPG images, this technique is used to extract part of image containing all the information required for a specific application. It can be generally arranged into two, specifically worldwide thresholding and nearby (versatile) thresholding, in view of the edge choice criteria. For an Image I, the worldwide thresholding strategy tries to discover a limit  $t$  to such an extent that pixels with power esteems more noteworthy than or equivalent to  $t$  will be assigned 1 and the rest of the pixels are assigned to 0 [6]. Likewise, thresholding creates binary images. The computational complexities of these algorithms are very low. Thresholding results in binary images where 1 represents the object of interest and 0 represents the background as shown in Fig. 5b for the input panoramic image.

### 2. *Region-/Area-Based Segmentation*

These calculations fundamentally comprise of split and merge calculation, region developing algorithm and watershed Segmentation.

#### 2.1 *Split and Merge*

In split and consolidation calculation, the entire picture is at first considered as one locale. This area will be part into four quadrants, if certain homogeneity measure is

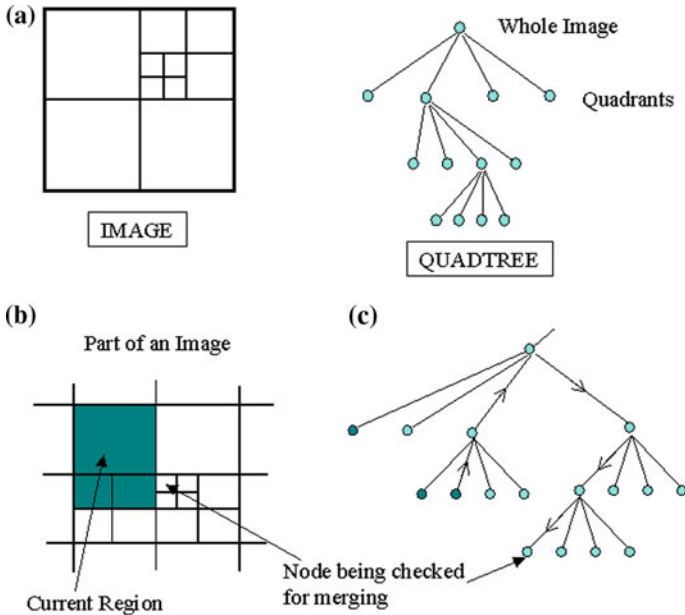


Fig. 2 The split and union procedure. a Division as quadrant. b, c The union of quad [12]

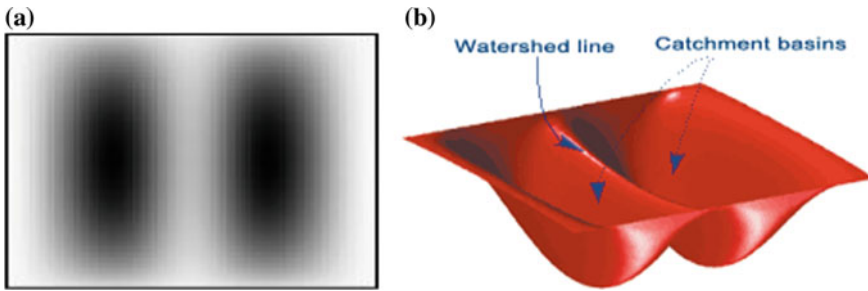
not met. The split procedure will be rehashed recursively until the point that every area contains just homogeneous pixels. The calculation at that point contrasts every one of the areas and their neighboring regions and consolidates the locales that are comparable as indicated by a few criteria. The homogeneity rule is typically in view of the estimations of pixel powers. Districts with standard deviation not as much as an edge are viewed as homogeneous. Split and merge algorithm is computationally fast [6].

### 2.2 Region Growing

Locale growing is inverse of the split and union procedure. At first, area developing begins by stipulating n seed pixels and each seed pixel is managed as a locale. The region developing calculation starts to locate some neighboring pixels to those locales which are like the first area and include these areas thereby developing the region [6] (Fig. 2).

### 2.3 Watershed Segmentation

Watershed segmentation is the most promising region-based approach in the field of mathematical morphology. In a watershed segmentation technique, image is regarded as topographic surface where altitudes are represented by the gray level of the image and flat zones in the image are represented by the constant gray levels. Watershed transformation consists of ridge lines and catchment basins [4]. Catchment basins are corresponded by the low gradient region and high gradient interiors or a water-



**Fig. 3** a Catchment basins. b The watershed ridge lines [7]

shed corresponds to the watershed ridge lines. Catchments basins are homogeneous comprising all pixels having a place with a similar catchment bowl are associated with the locale of least elevation, and these catchment basins represent the region of interest to be extracted [4]. The catchment basins separated by the ridge lines are shown in the Fig. 3.

The watershed algorithm works well if each local minimum corresponds to segmented object. If there are many local minima's in the image than the segmented, then the algorithm suffers from over-segmentation [4].

### 3. *Edge-based segmentation*

The purpose of detecting edges in image is to locate the abrupt changes in the intensities across the object boundary to differentiate it from the surrounding intensities (object/region) required for segmentation in terms of gray level, texture or color. In typical segmentation algorithms, extracting the boundary of an object is essential in identifying number of objects present in an image. Edges are detected using various edge detecting operators such Robert, Prewitt or Sobel [6] as depicted in Fig. 4. Customarily Sobel edge identifier contains a couple of 3 by 3 convolution bits, as shown in Fig. 5. It calculates (gradient) first-order derivatives along x and y directions of the original two-dimensional image. The extent of the subsequent edge is the inclination of the first picture. The first-order derivatives produce thick edges. The Laplacian computes the second-order derivatives of the image instead of first-order derivatives. The Laplacian is regularly consolidated with the Gaussian smoothing part, which is alluded to as the Laplacian of Gaussian capacity, and it is not specifically connected to the first picture since it is delicate to the clamor. Here the size of first request subordinates speaks to the edges shown in a picture, and indication of the second request subsidiary is utilized to discover whether an edge point lies on light or dim side of an edge [8].

#### 1. *Graph-based segmentation*

Graph-based segmentation algorithms are relatively new in the field of image segmentation. The principle thought behind this approach is the development of weighted diagram, where every vertex relates to a picture pixel or area and each edge is

**Fig. 4** Sobel kernel pair, **a** kernel x and **b** kernel y [6]

<b>(a)</b>	<b>(b)</b>																		
<table border="1"> <tr><td>-1</td><td>0</td><td>+1</td></tr> <tr><td>-2</td><td>0</td><td>+2</td></tr> <tr><td>-1</td><td>0</td><td>+1</td></tr> </table>	-1	0	+1	-2	0	+2	-1	0	+1	<table border="1"> <tr><td>+1</td><td>+2</td><td>+1</td></tr> <tr><td>0</td><td>0</td><td>0</td></tr> <tr><td>-1</td><td>-2</td><td>-1</td></tr> </table>	+1	+2	+1	0	0	0	-1	-2	-1
-1	0	+1																	
-2	0	+2																	
-1	0	+1																	
+1	+2	+1																	
0	0	0																	
-1	-2	-1																	

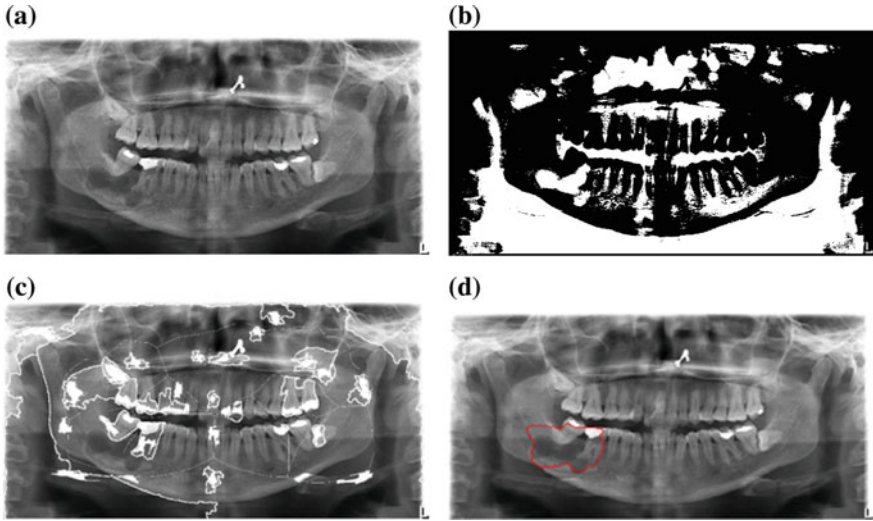
weighted concerning a few criteria. Chart-based division calculations tend to locate the worldwide ideal arrangements when contrasted with the district-based division calculations which depend on eager approach. This diagram cut division experiences over division and a disadvantage of computationally costly.

*2. Classification-based Segmentation*

The essential topic fundamental approach is a calculation used to prepare a classifier to group good segmentation and bad segmentation. Consequently, the criteria used for the classification include brightness similarity, texture similarity, curvilinear continuity; contour energy, etc. A preprocessing is done to reduce the dimensions of an image. The normalized cut algorithm is applied to these preprocessing steps. The image segmented manually by humans' represents positive examples where as negative examples represents the segmentation outputs matched randomly with some fuzzy logic algorithm. The Laplacian of Gaussian approach is employed to obtain the zero-crossing area of the original image. From those zero-crossing areas, fuzzy set is used to describe the direction and transition of intensity values. The fuzzy rules are obtained from the global knowledge presented by medical experts. Fuzzy reasoning methods are used to detect a rough boundary. The neural system is prepared in view of the arrangement of manual sectioned samples. Consequently, relevant guidelines can be educated and spatial consistency can be moved forward. Training is required for this type of segmentation. Subsequently, this algorithm requires learning and training parameter set to produce segmentation. The selected training samples largely determine the accuracy of this algorithm. Also, this algorithm is more tedious to use (Fig. 5).

**3 Comparison of Generic Segmentation Algorithms**

A comparison of generic segmentation algorithms mentioned below is made according to the performance, information used, computational complexity, whether training is required, whether they are sensitive to the noise, whether they are easy to use and whether manual initialization is required. The results are shown in Table 1. Thresholding uses the information based on a single pixel in an image while most of



**Fig. 5** **a** Input image; **b** result of thresholding on original image with a threshold value = 85; **c** application of watershed segmentation on dental panoramic images showing markers and object boundaries superimposed on original image; **d** result of level sets segmentation on original image

**Table 1** Comparison of generic segmentation algorithms

Features	Thresholding	Region-based	Edge-based	Graph-based	Classification-based
Information	Pixel	Patch	Patch	Patch	Patch
Complexity	$O(n)$	$O(n)$	$O(n)$	$O(n \log n)$	$O(n)$
Manual init	No	Yes	No	No	No
Training	No	No	No	No	Yes
Easy to use	Yes	Yes	Yes	Yes	No

the other segmentation depicts from information obtained from a region. The performance of the thresholding algorithm depends on the distribution of pixel intensities of an image. Edge-based segmentation tends to produce concavity of edges. Most of the algorithms discussed exhibit over-segmentation. The computational complexities of the graph-based and characterization-based division strategies are higher contrasted with the edge-based, thresholding and region-based approach tending to be linear. Region-based algorithm usually requires manual initialization. All the algorithms are sensitive to noise. Most of the algorithms are easy to use except classification-based algorithm which requires training. Table 1 compares and gives an understanding of segmentation methods applied on the Orthopantamogram images.



## 4 Feature Selection and Extraction

Feature Extraction constitutes an essential piece of machine vision frameworks. Customarily, the element extraction strategies are utilized after the image has gone division. It is designed to represent the objects or the region of interest in a informative manner like defining the features and attributes for the objects extracted after segmentation. The features obtained can be used in further processing as it convert the pictorial to non-pictorial data representation which can be used in pattern recognition to classify the objects into benign or malignant. The techniques for Classification require the estimation of image parameters such mark of the question limit, edge power variety and shape circularity and so on. Likewise, the highlights of shape descriptors like territory, roundness, solidness could aid cysts and tumor characterization [9]. After the information about the object of interest is known, some pattern recognition techniques are employed in machine vision. It represents the last stage in image processing approach. Many classifiers are available for region and object classification. Some of the pattern recognition techniques neural nets, k-NN, support vector machines (SVM) can be used for classifying tumor or non-tumor images.

## 5 Results and Discussion

Dental X-ray imaging is a standout among the most essential applications in the field of therapeutic image processing. In this paper, several image processing algorithms have been applied on the dental X-rays, and the results are compared with respect to peak signal-to-noise ratio (PSNR) and mean squared error (MSE) in order to interpret which enhancement algorithm produces better results which could help in the diagnosis and its early detection.

Measuring the picture quality is frequently a troublesome assignment on account of the huge number of factors required in arriving at the final result like ranging from the precision of display technology, observability, encompassing lighting to the individual's state of mind and many more. surveying of visual nature of an image is a subjective procedure that are generally assessed by target measures [10]. Objective quality measures provide accurate and repeatable results which are easy analysis since it relies on less controllable factors because of its mathematical convenience. Disregarding this component, target quality measures don't coordinate with the subjective experience of human watcher seeing the image.

*Peak signal-to-noise ratio:*

The term peak signal-to-noise ratio (PSNR) is an expression for the proportion between the most extreme conceivable esteem (power) of a signal and the energy of mutilating noise that influences the nature of its portrayal. It is usually represented in logarithmic scale decibel. It can be used as comparative measures for image enhancement methods to decide which enhancement algorithm produces good results. Suppose, if an algorithm enhances the degraded image which is closing matching with

**Table 2** MSE and PSNR for watershed and contrast manipulation algorithms using OPG Images

Algorithm	PSNR	MSE
Contrast manipulation	14.9440	2.0830e+03
Watershed transformation	17.1599	1.2505e+03

the original image then it is assumed that algorithm produces better results for a specific application [10].

Mathematically, PSNR is represented as

$$PSNR = 20 \log_{10} \left( \frac{MAX_f}{\sqrt{MSE}} \right)$$

B. Mean squared error (MSE) equation is given by

$$MSE = \frac{1}{xy} \sum_0^{x-1} \sum_0^{y-1} \|f(i, j) - g(i, j)\|^2$$

- f represents die matrix data cf our original image
- g represent the matrix data of our degraded iniaee in question
- x represent the numbers of rows of pixels of the images and i
- y represent the number of columns of pixels of the image and
- j represent the index of that column
- MAX<sub>f</sub> is the maximum signal value that exists in our original “known to be good” image

Image with high PSNR and value signifies better quality for the restored or processed image. Lower value of MSE results in higher PSNR. These image quality measures are employed for dental panoramic images for different image processing algorithms as shown in Table 2. From the table, it is observed that histogram modification yields good results compared to other algorithms with a PSNR value of 19.6742 and MSE of 700.9043.

## 6 Conclusion

A precise survey improved the situation number of images upgrade and division of cysts and tumors from OPG pictures. Edge- and region-based segmentation techniques have been analyzed based on the various features such as complexities, information content extracted, manual initialization and training required for the segmentation algorithm to extract the required region of interest. With the application of watershed segmentation, the optimal objects to be segmented from the whole OPG images have been discussed. Furthermore, the level sets algorithm has

been implemented for the segmentation of cysts/tumors. Usually the shapes of cysts and tumors are irregular; these levels sets algorithm precisely segmenting the cystic region with its proper contour. The experimental results suggest the efficacy of the proposed system when compared to the manual segmentation. The image processing approaches discussed help the dentists to classify cysts images from normal images and the enhanced images could help in highlighting the fine details like edges of cysts or lesion of Jaws. This could potentially assist in diagnosis for screening the early detection of dental anomalies improving the diagnostic accuracy and treatment outcomes.

## References

1. Veena Divya, K., Jatti, A., Joshi, R., Meharaj, S.: Image processing and parameter extraction of digital panoramic dental X-rays with ImageJ. In: International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), pp. 450–454. IEEE (2016)
2. Veena Divya, K., Jatti, A., Revan Joshi, P.: Appending active contour model on digital panoramic dental X-rays images for segmentation of maxillofacial region. In: 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES), Kuala Lumpur, Malaysia, 4–8 December 2016, 978-1-4673-791-1/16/\$31.00. IEEE (2016)
3. Veena Divya, K., Jatti, A., Revan Joshi, P.: Computer aided classification using support vector machines in detecting cysts of jaws. *Adv. Sci. Technol. Eng. Syst. J.* **2**(3), 674–677 (2017)
4. Mikulka, J., et al.: Classification of jawbone cysts viaorthopantomogram processing. In: 2012 35th International Conference on Telecommunications and Signal Processing (TSP). IEEE (2012)
5. Lin, P.L., Huang, P.Y., Huang, P.W.: An effective teeth segmentation method for dental periapical radiographs based on local singularity. In: 2013 International Conference on System Science and Engineering (ICSSE). IEEE (2013)
6. Decusara, M.: Use of orthopantomogram in dental practice. *Int. J. Med. Dentistry* **1** (2011)
7. Li, C., Xu, C., Gui, C., Fox, M.D.: Distance regularized level set evolution and its application to image segmentation. *IEEE Trans. Image Process.* **19**(12), 3243–3254 (2010)
8. Marques, O.: *Practical Image and Video Processing Using MATLAB*. Wiley, Hoboken, New Jersey (2011)
9. Liao, Q., Hong, J., Jiang, M.: A comparison of edge detection algorithm using for driver fatigue detection system. In: 2010 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), vol. 1, pp. 80–83. IEEE (2010)
10. Veena Divya, K., Jatti, A., Revan Joshi, P.: Characterization of dental pathologies using digital panoramic X-ray images based on texture analysis. In: Presented a Paper at 39th Annual International Conference of the IEEE Engineering in Medicine and Biological Society, 11–15 July 2017, JEJU Island, South Korea (2017)
11. <https://whitedentalcare.wordpress.com/tag/dental-cysts/>
12. <http://www.doc.ic.ac.uk/~dfg/vision/v02.html>

# Image Quilting for Texture Synthesis of Grayscale Images Using Gray-Level Co-occurrence Matrix and Restricted Cross-Correlation



Mudassir Rafi and Susanta Mukhopadhyay

**Abstract** Exemplar-based texture synthesis is a process of generating perceptually equivalent textures with the exemplar. The present work proposes a novel patch-based synthesis algorithm for synthesizing new textures that employs the powerful concept of gray-level co-occurrence matrix coupled with restricted cross-correlation. Furthermore, a simple and peculiar blending mechanism has been devised which avoids the necessity of retracing the path after ascertaining the minimum cut within the overlap region between the two neighboring patches. The method has been tested and executed for the samples derived from Brodatz album, the widely acceptable benchmark dataset for texture processing. The results are found to be comparable to Efros and Freeman for stochastic texture while outperforms the Efros and Freeman algorithm for semistructured texture.

**Keywords** Texture synthesis · Image quilting · Patch-based texture synthesis  
GLCM

## 1 Introduction

Texture is a collective effect produced as a result of aggregation of pixels present in the image forming definitive patterns of shape, scale, orientation, color, and spatial frequency. The surface characteristics of several entities like terrain, plants, minerals, fur, and skin are attributed to their textural surface. In computer graphics, the major

---

M. Rafi (✉) · S. Mukhopadhyay  
Department of Computer Science and Engineering, Indian Institute  
of Technology (ISM), Dhanbad, Dhanbad, India  
e-mail: mudassir.rafi23@gmail.com

S. Mukhopadhyay  
e-mail: msushanta2001@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent  
Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_4](https://doi.org/10.1007/978-981-13-1708-8_4)

concern is to create visual realism that depends upon the accurate localization and synthesis of natural texture. Texture synthesis is defined as a process of producing new textures that are perceptually equivalent to the input texture. Texture synthesis is important for fast scene generation, image inpainting, and texture restoration. Texture synthesis never guarantees the generation of textures that are replica of the exemplar; however, it guarantees the perceptual similarity of the synthesized image with the exemplar. In the literature, various classifications of the texture synthesis techniques have been suggested. The most recent one is suggested by Raad et al. [1]. They have classified all the techniques into two classes, namely statistics-based [2–4] techniques and nonparametric patch-based [5–11] techniques. Statistics-based methods are motivated by the pioneer work of Julesz [12], who stated that texture pairs having same second-order statistics are not pre-attentively discernible by humans. In order to perform synthesis, these methods need two sequential steps to be carried out, namely analysis and synthesis. Efros and Leung [7] introduced the pixel-based approach based on Shannon’s Markov random field model devised to simulate text. The value of pixel to be synthesized is determined by searching over square patches in the sample texture. The similarity criterion used in this case is L-2 norm. The method was found to be slow. Wei and Levoy [10] employed raster scan ordering to transform noisy pixels into the resultant texture. The performance of algorithm was improved using multi-scale framework and tree-structured vector quantization. The L-2 norm was also minimized in the RGB space without normalization. In 2001, Ashkhmin [5] improves Wei and Levoy technique and obtained satisfactory results for natural texture at which Weil and Levoy technique failed. Tonietto and Walter [13] synthesized the texture from a collection of sample at different resolution. Zhang et al. [14] introduced an image-based texture synthesis for rendering of progressively variant textures. Efros and Freeman [8] introduced patch-based approach and named it as image quilting, by stitching together random blocks of the sample and modified them in a consistent way to synthesize texture. Patch-based approaches are better than pixel based as it builds the texture at coarser scale while keeping the high frequencies of the sample intact. In the present work, a patch-based texture synthesis algorithm is proposed. The paper is organized as follows. Section 2 describes the mathematical concept of gray-level co-occurrence matrix. In Sect. 3, the motivation behind this technique is given. Section 4 throws some light on cross-correlation and describes the concept of restricted cross-correlation. Section 5 presents the central idea of the proposed technique the GLCM-based image quilting along with the proposed blending mechanism in a subsection. Section 6 presents the experimental prerequisites, procedure, and a comprehensive discussion on the results. The final section concludes with a summary of the entire process performed.

## 2 Gray-Level Co-occurrence Matrix (GLCM)

In texture description, the spatial arrangement of pixels is also important along with the intensity values. Any attempt not considering this facet may preserve the brightness information but would not be able to completely describe the textural

features. A gray-level co-occurrence matrix [15] is such an approach that considers both the intensity values in conjunction with the spatial arrangement of pixels. Let us suppose that  $G$  be a matrix whose elements  $f(i, j)$  are representing the frequencies of pixel pairs with intensity values  $z_i$  and  $z_j$  occurring in the image  $g(r,c)$  at the location specified by  $Q$ , where  $Q$  is an operator defining the positions of the two pixels relative to each other and  $1 \leq i, j \leq L$ . The matrix dimension depends upon the number of gray levels present in the image. Conventionally, the numbers of gray levels are quantized in order to reduce the size of the matrix. After computing all the values of the matrix, a normalization step is applied using the equation as given below:

$$N_{(p,q)} = \frac{f(p, q)}{\sum_{i=0}^L \sum_{j=0}^L f(i, j)}, p, q = 0, \dots, L \quad (1)$$

where  $N_{(m,n)}$  is the normalized matrix,  $f(i, j)$  is the number of pixel pairs having values  $z_i$  and  $z_j$ ,  $f(p, q)$  is the entry in the co-occurrence matrix, and  $L$  is the maximum number of gray levels. Haralick and Shanmugam [15] proposed a set of 14 statistical measures based on the co-occurrence matrix. These include angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, measures of correlation, and maximal correlation coefficient. Out of these 14 measures, contrast, correlation, energy, and homogeneity are the fundamental ones.

**Contrast** is a quantitative measure of the intensity contrast between a pixel and its neighbor over the whole size of the image. The values range from 0 to  $(L - 1)^2$ .

$$Contrast = \sum_{i=1}^L \sum_{j=1}^L (i - j)^2 N_{(i,j)} \quad (2)$$

**Correlation** is a measure of how a pixel is correlated to its neighbor over the entire size of the image. Values are ranges from  $-1$  to  $1$ . If the correlation is perfect  $1$  is assigned whereas for a perfect negative correlation  $-1$  is used.

$$Correlation = \sum_{i=1}^L \sum_{j=1}^L \frac{(i - m_r)(j - m_c)}{\sigma_r \sigma_c} \quad (3)$$

$$\text{where } \sigma_r \neq 0, \sigma_c \neq 0 \text{ and} \quad (4)$$

$$m_r = \sum_{i=1}^L i \sum_{j=1}^L N_{(i,j)} \quad (5)$$

$$m_c = \sum_{j=1}^L j \sum_{i=1}^L N_{(i,j)} \quad (6)$$

**Energy** is a measure of regularity and lies in the range  $[0, 1]$ . It is 1 for a constant image.

$$Energy = \sum_{i=1}^L \sum_{j=1}^L N_{i,j}^2 \quad (7)$$

**Homogeneity** measures that how much the distribution of elements in the matrix  $G$  is closer to its diagonal. The range of values lies in  $[0, 1]$ , with the highest being reached when  $G$  is a diagonal matrix

$$Homogeneity = \sum_{i=1}^L \sum_{j=1}^L \frac{N_{i,j}}{1 + |i - j|} \quad (8)$$

### 3 Restricted Cross-Correlation

In signal processing, the cross-correlation quantitatively assesses the similarity between two functions (signals) at all possible time shifts, or time lags. In the field of statistics, it is described in terms of the expected values, whereas, for deterministic signals, it is defined in terms of sums or integrals. Cross-correlation is commonly used for searching a long signal for a shorter known feature. Mathematically,

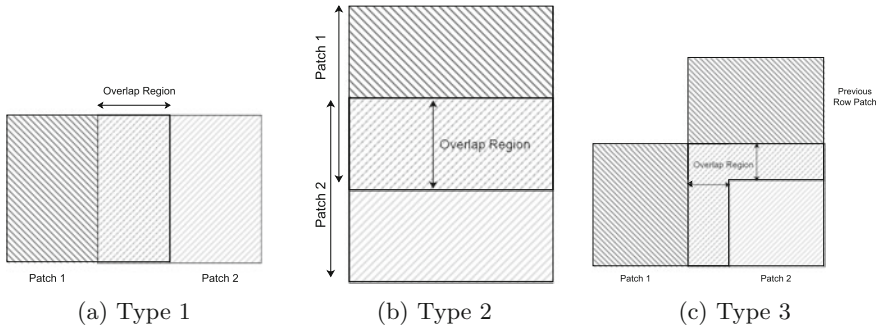
$$(f * g)(\tau) = \int_{-\infty}^{+\infty} f^*(t)g(t + \tau)dt \quad (9)$$

where  $f^*$  represents the complex conjugate of  $f$  and  $\tau$  represents the displacement /lag. In discrete form, it can be represented as,

$$(f * g)(\tau) = \sum_{-\infty}^{+\infty} f^*[m]g(m + n) \quad (10)$$

Cross-correlation and convolution are similar quantities. In the present work, authors have restrict the cross-correlation to allow the displacement of one function relative to the other in one direction only i.e either horizontally or vertically, corresponding to the boundary share with the neighboring patch. As shown in Fig. 1, the boundary share can be of three types, namely Type 1, Type 2, and Type 3. These can be defined as:

- Type 1: When the current patch shares the left boundary with the adjacent patch.
- Type 2: When the current patch shares the upper boundary with the patch present in the upper row.



**Fig. 1** Different types of boundary shares

- Type 3: When the current patch shares left boundary with the adjacent patch and upper boundary with the patch present in the upper row.

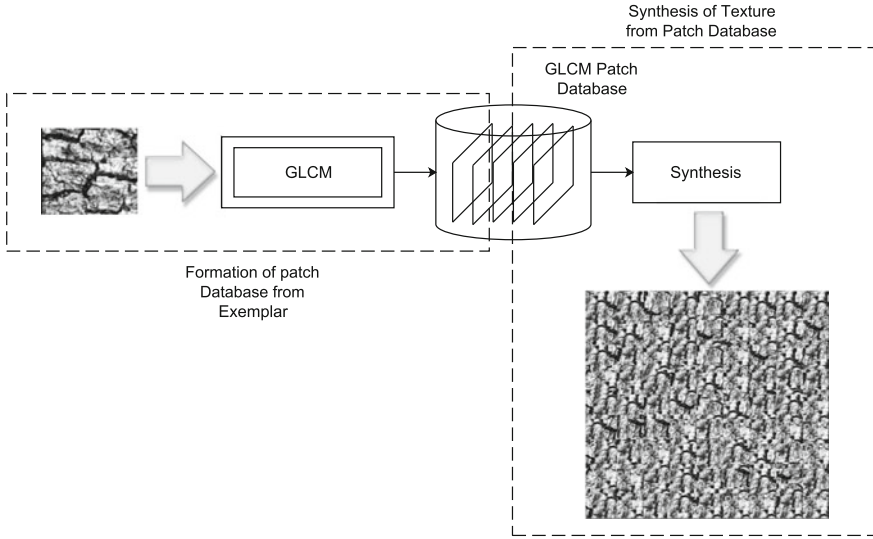
## 4 GLCM-Based Image Quilting

Gray-level co-occurrence matrix (GLCM) is a widely recognized, powerful tool for extracting information from the texture images. The co-occurrence matrices has an amazing property to keep not only the intensity information but also the spatial occurrence at a particular distance and direction. Thus, the statistical properties derived from these matrices could be used as an effective descriptor for searching a patch with in a database. The present method employs the input exemplar  $u$  and synthesize the output image  $w$  sequentially, in a raster scan order (from left to right and top to bottom). The method uses two sequential steps. The first step creates a patch database by employing the properties derived from gray-level co-occurrence matrix (GLCM) applied on all the overlapping patches taken from the exemplar whereas the second step synthesizes the required texture. The patch database comprises the square patches of size  $m$  in order to avoid the computational complexity. Each patch has been stored in the database along with its GLCM properties so that on the basis of these numerical values the corresponding patch can be retrieved when needed in the subsequent step. The authors have used here sequential search, however a fast searching algorithm could be use either, that would certainly reduce the overall running time of the algorithm. The second step involves the process of synthesis, at the topmost left corner of the image to be synthesize with its statistical (GLCM) value at hand. This value is used to search the database for the most similar patch adjacent to this using following equations (Fig. 2).

$$d(P_{prev}, P_{cur}) = (G_{prev}^{stat} - G_{cur}^{stat})^2 \quad (11)$$

$$d(P_{prev}, P_{cur}) \leq \delta \quad (12)$$





**Fig. 2** Proposed method

where  $d(P_{prev}, P_{cur})$  represents the distance between the statistical properties of the two adjacent patches.  $G_{prev}^{stat}$  is the GLCM property of previous patch  $G_{cur}^{stat}$  is the GLCM property of current patch. Equation 12 imposed the similarity criterion on the distance  $d$ . The patches satisfying this equation are determined, and out of these patches a patch at random has been chosen. The randomization is involved to avoid the verbatim copies of same patch. In the previous section, three types of border sharing have been defined. For border share of type 1, the lag between the two adjacent patches is computed employing cross-correlation and allowing it to move only in vertical direction along the left boundary. In this way, the computed lag may either be positive, negative, or zero. If the lag is zero, it signifies that the two patches are similar therefore no adjustment required for the current patch. The positive lag value tells that the current patch should be moved downward in order to synchronize it with the previous patch. When the current patch is shifted downwards corresponding to the computed lag, there arose a blank at the opposite side of the current patch, at the same time, extra portion equal to  $(m - lag)$  came out when compared with the edge of the previous patch. This extra portion is clipped and placed at the blank space. Additionally, when the computed lag is negative, the current patch is shifted upward, the extra portion is clipped and placed at the opposite side of the current patch as mentioned above. For boundary share of type 2, the lag is computed between the upper patch and the current patch. Furthermore, similar to type 1, it may either be positive, negative, or zero. Again, zero lag signifies the similarity between patches and nothing has to done, whereas for positive lag, current patch is shifted right side, extra portion is clipped and placed over the arose blank space. Similarly, for negative lag the shifting direction changes and the clipped extra portion is placed analogous to

the positive lag. Moreover, for boundary share of type 3, first the procedure described for type 1 is applied, followed by the procedure for type 2 is used, not to mention that it has overlap with both the left patch and the patch present above.

#### 4.1 Minimum Difference Transition Blending

The major goal of blending is to make the transition between the two overlapping blocks as smooth as possible. It can be done by allowing the transition at a point when the variation between the two overlapping surfaces is as minimum as possible. Within a textural region, the variation among the pixel intensity values is large, and finding a global minima that satisfies the minimum differencing requirement of the overlapping region is not possible. In lieu of this problem, authors have proposed a new and simple blending mechanism that neither requires retracing of the minimum cut path nor it needs to remember the selection at the previous step to make a cumulative decision for minimum path. The proposed method first defines the surface overlap error by

$$E_{ov} = (Ov_1 - Ov_2)^2 \quad (13)$$

where  $Ov_1$  is the overlapping share from patch 1 and  $Ov_2$  represents the overlapping share from patch 2. As shown in the figure, the method suggests the selection of pixel values at the overlap region from the two overlapping blocks. The method works row-wise/column-wise. Within a row/column (for vertical/horizontal trace), the values are chosen from patch 1 upto the minimum difference whereas from minimum difference onwards the values are selected from patch 2. In the Fig. 4, the minimum difference is shown by the black patch. As for each row/column, the minimum difference ( $E_{ov}$ ) would be at different place within a row/column. Thus, the transition border would be of zigzag in shape and provide a smooth transition between the two patches (Fig. 3).

## 5 Experimental Results and Discussion

The experimental works have been performed using MATLAB 2013 running on Windows operating system with intel core i7 processor and 4 GB RAM. Brodatz dataset, the widely acceptable and a benchmark dataset for texture processing, has been employed for experimentation. It can easily be found on outex site as *contrib TC 00004*. The images in the dataset are of higher dimension, i.e.,  $512 \times 512$  pixels whereas the proposed algorithm requires a lower dimensional exemplar in comparison to the dimension of the image to be synthesized. As a matter of fact, any image drawn from this database has been divided into 16 nonoverlapping block, each with dimension  $128 \times 128$  pixels. Additionally, out of these blocks one has been selected to be taken as exemplar.

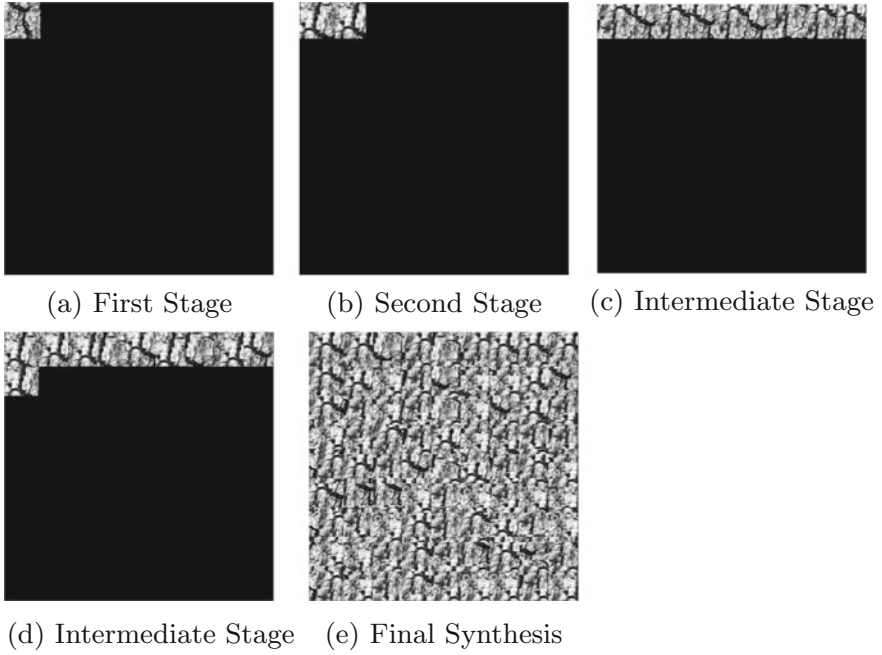


Fig. 3 Various stages of GLCM-based image quilting

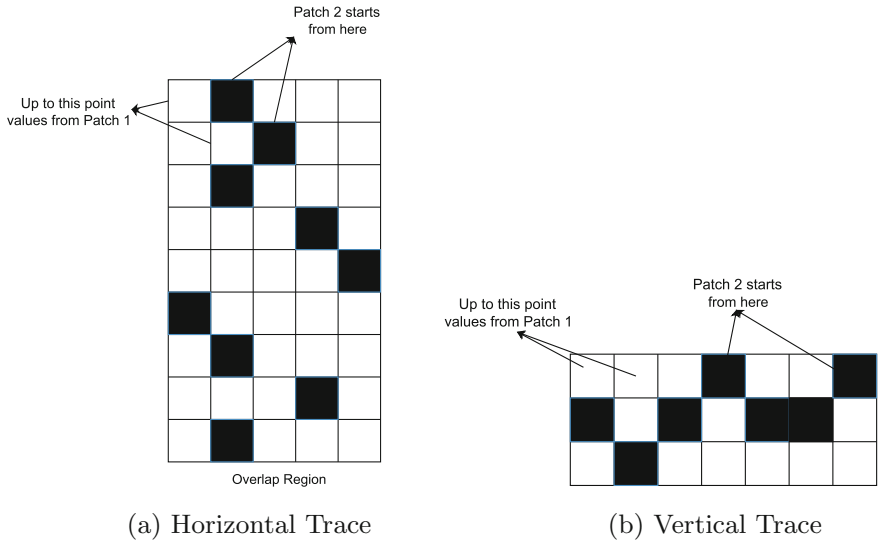
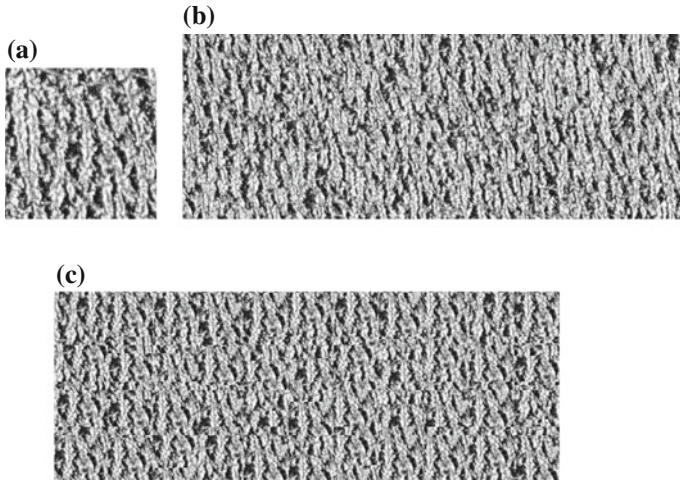


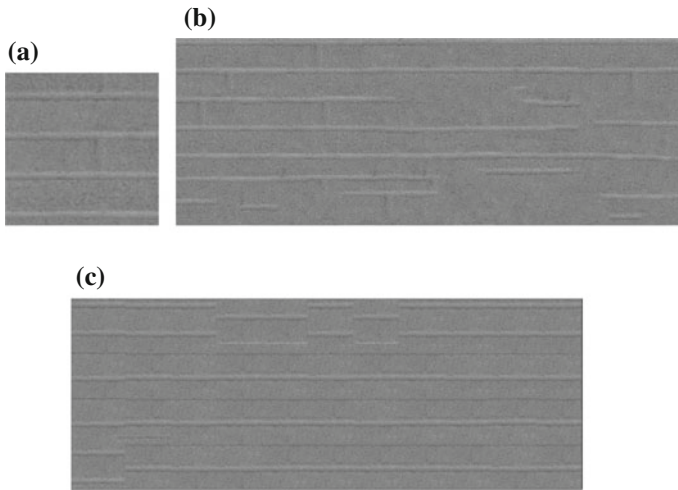
Fig. 4 Minimum difference transition blending



**Fig. 5** a Sample, b Efos and Freeman, c GLCM-based image quilting

The first step in the proposed method requires the formation of GLCM patch database from the selected exemplar. The exemplar is divided into overlapping patches each of dimension  $m$ . Here,  $m$  has been taken as 64 in order to consider a substantial number of textural primitives. For each of these patch, a GLCM matrix has been framed and corresponding GLCM statistical properties are computed. The optimized MATLAB functions *graycomatrix* and *graycoprops* have been used for this purpose. Moreover, the patches along with their GLCM properties as key have been stored, so as to make a patch database. Furthermore, authors have used here only homogeneity values in order to avoid the computational complexity. However, other Haralick properties either alone or in combination can equally be used. The *randi* function of MATLAB has been used to select the random patch from the patch database and placed it at the top most left corner of the image to be synthesized. The other user-defined parameter involves the size of the overlapping region and the size of the image to be synthesized. The size of the overlapping region has been taken as 1/6 of the size of the exemplar as suggested by Efos and Freeman. Moreover, the dimensions of the image to be synthesized have been taken as  $256 \times 512$  due to the visual constraint in the MATLAB image viewer.

The results of the synthesis process show that for stochastic texture, the method gives satisfactory results which are comparable to the results given by Efos and Freeman (Fig. 5) whereas the proposed method outperforms the Efos and Freeman for structured texture (Fig. 6). The whole process including the database formation, retrieval, and synthesis is taking approximately 34 s for unoptimized code. The execution time can further be reduced by making the database formation as a separate process and using some fast searching methods for patch search. The authors would like to improve this method further and want to apply it on the color images as well.



**Fig. 6** a Sample, b Efros and Freeman, c GLCM-based image quilting

## 6 Conclusion

The paper has introduced a new patch-based image synthesis algorithm employing gray-level co-occurrence matrix, a widely acceptable tool for texture description and restricted cross-correlation. In addition to this, the authors have also introduced a simple and effective blending technique, minimum difference transition blending, that neither requires to remember the selection at the previous step nor it requires to retrace the minimum cut path among two neighboring patches at the overlap regions. The experimental results prove the efficacy of the present method.

## References

1. Raad, L., Desolneux, A., Morel, J.-M.: A conditional multiscale locally Gaussian texture synthesis algorithm. *J. Math. Imaging Vis.* **56**(2), 260–279 (2016)
2. Galerne, B., Gousseau, Y., Morel, J.-M.: Random phase textures: theory and synthesis. *IEEE Trans. Image Process.* **20**(1), 257–267 (2011)
3. Heeger, D.J., Bergen, J.R.: Pyramid-based texture analysis/synthesis. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp. 229–238. ACM (1995)
4. Portilla, J., Simoncelli, E.P.: A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**(1), 49–70 (2000)
5. Ashikhmin, M.: Synthesizing natural textures. In: *Proceedings of the 2001 Symposium on Interactive 3D Graphics*, pp. 217–226. ACM (2001)
6. Kwatra, V., Essa, I., Bobick, A., Kwatra, N.: Texture optimization for example-based synthesis. *ACM Trans. Gr. (ToG)* **24**(3), 795–802 (2005)

7. Efros, A.A., Leung, T.K.: Texture synthesis by non-parametric sampling. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, vol. 2, pp. 1033–1038. IEEE (1999)
8. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 341–346. ACM (2001)
9. Liang, L., Liu, C., Xu, Y.-Q., Guo, B., Shum, H.-Y.: Real-time texture synthesis by patch-based sampling. *ACM Trans. Gr. (ToG)* **20**(3), 127–150 (2001)
10. Wei, L.-Y., Levoy, M.: Fast texture synthesis using tree-structured vector quantization. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 479–488. ACM Press/Addison-Wesley Publishing Co. (2000)
11. Kwatra, V., Schödl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures: image and video synthesis using graph cuts. *ACM Trans. Gr. (ToG) (ACM)* **22**, 277–286 (2003)
12. Julesz, B.: Visual pattern discrimination. *IRE Trans. Inf. Theory* **8**(2), 84–92 (1962)
13. Tonietto, L., Walter, M.: Towards local control for image-based texture synthesis. In: XV Brazilian Symposium on Computer Graphics and Image Processing, 2002 Proceedings, pp. 252–258. IEEE (2002)
14. Zhang, J., Zhou, K., Velho, L., Guo, B., Shum, H.-Y.: Synthesis of progressively-variant textures on arbitrary surfaces. *ACM Trans. Gr. (TOG) (ACM)* **22**, 295–302 (2003)
15. Haralick, R.M., Shanmugam, K.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**(6), 610–621 (1973)

# Tongue Recognition and Detection



Ravi Saharan and Divya Meena

**Abstract** In today's era, most of the data are converted into digital form and stored on cloud, and security of data is the main concern as more effort is put forth by researchers to increase security. High security will be provided by high authenticity of a person, only endorsed person. To achieve high authenticity, authentication of an individual can be performed using biometrics which uses the unique organ of a person like iris, fingerprint, DNA, speech recognition, tongue. Biometric of an individual itself justifies the presence of authenticated person. In authentication of a person, tongue can also be used because it is a unique organ of a person, which provides unique identity to a person. In this type of authentication, system accuracy is the main concern. In this paper, we will discuss about the different techniques, which are implemented for the authentication of somebody using tongue images.

**Keywords** SIFT · Tongue recognition · ROI of image · Gabor filter

## 1 Introduction

### 1.1 Image Processing

Image processing is tool or an algorithm to process an image in order to compress image, enhance image, or extract some useful information from the image. It is a type of signal dispensation in which input is image, like video frame or photograph, and output can be image or characteristics of that image. We can perform image segmentation, image enhancement, noise reduction, geometric transformations, and image registration on an image.

---

R. Saharan (✉) · D. Meena  
Central University of Rajasthan, Kishangarh, Ajmer 305817, Rajasthan, India  
e-mail: ravisaharan@curaj.ac.in

D. Meena  
e-mail: divyacse104@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_5](https://doi.org/10.1007/978-981-13-1708-8_5)

## 1.2 *Biometric*

In the current digital world, biometric plays an important role to authenticate a person identity. Biometric is becoming more and more common to every field to authenticate a person identity. Biometrics is the measurement and statistical analysis of people's physical and behavioral characteristics. The technology is mainly used for the identification and access control, or for identifying individuals and giving grant for access that is under surveillance. These body part of a person can be used as a Biometric like- Iris, fingerprint, facial geometry, voice, ear geometry, hand geometry, DNA etc. DNA have very less change throughout the age so this have high accuracy in authentication of a person [1].

There are mainly two type of biometrics

1. Physiological characteristics: In such type of biometric, physical shape of the object is considered.
2. Behavioral characteristics: In the behavioral type of biometric, behavior of the object is noticed such as typing rhythm, gait, gestures, and voice.

## 1.3 *Tongue*

In authentication of a person, tongue can also be used because it is a unique organ of a person, which provides unique identity to a person. Tongue image analysis is new in biometric and research in this field. Tongue has many different properties, which provide different factor to makes it unique for each person. Tongue also has its behavioral character and physiological character. Two person's tongue has different shape, surface textures, and color, which makes it unique for authentication. Movement of tongue can be used for its behavioral characteristics. For the investigation of a person, tongue can be used easily because it is easily exposed and it does not change its properties by reacting with the environment. It stays safe in mouth. It is not easy to forge and not conceivable to cheat another one. If any wound happens to the tongue, it gets rid soon; change in tongue structure is not possible, so it will be useful for biometric to make sure the identity of a person. To capture tongue image and its properties for analysis, a person's tongue should be stable and in a fix position, so that we can make comparison between two image using same parameter.

1. Different shape of tongue:  
See Figs. 1 and 2.
2. Different texture of tongue:  
See Fig. 3.

### **Applications of tongue biometric [2]:**

Day by day, all the money transaction and payment are made by online transaction. There is so many other biometrics available as I have discussed above, but now new



in biometric is the tongue recognition system, for authentication is the better way to provide security at high level.

These are some applications: Identification of criminals, account access, ATMs, online banking, access to personal information, patient identification, employee access, air travel are fields where biometrics can be very useful [3].

## 2 Literature Survey

### 2.1 Extraction of Spot on Tongue Print [4]

In this technique histogram, equalization technique is used for enhancement of image to get more information about image and can get better result by image processing. This method can be used on entire image or on part of image. To improve visual appearance, histogram equalization technique is used. It is based on pixel distribution of an image pixel; every image has three separate color values of the pixel: RGB [5]. To achieve more accurate matching, histogram technique is applied. After storing information about tongue, matching is performed with database and the matching score is calculated.

### 2.2 Shape Feature Extraction Algorithm (Control Points) for Tongue Images [4]

Shape of tongue can be measured by the control points which are represented by shape vector. Control points are used to bound region of interest or area of interest. Here  $p_1, p_2, p_3 \dots p_{11}$  points are creating the boundary of the tongue. Length, bend, thickness, width, curvature of to tongue tip are formed boundary. Here  $p_{tip}$  is tip of the tongue and  $p_m$  denotes the corner of mouth (Fig. 4).

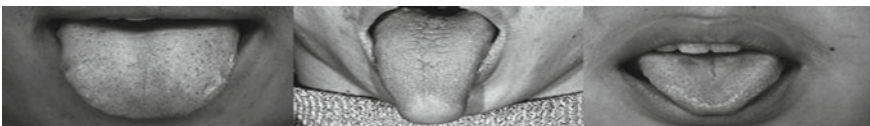


Fig. 1 Front view of tongue



Fig. 2 Profile view of tongue

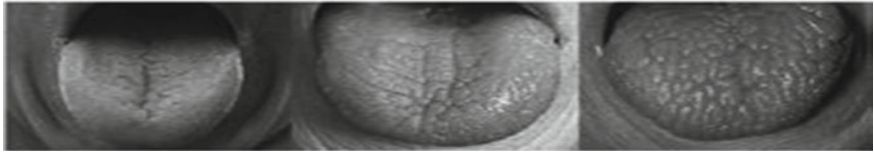
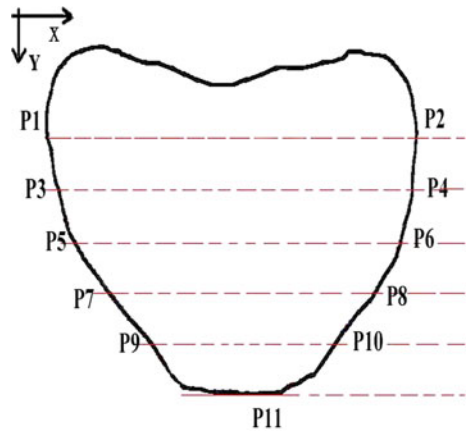


Fig. 3 Different texture of tongue

Fig. 4 Shape feature model for the frontal and profile view image



### 2.3 Map Relationship Between Feature and Tongue Type [4]

Relationship between different properties of tongue is created by calculating different factor of tongue, and this relationship gives uniqueness to a person and match with new input image, which comes for the authentication, and according to decision, the person is authenticated. Likewise, the color and shape of tongue are identified and these details are stored for a person; some quantitative tongue properties feature are calculated to check whether tongue is thick or thin and second is the coverage area of the tongue. Color of tongue can be calculated using HSL model. HSL stands for hue, saturation, and lightness. This model can identify the tongue color using two groups: tongue substance set  $P_s$  and tongue coating set  $P_c$ . Tongue substance color is of five types whitish, light red, dark red, regular red, and purple. Tongue coating color may be of three types: black, yellow, and white. By calculating average of HSL, value of substance, and coating, the characteristic of substance and coating color are

computed. In this, fissures and petechiae are detected. According to existence of fissures and petechiae, tongue is classified into three types: non-fissured petechiae tongue, fissured tongue, and petechiae tongue.

#### ***2.4 SIFT Feature Extraction Technique [6]***

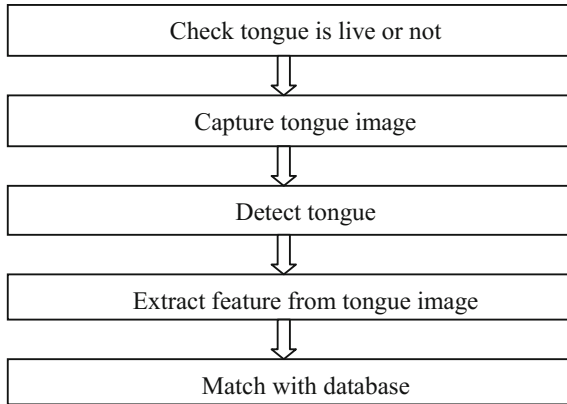
SIFT is scale-invariant feature transform. SIFT analysis detects silent features in an image and extracts descriptor that is different in viewpoint. SIFT standard interest point detector and standard SIFT histogram of gradient descriptor are used to detect local image features. They provide a set of features for an object that is not changed by several difficulties experienced in other methods such as object mounting and rotation. SIFT features are also very strong with respect to noise in the image. The SIFT approach for image feature generation takes an image and mutates it into a large group of local feature vectors. Each of these feature vectors is invariant to any scaling, rotation, or translation of the image. After changing scale, rotation, illumination, and viewpoint, we can obtain good result. For the extraction of these features, the SIFT algorithm applies a four-stage filtering approach. SIFT technique is used for surface texture information identification.

#### ***2.5 Gabor Filter for Feature Extraction [7]***

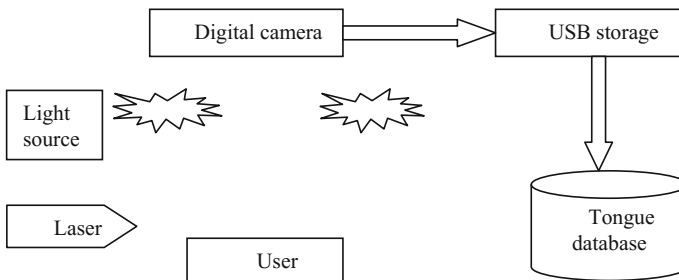
Gabor filters are bandpass filters, which are used in image processing for feature mining, texture analysis, and stereo inequality assessment. Gabor filter is generally used for describing textures. It performs very well in classifying images with different textures. From an image, if feature extraction is required, then different set of Gabor filter with different frequencies and orientation is helpful for extraction. It is generally used in pattern analysis. Under certain conditions, the phase of the response of Gabor filter is approximately linear. Before to perform extraction of features from image, preprocessing is required as Image stretching. Extraction of area of concentration is in the Gabor filter, is the tongue image normalization with respect to position, orientation, scale, reaction. Gabor filter can be used for object detection, image representation, color and pattern gradient, etc. Gabor filter is used in feature extraction for texture analysis, and it decomposes the image into components corresponding to different scale and different orientation.

### **3 Procedure for Tongue Biometric System**

See Fig. 5.



**Fig. 5** Procedure for tongue biometric system



**Fig. 6** Setup to capture tongue image

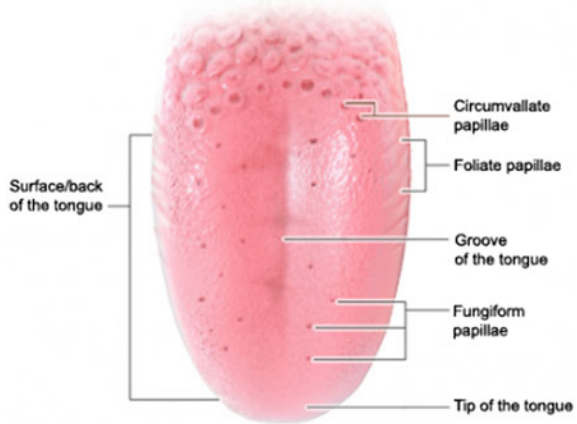
### 3.1 Setup to Capture Tongue Images

To capture proper tongue image, we should use two camera, so that proper image of tongue from profile and front view will be clear, and it should be in proper lightning so that complete image of tongue can be captured (Fig. 6).

### 3.2 Characteristic of Tongue Which Make It Unique

- Groove of Tongue: Length of groove of tongue matters because every one’s tongue has different length of groove of tongue with other different characteristic.
- Color of Tongue: Tongue has different color, which gives it unique identity. Every person has different tongue color because of different type of coating. Tongue color may be dark red, whitish, light red, purple, regular red with yellow, black and white coating.

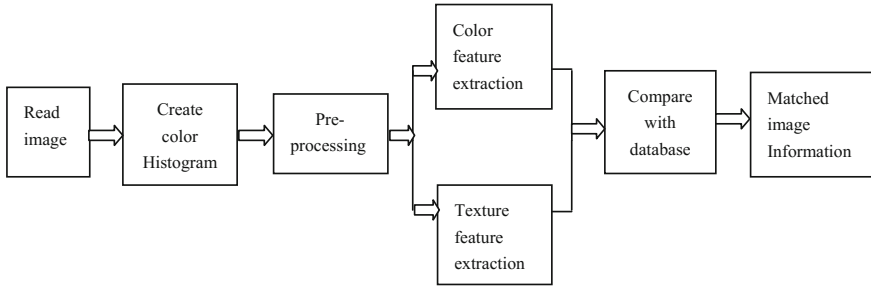
**Fig. 7** Characteristic of tongue, which make it unique



- **Tip of Tongue:** Different tongue has different type of tip. Tip of tongue may be of long, broad, and color of tongue also varies as it may dark red or different color of red with coating of different color making this characteristic of tongue unique.
- **Surface of Tongue:** Everyone's tongue has different surface, which we can call texture of tongue; it may be of smooth and rough.
- **Broadness of Tongue:** Tongue broadness varies person to person, and this characteristic of tongue also helps in making tongue unique.
- **Width of Tongue:** Tongue width also differs for all person, and it gives its contribution to make it unique (Fig. 7).

### 3.3 Different Tongue Type

- **Normal tongue:** In normal tongue, there is no mark of anything only shape is the key point.
- **Blood deficiency:** Blood deficiency tongue has a pale shape and it has minute or no coating.
- **Heat:** In heat type of tongue, thin yellow coating is there and some redness at corner present.
- **Damp retention:** In damp retention, tongue is swollen and white greasy layer is there.
- **Blood stasis:** In blood, stasis black spots are there and tongue color is purple.
- **Yang deficiency:** In yang deficiency thin white coating is there and pale swollen tongue.
- **Qi deficiency:** In Qi deficiency teeth marks are there thin white coating is there and pale tongue with red spots.



**Fig. 8** Tongue recognition system

- Yin deficiency: In Yin deficiency slight or no coating with fissures on tongue surface, also called fissure tongue and color of tongue is red.
- Damp heat: In damp heat tongue, greasy yellow coating with red color tongue.
- Qi stagnation: In Qi stagnation red tip of tongue is considered.

### 3.4 Methodology

The proposed work is meant to ensure efficient tongue recognition from already-created tongue database.

#### Algorithm for tongue detection:

See Fig. 8.

Step 1: Read image.

Step 2: Convert color image into binary image.

Step 3: Fill black hole.

Step 4: Create boundary of the binary image and convert into color image.

#### Tongue recognition system:

Algorithm:

Step 1: First, load the database into workspace.

Step 2: Load the query image.

Step 3: Create the color histogram of query image.

Step 4: Calculate mean value for these individual color histograms.

Step 5: Calculate the average of three histogram's mean value which is three (Red, Green, Blue).

- Step 6: Calculate standard deviation for these individual color histograms.
- Step 7: Calculate the average of three histogram’s standard deviation which is three (Red, Green, Blue).
- Step 8: Convert image into grayscale image.
- Step 9: Identify the ROI (Region of Interest).
- Step 10: Divide the ROI into four sub-blocks and reduce dimension of feature vector.
- Step 11: Apply Gabor filter on all four sub-blocks and extract important texture feature from the tongue image.
- Step 12: All extracted features’ mean and standard deviation are calculated, and then average of all four Gabor value’s mean and standard deviation is calculated and then compared with values stored database; if values for all are same, then matching result will pop up.

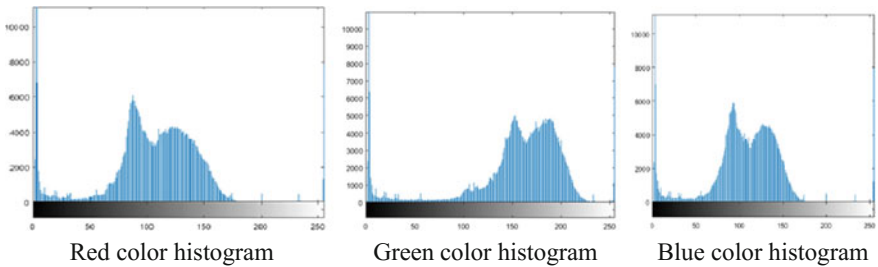


Image	Color component	Mean	Standard deviation
Tongue ROI 1	R, G, B	136.3157, 94.7152, 95.2004	70.3589, 54.2227, 54.0165
Tongue ROI 2	R, G, B	171.3370, 125.1681, 124.4985	59.1520, 55.3669, 50.0746
Tongue ROI 3	R, G, B	163.2757, 98.1515, 119.3497	76.8424, 57.0548, 55.5330

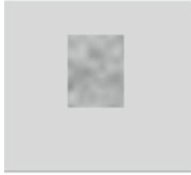
Mean and std of different color histogram

Image	Orientation	Mean (W, sigma, x, y)	Std (W, x, y, sigma)
Image 1	90	-0.0770, 0.1297, -0.1248, 0.0685	0.4155, 0.4259, 0.4145, 0.4281
Image 2	90	0.4787, -0.1012, -0.1350, -0.3057	0.26550, 0.4104, 0.4167, 0.4072
Image 3	90	-0.1284, 0.3206, 0.2828, 0.1380	0.3922, 0.2490, 0.1618, 0.3912

Mean and std for Gabor filter



ROI of Tongue Image



1st sub block



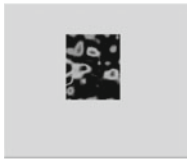
2nd sub block



3rd sub block



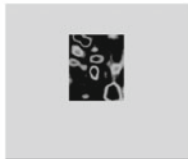
4th sub block



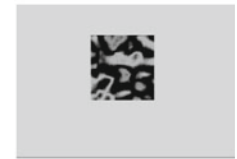
1<sup>st</sup>Sub block



2<sup>nd</sup> Sub block



3<sup>rd</sup> Sub block



4<sup>th</sup> Sub block

Gabor filtered Sub blocks



RESULT

## 4 Conclusion

A number of biometrics have been used and developed for unique authentication, and a very limited work has been done in tongue print recognition system and its use in any application. The human tongue promises to deliver a unique identification system than other as finger print and iris and other biometric cannot match in context of it well protected in mouth and difficult to forge. In this work, we have described how tongue biometric can be used for authentication and recognition. We have discussed different methods of feature extraction and implemented on tongue biometric. Tongue is a new topic in biometric research and for more analysis large database is required for the experiment to foster the research and to use tongue as a biometric for authentication.



## References

1. Zhi, L., Yan, J., Zhou, T., Tang, Q.: Tongue Shape Detection Based on B-Spline ICMLC 2006, August 2006, vol. 6, pp. 3829–3832 (2006)
2. Naaz, R., Kundra, S., Garg, P., Sharma, A.: Tongue biometric and its application in public use system. In: 2011 3rd International Conference on Machine Learning and Computing (ICMLC 2011) (2011)
3. Garibotto, G.; ElSagDatamat Spa: Video surveillance and biometric technology applications. In: Proceedings of Sixth IEEE International Conference Advanced Video and Signal Based Surveillance, p. 288 (2009)
4. Liu, Z., Yan, J.-Q., Zhang, D., Tang, Q.-L.: A tongue-print image database for recognition. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, August 2007, pp. 19–22 (2007)
5. Li, C., Yuen, P.: Tongue image matching using color content. *Pattern Recognit.* **35**(2), 407–419 (2002)
6. Diwakar, M., Maharshi, M.: An extraction and recognition of tongue-print images for biometrics authentication system. *Int. J. Comput. Appl.* (0975–8887) **61**(3) (2013)
7. Lahmiri, S.: Recognition of tongueprint textures for personal authentication: a wavelet approach. *Int. J. Comput. Appl.* **3**(3) Aug (2012)

# Sample Entropy Based Selection of Wavelet Decomposition Level for Finger Movement Recognition Using EMG



Nabasmita Phukan and Nayan M. Kakoty

**Abstract** This paper reports the recognition of five finger movements using forearm EMG signals. A relationship between the sample entropy (SampEn) of EMG signals at four wavelet decomposition levels and classification accuracy has been established. Experiments with the EMG at third level of wavelet decomposition can classify the finger movements with a maximum accuracy of 95.5%. These results show that EMG at the decomposition level which possess minimum SampEn produces the maximum classification accuracy. The experimental result shows that this relationship is a very useful criterion for selection of wavelet decomposition level to recognize EMG-based finger movements.

**Keywords** Wavelet transform · Sample entropy · SNR · EMG

## 1 Introduction

In the field of biomedical signal processing, EMG-based prosthesis control has been a significant contribution. Although number of reports have been available for control of prosthesis through EMG based recognition of hand movements [1, 2], recognition of finger movements have received lesser attention. This is mainly because of non-deterministic nature of EMG signals, which otherwise holds promise for more dexterous control of prosthesis.

---

N. Phukan · N. M. Kakoty (✉)  
Embedded Systems and Robotics Laboratory, Tezpur University,  
Tezpur 784028, India  
email: [nkakoty@tezu.ernet.in](mailto:nkakoty@tezu.ernet.in)  
URL: <http://tezu.ernet.in/erl>

N. Phukan  
e-mail: [nabasmitap@gmail.com](mailto:nabasmitap@gmail.com)

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_6](https://doi.org/10.1007/978-981-13-1708-8_6)

Tsenov et al. [3] reported the recognition of four finger movements using neural network classifier with 93% accuracy. Control of an underactuated prosthetic hand using EMG has been reported by Zhao et al. [4]. The control part of prosthetic hand was based on neural network learning techniques and the parametric autoregressive model. Tenore et al. [5] have reported the recognition of five fingers using 32 EMG electrodes with an accuracy of 98%. However, a reduction in the number of electrodes, without compromising with the classification accuracy, would significantly simplify the requirements for control. Towards this end, Ouyang et al. [6] have proposed an efficient feature projection method using linear discriminant analysis for EMG pattern recognition based on lower number of channels.

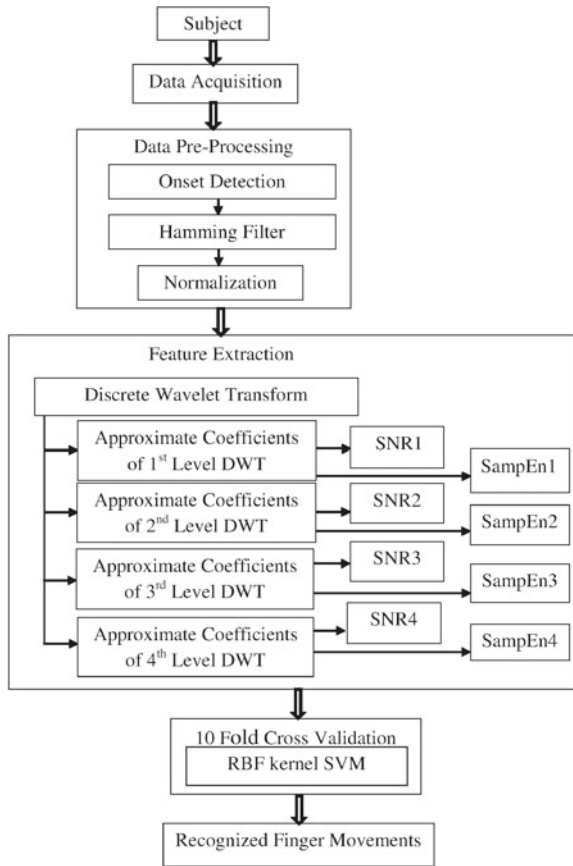
Cipriani et al. [7] reported real-time experiments, involving able-bodied and amputated participants, using eight EMG channels for classification of seven finger movements. Although there has been advances in analysing EMG, a focused methodology for the recognition of finger movements, i.e. both flexion/extension and abduction/adduction based on lower number of EMG channels need to be extensively explored for finer control of prosthesis.

This paper reports a methodology for recognition of five finger movements based on two-channel EMG, establishing a relationship among the level of EMG wavelet decomposition and sample entropy (SampEn) for higher recognition rate. The rest of the paper is arranged as follows: Sect. 2 describes the proposed methodology for finger movement recognition. Materials and methods including the recognition of ten class finger movements are presented in Sect. 3. Section 4 describes results and discussions followed by the concluding remarks in Sect. 5.

## 2 Finger Movements Recognition Architecture

Figure 1 shows the proposed architecture of finger movement recognition. The EMG signals are obtained from forearm muscles for the flexion–extension movements of the five fingers in the data acquisition stage. The EMG acquisition was in line with the permission of the Tezpur University Ethical Committee and with the informed consent from the volunteering participants. These EMG signals are preprocessed in the preprocessing stage to accurately record, visualize and analyse. This is done through sampling the EMG at a rate of 1 kHz, EMG onset detection and hamming filtering. The preprocessed EMG signals were normalized to remove subjectivity based noises. Following the noise removal, DWT coefficients were extracted in the feature extraction stage. The approximate coefficients, which contain the most important information of the original signal [8], were considered as features at four decomposition levels. Signal-to-noise ratio (SNR) and SampEn were calculated at each level of decomposition. The recognition of the finger movements was through a ten-fold cross validated support vector machine.

**Fig. 1** Proposed finger movements recognition architecture



### 2.1 Signal to Noise Ratio

SNR is a measure for signal strength in comparison to the noise in it. The SNR in an EMG signal ( $x_i$ ) is calculated as given by Eq. 1 wherein  $\bar{x}_i$  = average of  $x_i$  and  $S_x$  = standard deviation of  $x_i$ .

$$SNR = 20 \times \log[\bar{x}_i/S_x] \tag{1}$$

### 2.2 Sample Entropy

SampEn quantifies the complexity and regularity of the signal [9] and is given by the negative natural logarithm of an estimate of the conditional probability [10]. SampEn in the EMG signal gives us the randomness of the information content in the feature

set and is given in (2).

$$SampEn(k, r, N) = -\ln(A(k)/B(k - 1)) \quad (2)$$

where  $r = 0.2$  and  $k = 0, 1, \dots, m - 1$  with  $B(0) = N$ , the length of the input DWT feature set.

### 3 Materials and Methods

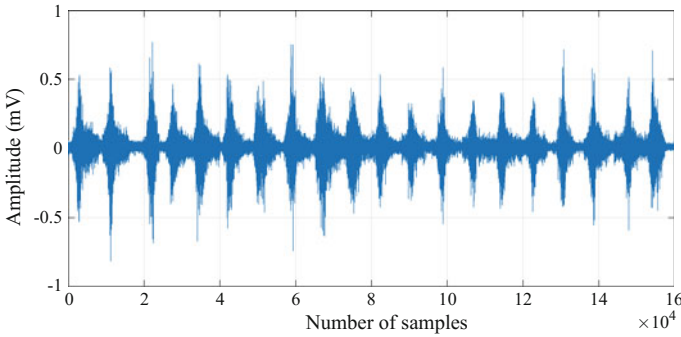
#### 3.1 Data Acquisition

Based on the fact that an amputee can produce EMG similar to that of a healthy subject [11], EMG signals were collected from four healthy subjects of age group between 18 and 35 years performing flexion and extension for ten trials. Two-channel Ag/AgCl surface electrodes were used for EMG acquisition. These were arranged along the longitudinal midline of the muscles to detect improved superimposed EMG signals [12]. The placement of the electrodes and muscle selection is tabulated in Table 1.

During data acquisition, subjects were comfortably seated and instructed to rest their forearm on the armrest of the chair. Subjects were instructed to perform the flexion–extension with a comfortable and consistent level of effort for approximately 8 s, and then relax. This was repeated for five times and recorded in one trial. Additionally, spontaneous EMG signal during hand relaxation was recorded as a different class. Sufficient relaxation time between trials ( $\approx 30$  min) and between repetitions of the same movement ( $\approx 12$  s) was allowed. A total of (4 subjects  $\times$  5 fingers  $\times$  2 movements  $\times$  10 trials) = 400 two-channel EMG signals have been considered for the experiment. These signals were sampled at 1 kHz followed by a band pass filter set at 10–200 Hz, and 50 Hz notch.

**Table 1** EMG electrodes placement on subject during acquisition

Electrodes	Muscles	Functions
Channel 1	Flexor digitorum	Finger flexion
Channel 2	Extensor digitorum	Finger extension
Reference	Ulnar styloid	Reference



**Fig. 2** Raw EMG of flexion/extension

### 3.2 EMG Preprocessing

#### 3.2.1 Onset Detection

The EMG signals acquired during hand relaxation or at resting position were of constant amplitude. These irrelevant EMG signals were discarded by the EMG onset detection technique. The EMG having amplitude more than three times of the standard deviation (SD) of the EMG at resting position were extracted [13]. The threshold value was fixed at three times the SD by observing the amplitude difference of the EMG signal during finger movements and at resting position. The extracted EMG signal is expressed as:

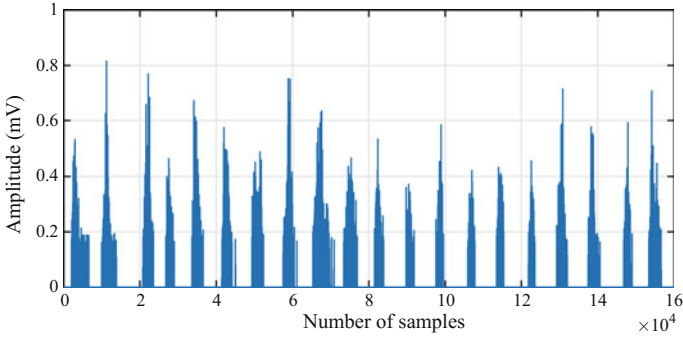
$$x = \sum_{i=1}^N (x_i) \tag{3}$$

Wherein  $x_i = x_i$ , if  $x_i \geq Th$  and  $x_i = 0$ , if  $x < Th$

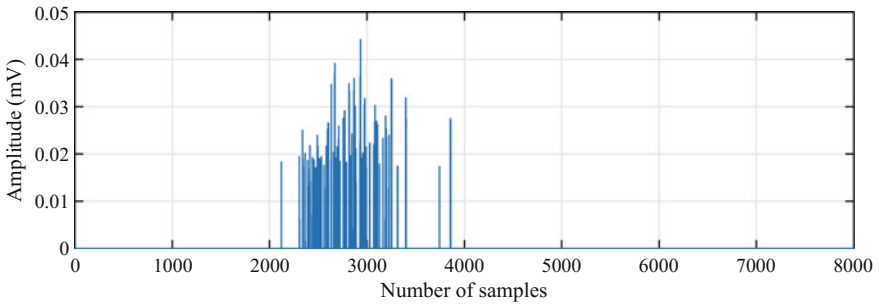
with  $x_i$  being the  $i$ th sample value of EMG signal  $x$  and  $Th$  is the threshold value for onset detection. Figure 2 shows the EMG signals during finger movements during for trials. Figure 3 shows the EMG signal following the onset detection in line with the Eq. 3.

#### 3.2.2 Hamming Filtering

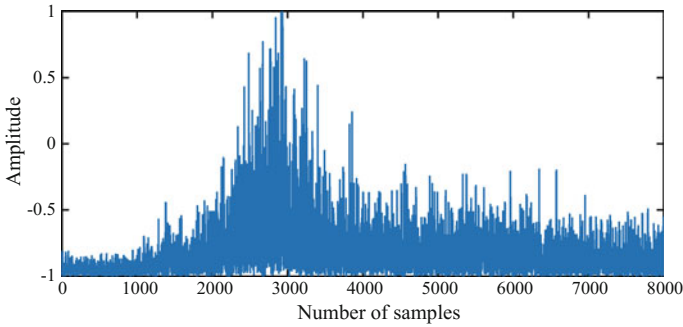
On onset detection, Hamming window was applied on the signal to extract EMG during either one flexion or one extension of finger movements. EMG signal obtained through hamming filter can be expressed as in Eq. 4. Figure 4 shows the EMG signal passed through the Hamming filter.



**Fig. 3** Onset detected raw EMG



**Fig. 4** Hamming filtered EMG



**Fig. 5** Normalized EMG signal

$$x = h(n) * x_i \tag{4}$$

where  $h(n) = 0.54 - 0.46 \cos((2\pi n)/N)$ ,  $0 \leq n \leq N$ , with  $N = L + 1$ , and  $x_i$  being the  $i$ th sample value of the EMG signal  $x$ ,  $h(n)$  is the windowed EMG signal obtained from the hamming window of size  $L$ .

### 3.2.3 Normalization

EMG signals are influenced by different artifacts like thickness of skin layer, crosstalk by other biosignals, electrode size and position [14]. To reduce the effect of these factors, normalization on the EMG was performed using Eq. 5 [15]. Figure 5 shows the normalized EMG.

$$x_{norm} = [(x_i - x_{min}) / (x_{max} - x_{min})] \times 2 - 1 \quad (5)$$

where  $x_{min}$  and  $x_{max}$  are minimum and maximum values of EMG signal  $x_i$ .

## 3.3 Feature Extraction

Feature extraction and selection are of the most significant for pattern recognition [16]. Time/frequency domain features, which represent both time and frequency domain information and suited for characterizing the non-stationary nature of signals, are better suited for EMG recognition [17]. Wavelet transform is the technique to transform a signal into time–frequency domain. Continuous wavelet transform (CWT) and discrete wavelet transform (DWT) are the two methods for wavelet transformation. DWT exhibits good frequency resolution at low frequencies and good time resolution at high frequencies [18]. Furthermore, for real-time signal processing issues, DWT is considered more efficient [16]. Based on these findings, DWT was chosen to extract the EMG features for recognition of finger movements in our experiment.

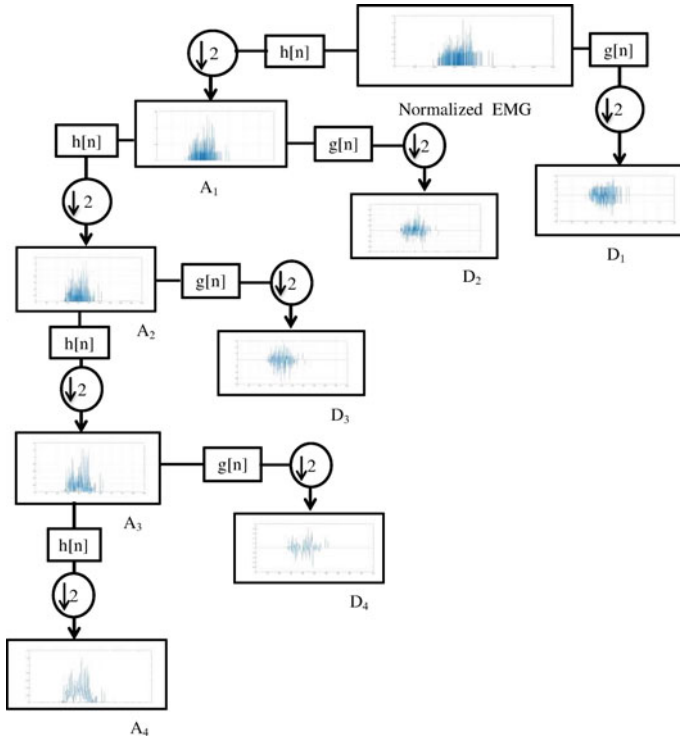
### 3.3.1 Discrete Wavelet Transformation

DWT technique iteratively transforms EMG signal into multi resolution subsets of coefficients by passing EMG through a high-pass and a low-pass filters. At each level of decomposition, a subset of detailed coefficient ( $D_j$ ) and approximation coefficient ( $A_j$ ) are obtained using an L-sample high-pass filter  $g$ , and an L-sample low-pass filter  $h$ . Both approximation and detail signals are downsampled by a factor of two. This can be expressed as follows:

$$A_j[n] = H \langle A_{j-1}[n] \rangle = \sum_{k=0}^{L-1} h[k] A_{j-1}[2n - k] \quad (6)$$

$$D_j[n] = H \langle D_{j-1}[n] \rangle = \sum_{k=0}^{L-1} g[k] A_{j-1}[2n - k] \quad (7)$$





**Fig. 6** Wavelet decomposition after four levels of wavelet transformation

where H and G represent the convolution/downsampling operators. Sequences  $g[n]$  and  $h[n]$  are associated with wavelet function  $\psi(t)$  and the scaling function  $\phi(t)$  as follows:

$$g[n] = \langle \psi(t), \sqrt{2} \cdot \psi(2t - n) \rangle \tag{8}$$

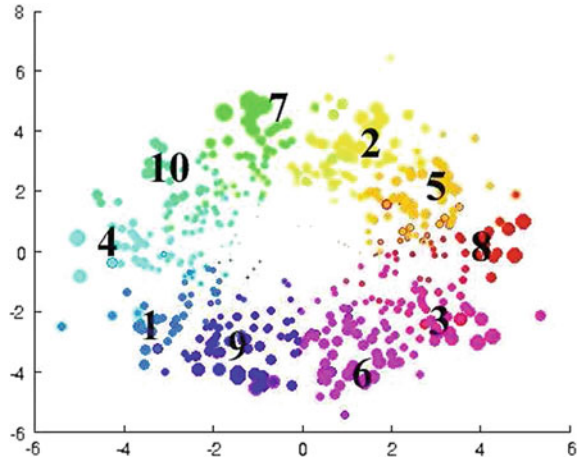
$$h[n] = \langle \phi(t), \sqrt{2} \cdot \phi(2t - n) \rangle \tag{9}$$

To achieve optimal performance in the wavelet analysis, Daubechies (db2) was selected as mother wavelet [19]. Figure 6 shows the wavelet decomposition of the normalized EMG for one flexion movement.

### 3.4 Recognition of Finger Movements

The approximate coefficients from each level of wavelet decomposition have been fed into a radial basis function (RBF) kernel SVM-based classifier. The multiclass SVM classify the ten classes of data for flexion and extension movements of five fingers—

**Fig. 7** Recognized features belonging to the five finger movements. 1: index flexion, 2: index extension, 3: middle flexion, 4: middle extension, 5: ring flexion, 6: ring extension, 7: little flexion, 8: little extension, 9: thumb flexion and 10: thumb extension



index, middle, ring, little and thumb. SVM maps the features to a higher dimensional feature space. The fundamental feature of a SVM is the separating maximum-margin hyperplane whose position is determined by maximizing its distance from the support vectors of different classes. Following [8], the decision function which classifies the feature vector is expressed as:

$$f(x) = b + \sum_{i=1}^F w_i \cdot k(y_i) \quad \text{satisfying} \quad \min[\phi(w_i)] = \frac{1}{2}(w_i \cdot w_i)$$

where b is bias term, F is total number of input features and  $w_i$  is normal to the  $i$ th feature space and

$$y_i = +1 \quad \text{if} \quad w_i \cdot \rho + b > 1 \quad \text{or} \quad y_i = -1 \quad \text{if} \quad w_i \cdot \rho + b < 1$$

with  $\rho$  as input feature set, i.e. DWT approximate coefficients. The RBF kernel  $k$  used in the experiment is given by:

$$k(y_i) = \exp(-\gamma(\|y_i - y\|^2))$$

Figure 7 shows the recognized features belonging to the ten class for five finger movements.

**Fig. 8** Confusion matrix with the first-level approximate coefficients

		Actual class label										Acc (%)
		1	2	3	4	5	6	7	8	9	10	
Predicted class label	1	15	2	2		1		1				75
	2		17		1	2					6	85
	3			17	1							85
	4				17							85
	5	2				15	4					75
	6			1			14	4				70
	7				1			11	8			55
	8		1			2		4	12	2		60
	9	2								17		85
	10	1					1			1	14	70
Average accuracy percentage											74.5	

**Fig. 9** Confusion matrix with the second-level approximate coefficients

		Actual class label										Acc (%)
		1	2	3	4	5	6	7	8	9	10	
Predicted class label	1	17		2						1		85
	2		18				1				1	90
	3			17	1			2				85
	4				18					2		90
	5					19	1					95
	6			1			17					85
	7						1	18	1			90
	8		1						16			80
	9					1			1	17	1	85
	10	1			1				2		18	90
Average accuracy percentage											87.5	

### 4 Results and Discussions

The approximate coefficients extracted at four levels of wavelet decomposition have been explored individually for recognition of finger movements. The confusion matrices in Fig. 8 through 11 shows the recognition of ten movements. Labels in confusion matrices corresponds to finger movements in line with Fig. 7 (Figs. 9 and 10).

It has been observed that the third level of wavelet decomposition coefficients produced the highest recognition rate, i.e. 95.5%. The SNR of the EMG features has been calculated at each level of decomposition and is shown in Fig. 12. It has been observed that the SNR values increase with an increase in decomposition levels. It is obvious as the noise decreases with each level of decomposition.

The SampEn values of the approximate coefficients at each level of decomposition is shown in Fig. 13. It has been observed that the SampEn is minimum for coefficients at third level of decomposition, i.e. coefficients resulting in the highest recognition rates. Based on these experimental results, it can be observed that although with the increase in the level of wavelet decomposition, the signal strength increases

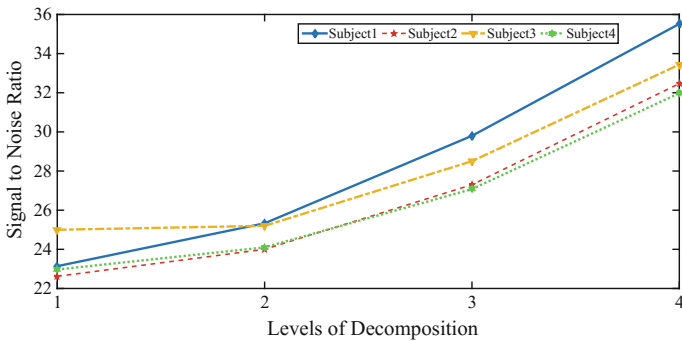
**Fig. 10** Confusion matrix with the third-level approximate coefficients

		Actual class label										Acc (%)	
		1	2	3	4	5	6	7	8	9	10		
Predicted class label	1	19											95
	2		19										95
	3			19	1								95
	4				19								95
	5					20	1						100
	6				1			19					95
	7								20	1			100
	8			1						18	1		90
	9									1	19	1	95
	10	1										19	95
Average accuracy percentage											95.5		

**Fig. 11** Confusion matrix with the fourth-level approximate coefficients

		Actual class label										Acc (%)	
		1	2	3	4	5	6	7	8	9	10		
Predicted class label	1	18								2			90
	2		17				1	1					85
	3			18	1								90
	4				16			1	2				90
	5					19	1						95
	6			2		1	17						85
	7							17	1				85
	8		2						16	1			80
	9								1	17	4		85
	10	2			1		1					16	80
Average accuracy percentage											86.5		

compared to the noise, establishing SampEn as a reliable paradigm to determine the decomposition level for highest recognition rates.



**Fig. 12** SNR values at four decomposition levels

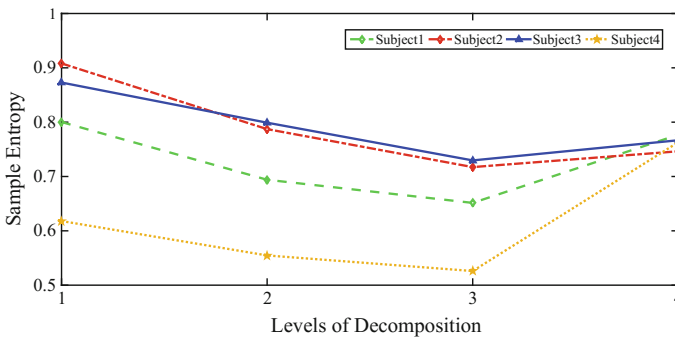
## 5 Conclusions

This paper presented recognition of five finger movements: flexion and extension using two-channel EMG. The approximate coefficients obtained through DWT at four decomposition levels consist of the feature sets. The classification was through a RBF kernel SVM. The SampEn and SNR of the feature sets at four decomposition levels have been evaluated. From the experimental results, it has been found that the approximate coefficients at third level of wavelet decomposition resulted in the highest recognition rate of 95.5%. The SNR of the approximate coefficients increases linearly with the increase in the decomposition level. The SampEn is lower with approximate coefficients at third level of decomposition. Based on these experimental results, it can be observed that although with the increase in the level of wavelet decomposition, the signal strength increases compared to the noise as indicated by the SNR values, SampEn is a reliable paradigm to determine the decomposition level for the highest recognition rates. These results show that the EMG at the decomposition level which possess minimum SampEn produces the maximum classification accuracy. The experimental results show that this relationship is a very useful criterion for selection of wavelet decomposition level for recognition of EMG based finger movements.

**Acknowledgements** Centre of Excellence in Machine Learning and Big Data Analysis, Tezpur University, funded by Ministry of HRD, Government of India.

## References

1. Kakoty, N.M., Hazarika, S.M., Gan, J.Q.: EMG feature set selection through linear relationship for grasp recognition. *J. Med. Biol. Eng.* **36**(6), 883–890 (2016)
2. Ravindra, V., Castellini, C.: A comparative analysis of three non-invasive human-machine interfaces for the disabled. *Front. Neurobot.* (2014)



**Fig. 13** Sample entropy values at four decomposition levels

3. Tsenov, G., Zeghib, A.H., Palis, F., Shoylev, N., Mladenov, V.: Neural networks for online classification of hand and finger movements using surface EMG signals. *Neural Netw. Appl. Electr. Eng.* **8** (2006)
4. Zhao, J., Xie, Z., Jiang, L., Cai, H., Liu, H., Hirzinger, G.: A EMG control for a five-fingered prosthetic hand based on wavelet transform and autoregressive model. In: *IEEE International Conference on Mechatronics and Automation*, pp. 1097–1102, USA (2006)
5. Tenore, F.V.G., Ramos, A., Fahmy, A., Acharya, S., Cummings, R.E., Thakor, N.V.: Decoding of individuated finger movements using surface electromyography. *IEEE Trans. Biomed. Eng.* **56**(5), 167–171 (2009)
6. Ouyang, G., Zhu, X., Z.J., Liul, H.: High-density myoelectric pattern recognition toward improved stroke rehabilitation. *IEEE Trans. Biomed. Eng.* **18**(1), 257–265 (2014)
7. Cipriani, C., Antfolk, C., Controzzi, M.: Online myoelectric control of a dexterous hand prosthesis by transradial amputees. *IEEE Trans. Neural Syst. Rehabil. Eng.* **19** (2011)
8. Kakoty, N.M., Hazarika, S.M.: Recognition of grasp types through PCs of DWT based EMG features. In: *IEEE International Conference on Rehabilitation Robotics*, pp. 478–482, Zurich (2011)
9. Richman, J.S., Moorman, J.R.: Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **278**(6), H2039–H2049 (2000)
10. Lake, D.E., Richman, J.S., Griffin, M.P., Moorman, J.R.: Sample entropy analysis of neonatal heart rate variability. *Am. J. Physiol.-Regul. Integr. Comp. Physiol.* **283**(3), 789–797 (2002)
11. Crawford, B., Miller, K., Shenoy, P., Rao, R.: Real-time classification of electromyographic signals for robotic control. Technical Report 2005-03-05, Department of Computer Science, University of Washington (2005)
12. Boostani, R., Moradi, M.H.: Evaluation of the forearm EMG signal features for the control of a prosthetic hand. *Physiol. Meas.* **24**(2), 309–319 (2003)
13. Atzori, M., Gijssberts, A., Kuzborskij, I., Elsig, S., Hager, A.G.M., Deriaz, O., Castellini, C., Miller, H., Caputo, B.: Characterization of a benchmark database for myoelectric movement classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* **23**(1), 73–83 (2015)
14. Reaz, M.B.I., Hussain, M.S., Mohd-Yasin, F.: Techniques of EMG signal analysis: detection, processing, classification and applications. *Biol. Proc. Online* **8**(1), 11–35 (2006)
15. Aung, Y.M., Al-Jumaily, A.: Estimation of upper limb joint angle using surface EMG signal. *J. Adv. Robot. Syst.* **10**(10), 369–376 (2013)
16. Oskoei, M.A., Hu, H.: Myoelectric control systems: a survey. *J. Biomed. Signal Process. Control* **2**, 275–294 (2007)
17. Sheean, L.G.: Application of time-varying analysis to diagnostic needle electromyography. *Med. Eng. Phys.* **34**(2), 249–255 (2012)
18. Phinyomark, A., Phukpattaranont, P., Limsakul, C.: Optimal wavelet functions in wavelet denoising for multifunction myoelectric control. *Trans. Electr. Eng. Electron. Commun.* **8**(1), 43–52 (2010)
19. Phinyomark, A., Limsakul, C., Phukpattaranont, P.: Evaluation of wavelet function based on robust EMG feature extraction. In: *The Seventh PSU Engineering Conference*, pp. 277–281 (2009)

# Skin Detection Using Hybrid Colour Space of RGB-H-CMYK



Ashish Kumar and P. Shanmugavadivu

**Abstract** Skin detection is an essential step in human face detection and/or recognition, using digital image processing techniques. This paper presents a new human skin detection technique, termed as RGB-H-CMYK that uses triple colour spaces of an image, namely RGB (Red, Blue and Green), H (Hue of HSV) and CMYK (Cyan, Magenta, Yellow and Black). In this proposed method, threshold-based rules are applied on RGB, H and CMYK for skin classification. The input image in these three hybrid colour schemes is explored in different combination such as RC (RGB and CMYK), RH (RGB and H) and RHC (RGB and H and CMYK). The RHC\_Vote qualifies the current pixel as skin pixel when at least two rules vote for it. The computational merit of this hybrid colour scheme-based skin detection is validated on the real-time dataset and ECU skin database. The average *Recall* and *Accuracy* of this method is recorded as 85% and 89%, respectively. This approach is confirmed to have an edge over its competitive methods, as it promises object localization, based on neighbourhood intensities without using the computationally complex approaches such as facial texture and geometric properties.

**Keywords** RGB · HSV · CMYK · RGB-H-CMYK · Skin detection  
Human skin classification · Face detection · Face recognition

## 1 Introduction

Human skin detection is a de facto pre-processing task for human face localization in a variety of human–computer interaction (HCI) applications such as face detection [1], face recognition [2], face tracking, object tracking and crowd analysis [3],

---

A. Kumar (✉) · P. Shanmugavadivu  
Department of Computer Science and Applications, Gandhigram Rural  
Institute – Deemed University, Gandhigram, India  
e-mail: ashishgru@gmail.com

P. Shanmugavadivu  
e-mail: psvadivu67@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_7](https://doi.org/10.1007/978-981-13-1708-8_7)

content-based image retrieval [4], steganography [5] and adult image filtering [6]. Though human skin colour detection through human visual perception is an effortless task, it is deemed as a complicated procedure through machine vision. In general, skin colour detection can be performed in two flavours as point processing and region bound. In the former, each pixel is considered as an independent entity and its candidature for skin or non-skin pixel is governed by a set of predefined rules, whereas in the latter, the skin pixel classification is performed based on the intensity/texture of the neighbourhood pixels.

## ***1.1 Skin Colour Modelling***

The colour model of a digital image depicts a significant role in representing its information content [7]. Many researchers have proposed newer computational solutions to explore the intensity profile of the input images, in order to classify the intensity details into skin or non-skin, based on a set of predefined rules. These techniques are broadly categorized into three as explicit threshold techniques, statistical techniques and machine learning techniques [7–9].

### **1.1.1 Explicit Threshold Techniques**

These methods primarily choose one or more intensity values to set the threshold, based on which the RGB-H-CMYK skin detection techniques are devised. The selection of threshold is varied with the choice of colour space. These techniques are computationally simple and assure faster execution [10].

### **1.1.2 Statistical Techniques**

The statistical measures of an image, namely mean, median, variance, standard deviation etc., describe the local and global intensities profile, which are used in numerous image processing computational methods. These techniques are generally classified into parametric and nonparametric methods. Parametric methods assume that sample data come from a population that follows a probability distribution of image intensity. Parametric techniques use a modelled colour space with a prescribed geometric shape. Gaussian [11] and elliptical boundary model [12] are the illustrations of parametric skin colour modelling. Parametric model performance varies significantly among the colour spaces. In nonparametric techniques, a histogram for the given colour space is built, and subsequently, probability density function (PDF) is computed. Either each pixel is classified as skin pixel or non-skin pixel, based on its PDF exceeds the predefined threshold. The main advantage of nonparametric models such as Bayes classifier is fast training and usage and is theoretically independent of the intensity distribution of skin pixels. The performance of nonparametric methods



directly depends on the representativeness of the training images set that demands more storage space [7].

### 1.1.3 Machine Learning Techniques

This category of techniques is also referred as dynamic classifier, self-learning techniques and semi-parametric methods. Such models are trained to classify the input images into either as positive or negative sets (skin and non-skin set) [13–15]. The emergence of new strategies using clustering, classification methods, namely KNN, adaptive Bayes classifier, self-organizing map etc., which fall under this category is mostly problem-specific, rather than being generic [16]. The proposed RGB-H-CMYK skin detection technique falls under explicit threshold technique category, and it uses the combination of three dominant colour schemes RGB, H of HSV and CMYK.

The rest of this article is organized as follow. In Sect. 2, literature review is given and in Sect. 3 the methodology of RGB-H-CMYK is described. The experimental results are discussed in Sect. 4. The conclusions are given in Sect. 5.

## 2 Literature Review

The RGB, HSV and YCbCr are broadly considered as the most popular by used colour spaces for skin detection. The researchers have suggested either single or combination of these colour spaces for the effective differentiation of skin pixels from non-skin pixels. RGB colour space is considered as the most common choice for many researchers [16]. However, the combined effect of chrominance and luminance information together in RGB channels limits its applications in certain cases.

The intuitiveness of the HSV (Hue-Saturation-Value) components and explicit discrimination between luminance and chrominance properties are well depicted in HSV colour space, which serves as a key factor for skin pixel detection. Hue is observed to be invariant to white light sources, matte surfaces, as well as to ambient light and surface orientation relative to the light source that makes this colour space as a competitive alternate to RGB in skin detection [17].

YCbCr also provides explicit distinction between the luminance and chrominance components and also can be transformed from RGB. Mahmoud and Phung et al. [18, 19] insisted on YCbCr colour space uses for skin detection. Moreover, Hsu et al. [20], Khan et al. [21] and Phung et al. [22] suggested that RGB and YCbCr colour space are more suitable for skin pixel detection.

It is observed that CMYK colour space is less explored in the light of skin detection and is less suggested for skin pixel detection. Recently, Dariusz and Weronika [23] has claimed that CMYK as a good alternative for skin detection and suggested a set of boundary equations for skin region detection in CMYK colour components. Based on the previous research work in this domain of research, the authors of this

article have combined the potentials of RGB and the Hue of HSV along with the least explored colour space CMYK. It is apparent from the obtained results that this proposed method outperforms its competitive and contemporary methods in terms of *Recall*. Additionally, the combination of these colour spaces is proposed for better skin pixel detection as well as to overcome the reported disadvantages. It is recorded that the combination of colour spaces yields better skin detection results than those obtained using single colour space [24–30]. Due to its computational merits, this method confirmed to have an edge over its recent competitive methods. Hence, hybrid colour space oriented skin detection readily finds place in face detection as well as recognition.

### 3 Hybrid Colour Space Skin Detection

In hybrid colour space skin detection methods, two or more different colour space skin detection methods are joined together to classify the current pixel as skin or non-skin pixel. Explicit thresholding values of RGB, H of HSV and CMYK for skin segmentation are given below.

#### 3.1 RGB to CMYK Colour Conversion

As suggested by Dariusz and Weronika [23], RGB is converted into CMYK using the following equations:

$$K = \min(255-R, 255-G, 255-B) \quad (1)$$

$$C = (255-R-K)/(255-K) \quad (2)$$

$$M = (255-G-K)/(255-K) \quad (3)$$

$$Y = (255-B-K)/(255-K) \quad (4)$$

It is apparently evident that addition of K component to CMY alters the properties of this colour space. The computed K component is radically different from the K (black) in CMYK. Skin detection through the computed K component in CMYK colour space greatly influences the process of face localization, attributing to faster convergence towards the solution domain [25].

### 3.2 RGB Colour Scheme

The boundary restriction in RGB colour space suggested by Rahman [10] is as follows:

$$Rule1 = (R > 95) \text{ AND } (G > 40) \text{ AND } (B > 20) \text{ AND } (\max\{R, G, B\} - \min\{R, G, B\} > 15) \text{ AND } (|R - G| > 15) \text{ AND } (R > G) \text{ AND } (R > B) \quad (5)$$

$$Rule2 = (R > 220) \text{ AND } (G > 210) \text{ AND } (B > 170) \text{ AND } (|R - G| \leq 15) \text{ AND } (R > B) \text{ AND } (G > B) \quad (6)$$

$$RULE\_RGB = Rule1 \cup Rule2 \quad (7)$$

### 3.3 HSV Colour Scheme

As Hue values play a vital role between the skin and non-skin pixels, Rahman [10] suggested a subspace H boundary as:

$$Rule3 = H < 25 \quad (8)$$

$$Rule4 = H > 230 \quad (9)$$

$$RULE\_HSV = Rule3 \cup Rule4 \quad (10)$$

### 3.4 CMYK Colour Scheme

In CMYK, colour space skin colour boundary equations suggested by Dariusz and Weronika [23] are:

$$Rule5 = K < 205 \quad (11)$$

$$Rule6 = 0 \leq C \leq 0.05 \quad (12)$$

$$Rule7 = 0.0909 < Y < 0.945 \quad (13)$$

$$Rule8 = 0.1 \leq Y/M < 4.67 \quad (14)$$

$$RULE\_CMYK = Rule5 \cap Rule6 \cap Rule7 \cap Rule8 \quad (15)$$

### 3.5 Hybrid Colour Scheme

This method explores the different combination of these three rules as:

$$RC = Rule\_RGB \cap Rule\_CMYK$$

$$RH = Rule\_RGB \cap Rule\_HSV$$

$$RHC = Rule\_RGB \cap Rule\_HSV \cap Rule\_CMYK$$

$$RHC\_Vote = \min\_2vote(Rule\_RGB, Rule\_HSV, Rule\_CMYK)$$

where  $\min\_2vote$  is a function which returns true if at least two rules vote for the current pixels as a skin pixel. In addition, the proposed algorithm ignores all the bright pixels ( $\{R, G, B\} > 250$ ) from the test images. This greatly attributes to the accuracy of skin classification.

The devised triple colour scheme of RGB-H-CMYK-based skin detection technique aims to provide feasible solutions for skin detection. Based on Eqs. (7), (10) and (15), three separate detectors for RGB, H (Hue component of HSV) and CMYK colour space are generated. The four different combinations of these rules are tested against the input colour images for the possible skin area detection. These hybrid rules are named as RC (RGB-CMYK), RH (RGB-H), RHC (RGB-H-CMYK) and RHC\_Vote (RGB-H-CMYK subject to min two votes).

The authors of this research article have obtained promising results for these new combination of rules, for the chosen colour spaces of RGB-H-CMYK.

## 4 Result and Discussion

This technique was tested on publicly available different profile images. Moreover, this method is experimented on the databases created from publicly available profile images. ECU skin database [31] is used for comparing results of the proposed methods with Rahman's [10] method in terms of *Recall*. It is a measure of the positive data been classified as being positive. ECU skin database organized in three series, namely HGR1, HGR2A and HGR2B, consists of more than 1500 skin images and their skin masks. The results of the randomly selected public profile sample are depicted as illustrations in Figs. 1 and 2. First row of these figures shows original image and skin pixel classification in RGB, HSV and CMYK colour space, respectively. Second rows displays the results obtained with different combination RC (RGB-CMYK), RH (RGB-H), RHC (RGB-H-CMYK) and RHC\_Vote, respectively. RC approach declares less true skin pixel, whereas RHC\_Vote almost classifies all skin pixel correctly. In Fig. 2, some non-skin pixels are also classified as skin pixels due to their utmost similarity with skin pixels. RC, RH and RHC miss some genuine skin pixels inside the skin region itself, and this issue is taken care by RHC\_vote approach as

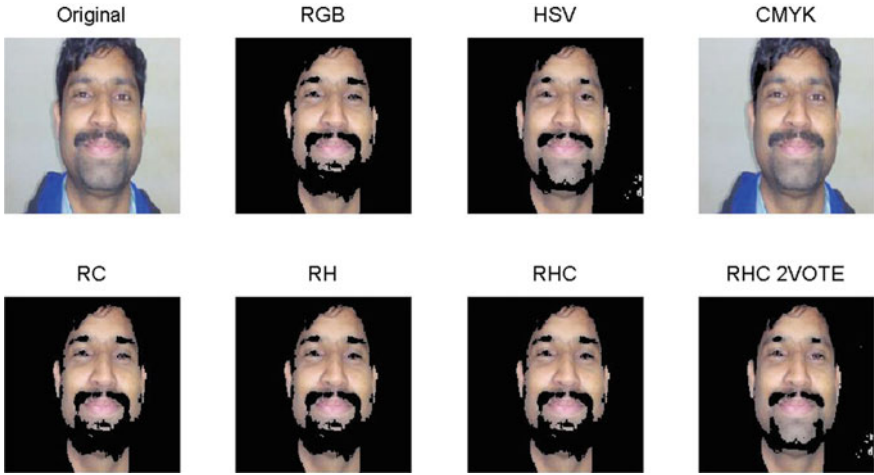


Fig. 1 Skin detection using single and hybrid colour schemes for sample 1

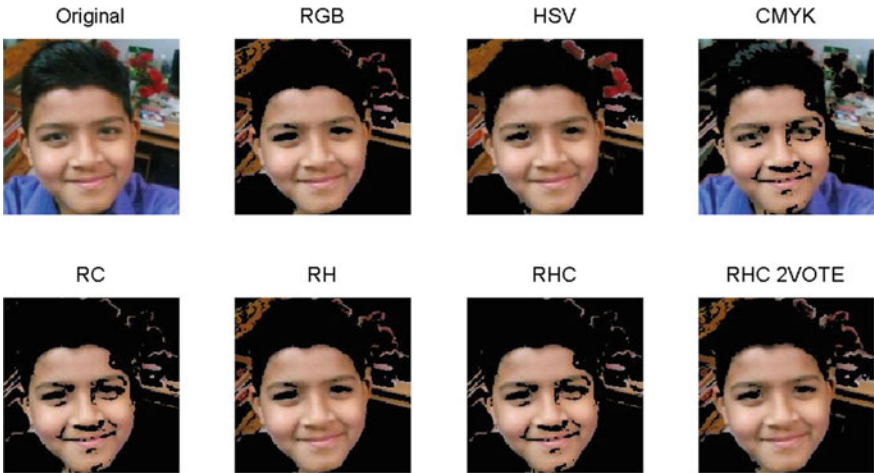


Fig. 2 Skin detection using single and hybrid colour schemes for sample 2

depicted in Figs. 1 and 2. Table 1 presents the *Recall* values of 10 random sample images chosen from ECU skin database and the average *Recall* value. The RGB-H-CMYK method outperforms Rahman’s [10] results with respect to *Recall* which plays important role in skin colour based face detection. The *Recall* values of RGB-H-CMYK method are consistently higher than Rahman’s method. The average *Recall* value of the proposed method is 85%, whereas it is 53% for Rahman’s scheme which illustrates the merit of RGB-H-CMYK technique in skin detection.

**Table 1** Performance analysis of hybrid colour scheme in terms of *Recall value*

Image sample	Rahman RGB-H-CbCr	RGB-H-CMYK (proposed)
Image 01	0.69642	0.77683
Image 02	0.37060	0.83299
Image 03	0.42660	0.82079
Image 04	0.57328	0.92205
Image 05	0.61265	0.78579
Image 06	0.43469	0.84023
Image 07	0.50932	0.88141
Image 08	0.56323	0.93833
Image 09	0.59478	0.88396
Image 10	0.59666	0.85083
Average	0.537823	0.853321

## 5 Conclusion

This research article presents a new hybrid colour scheme system for skin detection, which is a combination of RGB, HSV and CMYK colour scheme. The input image in RGB is transformed into H component of HSV as well as into CMYK colour space. It is concluded that RGB-H-CMYK and RHC\_Vote are confirmed to produce accurate results for skin identification. Due to its merits, this skin detection technique finds applications in face identification, face detection and emotion detection.

## References

1. Kovac, J., Peer, P., Solina, F.: Human skin color clustering for face detection. In: EUROCON 2003. Computer as a Tool. The IEEE Region 8, pp. 144–148. IEEE (2003)
2. Ibrahim, N.B., Selim, M.M., Zayed, H.H.: A dynamic skin detector based on face skin tone color. In: Proceedings of the 8th International Conference on INFormatics and Systems (INFOS2012) (2012)
3. Sobottka, K., Pitas, I.: A novel method for automatic face segmentation, face feature extraction and tracking. *Signal Process. Image Commun.* **12**(3), 263–281 (1998)
4. Mofaddel, M.A., Sadek, S.: Adult image content filtering: a statistical method based on multi-color skin modeling. In: 2nd International Conference on Computer Technology and Development (ICCTD), pp. 682–686 (2010)
5. Cheddad, A., Condell, J., Curran, K., Mc Kevitt, P.: A skin tone detection algorithm for an adaptive approach to steganography. *Int. J. Signal Process.* **89**, 2465–2478 (2009)
6. Jiann-Shu, L., Yung-Ming, K., Pau-choo, C.: The adult image identification based on online sampling. In: International Joint Conference on Neural Networks (IJCNN '06), pp. 2566–2571 (2006)
7. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: Proceedings of Graphicon, Moscow, Russia, pp. 85–92 (2003)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin color modeling and detection methods. *Pattern Recognit.* **40**, 1106–1122 (2007)
9. Al-Mohair, H.K., Mohamad-Saleh, J., Suandi, S.A.: Human skin color detection: a review on neural network perspective. *Int. J. Innov. Comput. Inf. Control (ICIC)* **8**(12), 8115–8131 (2012)

10. Rahman, N.A., Wei, K.C., See, J.: RGB-H-CbCr skin color model for human face detection. In: Proceedings of the MMU International Symposium on Information & Communications Technologies (2006)
11. Menser, B., Wien, M.: Segmentation and tracking of facial regions in color image sequences. In: Proceedings of the SPIE Visual Communications and Image Processing, pp. 731–740 (2000)
12. Lee, J.Y., Yoo, S.I.: An elliptical boundary model for skin color detection. Proc. Int. Conf. Imaging Sci. Syst. Technol. (2002)
13. Al-Mohair, H.K., Mohamad-Saleh, J., Suandi, S.A.: Hybrid human skin detection using neural network and K-means clustering technique. Appl. Soft Comput. **33**, 337–347 (2015)
14. Osman, M.Z., Maarof, M.A., Rohani, M.F.: Towards integrating statistical color features for human skin detection. In: 18th International Conference on Engineering and Applied Sciences (ICEAS), Kuala Lumpur, vol. 18, no. 2 IV, pp. 627–631 (2016)
15. Doukim, C.A., et al.: Combining neural networks for skin detection. Int. J. Signal Image Process. (SIPIJ) **1**(2), 1–11 (2011)
16. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recognit. **40**(3), 1106–1122 (2007)
17. Sigal, L., Sclaroff, S., and Athitsos, V.: Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. **2**, 152–159 (2000)
18. Mahmoud, T.M.: A new fast skin color detection technique. World Acad. Sci. Eng. Technol. **43**, 501–505 (2008)
19. Phung, S.L., Bouzerdoum, A., Chai, D.: A novel skin color model in YCbCr colour space and its application to human face detection. In: International Conference on Image Processing (ICIP'2002), vol. 1, pp. 289–292 (2002)
20. Hsu, C.L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 696–706 (2002)
21. Khan, R., Hanbury, A., Stottinger, J., Bias, A.: Color based skin classification. Pattern Recognit. Lett. **33**(2), 157–163 (2012)
22. Phung, S.L., Chai, D., -Bouzerdoum, A.: A novel skin color model in YCbCr color space and its application to human face detection. Proc. IEEE Int. Conf. Image Process. **1**, 1-289–1-292 (2002)
23. Dariusz, J.S., Weronika, M.: Human colour skin detection in CMYK colour space. IET Image Process. **9**(9), 751–757 (2015)
24. Tayal, Y., Lamba, R., Padhee, S.: Automatic face detection using color based segmentation. Int. J. Sci. Res. Publ. **2**(6), 1–7 (2012)
25. Atharifarid, A., Ghofrani, S.: Component-based face detection in color images. World Appl. Sci. J. **13**(4), 847–857 (2011)
26. Bin Ghazali, K.H., Ma, J., Xiao, R.: An innovative face detection based on skin color segmentation. Int. J. Comput. Appl. **34**(2), 6–10 (2011)
27. Anukrishnan, N., Ramya, B., Mohan, S.: Design and development of car ignition access control system based on face recognition technique. SAS TECH J. **9**(2), 63–70 (2010)
28. Samart, N., Chiechanwattana, S., Sunat, K.: A novel rule for face region detection based on RGB-HSV-YCbCr skin model. In: 3rd International Conference on Science and Technology for Sustainable Development of the Greater Mekong Sub-region, vol. 2, no. 1, pp. 330–337 (2011)
29. Chaves-González, J.M., Vega-Rodríguez, M., Gómez-Pulido, J., Sánchez-Pérez, J.M.: Detecting skin in face recognition systems: a colour spaces study. Digit. Signal Process. **20**(3), 806–823 (2010)
30. Frode, E.S., Levent, N., Yo-Ping, H.: Simple and practical skin detection with static RGB Color lookup tables: a visualization-based study. IEEE Int. Conf. Syst. Man Cybern. SMC 2371–2375 (2016)
31. Kawulok, M., Kawulok, J., Nalepa, J., Smolka, B.: Self-adaptive algorithm for segmenting skin regions. EURASIP J. Adv. Signal Process. (170) (2014)

# GPU-Based Approach for Human Action Recognition in Video



Ishita Dutta, Vikas Tripathi, Vaishali Dabral and Pooja Sharma

**Abstract** The power of graphic processing units (GPUs) can be harnessed to obtain an appreciable increase in computing performances by parallelizing various techniques. In view of this, the paper compares the performance of various descriptive statistical techniques like mean, variance and standard deviation on GPU for centroid calculation. Taking after this, the most productive procedure from the said methods has been contrasted with centroid calculation using k-means, processed on CPU. An appreciable increase in accuracy was achieved when we processed the above-mentioned techniques on GPU for centroid calculation in comparison with centroid calculation processed on CPU using k-means technique. The HMDB-51 dataset is used for computations. The aim of the paper is to find the most efficient and accurate approach for centroid and distance calculation in clustering. We attain an accuracy enhancement of 6.58% on comparing centroid calculation using variance method on GPU to centroid calculation using k-means on CPU.

**Keywords** GPU · Statistical techniques · HMDB-51

---

I. Dutta (✉)

Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India  
e-mail: ishitadutta69@gmail.com

V. Tripathi · P. Sharma

Graphic Era University, Dehradun 248002, Uttarakhand, India  
e-mail: vikastripathi.be@gmail.com

P. Sharma

e-mail: poojasharma3829@gmail.com

V. Dabral

Indraprastha Institute of Information Technology, Delhi, New Delhi, India  
e-mail: vaishalidabral@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_8](https://doi.org/10.1007/978-981-13-1708-8_8)



## 1 Introduction

Parallel processing of data dramatically improves the computing performance by faster delivery of results compared to sequential processing and considering the bulk of data that needs to be processed when handling computer vision problems such as object detection, recognition or distinction, parallel processing succor to provide faster solutions. GPUs have many cores which are helpful in operating pictures and graphics faster than CPU. Graphics processing units (GPUs) are high in demand as graphics application. CPU executes things sequentially, but GPU executes things parallel, hence they are more efficient for image processing. GPUs typically handle the computation for computer graphics, and traditional computation is handled by CPU. This thereby decreases the high load on CPU and increases the performance by reducing the time taken to execute large dataset. In this paper, we have evaluated the results for centroid calculation obtained through various techniques like mean, variance and standard deviation processed on GPU while also comparing the most efficient technique with the previous computations performed on CPU using k-means technique for centroid calculation. This paper's purpose was to propose an effective approach for calculating centroid and distance through parallel processing using GPU for clustering the data. Clustering helps in creating a sphere for similar kind of data and is useful in discovering knowledge from data. Our algorithm being parallelized on CUDA helps obtain a massive speedup in computation.

## 2 Literature Review

Detecting action in videos has grabbed major attention in computer vision [1, 2], it has many useful applications like in video surveillance, health care and human computer. Due to the increasing interest in Multimedia content, it is very important to improve the time required to categorize action from the large dataset present presently [3] and to improve the algorithm. The major challenge to the large set of data is that it becomes difficult to process large dataset for single processor, at once. But by the recent enhancements in parallel computing we can have a scalable and high-performance solution, by implementing parallel clustering algorithm through GPU [4]. GPU has multithreaded structure in multicore environment [5] and is very affordable platform for parallel computing, which has demonstrated that parallel computing algorithm can produce huge benefit to the present scenario of video classification. Cheap cost of a few thousand rupees multimedia content is way more expressive than other form of data present, so it is very important to categorize or classify the data. In this paper, we have used k-means clustering to classify. Dhillon and Modha have presented a parallel k-means clustering algorithm [6]. Clustering algorithm's efficiency can be improved by increasing the number of clusters. For large datasets, the use of adaptive can be done in order to achieve efficient computation of clusters [7, 8]. There are many other papers who have demonstrated scalability of parallel k-means

algorithm [9, 10]. Stoffel and Belkoniene have shown a linear speedup for large set of data [11]. Our motive in this paper is to parallelize the clustering techniques and to reduce overall time on various calculations by the big set of data. We have calculated GPU centroid and distance matrix through parallel processing so as to decrease the overall time taken by clustering. For any given dataset, computing distance matrix helps to determine the prime match of the cluster in the dataset [5]. The dataset was evaluated on CUDA [12]; for classification random forest has been used [13]. Tripathi et al. [14] has presented a framework that has a strong security framework system for ATM both using MHI and HU.

### 3 Methodology

Our model exploits GPU in two ways, firstly for centroid calculations and then for distance matrix computation. The basic motive was to decrease the computational time while further increasing the accuracy using certain modifications. Minimization in computational time was attained by processing the computer intensive code of centroid calculation in GPU. Whereas in order to obtain an increased accuracy in centroid calculation, three methods, namely: centroid calculation using mean, centroid calculation using variance and centroid calculation using standard deviation, described in Sect. 3.1, have been employed. Further, the distance matrix calculations, described in Sect. 3.2, have also been performed on GPU. We have compared the time and accuracy given by each combination of centroid and distance calculation methods, and the best results were given by centroid calculation using variance.

The methodology has been divided into two phases:

- Centroid calculation.
- Distance matrix calculation.

#### 3.1 Centroid Calculation Using GPU

It is required to work efficiently when handling large datasets like HMDB-51; in addition to this centroid calculation for this dataset is a time exhausting code when run on CPU although it produces good results. Therefore, if calculations for centroid are done parallel by dividing the dataset into groups, the computational time can be decreased substantially. This can be achieved by dividing the dataset into clusters and sending those clusters for parallel execution in GPU as native processor executes sequentially. Our algorithm calculates the results parallelly by using the concept of threads.

We have used different techniques to calculate centroid so as to improve overall accuracy like mean, variance and standard deviation. Algorithm 1, Algorithm 2 and Algorithm 3 show the working of mean method, variance method and standard deviation method, respectively.

In our Algorithm 1, we divided dataset into certain groups called clusters and for each cluster centroid was calculated by taking the mean of the values belonging to that cluster. Since the value of each cluster is independent of the other, so they can be calculated parallelly in GPU using threads. For each cluster, mean is calculated by adding the number of elements belonging to that cluster from each column and then taking its average to get the centroid. The formula for mean is shown in Eq. 1

$$\bar{X} = \frac{\sum_{m=0}^{range} X}{range}. \quad (1)$$

---

Algorithm 1

Centroid Calculation using Mean

Input: HMDB-51 dataset

Output: centroid matrix

---

```
t=i=m=0
i=threadIdx.x
clusters=499
limit=num_rows/clusters
LOOP k till clusters:
    Sum_temp=0
    z=0
    LOOP z till limit and m<num_rows:
        Sum_temp+=data_train[m][i]
        z++
    m++
END
centroid[k][i]=Sum_temp/limit
END
```

---

In Algorithm 2 instead of taking mean as centroid, the mean values obtained from Algorithm 1 for each cluster have been taken for variance calculation. Like mean the variance of each cluster is independent of the other, and hence can be calculated parallelly using threads in GPU. The variance of each cluster is calculated

by subtracting each element from the mean of that cluster and then squaring the result. Sum of all the values obtained so far is taken and then divided by one less than the total number of elements in that cluster. The formula for variance is shown in Eq. (2).

$$Vari = sd^2 = \sum_{m=0}^{range} (x - \bar{x})^2. \quad (2)$$

---

Algorithm 2

Centroid Calculation using Variance

Input: HMDB-51 dataset, mean matrix

Output: centroid matrix

---

```

k=m=0
i=threadIdx.x
clusters=499
limit=num_rows/clusters
LOOP if k<clusters:
    sum=0;
    t=0
    LOOP if t<limit&& m<row_size temp=mean[k][i]-
        data_train[m][i]
        sum+=temp*temp
        t++
        m++
    END
    centroid_matrix[k][i]=sum/(limit-1)
END
END

```

---

In Algorithm 3, we have shown the working of standard deviation. The standard deviation is the square root of variance. Equation (2) shows the formula for calculating standard deviation. Firstly, mean of each cluster is subtracted from each component of the corresponding column, and then, the square of resulting values is summated. The final value acquired is divided by one less than number of elements in each cluster. The square root of the value obtained is the final value of centroid. Since every standard deviation value is independent of each other, so they are computed independently.

---

 Algorithm 3

## Centroid Calculation using Standard Deviation

Input: HMDB-51 dataset, mean matrix

Output: centroid matrix

---

```

k=m=0
i=threadIdx.x
clusters=499
limit=num_rows/clusters
LOOP if k<clusters:
    sum=0;
    t=0
    LOOP if t<limit&& m<row_size temp=mean[k][i]-
        data_train[m][i] sum+=temp*temp

        t++
        m++
    END centroid_matrix[k][i]=sqrt(sum/(limit-
    1))
END
  
```

---

### 3.2 Distance Matrix Calculation Using GPU

Once computations for centroid calculation complete, distance matrix calculation using Algorithm 4 starts next. Distance is calculated using Euclidean distance formula. For distance calculation, an element from a column in the centroid matrix is taken one at a time and every remaining element in the corresponding column is subtracted from the selected element. Following this, the summation of square of each difference is computed and the square root of the final value obtained gives the final distance of the element from its nearest centroid.

---

**Algorithm 4**

---

Distance Calculation using GPU

Input: Centroid Matrix, Train File, Test

File Output: Train file and Test file

---

```

i=k=0
f=130
column=499
in= blockIdx.x*blockDim.x+threadIdx.x
  if index<threads:
LOOP till k<f:
    Sum_tem= 0.0
    LOOP till i<column: temp=train[in][k]-
                        centroid[i][k]
                        Sum_temp=temp*temp+Sum_temp
    END
    Distance_train[in][k]= Sum_temp
  END
END
END

```

---

Since centroid calculation and distance matrix calculation in CPU lead to huge amount computation, we used GPU for such computations. Algorithm 1, Algorithm 2 and Algorithm 3 described in Sect. 3.1 were processed on GPU for centroid matrix calculation. After centroid matrix calculation, centroid matrix along with train file is sent for distance calculation using Algorithm 3 which is described in Sect. 3.2. After clustering, the resultant train and test files were sent for classification using random forest.

## 4 Result and Discussion

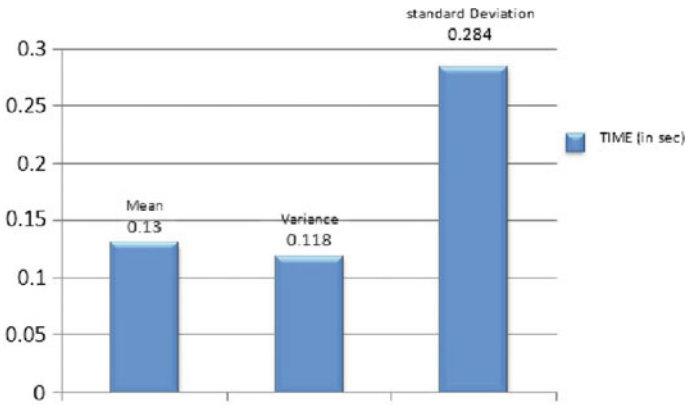
In this paper, we have used NVIDIA GeForce 610 M model and the system that we have used is Intel® core™ i3-2350 M CPU @ 2.30 GHz with 48 CUDA cores and 2048 MB of video memory. Our main objective was to present the runtime comparison between CPU and GPU for the same algorithm and also to improve the existing algorithm of clustering. For clustering, we have calculated centroid and distance matrix calculation method which is executed in GPU. The results were calculated on HMDB dataset. We have achieved overall accuracy of 53.047% of centroid calculation through variance.

Following Table 1 evaluates the performance of various techniques of centroid calculation, where variance method transfers a maximum accuracy of 53.047% and mean method transfers the least accuracy of 48.365%.

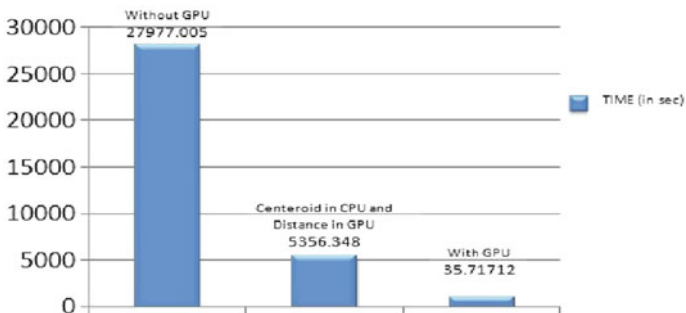
Figure 1 shows the time taken by the various techniques. A noticeable decrease in the processing time with variance method of centroid calculation can be observed in the graph, with processing time for mean method being 0.13 s as compared to processing time for variance method being 0.118 s.

**Table 1** Accuracy report

Parameter	Performance (%)		
	Mean	Variance	Standard deviation
Mean precision	52.87	56.32	51.49
Mean recall	48.16	52.53	46.90
Mean F1	48.12	52.15	47.00
Accuracy	48.365	53.047	46.960



**Fig. 1** Time comparison between different ways of calculation centroid



**Fig. 2** Processing time by various approaches

Figure 2 contains a bar graph indicating the total time taken by different approaches. The graph clearly reveals the computational time drastically drops to 11.852 s when the computations are done on GPU. The variance method is used for centroid calculation.

Table 2 presents an analysis of efficiencies of the techniques when run on CPU and GPU, respectively, where the significant rise in accuracy is apparent when centroid calculation and the distance calculations are done on GPU.

**Table 2** Comparison of accuracies in various approaches

Parameter	Performance (%)		
	Centroid using k-means in CPU and distance calculation in CPU	Centroid in GPU using mean method and distance calculation in GPU	Centroid in GPU using variance method and distance calculation in GPU
Mean precision	43.62	52.87	56.32
Mean recall	40.32	48.16	52.53
Mean F1	40.29	48.12	52.15
Accuracy	42.087	48.36	53.047

## 5 Conclusion

We have analyzed different methods like mean, variance and standard deviation for centroid calculation to decide the most proficient and exact technique. Initially, we compared the computational time of processing done on CPU to that of GPU, where a dramatic drop of 27985.157 s was accomplished. Following this, we compared the performance of various descriptive statistical techniques on GPU to find the quickest and most precise approach for centroid calculation from which we discovered that variance method for centroid calculation gives substantially better results compared to the other techniques, with variance method achieving a dramatic increase of 4.7% in accuracy. Results achieved by our framework conclusively demonstrate that it can be used to benefit several real-time applications like video surveillance. Our framework is open for advancement which will improve the viability of handling.

## References

1. Moeslund, T.: Summaries of 107 computer vision-based human motion capture papers Technical Report LIA 99-01, University of Aalborg (1999)
2. Pang, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition; comprehensive study and good practice. *Comput. Vis. Image Underst.* 109–125 (2016)
3. Ganti, V., Gehrke, J., Ramakrishnan, R.: Mining very large databases. *Computer* **32**, 38–45 (1999)
4. Judd, D., McKinley, P., Jain, A.: Large-scale parallel data clustering. In: *Proceedings of the International Conference on Pattern Recognition*, pp. 488–493 (1996)
5. Yadav, K., Mittal, A., Ansari, M.A., Vishwarup, V.: Parallel implementation of similarity measures on GPU architecture using CUDA. *Indian J. Comput. Sci. Eng. (IJCSE)*
6. Dhillon, I.S., Modha, D.S.: A data clustering algorithm on distributed memory multiprocessors. Large-scale parallel data mining. *Lect. Notes Artif. Intell.* **1759**, 245–260 (2000)
7. Goil, S., Nagesh, H., Chaudhary, A.: MAFIA: efficient and scalable subspace clustering for very large datasets. Technical Report No. CPDC-TR-9906-010 (1999)
8. Ester, M., Kriegel, H.P., Sanders, J., Xu, X.: A density based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference in Knowledge Discovery in Databases and Datamining* (1996)



9. Nagesh, H., Goil, S., Choudhary, A.: A scalable parallel subspace clustering algorithm for massive data sets. In: Proceedings International Conference on Parallel Processing, pp. 477–484. IEEE Computer Society (2000)
10. Ng, M.K., Zhaxue, H.: A parallel k-prototypes algorithm for clustering large data sets in data mining. *Intell. Data Eng. Learn.* **3**, 263–290 (1999)
11. Kilian, S., Belkoniene, A.: Parallel k/h-means clustering for large data sets. In: Euro-Par'99 Parallel Processing, pp. 1451–1454. Springer, Berlin, Heidelberg (1999)
12. Alsmirat, M.A., Jararweh, Y., Al-Ayyoub, M., Shehab, M.A., Gupta, B.B.: Accelerating compute intensive medical imaging segmentation algorithms using hybrid CPU-GPU implementations. *Multimed. Tools Appl.* **76**, 3537–3555 (2017)
13. Dabral, V., Tripathi, V., Khan, K.: Real time computation of clustering and distance matrix through GPU. *Int. J. Control Appl.* **9**, 583–590 (2017)
14. Tripathi, V., Gangodkar, D., Latta, V., Mittal, A.: Robust abnormal event recognition via motion and shape analysis at ATM installations. *J. Electr. Comput. Eng.* (2015)

**Part II**  
**Machine Learning and Data Mining**

# Protein Sequence in Classifying Dengue Serotypes



Pandiselvam Pandiyarajan and Kathirvalavakumar Thangairulappan

**Abstract** Dengue is the growing disease. It serves, especially in children. Different diagnosing methods like ELISA, Platelia, haemaocytometer, RT-PCR, decision tree algorithms and recommender system with fuzzy logic are used to diagnose the dengue by blood specimen. But these methods identify severe cases after five to ten days of the person infected by dengue. Some other methods require saliva and urine samples instead of blood specimen when a volume of blood samples cannot be obtained from person, especially from children. But from this sample, the correct result could not be identified. To overcome these problems, this paper proposes dengue diagnosis method based on amino acids or components in the protein sequence as it needs only skin cells or hair or nail which can be collected easily from the patients. The proposed method not only diagnoses the dengue but also identifies serotypes using statistical analysis of protein sequence. The experimental results prove that the proposed method identifies dengue and its serotypes correctly by amino acids and components of protein sequences. The proposed method is capable of finding deficiency or dominance of amino acids or components in the dengue-infected protein sequence by assessing entropy, relative and weighted average values of amino acids or components.

**Keywords** Dengue serotypes · Protein sequence · Diagnosing methods  
Protein classification

---

P. Pandiyarajan (✉)

Department of Computer Science, Ayya Nadar Janaki Ammal College,  
Sivakasi 626124, Tamil Nadu, India  
e-mail: pandiselvam.pps@gmail.com

K. Thangairulappan

Research Centre in Computer Science, V.H.N. Senthikumara Nadar College,  
Virudhunagar 626001, Tamil Nadu, India  
e-mail: kathirvalavakumar@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_9](https://doi.org/10.1007/978-981-13-1708-8_9)

## 1 Introduction

Dengue is a man-killing disease transmitted by *Aedes aegypti* mosquito in several regions, and it also spreads some other viral infections such as Chikungunya, yellow fever and Zika infection. The disease is spread through the tropics. The second highest state in dengue outbreak 2017 is Tamil Nadu, India. As per the state government health department report, 4400 dengue cases are recorded [1]. There are four distinct, but closely related serotypes, namely DENV I, DENV II, DENV III and DENV IV. Recovery from disease by one creates permanent immunity against that specific serotype. Succeeding infections by other serotypes enhance the risk of increasing severe dengue. The burden of dengue in the world is to classify dengue cases. Some existing methods are misclassified the dengue cases. The diagnosing of particular serotypes is a crucial task in a medical field. The paper suggests a useful method for so.

## 2 Literature Review

Guzman et al. [2] have assessed the performances of dengue diagnosis methods ELISA and Platelia. These methods detected the dengue virus protein NS1 (non-structural protein 1) in plasma/serum of the patients. Platelia method is more sensitive than ELISA method. In Platelia, NS1 or IgM is tested on the dengue-infected patients. The combination of NS1 and IgM detection increased a higher sensitivity of dengue diagnosis in collected blood samples. Tanner et al. [3] have proposed a dengue diagnosis method using decision tree algorithms with the parameters including platelet count, IgM and IgG antigen count and crossover threshold values of dengue patients. C 4.5 decision tree classifier has been used to classify the dengue from non-dengue fever. As the results of a classifier, blood samples were classified into three classes. Dengue hemorrhagic fever (DHF) cases were classified correctly.

Fried et al. [4] have found that the secondary diseases of dengue are strongly associated with more severe grades of DHF (DHF I, DHF II ...), and DENV II is highly associated with more severe secondary diseases of dengue. Singh et al. [5] have proposed recommender system for detection of dengue using fuzzy logic. They have developed an android application for detection of dengue using the factors such as fever, blood pressure, joint pain, skin rashes, pain behind the eyes, severe headache. This system analyzed these factors used to find whether the fever is dengue or not.

Andries et al. [6] have used the different diagnostic methods (RT-PCR, NS1 antigen and antibody detection DENV IgM/IgA ELISA) applied on saliva and urine. These methods are useful for the young children when the blood samples cannot be easily obtained. Grande et al. [7] have measured the quality of dengue diagnosis with antibody response by ELISA. Vongsouvath et al. [8] have evaluated the different diagnostic methods of dengue. Four serotypes were isolated and quantified by RT-PCR. Greater accuracy is obtained from RT-PCR than ELISA test. Prakash

et al. [9] have analyzed various dengue diagnosing methods such as RT-PCR, NS1, nucleic acid amplification, serological diagnosis and biosensor. They concluded that existing methods do not work well when having a low-level presence of IgM antibody. Existing methods are relying only on requirements of blood samples from the patients.

The classification and pattern recognition techniques of data mining can be used for diagnosing arbovirus dengue [10] and classifying the patient record data of dengue [11]. Rough set theory is also used for generating classification rules of dengue [12]. Arunkumar et al. [13] have proposed a dengue disease prediction system using decision tree and support vector machine (SVM). The decision tree is generated using fisher filtering method. SVM is applied to the decision tree for obtaining better classification result.

Pabbi [14] have provided the fuzzy rules for classifying dengue into three classes DF, DHF and dengue shock syndrome (DSS) by using the factors age, TLC, SGOT/SGPT, platelets count and BP. Fatima and Pasha [15] have proposed a method for classifying different dengue serotypes. Differences between dengue serotypes are identified using SVM classifier. Shaukat et al. [16] have analyzed the attack of dengue fever in different areas of Jhelum in Pakistan using k-means, k-medoids, DB scan and optics clustering algorithms.

The system designed by [17] has proposed three artificial neural network models for diagnosing and identifying the dengue-infected patient's data from Jalpaiguri Sadar hospital, North Bengal, India. The warning system of dengue made to predict the future outbreaks in Jember [18] based on risk factors. ANN-based dengue diagnosing system [19] used for identifying the severity of dengue virus in microscopic images of blood cells.

Dengue diagnosis based on moving of antibodies directed in blood against the virus. Existing methods need a volume of the blood specimen. These methods were not suitable when the patient was a child. The proposed method uses components/amino acids which obtained from skin cells, hair and nail. Any type of viral infection spreads by encoding specific amino acids in the protein sequence. The amino acids in the protein sequence may be either dominant or deficient when the person is infected with any type of diseases. The dominant and deficient of particular amino acid is also varying from one disease to another. In the proposed method, dengue is diagnosed by finding the dominant and deficient of amino acids/components of a protein sequence using entropy, relative and weighted averages.

### 3 Materials and Method

#### 3.1 Bio-sequences

Gene is a vehicle of genetic information which is used to decide the characteristics (eye color, hair color) of a human. The protein sequence is an organic component composed of amino acids, and this sequence of every person is conflicting from another person with only 0.5%.

#### 3.2 Amino Acids and Components

Amino acids are the building blocks of a protein sequence. They are classified into acidic, basic and neutral components based on the amino group and carboxylic group. Neutral components are classified into four subcomponents: aliphatic, aromatic, heterocyclic and sulfur. Deficiency or dominance of amino acids/components has led to disease. This leads to propose a method using the components for identifying the serotypes in dengue. This work classifies the protein sequence into five components such as sulfur, neutral, aliphatic, acidic and aromatic.

#### 3.3 Procedure

Collect skin cells from the patient and convert into DNA with nucleotides. Nucleotides are converted into a protein sequence. Read the protein sequence. Count each amino acid and component in the sequence. Dengue protein sequence may be of existing serotypes. For identifying the dengue serotypes, this system has to be followed.

##### **Entropy.**

Entropy is the quantity of probability of information. Calculate entropy for each amino acid and components in the protein sequence using Eq. (1) or Eq. (2).

$$H = - \int P(X) \ln P(X) dx \quad (1)$$

where  $P(X)$  is the probability of amino acids or components in the protein sequence.

$$\text{Entropy} = - \sum_{i=1}^m p_i \log p_i \quad (2)$$

where  $p_i$  is the probability of count of amino acids or components and  $m$  represents quantity of amino acids ( $=20$ ) or components ( $=5$ ). Entropy values extend from 0 to 1.

$$H_n(P_1, \dots, P_n) \leq H_n\left(\frac{1}{n} \dots \frac{1}{n}\right) = \log_b(n) \quad (3)$$

where  $H_n$  is an entropy value for all probabilities and  $1/n$  is the probabilities of information. If the particular amino acid is equally distributed, that amino acid gets a maximum entropy value which is specified in Eq. (3)

### Relative values.

The relative value of amino acid is the deviation of entropy values of infected person from the minimum entropy of normal person. The relative value of components is the deviation of the weighted average of an infected person from the weighted average of components of a normal person. Calculate relative values for every amino acid and component using Eq. (4). The relative value is defined as:

$$R = \frac{1}{n} \sum_{i=1}^{i=n} \frac{\text{Actual value} - \text{Expected value}}{\text{Expected value}} \quad (4)$$

where  $R$  is the relative value of amino acids or components. When an amino acid is considered, the actual value represents the entropy value of an infected person; expected value represents the minimum entropy value of normal person. When a component is considered, the actual value represents the weighted average of an infected person; expected value represents the weighted normal person.  $R$  value of any disease is unique for any patient as the dominant and deficiency of the amino acid and components is unique for the disease.

### Weighted average.

The weighted average of a component is calculated using Eq. (5).

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (5)$$

where  $w_i$  denotes the entropy of components of diseased person and  $x_i$  denotes the entropy of normal person's components.

### Identification of serotypes.

Dengue served can be identified using amino acids and can be classified. Select amino acids with negative relative values. If the person is infected by DENV I, then phenylalanine (F) and tryptophan (W) are with negative values; if the person is infected by DENV II, then phenylalanine (F), leucine (K), valine (V) and tryptophan (W) are with negative values; if the person is infected by DENV III, then phenylalanine (F) is with negative value. If the person is infected with DENV IV, then phenylalanine (F), glycine (G), leucine (K), valine (V), and tryptophan (W) are

with negative values. From the above conditions, we can identify that the person is infected with dengue if an amino acid, *phenylalanine (F)*, is with a negative value.

Dengue served can also be identified using components and can be classified. Calculate the weighted average for each component of normal human and diseased person using Eq. (5). Calculate the relative values for identifying deviation of the weighted average of an infected person from the weighted average of normal person's components using Eq. (4). Classify the components based on this R.

## 4 Results and Discussion

Experimental results were carried out by dengue-infected protein sequence collected from the National Center for Biotechnology Information (NCBI) [20]. This center is a national resource for molecular biology information funded by US government.

The protein sequence of dengue patients is used in this method. Obtained entropy values for components and amino acids are listed and shown in Tables 1, 2, 3 and 4 and Figs. 1 and 2. In general, the protein sequence of every human is differing with only 0.5%. The proposed method identifies the difference in those percentages. The entropy and relative values of normal sequence and infected sequences are shown in Table 5.

**Table 1** Entropy values of amino acids

Amino acid	Entropy values				
	Normal	DENV I	DENV II	DENV III	DENV IV
Alanine (A)	0.241	0.222	0.079	0.25	0.0768
Cysteine (C)	0.18	0.194	0.0729	0.2	0.0713
Aspartic acid (D)	0.132	0.143	0.0534	0.17	0.0477
Glutamic acid(E)	0.138	0.154	0.0546	0.17	0.0533
Phenylalanine(F)	0.137	0.081	0.0285	0.09	0.0275
Glycine (G)	0.186	0.118	0.0406	0.14	0.0378
Histidine (H)	0.157	0.218	0.0818	0.25	0.0748
Isoleucine (I)	0.222	0.241	0.0912	0.29	0.088
Leucine (K)	0.168	0.095	0.03	0.1	0.0275
Lysine (L)	0.179	0.191	0.0777	0.23	0.0687
Methionine (M)	0.323	0.268	0.1016	0.31	0.0958
Asparagine (N)	0.132	0.196	0.0757	0.24	0.0667
Proline (P)	0.11	0.142	0.05	0.16	0.0497
Glutamine (Q)	0.169	0.119	0.0413	0.13	0.0402
Arginine (R)	0.265	0.149	0.0583	0.18	0.0516

(continued)



**Table 1** (continued)

<i>Amino acid</i>	<i>Entropy values</i>				
	Normal	DENV I	DENV II	DENV III	DENV IV
Serine (S)	0.335	0.198	0.0712	0.21	0.0703
Threonine (T)	0.333	0.238	0.0891	0.28	0.0873
Valine (V)	0.058	0.109	0.0369	0.13	0.0338
Tryptophan (W)	0.149	0.084	0.0303	0.11	0.0298
Tyrosine (Y)	0.215	0.206	0.0768	0.24	0.0748

The components of a protein sequence are the combination of amino acids. The relative values for components aromatic (R), acidic (C), neutral (N), sulfur (S) and aliphatic (L) are calculated by weighted average of the protein sequence **0.20664** for a normal human. Based on the relative values, the protein sequence is classified as normal, DENV I, DENV II, DENV III and DENV IV. Table 6 reveals that if

**Table 2** Relative values of amino acids

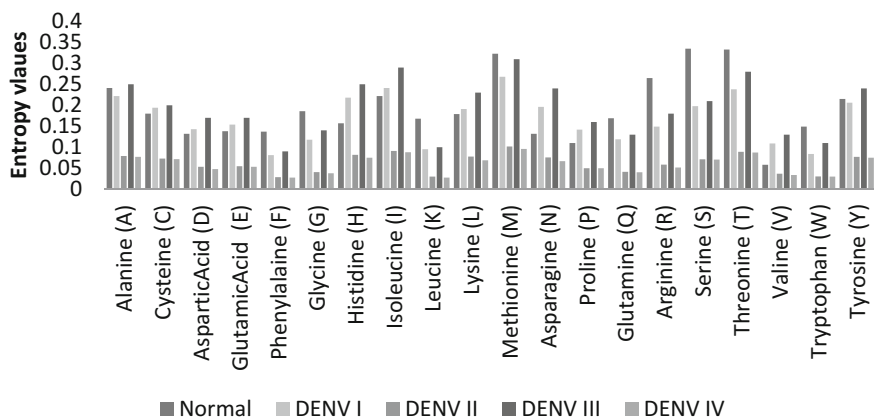
<i>Amino acid</i>	<i>Relative values</i>			
	DENV I	DENV II	DENV III	DENV IV
Alanine (A)	1.55632184	0.975	1.929885057	0.92
Cysteine (C)	1.23103448	0.8225	1.316091954	0.7825
Aspartic acid(D)	0.64712644	0.335	0.934482759	0.1925
Glutamic acid(E)	0.77011494	0.365	0.912643678	0.3325
Phenylalanine(F)	<b>-0.066667</b>	<b>-0.2875</b>	<b>-0.0045977</b>	<b>-0.3125</b>
Glycine (G)	0.36091954	0.015	0.645977011	<b>-0.055</b>
Histidine (H)	1.50574713	1.045	1.824137931	0.87
Isoleucine (I)	1.76436782	1.28	2.301149425	1.2
Leucine (K)	0.09195402	<b>-0.25</b>	0.181609195	<b>-0.3125</b>
Lysine (L)	1.19195402	0.9425	1.595402299	0.7175
Methionine (M)	2.08390805	1.54	2.570114943	1.395
Asparagine (N)	1.24712644	0.8925	1.770114943	0.6675
Proline (P)	0.62988506	0.25	0.791954023	0.2425
Glutamine (Q)	0.37011494	0.0325	0.493103448	0.005
Arginine (R)	0.7091954	0.4575	1.103448276	0.29
Serine (S)	1.27011494	0.78	1.356321839	0.7575
Threonine (T)	1.72988506	1.2275	2.227586207	1.1825
Valine (V)	0.25402299	<b>-0.0775</b>	0.46896517	<b>-0.155</b>
Tryptophan (W)	<b>-0.0344828</b>	<b>-0.2425</b>	0.208045977	<b>-0.255</b>
Tyrosine (Y)	1.36321839	0.92	1.806896552	0.87

**Table 3** Entropy values of components

Component	Entropy values				
	Normal	DENV I	DENV II	DENV III	DENV IV
Aromatic	0.2906	0.2405	0.0891	0.2750	0.0854
Acidic	0.2470	0.3082	0.1193	0.3391	0.1121
Neutral	0.6457	0.5120	0.2181	0.5599	0.210
Sulfur	0.3802	0.3842	0.1563	0.4131	0.1503
Aliphatic	0.6315	0.6161	0.2984	0.6687	0.2807

**Table 4** Weighted average value of components

Component	DENV I	DENV II	DENV III	DENV IV
Aromatic	0.0698893	0.0258925	0.079915	0.0248172
Acidic	0.0761254	0.0294671	0.0837577	0.0276887
Neutral	0.3305984	0.1408272	0.3615274	0.136049
Sulfur	0.1460728	0.0594253	0.1570606	0.0571441
Aliphatic	0.3890672	0.1871766	0.4222841	0.1772621
Average	0.4609354	0.201726	0.5032095	0.192693
WA of normal	0.20664	0.20664	0.20664	0.20664
R value	0.2542954	0.004914	0.2965695	0.013947

**Fig. 1** Entropy values of amino acids

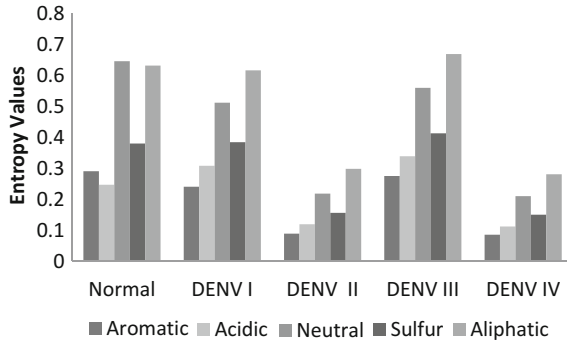


Fig. 2 Entropy values of components

Table 5 Relative values of dengue-infected and normal persons

Amino acids	Relative values							
	Patient I	Patient II	Patient III	Patient IV	Patient V	Patient VI	Patient VII	Patient VIII
Alanine (A)	1.55632184	0.241	0.92	1.55632184	1.55632184	1.55632184	0.975	1.929885057
Cysteine (C)	1.23103448	0.18	0.7825	1.23103448	1.23103448	1.23103448	0.822	1.316091954
Aspartic acid (D)	0.64712644	0.132	0.1925	0.64712644	0.64712644	0.64712644	0.335	0.934482759
Glutamic acid(E)	0.77011494	0.138	0.3325	0.77011494	0.77011494	0.77011494	0.365	0.912643678
Phenylalanine(F)	0.1066667	0.137	<b>-0.3125</b>	<b>-0.066667</b>	<b>-0.066667</b>	<b>-0.066667</b>	<b>-0.2875</b>	<b>-0.0045977</b>
Glycine (G)	0.36091954	0.186	<b>-0.055</b>	0.36091954	0.36091954	0.36091954	0.015	0.645977011
Histidine (H)	1.50574713	0.157	0.87	1.50574713	1.50574713	1.50574713	1.045	1.824137931
Isoleucine (I)	1.76436782	0.222	1.2	1.76436782	1.76436782	1.76436782	1.28	2.301149425
Leucine (K)	0.09195402	0.168	<b>-0.3125</b>	0.09195402	0.09195402	0.09195402	<b>-0.25</b>	0.181609195
Lysine (L)	1.19195402	0.179	0.7175	1.19195402	1.19195402	1.19195402	0.9425	1.595402299
Methionine (M)	2.08390805	0.323	1.395	2.08390805	2.08390805	2.08390805	1.54	2.570114943
Asparagine (N)	1.24712644	0.1322	0.6675	1.24712644	1.24712644	1.24712644	0.8925	1.770114943
Proline (P)	0.62988506	0.11	0.2425	0.62988506	0.62988506	0.62988506	0.25	0.791954023
Glutamine (Q)	0.37011494	0.169	0.005	0.37011494	0.37011494	0.37011494	0.0325	0.493103448
Arginine (R)	0.7091954	0.265	0.29	0.7091954	0.7091954	0.7091954	0.4575	1.103448276
Serine (S)	1.27011494	0.335	0.7575	1.27011494	1.27011494	1.27011494	0.78	1.356321839
Threonine (T)	1.72988506	0.333	1.1825	1.72988506	1.72988506	1.72988506	1.2275	2.227586207
Valine (V)	0.25402299	0.0548	<b>-0.155</b>	0.25402299	0.25402299	0.25402299	<b>-0.0775</b>	0.46896517
Tryptophan (W)	0.1344828	0.149	<b>-0.255</b>	<b>-0.0344828</b>	<b>-0.0344828</b>	<b>-0.0344828</b>	<b>-0.2425</b>	0.208045977
Tyrosine (Y)	1.36321839	0.215	0.87	1.36321839	1.36321839	1.36321839	0.92	1.806896552
Result of proposed system	<b>NOT DENGUE</b>	<b>NOT DENGUE</b>	<b>DENV IV</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV II</b>	<b>DENV III</b>
Target	<b>NOT DENGUE</b>	<b>NOT DENGUE</b>	<b>DENV IV</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV II</b>	<b>DENV III</b>

the relative value of components is **0.2542954**, then the patient is infected with DENV I; if the relative value of components is **0.004914**, then the patient is infected with DENV II; if the relative value of components is **0.2965695**, then the patient is infected with DENV III; if the relative value of components is **0.013947**, then the

**Table 6** Weighted averages of components of dengue-infected and normal persons

Component	Weighted averages							
	Patient I	Patient II	Patient III	Patient IV	Patient V	Patient VI	Patient VII	Patient VIII
Aromatic	0.0298893	0.0198893	0.024817	0.0698893	0.0698893	0.0698893	0.0258925	0.0248172
Acidic	0.00761254	0.00561254	0.027689	0.0761254	0.0761254	0.0761254	0.0294671	0.0276887
Neutral	0.1305984	0.1105984	0.136049	0.3305984	0.3305984	0.3305984	0.1408272	0.136049
Sulfur	0.4607284	0.4107284	0.057144	0.1460728	0.1460728	0.1460728	0.0594253	0.0571441
Aliphatic	0.28906715	0.22906715	0.177262	0.3890672	0.3890672	0.3890672	0.1871766	0.1772621
WA values	<b>0.18357916</b>	<b>0.15517916</b>	<b>0.192693</b>	<b>0.4609354</b>	<b>0.4609354</b>	<b>0.4609354</b>	<b>0.201726</b>	<b>0.5032095</b>
R values	<b>0.02306084</b>	<b>0.05126084</b>	<b>0.013947</b>	<b>0.2542954</b>	<b>0.2542954</b>	<b>0.2542954</b>	<b>0.004914</b>	<b>0.2965695</b>
Result	<b>NOT DENGUE</b>	<b>NOT DENGUE</b>	<b>DENV IV</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV II</b>	<b>DENV III</b>
Target	<b>NOT DENGUE</b>	<b>NOT DENGUE</b>	<b>DENV IV</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV I</b>	<b>DENV II</b>	<b>DENV III</b>

patient is infected with DENV IV; otherwise the patient is not infected with dengue. The proposed system is assessed by protein sequences of 8 patients. Among the 8 protein sequences, 5 sequences are infected with dengue and remaining 3 sequences are normal human sequences. From the observation in Table 5, it has been found that patient I and patient II are classified as not infected by dengue as their amino acid phenylalanine (F) is not a negative value; three patients, namely patients IV, V and VI, are classified as DENV I as phenylalanine (F) and tryptophan (W) are with negative values; patient VII is classified as DENV II as phenylalanine (F), leucine (K), valine (V) and tryptophan (W) are with negative values; patient VIII is classified as DENV III as phenylalanine (F) is with negative value and a patient III is classified as DENV IV as phenylalanine (F), glycine (G), leucine (K), valine (V) and tryptophan (W) are with negative values. From the observation in Table 6, it has been identified that three patients, namely patients IV, V and VI, are classified as DENV I as the relative value of component is 0.2542954; patient VII is classified as DENV II as the relative value of component is 0.004914; patient VIII is classified as DENV III as the relative value of component is 0.2965695; patient III is classified as DENV IV as the relative value of component is 0.013947; and two patients, namely patient I and patient II, are classified as the patients not infected with dengue as their relative values of the components are not any one of **0.2542954, 0.004914, 0.2965695 and 0.013947**. The obtained results are same as per the target of NCBI.

## 5 Conclusion

The entropy of protein sequences plays an important role in identifying dengue and its serotypes of the infected patients. Some existing methods cannot apply for dengue-infected young children as it needs a volume of blood. Some other methods particularly proposed for young children use urine and saliva when blood cells cannot be obtained, but its diagnosis is not perfect as in plasma. The proposed method does

not need urine, saliva or plasma, but it needs only protein sequences which obtained from any patients easily, and this method gives the correct result for identifying dengue virus and its serotypes in patients which are observed from the results of the experiment.

## References

1. India Times. [www.timesofindia.Indiatimes/city/Chennai/articleshow/59262649.cms](http://www.timesofindia.Indiatimes/city/Chennai/articleshow/59262649.cms)
2. Guzman, M.G., Jaenisch, T., Gaczkowski, R., Ty Hang, V.T., Sekara, S.D., Kroeger, A., Nazquez, S., Ruiz, D., Martinez, E., Mascado, J.C., Balmaseda, A., Harris, E., Dimano, E., Leano, P.-S.A., Villegas, E., Benduzu, H., Villalobos, I., Farrar, J., Simmon, C.D.: Multi-country evaluation of the sensitivity and specificity of two commercially available NS1 ELISA assays for dengue diagnosis. *PLoS Negl. Trop. Dis.* **8** (2010)
3. Tanner, L., Schreiber, M., Low, J.-G.H., Ong, A., Tolfvenstam, T., Lai, Y.L., Ching Ng, L., Leo, Y.S., Puong, L.T., Vasudevan, S.G., Simmons, C.P., Hibberd, M.L., Eong, E.: Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Negl. Trop. Dis.* **3** (2008)
4. Fried, R.J., Gibbons, V.R., Kalyanaraj, S., Thomas, S.J., Srikiachachorn, A., In-kyu, Y., Jarman, G.R., Green, S., Rothman, L.A., Cummings, A.-T.D.: Serotype-specific differences in the risk of dengue hemorrhagic fever: an analysis of data collected in Bangkok, Thailand from 1994 to 2006. *PLoS Negl. Trop. Dis.* **3** (2010)
5. Singh, S., Singh, A., Samson, Singh, M.: Recommender system for detection of dengue using fuzzy logic. *J. Comput. Eng. Technol.* **7**, 44–52 (2016)
6. Andries, A.C., Duong, V., Ly, S., Cappelle, J., Kim, K.S., Lorn Try, P., Ros, S., Ong, S., Huy, R., Horwood, P., Flamand, M., Sakuntaabhai, A., Tarantola, A., Buchy, P.: Value of routine dengue diagnostic tests in urine and saliva specimens. *PLoS Negl. Trop. Dis.* **9** (2015)
7. Grande, A.J., Reid, H., Thomas, E., Foster, C., Darton, T.C.: Tourniquet test for dengue diagnosis: systematic review and Meta-analysis of diagnostic test accuracy. *PLoS Negl. Trop. Dis.* **8** (2015)
8. Vongsouvath, M., Phommasone, K., Sengvilaipeaceuth, O., Kosoltanapiwat, N., Chantratita, N., Blacksell, S.D., Leesue, J., Lamballerie, X.D., Mayxay, M., Keomany, S., Newton, P.N., Dubotperes, A.: Using rapid diagnostic tests as a source of viral RNA for dengue serotype by RT-PCR—a novel epidemiological tool. *PLoS Negl. Trop. Dis.* **5** (2016)
9. Parkash, O., Shueb, R.H.: Diagnosis of dengue infection using conventional and biosensor based techniques. *Viruses* **7**, 5410–5427 (2016)
10. Fathima, S.A., Manimegalai, D., Hundewale, N.: A review of data mining classification techniques applied for diagnosis and prognosis of the arbovirus—dengue. *Int. J. Comput. Sci.* **6**, 322–328 (2011)
11. Tarle, B., Tajanpure, R., Jena, S.: Medical data classification using different optimization techniques: a survey. *Int. J. Res. Eng. Tech.* **5**, 101–108 (2016)
12. Mishra, S., Mohanty, P.S., Hota, R., Badajena, J.C.: Rough set approach for generation of classification rules for dengue. *Int. J. Comput. Appl.* **11**, 31–35 (2015)
13. Arunkumar, P.M., Chitradevi, B., Karthick, P., Ganesan, M., Madhan, A.S.: Dengue disease prediction using decision tree and support vector machine. *SSRG Int. J. Comput. Eng.* **1**, 60–63 (2017)
14. Pabbi, V.: Fuzzy expert system for medical diagnosis. *Int. J. Sci. Res.* **5**, 1–7 (2017)
15. Fatima, M., Pasha, M.: Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **9**, 1–16 (2017)
16. Shaikat, K., Masood, N., Shafaat, B.A., Jabbar, K., Shabbir, H., Shabbir, S.: Dengue fever in perspective of clustering algorithms. *Data Min. Genomics Proteomics* **6** (2015)

17. Saha, P., Mandal, R.: Detection of dengue disease using artificial neural networks. *Int. J. Comput. Eng.* **5**, 65–68 (2017)
18. Roziqin, C.M., Basuki, A., Harsono, T.: Parameters data distribution analysis for dengue fever breaks in Jember using Monte Carlo. *Int. J. Comput. Sci. Softw. Eng.* **5**, 45–48 (2016)
19. Subitha, N., Padmapriya, A.: Diagnosis for dengue fever using spatial data mining. *Int. J. Comput. Trends. Technol.* **4**, 2646–2651 (2013)
20. National Center for Biotechnology Information. [www.ncbi.nlm.nih.gov/genomes/virusvariation/database/nph-select.cgi](http://www.ncbi.nlm.nih.gov/genomes/virusvariation/database/nph-select.cgi)

# An Assistive Bot for Healthcare Using Deep Learning: Conversation-as-a-Service



Dhvani Shah and Thekkekara Joel Philip

**Abstract** Gone are the days when software was used only for complex mathematical calculations or graphical motions alone. Today, it is software that has exponentially grown to become more powerful and more human—most obviously in applications such as ‘Chatbots.’ The year 2017 marks the Chatbot revolution in various industries like health, career, insurance, customer care support. Artificial intelligence (AI), which is the key player in enabling human-like behavior intelligently, is dramatically changing business. Chatbots, fueled by AI, are becoming a viable option for human–machine interaction. Deep learning algorithms have made it possible to build intelligent machine. In this research, we have developed a HealthBot using TensorFlow and Natural Language Processing (NLP) techniques. There is no denying that efficient patient engagement is a key challenge for all healthcare organizations and any company that can unravel this challenge can effectively earn high returns of investments. Chatbots are one of the major overhauls that hospitals can easily provide more customized care for patients while cutting down on the waiting period. The proposed HealthBot lists the common symptoms; then, based on user’s health issue it gets deeper into the conversations predicting the health problem of the user. Such bots are needed for today’s fast-moving population where they have no time to keep a tab on their health. Neural network implementation adds more accuracy to the responses. The proposed Chatbot model is a retrieval-based bot and of closed domain. Finally, the HealthBot is deployed on the Flask, a Python web development framework.

**Keywords** Chatbot · Neural network · HealthBot · AI · Deep learning  
TensorFlow · Regression

---

D. Shah (✉)  
St. John College of Engineering and Management, Palghar, India  
e-mail: shahdhvani08@gmail.com

T. J. Philip  
Universal College of Engineering, Vasai-Virar, India  
e-mail: tjoelphilip@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_10](https://doi.org/10.1007/978-981-13-1708-8_10)

# 1 Introduction

It is a common tendency to have human interaction in almost all our daily activities; however, technology amplifies our abilities. Today, thanks to AI and NLP, we can use Chatbot technology to provide almost human-like conversations. The conversations are powered by AI (artificial intelligence). A good healthcare infrastructure is indeed essential for any nation's civic life, and it is vital that the sick be granted an indigenous provision to access better healthcare, without the need to wait for weeks or months just for a visit.

In majority of the below poverty line nations, the basic necessities to lead a healthy life are completely unknown due to various reasons ranging from ignorance to lack of information. It is near impractical to resolve these difficulties with a bot; however, a bot can certainly help to make the situation better. The real advantage of these Chatbots is the ability to provide proper guidance and information's for leading a healthy life, as there are many people who still lack the basic knowledge of proper healthcare. Many youngsters lack knowledge about safe-sex and have no awareness about disease transmitted sexually, because it is considered as a taboo in family.

It is well known that majority of the world population do not know the correct usage of basic drugs and antibiotics, which at a later stage leads to medical abuse and which indirectly renders the infused therapy more or less ineffective. These calculated issues can easily be solved with the help of the internet and the access to large chunks of medical resources—which are primarily free. These intelligent personal assistants (IPA) on our phones suddenly become definite responses for certain needs which are supported by machine learning and neural networks [1, 2].

Machine learning is the parent domain from which deep learning is derived, which when combined with algorithms of structure and functioning of the human brain paves way to artificial neural networks. Deep learning architecture consists of neural networks made up of neurons, activation functions and weights that learn on their own using learning algorithms [3].

Bots are nothing but intelligent agents residing on a server to communicate with humans or other bots to make human task much easier, without the need of any specific protocols or API's nor with any 'master bots' such as Google Assistant. They communicate in plain English, and deep learning makes them more accurate in throwing the appropriate response to the given query.

In this study, we have built a contextual Chatbot using TensorFlow and Python—to contribute in the health sector. Our bot is capable of diagnosing the health issue, suggesting the appropriate physician, giving reminders about prescription and also making an online appointment with the physician. The most important and primary benefit of Chatbots in healthcare domain is the supreme ability to provide advice and information for a healthy life to help those people who lack basic knowledge of healthcare.



## 2 Related Work

Conversational agents, or Chatbots, interact with the user in a human-understandable language. Their implementation has become increasingly sophisticated, and they are used in varied fields like education (e.g., [4, 10]), commerce (e.g., [5, 6]), entertainment (e.g., [7]) and the public sector (e.g., [8, 9]).

In [11], it has been proposed by the authors that neural responding machine (NRM), an innovative neural network-based response generator mechanism for short-text conversation. NRM enforces the general method of encoder–decoder’s effective framework which reinforces the generation of response as a decoding process based on the dormant representation of the input text, while both encoding and decoding are understood with recurrent neural networks (RNN). The authors have highlighted the drawbacks of retrieval-based bots and were able to achieve 75% accuracy with the proposed model.

OneStop Health is launched by Your.MD, which is a London-based healthcare organization which delivers assistance in the medical domain by the extravagant usage of artificial intelligence. It provides a bridging channel for clients via Chatbot which lucidly interprets the health symptoms and also provides them the best treatment. The design is as similar to clinic, with the only variation being that it is an online version of the same, which, after a free consultation with a doctor—who is connected remotely, can deliver prescriptions the following day. There is also Plus-Guidance—a unique online service focused at helping those suffering from mental health difficulties by offering a 24-hour video, voice of chat support service. It uses machine learning and natural language processing to learn from every conversation it has communicated. Using massive cloud-based servers, the application can crunch through every diagnosis, fine-tuning its technique and offering more subtle solutions. Finally, the natural language processing analyses how people speak, creating responses that feel human [12].

Chatbots are trending now, but they have their roots several years ago. As mentioned, Eliza was the first Chatbot created by MIT. If a patient said ‘my head hurts,’ Eliza would respond, ‘Why do you say your head hurts?’ Eliza, with just 200 lines of code, worked like a therapist [13].

Recently, Chatbots have gain acceleration in the health stream mainly due to technological progress in AI techniques. In the course of time, many bots have been developed to keep people healthy [14].

This research study caters to the meteoric rise in AI, which has been well projected in more than 100 startups to transform the healthcare industry. The main objective being to assist the users/patients to classify their symptoms into specific health issue based on their problems, to prescribe medicines to common and less severe problems like mild cold and cough, to suggest doctors’ name for consultation for serious problems like heart-related problems, to book an appointment with them and finally to remind them to take prescriptions on time. Also, at some time if our HealthBot does not have answers to certain queries, we have implemented a module to get the real-time details through scraping technology using BeautifulSoup in python.

### 3 Chatbot Architecture

The two types of Chatbots, based on their applications, are classified generally as: Chatbots for entertainment and Chatbots for business. The responses provided by the Chatbot's to a particular user's query should be smart enough for user to keep active on the Chatbot application expanding the conversation. It is not important for the Chatbot neither to understand what user is conveying nor to remember all the details of the conversation.

Another way to evaluate an entertainment bot is the Turing test which then can be used to compare the bot with a human. Other measurable metrics are mathematical calculation of the average length of time used for conversation between the bot and end users or else the average time spent by a user per week. If conversations are miniscule, then the bot may not be considered entertaining enough.

On the other hand, Chatbots for business are often transactional and they have a precise purpose. The conversation between the user and the bot is stereotypically focused on user's requirements. Travel Chatbots provides a brief data about tours, flights and hotels and helps to find the best available package according to user's norms. The infamous Google Assistant bot readily suggests information necessitated by the user instantly. Even the Uber makes use of a bot to take a ride request. Dialogues are usually short, spanning for less than 15 min. Each chat typically has a specific goal, and the quality of the bot can therefore be evaluated, as to how many users reach the goal.

#### 3.1 Models

Retrieval-based models use a predefined storehouse of responses and a unique empirical methodology to choose an appropriate response, based on the input and context. This heuristic methodology could be as simple as a rule-based expression match, or as intricate as an ensemble of machine learning classifiers. Such systems do not produce any new text, and they just choose a response from a static set. On the other hand, they too provide more expectable results. Due to the repository of predefined answers, retrieval-based methods do not make linguistic mistakes. However, they may be unable to handle new hidden cases for which no suitable predefined answer exists.

Generative models are 'cleverer' as compared to retrieval-based models. They can refer to entities in the input and give the impression that one is talking to a human. They are capable of formulating new responses from scratch based on the question asked from the user. These models are typically based on machine translation techniques, but instead of translating from one language to another, they 'translate' from an input to a smart response. It is problematic to, however, train these models, as grammatical mistakes are quite likely to happen (particularly on lengthier sentences) and typically require enormous amounts of training data.

Pattern-based heuristics—a viable methodology for choosing a reply—can be plotted in many different ways, from if-else conditional logic to machine learning classifiers. They follow the simple methodology based on a set of rules with patterns as conditions for the rules. This kind of models is very predominant for entertainment bots. Artificial Intelligence Markup Language (AIML), an XML dialect for creating natural language software agents—a widely used language for writing patterns and providing response prototypes used basically by bot developers. When the Chatbot receives a message, it iterates through all the patterns until finds a pattern which matches user query. Also, if the relevant match is not found, the Chatbot uses the equivalent template to generate a response intelligently.

### ***3.2 Implementation Steps/Methodology/Proposed Architecture***

As stated earlier, the proposed HealthBot uses deep learning and natural language processing techniques. This model is built on TensorFlow library which uses softmax as the activation function for the output layer and regression model as the learning algorithm. The function of softmax is significant in the field of machine learning as it can plot a vector to a probability of a given output in binary classification. Each neuron receives a vector of outputs from other neurons that fired each axon with its own weighting. These are then linearly combined and used in the softmax function to determine whether the next neuron fires or not. The python implementation of softmax function is shown below:

```
>>>import math
>>>y = [1.0, 2.0, 3.0, 4.0, 1.0, 2.0, 3.0]
>>>y_exp = [math.exp(i) for i in y]
>>>print([round(i, 2) for i in y_exp])
[2.72, 7.39, 20.09, 54.6, 2.72, 7.39, 20.09]
>>>sum_y_exp = sum(y_exp)
>>>print(round(sum_y_exp, 2))
114.98
>>>softmax = [round(i / sum_y_exp, 3) for i in y_exp]
>>>print(softmax)
[0.024, 0.064, 0.175, 0.475, 0.024, 0.064, 0.175]
```

The unique methodology that regression analysis reinforces of predictive modeling technique examines the relationship between a dependent (target) and independent variable (s) (predictor). It indicates the significant relationships between dependent variable and independent variable. It indicates the strength of impact of multiple independent variables on a dependent variable. Here, we have used 2-layer neural network and gradient descent algorithm to find the highest accuracy or to minimize the error rate, on the training data. All these are highly abstracted using tflearn, deep learning library, featuring a higher-level API for TensorFlow.

**Table 1** Input and output for spell corrector

Input	Output
Having sevre headage	Having severe headache
Having mild fevr	Having mild fear
Having mild fevrr	Having mild fever
Irrtation	Irritation
Transportibility	Transportability

### 3.2.1 Spelling Corrector

There may have been unique cases wherein the users enter wrong spellings for a particular word. It is a definite prerequisite that all the Chatbots should understand human level language and writing an incorrect spelling is part and parcel of human communication, e.g., the user may type, ‘having sever headage’ instead of ‘having severe headache.’ Based on the intensity of headache, our Chatbot will provide the necessary guidelines. To provide an accurate response, the intent classification should be precise and hence the spell check corrector needs to be implemented. This spell corrector which provides around 75% accuracy consist of repository which is a concatenation of public domain book excerpts from Project Gutenberg and lists of most frequent words from Wiktionary and the British National Corpus.

As illustrated in the Table 1, for second row, it was desired to type ‘fever’ but instead a erroneous spelling ‘fevr’ and the corrected word was given as ‘fear’ instead of ‘fever’. But when it was typed ‘fevrr’, the algorithm predicted correctly it as ‘fever’. This happened because ‘fevr’ being a 4-letter word, the algorithm predicted the possible 4-letter word, ‘fear,’ and when the input was a 5-letter word ‘fevrr’, it gave the expected word ‘fever’. This leads to the conclusion that it is important to understand the context too, i.e., ‘fear’ is a feeling and ‘fever’ is a symptom for certain health issue. Also, in the last row, the corrected word should have been ‘transportability’ instead of ‘transportibility’. These two concerns further motivate to implement a more accurate spell check algorithm in the near future and having bigger text file.

### 3.2.2 Natural Language Processing Techniques

This involves splitting the given text into sentences to analyze each sentence accurately and then further splitting the sentences into words. For this purpose, we have used the NLTK library—which is a leading platform for building Python programs to work with human language data [15]. Tokenization is an important concept which helps to break a sentence into respective multiple words. So, we have utilized tokenization and parts-of-speech (POS) tagging for understanding the human language precisely followed by lexicon normalization method named ‘stemming’ to remove textual noise caused by multiple representations exhibited by the same word.

Stemming helps the machine equate words like ‘have’ and ‘having.’ Also, it is imperative to understand each along with its synonyms. This helped us to classify our input and the intents correctly in order to choose the response because a user may either say ‘having severe headache’ or ‘having terrible headache’; for both these cases, our Chatbot should produce the same response. To achieve this purpose, we have used the WordNet library from the NLTK toolkit—which is a lexical database for English. Below is the small snippet of the same from Chatbot implementation. Alternatively, there is a library called WordAPI which achieves the same results [16].

```
import nltk
from nltk.corpus import wordnet
synonyms = []
for syn in wordnet.synsets("severe"):
    for l in syn.lemmas():
        synonyms.append(l.name())
print(set(synonyms))
>>>{'spartan', 'dangerous', 'grievous', 'knockout',
'austere', 'serious', 'grave', 'stern', 'life-
threatening', 'hard', 'severe', 'wicked', 'stark',
'terrible'}
```

After preprocessing is done, we have a precise yet clean list of sentences and lists of words inside each sentence. Each word is marked with a part of speech and concepts, and we have a lemma for every word. So, the next task is the creation and pattern classification of conversational intent which has a JSON-like structure followed by the building of deep neural network. The intents for the HealthBot have the following structure: a tag (a unique name), a pattern (sentence patterns for the neural network text classifier) and list responses (one will be used as a response).

Until now, we have created a list of sentences; then, each sentence is a list stemmed words and each sentence is associated with an intent (class). It is important now that we should format our data into a structure understandable by the TensorFlow library, i.e., we further need to transform it, which may have been derived from documents of words into tensors of numbers. At this point, we have implemented a bag of words function onto our training data, which is transformed for each sentence, i.e., each training sentence is reduced to an array of 0s and 1s against the array of unique words in the corpus (text classification technique). After all these processing steps on the training data, we shall be able to build the deep neural network.

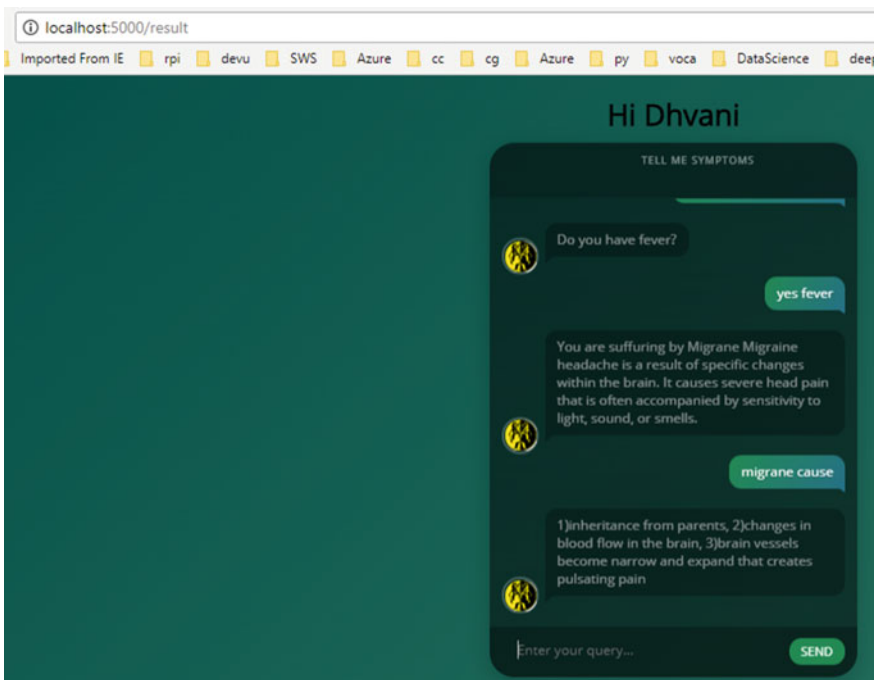
### 3.2.3 Building the Deep Neural Network

Below is the high-level implementation of NN model built using tflearn on TensorFlow.

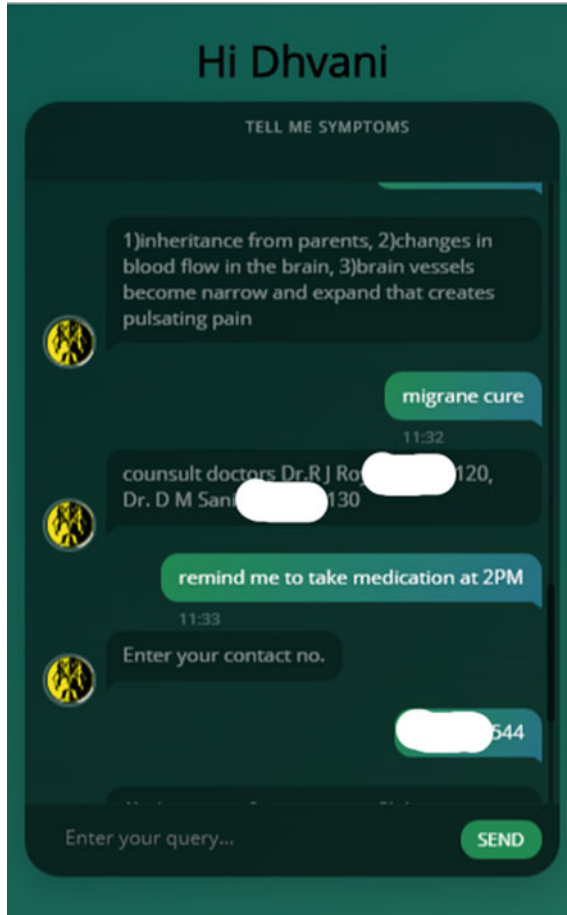
```
net = tflearn.input_data(shape=[None, len(train_x[0])])
net = tflearn.fully_connected(net, 8)
net = tflearn.fully_connected(net, 8)
net = tflearn.fully_connected(net, len(train_y[0]),
activation='softmax')
net = tflearn.regression(net)
model = tflearn.DNN(net, tensorboard_dir='tflearn_logs')
model.fit(train_x, train_y, n_epoch=2000, batch_size=8,
show_metric=True)
model.save('model.tflearn')
```

Thus, we have successfully implemented our HealthBot. To make our bot utilize the maximum computing power of conversation-as-a-service, we have deployed it on Flask web server. In the future, when this algorithm would be developed into a more efficient and accurate one, we plan to push our bot on a live server or public cloud-like Azure or AWS, so as to make the device independent, i.e., access the HealthBot anytime from anywhere. Below are the snapshots of our HealthBot deployed on Flask server (Figs. 1 and 2).

After the indication of a health problem, our bot has given certain causes for it and later gives him the list of doctors for the consultation.



**Fig. 1** (*HealthBot*) diagnosing migraine based on the given symptoms. Also, 'localhost://5000' is the indication that the HealthBot is deployed on flask web server



**Fig. 2** (*HealthBot*) giving list of doctors, their contact numbers and also the user is setting a reminder on bot about his medications. The bot is able to send the message to the user using Twilio python library at the given time

## 4 Conclusion

In this study, we have successfully implemented a robust HealthBot which is efficiently capable of effectively diagnosing the problem based on patient symptoms, give necessary precautions for the problem and if severe, can provide respective specialist in the field, i.e., doctors. The HealthBot is also capable to remind the user to take medications on the predefined time. For queries which are unknown, we have implemented a very simple scraping module using BeautifulSoup which has been wholly deployed on Flask web server.

In the future, we plan to implement generative models using deep learning architectures like sequence so as to sequence text prediction using recurrent neural networks. Further, more precise NLP techniques can be implemented for spell check and synonyms, to achieve better intent classification.

## References

1. Rannamagi, R.: Is AI evolved enough to build a life coach as a chatbot? (2017). <https://chatbotsmagazine.com/is-ai-evolved-enough-to-build-a-life-coach-as-a-chatbot-e41f3376c425>. Accessed 17 Aug 2017
2. Mesko Dr.: Chatbots will serve as health assistants (2017). <http://medicalfuturist.com/chatbots-health-assistants/>. Accessed 17 Aug 2017
3. LeCun, Y., Bengio, Y., Hinton, G.: REVIEW Deep learning, vol. 521. Macmillan Publishers Limited (2015)
4. Jia, J.: CSIEC (Computer Simulator in Educational Communication): an intelligent web-based teaching system for foreign language learning. In: ED-MEDIA (World Conference on Educational Multimedia, Hypermedia & Telecommunications). Lugano, Switzerland (2004)
5. De Angeli, A., Johnson, G.I., Coventry, L.: The unfriendly user: exploring social reactions to chaterbots. In: Helander, Khalid, Tham (eds.), International Conference on Affective Human Factors Design. Asean Academic Press, London (2001)
6. Creative Virtual. UK Lingubot Customers 2004–2006: Listing of major companies using Linubot technology. [www.creativevirtual.com/customers.php](http://www.creativevirtual.com/customers.php). Accessed 14 Aug 2017
7. Wacky Web Fun Ltd. RacingFrogz.org. c2005: Online game with chatbot capabilities. [www.racingfrogz.org](http://www.racingfrogz.org). Accessed 17 Aug 2017
8. West Ham and Plainstow NDC: New deal for communities: splodge new deal for communities chatbot assistant (2005). [www.ndfc.co.uk](http://www.ndfc.co.uk). Accessed 2 June 2017
9. Ellis, R., Allen, T., Tuson, A.: Applications and innovations in intelligent systems XIV. In: Proceedings of AI-2006, International Conference on Innovative Techniques and Applications of AI (2006)
10. Kerly, A., Hall, P., Bull, S.: Bringing chatbots into education: towards natural language negotiation of open learner models. In: Proceedings of AI-2006. International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer
11. Lifeng Shang, Zhengdong Lu, Hang Li.: Neural Responding Machine for Short-Text Conversation. Accepted as a full paper ACL (2015)
12. Pennic, J.: Your.MD Launches AI-powered doctor diagnosis on facebook messenger via chatbot (2016). <http://hitconsultant.net/2016/04/29/md-launches-ai-powered-doctor-diagnosis-facebook-messenger-via-chatbot/>. Accessed July 2017
13. Szymczak, A.: Introduction to chatbots in heathcare. <https://blog.infermedica.com/introduction-to-chatbots-in-healthcare/>. Accessed July 2017
14. Rigg, J.: UK health service to trial chatbot that gives medical advice (2017). <https://www.engageadget.com/2017/01/05/nhs-chatbot/>. Accessed July 2017
15. Natural Language Toolkit. <http://www.nltk.org/>. Accessed June 2017
16. WORDS API. <https://www.wordsapi.com/>. Accessed June 2017



# A Comprehensive Recommender System for Fresher and Employer



Bhavna Gupta, Sarthak Kanodia, Nikita Khanna and Saksham

**Abstract** Due to overwhelming data on social networking sites about jobs and candidates, it becomes a time-consuming task to generate a match between candidates and employers. Moreover, recruitment of a candidate, who has no work experience called as fresher, poses a two-way problem. Firstly, the candidate due to a lack of experience is not able to decide upon a job among various opportunities which could utilize his/her maximum potential, whereas the employer does not get any past referrals for the candidate to help in the process of recruitment. The proposed study addresses this problem by assisting both; a fresher with a recommended list of job openings which could interest him/her and the employer with a recommendation list of freshers which can be relied upon for the job. The study is assessed and validated with a series of experiments using real data from a social networking site, LinkedIn.

**Keywords** Attributes · Similarity · Ratings · Recommender system

## 1 Introduction

Social networking sites (LinkedIn) provide a platform to connect job seeker(s) and employer(s). While job seeker has a perfect job in his mind, employer also has a picture of an ideal candidate for its job. Their goals are difficult to achieve if the job seeker (fresher/student) has little or no work experience, as work experience gives benefits to both parties: job seekers as well as recruiters. It lays future path for job seekers to make them reach to their destination, whereas it generates referrals for

---

B. Gupta (✉) · S. Kanodia · N. Khanna · Saksham  
Department of Computer Science, Keshav Mahavidyalaya,  
University of Delhi, New Delhi, Delhi, India  
e-mail: guptac7@gmail.com

S. Kanodia  
e-mail: sarthakkanodia@outlook.com

N. Khanna  
e-mail: nikita\_khanna95@outlook.in

recruiters with the help of which they can rely on the credibility of the candidate and minimize risk on their resources. To minimize the effect of work experience on students, a lot of effort and time is devoted both on-campus and off-campus, such as internship opportunities are provided to them from campus and students are provided with help and suggestion from their friends/seniors and family. In addition, job fairs and placement drives are held regularly to channelize students and employers to understand each other's requirements.

It is often realized that different sources (social networking sites, job fairs, placement drives) generate huge amount of data for the fresher, about number of job descriptions and for the employer, about candidate profiles. So there is a need for an efficient mechanism, for the freshers and as well as employers, which can filter useful information from this huge amount of data.

The paper proposes a two-phase system, which serves both, the fresher and the employer, by providing each of them a recommendation list meeting their needs. For recommending employers to a fresher, similarity between the fresher metadata (skill set, internship program) and the graduates who are in the job is obtained. Using similarity measure and applying threshold according to the personalized choice of fresher, a list of employers is generated. This list is further refined on the basis of employer ratings. This accumulated list of top k employers is provided as recommendation list to fresher. To get the recommendation list of potential freshers for an employer, the data of their previously employed recent students are taken and similarity is computed between them and freshers who have applied for the job. This similarity measure helps to generate a recommendation list of all those potential freshers for the employer with which their company will get benefited.

The rest of the paper is organized as follows. Section 2 discusses the related work and Sect. 3 details our proposed work. The various experimental results are elaborated in Sect. 4 followed by conclusion and future work in Sect. 5.

## 2 Related Work

Recommender system helps in filtering huge amounts of data/information while making a decision. But unlike traditional recommender system, job recommender system recommends one type of user (e.g., job applicant) to another type of user (e.g., employer) [1, 2]. Job recommender system uses different approaches that are presented and discussed in [3]. Koh and Chew [4] proposed to use standard parameters which hold distinctive values for a job seekers while [5, 6] improved the results of job recommendation by providing weights to both job seekers' and jobs' different fields. Semantic and tree-based knowledge matching process is discussed to get a profile similarity score with jobs in [7]. It is also been reflected how different profile patterns, similarity patterns and users' interactions can improve the results altogether [8]. Job recommendation framework based on rating of employers and job seekers' nearest neighbor is discussed in [9]. Shi [10] used basic and knowledge attributes for employment and various psychological and social relationship attributes

to improve the similarity score between job seekers and employers. Both parties (job seekers, employers) needs are termed as reciprocal approach which is addressed in [11]. Pizzato et al. [12] used hybrid approach of content and collaborative filtering techniques. In [13], content-based and interaction-based relations are translated into edges connecting different entities (candidates, employers and jobs).

The paper is focused specifically for generating a match between freshers (who do not have any job experience) and employers. Work experience helps both candidates and employers as it reduces the risks on employers’ resources as well as for candidates, and it guides in laying the future path.

### 3 The Proposed System

The currently running recruiting systems are facing problem due to overloading of information from both the parties: job seekers and employers. This problem increases multifold, if job seeker is fresher, i.e., has no prior work experience which makes the employer helpless as he/she has no details of past experience of candidate to get the referrals and lack of experience makes the job seeker confused to decide among the job opportunities. To address this problem, a comprehensive job recommender system is proposed in this paper, which recommends (1) employers to the aspiring freshers (2) eligible freshers for the specified job to the employers.

The system is built in two phases which is also represented diagrammatically in Fig. 1. Various components involved in the system are described in the following subsection.

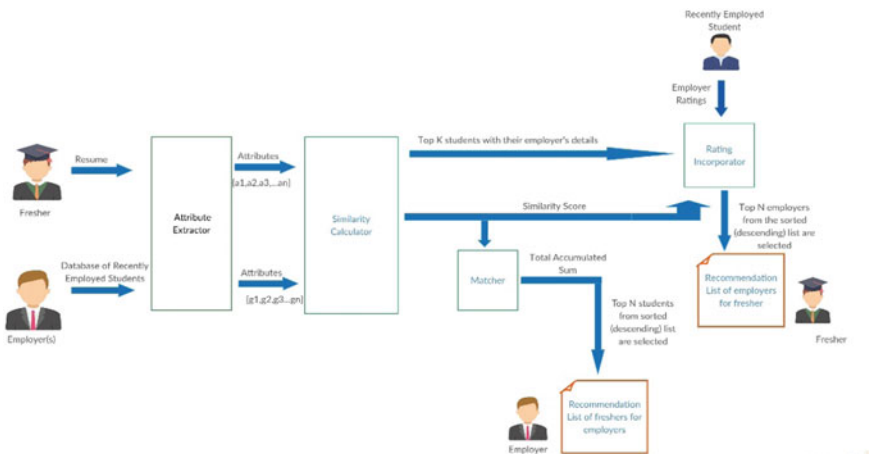


Fig. 1 Job recommender system for fresher and recruiter

### 3.1 Attribute Extractor

This component is supplied with either a fresher's profile or profile of the recently employed students at the employer depending upon the phase in which system is getting used. This module extracts relevant attributes from the input to generate a list of relevant attributes that will be further utilized. The attribute extractor module makes sure that relevant attributes should be selected.

To extract attributes from profile of the fresher, the attributes are defined in two categories, i.e., bool and discrete. Fresher's information such as internship company, internship title, institute and course is considered as bool, whereas internship duration is considered as discrete attribute. Each fresher's profile is formally defined as  $u = \{a_1, a_2, a_3, \dots, a_n\}$  where  $u$  is the student and  $a_1, a_2, \dots, a_n$  are the fresher's attributes as extracted by the module.

In second phase, the database of recently employed students with recruiter is supplied as input to module. The output for this module will be  $e = \{g_{i1}, g_{i2}, g_{i3}, \dots, g_{in}\}$  where  $e$  is the employer and  $g_{i1}, g_{i2}, g_{i3}, \dots, g_{in}$  are the attributes of  $i$ th employed student.

### 3.2 Similarity Calculator

In the first phase, this component computes the similarity between fresher and recently employed students at various employers. Similarity is calculated for both discrete and bool attributes.

For bool attributes, similarity is computed using Eq. 1.

$$S(u, g) = \sum_{i=1}^n w_i * sim(u_{a1}, g_{a1i}) \quad (1)$$

and

$$sim(u_{a1}, g_{a1i}) = \begin{cases} 1 & u_{a1} = g_{a1i} \\ 0 & u_{a1} \neq g_{a1i} \end{cases}$$

Whereas for discrete attributes, the similarity is computed using Eq. 2.

$$S(u, g) = \sum_{i=1}^n w_i * \left(1 - \frac{(|u_{a1} - g_{a1i}|)}{a_{imax} - a_{imin}}\right) \quad (2)$$

where  $S$  is the job applicant/fresher looking for job;  $u_{a1}$  is  $a_1$  attribute of the fresher;  $g_{a1i}$  is  $a_1$  attribute of  $i$ th recently employed student by the employer;  $w_i$  is the

adjustable weight assigned to attribute;  $a_{imax}$ ,  $a_{imin}$  are the maximum, minimum values of the  $i$ th attribute whose similarity is being compared.

### 3.3 Rating Incorporator

This module works for phase 1. The input of this module is the detailed list of top  $k$  recently employed students with their similarity scores above a threshold with fresher and ratings of their employers as given by whole of their staff. This collective rating of particular employer is taken as his/her reputation. The similarity score and the ratings are combined using appropriate weights for each selected employer as shown in Eq. 3 and final recommendation list is prepared.

$$sim_{acc} = w1 * S(u, g_i) + w2 * Repu(e_j) \quad (3)$$

where  $w1$  and  $w2$  are weights assigned to similarity score and ratings of  $j$ th employer of  $i$ th student present in the list prepared similarity module.

### 3.4 Matcher

For the second phase, after calculating the similarity between current fresher/job seeker and the recently employed student, a matching between a given fresher  $u$  and a potential employer  $e$  needs to be defined. This is the function of matcher module. An employer will be defined as a set of similarity score of its recently employed students as  $e = \{g_1, g_2, g_3, \dots, g_n\}$  with the fresher where  $e$  is the employer and  $g$  is its recently employed students. The matching is defined as in Eq. 3.

$$M(u_j, e) = \sum_{i=1}^m S(u, g_i) \quad (3)$$

where  $S(u, g_i)$  is the similarity of fresher with the  $i$ th recently employed students of the employer. The  $M(u_j, e)$  is computed for each  $j$ th fresher and placed in a list. The top matching freshers from obtained sorted list are the recommendation list for the employer.

Summarizing, in first phase, the similarity between  $u$  (the fresher/job seeker) and  $g$  (recently employed student who has offer and who will also provide employer rating) is computed by similarity calculator module and placed in set  $G$ . This similarity calculation is based on various bool and discrete attributes. The recently employed students in  $G$  set are then sorted and arranged according to their similarity weights. Then, we select the top  $k$  students from set  $G$  and also obtain their employers. Once the list is obtained then collective rating of each employer present in the list is taken and

combined with its similarity score in rating incorporator module. Finally, the sorting is done based on this average rating in descending order and the top N employers are recommended to the fresher u.

In second phase, the similarity between the freshers seeking job and employer's employed students is computed using similarity calculator. These similarity scores are supplied to the matcher module. The matcher module combined these similarity scores and generated a descending list of freshers with their scores.

Algorithmically, first phase is written as follows:

**Input:**  $G, u$

**Output:**  $E$

- 1: **for** each recently employed student  $g_i$  in  $G$  **do**
- 2: Calculate similarity with the fresher  $s_{u,g_i} = S(u, g_i)$
- 3: **end for**
- 4: Sort  $s_{u,g_i}$  in descending order
- 5: Remove all entries below the threshold value
- 6: Obtain top  $k$  recently employed students to form  $G_k$
- 7: Obtain employers to get employer set  $E$  related to  $G_k$
- 8: **for** each employer  $e_j$  in  $E$  **do**
- 9: **for** each student  $g_i$  in  $G$  corresponding to  $e_j$  **do**
- 10: **if**  $r_{g_i,e_j} > 0$  **then**
- 11:  $rate_{acc} = rate_{acc} + r_{g_i,e_j}$
- 12: **endif**
- 13:  $Rep(r_{u,e_j}) = rate_{acc}/i$
- 14:  $sim_{acc} = w1 * S(u, g_i) + w2 * Rep(e_j)$
- 15: **end for**
- 16: **end for**
- 17: Sort  $E$  in descending order of  $sim_{acc}$
- 18: return  $E$

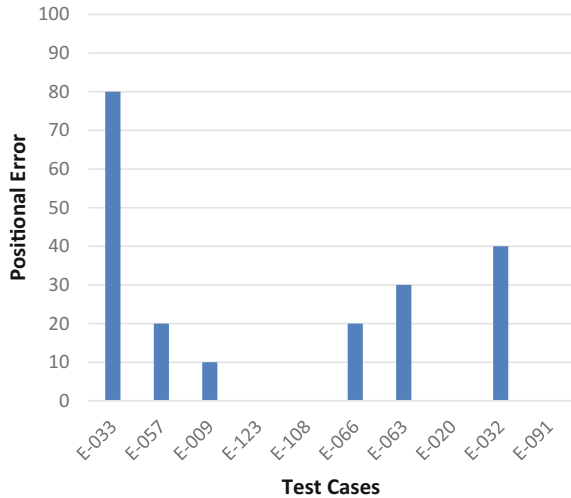
Algorithmically, second phase is written as follows:

**Input:**  $U, E$

**Output:**  $U'$

- 1: **for** each fresher  $u_i$  in  $U$  **do**
- 2: **for** each recently employed student  $g_j$  of  $e$  in  $E$  **do**
- 3: Calculate similarity  $S(u_i, g_j)$
- 4: **end for**
- 5:  $M(u_i, e) = (M(u_i, e) + S(u_i, g_j))$
- 6: **end for**
- 7: Sort  $M(u_i, e)$  in descending order
- 8: Obtain top N freshers to form  $U'$
- 9: return  $U'$

**Fig. 2** Percentage of error for employers' positions in the recommendation list for fresher



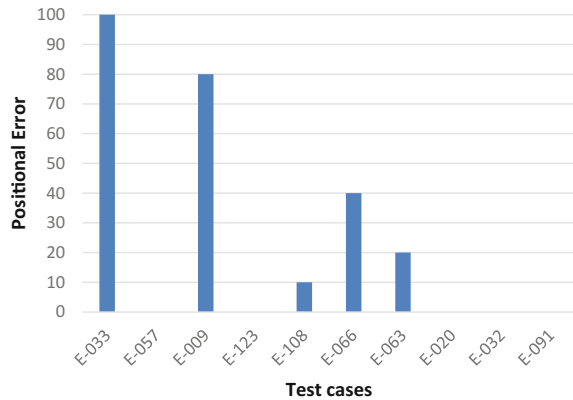
## 4 Results and Evaluations

To validate the proposed system, data from LinkedIn site are used. A dataset of 124 recently employed students at 20 companies having their profiles as courses (with a dictionary to map similar courses), 26 intern companies where a student have interned at, 23 intern titles, i.e., the job role a student had been assigned during the internship, 11 institutions, 46 skills. These students' ratings about their companies (<https://www.glassdoor.com>) are also taken. Each company has a set of 15 assigned skills for which job openings are there, and each of their employees can have minimum of four skills and a maximum of eight skills. Moreover, freshers dataset of 105 unemployed students is taken from the campus itself.

A prototype was implemented for the proposed system and run on a dataset of 105 freshers and 124 recently employed students with 20 companies. Recommendation list of 10 companies/employers for each fresher and recommendation list of 10 freshers for each employer is generated. To validate those lists, 10 test cases for first and second phase are executed and shown in Figs. 2 and 3, respectively.

Figure 2 represents the error present in the positional error of employer's position in each of list items of recommendation lists generated for fresher. Figure 3 represents the error present in the positional error of fresher's position in each of list items present in recommendation lists generated for employers. It is found that the system is **80%** accurate when recommending a company to a fresher and **75%** accurate when recommending a fresher to a company.

**Fig. 3** Percentage of error for freshers' positions in the recommendation list for company



## 5 Conclusion

Understanding the need of fresher for an ideal job and employer for an ideal candidate, a job recommender system is proposed in this paper. The system used the similarity measures among fresher and recently employed students by the employers. These similarity scores are further refined using the reputation of the employers based on ratings by recently employed students. The system is implemented as a prototype and validated on the real dataset as obtained from the LinkedIn site. The system performs better than similarity-based recommender systems as it accounts for employer feedback through the rating mechanism.

## References

1. Al-Otaibi, S.T., Ykhlef, M.: Job recommendation systems for enhancing e-recruitment process
2. Hong, W., Zheng, S., Wang, H.: A job recommender system based on user clustering. *J. Comput.* (2013)
3. Al-Otaibi, S.T., Ykhlef, M.: A survey of job recommender system. *Int. J. Phys. Sci.* (2012)
4. Koh, M.F., Chew, Y.C.: Intelligent job matching with selflearning recommendation engine. *Els. J.* (2015)
5. Diaby, M., Viennet, E., Launay, T.: Exploration of methodologies to improve job recommender systems on social networks, Springer, Wien (2014)
6. Liu, R., Ouyang, Y., Rong, W., Song, X., Xie, W., Xiong, Z.: Employer oriented recruitment recommender service for university students. *Spr. J.* (2016)
7. Musale, D.V., Nagpure, M.K., Patil, K.S., Sayyed, R.F.: Job recommendation system using profile matching and web-crawling. *Int. J. Adv. Sci. Res. Eng. Trends* (2016)
8. Gillet, D., Lu, Y., El Helou, S.: Analyzing user patterns to derive design guidelines for job seeking and recruiting website. *École Polytechnique Fédérale de Lausanne* (2012)
9. Liu, R., Ouyang, Y., Rong, W., Song, X., Tang, C., Xiong, Z.: Rating prediction based job recommendation service for college students. *Spr. J.* (2016)
10. Shi, S.: Real-time job recommendation engine based on college graduates' personal. *J. Res. Sci. Technol.* (2016)



11. Kille, B., Abel, F.: Engaging the crowd for better job recommendations, CrowdRec (2015)
12. Pizzato, L., Rej, T., Chung, T., Yacef, K., Koprinska, I., Kay, J.: Reciprocal Recommenders, University of Sydney (2006)
13. Gillet, D., Lu, Y., El Helou, S.: A recommender system for job seeking and recruiting website. École Polytechnique Fédérale de Lausanne (2013)

# A New Approach of Learning Based on Episodic Memory Model



Rahul Shrivastava and Sudhakar Tripathi

**Abstract** This paper presents a computational model of episodic memory that learns event in response to a continuous sensory input. The proposed model considered event (personal experience) as a collection of coactive activities, where it learns the activities in incremental manner (learns new activity without forgetting the old activities) with the help of fuzzy ART network and learns the event as a unique combination of different category field coactive activities, and also captures the occurred sequence of event in the form of sequence-dependent weights in an episode, which makes it more robust to recall with noisy cue. Also used Hebbian learning to make associations between coactive activities, which helps in pattern completion from the partial and noisy input. To validate the proposed model, an empirical study conducted, where the proposed episodic memory model is evaluated based on the recall accuracy using partial and erroneous cues. The analysis shows that the proposed model significantly associated with encoding and recalling events and episodes even with incomplete and noisy cues, and also our model is found to be more space efficient, and more robust in recalling with noisy cue in comparison with previous ART network-based episodic memory models.

**Keywords** Episodic memory · Encoding · Recalling · Forgetting · ART network

## 1 Introduction

Episodic memory (EM) is a collection of episodes where each episode is a sequence of autobiographic events, which is experienced by an agent. In each episodic memory, event stores with some context information of time place (when it happens, where it

---

R. Shrivastava (✉) · S. Tripathi  
Department of Computer Science & Engineering,  
National Institute of Technology, Patna, Patna, Bihar, India  
e-mail: rahul.vidishaa@gmail.com

S. Tripathi  
e-mail: stripathi.cse@nitp.ac.in

happens) [1, 2]. This context information works as a cue for the retrieval of an event, and later this recalled partial sequence of events recalls the whole stored episode. EM traces are highly distributed and redundant, which makes them highly robust against cell loss [3]. This EM learns event-specific reward which helps to take decision on the basis of prediction of reward, which is associated with the event that is similar to current situational attributes.

In contrast to semantic memory (SM), EM required only single exposition to store an event, to learn event-specific reward and generates the sparse memory presentation (minimum overlapping of traces between events) to reduce the cross talk between the similar events [4], whereas SM required lot of expositions to extract the facts, knowledge and strategies from the overlapping patterns of memory traces of several events using association rule mining and stores it in some structured way without any context of spatial/temporal information. As time passes, some of the memory traces of remote memories in EM loses its association with other traces of same event because of few recalling rate of the same, and some other traces transform into the generalized information of SM because of high recalling rate [4]. By this way, SM derives from the EM and collects the event-specific facts, and later it generates the generalized information from these collected facts of distinct events [5, 6].

The motivation of our paper is to provide the model which simulates the working of episodic memory by making use of ART network, where model can learn the activities in incremental manner which the previous EM-ART model do, also able to capture the experience (event) as a conjunction of different activities and able to capture the sequential pattern of events as an episode, where model needs to be capable to differentiate between two highly similar events/episodes which are semantically different [7] and also able to tolerate the minor differences in the sequence of events [8].

## 2 Background

ART network is a kind of unsupervised network which can adapt the new features without forgetting the old features (incremental learning) [9]. Some previous models (EM-ART model [10–12]) used ART network for EM learning. EM-ART model has ability to memorize the coactive activities as a single experience, but in case when all different field activities of an event do not pass the vigilance criteria, then it stores the similar activity with the weight vectors in different events separately, which creates the problem of redundancy and makes it less space efficient.

To create an activation pattern of events in an episode, each time EM-ART accesses the previous sequential events on occurrence of new event to update the activation value of each sequential event in  $O(n^2)$  time (where  $n$  is the number of constituent events of an episode) which makes it less time efficient.

### 3 Proposed Model

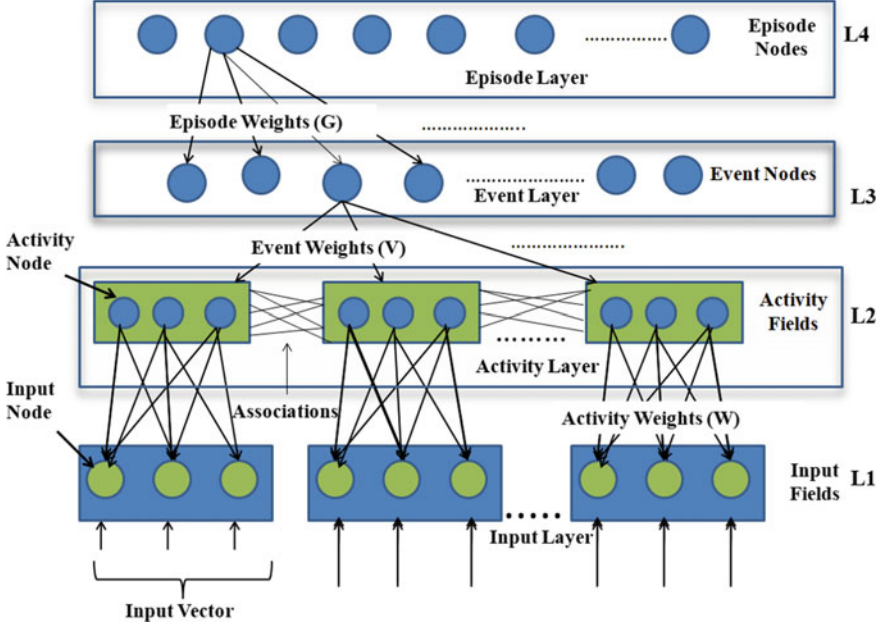
Our proposed model is based on fusion of fuzzy ART neural network [9], which offers a computational process for encoding, recognition and recalling of learned pattern. Here, we considered event as a collection of coactive activities which occurs at same time (same events) in different input field and episode as a collection of temporally correlated events, where each input field learns the activities of own fields in the form of weights and performs incremental learning (learn new activities without forgetting of old activities) by using ART (adaptive resonance theory network).

Our model (shown in Fig. 1) consists of four layers L1 (input layer), L2 (activity layer), L3 (event layer) and L4 (episode layer), where L1 and L2 both divided into input fields and activity fields belongs to different category, respectively, where each input field consists of number of input nodes equal to the size of input vector, where each input node takes the corresponding input value of the vector. Each activity field consists of activity nodes, where each activity node represents the activity, which stores the template of activity pattern generated as a result of sensory input vector. In this way, each activity node stores the activity in the form of weight vector connecting the activity node with all input nodes of similar category, later these weights help in recalling or replay of activity. The node which has highest similarity with input vector, and higher than the vigilance parameter will only be chosen to resonate. After resonance, it allows to change its weight vector to update according to Eq. 1.

Layer L3 is the event layer which contains the event nodes, where each event node links with an activity node of each activity field and represents the constituent activity of the event. This back-to-back input from the sensory field activates the sequence of event in response to the back-to-back activation of coactive activities, which tends to generate an activation sequence of events which is captured as an episode in the fourth layer by the help of an episode node in the form of episode weight vector.

#### 3.1 Activity Learning

It is the very first step which is required in encoding of an experience, where it learns different activities in response to continuous sensory input to different category fields, where each category field have own activities, e.g., tongue is a different category field, where taste of a different food creates a different sensation (activity) on tongue, and the activities of one field cannot occur in another field. Let us discuss for any particular field  $k$ , where input will come from the sensors which are connected to the field. Here, input data can be in the form of real-valued vector, which is collected by the input layer (F1) nodes, where the numbers of input nodes are equal to the size of input vector, and each node collects the single value of input vector. The input data memorize in the form of weight vector activity nodes of (F2), where a weight  $(w_{ij}^k)$  is the value between the activity node  $j$  and input node  $I$  of category field  $k$ . Whenever



**Fig. 1** Figure represents the proposed architecture of computational model of episodic memory, where it is divided into four layers: input layer, which is divided into input fields, where each input field contains the input nodes to receive the corresponding input vector value from sensors. Activity layer, which is divided into activity fields, which contains the activity nodes to store the activity weight vector ( $w$ ) for the corresponding activity pattern received from the input field. Event layer contains the event nodes to store activation pattern of activity in the form of event weight vector ( $V$ ) and episode layer contains the episode nodes to store the activation sequence of events in the form of episode weight vector ( $G$ )

any input  $X^k$  comes from the sensor in the  $k$ th field, it will try to match previously occurred activities by matching with the weight vector ( $W_{jk}$ ) of each activity node  $j$  of F2 of  $k$ th field according to Eq. 1.

Here, input is  $X = (x^1, x^2, x^3 \dots x^k)$ , where  $X$  is a set of  $k$  input vectors and each vector is an input to each input field.

Weight vector  $W_j^k = (w_j^1, w_j^2 \dots w_j^n)$  is the weight vector associated with the  $j$ th activity node of  $k$ th input field of size  $n$ .

$$m_j^k = \frac{|x^k \wedge w_j^k|}{|x^k|} > \mu^k \tag{1}$$

where  $m_j^k$  is the degree of match of input  $X^k$  with the weight vector of category  $k$ ,  $\mu^k$  is the vigilance parameter (minimum degree required for matching) of category  $k$  [12]. Activity node  $j$  will be select for weight learning only if degree of match of

its vector with input is greater than  $\mu^k$ , and weight vector will be updated according to Eq. 2, where  $\alpha^k$  controls the learning rate of category k.

$$w_j^k(\text{new}) = (1 - \alpha^k)w_j^k(\text{old}) + \alpha^k(x^k \wedge w_j^k(\text{old})) \quad (2)$$

where  $w_j^k(\text{old})$  is the weight vector associated with the  $j$ th activity node and  $k$ th input field, and  $w_j^k(\text{new})$  is the derived weight vector from the old,  $x^k$  is the input vector to the  $k$ th input field and  $\alpha^k \in [0, 1]$  is the learning rate parameter which controls the learning rate of the  $k$ th activity field.

### 3.2 Event Encoding

Event is a conjunction of coactive activities (experiences) of different fields, where each activity is shared among several events. At the time of encoding of an event, new event node recruits in the event layer which have links connected to each category field (one link for each category field), these links have some weight value which is equal to the activity number of a resonated activity of the corresponding category field. Suppose there is an event node  $j$  whose weight vector  $V_j$  size equal to the number of category fields, then weight value  $V_{ij}$  is the weight between the event node  $j$  and category field  $i$  (where  $1 \leq i \leq K$ ), which represents the activity number of an activity present in event  $j$  from the category field  $i$ . Also, each event is having some activation ( $EA_j$ ) value calculated according to Eq. 3, which is a function of activity numbers of all category fields.

$$EA_j = \sum_{i=1}^{i=K} V_{ij} \quad (3)$$

where  $EA_j$  is the activation value of the event  $j$ ,  $V_{ij}$  is the weight between event node  $j$  and category field  $i$ .

### 3.3 Episode Formation

Episode is a sequence of temporally spatially correlated events, where different sequence of similar set of events represents a different episode and its constituent events can be shared among episodes. To create an episode, a node recruit in episode layer which has some weighted links connected to the event nodes of the event layer, where each weight value of the link between an episode node and event node represents the activation value of the event in the episode. This activation weight value of an event in an episode (calculated in Eq. 4) is the sum of the activation of the

current event node (calculated in Eq. 3) and the weighted activations of the previous sequential event of the similar episode.

$$G_{pq} = EA_q + (\beta_{q-1} * EA_{q-1}) + (\beta_{q-2} * EA_{q-2}) + \dots + (\beta_1 * EA_1) \quad (4)$$

$$\beta_i = \frac{\frac{1}{t_i}}{\frac{1}{t_1} + \frac{1}{t_2} + \frac{1}{t_3} + \dots + \frac{1}{t_{q-1}}} \quad (5)$$

where  $G_{pq}$  is the activation weight of event  $q$  in the episode  $p$ ,  $EA_q$  is the activation value of event  $q$ ,  $EA_{p,q-1}$  is the activation of the event  $q - 1$  which comes earlier in the sequence in the episode  $p$  and  $\beta$  (calculated in Eq. 5) represents the weightage given to any event in calculation of the activation weight of any other event ( $q$ ), and  $t_i$  is the time gap between the event  $i$  and event  $q$ .

Here in calculation of  $G_{pq}$ , higher weightage is given to the events which come earlier and closer to the event ( $q$ ), and these calculated weights are too much dependent on the sequence of events instead of its positional weights like in another models.

### 3.4 Episode Recalling

Recalling is the mechanism of replay the whole stored sequence of events of a resonated episode on presentation of a cue, where cue can be noisy (noisy cue contains the distracting events which baffles to resonate a desired episode). To resonance check, it is required to match (according to Eq. 5) the extracted activation values of the cue events from the corresponding stored weighted activation value of events in all episodes, the episode which has highest degree of match and greater than the vigilance parameter  $\mu^e$  will only resonate to recall.

$$M_p = \frac{|G_p \wedge L|}{|G_p|} \quad (6)$$

where  $M_p$  is the degree of match between the weight vector of episode  $p$  and extracted weight value of from the input cue  $L$ . Whenever an episode recognized on coming of weak and partial input cue, then complete sequence of events of the recognized episode can be reproduced with the help of the weight vectors of the episode. The event which is having least weight value with the recognized episode select first to recall by reactivation of all its constituent activities in L2 layer by reproducing the stored activity pattern in the corresponding output field of activity.

### 3.5 Event Recalling

Event recalling is a mechanism of pattern completion on presentation of partial input or query. It is widely accepted that the hippocampus is responsible for pattern completion, which creates associations between the coactive cortical activities, like by smelling a particular dish we can easily recognize the dish, because of strong associations between the dish and smell. These associations can be generated with the help of Hebbian learning according to Eq. 6, where the associative weight between any two activity nodes increases by 1 when both are active in similar event at any instant of time. These associations stores the semantic knowledge [] and helps in long-range semantic inference.

$$S_{jk} = S_{jk} + (AA_j * AA_k) \quad (7)$$

where  $S_{jk}$  is the associative weight between activity nodes  $j$  and  $k$ , initially  $S_{jk} = 0$ , and  $AA_j$  is the activation of activity node  $j$ , where  $AA_j \in [0, 1]$ .

To perform recalling on presentation of partial input, we used a graph technique, where we evaluate a complete graph of highest degree (where all activities are connected with each other) from the partial input to recall an event. Firstly, it will take a set (let say  $U$ ) of nodes which are present in the cue and then evaluate a set (let say  $T$ ) which contains the intersection of adjacent nodes of the nodes belongs to  $U$ , and then evaluates a complete graph of highest possible degree (Degree is a edge weight of complete graph), and then activate the set of nodes of the resultant complete graph which tends to activate an event if its calculated event activation (according to Eq. 3) value is higher than the threshold. If none of the event resonates, then try for the next higher degree complete graph and repeat the process until any event resonates.

## 4 Case Study

To evaluate our model, we tested our model to perform different tasks (shown in Table 1), where each task required different sequence of events to perform, and also several events are shared among episodes, e.g., task of making tea and task of making coffee shared several events like add water, start stove and boil water. Whenever any partial event sequence is presented as a cue, then one of the stored episodes will resonate to replay according to Eq. 5. Here, cue can have varied level of noise, where noise is the event which is present in the event sequence of cue and distracts to match with stored episode.



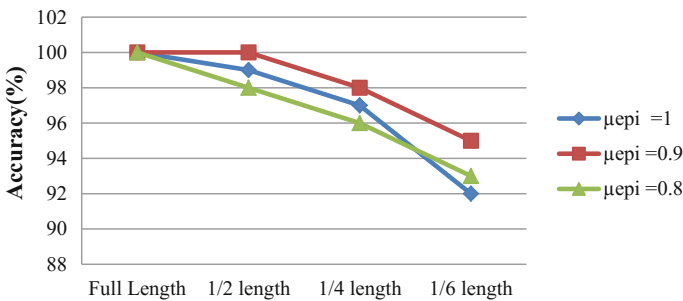
**Table 1** Sequential events in different tasks/episodes

Prepare tea	Prepare popcorn
Take pan	Take corn packet
Fill water	Take pressure cooker
Start stove	Put oil in pressure cooker
Grasp tea packet	Put corn in pressure cooker
Add tea	Start stove
Grasp sugar	Weight 5 min
Add sugar	Off stove
Grasp milk packet	–
Add milk	–
Boil	–
Off stove	–

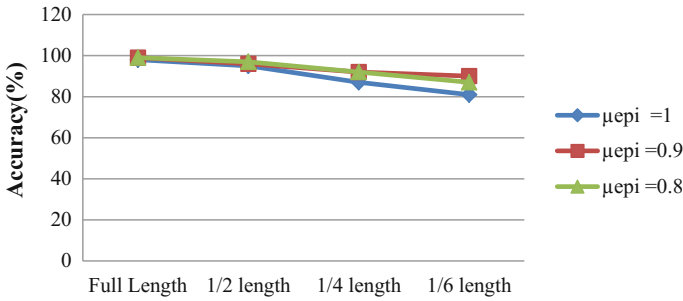
### 5 Results

We observed results based on retrieving accuracy of recalled events/episode on different type of cues. Model is evaluated on partial and full-length cue with varied level of vigilance parameter value ( $\mu_{epi}$ ) for episode resonance. We conduct the test on two type of cue, first on the cue which is retrieved from the beginnings of the episode (shown in Fig. 2) and second from which is extracted from the end of the episodes (shown in Fig. 3). We observed the retrieving accuracy is low at higher value of vigilance, because of the low tolerance level, and our model performed almost similar in both type of cue, because our model captures the sequence in the form of weights which gives higher weightage to the closer events in sequence, this is why our model can perform robustly in recalling.

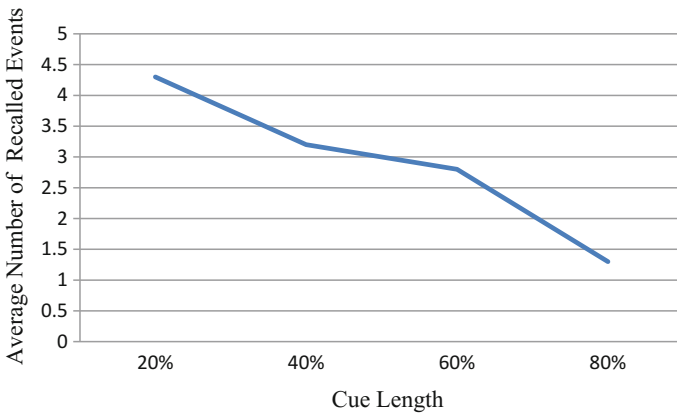
Another test is conducted on pattern completion from the partial input, here partial input is like a query for the model to extract the other associative activities (answer of query). We conducted the test on different level of partial input, containing different



**Fig. 2** Figure represents the recalling accuracy of episode under various cue length and vigilance parameter ( $\mu_{epi}$ ), where cues are extracted from the beginning of the episodes



**Fig. 3** Figure represents the recalling accuracy of episode under various cue length and vigilance parameter ( $\mu_{epi}$ ), where cues are extracted from the end of the episodes



**Fig. 4** Figure represents the average number of recalled events at varied level of partial input, where vertical axis in the graph represents the average number of recalled events, and horizontal axis represents number of activity present (in percentage) in a cue with respect to the total activity present in an event

number of activities (shown in Fig. 4) in cue (which is extracted from the events), which is shown in percentage in Fig. 4. We observed that the less number of cue activities (less specified detail) activates the large number of events, and the large number of cue activities (more specified detail) activates the almost single event.

## 6 Conclusion and Future Work

We presented a new approach to simulate the working of EM, where we have done modifications in EM-ART model by introducing a new activity layer, where learning of an activity is independent to the resonance process of an event and other coactive activities, which removes the redundancy and makes it more space efficient than the

other previous EM-ART models. Also, we used a different mechanism to extract an activation pattern for sequence of events of an episode, which makes it more robust to recall comparison with the other model. We performed empirical study on model in recalling/prediction of different sequences of events of different tasks/episode on presentation of noisy cue with varied level of cue length. Our results showed that our model performed robustly in recalling with noisy cue and comparatively better than other model. In future, we will try to derive the semantic and procedural memory from the episodic memory and will try to add the Ebbinghaus forgetting mechanism in the model.

## References

1. Tulving, E.: Multiple memory systems and consciousness. *Human Neurobiol.* **6**(2), 67–80 (1987)
2. Mizumori, SJY: Context prediction analysis and episodic memory. *Front. Behav. Neurosci.* **7**, 132 (2013)
3. Rolls, E.T.: A computational theory of episodic memory formation in the hippocampus. *Behav. Brain Res.* **215**(2), 180–196 (2010)
4. Grossberg, Stephen: Competitive learning: from interactive activation to adaptive resonance. *Cognit. Sci.* **11**(1), 23–63 (1987)
5. Yassa, M.A., Reagh, Z.M.: Competitive trace theory: a role for the hippocampus in contextual interference during retrieval. *Front. Behav. Neurosci.* **7**, 107 (2013)
6. Shastri, Lokendra: Episodic memory and cortico-hippocampal interactions. *Trends Cognit. Sci.* **6**(4), 162–168 (2002)
7. Shastri, L., Ajjanagadde, V.: From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behav. Brain Sci.* **16**(03), 417–451 (1993)
8. Wang, W., et al.: A self-organizing approach to episodic memory modeling. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE (2010)
9. Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Netw.* **4**(6), 759–771 (1991)
10. Wang, W., et al.: Neural modeling of episodic memory: encoding, retrieval, and forgetting. *IEEE Trans. Neural Netw. Learn. Syst.* **23**(10), 1574–1586 (2012)
11. Leconte, F., Ferland, F., Michaud, F.: Fusion adaptive resonance theory networks used as episodic memory for an autonomous robot. In: *International Conference on Artificial General Intelligence*. Springer, Cham (2014)
12. Subagdja, B., Tan, A.-H.: Neural modeling of sequential inferences and learning over episodic memory. *Neurocomputing* **161**, 229–242 (2015)

# A Hybrid Model for Mining and Classification of Gene Expression Pattern for Detecting Neurodegenerative Disorder



S. Geeitha and M. Thangamani

**Abstract** The exploration of gene expression data leads to various discovery of diseases in the human life. This research classifies gene expression pattern and detects the discriminative genes associated with neurodegenerative diseases by implementing the Naive Bayesian (NB) network model based on particle swarm optimization (PSO) techniques to reduce the disease dimension. Artificial neural network (ANN) is a traditional approach used to classify the disease type and produces either failure or non-failure based on the disease features. The integration of artificial neural network (ANN) and Bayesian logistic regression (BLR) model has been developed to pre-select the gene sample for feature selection, and those selected genes are then used to construct the ANN model. This hybrid model is mainly employed to reduce the time in gene classification and uncovers the diseased gene expression pattern that helps in selecting the victim genes for early detection of diseases in the medical era.

**Keywords** Data mining · Gene expression pattern · Neurodegenerative disorder  
Naive Bayesian network model · Particle swarm optimization technique  
Artificial neural network · Bayesian logistic regression

## 1 Introduction

Data mining plays a major role in the field of bioinformatics since enormous data are generated daily. Analyzing gene expression has become a recent trend in the medical area to retrieve some useful information about the gene expression especially in the detection of neurodegenerative disorders. Classification of gene expression pattern makes a remarkable role in the cancer diagnosis as it represents the state of

---

S. Geeitha (✉) · M. Thangamani  
Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India  
e-mail: geethu.neelu@gmail.com

M. Thangamani  
e-mail: manithangam2@gmail.com

cell at molecular level [1]. Micro-array technology is fundamental tool for studying the gene expression pattern [2]. It has the capacity to determine the thousands of genes at the same time. The data mining algorithm and tools are deployed in various gene expression analyses to find the specific feature of the gene expression of the diseased patients. This study proposed feature selection algorithms including logistic regression to select the candidate genes associated with neurodegenerative disorder to categorize and identify the discriminative samples which is then extracted by ANN model. The performance of this hybrid technique is then optimized by insertion of PSO-based method.

## 2 Related Works

The PSO algorithm performs searches using a population of particles corresponding to individuals in an evolutionary algorithm (EA). And it is specifically used for parameter optimization in continuous process. It swarms the behaviors observed in flocks of birds [3]. The feature selection algorithms such as support vector machine, random forests, Naive Bayes, artificial neural network, logistic regression and k-nearest algorithm were conducted to rank the genes according to the series of algorithms [4]. The clustering method namely low rank representation (LRR) algorithm is implemented to extract the essential information from noisy infrastructure and also capable of capturing the undiscovered gene patterns with similar features [5]. Geetha et al. proposed the usage of high throughput technologies in performing the exhaustive number of measurements over a short period of time giving access to individual DNA, transcribed RNA from genes over time [6]. Kranthi Kumar et al. proposed a probabilistic-based PSO for choosing the subset features of original attributes to find the optimized relationship in the selected features of medical data [7]. The proposed work implements the hybrid model to minimize the classification duration and identify the diseased genes for early detection of neurodegenerative disorders. Accurate detection of Alzheimer's disease (AD) at early stage is beneficial for managing the disease [8–12]. The combination of SVM and decision tree is used to classify the gene pattern but not suited for nonlinear data. The inductive bias reveals an instance that consistently tries to generalize the closest neighbor and does not perform with hyper-rectangle and prune if conflicts [13, 14].

## 3 Proposed Methodology

The human brain normally includes 25,866 single-nucleotide polymorphism (SNP) that accompanies the association pairs of 3709 genes. For every gene, regulated SNPs are extracted and these can be called as significant related SNPs [15]. The proposed data mining methodology employs a hybrid technique for classifying gene expression pattern by identifying the discriminative genes that may be relevant in

early detection of the neurodegenerative disorders. In this paper, we undergo five phases to analyze the gene expression pattern

### 3.1 Data Loading

Sample single-nucleotide polymorphisms (SNPs) are taken from NCBI source for analyzing the diseased gene with the normal gene. These SNPs are the main source that leads to susceptibility of certain disease by revealing their effects on gene expression at the post-transcriptional level; finally, those results in gene dysregulation. In our study, totally eight samples of SNPs are taken for experiments that focus mainly on victim gene.

### 3.2 Preprocessing

Each SNP sample comprises SNPID, chromosome position gene type, functional consequence and clinical significance. In this phase, noisy data are removed and filtered by normalization process using WEKA tool.

### 3.3 Classifying Gene Expression Using Naive Bayesian Network Model

At the initial phase, many molecular networks are constructed to characterize interactions between the biomolecules by implementing the Bayesian network. The gene data set is converted into frequency table. The class prior and predictor probability of gene data set is classified, likelihood of the diseased gene expression is determined, and then the posterior probability of the target gene is calculated. The target gene is obtained by the Naive Bayesian equation

$$P(GC|X) = \frac{P(X|GC)P(GC)}{P(X)} \quad (1)$$

In Eq. 1,

P(X) Predictor prior probability of sample gene data 'X' is true

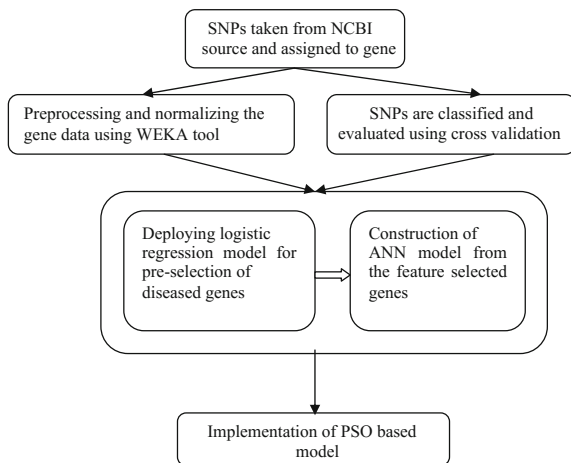
P(GC) Prior probability of gene class

P(X|C) Likelihood probability of diseased gene if occurrence is true

P(C|X) Posterior probability of gene expression given that attribute

X is true.

**Fig. 1** Proposed architecture



### 3.4 Construction of ANN and Logistic Expression

The proposed hybrid model is most probably employed in gene expression classification and for predicting purposes. This hybrid technique is introduced to classify the samples and is evaluated using cross-validation methods, and the result is then utilized to produce optimal model to construct feature gene selection model. After classifying the gene expression pattern by Naive Bayesian network, this hybrid methodology undergoes second classification for selecting diseased genes from the evaluated sample gene expression. The data flow diagram (Fig. 1) represents the construction of the proposed architecture.

### 3.5 Implementation of PSO Technique

Particle optimal solution (PSO) provides multiple potential solutions at one time. The first solution obtained is the best solution (pbest) that provides the fitness value. From the NB model, gene expression pattern is taken as test data set and during each iteration the fitness is determined by the objective function. Each of the individual gene expression data is updated, and global best (gbest) is evaluated. And finally the velocity and position of each diseased gene expression are updated using following Eq. (2)

$$v_{ge(t+1)} = \omega v_{ge(t)} + c_1 \gamma_1 [x_i(t) - x_g(t)] + c_2 \gamma_2 [(g(t) - x_i(t))] \quad (2)$$

- $v_{ge}(t)$       diseased gene particle's velocity at time t
- $x_i(t)$       is the gene particle's position at time t
- $x_g(t)$       is the best solution (Pbest) of gene expression pattern

$g(t)$  is the swarm's best solution as of time  $t$   
 $c_1$  and  $c_2$  are acceleration coefficients  
 $\gamma_1$  and  $\gamma_2$  are random variables.

**PSO Pseudocode:**

```

P= particle Initialization ();
For i = 1 to smax
  For each gene particle Gp in P do
    Fp=f (Gp);
    If fp is better than f(Gpbest)
      Pbest=Gp;
    end
  end
  gbest = best Gp in P;
  For each gene particle p in P do
    V=v+c1*random*(Gpbest-Gp) +c2*random*(gbest-Gp);
    P=Gp+v;
  End

```

**4 Results and Experiments**

Table 1 shows 8 samples of SNPs taken from NCBI source, and it is classified according to the chromosome position and gene expression. Each sample comprises 50–100 genes of various types mentioned in the table.

In this work, Naive Bayesian network and hybrid technique were performed to classify the gene expression data and also to remove noise in gene selection. The sample of 8 SNPs taken from NCBI source is represented in Table 1. These gene data are normalized for filtering the noisy data, and the normalized SNP sample is shown in Fig. 2.

**Table 1** SNPs from NCBI source

SNP-ID	Chromosome position	Gene type
rs906807	18:9117869	NDUFV2
rs1048971	1:207472977	CR2
rs104498	6:131851228	CR2
rs1050565	17:30249058	BLMH
rs1051730	15:78601997	CHRNA3
rs1061234	11:5249456	HBG1
rs1064651	1:155235727	GBA
rs16176640	7:100719675	EPO



Selected attribute

Name: SNP-ID=rs906807  
 Missing: 0 (0%)

Distinct: 2

Type: Numeric  
 Unique: 1 (11%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.111
StdDev	0.333

Fig. 2 Normalizing SNP sample

Naive Bayes Classifier

Attribute	Class			
	BLM (0.25)	CHRNA3 (0.25)	HB31 (0.25)	GBA (0.25)
=====				
SNP-ID				
rs1050565	2.0	1.0	1.0	1.0
rs1051730	1.0	2.0	1.0	1.0
rs1061234	1.0	1.0	2.0	1.0
rs1064651	1.0	1.0	1.0	2.0
[total]	5.0	5.0	5.0	5.0
Chromoposition				
17:30249058	2.0	1.0	1.0	1.0
15:78601997	1.0	2.0	1.0	1.0
11:5249456	1.0	1.0	2.0	1.0
1:155235727	1.0	1.0	1.0	2.0
[total]	5.0	5.0	5.0	5.0
Functional Consequence				
missense	2.0	1.0	2.0	2.0
nc transcript variance	1.0	2.0	1.0	1.0
[total]	3.0	3.0	3.0	3.0
Clinical Significance				
Benign	2.0	1.0	1.0	1.0
drug response	1.0	2.0	1.0	1.0
other	1.0	1.0	2.0	1.0
Pathogenic	1.0	1.0	1.0	2.0
[total]	5.0	5.0	5.0	5.0

Fig. 3 NB classification-based PSO model from NCBI source

By applying NB model, gene class references are classified (Fig. 3) except articles, WEKA tool. Clinical significance is based on the chromosome position and functional consequence of gene expression pattern.

To determine the significant category of genes, classification analysis is done by logistic regression (Fig. 4) and diseased genes are identified from the evaluation of artificial neural network.

Predictive analysis is performed to determine the relationship between the candidate genes and to estimate the probability of binary response based on one or more predictor gene expression (Fig. 5).

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8  
Coefficients...

Variable	Class		
	BLMH	CHRNA3	HGB1
Functional Consequence	-4.7313	31.6868	-4.7313
Intercept	0	-15.0958	0

Odds Ratios...

Variable	Class		
	BLMH	CHRNA3	HGB1
Functional Consequence	0.0088	5.773175400245761E13	0.0088

Fig. 4 Logistic regression model

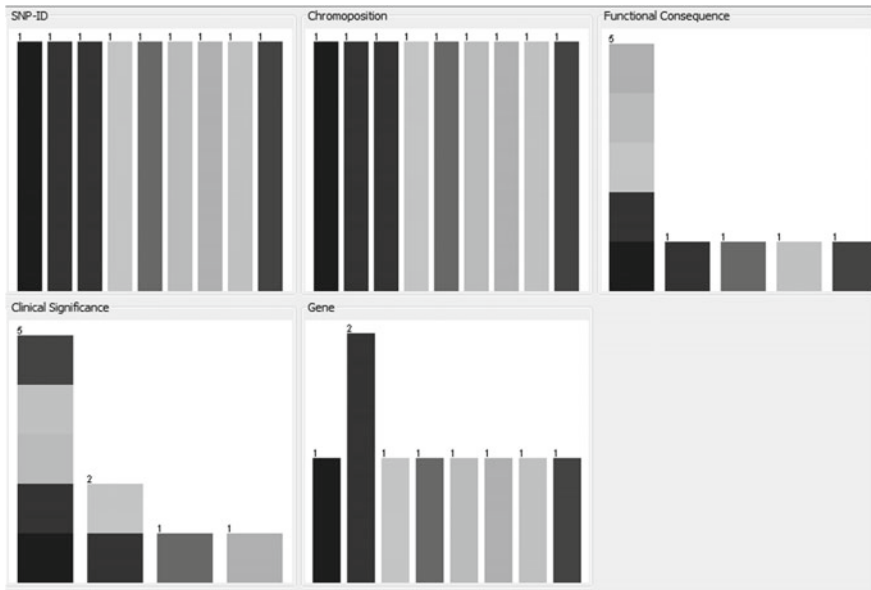


Fig. 5 Comparison of victim gene with normal gene based on functional consequence and clinical significance in WEKA platform

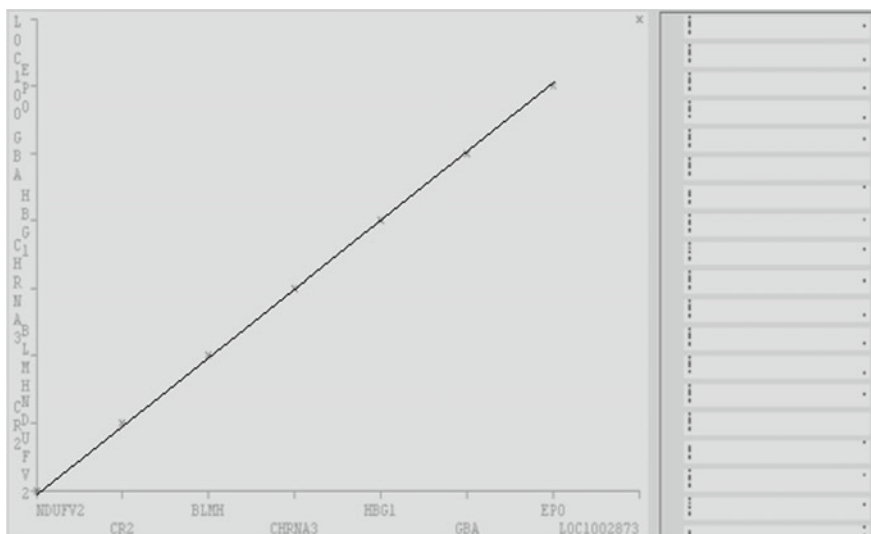
The dark shades in the chromosomal position and gene attributes represent the victim gene affected by neurological disorders. Finally after classification of gene expression pattern using hybrid method, instance-based learning classifier is performed to facilitate better classification. In the proposed method, performance of the hybrid model is represented in the table format.

Table 2 describes the performance analysis of the proposed hybrid model (PSO-based ANN with logistic regression) and compares the results with the traditional models in terms of F-measure. The true positive rate in the proposed model is compared with existing models.

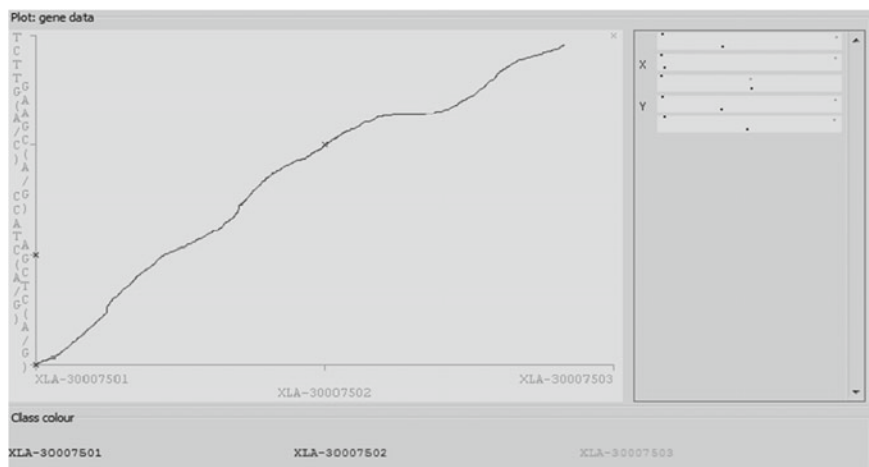
In this proposed work, classification is performed using WEKA platform by implementing hybrid model and then PSO algorithm is then added to this platform to determine the optimal fitness. The evaluated gene pattern is taken as test data and to obtain accuracy and optimal solution test cases is taken as pbest and global data is evaluated from the training data. The velocity and position of normal genes and diseased gene expression are compared (Figs. 6 and 7) in the proposed work.

**Table 2** Performance of hybrid model

Gene data size	Log.Reg	ANN LOG.Reg	NBSO + ANNLOGREG
50	0.73	0.76	0.80
100	0.65	0.70	0.81
500	0.71	0.74	0.78
1000	0.63	0.69	0.86



**Fig. 6** Representation of normal genes



**Fig. 7** Representation of victim genes

The normal gene expressions are represented by a linear data points, whereas the data points are nonlinear in the victim genes.

## 5 Conclusion

The PSO-based hybrid technique produces the better accuracy in classifying the diseased gene from the normal gene. Naive Bayesian model constructs an effective classification analysis for determining the gene expression pattern. It also removes certain noise in the gene data. The PSO model is regarded as best technique for obtaining optimal fitness from the test data received from the NB model, and it provides best solution by minimizing classification time and also to identify the discriminative gene data. This system inhibits various methods to categorize the diseased gene from the normal gene which leads to better diagnosis in the medical field. The future work can be done to improve the computational efficiency, and the hybrid methodology can be proved on the big data.

## References

1. Singh, R.K., Sivabalakrishnan, Dr. M.: Feature selection of gene expression data for cancer classification: a review. In: 2nd International Symposium on Big Data and Cloud Computing (ISBCC '15), No. 50, pp. 52–57. Elsevier B.V, Procedia Computer Science (2015)
2. European Bioinformatics Institute, EMBI-EBI Annual Scientific Report 2013 (2014)

3. Chen, Y., Chen, F., Wu, Q.: An artificial neural network based dynamic decision model for time-series forecasting. In: Proceedings of International Joint Conference on Neural Networks, pp. 12–17. Orlando, Florida, USA (2007)
4. Guo, P., Zhang, Q., Zhu, Z., Huang, Z., Li, K.: Mining gene expression data of multiple sclerosis. *PLOS One* **9**(6), 1–9 (2014)
5. Cui, Y., Zheng, C.-H., Yang, J.: Identifying subspace gene clusters from microarray data using low-rank representation. *Plos One* **8**(3) (2013)
6. Geeitha, Dr., Thangamani, M.: Omics technology in big data, scope. *Int. J. Sci. Human. Manage. Technol.* **2**(1) (2015)
7. Kranthi Kumar, G.: An optimized particle swarm optimization based ANN model for clinical disease prediction. *Ind. J. Sci. Technol.* **9**(21) (2016)
8. Chaves, R., Ramirez, J., Gorriz, J.M.: Alzheimer's Dis, N., Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis. *Expert Syst. Appl.* (40), 1571–1578 (2012)
9. Chen, Y., Liu, Z., Zhang, J., Xu, K., Zhang, S., Wei, D., et al.: Altered brain activation patterns under different working memory loads in patients with Type 2 diabetes. *Diab. Care* (37), 3157–3316 (2014)
10. Cohen, A.D., Klunk, W.E.: Early detection of Alzheimer's disease using PiB and FDG PET. *Neurobiol. Dis.* (72), 117–122 (2014)
11. Collins, M.P., Pape, S.E.: The potential of support vector machine as the diagnostic tool for schizophrenia: a systematic literature review of neuroimaging studies. *Eur. Psych.* (22), 117–122 (2011)
12. Colloby, S.J., O'Brien, J.T., Taylor, J.P.: Patterns of cerebella volume loss in Dementia with Lewy bodies and Alzheimer's disease: a Vbm-Dartel study. *Psych. Res.* (223), 187–19 (2014)
13. Ozbakır, L., Delice, Y.: Exploring comprehensible classification rules from trained neural networks integrated with a time-varying binary particle swarm optimizer. *Eng. Appl. Artif. Intell.* **24**(2), 12–19 (2011)
14. Kiranyaz, S., Once, T., Gabbouj, M.: Stochastic approximation driven particle swarm optimization with simultaneous perturbation—who will guide the guide? *Appl. Soft Comput.* **11**(2), 102–121 (2011)
15. Li, J., Wang, L., Guoa, M., Zhang, R., Dai, Q., Liu, X., Wang, C., Teng, Z., Xuan, P., Zhang, M.: Mining disease genes using integrated protein–protein interaction and gene–gene co-regulation information, *Febs Open Bio*, Elsevier Publication, vol. 5, pp. 251–256 (2015)

# A New Deterministic Method of Initializing Spherical K-means for Document Clustering



Fatima Gulnashin, Iti Sharma and Harish Sharma

**Abstract** Document clustering is required when the possible categories into which text data are to be organized are not known. Standard clustering algorithms do not suit well due to high sparsity of term matrices of document corpus. Use of cosine similarity among document vector has proved to give good results. Its use with k-means is referred as spherical k-means. The performance of spherical k-means highly depends on its initialization. This paper proposes a deterministic initialization technique for spherical k-means that considers the distribution of vectors within the space. Experiments on real-life data with skewed distributions are done to compare performance with other initialization methods. A related technique to avoid generation of empty clusters is also proposed.

**Keywords** Document clustering · Spherical k-means · Initializing k-means  
Deterministic initialization · Clustering

## 1 Introduction

Document classification or text classification has long been in practice. It needs supervised learning approaches that need to be trained from collected samples of known category labels. Many successful classifiers based on neural networks and support vector machines have been proposed. The only drawback is that these assume existence of training samples from each possible category. It is not the case in

---

F. Gulnashin (✉)

R.N. Modi Engineering College, Kota, India  
e-mail: fatima.gulnashin01@yahoo.co.in

I. Sharma

Career Point University, Kota, India  
e-mail: itisharma.uce@gmail.com

H. Sharma

Rajasthan Technical University, Kota, India  
e-mail: harish.sharma0107@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_14](https://doi.org/10.1007/978-981-13-1708-8_14)

real-life applications where text documents may arrive during use rather than training phase. Some categories evolve. Certain applications may decide to have hierarchical categories, or the possible categories are not known at all. In all such situations, an unsupervised technique is preferred. It is called document clustering. The purpose is to group similar documents together.

The documents are conveniently represented as numbers and hence can be clustered using standard k-means algorithm [1]. It uses Euclidean distance which is not suitable for text documents [2]. Similarity among text documents is best achieved through cosine similarity measure when corpus is represented through vector space model. In a vector space model, the documents are stored as vector of numbers indicating word count for each word occurring in the corpus. These tend to get affected by length of documents. Hence, document vectors are normalized such that they fall within a unit hypersphere. Clustering these transformed values using k-means gives a modified optimization objective and is called spherical k-means [3]. Like standard k-means, spherical k-means is also plagued by the marring effect of bad seeding. Initialization of this method can be attempted differently from the initialization of k-means. This is because some characteristics of document data are very different from general numeric data. The distribution is not Gaussian, and geometric interpretation is not possible for it. Hence, popular initialization techniques of k-means based on principal component analysis like [4] and distance based like k-means++ [5] are not appropriate. Authors in [6] suggest that the first concept vector is computed as the concept of entire dataset; thereafter this unified concept is perturbed randomly to obtain k different concept vectors. Duwairi and Abu-Rahmeh [7] present a deterministic technique for initialization that places the concept vectors uniformly in all directions within the unit hypersphere.

This paper presents a deterministic technique for initializing spherical k-means. A drawback of the method suggested by Duwairi and Abu-Rahmeh [7] is discovered, and the proposal is to overcome it.

## 2 Initialization Method by Duwairi et al.

Like the popular k-means algorithm, the spherical k-means too is very sensitive to the initial conditions. The initial values of centroids need to be set carefully for good results. Duwairi and Abu-Rahmeh [7] have proposed a deterministic method for initialization. Suppose, k number of clusters are desired in output then, k centroids have to be set initially. They suggest to uniformly allocating these centroids in the object space. Every dimension in the data from 1 to d has different range. Its range is divided into k parts, and k values of that dimension are computed separately. Let the centroids be denoted by  $m_i$  for  $i = 1$  to  $k$ . For every dimension, the minimum

and maximum values are stored in vectors  $l$  and  $h$ , respectively. Mathematically, the centroids or concept vectors are computed as

$$m_{ij} = l_j + j * \frac{h_j - l_j}{k + 1}. \quad \forall i = 1, 2 \dots k \text{ and } j = 1, 2 \dots d \quad (1)$$

Thus, this method tries to place  $k$  centroid uniformly in the data space. To have them within the hypersphere, the remapping using normalization by its length is done as

$$m'_i = \frac{m_i}{|m_i|} \quad (2)$$

### 3 Drawbacks of Duwairi Method

The initialization proposed by Duwairi and Abu-Rahmeh [7] aims at deterministic and uniform distribution within the object space. It is a simple and fast technique as desirable in high-dimensional data like that of documents. Yet it has a flaw from practical point of view and a major drawback from point of view of generality. The drawback is a uniform placing of centroids in the unit hypersphere will give good results only when the documents themselves are also uniformly distributed in the space. The Duwairi method considers only the range through minimum and maximum values. No information concerning actual distribution is collected. This may lead to empty cluster in the initial step of assigning objects to clusters. The process of spherical k-means is such that if any cluster is empty in the initial step, it remains empty in the output. This indicates that Duwairi method cannot produce good results if the data contain some outliers or have too many variations in the size of clusters. If more than 50% of objects are concentrated in a region, the Duwairi initialization will be very poor. This adverse effect of outlier values cannot be overlooked.

### 4 Proposed Initialization Method

The deterministic method of [7] can be improved to be more generalized by making the distribution of centroids according to density of objects in the space. If it is achieved solely through statistical methods, then the advantage of simplicity and scalability can be retained. We propose to initialize centroids by  $k$  medians in every dimension. Let  $X$  be the matrix of size  $N \times d$  representing the data of  $N$  documents. Then form  $X_s$  as a matrix where each column of  $X$  is sorted individually and stored. That is,

$$X_{s_j} = \text{sort}\left(\left(x_{1j}, x_{2j} \dots x_{Nj}\right)^T\right), \quad \forall j \in [1, d] \quad (3)$$



Now, the centroids can be decided as

$$m_{1j} = \text{median}\left(\left\langle xs_{1j}, xs_{2j} \dots xs_{\frac{N}{k}j} \right\rangle\right), \quad \forall j \in [1, d]$$

$$m_{2j} = \text{median}\left(\left\langle xs_{(\frac{N}{k}+1)j}, xs_{(\frac{N}{k}+2)j} \dots xs_{(\frac{2N}{k})j} \right\rangle\right), \quad \forall j \in [1, d]$$

And so on. Thus, the sorted values are divided into  $k$  groups, and median of group  $i$  is the value of that dimension of centroid  $m_i$  in generalized form,

$$m_{ij} = \text{median}\left(\left\langle xs_{(\frac{(i-1)N}{k}+1)j}, xs_{(\frac{(i-1)N}{k}+2)j} \dots xs_{(\frac{iN}{k})j} \right\rangle\right), \quad \forall i \in [1, k], \forall j \in [1, d] \quad (4)$$

In order to conform with  $X'$ , the data mapped on unit hypersphere, the centroids are relocated through normalization by their vector lengths as

$$m'_i = \frac{m_i}{|m_i|} \quad (5)$$

## 5 Proposed Clustering Method

The proposed initiation techniques as discussed above are used with spherical  $k$ -means to cluster the documents. Let  $\mathbf{X}$  be the representation of corpus in vector space model  $\mathbf{X} = X_1, X_2, \dots, X_N$ , where any  $i$ th document is a vector  $X_i = x_{i1}, x_{i2}, \dots, x_{id}$ ,  $d$  being the number of terms. It is transformed by normalizing each vector by its length to  $X'$ , as

$$x'_i = \frac{x_i}{|x_i|}$$

Sort the original matrix  $\mathbf{X}$  column wise without preserving rows to obtain  $X_s$  as given in (3). Decide the  $k$  initial centroids using (4). For each of the  $N$  document vectors assign a cluster label according to nearest centroid. Instead of computing distance, the cosine similarity is checked. Document is assigned the cluster label if the centroid is with maximum similarity to the document. When all objects (documents) have been assigned cluster labels, the centroids are updated as means of the objects in their cluster. In case, any cluster is empty assign the object closed to its centroid as new centroid. Since this step may take much computation  $(p - 1)$ th and  $(p + 1)$ th centroids. This preserves the concept of medians of initialization process. A random assignment of centroid for an empty cluster is wrong as it may cancel the effect of good initialization.

## 6 Results

The experiments are performed on two very popular real-life datasets—Reuters and Newsgroups, taken from the UCI repository. Instead of taking the whole corpus as entire data, we have picked documents from the corpus such that only 5 prominent categories are formed. Besides this, the words that are very rare have been removed from the Reuters dataset, while the words which have no discriminative power have been removed from the Newsgroup dataset. Both term frequency and tf-idf representations of the datasets are used. The tf-idf conversion is performed before feature selection.

We have implemented a random initialization method, Duwairi and Abu-Rahmeh method (now referred as DAR) and the proposed initialization as MATLAB programs. The performance of the algorithms is measured on the basis of Adjusted Rand Index (ARI) which measures the correlation between actual labels and output cluster labels of the documents. As an indication of speed of convergence, total iterations required for convergence are also recorded.

Figure 1 shows comparison of ARI values over Reuters datasets and Fig. 2 for newsgroups datasets. The ARI values of proposed technique are higher in all datasets and maintain the level as number of features selected is decreased, while the random initialization shows much variation in the values, but never goes as high as the proposed technique. A zero value means only one cluster containing all documents was produced by DAR (Figs. 3 and 4).

As a measure of speed of convergence, we measured the number of iterations taken by algorithms to cluster the datasets. The proposed algorithm takes more time to converge than DAR technique.

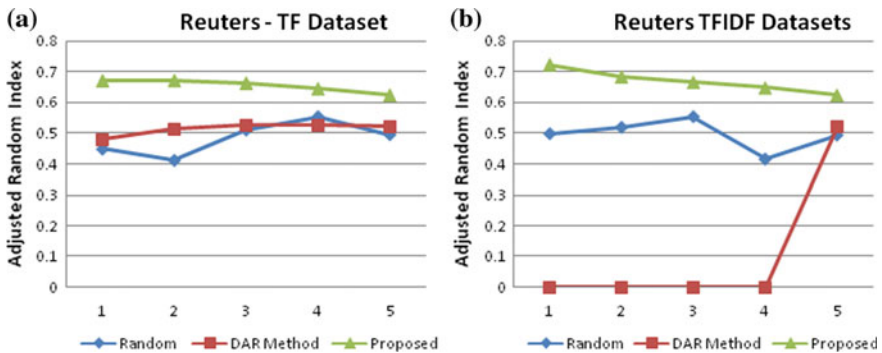


Fig. 1 ARI values for reuters datasets

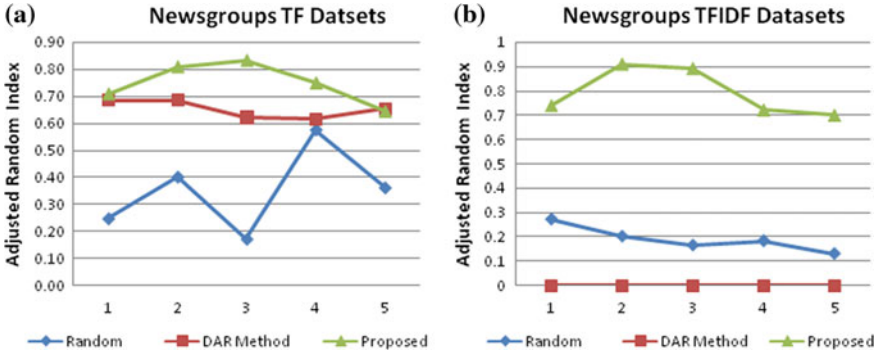


Fig. 2 ARI values for newsgroups datasets

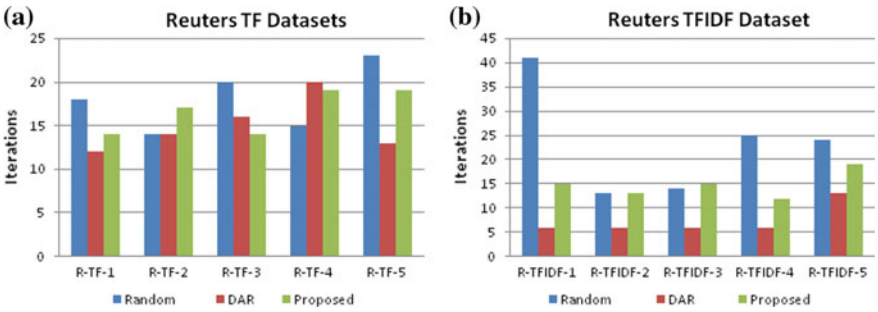


Fig. 3 Iterations for reuters datasets

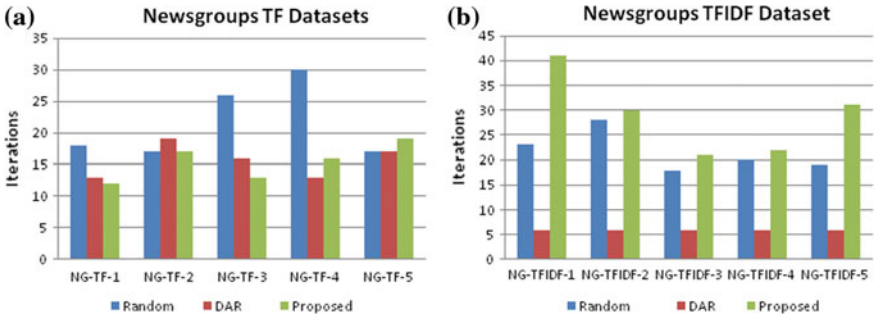


Fig. 4 Iterations for newsgroups datasets

## 7 Conclusion

Spherical k-means is an appropriate method for clustering documents as it nullifies the biasing effect of length of documents and considers the cosine similarity among documents which is by far considered the most suitable for such directional data.

Yet, it suffers from the drawback of being affected adversely by poor initializations. Besides established techniques of initializing k-means, some dedicated methods to initialize spherical k-means are also available in the literature, but very few. This paper highlights drawbacks of a recently proposed initialization method by Duwairi and Abu-Rahmeh [7] and a technique to improve this is suggested. Also, a method to avoid generation of empty clusters is proposed. Through experiments we show how Duwairi method produces empty clusters in corpus where documents are not uniformly distributed. Our method produces output of good quality even in such conditions.

## References

1. Forgy, E.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**, 768 (1965)
2. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: *Proceedings of the AAAI Workshop on AI for Web Search*, pp. 58–64 (2000)
3. Dhillon, I., Modha, D.: Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42**(1), 143–175 (2001)
4. Su, T., Dy, J.: A deterministic method for initializing K-means clustering. In: *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 784–786 (2004)
5. Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding. In: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete algorithms*, pp. 1027–1035 (2007)
6. Dhillon, I., Fan, J., Guan, Y.: Efficient clustering of very large document collections. In: *Data Mining for Scientific and Engineering Applications*, pp. 357–381. Kluwer Academic Publishers (2001)
7. Duwairi, R., Abu-Rahmeh, M.: A novel approach for initializing the spherical K-means clustering algorithm. *Simul. Model. Pract. Theory* **54**, 49–63 (2015)

# Learners' Player Model for Designing an Effective Game-Based Learning



Lamyae Bennis, Said Benhlima and M. Ali Bekri

**Abstract** Numerous opinions highlight the fact that adaptability to diverse learners' profiles and needs is an ability that is hard to afford by human teachers in enormous students classes, and as adaptation is a necessary part in education systems, as long as various dissimilarities exist among learners in terms of knowledge, abilities, favorites, and motivation. On the other hand, Game-based Learning (GBL) or educational serious game stimulates learner motivation and draws his attention to a learning subject. However, there is a huge lack of GBL authoring tool, which takes into account the learners' needs into the game design. To this end in this paper, we present the proposed new architecture and its implementation of the logical model of the chosen GBL authoring tool, also known as eAdventure2.

**Keywords** Serious game · Game-based learning · Serious game design · IMS learner information package (IMS LIP) · Public and private information for learner (PAPI)

## 1 Introduction

On July 4, 2002, a free serious game entitled “Americas Army” has been running on the Internet, developed for the army of the USA. Chen and Michael [1] defined SG as “every game whose primary purpose is other than simple entertainment”. “In the education area, games in general have been recognized to help the development of strategic thinking, planning, communication, collaborative, decision-making, and

---

L. Bennis (✉) · S. Benhlima · M. A. Bekri  
Faculty of Science, Department of Mathematics and Computer Science,  
Moulay Ismail University, Meknes, Morocco  
e-mail: lamyabennis@gmail.com

S. Benhlima  
e-mail: saidbenhlima@yahoo.fr

M. A. Bekri  
e-mail: ali.bekri@gmail.com

negotiation skills of the player” [2, 3]. There is a big lack of methods that lead to a good SG design. That is why we have suggested a new game design that integrates the learner needs in the conception of our generated learning games. To discuss this issue, we structure the paper as follows: (1) Introduction, (2) serious game design, (3) learners’ player model and profiling, (4) the standards of the learners’ player model, (5) result and implementation, (6) conclusion and future work.

## 2 Serious Game Design

Designing a serious game is a complex task as it demands consistency in implementation of two antagonistic components: playful component and the learning scenario. Researchers have proposed different serious game design methodologies to coexist both as component, for example, the model of Marfizi Schottman, the model DOD-DEL, the model KTM Advance, the EMERGO creation methodology, the generic model DICE [4]. Salen and Zimmerman [5] have defined “learning game design” as the procedure by which a conceptor produces a learning game to be used by entertainer”. On the other hand, in adaptive hypermedia systems and adaptive EIAH (IT environments for human learning), the learner has been always the main focus. Therefore, adaptability involves the integration of learner model in the system and the use of this model to adapt the navigation, content, and interaction. Currently, with the advent of informal learning (serious games), the learner is placed at the center of the educational process [6], and then the fundamental principle implemented is to estimate the needs of students to adapt teaching content, hence arise the necessity to involve learners players in SGD and specifically learners needs. The work reported in this paper follows this line of reasoning and involves learners’ player needs in game design.

## 3 Learners’ Player Model and Profiling

The learner profile draws the attention of trainers, as they have always the devotion to individualize learning. To do this, they need to extract the information of each student. Research has demonstrated the value of the learners information about the state of their knowledge, their goals, interest, preferences to help them in developing thinking skills and enhance their learning motivation [7, 8]. The learner profiles can be created at the request of various actors of the learning situation: the teacher to monitor the learning of students in the year; the institution, to follow the learning evolution of their students [9]. Furthermore, the profiles are made in order to be operated by different recipients, human, or software. Profiles created by a teacher are to be operated by the same teacher, by institutions, sometimes by the learner concerned or his family. The profiles created by a computer system are mostly intended for operation by the system itself. However, some software outsource their profiles, in order to make them

visible from the outside, mainly to the learner and the teacher [10]. Others even create profiles with the main goal to communicate with the human actors, and that is the approach adopted by the research on models of open learning [11]. The term learner model correspond to the generic modelization of learners in a computer system. The learner model is a crucial element in an adaptive intelligent tutoring system, and it allows the system to maintain a deep awareness of each learner, by the withdrawal of relevant features that could give a detailed description of its abilities, motivation, identify its level knowledge, define its interests, emotions, and learning style. Based on this studies, we have built our approach: integrating the learner player model in the game design of learning game with the purpose to have an effective learning game.

### 4 The Standard of the Learner Player Model

In the following, we present the different standards describing the learner model: PAPI (public and private information for learner) and IMS learner information package (IMS LIP).

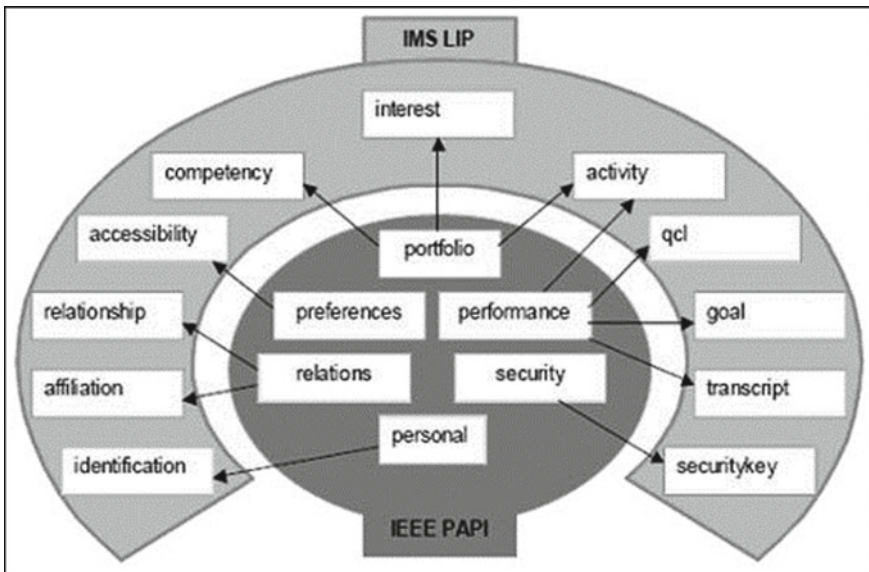


Fig. 1 The IMS learner information package (IMS LIP)

## 4.1 PAPI

The PAPI (public and private information for learner) is a combined standard that identifies the arrangement of learner data. The PAPI describes entities for registering descriptive information about: knowledge attainment, talents, capabilities, security parameter, apprentice favorites, and styles performance. The main purpose of this Standard is:

- To assist learners to construct individual learner information that can be use it through their instruction.
- Supporting manageability and portability of apprentice.
- Allowing learning content to deliver more adapted and effective learning experiences.

However, the learner data, specifically learning pedagogic, are not taken into account. This is the new evolution in IMS standard IMS LIP.

## 4.2 IMS LIP

IMS learner information package(IMS LIP)is an organized information model [12]. This one comprises both data and metadata. The IMS LIP outlines fields into which the data can be sited and the type of data that may be placed into these fields. The latter is divided into 11 basic categories (See Fig. 1). We clearly describe each component of the IMS LIP:

- The identification: It describes the demographic and biographic data learner (e.g., name, age, address, email.).
- The Purpose: It defines the purpose of the learning task, the expectation of career, and other goals.
- QCL (Qualifications): licenses and certifications describes all diplomas of the learner.
- The activity describes any activity related to learning in any execution state (e.g., training, work experience.).
- The interests maintain all information describing the learner's hobbies and recreational activities.
- Skills: It describes the skills, experience, and knowledge.
- Transcription: A file that is used to provide a summary of the school.
- Affiliation: It delivers info of membership in qualified organizations.



- Accessibility: It describes the general accessibility as: linguistic abilities, disabilities, eligibility requirements, and learning preferences.
- Security: all passwords and learner's security keys.
- Relationship: all relations between the basic elements.

## 5 Result and Implementation

### 5.1 eAdventure and Its Current Game Design

eAdventure (formerly “eAdventure”) is an open-source advanced game authoring tool, which is written in Java. It was made as a research project of “e-UCM” e-learning research group at University Complutense of Madrid. Thanks to eAdventure, everyone can create a 2D point and click conversational adventure game. This kind of game is characteristically considered more suitable for instructive settings due to the attentiveness given to exploration and reflection as opposite to time stress or fast-paced action [13]. eAdventure is developed to be easily used by novices users, through it we can generate a LG which is supported by all universal operational systems, like macOS, Windows, and Linux. The eAdventure architecture contains two main application, the eAdventure data model (the eAdventure editor) and the eAdventure engine core [14] (See Figs. 2 and 3). Through the eAdventure editor (the description of the game), the novice user can design his own LG by choosing the game components. This includes characters, items, or game scenarios but also effects triggered in the game. After the user completes the description of the game, he clicks

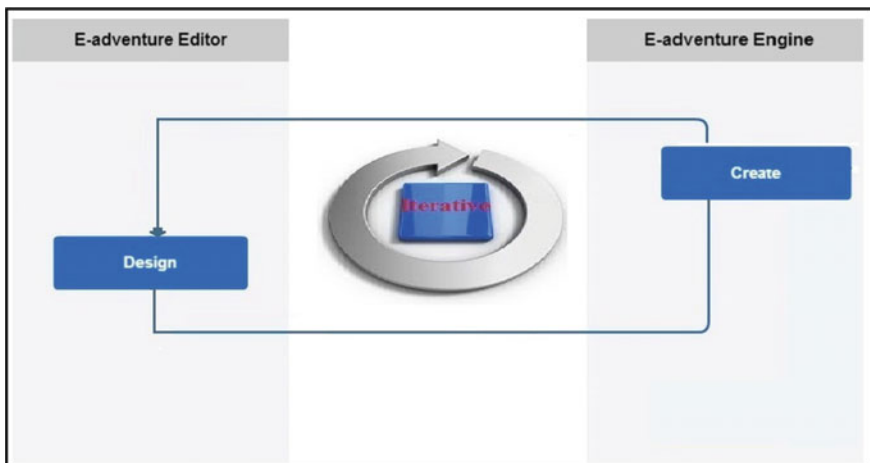
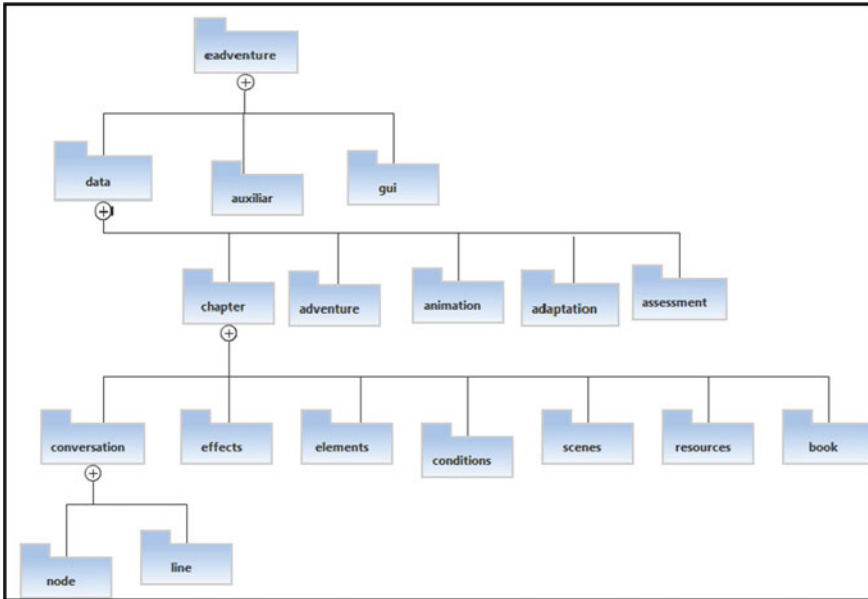


Fig. 2 The current eAdventure game design



**Fig. 3** The current logic model design of the eAdventure

the button run, and the eAdventure engine reads the EAD elements from a XML file and converts them to GameObjects in order to have a functional LG (See Fig. 2).

## ***5.2 Our Approach Based on the Proposed eAdventure Game Design and Its Implementation***

The second version of the platform has improved to respond to the needs and desires of consumer. This one was created for the development of adventure plays and was experimented by several teachers and students [15]. The use of this platform has illustrated that there are many limits that cannot be attended with the current adventure architecture; for this reason, we suggest a new structural design (see Fig. 4) where learner model was integrated into the eAdventure game editor in order to generate an adaptable and flexible LG. The new model extends the current version by adding new packages named “Learners’ need” and “adaptability” in order to generate a game that respond to learners needs (see Figs. 5 and 7). However, developing a LG authoring tool which creates a game that is amusing for players with diverse types of profiles necessitates more than integrating learner players’ personal needs into the game design but also adding a new class in our case we named “learner profile”, and the attributes of this class are defined based on IMS learner information package (IMS LIP) model (see Fig. 6) .

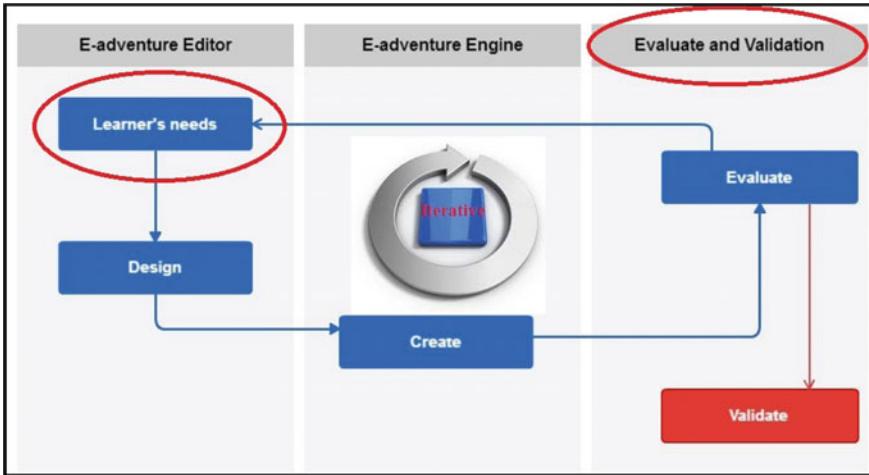


Fig. 4 The proposed new eAdventure game design

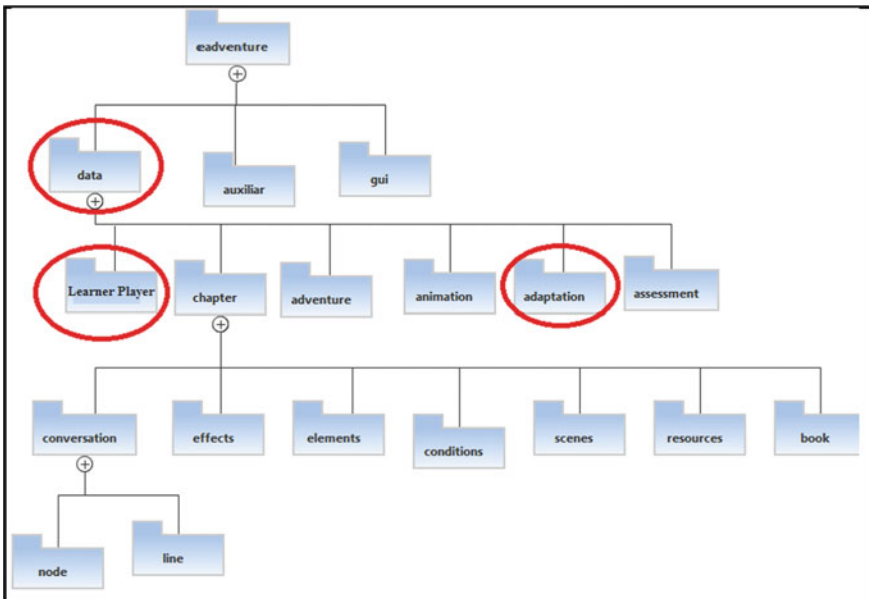


Fig. 5 The proposed new logic model design of the eAdventure

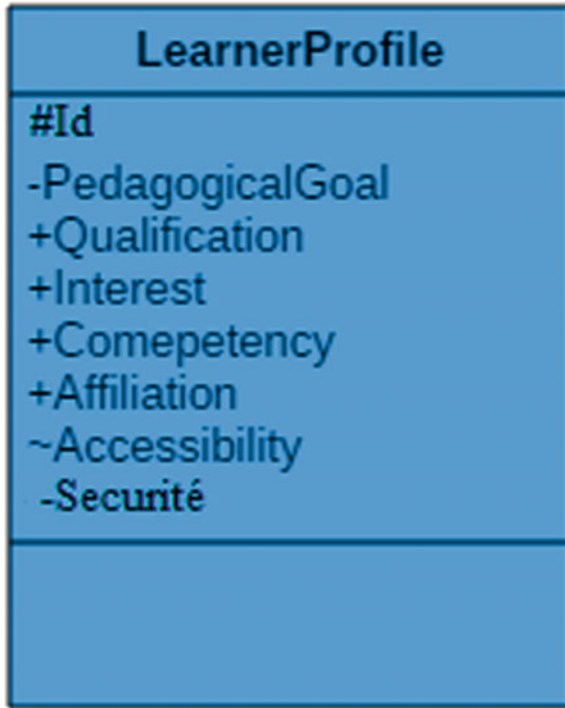


Fig. 6 Learner player profile class diagram

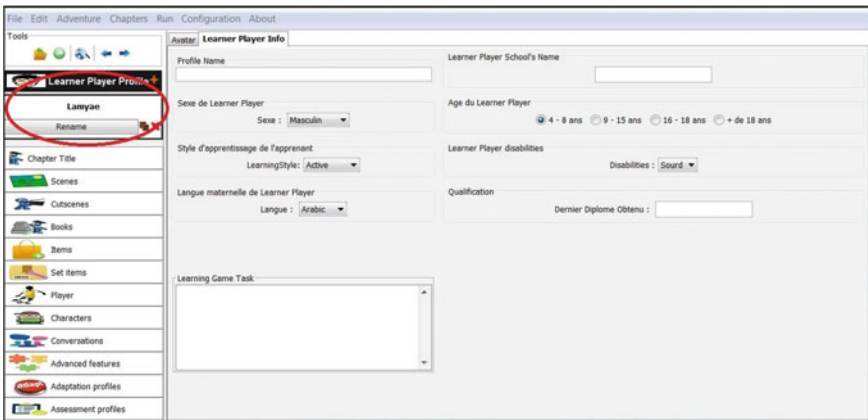


Fig. 7 Learner player profile based on IMS learner information package (IMS LIP) model

## 6 Conclusion and Future Work

This paper presents the different learner's player model and the current eAdventure game design and its disadvantages and introduces our proposed game design based on the integration of the learner model in metalayer as solution. In the near future, we are looking to develop a system that runs perfectly and allows to offer a dynamic learning adaptation-centered learners player characteristic, such as learning style and cognitive status learner, in order to meet its needs and expectations depending on its characteristics. The next step will be the creation of LG using the new eAdventure game design.

## References

1. Michael, D.R., Chen, S.L.: *Serious games: games that educate, train, and inform*. Muska & Lipman/Premier-Trade (2005)
2. Squire, K., Jenkins, H.: Harnessing the power of games in education. *Insight* **3**(1), 5–33 (2003)
3. Kirriemuir, J., McFarlane, A.: Literature review in games and learning (2004)
4. Bennis, L., Benhlima, S.: Comparative study of the process model of serious game design through the generic model dice. In: *Intelligent Systems and Computer Vision (ISCV)*, 2015, pp. 1–5. IEEE (2015)
5. Salen, K., Zimmerman, E.: *Rules of play: game design fundamentals*. MIT press (2004)
6. Laroussi, M.: *Conception et réalisation dun système hypermédia adaptatif didactique: Le système CAMELEON*. PhD Thesis, Ecole National des Sciences Informatiques. Tunis (2001)
7. Bennis, L., Benhlima, S.: Building an adaptive game-based mobile learning using the Felder-Silverman learning style model (FSLSM) approach. *Adv. Inf. Technol. Theory Appl.* **1**(1) (2016)
8. Bennis, L., Benhlima, S.: A new approach to design an attractive game based learning in various domains. *J. Theor. Appl. Inf. Technol.* **85**(3), 352 (2016)
9. Jean-Daubias, S., Eyssautier-Bavay, C., Lefevre, M., Liris, L.: Modèles et outils pour rendre possible la réutilisation informatique de profils d'apprenants hétérogènes. *Revue Sticef. org* **11**, 20 (1910)
10. Paiva, A., Self, J., Hartley, R.: Externalising learner models. In: *Proceedings of World Conference on Artificial Intelligence in Education*, pp. 509–516 (1995)
11. Bull, S., Kay, J.: Student models that invite the learner in: the smili: () open learner modelling framework. *Int. J. Artif. Intell. Educ.* **17**(2), 89–120 (2007)
12. IMS LIP. *IMS learner information package specification* (2008)
13. Garris, R., Ahlers, R., Driskell, J.E.: Games, motivation, and learning: a research and practice model. *Simul. Gaming* **33**(4), 441–467 (2002)
14. Bennis, L., Benhlima, S.: Toward a new approach: extending a game-based learning authoring tool eadventure to multiple mobile devices. In: *Europe and MENA Cooperation Advances in Information and Communication Technologies*, pp. 47–56. Springer (2017)
15. Moreno-Ger, P., Martinez-Ortiz, I., Fernández-Manjón, B.: The <e-game> project: facilitating the development of educational adventure games. In: *Cognition and Exploratory Learning in the Digital Age (CELDA 2005)*, pp. 353–358 (2005)

# Reducing Time Delay Problem in Asynchronous Learning Mode Using Metadata



Barsha Abhisheka and Rajeev Chatterjee

**Abstract** Asynchronous learning mode is a popular E-learning mode. It provides flexibility in terms of geographical location and time for learning. At present there are issues related to implementation of asynchronous E-learning techniques. A number of issues or problems are identified in this article, and their related solutions are proposed. This proposed solution is being promoted to enhance learner's interest, motivation and intern performance of the learner. A good system always has less human intervention, and the problems should be robust in nature. In this proposed research work, we have identified problems regarding time delay, for the learning material delivered such as videos. A new framework has been proposed to alleviate this problem with the help of metadata and instructional objective (IO). The objective of this work is to support proper learning application. The paper proposed a technique that shows how this problem may be resolved. Progress of performance has been shown in the result.

**Keywords** E-learning · Asynchronous learning mode  
Instructional objective (IO) · Metadata

## 1 Introduction

In the dynamic and rapid changing world, asynchronous learning mode [1, 2] has played a great role in distance education. This learning mode is not dependent of geographical distances and has little time constrains. Learners prefer asynchronous learning instead of synchronous because learners can take online courses to learn at their convenient time without hampering their normal activities.

---

B. Abhisheka (✉) · R. Chatterjee  
Department of Computer Science and Engineering, NITTTR, Kolkata, India  
e-mail: abhishekabarsha93@gmail.com

R. Chatterjee  
e-mail: chatterjee.rajeev@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_16](https://doi.org/10.1007/978-981-13-1708-8_16)

The asynchronous learning mode is gaining popularity at present. However, there are certain issues related to this mode. The proposed research article will provide a novel framework so that the issues can be handled technically.

During learning in asynchronous mode when student has doubt in video lectures by used for the studying, at that time, students have to send their queries to the expert and wait till answers are provided and time of delivering answers totally depends on expert. Hence, learner has to wait and this waiting time may diminish the interest or concentration of a learner. To provide a comprehensive solution, a new technique has been proposed and with the help of this technique learner does not have to wait for his queries, they will get immediate answers without wasting their time.

This research article is categorized as follows. Section 2 propounds related work. In Sect. 3, the overall concept of IO and metadata is discussed. In Sect. 4, the proposed methodology of video based on IO and metadata is discussed. The consequence and comparisons are discussed in Sect. 5. Section 6 presents the conclusion and future works.

## 2 Review on Existing Work

In this section we reviewed some of the related previous works.

A work by Abdelali [3] represents a formal strategy regarding web mining and web videos, using metadata-based classification and clustering procedure to deliver learners with better search results.

Wei et al. [4] showed a process to retrieve video with the help of video metadata knowledge-based method. The authors divided the key frames into grayscale distribution and probability density function of an image. Through this method learner can search the details they want exactly and in much faster way.

Agarwal et al. [5] described a technique by which users can find video of interest on YouTube. They have divided the videos into multiple labels, using their text-based metadata features to make search faster. With the help of their proposed technique, users can easily separate interested and unwanted videos on the internet.

Das and Chatterjee [6] proposed a methodology for designing the user interface framework, in order to save time to learn the system's user interface which is constructed on synchronous and asynchronous learning mode.

Podder et al. [7] described about user-friendly and good user interface design framework for synchronous and asynchronous learning to alleviate cognitive load of a student at the time of learning through mobile device.

## 3 Concept of IO and Metadata

In this section we describe about concept of IO and metadata that is being used in the framework. In section A, details about IO are given and Section B deals with metadata.

### 3.1 *Instructional Objective*

An instructional objective is a statement that provides clear direction on how learner can learn. Instructional theories focus on the architecture for boosting education of the learner [8]. The learners may in most cases want instant solutions for their problem. In most learning programs a predetermined, fixed amount of content in a set amount of time is being taught. However, the capabilities of various learners are different; they learn at different speed and have different learning requirements.

Learning materials should have enough content to understand a topic very lucidly. Bloom proposed an instructional theory [9] in the form of Bloom's taxonomy. This taxonomy defines a knowledge pyramid and divides the cognitive domain of learning into various levels. The knowledge pyramid is shown in Fig. 1.

### 3.2 *Metadata*

Metadata is defined as the data that deliver information about one or more aspects of the data; it is being provided to simplify and encapsulate primary information about data which can make searching and indexing easier [10].

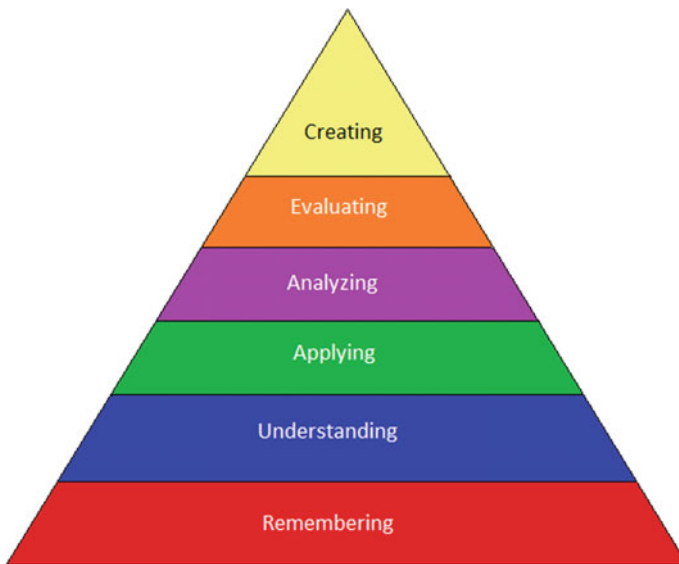


Fig. 1 Knowledge pyramid



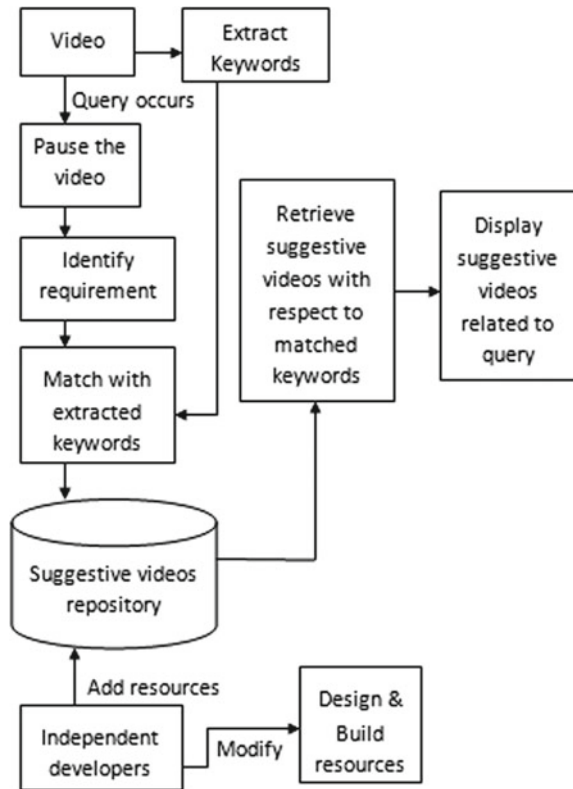
In our proposed work, we have used the concept of metadata to store keywords of a video so that we can easily retrieve required video clips. This can be done by inserting metadata elements using XML for the learning object.

## 4 Proposed Technique

### 4.1 Block Diagram

The block diagram explains the flow of process. It contains video, video pausing time, keyword extraction, requirement identification, matching technique, video repository and resources developers. Resource developers can add and modify the resource into the repository. The overall block diagram of the proposed technique is shown in Fig. 2.

Fig. 2 Block diagram for the proposed methodology



The proposed block diagram uses web video as a study material. This extracts keywords from the video and stored in XML file as a metadata. Whenever a query occurs regarding video, learner can pause the video and system checks at which IO (duration) student pauses the video. As keywords are stored with respect to IO, it matches with keywords, and at least one matching will be found, it goes to suggestive videos repository. Suggestive videos with respect to matched keywords are retrieved and displayed to learner related to their query. Developers can add, design and modify these resources according to requirement. In Sect. 4.2, the detail about this technique with the help of algorithm is discussed.

## 4.2 Algorithm

This section provides the detail technique about proposed methodology. This is divided into a number of steps. They are listed below:

- Step 1: With the help of the HTML file the default video tag with video type is declared. HTML file contains the video type and ID to extract the keywords of that particular video and also those keywords are written in Meta tag.
- Step 2: “Duration” and “Keyword” are declared in the form of metadata. The “Duration” divides the length of the video into time slot, and the “Keyword” asserts available identified keywords in the video content. These identified keywords are used for searching significant videos in the compulsion.
- Step 3: The entire video is divided into different time slot, and each slot is covered by one IO. According to the requirements each time slot assigns different keywords and it generates repository for significant videos called as Div Id. After completion of one IO, the number of div ids will create on the basis of required suggestion.
- Step 4: In the end, XML file is created which contains extracted keywords and path of significant videos. Figure 3 describes the implementation of metadata using XML.

*//This algorithm describes the working principle of video. When learner pause the //video, system stores the current time. There is some metadata stored for each time //slot, using mapping technique to read data from xml file. After successfully reading, //it matches the keywords and content of xml file. When match found it display the //suggestive videos related to queries.*

1. BEGIN/\* BEGINNING OF THE ALGORITHM\*/
  - {
2. On video PAUSE event get the current time of the video.
  - VAR TIME = vid.currentTime;
3. For each element of metadata name "DURATION"
  - BEGIN
    - {
    - if(time > 0 minTime && time <= 120 maxTime )
      - {
      - contentData = RR;
      - }
    - else if(time > 121 minTime && time <=253 maxTime)
      - {
      - contentData = SJF;
      - }
    - else if(time > 254 minTime && time <=473 maxTime)
      - {
      - contentData = FIFO;
      - }
    - else
      - {
      - contentData = Default value;
      - }
    - }
  - End
4. Use AJAX get method to read data from xml.
5. On Successful read of data from step 4.
6. Iterate through xml, search data on the basis of keyword content in xml.
  - \$(xml).find('Video').each(function ( ))
    - {
    - var data = \$(this).find("Content").text()
    - }
7. Match element of metadata and content of xml
  - if (data.indexOf(contentData) >0)
    - {
    - bind the videoID and VideoPath in HTML div to display at run time.
    - }
    - else
      - {
      - return false;
      - }
    - }
  - END

```

<?xml version="1.0" encoding="utf-8"?>
<Data>
<Video>
<Name>Video 1</Name>
<Id>vid1</Id>
<Content>RR</Content>
<VideoPath>Videos/RR example 1.mp4</VideoPath>
</Video>
<Video>
<Name>Video 2</Name>
<Id>vid2</Id>
<Content>RR</Content>
<VideoPath>Videos/RR example 2.mp4</VideoPath>
</Video>
<Video>
<Name>Video 3</Name>
<Id>vid3</Id>
<Content>SJF</Content>
<VideoPath>Videos/SJF example 1.mp4</VideoPath>
</Video>
<Video>
<Name>Video 4</Name>
<Id>vid4</Id>
<Content>FIFO</Content>
<VideoPath>Videos/Analysis of FIFO .mp4</VideoPath>
</Video>
</Data>
    
```

**Fig. 3** Implementation of metadata using XML

### 4.3 Metadata Implementation Using XML

Metadata is a very important component in multimedia system; it helps in searching and tracking of different media objects. XML is used to interchange data over network as it compatible and works on any platform. Implementation of metadata using XML is represented in Fig. 3.

## 5 Results

The performance of the proposed technique is examined with the existing technique, on the basis of the parameters as shown in Table 1. It is compared that the proposed technique has certain features over the existing system. The new technique reduces waiting time which may not be possible with the existing system. It is also maintaining the concentration and interest of the learner. It enhances the learning efficiency of learner by solving their doubts with one click of pause button. We have proposed an algorithm for reducing time to search answers of a particular learner in e-learning system.

**Table 1** Comparison of existing and proposed techniques

S. no.	Parameters	Existing technique	Proposed technique	Result
1	Reduce waiting time	No	Yes	Students do not need to wait for their answers
2	Time taken to clear doubts	Depends on expert	By clicking pause button learner can clear their doubts	Learners get their answers within fraction of second
3	Maintaining the concentration and interest of learner	No	Yes	Provide continuity and helps to prevent mind from distraction
4	Enhance learning efficiency of learners	Poor efficiency	More efficiency	Getting answers on time increases learning efficiency of learners

With the help of proposed technique learners do not need to wait for their answers; they get immediate suggestions with respect to query. It also provides continuity in learning process without any barriers and helps to prevent mind from diversion.

The proposed idea may generate a path of success for better performance in asynchronous learning mode, and it may be used as a standard guideline for many learning applications.

## 6 Conclusions and Future Work

In this research article a technique is being proposed to improve the performance as well as motivation of the learner in technical course. Analytically, it also provides improvement.

However, this research activity may not be applicable for all the test cases. For non-technical courses this may not provide the fruitful results.

In future this framework may be augmented so that it becomes content neutral.

**Acknowledgements** The authors acknowledge the support provided by the members of faculty and staff of Department of Computer Science & Engineering, NITTTTR, Kolkata, for successful conduction of research activities.

## References

1. Chuanwei, Q., Yanfei, R.: Design and research of distance education platform based on virtual reality. In: Intelligent Transportation, Big Data and Smart City, pp. 290–293. IEEE, Halong Bay Vietnam (2015)
2. Badrinath, V., Balasubramanian, S.: Learners' preferences and influencing parameters in e-learning. In: Management Issues in Emerging Economies, IEEE. Thanjavur Tamilnadu India, pp. 74–78 (2012)
3. Abdelali, S.: Education data mining: mining MOOCs videos using metadata based approach. In: Information Science and Technology, IEEE, pp. 531–534. Tangier Morocco (2016)
4. Wei, X., Shen, W., Jiang, S.: A novel algorithm for video retrieval using video metadata information. In: Education Technology and Computer Science, IEEE, vol. 2, pp. 1059–1062. Wuhan Hubei China (2009)
5. Agarwal, N., Gupta, R., Singh K.S.: Metadata based multi-labelling of YouTube videos. Cloud Computing, Data Science & Engineering, IEEE, pp. 697–711. Noida, India (2017)
6. Das, S., Chatterjee, R.: A proposed systematic user-interface design framework for synchronous and asynchronous e-learning systems, Springer India. In: Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, vol. 340, pp. 337–347. New Delhi India (2015)
7. Podder, A., Bhadra, T., Chatterjee, R.: User-interface design framework for e-Learning through mobile devices, Springer India. In: Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, vol. 434, pp. 227–236. New Delhi (2016)
8. Wikipedia, “Instructional theory”. [https://en.wikipedia.org/wiki/Instructional\\_theory](https://en.wikipedia.org/wiki/Instructional_theory). Accessed 11 Jan 2017
9. Bhargav, H.S., Akalwadi, G., Pujari, N.V.: Application of blooms taxonomy in day-to-day examinations. Adv. Comput. IEEE, 825–829 (2016)
10. Liu, Y., Huang, T., Zhao, J.: Study and application of metadata management based on XML. In: Genetic and Evolutionary Computing, IEEE, pp. 252–255. Guilin, China (2009)

# Improved Forecasting of CO<sub>2</sub> Emissions Based on an ANN and Multiresolution Decomposition



Lida Barba and Nivaldo Rodríguez

**Abstract** The sustainability of the environment is a shared goal of the United Nations. In this context, the forecast of environmental variables such as carbon dioxide (CO<sub>2</sub>) plays an important role for the effective decision making. In this work, it is presented multi-step-ahead forecasting of the CO<sub>2</sub> emissions by means of a hybrid model which combines multiresolution decomposition via stationary wavelet transform (SWT) and an artificial neural network (ANN) to improve the accuracy of a typical neural network. The effectiveness of the proposed hybrid model SWT-ANN is evaluated through the time series of CO<sub>2</sub> per capita emissions of the Andean Community (CAN) countries from 1996 to 2013. The empirical results provide significant evidence about the effectiveness of the proposed hybrid model to explain these phenomena. Projections are presented for supporting the environmental management of countries with similar geographical features and cultural diversity.

**Keywords** Carbon dioxide · Multiresolution decomposition  
Stationary wavelet transform · Artificial neural network · Forecasting

## 1 Introduction

The carbon dioxide emissions are part of the threats that affect the environment. One of the Millennium Development Goals of the United Nations declares the incorporation of principles of sustainability development into the policies and programs of the nations. Unfortunately, according to the data located in the repositories of the World Bank Group [1], the carbon dioxide emissions present an upward trend. In 2011,

---

L. Barba (✉)

Facultad de Ingeniería, Universidad Nacional de Chimborazo, 060102 Riobamba, Ecuador  
e-mail: lbarba@unach.edu.ec

N. Rodríguez

Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, 2362807 Valparaíso, Chile

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_17](https://doi.org/10.1007/978-981-13-1708-8_17)

177

32.3 billions of metric tons of CO<sub>2</sub> emissions were observed, which were increased to 48.9 in comparison with the emissions in year 1990.

Several investigations have determined that high CO<sub>2</sub> emissions increase the plant photosynthesis and reduce the transpiration [2]. The studies of Tao et al. [3] show that the effects of CO<sub>2</sub> vary with the temperature, water availability and solar radiation. The simulations show that in 2020 as effect of the wheat productivity in China the CO<sub>2</sub> emissions will increase significantly, while it will decrease by the increase of O<sub>3</sub>. Consequently, interactive and not only negative effects on climate changes are observed through carbon dioxide emissions.

Given the importance of analysis related to behavioral patterns of the carbon dioxide emissions, various forecasting works have been developed with the aim of providing useful projections to improve decision making. For example, Pérez-Suárez and López-Menéndez [4] present the CO<sub>2</sub> forecast of 150 countries based on the Kuznets environmental curve. The study shows an explained variance over 80% for 78 countries, including the CAN members (Ecuador, Colombia, Peru and Bolivia), and an absolute average percent error near of 7%. On the other hand, Pao and Tsai [5] applied the Gray model in comparison with the ARIMA model to predict the total CO<sub>2</sub> emissions in Brazil. The study presents MAPEs (average absolute percentage error) among 2.46 and 4.22%. Wu et al. [6] presented the forecast of CO<sub>2</sub> emissions for BRICS countries (Brazil, Russia, India, China and South Africa) by means of the Gray model, the study showed the relationship among the GDP and the energy with respect to the CO<sub>2</sub> emissions. The prediction reached average MAPEs of 2.36%.

In this paper, it is presented a hybrid model based on the stationary wavelet transform and an artificial neural network to improve the average accuracy observed in the works cited previously and the accuracy reached by a typical neural network. The model is evaluated through the annual time series from 1996 to 2013 of the CO<sub>2</sub> emissions of the Andean Community countries (Colombia, Ecuador, Peru and Bolivia) located in the repositories of the World Bank Group.

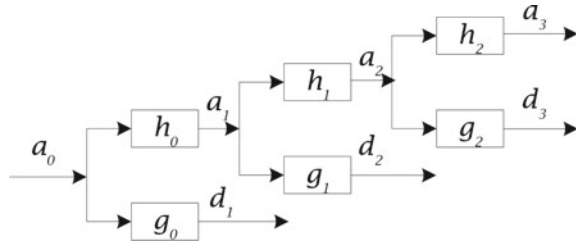
The article is organized as follows. Stationary wavelet transform and the artificial neural network are explained in Sect. 2. The forecasting accuracy metrics are explained in Sect. 3. Case Studies are shown in Sect. 4. Results and Discussion are described in Sect. 5. Finally, Conclusions close the work in Sect. 6.

## 2 Forecasting Methodology

The forecasting methodology is based on multiresolution decomposition via stationary wavelet transform and prediction through an artificial neural network.



**Fig. 1** Decomposition scheme of SWT with  $J = 3$



### 2.1 Decomposition Based on Stationary Wavelet Transform

Stationary wavelet transform is applied for decomposing a discrete time series in coefficients of approximation and detail. The implementation processes of SWT are described in the algorithm of Shensa [7]. SWT is based on discrete wavelet transform [8], but the down-sampling procedure is omitted and the filters are up-sampled [9]. The up-sampling process gets components that have length equal to the original signal.

In SWT, the length of the observed signal must be an integer multiple of  $2^j$ , where  $j = 1, 2, \dots, J$  is the scale number. The signal is separated in approximation coefficients and detail coefficients at different scales, this hierarchical process is called multiresolution decomposition [10].

The filtering process uses low pass filters and high pass filters, and each one is used at different decomposition levels, as it is shown in Fig. 1. At first decomposition level, the observed time series  $a_0$  is convoluted with low pass filter  $h_0$ , then the first approximation signal  $a_1$  is obtained. At the same first decomposition level, the high pass filter  $g_0$  is applied to obtain the first detail signal  $d_1$ . The filtering process at the first level is illustrated as follows

$$a_1(n) = \sum_i h_0(i)a_0(n - i), \tag{1}$$

$$d_1(n) = \sum_i g_0(i)a_0(n - i), \tag{2}$$

The next decomposition levels  $j = 1, \dots, J - 1$  obtain new signals of approximation and detail; it is given as

$$a_{j+1}(n) = \sum_i h_j(i)a_j(n - i), \tag{3}$$

$$d_{j+1}(n) = \sum_i g_j(i)a_j(n - i), \tag{4}$$

SWT obtains sub-bands of frequency, the approximation signal obtained in the last level and the detail signals must be reconstructed by means of inverse stationary wavelet transform (iSWT). The implementation of iSWT consists in applying a set

of reconstruction filters in inverse order. The component of low frequency  $c_L$  is computed with the last approximation signal  $a_J$ , whereas the component of high frequency  $c_H$  is computed with the addition of all reconstructed detail signals  $d_j$ .

## 2.2 Prediction Based on an Artificial Neural Network

Some forecasting solutions based on artificial neural networks (ANN) have been observed in diverse areas of knowledge. ANNs have demonstrated high capability of approximation and universal generalization of nonlinear problems [11, 12]. The effective calibration of an ANN contributes with its convergence. Diverse approaches present variations in the transfer and activation functions [13, 14], time delay [15], number of hidden nodes [16] or modifications of the learning algorithm [17].

A conventional MLP of three layers is implemented and improved by the use of components as inputs instead of raw data. This strategy avoids the setting processes previously described related to the structure. The ANN uses the lagged terms  $z_i$  of the components at the input layer, they are weighted with respect to the hidden layer, and at the output of the hidden layer is applied the activation function  $f(\cdot)$ :

$$\hat{x}(n+1) = \sum_{j=1}^Q b_j Y_{Hj}, \quad (5)$$

$$Y_{Hj} = f\left(\sum_{i=1}^P w_{ji} z_i\right), \quad (6)$$

where  $\hat{X}(n+1)$  is the predicted value,  $w_{11}, \dots, w_{P1}, \dots, w_{PQ}$  are the nonlinear weights of the connections between the inputs and the hidden neurons. Whereas  $b_1, \dots, b_Q$  are the weights of the connections between the hidden neurons and the output (under the assumption that there is a unique output). In this case, the common activation function is logistic ( $f(x) = 1/(1 + e^{-x})$ ).

## 3 Forecasting Accuracy Metrics

The accuracy of the prediction is computed with the metrics: mean absolute percentage error (MAPE), root mean squared error (RMSE) and the modified Nash–Sutcliffe efficiency (mNSE).

$$MAPE = \left[ \frac{1}{N_v} \sum_{i=1}^{N_v} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \right] \times 100 \quad (7)$$

$$RMSE = \sqrt{\frac{1}{N_v} \sum_{i=1}^{N_v} (x_i - \hat{x}_i)^2} \quad (8)$$

where  $N_v$  is the testing sample size,  $x_i$  is the  $i$ -th observed value and  $\hat{x}_i$  is the  $i$ -th estimated value.

$$mNSE = 1 - \frac{SAE}{SAD}. \quad (9)$$

where  $SAE$  and  $SAD$  are defined with

$$SAE = \sum_{i=1}^N |x_i - \hat{x}_i|, \quad (10)$$

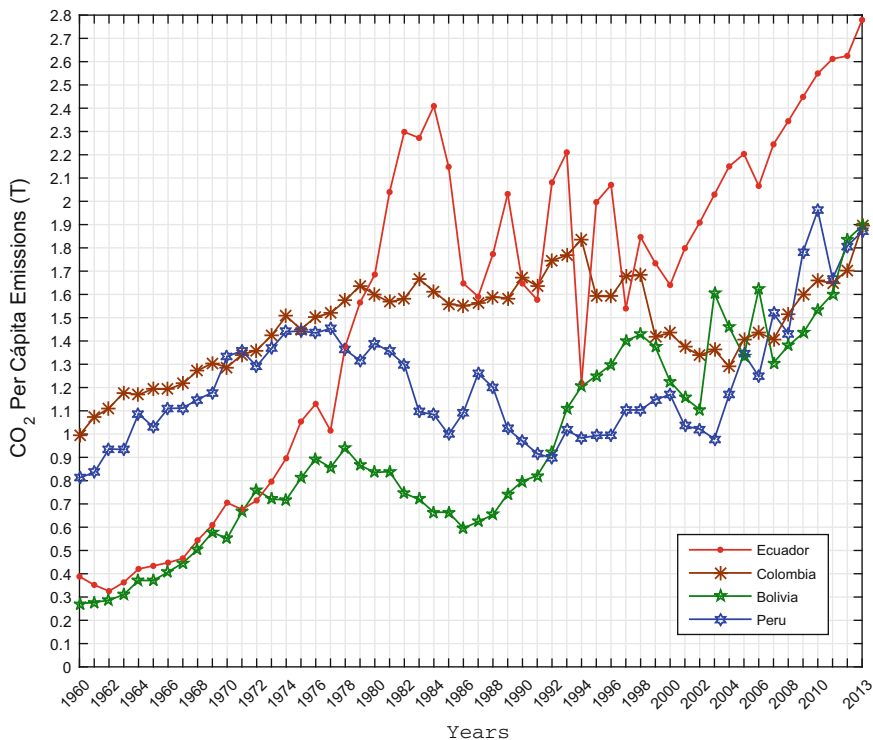
$$SAD = \sum_{i=1}^N |x_i - \bar{x}|, \quad (11)$$

## 4 Case Studies

The open repositories of the World Bank Group contain development data of several countries and a variety of topics. Among the time series are those related to carbon dioxide emissions in metric tons per capita of the countries.

The CO<sub>2</sub> emissions per capita of the four countries members of the Andean Community: Ecuador, Colombia, Bolivia and Peru are presented in Fig. 2. The presented values are calculated by means of the ratio between the total CO<sub>2</sub> emissions and the population of each country. In all cases, the samples have an annual collection interval, with records from year 1960 to 2013.

The emissions in the last decade show an upward trend in the four CAN countries. In the case of Ecuador, there is a considerable growth from 1977 with several peaks until 1998. From the year 2000, a more linear behavior, similar to 1960–1976, is observed. CO<sub>2</sub> emissions from Colombia, Peru and Bolivia show similar behavior in terms of variability, which is most evident in recent decades. Table 1 shows statistical and dispersion measurements of the observed data. The highest arithmetic mean of emissions is observed for Ecuador, followed by Colombia, Peru and Bolivia. The maximum value is reached by Ecuador with 2.779 metric tons, followed by Peru, Bolivia and Colombia with 1.961, 1.895 and 1.893 metric tons, respectively. In terms of dispersion measures, it is observed that Ecuador has a historical behavior of greater variability, with a standard deviation of 0.737 and a variance of 0.533, followed by Bolivia with a standard deviation of 0.429 and a variance of 0.181, while Colombia and Peru show a minimum variance of 0.039 and 0.068, respectively.



**Fig. 2** Annual emissions of carbon dioxide (metric tons)

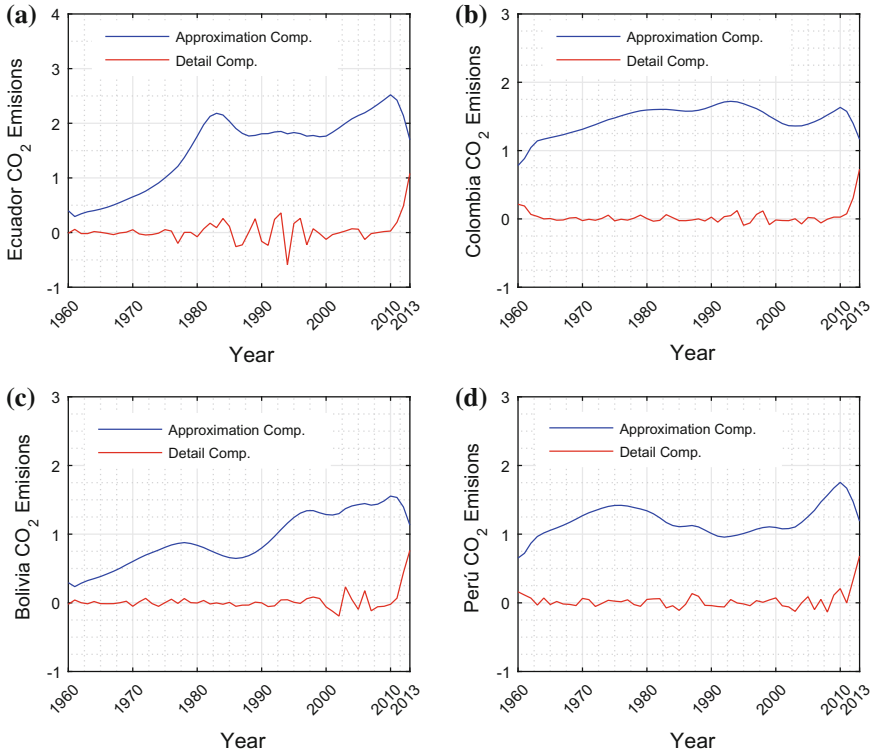
**Table 1** Statistical analysis of data

	Min	Max	Mean	$\sigma$	$\sigma^2$
Ecuador	0.325	2.779	1.546	0.737	0.533
Colombia	0.996	1.893	1.479	0.200	0.039
Bolivia	0.272	1.895	0.940	0.429	0.181
Perú	0.812	1.961	1.221	0.262	0.068

## 5 Results and Discussion

### 5.1 Components Extraction

The components of approximation and detail were extracted through the application of SWT through a Haar function with two decomposition levels  $J = 2$ . The components of the time series of Ecuador are shown in Fig. 3a, while for the rest of series are shown in Fig. 3b–d for Colombia, Bolivia and Perú, respectively. The approx-



**Fig. 3** Approximation components and detail components extracted via SWT

imation components show long-duration fluctuations, while the detail components show short-duration fluctuations.

### 5.2 Prediction

The prediction of CO<sub>2</sub> emissions per capita was developed through the ANN model which was described in Forecasting Methodology Section. The learning algorithm Levenberg–Marquardt was implemented for the weights adjusting [18, 19]. The number of inputs of the ANN model in all cases was set in  $P = 12$ , in attention to the information given by the fast Fourier transform algorithm [20]. The periodogram shows relevant periods of 12 years at 5% of significance level. The inputs of the ANN model were the lagged values of the SWT components. The number of hidden nodes was chosen after trial-and-error tests, and only one level was enough for reaching the lowest error. The output is the tons of carbon emissions per capita for the next year of each country. Consequently the ANNs were denoted with (12, 1, 1).

Table 2 shows the results of the forecast by means of the testing sample, which corresponds to 30% of the observed data. The evaluation is performed by calculating the efficiency metrics MAPE, RMSE and mNSE. The results show high accuracy, with MAPE values lower than 1%, average MSEs of 0.005 and average efficiencies of 97.7% for multi-year-ahead forecasting. The highest average accuracy is achieved with data from Bolivia with an average mNSE of 0.2% and an average mNSE of 98.3%, followed by Perú with an average mNSE of 98.33%, and Colombia with an average mNSE of 96.66%, and Ecuador with an average mNSE of 97.4%.

The conventional ANN model, which is not based on components, shows lower accuracy with respect to the hybrid SWT-ANN proposed model. In this solution, the data have been basically preprocessed by means of a conventional moving average smoothing. The multi-year-ahead forecasting results of the four series of CO<sub>2</sub> emissions are shown in Table 3; extended forecast horizons present poor results.

The highest accuracy by means of the typical ANN (Table 3) was obtained with the forecast of Perú emissions for one-step-ahead forecasting, with a MAPE of 1.55%, a RMSE of 0.0142 and a mNSE of 91.39%. The lowest accuracy was obtained for the time series of Ecuador.

From Tables 2 and 3, it is observed that the best results were reached by the application of the proposed forecasting model. Both models present the best results after 30 iterations (also implies poor results). SWT-ANN shows high accuracy for three-step-ahead forecasting, whereas conventional ANN obtains good accuracy only for one-step-ahead forecasting of Peru emissions. In that case, the gain of SWT-ANN over the conventional ANN model is of 9.3% for mNSE. The observed and predicted values via SWT-ANN hybrid model for three-step-ahead forecasting related to the testing sample are presented in Fig. 4. From Figure, it was observed a good fit.

## 6 Conclusions

In this work, it was presented the forecast of CO<sub>2</sub> emissions of four countries with similar conditions in terms of geographic and cultural diversity. The forecasting methodology was based on components hierarchically extracted from the observed time series and a conventional artificial neural network which inputs were those components. The results obtained with the testing sample demonstrated that the SWT-ANN method improves the accuracy of the conventional ANN model as well as the accuracy level of other approaches observed in the literature review. The average accuracy achieved for three-year-ahead forecasting via testing sample was of 0.22% for MAPE, 0.0054 for RMSE and 97.75% for mNSE. Extended horizons present a significative decreasing of accuracy.

A conventional ANN presents lower accuracy with an average MAPE of 8.14%, an average RMSE of 0.079 and an average mNSE of 43.7% for three-year-ahead forecasting, extended forecast horizons present poor results.

Given the effectiveness of the method, new forecasting simulations will be performed with time series coming from other countries and other areas of knowledge.

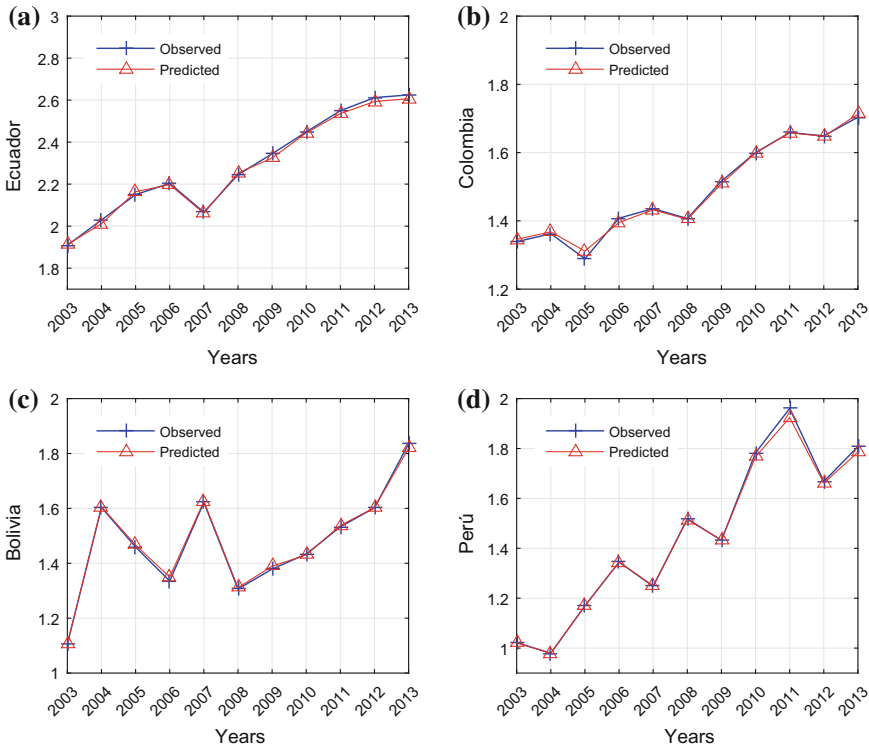
**Table 2** Prediction results via SWT-ANN

	One step			Two steps			Three steps		
	MAPE (%)	RMSE	mNSE (%)	MAPE (%)	RMSE	mNSE (%)	MAPE (%)	RMSE	mNSE (%)
Ecuador	0.0096	0.0003	99.9	0.177	0.0051	98.23	0.513	0.0134	94.21
Colombia	0.0042	0.0001	99.9	0.40	0.0071	95.36	0.469	0.009	94.72
Bolivia	0.0074	0.0003	99.9	0.201	0.0064	98.73	0.456	0.014	97.0
Perú	0.0078	0.0001	99.9	0.076	0.0017	99.27	0.421	0.008	95.84
Min	0.0042	0.0001	99.9	0.076	0.0017	95.36	0.421	0.008	94.21
Mean	0.0072	0.0002	99.9	0.213	0.005	97.89	0.465	0.011	95.44

**Table 3** Prediction results via conventional ANN

	One step			Two steps			Three steps		
	MAPE (%)	RMSE	mNSE (%)	MAPE (%)	RMSE	mNSE (%)	MAPE (%)	RMSE	mNSE (%)
Ecuador	6.075	0.0549	41.16	7.397	0.0706	17.89	12.231	0.1123	-
Colombia	1.611	0.0146	72.12	3.875	0.0379	47.53	7.048	0.063	20.11
Bolivia	3.287	0.0318	66.53	11.25	0.0969	-	15.542	0.1494	-
Perú	1.553	0.0142	91.39	11.373	0.1289	33.84	16.424	0.1827	40.33
Min	1.553	0.0142	41.16	3.875	0.0379	17.89	7.048	0.063	20.11
Mean	3.131	0.0288	67.8	8.473	0.0835	33.087	12.811	0.126	30.22





**Fig. 4** Prediction results of CO<sub>2</sub> per capita emissions

**Acknowledgements** Thanks to Animal Production and Industrialization (PROANIN) Research Group of the Universidad Nacional de Chimborazo for supporting this work through the project Artificial Neural Networks to predict the carcass tissue composition of guinea pigs.

## References

1. World Bank Group repository. <http://databank.worldbank.org/data/home.aspx> (2017)
2. Kimball, B.A., Pinter Jr., P.J., Garcia, R.L., LaMorte, R.L., Wall, G.W., Hunsaker, D.J., Wechsung, G., Wechsung, F., Kartschall, T.: Productivity and water use of wheat under free-air CO<sub>2</sub> enrichment. *Glob. Change Biol.* **1**(6), 429–442 (1995)
3. Tao, F., Feng, Z., Tang, H., Chen, Y., Kobayashi, Z.: Effects of climate change, CO<sub>2</sub> and O<sub>3</sub> on wheat productivity in Eastern China, singly and in combination. *Atmos. Environ.* **153**, 182–193 (2017)
4. Pérez-Suárez, R., López-Menéndez, A.: Growing green? Forecasting CO<sub>2</sub> emissions with environmental Kuznets Curves and logistic growth models. *Environ. Sci. Policy* **54**, 428–437
5. Pao, H-T., Tsai C-M.: Modeling and forecasting the CO<sub>2</sub> emissions, energy consumption, and economic growth in Brazil. *Energy* **36**, 2450–2458

6. Wu, L., Liu, S., Liu, D., Fang, Z., Xu, H.: Modelling and forecasting CO<sub>2</sub> emissions in the BRICS (Brazil, Russia, India, China, and South Africa) countries using a novel multi-variable grey model. *Energy* **79**, 489–495 (2015)
7. Shensa, M.: The discrete wavelet transform: wedding the a Troun and Mallat algorithms. *IEEE Trans. Signal Process.* **40**(10), 2464–2482 (1992)
8. Grossmann, A., Morlet, J.: Decomposition of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.* **15**(4), 723–736 (1984)
9. Nason, G., Silverman, B.: Wavelets and statistics, the stationary wavelet transform and some statistical applications. In: *Wavelets and Statistics*, pp. 281–299. Springer, New York (1995)
10. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11**(7), 674–693 (1989)
11. Hornik, K., Stinchcombe, X., White, H.: Multilayer feedforward networks are universal approximators. *Neural Netw.* **2**(5), 359–366 (1989)
12. Svozil, D., Kvasnicka, V., Pospichal, J.: Introduction to multi-layer feed-forward neural networks. *Chemometr. Intell. Lab. Syst.* **39**(1), 43–62 (1997)
13. Rojas, I., Pomares, H., Bernier, J.L., Ortega, J., Pino, B., Pelayo, F.J., Prieto, A.: Time series analysis using normalized PG-RBF network with regression weights. *Neurocomputing* **42**(1–4), 267–285 (2002)
14. Roh, S.B., Oh, S.K., Pedrycz, W.: Design of fuzzy radial basis function-based polynomial neural networks. *Fuzzy Sets Syst.* **185**(1), 15–37 (2011)
15. Liu, F., Ng, G.S., Quek, C.: RLDDE: A novel reinforcement learning-based dimension and delay estimator for neural networks in time series prediction. *Neurocomputing* **70**(7–9), 1331–1341 (2007)
16. Scarselli, F., Chung, A.: Universal approximation using feedforward neural networks: a survey of some existing methods, and some new results. *Neural Netw.* **11**(1), 15–37 (1998)
17. Gheyas, I.A., Smith, L.S.: A novel neural network ensemble architecture for time series forecasting. *Neurocomputing* **74**(18), 3855–3864 (2011)
18. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Quart. J. Appl. Math.* **2**(2), 164–168 (1944)
19. Marquardt, D.: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**(2), 431–441 (1963)
20. Hahn, B., Valentine, D.: *Essential MATLAB for engineers and scientists*, 6th edn, pp. 333–339. Academic Press, Elsevier (2013)

# Clustered Support Vector Machine for ATM Cash Repository Prediction



Pankaj Kumar Jadwal, Sonal Jain, Umesh Gupta and Prashant Khanna

**Abstract** Optimal prediction of cash in ATMs is a critical task. This research paper is concerned with the application of cash requirement forecasting of NN5 dataset by most promising machine learning technique support vector machine (SVM). Primary objective of this research paper is time series prediction of NN5 data with support vector regression at the first stage and further root mean square error (RMSE) is computed. Furthermore, the same study was conducted by clustering ATMs using k means clustering technique on NN5 data before applying support vector regression. Root mean square error (RMSE) is calculated for the clusters of ATMs, and average of RMSE retrieved from clusters is compared with accuracy obtained from single baseline SVM. RMSE indicates the application of unsupervised learning (clustering) used as a preprocessing step towards increases precision in the prediction of cash in ATMs.

**Keywords** Preprocessing · Clustering · Prediction · Support vector machine

---

P. K. Jadwal (✉)  
JK LakshmiPat University, Jaipur, India  
e-mail: pankajjadwal@gmail.com

S. Jain  
Department of Computer Science Engineering, JK LakshmiPat University, Jaipur, India  
e-mail: sonaljain@jklu.edu.in

U. Gupta  
Department of Mathematics, JK LakshmiPat University, Jaipur, India  
e-mail: umeshgupta@jklu.edu

P. Khanna  
Wintec, Hamilton, New Zealand  
e-mail: perukhan@gmail.com

## 1 Introduction

The banking industry is the backbone of the economy of any country. Major banks are having a competition among them to obtain most of the customers and financial transactions. Banks are trying hard to attract and retain their customers. ATMs are the most prominent medium of distribution of cash between clients and server (banks). The quantity of clients in banks is raising quickly so it is apparent that quantity of ATMs should be improved. After setting up and opening ATMs, prediction of optimal usage of cash is definitely an important and essential concern. If the total amount of money having the ATM is more than the requirement of the customer, then unused cash will be there and security concern will arise and if the amount of money in that ATM is much lower than the requirement, after that it will result into client dissatisfactions. Studies of ATMs cash replenishment focuses on options regarding time frames that every ATM ought to be replenished along the total cash that ought to be filled. Optimal prediction of cash in ATMs should be there so that a balance may be created between both sides.

Support vector machine (SVM) is probably the most famous and trusted machine learning model that may be used for both classification and regression perspective. Different machine learning and statistical techniques have been applied on ATM cash withdrawal data for optimal cash prediction in the last decade. Some of the statistical techniques are exponential smoothing (ES) and autoregressive integrated moving average (ARIMA) and some famous supervised models are support vector regression (SVR) and artificial neural network (ANN). These machine learning algorithms may be utilized for linear and nonlinear function approximations. These techniques vary in their accuracy, prediction efficiency, robustness and transparency [1].

## 2 Review of Literature

There are two domains where research has been done for cash demand forecasting. In the first domain, researchers work on demand forecasting at everyday level. In second domain, studies have been done on cash replenishment. This work is concerned with first domain. As per research in the first domain, the journey starts from forecasting competition (NN5). The motive behind organizing NN5 competition was to assess the precision of computational intelligence (CI) strategies in the forecasting of time series. This competition made this problem (Cash demand forecasting) very popular. Different researchers came with different ideas for getting the optimal solution to this problem. The accuracy of the approaches proposed by the researchers was measured by MAPE. The data contain daily cash withdrawals from different 111 ATMs which were located in different locations of England. The dataset was separated into the training dataset and testing dataset. Training data contain transactions of 2 years.

The objective of the competition was to forecast cash withdrawal from different 111 ATMs for last 56 days. Researchers [2] got the first place among all computational models. They suggested the final model consisting average of the predicted output of different models like linear model, Gaussian process regression and neural networks. Researchers applied different algorithms on the NN5 dataset to obtain optimal results.

A hybrid model [3] was proposed which is the cascaded group of neural network model and nearest trajectory ensembles. Self-organizing fuzzy network model [4] was employed on the NN5 dataset in order to obtain more prediction accuracy with a single model.

A novel machine learning model [5] (PSECMAC) was introduced to generate more precise outputs from the NN5 dataset. A novel model for multi-step-ahead forecasting [6] was also proposed and seasonality effects were also considered in that model.

Trafalis and Ince [7] compared SVR with radial basis functions and traditional neural network architectures to predict stock price indices and illustrated that SVR has worked better than neural networks. Tay and Cao [8] elaborated the utility of SVR for forecasting of five particular financial time series (S&P500 and a number of international bond indices particularly). SVR and backpropagation neural networks were applied on the data sets and compared the results on the basis of mean square error (NMSE) and mean absolute error (MAE). Tay and Cao [9] suggested an improved variant of SVR for the forecasting of financial series known as C-ascending SVMs. The overall performance of prediction is examined based on normalized mean squared error. Tay and Cao came to the conclusion that overall better performance can be achieved using this technique as in comparison with a regular SVR implementation.

Cao et al. [10] suggested another adaptive strategy and customization to the SVM called as descending SVM for modelling of time series which is not stationary. According to authors, there are two advantages of using the technique. First one is the superior performance and the second one is a sparser solution. Van Gestel et al. [11] recommended the integration of an LS-SVM utilized in a Bayesian framework. They applied point time series forecasting and volatility machine learning models had been created for forecasting of the economical stock index. A minor advancement in mean square error, mean absolute error and negative log likelihood (NLL) was identified using this technique. Pai et al. [12] proposed a seasonal support vector regression (SSVR) technique to predict time series dataset having seasonality.

Hung et al. [13] proposed a hybridized model, which integrates support vector regression with the classical moving average model for the prediction of cash withdrawals from ATM placed at various locations in England. Harris [14] proposed the use of clustering as a preprocessing step towards support vector machine for credit risk evaluation. The author compared clustered SVM with nonlinear SVM-based techniques in terms of performance and showcased better performance. P. Khanarsa et al. proposed multiple ARIMA subsequences aggregation (MASA) model [15] and model outperforms SARIMA and ETS exponential smoothing model on the basis of SMAPE.

Neural networks along with hierarchical clustering are used for cash withdrawal forecasting from the dataset (NN5) and clustered neural network outperformed neural network (multilayer perceptron) in terms of RMSE [16].

### 3 Data Set

In this research paper, NN5 dataset [17] is used which is taken from UCI repository. The dataset contains money transactions statistics of two years from various ATMs situated at various places in England. The motive behind organizing the NN5 competition was to forecast transactions of 101 ATMs, situated in various locations in England. All ATMs contain the transaction data of two years, and the prediction horizon is from 23 March 1998 to 17 May 1998. There are a couple of lacking figures in the data.

### 4 Flowchart of Proposed Approach

Figure 1 shows the proposed approach regarding flowchart.

### 5 C-SVM (Clustered SVM) Algorithm

*Input*-NN5 complete dataset

- Step I: **Preprocessing of the dataset:** There are some missing values in the dataset. Preprocessing of the dataset is done via replacing missing values with trend estimated values.
- Step II: **Normalization of the dataset:** Next process in the normalization of the dataset. In normalization process, the minimum transaction value of time series is used as centre and subtraction of maximum transaction amount with minimum transaction amount is used as the range of dataset.
- Step III: **De-seasonalization of normalized dataset:** Normalized dataset is de-seasonalized. Time series of all ATMs is converted into “day of the week” cash withdrawal seasonality parameters.
- Step IV: **Division of dataset into training and testing dataset:** NN5 dataset is separated into training and testing set. Dataset from 18 March 1996 to 22 March 1998 is used as training dataset and dataset from 23 March 1998 to 17 May 1998 is used as the testing set.
- Step V: **Repeat step 5 for all ATM’s time series from NN5001 to NN5101**
  - (a) **Model Creation:** Support vector regression is implemented on preprocessed, normalized and de-seasonalized dataset. Optimal values for

cost and epsilon is to be chosen. The darkest shade of plot represents the optimal value of cost and epsilon, shown in Fig. 2.

- (b) **Tuning of Model:** For getting optimal values of cost and epsilon, tuning of the model is done.
- (c) **Accuracy testing via Root Mean Square Error:** Accuracy of model is tested via applying model on testing dataset and RMSE (root mean square error) is calculated. Table 1 shows RMSE of all time series of the dataset.

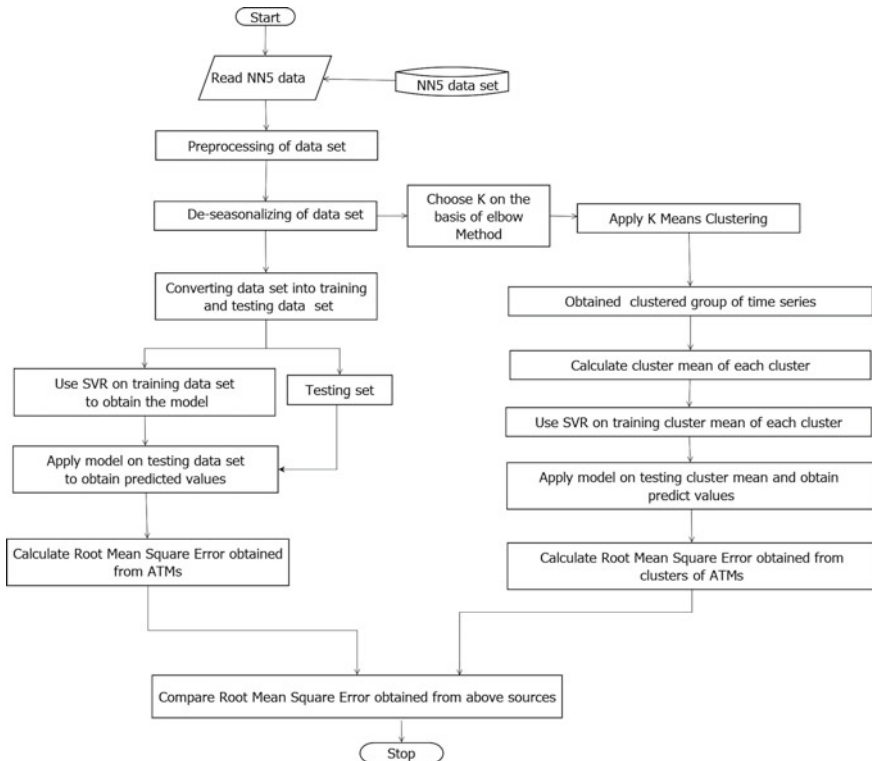
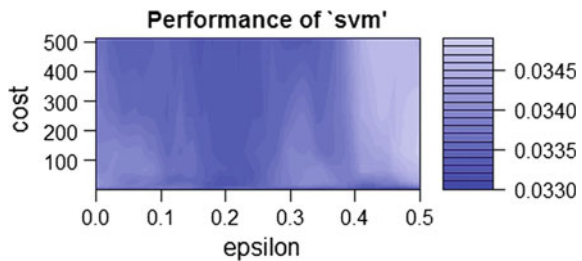


Fig. 1 Flow diagram of clustered support vector machine algorithm (C-SVM)

Fig. 2 SVM model of time series NN5001



**Table 1** Root mean square error for all time series

Time series	NN5001	NN5002	NN5003	NN5004	NN5005	NN5006	NN5007	NN5008	NN5009	NN5010
RMSE	0.12	0.17	0.16	0.26	0.3	0.31	0.18	0.3	0.19	0.22
Time series	NN5011	NN5012	NN5013	NN5014	NN5015	NN5016	NN5017	NN5018	NN5019	NN5020
RMSE	0.21	0.19	0.16	0.17	0.26	0.2	0.29	0.25	0.2	0.16
Time series	NN5021	NN5022	NN5023	NN5024	NN5025	NN5026	NN5027	NN5028	NN5029	NN5030
RMSE	0.16	0.14	0.33	0.26	0.17	0.24	0.23	0.19	0.26	0.3
Time series	NN5031	NN5032	NN5033	NN5034	NN5035	NN5036	NN5037	NN5038	NN5039	NN5040
RMSE	0.23	0.17	0.21	0.18	0.31	0.31	0.31	0.16	0.16	0.18
Time series	NN5041	NN5042	NN5043	NN5044	NN5045	NN5046	NN5047	NN5048	NN5049	NN5050
RMSE	0.23	0.16	0.15	0.17	0.25	0.17	0.12	0.3	0.23	0.21
Time series	NN5051	NN5052	NN5053	NN5054	NN5055	NN5056	NN5057	NN5058	NN5059	NN5060
RMSE	0.12	0.28	0.17	0.17	0.18	0.3	0.21	0.23	0.26	0.28
Time series	NN5061	NN5062	NN5063	NN5064	NN5065	NN5066	NN5067	NN5068	NN5069	NN5070
RMSE	0.21	0.27	0.26	0.3	0.22	0.25	0.16	0.19	0.36	0.16
Time series	NN5071	NN5072	NN5073	NN5074	NN5075	NN5076	NN5077	NN5078	NN5079	NN5080
RMSE	0.17	0.18	0.31	0.22	0.29	0.19	0.16	0.14	0.19	0.24
Time series	NN5081	NN5082	NN5083	NN5084	NN5085	NN5086	NN5087	NN5088	NN5089	NN5090
RMSE	0.23	0.21	0.21	0.15	0.3	0.29	0.22	0.23	0.19	0.21
Time series	NN5091	NN5092	NN5093	NN5094	NN5095	NN5096	NN5097	NN5098	NN5099	NN5100
RMSE	0.22	0.23	0.1	0.24	0.1	0.21	0.24	0.24	0.13	0.17



**Table 2** Root mean square error obtained for 15 cluster centres

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RMSE	0.19	0.27	0.33	0.37	0.08	0.18	0.19	0.04	0.14	0.24	0.05	0.2	0.27	0.09	0.1

Mean of RMSE obtained from all NN5 time series is calculated.

**Step VI: *K Means Clustering***

- (a) ***Choosing the number of clusters(K)***: Choosing of K: Before applying K means clustering, the value of k is to be chosen. Value of k is chosen by elbow point. The elbow technique concentrates at the percentage of variance described as a function of the accurate quantity of clusters. As per elbow method, number of clusters are 15 (k = 15).
- (b) ***K Means Clustering***: Now unsupervised learning (K Means) is used to produce optimal segments of ATMs.
- (c) Centre of each cluster is obtained. Now each centre will represent all the time series which are the part of that cluster.

**Step VII: *Repeat step 6 for each cluster (Cluster 1 to Cluster 15)***

- (a) Support vector regression is applied to cluster centre of each cluster. Optimal values for cost and epsilon is to be chosen. The darkest shade of plot represents the optimal value of cost and epsilon.
- (b) For getting optimal values of cost and epsilon, tuning of the model is done.
- (c) The accuracy of model is tested via applying model on testing data and RMSE is calculated.
- (d) RMSE is calculated for getting the performance of the model. Table 2 shows the RMSE for all cluster centres of the dataset.

**Step VIII: *SVM with clustering and without clustering is compared***

**Output:** *Prediction of ATM Cash requirements.*

## 6 Results

Intermediate results of the whole procedure have been exhibited in Tables 3, 4, 5, 6, 7, 8 and Fig. 3a, b. The final result of the algorithm is showcased as the comparison of RMSE between SVM having clustering and SVM without having clustering in Table 9.

Here the number of clusters are decided by elbow method. As the number of clusters is increased, then the division of within sum of squared error (WSS) and between sum of squared error (BSS) is decreased. Value of k is to be picked where WSS/BSS is less than 2. So the number of clusters is 15 on the basis of elbow method.

**Table 3** Dataset having no missing values

Time series	NN5001	NN5002	NN5003	NN5004	NN5005	NN5006	NN5007	NN5008	NN5009	NN5010
18 March 96	13.41	11.55	5.64	13.18	9.78	9.24	14.94	2.89	7.34	10.29
19 March 96	14.73	13.59	14.40	8.45	10.81	11.64	16.28	12.36	9.16	12.71
20 March 96	20.56	15.04	24.42	19.52	21.61	12.10	16.67	16.38	10.59	14.44
21 March 96	34.71	21.57	28.78	28.88	38.52	21.41	23.57	30.16	12.50	19.40
22 March 96	26.63	19.44	20.62	19.47	24.75	24.67	26.30	31.18	7.16	21.54
23 March 96	16.61	0.00	13.80	0.00	12.33	5.23	14.90	19.81	5.64	15.22
24 March 96	15.32	9.72	11.54	7.36	13.00	11.38	16.04	17.49	7.70	11.35
25 March 96	11.61	12.25	10.74	10.83	11.04	9.55	13.89	8.72	5.73	11.14
26 March 96	19.88	15.51	14.82	15.62	7.95	15.72	20.39	15.51	5.51	13.93
27 March 96	23.77	18.93	25.21	21.16	19.52	14.87	19.35	21.49	8.42	17.35
28 March 96	34.03	26.08	35.13	33.49	33.77	26.89	28.42	35.29	13.31	22.48
29 March 96	33.79	27.25	24.79	27.91	26.91	30.34	31.35	47.49	7.19	30.02
30 March 96	18.25	11.71	12.90	13.22	7.99	3.83	2.25	10.06	4.20	14.87
31 March 96	19.39	14.94	15.12	14.75	6.25	13.65	14.91	18.81	7.20	14.47
1 April 96	17.26	12.94	11.91	14.30	18.67	13.08	20.03	14.24	7.74	10.80
2 April 96	23.81	18.17	17.73	10.67	15.87	16.65	20.59	20.27	9.51	13.95
3 April 96	36.13	24.89	36.14	32.92	28.19	19.46	26.64	28.26	11.07	25.33
4 April 96	33.60	15.72	22.70	20.34	32.36	16.19	18.16	31.93	6.21	25.31
5 April 96	32.63	25.21	24.94	22.48	0.17	25.40	29.17	34.85	5.61	27.37
6 April 96	7.48	11.51	2.57	15.24	20.32	4.51	11.64	6.34	3.83	15.14
7 April 96	28.82	9.01	0.00	16.62	20.32	4.71	10.02	5.39	5.27	13.69

**Table 4** Normalized NN5 dataset

Time Series	NN5001	NN5002	NN5003	NN5004	NN5005	NN5006	NN5007	NN5008	NN5009	NN5010
18 March 96	0.18	0.25	0.09	0.24	0.14	0.15	0.31	0.04	0.34	0.19
19 March 96	0.19	0.29	0.23	0.16	0.15	0.19	0.34	0.19	0.43	0.25
20 March 96	0.27	0.32	0.38	0.36	0.3	0.2	0.35	0.25	0.5	0.3
21 March 96	0.46	0.46	0.45	0.53	0.54	0.36	0.5	0.46	0.58	0.43
22 March 96	0.35	0.42	0.32	0.36	0.35	0.41	0.55	0.48	0.33	0.48
23 March 96	0.22	0	0.22	0	0.17	0.09	0.31	0.3	0.26	0.32
24 March 96	0.2	0.21	0.18	0.14	0.18	0.19	0.34	0.27	0.36	0.21
25 March 96	0.15	0.26	0.17	0.2	0.16	0.16	0.29	0.13	0.27	0.21
26 March 96	0.26	0.33	0.23	0.29	0.11	0.26	0.43	0.24	0.26	0.28
27 March 96	0.31	0.41	0.4	0.39	0.27	0.25	0.41	0.33	0.39	0.37
28 March 96	0.45	0.56	0.55	0.62	0.47	0.45	0.6	0.54	0.62	0.51
29 March 96	0.44	0.58	0.39	0.51	0.38	0.5	0.66	0.73	0.34	0.7
30 March 96	0.24	0.25	0.2	0.24	0.11	0.06	0.05	0.15	0.2	0.31
31 March 96	0.26	0.32	0.24	0.27	0.09	0.23	0.31	0.29	0.34	0.3
1 April 96	0.23	0.28	0.19	0.26	0.26	0.22	0.42	0.22	0.36	0.2
2 April 96	0.31	0.39	0.28	0.2	0.22	0.28	0.43	0.31	0.44	0.28
3 April 96	0.48	0.53	0.57	0.61	0.4	0.32	0.56	0.43	0.52	0.58
4 April 96	0.44	0.34	0.36	0.37	0.45	0.27	0.38	0.49	0.29	0.58
5 April 96	0.43	0.54	0.39	0.41	0	0.42	0.61	0.54	0.26	0.63
6 April 96	0.1	0.25	0.04	0.28	0.29	0.07	0.24	0.1	0.18	0.31
7 April 96	0.38	0.19	0	0.31	0.29	0.08	0.21	0.08	0.25	0.28

**Table 5** De-seasonalized training dataset

Times series	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
NN5001	0.66	0.87	1.09	1.55	1.21	0.83	0.78
NN5002	0.80	0.88	1.03	1.44	1.27	0.76	0.82
NN5003	0.60	0.86	1.29	1.73	1.18	0.67	0.66
NN5004	0.70	0.79	1.11	1.62	1.24	0.79	0.76
NN5005	0.69	0.70	1.08	1.78	1.28	0.78	0.69
NN5006	0.74	0.89	0.97	1.56	1.61	0.44	0.79
NN5007	0.86	0.86	0.95	1.36	1.35	0.72	0.90
NN5008	0.64	0.92	0.99	1.58	1.56	0.56	0.74
NN5009	0.96	0.96	1.09	1.45	0.91	0.69	0.94
NN5010	0.70	0.77	0.95	1.56	1.36	0.87	0.80

**Table 6** De-seasonalized testing dataset

Times series	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
NN5001	0.68	0.77	1.17	1.53	1.29	0.82	0.74
NN5002	0.78	0.92	1.12	1.42	1.35	0.69	0.71
NN5003	0.62	0.85	1.37	1.70	1.21	0.64	0.62
NN5004	0.59	0.85	1.27	1.79	1.28	0.69	0.54
NN5005	0.52	0.56	1.20	1.97	1.43	0.73	0.59
NN5006	0.70	0.91	1.09	1.67	1.54	0.40	0.70
NN5007	0.82	0.93	1.03	1.37	1.36	0.69	0.81
NN5008	0.69	0.92	1.07	1.66	1.58	0.42	0.66
NN5009	0.94	1.06	1.11	1.51	0.83	0.77	0.79
NN5010	0.72	0.67	1.17	1.62	1.32	0.78	0.72

Figure 3a, b shows the benefits of clustering used in time series data. Figure 3a shows the comparison of actual output and predicted the output of de-seasonalized testing data of time series NN5001. Figure 3b shows the comparison of actual output and predicted output of cluster 3 mean. Cluster 3 is taken because NN5001 is associated with cluster 3.

## 7 Conclusion

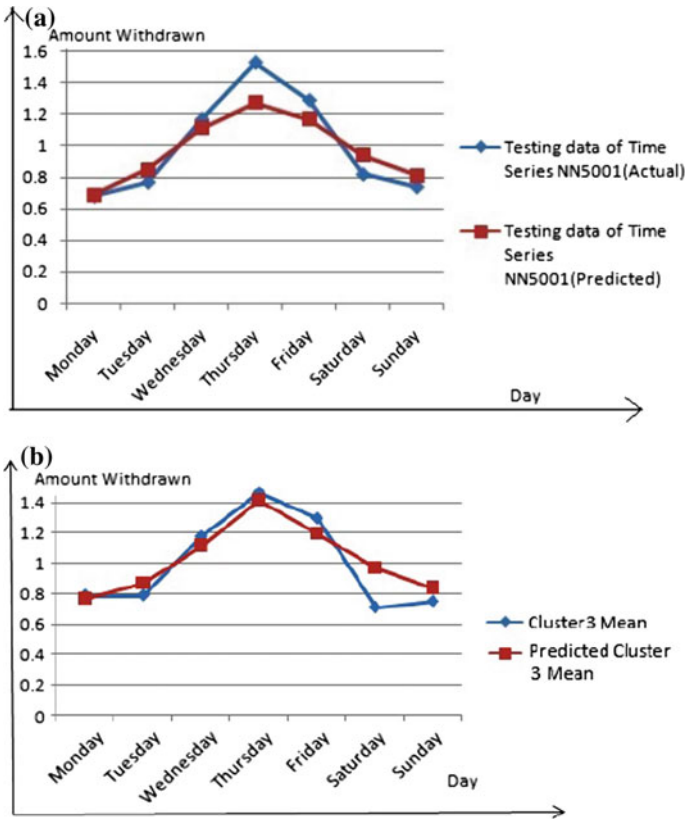
Prediction of funds in ATMs in an ideal way is an essential and complicated process. After setting up and putting ATMs, balancing of cash in ATMs should be there to avoid overuse and underuse of money. Money in the ATMs should always be well balanced to prevent both dissatisfaction among customers and wasting money. The optimal and precise prediction may be obtained using support vector regression on the

**Table 7** Cluster means of training data

Cluster	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	0.73	0.75	1.08	1.37	1.60	0.69	0.80
2	1.05	1.05	1.24	1.56	0.63	0.41	1.06
3	0.90	0.91	1.02	1.37	1.23	0.63	0.93
4	0.74	0.81	1.15	1.50	1.23	0.76	0.81
5	0.82	0.85	0.96	1.31	1.55	0.64	0.86
6	0.73	0.75	1.00	1.57	1.37	0.81	0.78
7	0.65	0.79	1.25	1.78	1.19	0.60	0.74
8	0.69	0.74	1.09	1.69	1.28	0.78	0.73
9	0.72	0.91	0.98	1.59	1.40	0.60	0.79
10	0.87	0.85	0.95	1.30	1.23	0.87	0.93
11	0.79	0.83	1.02	1.43	1.32	0.76	0.85
12	0.73	0.78	1.15	1.66	1.25	0.67	0.78
13	0.63	0.67	1.23	1.81	1.38	0.63	0.67
14	0.93	0.94	1.10	1.42	0.95	0.72	0.95
15	0.80	0.80	0.96	1.51	1.57	0.53	0.83

**Table 8** Cluster means of testing data of each cluster

Cluster	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	0.78	0.83	1.33	1.47	1.23	0.63	0.74
2	0.8	0.82	1.14	1.47	1.32	0.7	0.75
3	0.88	0.91	1.02	1.43	1.64	0.38	0.75
4	1.1	1.16	1.35	1.49	0.69	0.33	0.88
5	0.85	0.8	0.97	1.33	1.23	0.93	0.88
6	0.7	0.73	1.06	1.63	1.36	0.77	0.75
7	0.71	0.81	1.37	1.73	1.18	0.54	0.67
8	0.79	0.86	1.06	1.61	1.47	0.48	0.73
9	0.87	0.97	1.1	1.35	1.27	0.66	0.77
10	0.82	0.89	1.09	1.33	1.49	0.6	0.77
11	0.72	0.81	1.2	1.71	1.31	0.63	0.62
12	1.02	0.97	1.19	1.43	0.94	0.63	0.83
13	0.6	0.63	1.27	1.98	1.41	0.56	0.56
14	0.68	0.92	1.37	1.84	1.26	0.29	0.66
15	0.76	0.7	1.22	1.49	1.6	0.54	0.69



**Fig. 3** a Original versus predicted time series (without clustering). b Original versus predicted time series (with clustering)

**Table 9** RMSE obtained from SVM having clustering and SVM not having clustering

Error	Clustered time series	Non-clustered time series
RMSE	0.18	0.22

ATMs transactions. This paper exhibits the enhancement in precision by indicating the decrease in RMSE. The RMSE was decreased when ATMs had been clustered before applying SVR on the dataset. Training of machine learning algorithm is more adequate when similar ATMs are segmented into optimal clusters and leading towards the precise forecasting. Authors may use different clustering algorithms and suggest optimal clustering algorithm which may be used for generating segmented groups of ATMs.

## References

1. Ojemakinde, B.T.: Support vector regression for non-stationary time series (2006)
2. Andrawis, R.R., Atiya, A.F., El-Shishiny, H.: Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. *Int. J. Forecast.* **27**(3), 672–688 (2011)
3. Wichard, J.D.: Forecasting the NN5 time series with hybrid models. *Int. J. Forecast.* **27**(3), 700–707 (2011)
4. Coyle, D., Prasad, G., McGinnity, T.M.: On utilizing self-organizing fuzzy neural networks for financial forecasts in the NN5 forecasting competition. In: *Proceedings of the International Joint Conference on Neural Networks* (2010)
5. Teddy, S.D., Ng, S.K.: Forecasting atm cash demands using a local learning model of cerebellar associative memory network. *Int. J. Forecast.* **27**(3), 760–776 (2011)
6. Ben Taieb, S., Bontempi, G., Atiya, A.F., Sorjamaa, A.: A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Syst. Appl.* **39**(8), 7067–7083 (2012)
7. Boyd, W.: Support vector machine for regression and applications to financial forecasting. *IJCNN, IEEE (x)*, 6348 (2000)
8. Tay, F.E.H., Cao, L.J.: Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intel. Data Anal.* **5**, 339–354 (2001)
9. Cao, L., Tay, F.E.H.: Modified support vector machines in financial time series forecasting. *Neurocomputing* **48**(1–4), 847–861 (2002)
10. Cao, L.J., Chua, K.S., Guan, L.K.: e-descending support vector machines for financial time series forecasting. *Neural Process. Lett.* **15**, 179–195 (2002)
11. Van Gestel, T. et al.: Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Trans. Neural Netw.* **12**(4), 809–821 (2001)
12. Pai, P.-F., Lin, K.-P., Lin, C.-S., Chang, P.-T.: Time series forecasting by a seasonal support vector regression model. *Expert Syst. Appl.* **37**(6) (2010)
13. Hung, C., Hung, C., Lin, S.: Predicting Time series using integration of moving average and support vector regression. *Int. J. Mach. Learn. Comput.* **4**(6), 491–495 (2014)
14. Harris, T.: Credit scoring using the clustered support vector machine. *Expert Syst. Appl.* **42**(2), 741–750 (2015)
15. Khanarsa, P., Sinapiromsaran, K.: Multiple ARIMA subsequences aggregate time series model to forecast cash in ATM. In: *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*, pp. 83–88 (2017)
16. Jadwal, P.K., Jain, S., Gupta, U., Khanna, P.: K-Means clustering with neural networks for ATM cash repository prediction. In: *Satapathy, S., Joshi, A. (eds.), Information and Communication Technology for Intelligent Systems (ICTIS 2017), ICTIS 2017*, vol. 1. *Smart Innovation, Systems and Technologies*, vol. 83. Springer, Cham (2018)
17. <http://www.neural-forecastingcompetition.com/downloads/NN5/datasets>

# An Effective Intrusion Detection System Using Flawless Feature Selection, Outlier Detection and Classification



Rajesh Kambattan Kovarasan and Manimegalai Rajkumar

**Abstract** Intrusion detection system (IDS) is playing crucial role to provide the security in the fastest world by protecting the internet applications such as health-care applications, government secret information, secret banking data and intellectual properties of various scientists. In this paper, we propose new intrusion detection system for improving the detection rate. The proposed system is the combination of feature selection, outlier detection and classification. First, a newly proposed feature selection algorithm called intelligent flawless feature selection algorithm (IFLFSFA) is used for selecting optimal number of features which are most useful for identifying the attacks. Second, the proposed entropy-based weighted outlier detection (EWOD) technique is used to identify the outliers from the data set. Third, the existing classification algorithm called intelligent layered approach for effective classification is used. The experiments have been conducted for evaluating the proposed model using the KDD data set. The proposed system achieved better detection accuracy in terms of high detection accuracy and low error rate.

**Keywords** Intrusion detection system · Feature selection · Outlier detection · Classification · Flawless feature selection · Layered approach

---

R. K. Kovarasan (✉)

Department of Computer Science and Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India  
e-mail: rajesh.kambattan@gmail.com

M. Rajkumar

Department of Information Technology, PSG College of Technology, Coimbatore, Tamil Nadu, India  
e-mail: mmegalai@yahoo.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713, [https://doi.org/10.1007/978-981-13-1708-8\\_19](https://doi.org/10.1007/978-981-13-1708-8_19)



# 1 Introduction

In recent decades, intrusions in internet and local network become more tedious task to detect. Intrusions are available with malicious source code, includes virus and worms. Latest attacks lead to very big damage at organization level and also distributed architecture level [1]. To save computer users from malicious effects, IDS (intrusion detection system) is designed to look out network activities and produce alerts to respective persons like administrative and others. IDS is used for two purposes: one method is used to identify known attacks, and the other method is used for unknown attacks. The implementation of second technique is not easy, and the system should go with proper learning and testing process [2]. As per discussion [3] about MANET (mobile ad hoc network), it has dynamic nature and very easy to get harm by malicious nodes.

The proposed work is to reduce the malicious activities by identifying the intruders early in network done through monitoring the node behaviour/features. Features meant for nodes behaviour and very easy to track when they moved from normal activities. Normally, IDS is used to take a very few attributes from nodes and applied to testing process. Nowadays, it is not an easy process to identify attacks, so considering feature(s) is also very important. Based on natures [4] of features, the methods could be chosen by expert members. These features selection comes under pre-process techniques in data mining. Advantages of pre-processing are time will be reduced for calculation and also investment cost will be reduced with high performance to output.

Finally, intrusion detection systems end with nodes identification, even though nodes are malicious or not. This process is called as classification which is able to segregate good nodes and misbehave nodes. Outlier detection has been focused on many recent research fields. Outlier detection is an important work in big data and data mining with enormous applications such as video surveillance and credit card misuse detection etc. An outlier is an abnormal activity that deviated from normal work or normal attribute.

In this paper, a new intrusion detection system has been proposed for effective intrusion detection. The proposed system contains three phases such as feature selection, outlier detection and classification. The first contribution of this paper is the introduction of a new feature selection algorithm called intelligent flawless feature selection which is useful for recommending the useful features. The second contribution of this paper is the introduction of new outlier detection method called entropy-based weighted outlier detection method for removing the useless records. Third contribution of this paper is the uses of the existing classification algorithm called intelligent layered approach for effective classification. The main advantage of this proposed work is to select the useful features which are useful to improve the classification (intrusion detection) accuracy.

The rest of this paper is organized as follows: Sect. 2 provides the literature survey. Section 3 explains the proposed work. Section 4 demonstrates the results and discussion. Section 5 gives conclusion and the future works.

## 2 Literature Survey

There are many algorithms have been proposed in the areas of feature selection, classification and outlier detection for effective intrusion detection by various researchers in the past. False alarm is generated based on outlier movements in machine learning and knowledge discovery field. This scenario was explained by Ru et al. [1]. Bai et al. [3] proposed outlier detection model for large data set using local outlier factor. Here, each and every tuple is considered to be a small degree of outliers. Based on density outlier concepts, outliers are identified through two methods, namely grid-based partition algorithm and distributed LOF.

Bandyopadhyay and Santra [4] proposed grid count tree (GCD) which is a new type of data structure to identify outlier detection. It is very effective factor to segregate useful messages from abnormal. Subspace-based outlier detection for high-dimensional massive data set was proposed by Zhang et al. [5]. Here, local subset is identified and then respective outlier factor was assigned the same for identifying the distribution that does not respond. In recent days, to avoiding unwanted parameters or records, Bouarfa and Dankelman [6] used techniques like work flow mining to detect work flow outliers. To achieve the result, the authors used NW alignment algorithm.

Pai et al. [7] proposed a model for categorical data to identify relevant pattern and discard outlier. In this paper, a new relative pattern discovery is used for association analysis which is mainly applied at distortion problems. Kuna et al. [8] achieved the best result in the audit log of application system by data mining concepts. Here, outlier detection and classification algorithms are merged for better result. The authors conducted many experiments to identify unwanted logs in the multidimensional database with the support of data mining advantages. The problem of multivariate outlier is discussed by Muiioz and Muruzbal [9]. The authors are taken self-organizing map for detecting outliers. Based on the neuron's distance matrix, outlier is detected in statistics and graphics domains. They were concentrated about data selection as an input for further work.

Fraiman et al. [10] proposed a feature selection for functional data, which gives constantly good results for reduced data set. This reduced set gives better explanation rather than entire data set. Wang et al. [11] proposed a technique for classification to produce better identification among two different cancer cells with the help of textual extraction methods. It gives best result in the field of classification of single cells by label-free classification.

Zhou et al. [12] proposed a method for feature selection in the field of neurocomputing. Here, prediction model was developed to identify better feature set which contains 34 features for further process. A trust-based collaborative decision framework for IDS networks was proposed by Fung et al. [13]. Ganapathy et al. [14]

proposed a new outlier detection technique called weighted distance-based outlier detection for intrusion detection in mobile ad hoc network. This proposed algorithm is best to identify intruders than previous methods. Ganapathy et al. [15] proposed an intelligent layered approach for effective classification.

Subba et al. [16] proposed intrusion detection system for mobile network. This IDS is comprised of cluster head which provides intrusion detection service, and hybrid IDS is a kind of anomaly detection service. This scheme has reduced IDS traffic and overall power consumption. Zorarpac et al. [17] discussed data dimensions, because it is the one of the major issues in machine learning and data mining. This scheme proposed new hybrid concept which combines optimized bee colony and evolution algorithm for feature selection for classification work. This system improved good accuracy and run-time performance.

Pölsterl et al. [18] proposed a technique for feature extraction as an alternate to feature selection to identify local neighbourhood relations from survival data. For large samples, feature extraction is carried out without any problem. Muhammad Raza et al. [19] proposed an incremental dependency class (IDC) for feature selection to calculate dependency without positive region. This approach gives great advantage to rough set theory. Krawczyk et al. [20] proposed one class classification. Ensemble methods are best to estimate classification accuracy. This ensemble method is avoided to choose weak set and improve the robustness. This one class classification is used for three kinds of measurements to improve the system performance.

### 3 Proposed Work

This section presents the proposed system. In this section, we have discussed in detail the existing feature selection, the proposed outlier detection model and the existing classification algorithm.

#### 3.1 *Intelligent Flawless Feature Selection*

This paper introduced a new feature selection algorithm called intelligent flawless feature selection algorithm (IFLFS) which selects the optimal number of features that are used for effective classification. This algorithm applies all the pre-processing activity such as removal of noise data and null values and the selection of more relevant data. Here, we have used intelligent agent for decision-making. The steps of the algorithm are as follows:

Input : Dataset  
 Output: Reduced features

Step 1: Read the dataset  
 Step 2: for  $i = 1$  to  $n$  do  
 Remove the noisy and null values data  
 Remove the redundant records  
 Calculate the Information Gain Ratio value  
 Step 3: Find the mean value for IGR of dataset  
 Step 4: Remove the features which are less than the mean value of IGR and store into the set Reduced Feature Set (RFS).  
 Step 5: Apply ICRFFSA [14] on RFS for selecting optimal features and stored in to Selected Feature Set (SFS).  
 Step 6: Fix the *Threshold* ( $Th$ ) which feature IGR value is above 10% of the mean value.  
 Step 7: for  $j = 1$  to  $n$  do  
     If  $SFS_j > Th (RFS_j)$  Then  
          $SF_j = SFS_j$   
     Else  
          $IRF_j = SFS_j$   
 Step 8: Agent takes final decision on dataset to select effective features.  
 Step 9: Display the selected feature set  $SF_j$  and write into an input dataset file.

Here, IRF indicates irrelevant features and SF means selected features.

The proposed algorithm helps to select optimal number of features by applying the basic pre-processing activities and intelligent agent. Information gain ratio is calculated for all the features which are available in the data set. Before that, the noisy data, null value and the redundant data are removed. Find the mean value for the IGR value which is calculated for the attributes in the data set. Remove the features which are less than the mean value of IGR and store in the reduced feature set (RFS), and it is considered for applying ICRFFSA for selecting optimal features and then stored in the selected feature set (SFS). Here, we have not chosen the mean value as threshold. The threshold value is fixed only by considering the features which IGR is above 10% of the mean value.

Now, check whether the IGR of features which are available in RFS is above the threshold or not. If the IGR of feature is above the threshold, then store in selected features (SF), otherwise store in the irrelevant feature (IRF) set. Finally, apply the agent for making final decision over the selected features. Features are finalized and it is recommended for further process. The set of selected features which are available in the set SF are forwarded into the next phase for outlier detection.

### 3.2 Entropy-Based Weighted Outlier Detection

In this section, we discussed in detail the proposed information gain ratio-based outlier detection method for effective grouping and identifying the useful records in the given data set. Here, a new weight is assigned for each feature of the records in the data set. The weighted entropy value is calculated for each feature by using Eq. (1) which is used to measure the uncertainty level and the information level.

$$WE(x) = - \sum_{i=1}^n p(x_i) \log p(x_i) \times W_i, \tag{1}$$

where  $x$  indicates the random variable (feature) and  $WE(x)$  indicates the weighted entropy which demonstrates the probability distribution on  $x$ . Weight is assigned based on the importance of the features.

The input data set consists of  $m$  records with  $n$  number of attributes; the weighted entropy  $WE(x)$  of each feature (multi-variable) vector value  $\vec{x} = \{X_1, W_1, \dots, X_m W_m\}$ , where  $X_i$  is a random variable whose realizations belong to the set of  $\{x_i^1, \dots, x_i^m\}$ , can be calculated as Eq. (2).

$$WE(\vec{x}) = \sum_{x_1 \in \{x_{11} w_{11}, \dots, x_{1n} w_{1n}\}} \dots \sum_{x_m \in \{x_{m1} w_{m1}, \dots, x_{mn} w_{m1}\}} \dots (x_i w_i, \dots, x_m w_m) \lg p(x_i x_i, \dots, x_m w_m) \tag{2}$$

$WE(\vec{x})$  is calculated as the sum of weighted entropies of attributes using Eq. (3). Here, weighted entropy is calculated for all the features by considering the dependencies with other features of the data set.

$$WE(\vec{x}) = E(X_1) \times W_1 + E(X_2) \times W_2 + \dots + E(X_n) \times W_n. \tag{3}$$

The different weights are assigned for the different random variables (features) in each record of the whole data set. Finally, all the records contain different weights based on the value of selected features. These weights are considered for extracting the outliers.

#### Extracting the Outliers

The major task of this subsection is to extract the negative records from positive samples and unlabelled data. The weighted entropy of posterior probabilities of each record is calculated as  $d_i$  in the testing data set which is unlabelled. The enhanced Eq. (4) of the proposed weighted entropy is given below:

$$WE(d_i) = - \sum_{j=1}^{|C|} p(C_j | d_i) \lg(C_j | d_i) \times W_i, \tag{4}$$

where  $p(C_j|d_i)$  indicates the posterior probability of the record set  $d_i$  which belongs to the  $j$ th class of  $C_j$ ;  $|C|$  indicates the training set. To calculate the posterior probability which belongs to the class  $C_j$ , each positive record set in the training data sets which are considered as a centre of a positive class  $C_j$  and  $W_i$  indicates the  $i$ th record weights which are calculated using Eqs. (1)–(3). Then, the distance between each record set  $d_i$  and the centre is measured using the Minkowski distance. These distances for each record will be normalized, and estimating the probability  $p(C_j|d_i)$  is shown in Eq. (5).

$$p(C_j|d_i) = \frac{distance(di, pj)}{\sum_{j=1}^{|C|} distance(di, pj)}. \tag{5}$$

The negative examples are acquired using Eq. (6).

$$d_j = argmax_{d_j \in U} (WE(d_i)). \tag{6}$$

**Outliers (Negative Set) Selection Algorithm**

Input : Positive samples  $ps$ , Test data  $T$

Output : Negative Samples set  $O$  & Most contributed featured data set  $D$ .

Step 1: Class  $C = \{c_1, \dots, c_n\}$  are all positive class labels in  $ps$

Step 2: Outlier set  $O = \{\}$

Step 3: For each samples  $d_i \in T$  do

$$WE(d_i) = - \sum_{j=1}^{|C|} p(C_j|d_i) \lg(C_j|d_i)$$

End for

Step 4: Sort  $WE(d_i)$  in ascending order for each  $(d_i)$

Step 5: Weighted entropy is sorted from the first in the list

Step 6: Pick  $k$  number of top records  $d_i$  from the sorted list

Step 7: Outliers  $O = \{k \text{ top records } d_i\}$

Step 8: Display and store the outliers

Step 9: Finally, removed the outliers and make it into new dataset  $D$ .

In this algorithm, the top  $k$  number of records with maximum  $WE(d_i)$  have been selected as outlier records. Finally, we can extract the most contributed featured data set for further process. Here, test data  $T$  contain only the selected features which are available in  $SF$ . Finally, the final most contributed features are stored in the data set  $D$ .

**3.3 Classification**

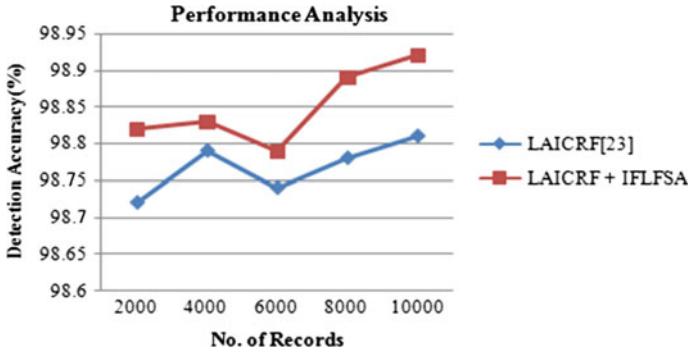
In this paper, an effective classification approach called intelligent layered approach [15] is used for effective classification. After performing the outlier detection algorithm, the resulted data set will be given as input to the existing intelligent layered

**Table 1** List of 15 selected features

---

protocol\_type, src\_byte, wrong\_fragment, Hot, root\_shell, su\_attempted, num\_access\_shells, error\_rate, diff\_srv\_rate, srv\_error\_rate, dst\_host\_srv\_count, dst\_host\_same\_srv\_count, dst\_host\_diff\_srv\_count, dst\_host\_srv\_diff\_host\_rate and dst\_host\_error\_rate

---

**Fig. 1** Performance analysis

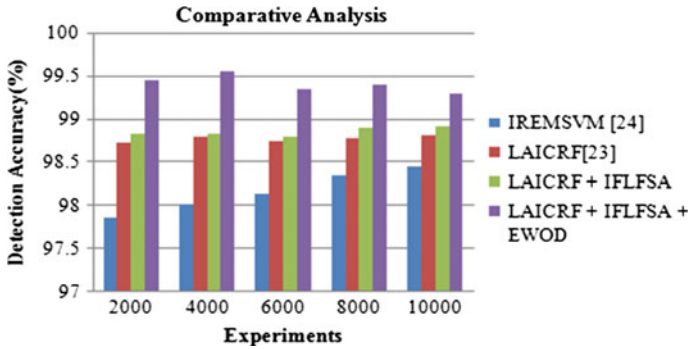
approach for classification. This layered approach works based on the attacks classification such as Probe, DoS, R2L, U2R and Normal. Here, four layers are available for each attack. The final reduced data set  $D$  is given as input to this layered approach. The record set is classified into four types by using an intelligent agent. This intelligent agent is used for making effective final decision over the reduced set  $D$ . The final classified resulted records are stored in the separate file for the different attacks and normal.

## 4 Results and Discussion

The proposed system has been implemented by using JAVA and also tested with WEKA tool. The KDD'99 Cup data set [21] is used for carrying out the experiments. This data set contains five million records with 41 features. Here, we have selected only 10,000 records randomly from the data set for carrying out the experiments. The different sets of records are used for conducting the various experiments, respectively 2000, 4000, 6000, 8000 and 10,000.

Table 1 lists the 15 selected features by the proposed feature selection algorithm from the 41 features which are available in the bench mark data set.

Figure 1 shows the performance of the feature selection with the existing classification approach called intelligent layered approach [15]. Five experiments have been conducted with different numbers of records, such as 2000, 4000, 6000, 8000 and 10,000.



**Fig. 2** Comparative analysis

From Fig. 1, it can be observed that the performance of the proposed feature selection is better when it is compared with the existing model called LAICRF which is the combination of the existing intelligent layered approach and intelligent CRF-based feature selection. This is due to the use of efficient features for classification. Here, we have considered the overall detection accuracy over the data set. The overall detection accuracy covers the detection accuracy on four types of attacks such as Probe, DoS, R2L and U2R. The reason for the significant accuracy in the proposed and existing systems is the records considered for the experiments.

From Fig. 2, it can be observed that the performance of the proposed model which is combining the proposed feature selection, weighted IGR-based outlier detection and the existing intelligent layered approach is better when it is compared with the existing intrusion detection models, namely IREMSVM and LAICRF, which is the combination of the existing intelligent layered approach and intelligent CRF-based feature selection. The reason for the improvement is the use of the proposed flawless feature selection technique and the proposed weighted outlier detection approach. Here, the overall detection accuracy of the proposed system is high (99.45%) when it is compared with the existing systems, namely LAICRF [15] (98.6%) and IREMSVM [24] (98.3%), and also considers the different combinations of the existing and proposed algorithms such as the existing LAICRF [15] and the proposed IFLFSA.

## 5 Conclusion and Future Works

In this paper, a new intrusion detection system has been proposed and implemented for effective intrusion detection by improving the classification accuracy. The proposed system is the combination of feature selection, outlier detection and classification. First, the proposed intelligent flawless feature selection algorithm is used for selecting an optimal number of features which are most useful for identifying the attacks. Second, the proposed entropy-based weighted outlier detection technique is



used to identify the outliers and the useful data from the data set. Third, the existing classification algorithm called intelligent layered approach is used for effective classification. The experiments have been conducted for evaluating the proposed model using the KDD data set. The scientific contributions of this work are the introduction of new feature selection algorithm and weight-based outlier detection through outlier selection algorithm. The proposed system achieved better detection accuracy (99.45%) in terms of high detection accuracy when it is compared with the existing algorithms (98.3 and 98.6%). The overall detection accuracy of the proposed work is 99.45% which is around 1% more accuracy than the existing systems.

## References

1. Ru, X., Liu, Z., Huang, Z., Jiang, W.: Normalized residual-based constant false-alarm rate outlier detection. *Pattern Recogn.* **69**, 1–7 (2016)
2. Aljawarneh, S., Aldwairi, M., Yassein, M.B.: Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *J. Computat. Sci.* Elsevier (2017)
3. Bai, M., Wang, X., Xin, J., Wang, G.: An efficient algorithm for distributed density-based outlier detection on big data. *Neurocomputing* **181**, 19–28 (2016)
4. Bandyopadhyay, S., Santra, S.: A genetic approach for efficient outlier detection in projected space. *Pattern Recogn.* **41**, 1338–1349 (2008)
5. Zhang, J., Yu, X., Li, Y., Zhang, S., Xun, Y., Qin, X.: A relevant subspace based contextual outlier mining algorithm. *Knowl. Based Syst.* **99**, 1–9 (2016)
6. Bouarfa, L., Dankelman, J.: Workflow mining and outlier detection from clinical activity logs. *J. Biomed. Inform.* **45**, 1185–1190 (2012)
7. Pai, H.T., Wua, F., Hsueh, S.P.Y.: A relative patterns discovery for enhancing outlier detection in categorical data. *Decis. Support Syst.* **67**, 90–99 (2014)
8. Kuna, H.D., García-Martínez, R., Villatoro, F.R.: Outlier detection in audit logs for application systems. *Inf. Syst.* **44**, 22–33 (2014)
9. Muiioz, A., Muruzhbal, J.: Self-organizing maps for outlier detection. *Neurocomputing* **18**, 33–60 (1998)
10. Fraiman, R., Gimenez, Y., Svarc, M.: Feature selection for functional data. *J. Multivar. Anal.* **146**, 191–208 (2016)
11. Wang, H., Feng, Y., Sa, Y., Lu, J.Q., Ding, J., Zhang, J., Hu, X.H.: Pattern recognition and classification of two cancer cell lines by diffraction imaging at multiple pixel distances. *Pattern Recogn.* **61**, 234–244 (2016)
12. Zhou, Y., Huang, T., Huang, G., Zhang, N., Kong, X.Y., Cai, Y.D.: Prediction of protein N-formulation and comparison with N-acetylation based on a feature selection method. *Neuro Comput.* **217**, 53–62 (2016)
13. Fung, C.J., Zhu, Q.: FACID: a trust-based collaborative decision framework for intrusion detection networks. *Adhoc Netw.* **53**, 17–31 (2016)
14. Ganapathy, S., Jaisankar, N., Yogesh, P., Kannan, A.: An intelligent system for intrusion detection using outlier detection. In: *IEEE Conference on Recent Trends in Information Technology*, pp. 3–5 (2011)
15. Ganapathy, S., Vijayakumar, P., Yogesh, P., Kannan, A.: An intelligent CRF based feature selection for effective intrusion detection. *Int. Arab J. Inf. Technol.* **13**(1), 44–50 (2016)
16. Subba, B., Biswas, S., Karmakar, S.: Intrusion detection in mobile Ad-hoc networks: Bayesian game formulation. *Eng. Sci. Technol.* **19**, 782–799 (2016)
17. Zorarpac, E., Ozel, S.A.: A hybrid approach of differential evolution and artificial bee colony for feature selection. *Exp. Syst. Appl.* **62**, 91–103 (2016)

18. Pölsterl, S., Conjeti, S., Navab, N., Katouzian, A.: Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection. *Artif. Intell. Med.* **72**, 1–11 (2016)
19. Raza, M.S., Qamar, U.: An incremental dependency calculation technique for feature selection using rough sets. *Inf. Sci.* **343–343**, 41–65 (2016)
20. Krawczyk, B., Wozniak, M.: Dynamic classifier selection for one-class classification. *Knowl. Based Syst.* **107**, 43–53 (2016)
21. KDD Cup 1999 Intrusion Detection Data (2010). <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
22. Usha, G., Rajesh Babu, M., Saravana Kumar, S.: Dynamic anomaly detection using cross layer security in MANET. *Comput. Electr. Eng. Elsevier* **59**, 231–241 (2017)

# A Novel LtR and RtL Framework for Subset Feature Selection (Reduction) for Improving the Classification Accuracy



Sai Prasad Potharaju and M. Sreedevi

**Abstract** Preprocessing is one of the data mining steps after data collection. There are several issues need to be addressed in preprocessing stage of data mining. One among them is feature selection (FS) or feature reduction (FR). There are several approaches available for handling issues of FS and FR. Those methods are categorized as filter, wrapper, and embedded modes. In this research, we introduce a novel filter-based feature selection framework called LtR (left to right) and RtL (right to left) based on symmetrical uncertainty (SU). Our method generates K-subset of features such that each subset has the finite number of unique features in it. Each subset is analyzed using various classifiers (Jrip, OneR, Ridor, J48, SimpleCart, Naive Bayes, IBk) and compared with the existing filter-based FS methods: information gain (IG), ReliefF (Rel), chi-squared attribute evaluator (Chi), and gain ratio attribute evaluator (GR). Experimental analysis revealed that minimum one of the subsets performs better than some of the existing methods.

**Keywords** Data mining · Preprocessing · Feature selection · Filter Symmetrical uncertainty

## 1 Introduction

The concept of data mining (DM) is gaining popularity in the world of commerce, business activities, health care industry, education institutes, and much more [1]. It is impractical to manage the data by traditional methodology due to the rapid and huge production of data by various sources. DM enables the decision-makers to take more accurate decision for improving the day-to-day activities in their respective

---

S. P. Potharaju (✉) · M. Sreedevi  
Department of CSE, K L University, Guntur 522502, Andhra Pradesh, India  
e-mail: psaiprasadcse@gmail.com

M. Sreedevi  
e-mail: msreedevi27@kluniversity.in

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_20](https://doi.org/10.1007/978-981-13-1708-8_20)

field. DM can be defined as the collection of intelligent algorithms to get interesting patterns from the dataset.

Before applying DM algorithms, a lot of work is expected for better results. One of such activity is data preprocessing [2]. In general, 10% of work involved in DM and remaining 90% of work involved in data preprocessing and post-processing. Data preprocessing is nothing but preparing the data in a suitable format which is required for DM. Sometimes, data present in different places need to be gathered into single location [3].

After collecting the data, they need to be cleaned. Several researchers have worked on data cleaning and presented nice algorithms to clean the data automatically, i.e., fixing inconsistent values, missing values, and mislabeled data to avoid manual intervention. After cleaning the data, they have to be represented in a suitable format according to mining techniques (association rule mining—ARM, classification, clustering, regression). After formatting, data need to be normalized in few cases to avoid misclassification results. In most of the cases, these collected data undergo with noise in it. Feature reduction is a technique to remove the noisy features and redundant features [4]. Feature selection is also a key component in data preprocessing. FS is also called as attribute selection or variable selection or variable subset selection or relevant feature selection for the construction of model [5].

FS technique is generally applied for simplification of model interpretation, to make them easy to understand by various users, to reduce the training time, and also to enhance the model generation by reducing over-fitting. The central idea for FS technique is to reduce the training time and memory consumption as the collected data may encounter irrelevant and duplicated features in it, because those features do not give an extra strength or influence the model.

Out of 'N' features in a dataset, not all features are useful for model generation. Only a few features can influence the learning model. To identify these useful features, there are few traditional techniques available in research. Those are filter, wrapper, and embedded approaches [5]. FS problem can be defined as a method of drawing a subset of features which can increase the learning capacity of any classifier. If there are N features in whole data space, then  $2^N - 1$  proper subsets will be formed. Out of them, which subset is the best one? For this, a classifier can be used to test each subset. Then, the subset which gives the best result can be considered as the best subset of features [6]. It is not an easy task if the number of features is more, especially in the analysis of microarray dataset.

Feature subset can be evaluated using wrapper, filter, and hybrid methods. Wrapper can be called as a supervised approach. In this, selected subset is used for training and to find out the error rate on validation dataset. The subset which gives minimum error rate will be considered as the best subset. Filter can be called as unsupervised technique. In this, weight of each feature is measured using ranker algorithm. Based on threshold value, top-ranked features will be selected for model generation. The hybrid method combines both these mechanisms [7]. There are various FS techniques available in the literature which falls in either filter or wrapper or hybrid. Those are chi-square, mutual information (MU), symmetric uncertainty (SU), F-score. Our current article is based on SU.

Subsequent section of this article has some existing literatures, proposed methodology, experimental analysis with discussion of results, and finally concluded with the future suggestions.

## 2 Literature Review

In the literature, many researchers contributed various innovative methods for feature selection. For producing compatible classification results, a FAST clustering algorithm is proposed, which works in two phases. In the first phase, whole feature space is split into clusters (subsets) using graph theory concepts. In the next phase, strong features are derived by applying minimum spanning tree on each cluster formed in the first phase [8]. A novel technique called feature unionization is proposed by the researchers [9]. In their research, instead of removing the features, most dominant features are combined and formed new features from it. FS concept is applied in social networks for determining the criminals with their tweets or posts [10]. FS technique is applied in Web mining by the authors and proposed a new framework for classification of the Web page [11].

Authors presented SFS-LW (LW index combined with sequential forward search) algorithm to reduce the complexities of cross-validation scheme in the case of wrapper approach. The experimental results exhibited the almost equal performance as wrapper approach [12]. In recent days, ensembling approaches are getting encouragement in DM. FS has also been applied in cloud computing and DDoS detection also. Multi-filter FS approach is proposed for DDoS detection in cloud computing. This approach is based on the ensemble technique. Researchers applied this technique on intrusion detection dataset and reduced 60% of features, thereby increasing the decision tree classification accuracy [13]. FS methods are gaining popularity in biological science and healthcare industries in recent days. To increase the performance of classifiers, it is applied on the cardiocograms dataset. For this, ReliefF, correlation-based technique, IG, and consistency-based methods are applied. For analyzing the subset of features by those methods, SVM is applied [14].

FS has also been applied in the field of biomedical case studies. Researchers applied correlation-based technique for classifying cancer data. The intention behind their research is that microarray dataset requires huge computational time, and it is unavoidable to find out the best and small subset of features. For their work, authors considered SU as the primary condition and SVM for analyzing the subset formed [15]. Several FS methods applied for analyzing microarray datasets and those are presented in the article [16]. Many available FS techniques targeted on identifying relevant features, but identifying relevant features is not sufficient for high-dimensional data. It is required to identify redundant features also [17].

In recent days, for the better diagnostic system, microarray technology has been used by many medical practitioners to identify the different tumors and to differentiate various types of cancers. But, thousands of features in microarray dataset affect the accuracy of classification. For better classification, SU-HSA (symmetrical

uncertainty-based harmony search algorithm) is proposed [18]. Recently, FS has been applied for intrusion detection systems for reducing the computational complexity. FS is applied for identification of Dos, Probe, and R2L attacks using chi-squared and IG. For the analysis of the subsets formed, Adaboost, NB, and J48 are considered [19].

However, for the current research, we considered SU as the main criterion. SU can be defined as follows:

$$SU = 2 * IG / (H(Y) + H(X)),$$

$H(X)$  is the entropy of X

$H(Y)$  is the entropy of Y

SU takes the value in the range [0, 1]. SU value 1 indicates that one attribute can predict completely others, and 0 indicates two attributes are uncorrelated. Our proposed methodology is inspired from ensembling (Adaboost, Boosting) approach. If a weak or average attribute is combined with strong feature, there is a chance for increasing the accuracy of classifier with those ensembled features [20]. Our method was tested with the real-time dermatology dataset available at UCI machine learning repository [21].

### 3 Proposed Methodology

The objective of our proposed methodology is to reduce the feature space. If there are 'N' features in a dataset and we want to select top 'K' features without any repetitions from those 'N,' in such situation, total  $C(N, K)$  number of subsets can be formed. Analyzing those many subsets over high-dimensional dataset is not a simple job. Otherwise, filter-based ranking techniques can be applied to generate the rank for each feature and then top 'K' features can be selected. Other than the features generated by the existing techniques, we have proposed a novel LtR and RtL framework for generating subset of features. Proposed method is as follows.

1. Find out the symmetric uncertainty (SU) value (Weight) of every feature and arrange them in descending order as per its weight.
2. Define the total # attributes (TN), whose SU value is greater than zero.
3. Define the # subsets (S) to be formed, such that each subset has TN/S features.
4. Store the first TN/S features from LtR (left to right) in descending order.
5. Store the next TN/S features from RtL (right to left) in descending order.
6. Repeat Step 4 and then Step 5 for remaining features until all the features are stored.
7. Store all the vertically first-level features in the first subset, and then second-level features in the second subset, and so on.
8. If all the subsets have an equal number of features, then stop. Otherwise, remove the last feature from the subset which has an extra feature.

Sample CPP code to form the subset of features can be found here (open the URL: <https://saiprasadcomp.files.wordpress.com/2016/10/cpp-progrmacode.pdf>).

For experimenting the above method, we considered dermatology dataset available at UCI machine learning repository. Table 1 describes the dataset with SU value of each feature and rank of each feature derived by IG, Chi, Rel, GR.

From Table 1, it is found that subset S41 has an extra feature which has to be discarded. After this process, group all first-order features in subset S41, second-order features in subset S42, third-order features in subset S43, and fourth-order features in subset S44.

Table 2 lists the features in each subset.

## 4 Experiment

For experimenting the proposed methodology, we considered  $k=3, 4, 5$ , i.e., we formed 3 subsets of features (Table 3), 4 subsets of features (Table 4), and 5 subsets of features (Table 5).

For testing the strength of each subset of features, an equal number of top features derived by the existing techniques (IG, Chi, Rel, GR) are taken. S31, S32, S33 subsets of features have 11 features in it. So, top 11 features derived by the existing techniques are considered to measure the strength of those subsets. In the same fashion, all other subsets are measured by analyzing with Jrip, OneR, Ridor, J48, SimpleCart, Naive Bayes, IBk classifiers.

## 5 Results and Discussion

Accuracy of each classifier with the each subset of features is given in this section.

From Table 6, it is clear that S31 subset of features recorded enhanced accuracy with the Jrip, OneR, Ridor. It is found that S32 subset of features also displayed increased performance with all classifiers when compared with the existing feature selection techniques. With this 3-subset approach, maximum 33% of features can be trained for model generation.

From Table 7, it has been observed that almost all subsets of features recorded enhanced performance with all the classifiers when compared with the existing feature selection techniques. With this 4-subset approach, maximum 25% of features can be trained for model generation.

From Table 8, subsets S51 and S53 performed better than all the existing techniques, and S54 performed better than IG, Chi, GR when analyzed with Jrip. S51 recorded better accuracy than all and S52 and S55 better than Chi and GR when analyzed with OneR. Remaining subsets strength can also be interpreted in similar fashion. With this 5-subset approach, maximum 20% of features can be trained for model generation. To prove the strength of the proposed method, same framework is

**Table 1** Dataset description with SU value of each feature and rank of each feature derived by IG, Chi, Rel, GR

Rank	SU value	Feature no. by SU	Feature no. by IG	Feature no. by GR	Feature no. by Chi	Feature no. by Rel
1	0.4778	21	21	12	33	21
2	0.4672	22	20	29	29	33
3	0.4489	20	22	33	27	22
4	0.4328	33	33	15	12	20
5	0.4291	29	29	27	15	28
6	0.427	27	27	31	31	27
7	0.426	12	12	6	25	29
8	0.4188	25	25	25	6	6
9	0.4147	6	6	8	22	12
10	0.3944	8	8	22	20	16
11	0.3739	15	9	21	8	25
12	0.3288	9	16	30	21	8
13	0.3197	28	15	20	30	15
14	0.2979	16	28	7	16	9
15	0.2904	10	10	24	9	4
16	0.28	24	24	10	7	14
17	0.2505	14	14	28	10	10
18	0.2244	5	5	34	34	5
19	0.2159	31	26	9	28	24
20	0.2094	26	3	14	24	3
21	0.1868	7	31	16	26	26
22	0.1825	30	19	5	14	19
23	0.1726	23	23	23	3	7
24	0.1692	3	7	26	5	11
25	0.1447	34	30	11	19	2
26	0.1441	19	2	4	23	31
27	0.1341	4	4	3	2	18
28	0.1301	2	34	19	4	23
29	0.1066	11	11	2	11	30
30	0.0641	1	1	13	1	17
31	0.0597	13	13	1	13	34
32	0.0495	17	18	17	18	1
33	0.0483	18	17	18	17	13
34	0	32	32	32	32	32

# Total features is 34

# Total features whose SU value is greater than zero (TN) is 33

*Note* Feature no. 32 has SU value zero. It has to be discarded

Assume # subsets (S) 4; then each subset has  $33/4 = 8$  features in it

According to the proposed methodology, features in each subset will be formed as given in Table 2



**Table 2** Subsets of features when S = 4

	First-order features	Second-order features	Third-order features	Fourth-order features	Direction
	21	22	20	33	LtR
	25	12	27	29	RtL
	6	8	15	9	LtR
	24	10	16	28	RtL
	14	5	31	26	LtR
	3	23	30	7	RtL
	34	19	4	2	LtR
	17	13	1	11	RtL
	18				LtR
Subset ID	S41	S42	S43	S44	

**Table 3** Features with 3 Subsets

Subset ID	Features in it
S31	21, 27, 12, 9, 28, 5, 31, 3, 34, 1, 13
S32	22, 29, 25, 15, 16, 14, 26, 23, 19, 11, 17
S33	20, 33, 6, 8, 10, 24, 7, 30, 4, 2, 18

**Table 4** Features with 4 Subsets

Subset ID	Features in it
S41	21, 25, 6, 24, 14, 3, 34, 17
S42	22, 12, 8, 10, 5, 23, 19, 13
S43	20, 27, 15, 16, 31, 30, 4, 1
S44	33, 29, 9, 28, 26, 7, 2, 11

**Table 5** Features with 5 Subsets

Subset ID	Features in it
S51	21, 8, 15, 26, 7, 1
S52	22, 6, 9, 31, 30, 11
S53	20, 25, 28, 5, 23, 2
S54	33, 12, 16, 14, 3, 4
S55	29, 27, 10, 24, 34, 19

applied on 5 more real-time datasets. Those result analyses can be found here (open the URL: <https://saiprasadcomp.files.wordpress.com/2016/10/result-analysis.pdf>).

**Table 6** Performance analysis with 3 subsets

	IG	CHI	GR	REL	S31	S32	S33
Jrip	64.48	83.60	82.78	72.40	<b>86.06</b>	85.24	82.24
OneR	49.72	48.90	49.72	49.72	<b>49.72</b>	49.45	<b>49.72</b>
Ridor	78.68	83.06	83.06	77.04	<b>88.79</b>	87.43	82.78
J48	78.68	83.33	81.69	74.86	87.43	<b>88.52</b>	83.06
SC	79.50	83.33	83.06	75.95	87.97	<b>89.61</b>	82.78
NB	79.23	85.51	83.87	79.23	90.43	<b>90.71</b>	84.42
IBK	80.05	85.79	83.06	80.32	84.42	<b>87.43</b>	85.71

**Table 7** Performance analysis with 4 subsets

	IG	CHI	GR	REL	S41	S42	S43	S44
Jrip	59.83	68.03	68.03	75.13	84.15	68.57	<b>87.97</b>	82.51
OneR	49.72	48.90	48.90	50.27	<b>50.27</b>	49.72	47.54	49.72
Ridor	75.13	68.57	68.57	76.22	80.05	78.68	<b>88.79</b>	80.32
J48	75.95	68.57	68.57	76.22	86.06	80.87	<b>91.53</b>	84.15
SC	74.86	68.57	68.57	77.59	85.24	80.60	<b>90.98</b>	85.24
NB	74.86	69.12	69.12	78.41	86.61	80.32	<b>91.25</b>	86.33
IBK	75.95	69.12	69.12	78.14	82.51	80.60	<b>88.25</b>	84.15

**Table 8** Performance analysis with 5 subsets

	IG	CHI	GR	REL	S51	S52	S53	S54	S55
Jrip	59.28	69.12	69.12	71.85	85.51	54.64	<b>86.06</b>	70.49	53.82
OneR	50.27	49.18	49.18	50.27	<b>50.27</b>	49.45	49.18	49.18	49.45
Ridor	74.59	68.57	68.57	75.95	86.61	70.21	<b>87.15</b>	74.04	59.28
J48	76.22	68.85	68.85	76.77	87.70	70.76	<b>87.70</b>	74.31	65.40
SC	75.13	68.85	68.85	75.40	<b>88.25</b>	70.21	87.43	74.31	63.93
NB	74.86	69.12	69.12	78.41	<b>87.97</b>	70.21	87.43	76.5	65.30
IBK	76.22	69.12	69.12	77.59	85.51	69.67	<b>86.61</b>	76.77	64.20

## 6 Conclusion

In this study, a novel LtR and RtL feature subset selection framework has been proposed. With this framework, ‘K’ number of subsets of features are formed; each subset has minimum number of features without any repetition. All the subsets of features are tested using Jrip, OneR, Ridor, J48, SimpleCart, Naive Bayes, IBk classifiers, and respective results are compared with the existing feature selection techniques. Displayed result shows that one of the subsets and in some cases more than one subset recorded improved results than the existing approaches. With this, we conclude that instead of selecting features using the existing methods, depending on the requirement, the proposed technique can be used to form the subset of features for

improved prediction. The same framework can be tested using Hadoop framework to minimize the comparison time.

## References

1. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (2016)
2. Goswami, S., Chakrabarti, A.: Feature selection: a practitioner view. *Int. J. Inf. Technol. Comput. Sci.* **6**, 66–77 (2014). <https://doi.org/10.5815/ijitcs.2014.11.10>
3. Amarnath, B., Balamurugan, S., Alias, A.: Review on feature selection techniques and its impact for effective data classification using uci machine learning repository dataset. *J. Eng. Sci. Technol.* **11**, 1639–1646 (2016)
4. Tang, J., Alelyani, S., Liu, H.: Feature selection for classification: a review. *Data Classificat. Algor. Appli.* **37** (2014)
5. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014). <https://doi.org/10.1016/j.compeleceng.2013.11.024>
6. Kumar, V.: Feature selection: a literature review. *Smart Comput. Rev.* **4**. <https://doi.org/10.6029/smartcr.2014.03.007>
7. Singh, B., Kushwaha, N., Vyas, O.P.: A feature subset selection technique for high dimensional data using symmetric uncertainty. *J. Data Anal. Inf. Process.* **02**, 95–105 (2014). <https://doi.org/10.4236/jdaip.2014.24012>
8. Song, Qimao, Ni, Jingjie, Wang, Guangtao: A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Trans. Knowl. Data Eng.* **25**, 1–14 (2013). <https://doi.org/10.1109/TKDE.2011.181>
9. Jalilvand, A., Salim, N.: Feature unionization: a novel approach for dimension reduction. *Appl. Soft Comput.* **52**, 1253–1261 (2017). <https://doi.org/10.1016/j.asoc.2016.08.031>
10. Cesur, R., Ceyhan, E.B., Kermen, A., Sağiroğlu, Ş.: Determination of potential criminals in social network. *Gazi Univ. J. Sci.* **30**, 121–131 (2017)
11. Mangai, J.A., Santhosh Kumar, V., Appavu alias Balamurugan, S.: A novel feature selection framework for automatic web page classification. *Int. J. Automat. Comput.* **9**, 442–448. <https://doi.org/10.1007/s11633-012-0665-x> (2012)
12. Liu, C., Wang, W., Zhao, Q., Shen, X., Konan, M.: A new feature selection method based on a validity index of feature subset. *Pattern Recogn. Lett.* **92**, 1–8 (2017). <https://doi.org/10.1016/j.patrec.2017.03.018>
13. Osanaiye, O., Cai, H., Choo, K.-K.R., Dehghantaha, A., Xu, Z., Dlodlo, M.: Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *EURASIP J. Wirel. Commun. Netw.* <https://doi.org/10.1186/s13638-016-0623-3> (2016)
14. Silwattananusarn, T., Kanarkard, W., Tuamsuk, K.: Enhanced classification accuracy for cardiocogram data with ensemble feature selection and classifier ensemble. *J. Comput. Commun.* **04**, 20–35 (2016). <https://doi.org/10.4236/jcc.2016.44003>
15. Piao, Y., Piao, M., Park, K., Ryu, K.H.: An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics* **28**, 3306–3315 (2012). <https://doi.org/10.1093/bioinformatics/bts602>
16. Bolón-Canedo, V., Sánchez-Marroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. *Inf. Sci.* **282**, 111–135 (2014). <https://doi.org/10.1016/j.ins.2014.05.042>
17. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* **5**, 1205–1224 (2004)
18. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.* **5**, 1205–1224 (2004)

19. Patil, P., Attar, V.: Intelligent detection of major network attacks using feature selection methods. In: Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20–22, 2011. Springer, pp. 671–679 (2012)
20. Potharaju, S.P., Sreedevi, M.: Ensembled rule based classification algorithms for predicting imbalanced kidney disease data. *J. Eng. Sci. Technol. Rev.* **9**(5), 201–207 (2016)
21. <https://archive.ics.uci.edu/ml/machine-learning-databases/dermatology/>

# Gradient-Based Swarm Optimization for ICA



Rasmikanta Pati, Vikas Kumar and Arun K. Pujari

**Abstract** Blind source separation (BSS) is one of the most interesting research problems in signal processing. There are different methods for BSS such as principal component analysis (PCA), independent component analysis (ICA), and singular value decomposition (SVD). ICA is a generative model of determining a linear transformation of the observed random vector to another vector in which the transformed components are statistically independent. Computationally, ICA is formulated as an optimization problem of contrast function, and different algorithms for ICA differ among themselves on the way the contrast function is modeled. Several optimization techniques such as gradient descent and variants, fixed-point iterative methods are employed to optimize the contrast function which is nonlinear, and hence, determining global optimizing point is most often impractical. In this paper, we propose a novel gradient-based particle swarm optimization (PSO) method for ICA in which the gradient information along with the traditional velocity in swarm search is combined to optimize the contrast function. We show empirically that, in this process, we achieve better BSS. The paper focuses on the extraction of one by one source signal like deflation process.

**Keywords** ICA · Contrast function · Optimization · Gradient Particle swarm optimization

---

R. Pati (✉)  
SUIIT, Sambalpur University, Sambalpur, India  
e-mail: rkpati@suiit.ac.in

V. Kumar · A. K. Pujari  
School of CIS, University of Hyderabad, Hyderabad, India  
e-mail: vikas007bca@gmail.com

A. K. Pujari  
e-mail: akpujari@curaj.ac.in

A. K. Pujari  
Central University of Rajasthan, Kishangar, Ajmer 305817, Rajasthan, India

## 1 Introduction

The independent component analysis (ICA) is one of the most prominent methods of data analysis and has been widely used in signal processing, pattern recognition, and machine learning. In signal processing, ICA is essentially viewed as a computational method for separating multivariate signals into additive components and has been applied in many contexts. ICA is extensively used in pattern recognition and image analysis mainly in applications like face recognition, object recognition, image filtering, embedding in feature space. ICA is similar to PCA in many respect, but unlike PCA, ICA attempts to determine which are independent. The inherent advantage of ICA is its ability to recover source (or unobserved signal) from observed mixture. ICA is also popularly known as a method of blind source separation (BSS). It is called blind because we do not have information about the source signals or the mixing method. For BSS, it is assumed that signals from source can be mixed linearly or nonlinearly. ICA attempts to separate source by some simple assumptions of their statistical properties. In ICA, data are represented by the random vector  $x$  and the components as the random vector  $s$ . It is to determine a transformation that maps observed data  $x$  into maximally independent components  $s$  for some measure of independence. The transformation is usually assumed to be linear, and the measure of independence can be a measure of non-Gaussianity.

Let us consider an observed  $M$ -dimensional discrete time signal where the  $n$ th sample is denoted by the column vector  $x(n)$ . The observed signal is the mixture of unknown  $N$ -dimensional source vector  $s(n)$  given by

$$x(n) = As(n) \tag{1}$$

where  $A$  is a linear mixing matrix. The input components are usually statistically dependent due to the mixing process, whereas the sources are not. If one succeeds in finding a matrix  $W$  that yields statistically independent output components  $y(n)$ , given by

$$y(n) = Wx(n) = WAs(n) \tag{2}$$

one can recover the original sources up to a permutation and constant scaling of the sources.  $W$  is called the demixing matrix, and finding the matrix is referred to as independent component analysis (ICA).

ICA computation involves determining the matrix  $W$  by a process of optimization of a non-convex optimization, and the widely adopted gradient descent algorithms [1] usually converge to a local optimizing point and seldom find the global optimizing point. As no global solution is guaranteed, most of ICA techniques exhibit random behavior yielding different results for different initial conditions and initial values of parameters. There are different approaches of estimating  $W$ . Maximization algorithm based on singular value decomposition (SVD) [2], gradient optimization of kurtosis function [3], and iterative method [5] of approximating  $W$ . In this paper, we are particularly examining the deflation-based source separation.

In this paper, a particle swarm optimization (PSO)-based ICA algorithm is presented to overcome the above problem. As an evolutionary computation technique and general global optimization tool, PSO was first proposed by Kennedy and Eberhart [4] which simulates the simplified social life models. Since PSO has many advantages over other heuristic techniques such as it can be easily implemented and has a great capability of escaping local optimal solutions [5], PSO has been applied successfully in many computer science and engineering problems. Another drawback of gradient-based methods [9, 14] is slow speed of convergence. PSO search is preferred over gradient search when the nonlinear objective function is multimodal and there are a large number of local optimizing solutions. In such a situation, gradient-based search gets stuck at a local optimizing point where as population based technique search through a broader area ensuring t possibility of reaching global optimizing solution. So an obvious question is whether one can combine gradient information of search direction together with the velocity computed by local/global best solution to enhance the search. Taking advantages of both gradient search and population based search, we propose a method which blends gradient search with PSO. We show empirically that by this process, we can have an efficient method of ICA computation.

The rest of the paper is organized as follows. In Sect. 2, we briefly review the cumulants, reference signal, contrast function. Section 3 discusses the optimization technique and iterative procedure for ICA. A brief introduction about particle swarm optimization (PSO) is given in Sect. 4. Section 5 describes our proposed methods termed as PSOAS for ICA. Experimental analysis of the proposed method is reported in Sect. 6. Finally, Sect. 7 concludes and indicates several issues for future work.

## 2 Cumulant-Based Contrast Optimization

A contrast function is any nonlinear function which is invariant to permutation and scaling matrices and attains its minimum value in correspondence of the mutual independence among the output components. Many contrast functions for ICA has been proposed in the literature, mainly based on information theoretical principles such as maximum likelihood, mutual information, marginal entropy, and negentropy, as well as related non-Gaussianity measures. Among them, the kurtosis (normalized fourth-order marginal cumulant) is arguably the most common statistics used in ICA, even if skewness has also been proposed.

Statistical properties of the output dataset can be described by its moments or, more conveniently, by its cumulants. Since the data have zero mean, the sample cumulants up to order four can be written in the following way.

$$\begin{aligned}
 C_i^{(y)} &= 0 \\
 C_{ij}^{(y)} &= \langle y_i y_j \rangle \\
 C_{ijk}^{(y)} &= \langle y_i y_j y_k \rangle \\
 C_{ijkl}^{(y)} &= \langle y_i y_j y_k y_l \rangle - \langle y_i y_j \rangle \langle y_k y_l \rangle - \langle y_i y_k \rangle \langle y_j y_l \rangle - \langle y_i y_l \rangle \langle y_j y_k \rangle
 \end{aligned}$$

with  $\langle \cdot \rangle$  indicating the mean over all data points.

Cumulants of a given order form a tensor. The diagonal elements characterize the distribution of single component and the fourth-order autocumulant,  $C_{iiii}^{(y)}$  is kurtosis of  $y_i$ . The cross-cumulants characterize the statistical dependencies between components. Thus, if and only if, all components are statistically independent, the off-diagonal elements (or the cross-cumulants) vanish. ICA is equivalent to finding an unmixing matrix  $W$  that diagonalizes the cumulant tensors of the output data at least approximately. Though it is easy and trivial to achieve diagonalization of first- and second-order cumulants, there is no obvious way of diagonalizing higher-order cumulant tensors. The diagonalization of these tensors can only be done approximately, and we need to define an optimization criterion for this approximate diagonalization

The approximate diagonalization of the cumulant tensors of order three and order four is achieved by minimizing an objective function which is the sum of the squared third- and fourth-order off-diagonal elements. Since the sum of square of all elements of a cumulant tensor is preserved under any orthogonal transformation of the underlying data, one can equivalently maximize the sum over the diagonal elements instead of minimizing the sum over the off-diagonal elements. This is a contrast function as defined in [6]. Thus, the process can be viewed as an optimization problem with the following objective function.

$$J(y) = \frac{1}{\mu} \sum_{\alpha} (C_{\alpha\alpha\alpha}^{(y)})^2 + \frac{1}{\tau} \sum_{\alpha} (C_{\alpha\alpha\alpha\alpha}^{(y)})^2, \tag{3}$$

The objective function  $J$  is kurtosis [7, 8] and can be rewritten as a function of an orthogonal matrix  $U$  which is to be determined through the optimization process. Expressing the above criterion function in terms of  $U$  is not straightforward, and hence, another cumulant-based contrast function is defined as follows. This definition uses cumulant of order four only. Recently, *reference-based* contrast functions are proposed based on cross-statistics or cross-cumulants between the estimated outputs and reference signals. Reference signals are nothing but artificially introduced signals for facilitating the maximization of the contrast function. Due to the indirect involvement of reference signals in the iterative optimization process, these reference-based contrast functions have an appealing feature in common: The corresponding optimization algorithms are quadratic with respect to the searched parameters.

$$\begin{aligned}
 C_z\{y\} &\triangleq \text{Cum}\{y, y, z, z\} \\
 &= E\{y^2 z^2\} - E\{y^2\}E\{z^2\} - 2E^2\{yz\}
 \end{aligned} \tag{4}$$



where  $E\{\cdot\}$ , denotes the expectation value and  $z$  is the reference signal. We consider another(reference) separation matrix  $V$  and  $z(n) = Vx(n)$ . We now define the contrast function explicitly in terms of  $W$  and  $V$  as follows.

$$I(W, V) = \left| \frac{C_z\{y\}}{E\{(y)^2\}E\{(z)^2\}} \right|^2 \quad (5)$$

where  $y(n) = Wx(n)$  and  $z(n) = Vx(n)$ .

### 3 Optimization Method

There have been a umpteen number of proposals to optimize the contrast function defined previously. In order to avoid an exhaustive search in the whole space of orthogonal matrices, a gradient ascent on  $J(U)$  is normally used. The gradient of a function is the vector of its partial derivatives. It gives a direction of the maximum increase in the function leading to an update rule for  $U$  that looks like

$$U(k+1) \leftarrow U(k) + \lambda(k)\nabla J|_{U(k)} \quad (6)$$

where  $\nabla J|_{U(k)}$  denotes either the natural, or relative gradient of  $J$  with respect to  $U$ , evaluated at  $U = U(k)$ .

Gradient ascent and its variants start with a random seed point and move from one point to another in the gradient direction. The performance of all gradient-based approach depends on a factor such as step size  $\lambda$  and initial seed point  $U(0)$ . The rate of convergence highly depends on the selection of step size, and an improper step size may lead to the poor performance and stability of the algorithm.

Use of a gradient-based maximization supposes that the algorithm will not be trapped in a spurious maximum, leading to  $U^*$ , that does not correspond to a satisfactory solution for the BSS problem (still mixing). Various authors such as [9, 10] have noted that the usual ICA contrast functions may have such spurious maxima if several source distributions are multimodal. For instance, Cardoso in [11] shows this phenomenon for the likelihood-based contrast function. More recently, Vrins et al. [12] have given an intuitive justification regarding the existence of spurious maxima when the opposite of the output marginal entropies is used for the contrast function.

A simple gradient search algorithm for the maximization of kurtosis-based contrast  $J$  is given in Algorithm 1. It is shown in [3] that Algorithm 1 may diverge unacceptably leading to a numerical overflow if a great number of iterations are required by the algorithm.

**Algorithm 1:**


---

**input** :  $x(n)$ : Observed signal  
**output**:  $v$ : Separation vector  
Initialize randomly  $U_0$   
**for**  $k = 0, 1, \dots, k_{max} - 1$  **do**  
     $d_k = \nabla I(U_k, U_k)$   
     $\alpha_k = \underset{\alpha}{\operatorname{argmax}} I(U_k + \alpha d_k, U_k)$   
     $U_{k+1} \leftarrow U_k + \alpha_k d_k$   
**end**

---

The separating property is not affected by a scaling factor, because of unavoidable scaling ambiguity in BSS [3]. It is common in BSS to impose the unit-power constraint  $E\{|y(n)|^2\} = 1$ . It is known that the unit-power constraint is equivalent to a unit-norm constrain on the separating vector  $v$ . A modified algorithm to avoid the drawback of Algorithm 1 is proposed in [3] by normalization of the separating vector  $v$  after every gradient iteration update. The points found after renormalizing the above algorithm belong to unit sphere. The main flow of the modified algorithm can be found in Algorithm 2. We use this method in our comparative studies in the later section.

The output obtained after the maximization process should be closer to the source signal rather than the reference one. With this aim, a modification is proposed in [9] where the reference vector is updated after each iteration by the output signal computed in the previous iteration.

**Algorithm 2:**


---

**input** :  $x(n)$ : Observed signal  
**output**:  $v$ : Separation vector  
Initialize  $U_0$  .  
**for**  $k = 0, 1, \dots, k_{max} - 1$  **do**  
     $d_k = \nabla_1 I(U_k, U_k)$   
     $\alpha_k = \underset{\alpha}{\operatorname{argmax}} I(U_k + \alpha d_k, U_k)$   
     $\tilde{U}_{k+1} \leftarrow U_k + \alpha_k d_k$   
     $\hat{U}_{k+1} \leftarrow \frac{\tilde{U}_{k+1}}{(E\{|\tilde{U}_{k+1}\{x(n)\}^2\})^{\frac{1}{2}}}$   
     $U_{k+1} \leftarrow \hat{U}_{k+1}$   
**end**

---

## 4 Particle Swarm Optimization

Particle swarm optimization (PSO) is a well-known population-based search. The PSO algorithm works by simultaneously maintaining several candidate solutions in the search space to find the global optimum, where the movement is influenced by the social component and the cognitive component of the particle. The most characteristic feature of PSO and its variants is that the search trajectory is influenced by the best solutions (local best and global best) obtained so far in the search to determine the next solution. Each individual particle has a velocity vector  $v_i$ , a position vector  $x_i$ , personal best  $pb_i$  that the particle encountered so far and neighborhood best  $lb_i$  means the best position that all particles have encountered so far among the neighborhood  $N_i$  of particle  $i$ . The position and velocity of each particle are updated as follows.

$$\begin{aligned}v_i^{t+1} &= v_i^t + c_1 r_1 (pb_i^t - x_i^t) + c_2 r_2 (lb_i^t - x_i^t) \\x_i^{t+1} &= x_i^t + v_i^{t+1}\end{aligned}$$

where  $c_1, c_2$  are acceleration coefficient and  $r_1, r_2 \in [0, 1]$  are uniformly distributed random numbers.

PSO is particularly attractive for its ability to yield global optimizing point with the fast converging rate. However, it does not use the gradient information which is very crucial for optimization.

## 5 PSOAS: The Proposed Method

In this section, we discuss the method of blending swarm search with gradient-based optimization for ICA. Unlike PSO, in the proposed algorithms, the velocity component of the particle is updated in every iteration with gradient direction along with the social influence. The search direction of the particle is a combination of gradient direction and the direction of global best. There have been some earlier proposals which use PSO to solve the ICA problem [6, 13, 14]. In the literature, many variants of gradient-based PSO exist [15–17]. Some researchers has combined a gradient factor with search direction computed by personal best and global best whereas in [18] terminates gradient search is initiated after termination of PSO.

Algorithm 3 describe the detail procedure related to applicability of PSO on Algorithm 1.

**Algorithm 3:** Gradient-based PSO

**input** :  $x(n)$ : Observed signal,  $S$ : Swarm size and  $\delta$ : Trade-off parameter  
**output**:  $U^{best}$ : Separation vector

Initialize  $U_0$  and the corresponding reference signal  $z_0^p(n) = U_0^p x(n), \forall 1 \leq p \leq S$ .

**for**  $k = 0, 1, \dots, k_{max} - 1$  **do**  
     $I_p = I(U_k^p, U_k^p), \forall 1 \leq p \leq S$   
     $best = \underset{p}{argmax} I_p$   
     $d_k^p = \nabla I(U_k^p, U_k^p)$   
     $\alpha^p = \underset{\alpha}{argmax} I(U_k^p + \alpha d_k^p, U_k^p)$   
     $\tilde{U}_{k+1}^p \leftarrow U_k^p + \alpha^p (\delta d_k^p + (1 - \delta)(U^{best} - U_k^p))$   
     $\tilde{U}_{k+1}^p \leftarrow \frac{\tilde{U}_{k+1}^p}{(E\{|\tilde{U}_{k+1}^p x(n)|^2\})^{\frac{1}{2}}}$   
     $U_{k+1}^p \leftarrow \tilde{U}_{k+1}^p$   
**end**

We modify Algorithm 3 with iterative updates to get another alternative, Algorithm 4 as follows.

**Algorithm 4:** Gradient-based PSO with Fixed-point update

**input** :  $x(n)$ : Observed signal,  $S$ : Swarm size and  $\delta$ : Trade-off parameter  
**output**:  $U^{best}$ : Separation vector

Initialize  $U_0$  and the corresponding reference signal  $z_0^p(n) = U_0^p x(n), \forall 1 \leq p \leq S$ .

**for**  $k = 0, 1, \dots, k_{max} - 1$  **do**  
     $\tilde{U}_0^p = U_k^p, \forall 1 \leq p \leq S$   
    **for**  $l = 0, 1, \dots, l_{max} - 1$  **do**  
         $I_p = I(\tilde{U}_l^p, U_k^p), \forall 1 \leq p \leq S$   
         $best = \underset{p}{argmax} I_p$   
         $\tilde{d}_l^p = \nabla_1 I(\tilde{U}_l^p, U_k^p)$   
         $\tilde{\alpha}^p = \underset{\alpha}{argmax} I(\tilde{U}_l^p + \alpha \tilde{d}_l^p, U_k^p)$   
         $\tilde{U}_{l+1}^p \leftarrow \tilde{U}_l^p + \tilde{\alpha}^p (\delta \tilde{d}_l^p + (1 - \delta)(U^{best} - \tilde{U}_l^p))$   
         $\tilde{U}_{l+1}^p \leftarrow \frac{\tilde{U}_{l+1}^p}{(E\{|\tilde{U}_{l+1}^p x(n)|^2\})^{\frac{1}{2}}}$   
    **end**  
     $U_{k+1}^p \leftarrow \tilde{U}_{l_{max}}^p$   
**end**

## 6 Simulation

This section discusses the experimental setup and reports the results. We conducted experiments on a variety of synthetic datasets. Complex-valued, independent, and identically distributed (i.i.d) QAM4 has been generated taking their values in

**Table 1** Experimental results of each comparing algorithm in terms of average and median MSE

	Parameters		Number of samples							
			500		1000		5000		10000	
	$k_{max}$	$l_{max}$	PSOAS	GAS	PSOAS	GAS	PSOAS	GA	PSOAS	GAS
Average MSE	1000	1	<b>0.0066</b>	0.0114	<b>0.0007</b>	0.0047	<b>0.0003</b>	0.0039	<b>0.0003</b>	0.0043
	200	5	<b>0.0005</b>	0.0046	0.0316	<b>0.0214</b>	<b>0.0007</b>	0.0038	<b>0.0007</b>	0.0049
	100	10	0.0161	<b>0.0098</b>	<b>0.0006</b>	0.0049	<b>0.0008</b>	0.0044	<b>0.0005</b>	0.0034
	50	20	<b>0.0010</b>	0.0050	<b>0.0236</b>	0.0301	<b>0.0005</b>	0.0049	<b>0.0007</b>	0.0048
	25	40	0.0171	<b>0.0109</b>	<b>0.0006</b>	0.0065	<b>0.0004</b>	0.0035	<b>0.0005</b>	0.0047
	10	100	<b>0.0010</b>	0.0035	<b>0.0342</b>	0.0407	<b>0.0004</b>	0.0045	<b>0.0005</b>	0.0040
	8	125	0.0271	<b>0.0099</b>	<b>0.0005</b>	0.0061	<b>0.0007</b>	0.0051	<b>0.0004</b>	0.0043
	5	200	<b>0.0007</b>	0.0038	<b>0.2894</b>	0.3538	<b>0.0004</b>	0.0053	<b>0.0004</b>	0.0045
Median MSE	1000	1	<b>0.0001</b>	0.0009	<b>0.0001</b>	0.0005	<b>0.0000</b>	0.0006	<b>0.0000</b>	0.0006
	200	5	<b>0.0000</b>	0.0008	<b>0.0002</b>	0.0010	<b>0.0000</b>	0.0006	<b>0.0000</b>	0.0007
	100	10	<b>0.0002</b>	0.0009	<b>0.0001</b>	0.0007	<b>0.0000</b>	0.0005	<b>0.0000</b>	0.0005
	50	20	<b>0.0001</b>	0.0007	<b>0.0002</b>	0.0012	<b>0.0000</b>	0.0006	<b>0.0000</b>	0.0006
	25	40	<b>0.0002</b>	0.0006	<b>0.0001</b>	0.0008	<b>0.0000</b>	0.0005	<b>0.0000</b>	0.0006
	10	100	<b>0.0001</b>	0.0004	<b>0.0003</b>	0.0018	<b>0.0000</b>	0.0006	<b>0.0000</b>	0.0006
	8	125	<b>0.0002</b>	0.0006	<b>0.0001</b>	0.0010	<b>0.0000</b>	0.0007	<b>0.0000</b>	0.0006
	5	200	<b>0.0001</b>	0.0006	<b>0.1834</b>	0.2570	<b>0.0000</b>	0.0006	<b>0.0000</b>	0.0006

$\{e^{i\pi/4}, e^{-i\pi/4}, e^{+i3\pi/4}, e^{-i3\pi/4}\}$  with equal probability  $\frac{1}{4}$ . For a different number of sample, a set of  $N = 3$  mutually independent and temporally i.i.d source has been generated. They have been mixed by a QL finite impulse response (FIR) filter with randomly driven coefficients of length 3 and with  $Q = 4$  sensors. The separating FIR separator has been searched with length  $D = N(L - 1) = 6$ .

To measure the performance of different algorithms, we have employed mean squared error (MSE) as an evaluation metric popularly used in blind source separation [3]. We report the equalization performance of the proposed method by taking the average and median values of MSE of 1000 trials. We compare our proposed methods with two well-known algorithms: Algorithm 1 and Algorithm 2 [3] with our Algorithm 3 and Algorithm 4, respectively.

Table 1 gives the comparative analysis of proposed method against state-of-the-art algorithms on different datasets. The best results among all comparing algorithm are highlighted in boldface. The row corresponding to value of  $k_{max} = 1000$  and  $l_{max} = 1$  reports the results provided by Algorithm 1 [3] and Algorithm 3 proposed in the present work. The remaining rows show the results provided by Algorithm 2 [3] and Algorithm 4 proposed in the present work. It can be seen from the Table 1 that the proposed method achieves better performance consistently than other comparing algorithms in terms of each evaluation metric. The following tables denote the algorithms, Algorithm 1 and Algorithm 2 of [3], as general algorithms (GAS) and our proposed algorithms, Algorithm 3 and Algorithm 4, as PSO algorithms (PSOAS).

## 7 Conclusion

In this paper, two new algorithms have been developed with PSO-based search for the purpose of maximizing the kurtosis contrast function. Particularly, Algorithm 4 allows two parameters to improve performance for practical purpose. The work also opens for future works with respect to genetic algorithm and source separation of complex-valued signals based on nonlinear autocorrelation. One can use genetic algorithm to see its practical purpose. As well as PSO may use in the scenario of nonlinear autocorrelation.

## References

1. Castella, M., Moreau, E.: A new method for kurtosis maximization and source separation. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2670–2673. IEEE (2010)
2. Kawamoto, M., Kohno, K., Inouye, Y.: Eigenvector algorithms incorporated with reference systems for solving blind deconvolution of mimo-iir linear systems. *IEEE Signal Process. Lett.* **14**(12), 996–999 (2007)
3. Castella, M., Moreau, E.: New kurtosis optimization schemes for miso equalization. *IEEE Trans. Signal Process.* **60**(3), 1319–1330 (2012)
4. Kennedy, J., Eberhart, R.: Particle swarm optimization (ps). In: Proceedings of IEEE International Conference on Neural Networks, Perth, Australia, pp. 1942–1948 (1995)
5. Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. *Nat. Comput.* **1**(2–3), 235–306 (2002)
6. Igual, J., Ababneh, J., Llinares, R., Miro-Borras, J., Zarzoso, V.: Solving independent component analysis contrast functions with particle swarm optimization. *Artif. Neural Netw. ICANN 2010*, 519–524 (2010)
7. Simon, C., Loubaton, P., Jutten, C.: Separation of a class of convolutive mixtures: a contrast function approach. *Signal Process.* **81**(4), 883–887 (2001)
8. Tugnait, J.K.: Identification and deconvolution of multichannel linear non-gaussian processes using higher order statistics and inverse filter criteria. *IEEE Trans. Signal Process.* **45**(3), 658–672 (1997)
9. Castella, M., Rhioui, S., Moreau, E., Pesquet, J.C.: Quadratic higher order criteria for iterative blind separation of a mimo convolutive mixture of sources. *IEEE Trans. Signal Process.* **55**(1), 218–232 (2007)
10. Boscolo, R., Pan, H., Roychowdhury, V.P.: Independent component analysis based on nonparametric density estimation. *IEEE Trans. Neural Netw.* **15**(1), 55–65 (2004)
11. Haykin, S.S.: *Unsupervised Adaptive Filtering: Blind Source Separation*, vol. 1. Wiley-Interscience (2000)
12. Vrins, F., Archambeau, C., Verleysen, M.: Entropy minima and distribution structural modifications in blind separation of multimodal sources. In: AIP Conference Proceedings, vol. 735, pp. 589–596. AIP (2004)
13. Krusienski, D.J., Jenkins, W.K.: Nonparametric density estimation based independent component analysis via particle swarm optimization. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05), vol. 4, pp. iv–357. IEEE (2005)
14. Li, H., Li, Z., Li, H.: A blind source separation algorithm based on dynamic niching particle swarm optimization. In: MATEC Web of Conferences. Volume 61., EDP Sciences (2016)
15. Borowska, B., Nadolski, S.: Particle swarm optimization: the gradient correction. (2009)

16. Noel, M.M., Jannett, T.C.: Simulation of a new hybrid particle swarm optimization algorithm. In: Theory, System (ed.) 2004, pp. 150–153. IEEE, Proceedings of the Thirty-Sixth Southeastern Symposium on (2004)
17. Vesterstrom, J.S., Riget, J., Krink, T.: Division of labor in particle swarm optimisation. In: Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on. Volume 2., IEEE (2002) 1570–1575
18. Szabo, D.: A study of gradient based particle swarm optimisers. PhD thesis, Masters thesis, Faculty of Engineering, Built Environment and Information Technology University of Pretoria, Pretoria, South Africa (2010)

# Empirical Evaluation of Inference Technique for Topic Models



Pooja Kherwa and Poonam Bansal

**Abstract** Topic modelling is a technique to infer themes and topic from a large collection of documents. Latent Dirichlet allocation is the most widely technique used in topic modelling literature. It is a model generative in nature with multinomial distribution to produce document, and then, again LDA is used as reverse process by estimating parameters to deduce topic and themes from unstructured documents. In topic modelling, many approximate posterior inference algorithms exist, and the most dominating inference techniques in LDA (latent Dirichlet allocation) are variational expectation maximization (VEM) and Gibbs sampling. In this paper, we are evaluating the performance of VEM and Gibbs sampling techniques on an Associated Press data set and Accepted Papers data set by fitting the topic model using latent Dirichlet allocation. In this experiment, we consider perplexity and entropy as significant metrics for the performance evaluation of topic models. In this, we found that for large data set like Associated Press data set with 2000 documents, variational inference is good inference technique and for small data set like Accepted Papers Gibbs sampling is the best choice for inference. Another advantage of Gibbs sampling is that it runs Markov chain and avoids getting trapped in local minima. Variational inference provides fast and deterministic solutions.

**Keywords** Topic model · Latent Dirichlet allocation · Inference · Gibbs sampling  
Multinomial distribution

---

P. Kherwa (✉) · P. Bansal

Maharaja Surajmal Institute of Technology, C-4, Janak-Puri, New Delhi 110058, India  
e-mail: Poona281280@gmail.com

P. Bansal

e-mail: pbansal89@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_22](https://doi.org/10.1007/978-981-13-1708-8_22)

237



## 1 Introduction

From last one and half decades, research work using topic model with latent Dirichlet allocation (LDA) is very popular in machine learning and natural language processing community for handling huge amount of unstructured data and annotating these data with themes and topic. Topic model captures meaningful co-occurrence of words and can reveal the underlying hidden semantic structure of corpus. They can be used to facilitate efficient browsing of a collection as well as large-scale analysis of text [1].

Topic models are different from other natural language processing approaches because they are based on the theory that every document contains a mixture of topics [2]. For example, a news article about a natural disaster may include topics about causes of such disasters, relief aid efforts and the damage/death toll. Probabilistic latent semantic [3] is the first mixed membership-based language model; under this model, the probability of appearance of the ‘ith’ word in a document is:

$$P(w_i|d) = \sum_{z \in Z} P(w_i|z)P(z|d) \quad (1)$$

One of the main drawbacks of PLSI is that it is not feasible to label unseen documents. This issue is resolved by latent Dirichlet allocation (LDA) [1]. In LDA, documents exhibit multiple topics and topic distribution also differs over documents. A major problem in using topic models and developing new models is the computational cost of calculating the posterior distribution. Therefore, a large body of work has considered approximate inference methods; the most popular methods are variational methods, specifically mean field methods, automatic differential variational inference [4] and Gibbs sampling based on Markov chain Monte Carlo.

In this paper, we are trying to analyse the performance of both approximate inference algorithms for topic modelling. The paper is organized as follows. In Sect. 2, we summarized the relevant previous work and give a detailed description of latent Dirichlet generative process of documents. In Sect. 3, various inference algorithm available with latent Dirichlet allocation is discussed. In Sect. 4, experimental set-up for evaluating the performance of inference on data set is described with evaluation measures. In Sect. 5, results and evaluation of inference algorithms are presented. Finally, the paper is concluded with future work in Sect. 6.

## 2 Background

To explore and browse modern digital libraries and World Wide Web, we need to develop necessary tools and automated methods. Topic models are such probabilistic models for exploring and revealing the semantic structure of document collection based on hierarchical Bayesian analysis [5]. The original text-motivated topic model

is given by Hofmann [3] who describes its mixed membership likelihood as a probability model for the latent semantic indexing [6]. LDA was developed in 2002/2003 in collaboration with Blei and Lafferty [1], and the term LDA has since become nearly synonymous with topic modelling in general [5]. In last two decades, LDA has given enough contribution in various fields of language technology and search technology for managing unstructured data. Popular contribution includes hierarchical formulation to know an unknown number of topics in LDA, topics that change over time [7] and correlated topic model [8].

Topic modeling has a wide range of applications in various fields such as text mining [9], image retrieval [10], social network analysis [11] and bioinformatics analysis [12, 13].

The computational cost of computing the posterior distribution is the major problem of topic modelling. Therefore, a large body of work has considered approximate inference methods, the most popular methods being variational methods, specifically mean field methods, and Markov chain Monte Carlo, particularly methods based on collapsed Gibbs sampling.

For decades, the dominant paradigm for approximate inference has been MCMC [14, 15]. MCMC sampling has evolved into an indispensable tool to the modern Bayesian statistician. Landmark developments include the Metropolis-Hastings algorithm [16, 17], the Gibbs sampler, [18] and its application to Bayesian statistics [14]. Neural network-based inference method has also a significant contribution in topic modelling literature [19–22].

Variational inference portrays Bayesian inference as an optimization problem. The advantage of variational inference is maximizing an explicit objective and being faster in most cases. Other work on variational inference includes mean field variational inference [5], collapsed variational inference [23], automatic differentiation variational inference [4] and expectation propagation [1].

Both the inference techniques have a wide variety of applications in the various fields of natural language processing.

### 3 Latent Dirichlet Allocations

Latent Dirichlet allocation is an attempt to infer semantic meaning from vocabulary of documents. LDA is based on generative process of multinomial distribution. It means distribution over distribution. Every corpus is a collection of multiple documents, and each document consists of multiple words. Each document has its own vocabulary. These documents and words are observed variables, and topic or themes of document are hidden variables. It is a complete unsupervised approach.

LDA is a probabilistic generative model with three-level structures as word, topic and document. In LDA, documents are distribution over topics and each topic is a distribution over words.

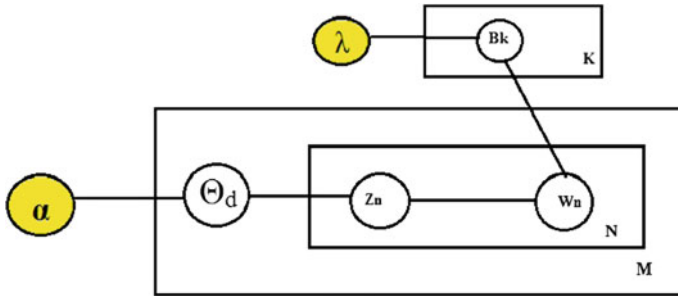


Fig. 1 Dirichlet plate notation [5]

In this generative model, a word  $w$  is an element of dictionary  $\{1, \dots, v\}$ , a document is represented with the sequence of  $N$  words and each word  $(W_1, \dots, W_N)$ ,  $W_n \in \{1, \dots, v\}$ . A corpus  $D$  is a collection of  $M$  documents.

Given an appropriate number of topics  $K$ , the generative process for a document  $d$  is as follows [5] (Fig. 1):

1. Sample a  $K$ -vector  $\theta_d$  from the Dirichlet distribution  $p(\theta|\alpha)$ , where  $\theta_d$  is the topic mixture proportion of document  $d$ .
2. For  $i = 1 \dots N_d$ , sample word  $w_i$  in the  $d$  from the document special multinomial distribution  $p(w_n|\theta_d, \beta)$ , where  $\alpha$  is a  $K$ -vector of Dirichlet parameter, and  $p(\theta|\alpha)$  is given as follows:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1} \tag{2}$$

Here,  $\beta$  is a  $K * V$  matrix of word probability, where  $\beta_{ij} = p(w_j = 1 | z_i = 1)$ ,  $i = 0, 1, \dots, K$ ;  $j = 0, 1, \dots, V$ .

Topics generated by LDA are independent because they are generated from Dirichlet distribution. LDA contains two Dirichlet random variables: the topic proportions  $\theta$  are distributions over topic indices  $\{1, \dots, k\}$ ; the topic  $\beta$  are distributions over vocabulary.

**Topic Distribution:** LDA uses Dirichlet distribution to generate multinomial distribution over topics. A Dirichlet distribution is a distribution over distribution; that is, instead of a single value, it provides a whole distribution for each sample drawn. The  $\alpha$  parameter influences the shape of this distribution.

**Term Distribution:** The  $\beta$  parameter influences the shape of this distribution. This parameter will work in the same way as  $\alpha$  did in topic distribution.

### 3.1 Calculating Posterior Distribution

This is the reverse process of LDA generative process for finding the hidden variables from observed variables. In this from a set of  $D$  documents and the observed words within each document, we want to infer the posterior distribution.

$$P(\vec{\theta}1 : D, z1 : D, 1 : N, \vec{\beta}1 : K | w1 : D, 1 : N, \alpha, \eta) = \frac{P(\vec{\theta}1 : D, \vec{z}1 : D, \vec{\beta}1 : k | \vec{w}1 : D, \alpha, \eta)}{\int \vec{\beta}1 : k \int \vec{\theta}1 : D \sum \vec{z} P(\vec{\theta}1 : D, \vec{z}1 : D, \vec{\beta}1 : k | \vec{w}1 : D, \alpha, \eta)} \quad (3)$$

These are most widely used inference techniques for solving this Bayesian inference.

1. Variational inference
2. Gibbs sampling (Markov chain Monte Carlo).

**Variational Inference:** Variational approximation can be very useful for Bayesian inference where intractable calculus problem occurs. Variational inference approximate is an intractable posterior distribution over hidden variables, such as Eq. (3). Theses hidden parameters are then approximated as close as possible to the true posterior. Variational inference has the advantage of maximizing an explicit objective and being faster in most cases.

**Gibbs Sampling:** The Gibbs sampler is type of Markov chain Monte Carlo distributions [18]. Let  $Z_i = (x_i, y_i)$  be a Markov chain, and the Gibbs sampler is used to generate specific multivariate distribution. Markov chain Monte Carlo has the advantage of being independent of modelling assumptions and good for small sample size.

## 4 Experimental Set-up

In this paper, we worked on Associated Press data set and Accepted Papers data set. Topic modelling using Associated Press data set is also done by inventor of topic modelling David M Blei in 2003. Associated Press data set is from the First Text Retrieval Conference (TREC-1) 1992. It is also part of ‘topicmodel’ package in R software. It consists of 2246 documents, and the vocabulary was already selected to contain only words which occur in more than 5 documents. So in total we have a document term matrix of 2046 documents and 10476 terms. Another Accepted Papers data set (<http://archive.ics.uci.edu/ml/datasets/AAAI+2014>) comprises the papers submitted to the AAAI 2013 main track. For each paper, we have the abstract, title and one or more high-level keywords selected by the authors during submission from a fixed list.

In the following, we fit an LDA model with topics using (1) VEM with estimated  $\alpha$ , (2) VEM with fixed  $\alpha$  and (3) Gibbs sampling with a burn-in of 1000 iterations.

### ***4.1 Selecting Optimal No. of Topics***

**Case 1:** In this experiment, we fix the number of topics as 40 for both the data sets.

**Case 2:** In this, we have taken different range of topics for both data sets as per their sizes; for example, in Associated Press data set, we take 30–70 and for Accepted Papers data set we have taken 10–30.

### ***4.2 Evaluation of Three Estimated Models on Different Data Sets***

For comparing the performance of fitted model, we use different evaluation mechanisms.

#### **4.2.1 Entropy Measure**

Entropy calculates the topic distribution in the document, low value indicates that distribution of topic in document is not even, and high value of entropy indicates that the topic distributions are more evenly spread over the topics. So entropy should be high for a good fitted model.

#### **4.2.2 Perplexity on Held-Out Data Set**

Perplexity is the measurement of how efficiently a probability model predicts a sample. In other words, perplexity is the inverse of geometric mean of per-word likelihood. A lower value of perplexity is considered optimal for a fitted model. The most common way to evaluate the probabilistic model is log-likelihood of a held-out test set. In this data set, it is divided into two parts: training set with 75% of data and test set with remaining 25% of data. Then, log-likelihood of the fitted model is calculated on held-out data set, i.e. test set. The higher value of log-likelihood is best for good model. Perplexity is reverse of log-likelihood measure, and low score of perplexity is desired for a best model on held-out data set.

**Table 1** Entropy values for three LDA models

Model	Entropy values (accepted papers)	Entropy values (associated press)	No. of topics
VEM with estimated alpha ( $\alpha$ )	2.302	0.7467564	40
VEM with fixed alpha ( $\alpha$ )	2.3022	2.5653961	40
Gibbs sampler	2.2964	2.8545954	40

## 5 Result & Evaluations

### 5.1 Results Case 1

There are Associated Press data set and Accepted Papers data set with  $K = 40$ .

In first model, we estimated by default parameter defined in ‘topicmodel’ package in R software and the starting value of  $\alpha = 50/k$ , where  $k$  is no. of topics. In second model,  $\alpha$  is held at fixed at the initial value. In third model, we use Gibbs sampling technique for posterior distribution approximation. The entropy value of all three models for both data sets is provided in Table 1.

As per evaluation measure described for entropy metrics, in this experiment for large data set with approx 2000 documents, variational expectation maximization with both the variants of  $\alpha$  (estimated and fixed) has highest entropy. So variational inference is best in this experiment. For small data set, Accepted Papers has only 300 documents, so in this data set Gibbs sampling inference topic model has the highest value of entropy revealing the highest performance.

Perplexity is the probability of the test data set, normalized by the number of words. In this experiment, we calculated perplexity for all three topic models for both training and test data sets. The results with perplexity values are shown in Table 2. For estimated alpha variational inference, perplexity value for test data set is lower than for training data set for both the data sets. Fixed alpha variation model has low perplexity for held-out data set in both the data sets. So results are considered nice. But when we compare perplexity value range in both the fitted topic models, variational inference with fixed alpha ( $\alpha$ ) will be considered as bad one with high range of perplexity. So fixed alpha model is not considered as good fitted topic model. In Gibbs sampling model, perplexity values are much lower than in both variational inference models for Accepted Papers data set. For Associated Press data set, variational inference with estimated alpha model has the lowest perplexity. So in the whole experiment it can be concluded that for large data set, variational inference-based inference technique is best for topic modelling and for small data set Gibbs sampling inference can be used as optimal inference technique in topic modelling.

**Table 2** Perplexity of three LDA models on training and test data sets

Model	Accepted papers data set		Associated press data set		No. of topics
	Perplexity (training data set)	Perplexity (test data set)	Perplexity (training data set)	Perplexity (test data set)	
VEM with estimated alpha ( $\alpha$ )	285.2289	278.1307	1622.632	1603.135	40
VEM with fixed alpha ( $\alpha$ )	304.0014	296.4071	2264.034	2219.028	40
Gibbs sampler	260.3376	255.0464	1959.146	1948.904	40

**Table 3** Perplexity values for associated press data set with  $K = 30-70$ 

Number of topics (K)	VEM with estimated alpha	VEM with fixed alpha	Gibbs sampling
30	1439.992	2088.584	2035.411
40	1267.037	1905.874	1918.329
50	1145.913	1788.225	1857.26
60	1049.869	1662.256	1802.07
70	969.0979	1558.179	1776.043

**Table 4** Perplexity values for accepted papers data set with  $K = 10-30$ 

Number of topics (K)	VEM with estimated alpha	VEM with fixed alpha	Gibbs sampling
10	249.76	268.49	234.56
15	191.32	282.64	240
20	148.356	297.948	238.3
25	170.54	314.57	242.33
30	173.413	332.49	244.94

## 5.2 Result Case 2

Associated Press data set with  $K = 30-70$  and Accepted Papers data set with  $K = 10-30$  are taken (Tables 3 and 4).

(a): In this experiment, for Associated Press data set, VEM with estimated alpha is the best model having lowest values of perplexities for different number of topics on test data set. Also able to find optimal number of topics for fitting best model at around 70, where all three fitted topic model has lowest perplexity.

(b): For Accepted Press data set, in which we experiment with topic ranges from 10–30, VEM with estimated alpha is the best model having lowest perplexity on test data set. Also it is able to find optimal no. of topics for fitting best model, i.e. 10–20.

So any number between 10 and 20 can be taken as optimal number of topics for best fit model.

## 6 Conclusion and Future Work

In this study, we use two inference methods: variational expectation maximization (VEM) and Gibbs sampling. Gibbs sampling is a type of Markov chain Monte Carlo (MCMC) inference technique and typically used for inference in LDA. But selecting optimal number of topics is a big problem in topic modelling literature. In this experiment, we select a range for different number of topics for both the data sets. In this experiment, through perplexity measure we are able to find the best model with optimal no. of topics. The best model is the model with the lowest perplexity. So in this, we find that in all the three inference techniques variational inference with estimated alpha is the best model with the lowest perplexity for both training and test data sets in Associated Press data set and in Accepted Press data set. So it means for smaller data set like Accepted Papers with 300 instances, the optimal no. of topic is 10–20 and Gibbs sampling is the best inference technique for these data sets. As the number of topics increases, perplexity values also increase and the performance of topic models starts decreasing. So in this experiment, we can conclude that with fixed topic numbers, as well as different range of topics for larger data set like Associated Press variational inference with estimated alpha gives the best solution model. And for smaller data set, Gibbs sampling gives the best performance with optimal number of topics, and as the number of topics increases for smaller data set, the performance degrades accordingly. So in totality, variational inference can optimize Bayesian computation and provide fast solution for massive data. Gibbs sampling has the advantage: a distribution free method and being so good for small sample size. So in future we are planning to find a heuristic approach to select the best optimal value for number of topics in latent Dirichlet allocation.

## References

1. Blei, D.M., Lafferty, J.D.: Topic models. Int. J. CRC Press Text Mining Classif. Cluster. Appl. 71–89 (2009)
2. Mitchel, T.M.: Machine Learning. McGraw-Hill, Inc. New York, USA (1997)
3. Hofmann.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR conference on Research and Development in Information Retrieval. New York, NY, USA, ACM, pp. 50–57 (1999)
4. Kucukelbir, A., Tran, D., Rangnath, R., Gelman, A., Blei, D.M.: Automatic Differential Variational Inference (2016). [arXiv:1603.00788](https://arxiv.org/abs/1603.00788)
5. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. **41**(6), 391–407 (1990)



7. Blei, D. and Lafferty, J.: Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning (2006) 113–120
8. Blei, D., Lafferty, J.A.: Correlated topic model of science. *Ann. Appl. Statist.* **1**, 17–35 (2007)
9. Rogers, S., Girolami, M., Campbell, C., Breitling, R.: The latent process decomposition of CDNA microarray data set. *ACM Trans. Computat. Biol. Bioinform.* **2**(2), 143–156 (2005)
10. Shivshankar, S., Srivathsan, S., Ravindran, B., Tendulkar, A.V.: Multi-view methods for protein structure comparison using latent Dirichlet allocation. *Bioinformatics* **27**(13), 161–168 (2011)
11. Zhao, W., Zou, W., Chen, J.J.: Topic modelling for cluster analysis of large biological and medical datasets. *BMC Bioinform.* **15**(S11) (2014)
12. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of pollution structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000)
13. Griffiths, T.L., Styvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences, vol. 101. PNAS, pp. 5228–5235 (2004)
14. Gelfand, A., Smith, A.: Sampling based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409 (1990)
15. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* **39**, 1–38 (1977)
16. Hastings, W.: Monte carlo sampling methods using Markov chains and their applications. *Biometrika* **57**97, 97–109 (1970)
17. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, M., and Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953)
18. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984)
19. Miao, Y., Yu, L., Blunsom, P.: Neural Variational Inference for Text Processing, pp. 1727–1736 (2016)
20. Hinton, G.E., Salakhutdinov, R.R.: Replicated softmax: an undirected topic model. In: Advances in Neural Information Processing System, pp. 1607–1614 (2009)
21. Larochelle, H., Lauly, S.: A neural autoregressive topic model. *Advanc. Neural Inf. Process. Syst.* 2708–2716 (2012)
22. Mnit, A., Gregor, K.: Neural Variational Inference and Learning Belief Network, pp. 1791–1799 (2014)
23. Teh, Y.W., Newman, D., Welling, A.M.: Collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Neural Inf. Process. Syst.* 1–8 (2006)

# Action Recognition Framework Based on Normalized Local Binary Pattern



Shivam Singhal and Vikas Tripathi

**Abstract** Human action recognition in computer vision has become dexterous in detecting the abnormal activities to fortify safe events. This paper presents an efficient action recognition algorithm which is based on local binary pattern (LBP). The implementation of this approach can be used for action recognition in small premises such as ATM rooms by focusing on the LBP feature extraction via spatiotemporal relations. We also focus on decreasing the descriptor values by normalizing computed histogram bins. The results through ATM dataset demonstrate the enhancement in action recognition problem under different extensions. The normalized features obtained are classified using random forest classifier. In our study, it is shown that normalized version of LBP surpasses the conventional LBP descriptor with an average accuracy of 83%.

**Keywords** Motion detection · Action recognition · Local binary pattern (LBP) Feature extraction · Texture features · Optical flow

## 1 Introduction

Action recognition algorithms are currently engaged in many computer vision applications. Video-based surveillance system robustly addresses motion detection and action recognition from image sequences. Advancements in video-based surveillance have offered a better technology to review moving object detection. Motion pattern detection and extraction present theoretical and practical advances in the area of video processing for advanced surveillance. Vision-based activity recognition [1] is a challenging problem to understand the behavior of moving objects through videos.

---

S. Singhal (✉) · V. Tripathi  
Department of Computer Science and Engineering, Graphic Era University,  
Dehradun 248002, Uttarakhand, India  
e-mail: singhal.shivam58@gmail.com

V. Tripathi  
e-mail: vikastripathi.be@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_23](https://doi.org/10.1007/978-981-13-1708-8_23)

In this paper, there are several algorithms that have attempted to build activity models such as Kalman filtering, hidden Markov models, Gaussian mixture models and the conditional random field. The proposed algorithms are pixel dependent to build the model. Motion analysis is applicable in recognizing an object activity using object motion trajectory. Motion analysis algorithms could be based on extracting activity from intensity of each pixel resulting in the motion flow. Hence, the motion sequence focuses on detecting regions with activity. The motion trajectory images temporally identify general model of movement by using the intensity of pixels. Motion energy image (MEI) describes the motion shape and spatial distribution of a motion, and motion history image (MHI) describes spatiotemporal distribution in the image sequences [2]. Motion detection extracts meaningful information which could be low-level quantitative features such as color histograms as well as high-level information, i.e., local feature descriptors around interest points. These feature descriptors are one of the key factors in describing visual content. Feature representation can be interest point-based representation [3] or appearance-based representation [4, 5]. The general descriptor extracts local feature descriptors around interest points and in interest points and is used in many different recognition problems. The features were extracted using this motion information, and to encode it in quantitative values, several authors have proposed effective descriptors like scale-invariant feature transform (SIFT), histograms of oriented gradients (HOGs), speed-up robust features (SURF), histograms of oriented flow (HOF), maximally stable extremal regions (MSERs) [6], local binary pattern (LBP) [7]. In SIFT [8, 9], various scales of an image are analyzed to extract features which compute descriptor values. SIFT descriptors are computed by using input as a keypoint frame, i.e., descriptor center. The points that are detected are called STIPs (spatiotemporal interest points) [10]. SURF [11, 12] approximates Laplacian of Gaussian with box filter and is calculated by applying integral images. SURF with the help of this integral image calculates feature descriptor based on the sum of the Haar wavelet response around the interest point. HOG [13] technique calculates the histogram of gradients and computes interest points. HOF provides information about pattern of relative motion between image sequences using optical flow [14, 15]. Maximally stable extremal regions (MSERs) are a blob detection feature descriptor. Matas et al. [6] proposed this descriptor for tracking colored objects and matched interest points [3] between images. LBP is a texture feature descriptor that computes histogram based on binary value and other local features. Accordingly, many researchers have proposed efficient texture descriptors focusing on orientation, histogram, optical flow and other factors to work out on analysis problems. In the significance of texture analysis and action recognition, local binary pattern (LBP)-based features have been introduced as an elementary technique.

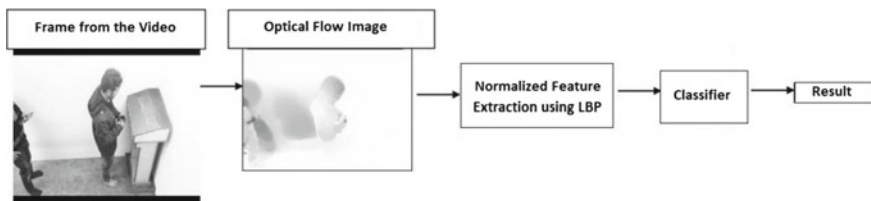
We present an extended approach of rotation invariant texture operator based on local binary patterns. Local binary pattern (LBP) [16, 17] was introduced as a simple texture descriptor [16] based on the algorithm of using a pixel's relationship with its neighborhood region. Ojala et al. [7] proposed the local binary pattern (LBP) method for rotation invariant texture classification. LBP [18] is defined as a grayscale texture descriptor [19], which deals with the sign of the neighboring pixels. Various approaches have been described to propose texture-based algorithms. LBP was orig-

inally described to compute histograms, extracted from thresholding neighborhood region. The local binary pattern [20] histograms are based on a uniformly fixed set of rotation invariant patterns [19, 21]. Some other approaches have been applied by several authors to extend original LBP. Like Zhao and Pietikäinen introduced volume LBP (VLBP) [22], this extended the approach to spatiotemporal data. Huang et al. [23] introduced an extension in the LBP approach with 3D LBP. Fehr et al. [24] approached the spherical harmonic transform for full 3D volume texture analysis to compute LBP. Most approaches to recognize texture have inspired a collection of extended studies, which generally integrate invariance with respect to grayscale spatial properties. The proposed feature is an extension of the LBP texture descriptor and its performance confirms the efficiency of the LBP-based approaches. In this paper, we attempt an extended method of the binary patterns by focusing on the feature extraction via spatiotemporal relations. We also focus on decreasing the descriptor values by normalizing computed histogram bins. LBP is a generalized texture descriptor to illustrate local image pattern and, our proposed work has achieved notable classification results.

We have analyzed four categories of human actions which are classified as single, single abnormal, multiple and multiple abnormal. This paper is further divided into three sections: Sect. 2 describes the proposed framework; Sect. 3 describes results and analysis; and Sect. 4 concludes the paper.

## 2 Methodology

The proposed method makes use of computer vision-based framework to recognize various activities in given video. Figure 1 represents the working of our proposed framework. It represents that this method consists of the camera feed in the form of video. The extracted frame from this video feed is further used to extract temporal information using optical flow. LBP and extended LBP descriptors individually are then used to compute histogram bins and further normalized to obtain features. These features computed are then classified using random forest classifier. Algorithm 1 represents the analytical representation of our proposed framework.



**Fig. 1** Framework for the proposed method

## 2.1 Feature Descriptor

To extract relevant information from given sequence of images, we have used LBP as a feature descriptor. Enhancement in LBP is performed by applying L2 normalization in descriptor calculation.

### Algorithm 1: Proposed LBP Method

**Input: Video with Resolution: 320 × 240**

1. Initialize $x=0$	1. Initialize $x$
2. While $x < \text{frames}$ do	2. Frames = Number of Frames
3. Initialize $y=0$	3. Initialize $y$
4. While $y < n$	4. $n = \text{buffer size}$
5. $y = y + 1$	5. Increment $y$
6. Compute Optical Flow between $I(x, y)$	6. Optical Flow b/w frames
7. Compute LBP on Optical Flow Image	7. Compute LBP descriptors
8. Normalize computed LBP features	8. Compute normalized features
9. $x = x + 1$	9. Increment $x$

Equation (1) represents formula for LBP descriptor generation at the center pixel  $(x_{cp}, y_{cp})$ . LBP notation is given by  $(N, r)$  to designate the neighborhood of radius 'r' and the use of 'N' points involved in the neighborhood.

$$\text{LBP}_{N,r}(x_{cp}, y_{cp}) = \sum_{n=0}^{N-1} f(m_n - m_c) 2^N \quad (1)$$

where  $m_n$  is magnitude of neighbor pixel around the center pixel,  $m_c$ , and Eq. (2) defines  $f(x)$ .

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The rotation invariant LBP operator  $\text{LBP}_{N,r}^{\text{RI}}$  is given by Eq. (3)

$$\text{LBP}_{N,r}^{\text{RI}}(x_{cp}, y_{cp}) = \min \{ \text{RS}(\text{LBP}_{N,r}(x_{cp}, y_{cp}), q) \mid q \in [0, N - 1] \} \quad (3)$$

where RS performs a circular right-shift operation of the bit sequence.

For a complex vector 'z', Eq. (4), the normalized form is given by Eq (5)

$$z = (z_1, z_2, z_3, z_4 \dots z_n) \quad (4)$$

$$|z| = \sqrt{\sum_{i=1}^n |z_i|^2} \quad (5)$$

### 3 Result and Accuracy

The framework has been trained and tested using MATLAB on computer having Intel i3, 2.0 GHz processor with an 8 GB RAM on the videos for computing our extended LBP descriptors. To classify the actions, we used Weka as the classification tool. The videos for the proposed methodology have the resolution of  $320 \times 240$ , which are recorded in an indoor environment; these are ATM surveillance videos. The dataset [25] provided by these ATM surveillance videos are classes classified under the following four classes: (i) single: when a single person is in the video frame performing normal activities; (ii) multiple: when multiple people are in the video frame and performing normal activities; (iii) single abnormal: when a single person is in the video performing abnormal activity; (iv) multiple abnormal: when multiple people are in the video performing abnormal activity over the 39 videos (9 single, 10 single abnormal, 12 multiple and 8 multiple abnormal). We have made our dataset of frame resolution  $320 \times 240$  for training and testing purposes. The framework is trained using these videos for extracting features from the motion images. Testing is done on different videos from the one used for training. The algorithm has been tested for multiple frames provided by the LBP descriptor.

In our proposed framework, we have used ATM video feed as input. Then, step-by-step frames are extracted to calculate temporal data from these inputs. First, we define a buffer size to calculate temporal information, i.e., optical flow between the extracted frame and a reference frame one at a time. This is followed by the original LBP approach on this analyzed optical flow image, and result is computed accordingly. To enhance this temporal information, we fused the computed LBP descriptors with additional spatial data which is computed from the LBP method on the reference frame. Table 1 shows the statistics of temporal framework based on optical flow and spatiotemporal framework. It shows that fused spatiotemporal information generates higher accuracy as compared to only temporal method. Furthermore, we used spatial relations in the original LBP and encountered that when LBP descriptors are computed by dividing the extracted frame into cells, the result showed some variations. These changes are actually the outcome of descriptor values computed which changes according to the preferred CELLSIZE. Table 2 clearly illustrates that when we use various values of CELLSIZE the produced accuracies vary. Further in our study we enhanced this result by normalizing the descriptor to 58 values.

**Table 1** Accuracy of LBP descriptor in percentage (%)

Number of frames (buffer size)	Optical flow (temporal data) (%)	Fusion of optical flow with reference frame (spatiotemporal) (%)
5	66.5	80.8333
10	64.1667	82.3333
15	62.6667	84.6667

**Table 2** LBP spatial results (in percentage %)

CELLSIZE	Original LBP descriptor values	Original spatial accuracy(%)	Proposed LBP descriptor values	Proposed spatial accuracy (%)
64 × 64	58	81.5	58	81.5
32 × 32	232	82.6667	58	78.5
16 × 16	928	72.8333	58	81.5
8 × 8	3712	66.5	58	80.1667

**Table 3** LBP temporal results (in percentage %)

CELLSIZE	Original LBP descriptor values	Original temporal accuracy (%)	Proposed LBP descriptor values	Proposed temporal accuracy (%)
64 × 64	58	63.8333	58	63.8333
32 × 32	232	68.5	58	62.1667
16 × 16	928	68.8333	58	64.1667
8 × 8	3712	71	58	64.5

Our method normalizes the cells into 58 descriptor values irrespective of the value of CELLSIZE. Since buffer size 10 gives the most stable result when both optical flow and fusion methods are considered (Table 1), we used buffer size for frames as 10.

Table 2 also depicts that the highest accuracy achieved in original LBP method is 82.667%, but the same method also achieves the lowest accuracy of 66.5%. With the variation in CELLSIZE, the result varies drastically and conversely in our normalized approach; the statistics are more stable and generates the lowest accuracy of 78.5%. Table 3 shows the same variations as Table 2 but in a temporal manner. We generated a constant 58 descriptor values by normalization LBP method on the computed optical flow image which is similar to our initial approach (Table 1). On analyzing Table 3, we encountered that our proposed result still shows stability. Opposite to this original LBP produces slightly higher results but still vary with the provided CELLSIZE. The original LBP produces immense descriptor values which consume a significant amount of time, whereas in our normalized method only 58 values are produced swiftly and hence are more efficient. Moreover, to produce spatiotemporal data we used LBP on optical flow image to compute temporal information and original LBP operator on the reference frame for spatial information. Then, we used our proposed normalization method to attain higher statistics. Table 4 computes a combined spatiotemporal data.

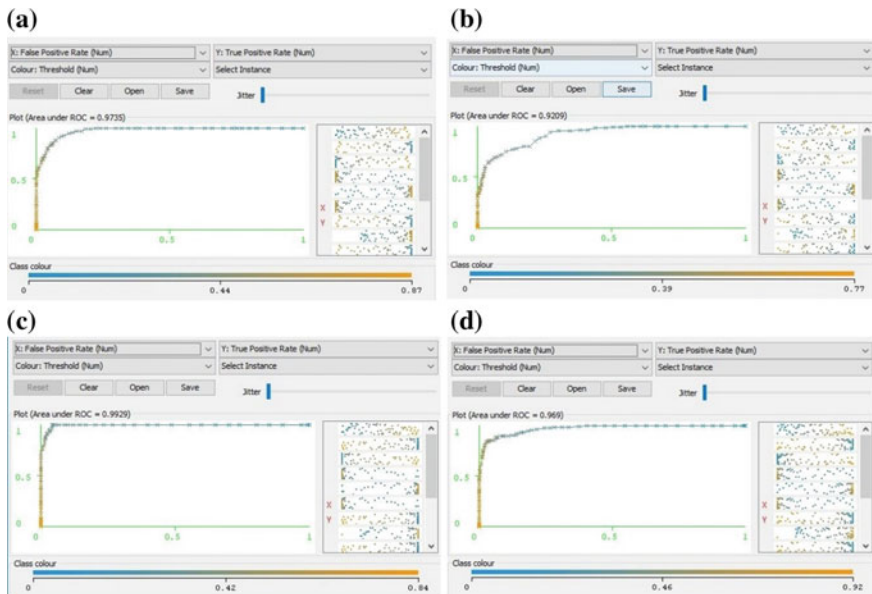
Table 4 shows that our approach produces more stable result irrespective of the CELLSIZE and computes only 116 spatiotemporal descriptor values, whereas original LBP produces 7424 values which are not time efficient and also computes lowest result of only 55.333%. Table 5 gives the confusion matrix obtained from normalized LBP method when we used spatiotemporal framework as per our fusion approach and when CELLSIZE of the frame is 64 × 64. Figure 2 represents the ROC curves of

**Table 4** LBP spatiotemporal results (in percentage %)

CELLSIZE	Original LBP descriptor values	Spatiotemporal accuracy (%)	Proposed LBP descriptor values	Spatiotemporal accuracy (%)
64 × 64	58 + 58	84	58 + 58	84
32 × 32	232 + 232	82.1667	58 + 58	82.6667
16 × 16	928 + 928	82.3333	58 + 58	82.3333
8 × 8	3712 + 3712	55.3333	58 + 58	83

**Table 5** Confusion matrix

a	b	c	d	← classified as
180	13	0	7	a = multiple
31	64	0	5	b = multiple abnormal
0	0	126	24	c = single
2	0	14	134	d = single abnormal



**Fig. 2** a ROC curve of multiple class; b ROC curve of multiple abnormal class; c ROC curve of single class; d ROC curve of single abnormal class

all the four classes when the achieved accuracy is 84% in reference to spatiotemporal information.



## 4 Conclusion

In this paper, we proposed an effective extension in texture descriptor LBP to encode normalized features occurring in images. The proposed framework effectively recognizes normal and abnormal events in given video by utilizing spatial and temporal LBP feature descriptors. The result shows that normalized LBP outperforms traditional LBP. Original LBP method presented an average accuracy of 75.96%, and our method enhances this to an average accuracy of 83.0% for ATM dataset. We also produced more stable and productive result than the conventional LBP by normalizing effective descriptor values, which was the major focus in this paper. This research is wide open for more enhancement of LBP. Subsequently, a more efficient normalization technique can be used in the future to obtain higher accuracy. Also, a better technique than optical flow can be used to enhance temporal and spatiotemporal LBP descriptors.

## References

1. Damaševičius, R., Vasiljevas, M., Šalkevičius, J., Woźniak, M.: Human activity recognition in AAL environments using random projections. *Comput. Math. Methods Med.* **2016**, Article ID 4073584, 1–17 (2016)
2. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. *Mach. Vis. Appl.* **23**(2), 255–281 (2010)
3. Klaser, A., Marszalek, M., Schmid, C.: A spatiotemporal descriptor based on 3D-gradients. In: *Proceedings of British Machine Vision Conference*, pp. 995–1004 (2008)
4. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 221–231 (2013)
5. Niebles, J.C., Li, F.-F.: A hierarchical model of shape and appearance for human action classification. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)
6. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: *Proceedings of British Machine Vision Conference*, pp. 384–396 (2002)
7. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study for texture measures with classification based on feature distributions. *Pattern Recogn.* **29**(1), 51–59 (1996)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* **60**(2), 91–110 (2004)
9. Guo, Z. et al. Accurate, pupil center location with the SIFT descriptor and SVM classifier. *Int. J. Patt. Recogn. Artif. Intell.* **30**(4), 1–15 (2016)
10. Chakraborty, B., Holte, M.B., Moeslund, T.B., González, J.: Selective spatio-temporal interest points. *Comput. Vis. Image Underst.* **116**(3), 396–410 (2012)
11. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: speeded up robust features. In: *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, vol. 3951, pp. 404–417 (2006)
12. Abedin, Md.Z., Dhar, P., Deb, K.: Traffic sign recognition using SURF. In: *Speeded up Robust Feature Descriptor and Artificial Neural Network Classifier*, pp. 198–201 (2016)
13. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Comput. Vis. Image Understand.* **117**(7), 790–806 (2013)

14. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1932–1939 (2009)
15. Mahbub, U., Imtiaz, H., Ahad, M.A.R.: An optical flow based approach for action recognition. In: Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 646–651 (2011)
16. Ojala, T., Pietikäinen, M., Mäenpää, T.T.: Multiresolution gray-scale and rotation invariant texture classification with local binary pattern. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
17. Zhang, Y.X., Zhao, Y.Q., Liu, Y., Jiang, L.Q., Chen, Z.W.: Identification of Wood Defects Based on LBP Features, pp. 4202–4205 (2016)
18. Maksymiv, O., Rak, T., Peleshko, D.: Video-Based Flame Detection using LBP-Based Descriptor: Influences of Classifiers Variety on Detection Efficiency, pp. 42–48 (2017)
19. Pietikäinen, M., Ojala, T., Xu, Z.: Rotation-invariant texture classification using feature distributions. *Pattern Recogn.* **33**(1), 43–52 (2000)
20. Pietikäinen, M., Zhao, M.G.: Two decades of local binary patterns: a survey. In: Bingham, E., Kaski, S., Laaksonen, J., Lampinen, J., (eds.), *Advances in Independent Component Analysis and Learning Machines*, Elsevier, pp. 175–210 (2015)
21. Pietikäinen, M., Ojala, T., Nisula, J., Heikkinen, J.: Experiments with two industrial problems using texture classification based on feature distributions. *Intelligent Robots and Computer Vision XIII: 3D Vision, Product Inspection, and Active Vision*, vol. 2354, no. 1, pp. 197–204 (1994)
22. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using volume local binary patterns. In: *ECCV, Workshop on Dynamical Vision*, pp. 165–177 (2006)
23. Huang, Y., Wang, Y., Tan, T.: Combining statistics of geometrical and correlative features for 3D face recognition. In: *Proceedings of the British Machine Vision Conference*, pp. 879–888 (2006)
24. Fehr, J.: Rotational Invariant Uniform Local Binary Patterns for Full 3D Volume Texture Analysis. *FINSIG* (2007)
25. Sanserwal, V., Tripathi, V., Pandey, M., Chen, Z.: Comparative Analysis of Various Feature Descriptors for Efficient ATM Surveillance Framework, vol. 10, no. 13, pp. 181–187 (2017)

# Enhancements to Randomized Web Proxy Caching Algorithms Using Data Mining Classifier Model



P. Julian Benadit, F. Sagayaraj Fancis and A. M. James Raj

**Abstract** Web proxy caching system is an intermediary between the users and servers that tries to alleviate the loads on the servers by caching selective web pages, behaves as the proxy for the server, and services the requests that are made to the servers by the users. In this paper, the performance of a proxy system is measured by the number of hits at the proxy. The higher number of hits at the proxy server reflects the effectiveness of the proxy system. The number of hits is determined by the replacement policies chosen by the proxy systems. Traditional replacement policies that are based on time and size are reactive and do not consider the events that will possibly happen in the future. The outcomes of the paper are proactive strategies that augment the traditional replacement policies with data mining techniques. In this work, the performance of the randomized replacement policies such as LRU-C, LRU-S, HARM, and RRGVF are adapted by the data mining classifier based on the weight assignment policy. Experiments were conducted on various data sets. Hit ratio and byte hit ratio were chosen as parameters for performance.

**Keywords** Web proxy caching · Data mining classifier  
Weight assignment policy · Randomized replacement

---

P. Julian Benadit (✉)

Department of Computer Science Engineering, Faculty of Engineering, Kengeri campus, CHRIST (Deemed To be University), Bangalore 560074, India  
e-mail: julianben1983@hotmail.com

F. Sagayaraj Fancis · A. M. James Raj

Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry 605014, Puducherry, India  
e-mail: fsfrancis@pec.edu

A. M. James Raj

e-mail: jamdiva05@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_24](https://doi.org/10.1007/978-981-13-1708-8_24)

## 1 Introduction

The World Wide Web and its usage are growing at a rapid rate which has resulted in overloaded web servers, network congestion, and consequently poor response time. Multitudes of approaches are continuously being made to overcome these challenges. ‘Web caching’ is one of the approaches that can enhance the performance of the web [1]. A web cache is a buffered repository of the web pages that are most likely to be requested frequently and in the near future. The general architecture of the Word Wide Web consists of the client users, the proxy server, and the origin server. Whenever the client requests the web object, it can be retrieved either from the intermediate proxy server immediately or from the origin server. Therefore, whenever a user’s request is satisfied from the proxy server, it minimizes the response time and it reduces the overload of the web origin server. Typically, the web cache may be located [1] at the origin server cache, at the proxy server cache, or at the client cache.

Section 2 addresses the previous work in traditional randomized web proxy caching algorithms and data mining methods. Section 3 details the overall working model for the web proxy caching based on data mining classifier model, and Sect. 4 presents the generic model for web proxy caching algorithm using data mining classifier model and the performance metrics for web proxy cache replacement algorithms.

## 2 Related Work

The methodologies for web caching may be categorized into two groups. The first category of methodologies is ‘*traditional random*’; in the sense, they use computationally simple parameters for cache replacement. The second category of the methodologies combines data mining techniques with the traditional approaches for the enhancement of the web caching system.

### 2.1 Summary of Traditional Randomized Web Proxy Caching Algorithms

This section summarizes the traditional randomized web proxy caching algorithm [2] based on the key parameters, and Table 1 summarizes the time line, key factors for the eviction, and its limitations for the traditional replacement policies least recently used based on cost (LRU-C), least recently used based on size (LRU-S), harmonic replacement (HARM), and randomized replacement general value mean function (RRGVF) [2, 3].

**Table 1** Summary of traditional randomized web proxy caching algorithms

Algorithms	Parameters	Evictions
LRU-C	Object cost $C_p$	Web objects at the bottom of the stack are removed based on the probability. $P(i) \propto \frac{c_i}{\max\{c_1, c_2, c_3, \dots, c_N\}}$
LRU-S	Object size $S_p$	Web objects at the bottom of the stack are removed based on the probability. $P(i) \propto \frac{\min\{s_1, s_2, s_3, \dots, s_n\}}{S_i}$
HARM	Object size $C_p$ Object cost $S_p$	In this case, the web objects are removed randomly with probability inversely proportional to cost of its web object. $\text{cost}(p) \propto \frac{S_p}{c_p}$
RRGVF	Number of samples N. Number of samples to keep from a previous iteration M	It evicts randomly the least useful object in the sample $M = N - \sqrt{\frac{(N+1)100}{n}}$

## 2.2 Classification-Based Web Caching

In this method, a new replacement was proposed called Khalid Obaidat Replacement Algorithm (KORA) to provide a better cache performance [4]. This algorithm uses a neural network method to identify and distinguish transient and shadow cache lines. Hence, this algorithm aims for the shadow lines as identified by the neural network. The features used in this method for training the data sets are line access frequency, recently accessed lines, reference pattern, etc. The KORA algorithm performs well compared to the conventional algorithm by having lower miss ratio. However, the algorithm improves only 8% performance compared to LRU. The next approach uses an adaptive web cache access predictor using neural network as one of the approaches to web caching [5]. In this approach, back-propagation neural network (BPNN) is used to improve the performance of web caching by predicting the most likely re-accessed objects. Another approach studies the web cache optimization with a nonlinear model using object features [6]; this approach utilizes the multilayer perceptron (MLP) network for predicting the value of the object based on the syntactic features of the HTML document. The next strategy is based on neural network proxy cache replacement (NNPCR) [7], which integrates the neural network. In this method, a back-propagation adjusts the weight factors in the network. Here, an object is selected for replacement based on the ratings returned by the back-propagation neural network (BPNN). The next approach uses intelligent Naïve Bayes approach for web proxy caching [8]. In this method, the Naïve Bayes approach is used to classify whether the web object can be re-accessed in the future or not. The next method improves the performance of a proxy cache using Tree augmented Naïve Bayes approach followed by very fast decision tree algorithm for improving the web proxy cache and data mining classification performance [9–11]. This method is integrated with traditional replacement algorithm LRU, GDS, GDSF, and GD\* to form a novel web caching.

### 3 Working Principle of Web Proxy Caching Based on Data Mining Classifier Model and Weight Assignment Policy

Web proxy caching aims at enhancing the performance of the proxy server by increasing the hit and byte hit ratios. One class of strategies augments the traditional replacement policies with data mining technique. The strategy uses a data mining classifier model based on the sliding window mechanism and weight assignment policy. The overall working flow model consists of different phases as shown in Fig. 1. These working methods are classified as given below:

#### 3.1 Data Pre-processing

In this method, the proxy log data sets are transformed into a suitable format. Here, the pre-processing methods are mainly used to remove the irrelevant field in the proxy

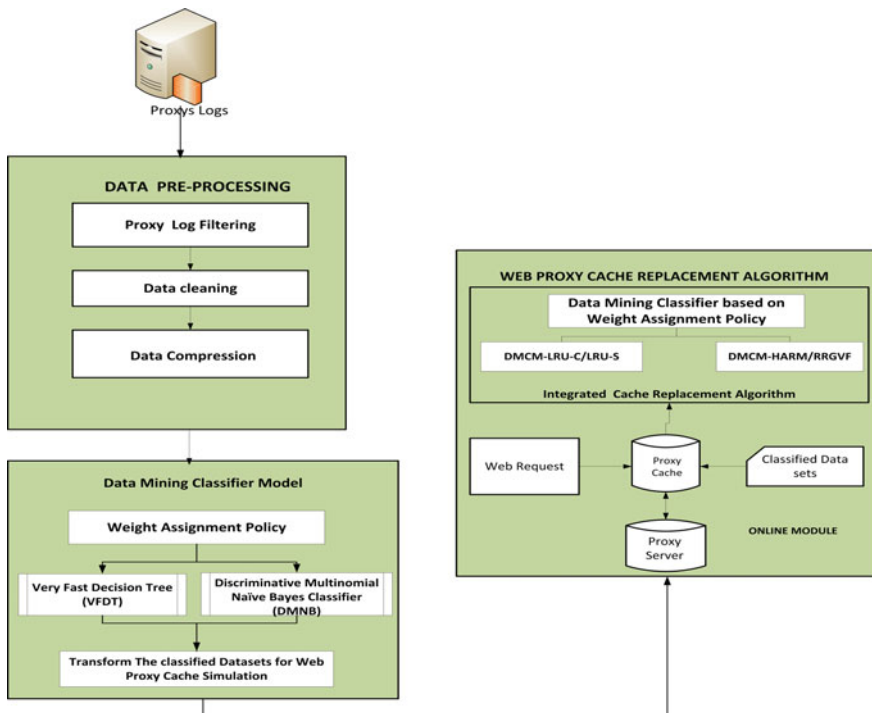


Fig. 1 Overall working model for the web proxy caching using data mining classifier model based on weight assignment policy

data sets to abstract the relevant features. The steps involved in data pre-processing are proxy logs filtration, data cleaning, and data compression.

### 3.2 Proxy Logs Filtration

In this technique, the recorded proxy log files obtained from the NLANR have undergone the basic filtration process in order to reduce the size of the log data sets as well as the running time of the simulation. This filtration module is based on the methods such as a latency-based method, size-based method, dynamic-based method, content-based method, and GET, ICP Type method.

### 3.3 Data Cleaning

Data cleaning is the process of removing the irrelevant entries in the proxy log file. Here, only the relevant HTML file is considered, and all other irrelevant log entries that were recorded by requesting graphics, sound, and other multimedia files, etc., are discarded.

### 3.4 Data Compression

In this method, the cleaned proxy data sets are reduced in size for efficient data mining in order to reduce the simulation time. In this case, the entries in the proxy log files may be irrelevant for statistical analysis, and they rarely are used for data mining classifier model. Finally, the required data are converted into structured format.

### 3.5 Weight Assignment Policy for Cache Replacement

The weight assignment policy of the web object  $p$  in the proxy cache  $t$  is expressed in Eq. 1, and the parameters are shown in Table 2. From the above strategy, the key value used in the caching system is applied to the randomized family replacement algorithm and the key factor of the replacement algorithm is modified according to Eq. 1 as shown above.

$$K_n(p) = L + \frac{F(p) + k_{n-1}(p) \times \left( \frac{\Delta T_t}{C_{ct} - T_{Lt}} \right)}{S(p)} \quad (1)$$

**Table 2** Parameters for weight assignment policy

Parameters	Description
$L$	Inflation factor to avoid cache pollution in the proxy cache $t$
$F(p)$	Previous frequency access of web object $p$ in the proxy cache $t$
$K_{n-1}(p)$	Previous key value of the web object $p$ in the proxy cache $t$
$\Delta T_t$	Difference in time between the current requests and previous requests for the web object $p$ in the proxy cache $t$
$T_{ct}$	Current reference time of the web object $p$
$T_{Lt}$	Last reference time of the web object $p$
$S(p)$	Size of the web object $p$
$k_n(p)$	Current key value

Therefore, whenever the cache replacement occurs, the replacement algorithm replaces the web object based on the key value used by the weight assignment policy.

### 3.6 Very Fast Decision Tree

It is a decision tree method based on Hoeffding tree. Very fast decision tree (VFDT) was developed by Domingo's and Hulten [9, 12]. It is one of the most effective and widely used classification methods. The tree and necessary statistics are stored in memory, and the examples can be processed faster than they can be read from the proxy disk. The working flow for the VFDT and the steps involved in the constructing of the decision tree are shown in Fig. 2. The main idea of Hoeffding trees is to find the best attribute at a given node by considering only a small subset of the training examples that pass through the node. The statistical result that can decide how many examples 'n' are used by each node is called Hoeffding bound.

### 3.7 Discriminative Multinomial Naïve Bayes Classifier

Here, we introduce an advanced supervised machine learning classifier called discriminative multinomial Bayes (DMNB) [11, 13] which uses an efficient discriminative parameter learning method. This algorithm combines the features of Naïve Bayes and a filtering approach called principal component analysis (PCA), which results in better classification accuracy compared to other existing approaches. The algorithm for the proposed DMNB is given in Fig. 3.



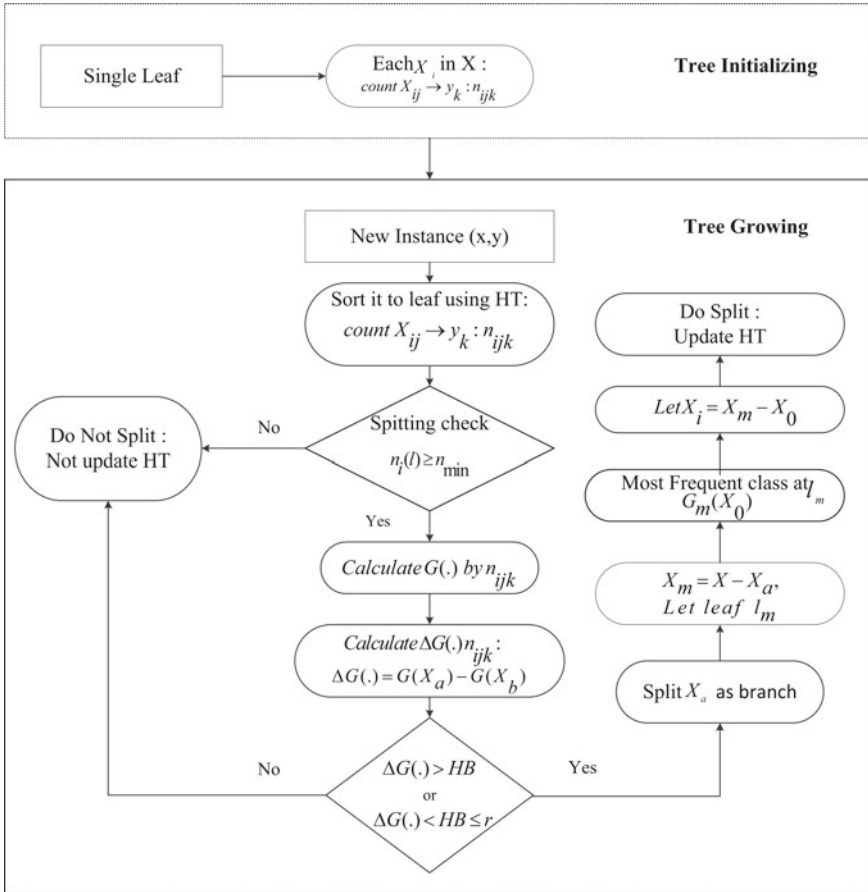


Fig. 2 Working flow of VFDT algorithm

### 3.8 Experimental Setup and Performance Measures for the Data Mining Classifier Model

The methods which are presented were simulated using the data mining software tool. For the simulation of DMNB, a data mining software tool called WEKA is used. For the experiments to be conducted, the proxy data sets were randomly divided into training and testing phases, i.e., 70% for training data sets and 30% for testing data sets, respectively. In the next method, for the simulation of VFDT, a high-speed data mining software tool called very fast machine learning (VFML) is used. The performance measures are suitable for supervised machine learning applications, so as to measure how accurately the data sets are classified. Experiments were conducted,

**Procedure Discriminative Multinomial Naïve Bayes Classifier**

1. Initialize the frequency word count as  $f_{ic} = 0$  .
2. for  $t$  from 1 to M do
  - i. The training document  $d^t$  is randomly chosen from the training data set T.
  - ii. Calculate the probabilities parameters using the equation as shown,  $\hat{P}(w_i | C) = \frac{f_{ic}}{f_c}$  Where,  $f_{ic}$  is the number of occurrences of  $w_i$  in documents of Class C and  $f_c$  is the total number of word occurrences in documents of Class C.
  - iii. Estimate the current frequencies  $f_{ic}^t$
  - iv. Compute the posterior probability  $\hat{P}(c | d^t)$
  - v. Compute the loss  $L(d^t)$  using  $L(d) = P(c | d) - \hat{P}(c | d)x$
3. for each non-zero word  $w_i$  in the document  $d^t$ 
  - i. Let  $f_{ic}^t$  is the frequency of the word  $w_i$  in the  $t^{\text{th}}$  document  $d^t$
  - ii. Let  $f_{ic}^{t+1} = f_{ic}^t + L(d^t) * f_i^t$  .

**Fig. 3** DMNB algorithm

and the results are obtained. Correct classification ratio (CCR) is a good measure for evaluating classifier which is given in Eq. 2.

$$CCR = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

where

TP True Positive,  
 TN True Negative,  
 FP False Positive,  
 FN False negative

The above-simulated experiments are evaluated based on the performance measure, which is shown in Fig. 4.

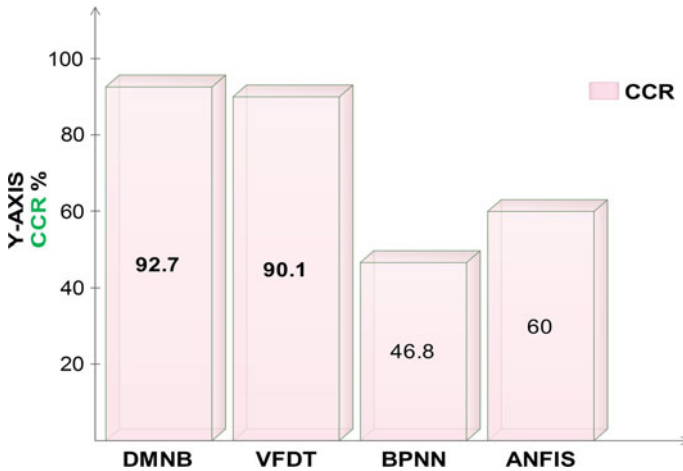


Fig. 4 Comparison of CCR for DMCM

#### 4 Generic Model for Web Proxy Caching Algorithm Using Data Mining Classifier Models

In this section, a generic model for web proxy caching strategies integrated with data mining classifier model is introduced. The algorithm for generic web proxy caching algorithm integrated with data mining classifier model is shown in Fig. 5. In line 1, an initial data mining classifier model (DMCM) is built on history weblogs. Each web object  $p$  is requested from the proxy cache  $t$  that contains the web object  $p$ , and then, the web object  $p$  is returned to the web client. Concerning these performance measures (lines 7–11), in this case, it is considered as *cache hit*. In addition, the number of bytes transferred back to the client is counted for the weighted hit rate measure. Once the data are transferred back to the client, the proxy cache is updated (line 12) by the data mining classifier model (VFDT and DMNB).

On the contrary, if the requested web object  $p$  is not available in the proxy cache  $t$  or it is stale, *cache miss occurs*, i.e., the web object  $q$  is deleted from the cache (line 16); in this case, the proxy server forwards the request to the origin server  $S$  (line 17), and a fresh copy of the web object  $p$  is retrieved from the origin server  $S$  and pushed into the proxy cache  $t$  (line 18). The push method consists of assigning the class value of the web object  $p$  by the data mining classifier model. Similarly, if the cache space  $t$  exceeds the maximum cache size  $N$  (line 19), the web object  $q$  from the cache is popped out from the cache (line 20) based on the class assigned by the data mining classifier model. Such an approach is known as cache replacement policy (lines 21–24), i.e., each time when the cache gets overflows. Finally, the data mining classifier model periodically updates the key value of the remaining web objects stored in the proxy cache. This process continues iteratively when the cache

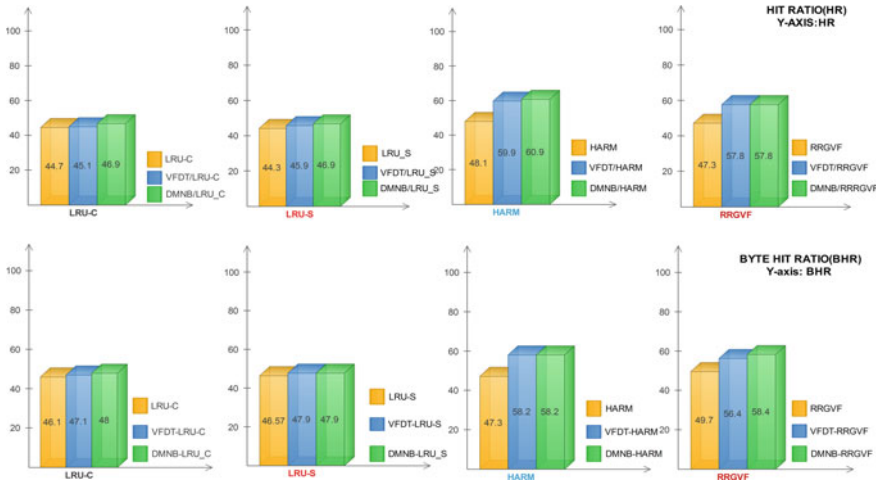
```

Procedure Data Mining Classifier Model (DMCM (VFDT, DMNB))
Proxy cache entry  $t$ ;  $t\_fresh$ ; int Hits = 0; int Byte. Hits = 0; int Cache
Max_Size N
1. begin
2.   DMCM.Build ( );
3.   begin
4.     loop forever
5.       begin
6.         do
7.           Get request the Web object  $p$  from the Proxy cache  $t$  .
8.           if (Proxy cache ( $t$  ).Contains_fresh_copy (Web object  $p$  )
9.             begin
10.              Hits = Hits++;                               /*Cache Hit*/
11.              Byte.Hits = Byte. Hits + t. Bytes-Retrieved-to-client.
12.              Cache.Update (  $t$  , DMCM (VFDT, DMNB)).
13.            end;
14.          else
15.            begin
16.              Cache Delete (  $t$  ).                          /*Cache Miss*/
17.              Retrieve_Fresh Copy of Web object  $p$  from origin Server  $S$  .
18.              Cache.Push (  $t$  , DMCM (VFDT, DMNB));
19.              While (Proxy cache (  $t$  ).Size > Max_size (N)
20.                Cache. Pop (  $q$  );                            /*Cache Replacement*/
21.                Switch ( )
22.                  Case 1 : “DMCM-LRU-S”.
23.                  Case 2 : “DMCM-LRU-C”.
24.                  Case 3 : “DMCM-HARM”.
25.                  Case 4 : “DMCM-RRGVF”.
26.                end;
27.                While (Condition);
28.              end;
29.              DMCM. update model( );
30.            end;
31.          end;

```

**Fig. 5** Generic model for web proxy caching replacement algorithms using DMCM

performance decreases. Also, notice that update of the data mining classifier model (line 31) is dissociated from the online caching of web object, and it can be performed in parallel.



**Fig. 6** Comparison of the overall hit and byte hit ratio of LRU-C, LRU-S, HARM, and RRGVF versus DMCM-based replacement

### 4.1 Experimental Results for Web Proxy Cache Replacement Algorithms

In this section, the results obtained for the data mining classifier model-based web proxy cache replacement algorithms are compared with the traditional cache replacement algorithms. The most commonly used metrics involved in the proxy cache simulation are hit ratio and byte hit ratio [2]. The simulations are carried out using the window-based cache simulator [14], and the experimental results are evaluated based on the packet cost model.

The experimental result compares randomized replacement based on the data mining classifier model which is shown in Fig. 6, respectively. From the figure, we infer that the randomized replacement policies like LRU-C, LRU-S, HARM, and RRGVF improved the hit ratio to 1.6, 0.4, 9.1, and 10.5% by the VFDT model. The DMNB model improves to 2.2, 2.6, 12.1, and 10.5%. The byte hit ratio has been improved to 1, 1.3, 13.1, and 13.3% by the VFDT model. The DMNB model improves to 2.1, 1.4, 13.1, and 11.3%.

## 5 Conclusion and Future Work

The various working modules of the overall working flow such as the data pre-processing, data mining classifier model based on weight assignment policy and generic model for web proxy caching algorithms were described in detail. The VFDT/DMNB classifier outperforms the other machine learning classifier by

improving the classification accuracy comparatively higher than the other machine learning algorithms. The integration of data mining classifier methods with traditional randomized web proxy caching improves the performance of overall hit and byte hit ratio. In the future, the data mining classifier model can also be integrated in content distribution network.

## References

1. Baentsch, M., Baun, L., Molter, G., Rothkugel, S., Sturn, P.: World wide web caching: the application-level view of the internet. *IEEE Commun. Mag.* 170–178 (1997)
2. Balamash, A., Krunz, M.: An overview of web caching replacement algorithms. *IEEE Commun. Surv. Tutor.* 44–56 (2004)
3. Gonzalez-Canete, F.J., Sanz-Bustamante, J., Casilari, E., Trivino- Cabrera, A.: Evaluation of randomized replacement policies for web caches. In: *Proceedings of IADIS International Conference WWW/Internet*, pp. 227–234 (2007)
4. Khalid, H., Obaidat, M.: KORA: A New Cache Replacement Scheme, *Computers and Electrical Engineering*, pp. 187–206 (2000)
5. Tian, Wen, Choi, Ben, Phoba, Vir: An Adaptive Web cache access predictor using network, pp. 450–459. *Lecture Notes in Artificial Intelligence, Developments in Applied Artificial Intelligence* (2002)
6. Koskela, T., Heikkonen, J., Kaski, K.: Web cache optimization with non- linear model using networks object features. *Comput. Netw.* 805–817 (2003)
7. Cobb, J., ElAarag, H.: Web proxy cache replacement scheme based on back-propagation neural network. *J. Syst. Softw.* 450–459 (2008)
8. Ali, W, Shamsuddin, S.M., Ismail, S.: Intelligent Naïve Bayes approaches for web proxy caching. *Knowl. Based Syst.* 162–175 (2012)
9. Benadit, P.J., Francis, F.S.: Improving the performance of a proxy cache using very fast decision tree classifier. *Procedia Computer Science* **48**, 304–312 (2015); *International Conference on Computer, Communication and Convergence (ICCC 2015)*. <http://www.sciencedirect.com/science/article/pii/S187705091500695X>
10. Benadit, P.J., Francis, F.S., Muruganatham, U.: Improving the performance of a proxy cache using tree augmented Naive Bayes classifier. *Proc. Comput. Sci.* **46** (2015)
11. Benadit, P.J., Francis, F.S., Muruganatham, U.: Enhancement of web proxy caching using discriminative multinomial Naive Bayes classifier. *Int. J. Inf. Commun. Technol. Inderscience Publisher* (2017)
12. Hulten, G., Spencer, L., Domingos, P.: Mining time changing data streams. In: *Proceedings of 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, ACM Press, pp. 97–106 (2001)
13. Mouratis, T., Kotsiantis, S.: Increasing the accuracy of discriminative of multinomial Bayesian classifier in text classification. In: *Proceedings of 4th International Conference on Computer Sciences on Convergence Information Technology*, pp. 1246–1251 (2009)
14. Gonzalez-Cante, J. Casilari, E, Trivino-cabrera, A.: A Windows based web cache simulator tool. In: *Proceedings of the 1st International conference on Simulation tools and Techniques for Communications, Networks and Systems & Workshops*, pp. 1–5 (2008)

# Extraction and Classification of Liver Abnormality Based on Neutrosophic and SVM Classifier



Jayanthi Muthuswamy

**Abstract** Liver is the important organ and common site for a variety of cancer diseases. The most important steps in treatment planning and evaluation of liver cancer are to identify the presence of liver cancer and to determine the various stages of liver cancer. This paper proposes an automatic method to segment the liver from abdominal computer tomography imaging and classify the liver as normal or abnormal liver. The aim of this work is to develop computer-aided liver analysis to segment the liver and classify the liver, thereby helping the physician for treatment planning and surgery. The method uses median filter for preprocessing and neutrosophic (NS) domain with FCM thresholding for segmenting the liver. In post processing, morphological operation is done to obtain liver contour. Features are extracted from the segmented liver using gray-level co-occurrence matrix (GLCM). These feature vectors are given as input to train the support vector machine (SVM) classifier, to classify healthy or unhealthy liver. The classifier performances are assessed and analyzed using various quality metrics like accuracy, sensitivity, specificity and misclassification rate.

**Keywords** Median filtering · Neutrosophic logic · Fuzzy C means Adaptive thresholding · Gray-level co-occurrence matrix (GLCM) Support vector machine (SVM)

## 1 Introduction

Liver [1–3, 5] is a largest organ in the human body and plays a vital role to keep the human body free of toxins and harmful substances, thereby maintaining body metabolic balance. The weight of the liver is approximately one and half kg, and it is located in the upper right quadrant of the abdominal cavity, just below the diaphragm. The abnormal mass of tissue in the liver is called liver tumor, and it is also called liver neoplasm which may be solid- or fluid-filled. Liver tumor can be of two types:

---

J. Muthuswamy (✉)

Department of ECE, New Horizon College of Engineering, Bengaluru, India  
e-mail: jayanthi.sathishkumar@live.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_25](https://doi.org/10.1007/978-981-13-1708-8_25)

269

**Table 1** Advantage and disadvantage of various imaging modalities

Modalities	Function	Advantage	Disadvantage
Ultrasonography (US)	Uses high-frequency sound waves to make internal structure	Does not emit radiation, noninvasive	User dependent and not very accurate
Computer tomography (CT)	Uses x ray beam to make internal structure	Emit radiation, not user dependent, noninvasive and used for soft tissues. Low cost	Moderate accurate
Magnetic resonance imaging (MRI)	Uses magnetic field and radio waves to make internal structure	Emit radiation, not user dependent, noninvasive and used for both soft and hard tissues. More accurate	More expensive

benign and malignant tumor. Benign is non-cancerous tumor, whereas malignant is cancerous tumor. In medical science, extracting the liver and tumor masses is a difficult task. This paper focuses on extraction and classification of liver.

Medical imaging modalities [3, 5] are used by the radiologists, to study the internal structures of abdominal organs in the human body. Table 1 shows the advantage and disadvantage of various imaging modalities. Among all, CT imaging is more preferable imaging modality used by the radiologists to find the liver disorder. In the proposed work, abdominal CT image is taken for discussion and the physicians use this technique to diagnosis the disease in earlier stages.

Automatic and accurate segmentation of liver is a challenging task. Some of the challenges are intensity of liver same as other organs, shape of the liver keeps on changing, and there is more noise due to patient movement. To increase the accuracy of segmentation, preprocessing is essential. The objective of preprocessing is to enhance, improve the quality of medical image and also remove noise caused by external factor. For better accuracy of segmentation, one or more segmentation algorithms are combined to form hybrid techniques [4]. This paper used hybrid techniques to extract the liver contour.

The main objective of the proposed work is to develop computer-aided liver analysis system (CAL) that classify the given abdominal CT image into normal liver or abnormal Liver. CAL analysis system [10] assists the radiologists to diagnose the diseases, thereby reducing the inter- and intra-observer variation involved in diagnosis. This CAL system supports medical imaging using machine learning process for automatic classification of region of interest. Machine learning process aims to assign a label to a given input vector. This technique uses feature vector, and these vectors are transformed into representation which is used for classification.

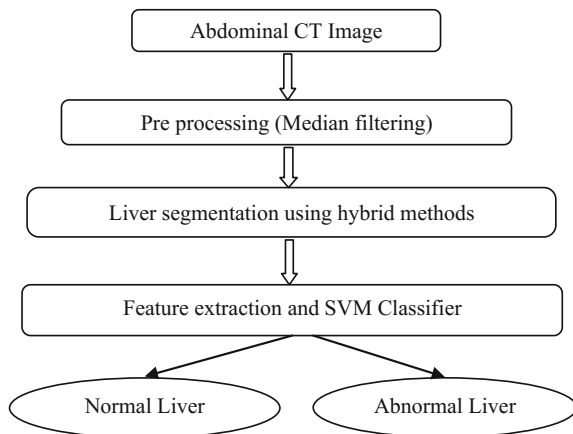


This paper is organized as follows. Section 2 deals with the literature survey. Section 3 describes the overall design of the proposed system. Section 4 provides the description about the data set, experiments conducted, results and discussion. Section 5 concludes the work along with future enhancements.

## 2 Related Works

Many researchers have proposed different methods and techniques for liver segmentation and classification. The author in [1] discussed different techniques and methods for segmenting the liver. They have quoted future direction of liver segmentation for liver volumetry. In [2], author employed an edge-preserving filter which preserves the edges and boundaries that suitable for further analysis of segmentation of liver. The author in [3] proposed a method for liver segmenting the liver using label connected component. They discussed how to measure the liver volume from the segmented output. Priyadarsini et al. [8] have presented a comparative analysis of various segmentation techniques for liver abdominal CT images. The author highlighted automated novel techniques for liver segmentation in order to help the radiologists for solving the healthcare problem. Mala et al. [12] proposed adaptive threshold-based morphological processing for liver segmentation. The author presented probabilistic neural network (PNN) for classifying the liver diseases. Disadvantage of PNN is slow performance and takes more memory space. In [6], author proposed an approach for extracting the liver and tumor from abdominal CT images and used for computer-aided diagnosis. The author used seeded region growing method to extract the liver contour and used limited amount of image samples for discussion. Kumar et al. [13] have developed a CAD system for segmenting the liver and tumor extraction. The drawback is the selection of seed point.

**Fig. 1** Flowchart of the proposed work



In [4], author elaborates comparative study of different segmentation methods for segmenting the liver and obtained the accuracy of 83% for hybrid method. Gehad et al. [5] proposed hybrid approach based on watershed method for segmenting the liver and achieved the accuracy of 92%. Selvathi et al. [16] proposed extreme learning machine for the diagnosis of liver diseases and achieved good degree of accuracy. In [9], author has discussed neural network-based classification and used various performance metrics to find the accuracy of classifier. The problems in existing systems are high processing time, over-segmentation, difficulty in identifying the tumor and slowdown in the training process. The above problems are overcome by hybrid technique. In the proposed work, one or more segmentation methods are used to improve the accuracy of the system.

### 3 Proposed Methodology

The objective of proposed method is to develop computer-aided diagnosis system for extraction of liver from abdominal CT images and classify the region of interest as normal or abnormal liver. The entire work is divided into four phases: preprocessing, liver segmentation using hybrid techniques, feature extraction using GLCM and classification based on SVM classifier. The flowchart of the proposed work is shown in Fig. 1.

#### 3.1 CT Imaging of Liver

Computer tomography imaging [6, 7] is a noninvasive, accurate method for diagnosis of abdominal diseases. CT scanning gives information about the internal body structure of the patients and has many slices. Not all slices that include relevant cancer tissue. Middle slice gives more information about cancer tissue. Figure 2 shows abdominal CT image of liver.



Fig. 2 Abdominal CT image of liver

### 3.2 Preprocessing

Imaging modality gives detail view of internal organs of the human body. This imaging process may contain various noises due to patient movement, and it is not suitable for direct processing. Non-processed image leads to poor segmentation results and increases the false positive error. So preprocessing [2, 6] is an important step in liver segmentation and classification task. In the proposed work,  $3 \times 3$  median filtering is used as edge-preserving filter where each pixel is replaced with median value of the neighboring pixel.

### 3.3 Segmentation

Neutrosophy [4, 5, 16] is a branch of philosophy which deals with indeterminacy, scope of neutralities. In classical set, indeterminacy is not described and evaluated. Fuzzy system has been applied to uncertainty. In medical application, doctor explains to the patient about the diseases based on sign of diseases, no sign of diseases and neutral statement (possibilities of sign). Beyond the level of fuzziness is neutrosophic set (NS) and carries more information. Few problems cannot be solved by fuzzy logic that can be resolved by NS. Neutrosophic image has a membership of 3 subsets ( $\langle A \rangle$ ,  $\langle \text{Not } A \rangle$  and  $\langle \text{Neut } A \rangle$ ) which are defined in different domain. The preprocessed image is converted to NS domain using the following formulas (Table 2).

**Table 2** Algorithm for converting neutrosophic image to binary image

<p><b>Input: Preprocessed Image</b>  <b>Output: Liver region</b></p>
<ol style="list-style-type: none"> <li>1. From the preprocessed image, calculate the histogram.</li> <li>2. From the histogram of preprocessed image, find the local maxima and mean of local maxima.</li> <li>3. Find <math>g_{min}</math> and <math>g_{max}</math>. <math>g_{min}</math> is first peak greater than the mean of local maxima and <math>g_{max}</math> is the last peak.</li> <li>4. Using mean of local maxima, <math>g_{min}</math> and <math>g_{max}</math>, calculate <math>Tr(m,n)</math> using Eq. 1.</li> <li>5. Calculate <math>In(m, n)</math> from the homogeneity value of <math>Tr</math> using Eq. 2.</li> <li>6. Calculate <math>Fa(m, n)</math> using Eq. 3.</li> <li>7. Apply 3 class FCM thresholding [6] to <math>Tr, In, Fa</math>.</li> <li>8. Map fuzzy neutrosophic image into binary image.</li> <li>9. From the binary image, find the largest connected component.</li> <li>10. Perform morphological opening with the structuring element with radius 3; then the resultant image is mask.</li> <li>11. To obtain the liver region, liver mask is multiplied with original image.</li> </ol>

$$Tr(m, n) = \frac{\overline{g(m, n)} - \overline{g_{min}}}{\overline{g_{max}} - \overline{g_{min}}} \tag{1}$$

$$In(m, n) = 1 - \frac{\overline{Ho(m, n)} - \overline{Ho_{min}}}{\overline{Ho_{max}} - \overline{Ho_{min}}} \tag{2}$$

$$Fa(m, n) = 1 - Tr(m, n) \tag{3}$$

$$Ho(m, n) = abs(\overline{g(m, n)} - \overline{g(m, n)}) \tag{4}$$

### 3.4 Feature Extraction

Features [12] like size, shape and texture are used to describe and understand the content of image. It is difficult to classify the human body organs using shape and gray-level information. So texture is one of the most used features in medical image processing problems. Difficulties in medical imaging problems are that shape of the organ is not same throughout all 2D slices and gray-level intensities are overlapped with soft tissue. So texture feature is used to discriminate the repeating pattern among different organ tissue. In the proposed work, second-order statistical features are used. Gray-level co-occurrence matrix is used to extract the second-order statistical texture features. Table 3 gives GLCM features.

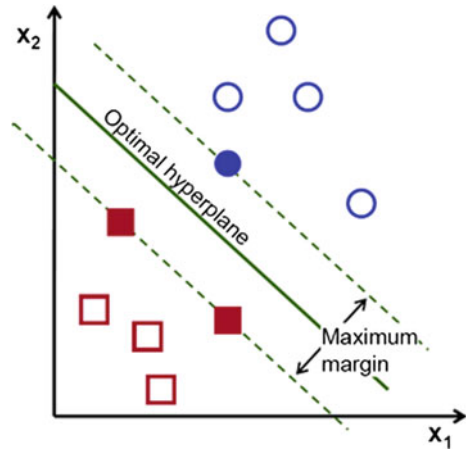
### 3.5 SVM Classifier

SVM classifier [12, 13] is one of the classifier algorithms that uses optimal hyper plane which separate different classes from each other. The two sections in SVM classifier are training section and classification section. In the first section, training samples are used to find the hyperplane. Hyperplane is maximum distance between

**Table 3** GLCM features

Features	Description	Formula
Energy	Measure of repeated pixels. The energy is high, if the occurrence of repeated pixels is high	$Energy = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i, j)^2$
Entropy	Measure of information present in the image	$Entropy = - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P(i, j) \log P(i, j)$
Homogeneity	Extracting the feature at various resolutions	$Homogeneity = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{P(i, j)}{1+ i-j }$
Contrast	Local intensity variation of the image	$contrast = \sum_{n=0}^{N-1} n^2 \left\{ \sum_{i=1}^N \sum_{j=1}^N P(i, j) \right\}$

**Fig. 3** Illustration of hyperplane



two classes. The points which are close to hyperplane are called support vectors. With the help of support vector, boundary is created for the hyper plane margin.

In second section, hyperplane is used to classify network data points. Illustration of hyperplane is shown in Fig. 3. MATLAB software has an application for classification learner and used for SVM classification. Using this, train set is created.

## 4 Experimental Results and Discussion

### 4.1 Experimental Results

In this work, 60 abdominal CT images are considered for experimental results. These images are collected from various sources from Internet [17]. Ten images are normal image without diseases. Remaining 10 images are abnormal images. Those images are taken for training set. Another 20 images of normal and 20 images of abnormal are taken for testing set. The proposed work was implemented in MATLAB. The proposed method output is shown in Fig. 4. Preprocessed image is shown in Fig. 4a. Preprocessed image is converted to NS domain, and the results are shown in Fig. 4b–e. Segmented liver and extracted tumor are shown in Fig. 4g, h.

### 4.2 Performance Measures

To understand and evaluate the performance of proposed method, several quality measures [12] can be used.

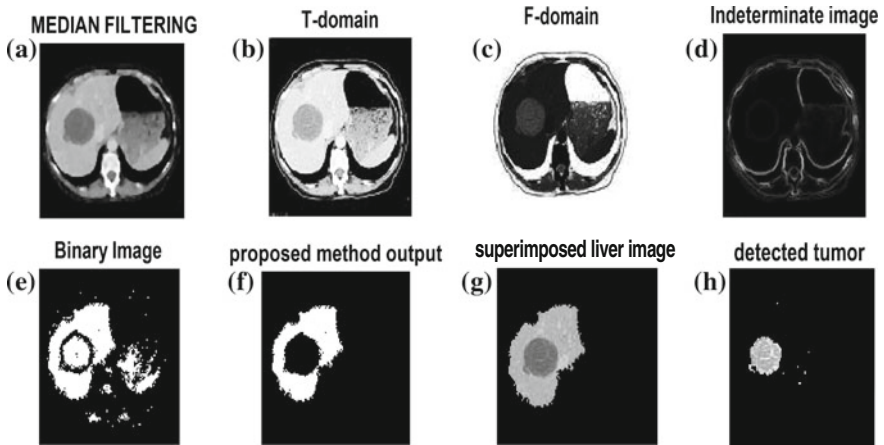


Fig. 4 Output of proposed method

$$\text{Dice coefficient} = \frac{2|A \cap M|}{|A + M|} \quad (5)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (6)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (7)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (8)$$

$$\text{Misclassification rate} = (\text{FP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (9)$$

TP is true positive: Patient has the disease, and the test is positive. FP is false positive: Patient does not have the disease, but the test is positive. TN is true negative: Patient does not have disease, and test is negative. FN is false negative: Patient has the disease, but the test is negative.

### 4.3 Experimental Discussion

The objective of the proposed method is to obtain good level of accuracy in liver extraction, quantitative evaluation utilizes Dice coefficient to evaluate the performance of segmentation results. This hybrid segmentation output is compared with existing method [1], and the proposed method obtained segmentation efficiency of 97%. Table 4 shows the Dice coefficient of hybrid method.

After extracting the liver, 6 features were extracted. The excessive features will slow down the training process, and sometime it will mislead the classifier. However, more features increase the computational space and time complexity. Hence, it is essential to use limited features. These features are used to form feature vector. These feature vectors (size is  $20 \times 6$ ) are given as input to SVM classifier. Feature

**Table 4** Dice coefficient for hybrid technique

Samples	Proposed method	Ground truth	Dice coefficient
Image 1	1344	1942	0.986
Image 2	13,309	13,886	0.9788

**Table 5** Feature extracted using GLCM

Samples	Mean	Standard	Entropy	Contrast	Energy	Homogeneity
Image 1	32.7	62.4	1.6217	0.1098	0.6254	0.9779
Image 2	81.1	114.032	1.3482	1.6576	0.4989	0.9550
Image 3	43.7	80.18	2.015	0.3636	0.5991	0.9662
Image 4	37.5	74.1	1.988	0.4966	0.6207	0.9476
Image 5	31.15	52.12	2.3194	0.1281	0.5649	0.9649

**Table 6** Results of SVM classification

Actual class	Predicted		
	Normal	Abnormal	Total
Normal	19 (TP)	1 (FP)	20
Abnormal	2 (FN)	18 (TN)	20
	21	19	40

**Table 7** Results of specificity, sensitivity, accuracy, misclassification rate

Classifier	Measures			
	Specificity	Sensitivity	Accuracy	Misclassification rate
SVM (%)	95	90	95	7.50

vector sample is listed in Table 5. Generally, normal liver image have less energy, less homogeneity and more entropy value.

There are three steps to perform the classification: Load the training data, create SVM model, and test and classify the new inputs. In this work, classification is done by using binary class SVM. The selected feature vectors are given as input, and group is assigned for two classes. Class 0 symbolizes normal liver, whereas class 1 symbolizes abnormal liver. The quantitative metrics used for analyzing the classifier performance are specificity, sensitivity, accuracy and misclassification rate. Accuracy is closeness to the actual output. Sensitivity is ability to identify the patient with the diseases. Specificity is ability to identify the patient without the diseases. Tables 6 and 7 show the results of SVM classifier.

After classification, abnormal images are identified. For those abnormal images, dynamic thresholding were performed to extract the tumor. Based on tumor size, it is further classified into benign or malignant tumor

## 5 Conclusion

The proposed method uses hybrid technique which is the combination of neutrosophic with largest connected component. This hybrid technique achieved segmentation efficiency of 97 and 93% of tumor segmentation. Features were extracted from segmented liver, and these feature vectors were used to train classifier. The experimental results show that SVM classifier achieved 95% accuracy. By increasing the number of samples, performance measures can be improved. This method is used to diagnosis the liver diseases and also able to find abnormality in abdominal CT images.

**Acknowledgements** I would like to thank Professor Dr. B. Kanmani, Dean of Academics, BMS college of Engineering, for guiding me and providing necessary resources.

## References

1. Gotra, A., Sivakumaran, L.: Liver segmentation: indications, techniques and future directions. *Insight Imag.* **8**, 377–392 (2017)
2. Jayanthi, M.: New edge preserving filter for better enhancement of liver CT images. *Indian J. Sci. Technol.* **10**(10), 1–7 (2017)
3. Jayanthi, M.: Segmentation of liver abnormality using label connected algorithm. *IJSET* **6**(7), 247–249 (2017)
4. Jayanthi, M.: Comparative study of different techniques used for medical image segmentation of liver from abdominal CT scan. In: *IEEE WiSPNET 2016 Conference*, pp. 1462–1465 (2016)
5. Sayed, G.I., Ali, M.A., Gaber, T., Hassanien, A.E., Snasel, V.: A Hybrid Segmentation Approach Based on Neutrosophic Sets and Modified Watershed: A Case of Abdominal CT Liver Parenchyma. *IEEE*, pp. 144–149 (2015)
6. Kumar, S.S., Moni, R.S., Rajeesh, J.: Contourlet transform based computer-aided diagnosis system for liver tumor on computed tomography images. In: *International Conference on Signal Processing, CCN Technologies* (2011)
7. Joshi, D., Londhe, N.D.: Automatic liver tumour detection. *IJCTEE* **3**(1) (2013)
8. Priyadarsini, S., Selvathi, D.: Survey on segmentation of liver from CT images. In: *IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pp. 234–238 (2012)
9. Gunasundari, S., Suganya Ananthi, M.: Comparison and evaluation of methods for liver tumor classification from CT Dataset. *Int. J. Comput. Appl.* **39**(18) (2012)
10. Kumar, S.S., Moni, R.S., Rajeesh, I.: Automatic liver and lesion segmentation: a primary step in diagnosis of liver diseases. *Signal Imag. Video Process.* <https://doi.org/10.1007/s11760-011-0223-y> (2011)
11. Fujita, H., Zhang, X., Kido, S., Hara, T., Zhou, X., Hatanaka, Y., Xu, R.: Introduction to CAD System. In: *International Conference on Future Computer, Control and Communication 2010*, pp. 200–205 (2010)
12. Mala, K., Sadasivam, V., Alagappan, S.: Neural network based texture analysis of liver tumor from computed tomography images. *Int. J. Biol. Biomed. Med. Sci.* **2**(1), 33–37 (2007)
13. Kumar, S.S., Moni, R.S., Rajeesh, I.: Automatic liver and lesion segmentation: a primary step in diagnosis of liver diseases. *Signal Imag. Video Process.* <https://doi.org/10.1007/s11760-011-0223-y> (2011)
14. Gao, L., Heath, D., Kuszyk, B.: Automatic liver segmentation technique for three- dimensional visualization of CT data. *Radiology* **201**, 359–364 (1996)



15. Zhang, B.: A Novel Approaches to Image Segmentation Based on Neutrosophic Logic (2010)
16. Priyadarsini, S., Selvathi, D.: Survey on segmentation of liver from CT images. In: IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 234–238 (2012)
17. [www.mayoclinic.org](http://www.mayoclinic.org)

# Improving Accuracy of Short Text Categorization Using Contextual Information



V. Vasantha Kumar, S. Sendhilkumar and G. S. Mahalakshmi

**Abstract** Categorization plays a major role in information retrieval. The abstracts of research documents have very few terms for the existing categorization algorithms to provide accurate results. This limitation of the abstracts leads to unsatisfactory categorization. This paper proposed a three-stage categorization scheme to improve the accuracy in categorizing the abstracts of research documents. The abstracts on most cases will be extending the context from the surrounding information. Initially, the context from the environment in which the abstract is present is extracted. The proposed system performs context gathering as a continuous process. In the next stage, the short text is subjected to general NLP techniques. The system divides the terms in the abstract into hierarchical levels of context. The terms contributing to the higher levels of context are taken forward to the further stages in categorization. Finally, the system applies weighted terms method to categorize the abstract. In case of uncertainties arising due to the limited number of terms, the context obtained in the initial stage will be used to eliminate the uncertainty. This relation of the context to the content in the short text will provide better accuracy and lead to effective filtering on content in information retrieval. Experiments conducted on categorization of short texts with the proposed method provided better accuracy than traditional feature-based categorization.

**Keywords** Context · Short text · Categorization

---

V. Vasantha Kumar (✉)

Department of Computer Science and Engineering, KCG College of Technology, Chennai, India  
e-mail: vasuvasu.aroma@gmail.com

S. Sendhilkumar

Department of Information Science & Technology, Anna University, Chennai, India  
e-mail: thamaraikumar@annauniv.edu

G. S. Mahalakshmi

Department of Computer Science and Engineering, Anna University, Chennai, India  
e-mail: gsmaha@annauniv.edu

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_26](https://doi.org/10.1007/978-981-13-1708-8_26)

## 1 Introduction

Categorization of content retrieved from the web is an integral part of all Information Retrieval activities. The information available on the Web is enormous and grows every day. Categorization is the need of the hour to filter out the required content from all the information that was retrieved. This problem of categorization becomes further more challenging when it comes to short texts. Being short is only half of what conciseness is all about. However, in digital bibliographies, short text snippets are the face value of the underlying research article. We mean: Abstracts, article title, author details, publication information, references, citations, etc., form the short text snippets in digital bibliographic libraries. Among these, 'abstracts' outline the essence of what is actually said in the research article and, therefore, is the major factor in determining the quality of research article.

Current digital libraries and other bibliographic repositories of technical information provide researchers with useful information to obtain citation details, author-specific details, reviews, comments, etc. However, most of the repositories and libraries do not provide free access to the entire article. With the title and the abstract available, researchers deal with the problem of imprecise categorization due to the low term frequencies. Even the keywords provided by the authors do not help much in the categorization [8]. In most cases, the keywords added to the confusion in categorization. The absolute frequencies of terms in the case of titles and abstracts mostly one or two, and thus the change in the occurrence of a keyword by one, will produce a change in the categorization.

The proposed system relates the content in the short text to the context. By context, we mean the semantic context as well as the surrounding information related to the research abstract. Semantic context is measured in terms of evaluation measures, like originality, novelty, relevance, clarity, extendibility, support, and conclusive information. Other contextual information includes the meta-details associated with the research abstracts. This context information is used to eliminate uncertainties that occur due to the limitations of short texts.

## 2 Related Work

### 2.1 *Short Texts in Bibliographic Repositories*

Digital bibliographic repositories serve as a source of collective information for researchers. These repositories are very much helpful since they portray every detail about the research article and its author(s). These meta-information about research articles contribute more than what is expected. A researcher while downloading the abstract of an article may be well convinced with the idea presented in the abstract, so that, he/she further proceeds to download other articles written by the author(s).

However, this comfort arises from the style and appeal with which the abstract is written. Also, there is less or almost no guarantee that the entire research article will also be the same as that of the abstract. In other words, the interest created in the minds of the reader by the abstract may have to be maintained while the reader is exposed to the rest of the article. Though there is no guarantee for this to happen, many times, we, the researchers, rate an article by means of its face value, the abstract. The semantic richness, style of the author(s), message, innovation, novelty, significance, empirical support, etc., attract us to get into the rest of the research article. Till date, there are many works regarding the analysis of text documents from the semantic perspective. However, there are very few for short texts, i.e., abstracts. The reason may be that it is tough to evaluate clarity, soundness, completeness of the abstract which is hardly 150–300 words.

Though many controversies arise with the above idea, it is very pleasing to note that there is only one or uniform variety of representing research abstracts, that too, in the form of texts. Hardly anybody has attempted it to be a visual abstract!! Recently, Elsevier had started an initiative to consider submissions which embed audio/video inside the research publications. In a nutshell, bibliographic repositories are indispensable to academicians and researchers. Bibliographic repositories like Digital Bibliography and Library Project (DBLP) (<http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php>), Cite Seer (<http://citeseerx.ist.psu.edu/>), Springer link (<http://www.springer.com>) provide citation details, author-specific details, reviews, comments, etc. However, most of the online digital libraries and repositories provide access only to abstracts and titles. Still, the amount of information available via these repositories is enormous. Therefore, the knowledge conveyed by the research abstracts need to be measured for variety of interesting applications like, analyzing citation statistics and evolution of research networks. Processing of short texts like abstract, article title, keywords, and authors [5, 18, 19]. Analysis of the content available in the digital libraries in order to generate customized results is another area of research. Other fields of work on the digital libraries include coauthor network analysis [18, 22] and ranking [5, 6]. Effective categorization of these short texts implicitly contributes toward finding knowledge excellence in a social network environment. This paper is an attempt to analyze the contextual information of short texts (research abstracts) to investigate how they contribute to improving the accuracy of text categorization.

## ***2.2 Short Text Analysis in the Evolution of Knowledge Networks***

Evolution of knowledge network from digital repositories helps researchers to pursue their work in a smart and intelligent way. Social networking elements like Blogs and Discussion Forums are not valued much due to their inconsistent standards. Though semantics exist in digital libraries and repositories for establishing meaningful relationship among authors and their work, it does not provide certain features like the

author bonding, knowledge transition of authors, domain-specific author quality, levels of author contribution, author centrality. Hence, researchers obviously need to put additional manual effort in standard bibliographic repositories like DBLP, Cite seer for better knowledge acquisition. This implicitly demands for a system to facilitate categorization of authors and the works of individuals. The feasibility for the transition of social network to knowledge network is not simple because of the limited availability of full content for free. Most digital libraries provide free access only to abstracts. In [20] a method for building a knowledge network based on ontology and Vector-Space model was mentioned. Also, unsupervised, distributed, and self-organizing approach to build knowledge networks is feasible [12]. This self-organizing approach helps in tacking the enormous amount of information available from the Web.

### ***2.3 Short Text Classification***

Categorization of short texts is fast growing as an important area of research lately. Categorization of short texts in Social Networking Sites is of interest with the growing trend of social networking [4, 11]. Various methods for improving the short text categorization including having a large secondary corpus to improve classification [13, 21], Kullback–Leibler distance [7], measuring similarity between short texts [16], and categorization by expanding the keywords and other ingenious methods have been proposed earlier [9, 14].

### ***2.4 Text Mining Techniques and Short Texts***

The structure of the Web documents is not consistent. This heterogeneous nature of the Web is what makes the task of mining from the Web a challenging task. Further, short texts like abstracts compound to the problem. The limited amount of terms available in short texts makes it difficult to categorize, analyze, and evaluate them. Due to the huge amount of information available in social media, there is a clear need for mining useful information from the available short texts in order to discover knowledge about the collective thinking of the various individuals. Here comes a focus on text mining processes. The need for text mining requires a place in the construction of knowledge network. Text mining techniques like clustering and classification are used to categorize and group bibliographic database in a domain-specific way.

Various algorithms like naive Bayesian, feature extraction, pattern matching prevail to carry over such domain-specific categorization. Researches on improving the existing methods are aimed at improving the effectiveness of the categorization [15]. In general for all these classification algorithms, the predefined class labels are assigned to each distinctive class. Then the given input is categorized on the basis of

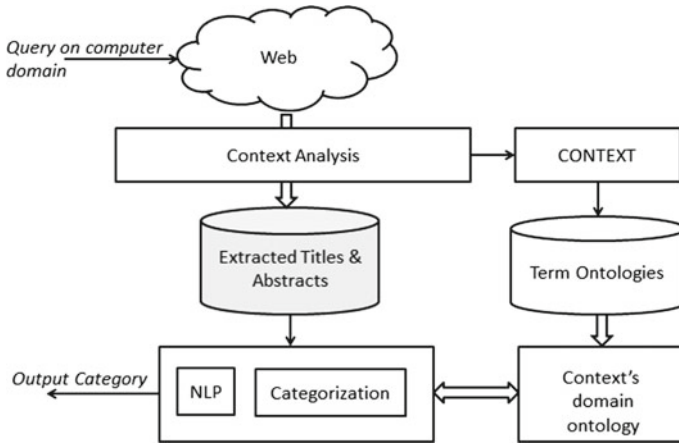
the labels. Similarly, clustering algorithms like k-means, agglomerative clustering, MAJORCLUST [1] have been extensively studied and improved. Mostly, in these algorithms, the centered value is assigned at random to choose for clustering distance. The distance measure then determines the category of the document. Both the traditional algorithms and the similarity clustering methods result in uncertainty when applied to short texts. This is because of the very low term frequencies in the case of the short texts. Due to this nature, the occurrence or absence of a single word may impact the classification of the short text to a great extent. More intelligent and unsupervised methods applying topic models for short text classification are discussed in the literature [24–26].

## ***2.5 Contextual Information***

With the growth of pervasive and mobile computing, contextual information is used in the content filtering for applications where context information is available. Applications make use of contextual information to personalize content based on user interests [2, 17]. Contextual information is also used in other domains like health care. Research on context gathering aims in eliminating the ambiguity in the collected context [3] and in personalization of mobile applications to the context. Analyzing context provides lots of information that would otherwise go unnoticed. Though the access in the digital repositories is mostly limited to the abstracts of documents, other information is available in the repositories which include keywords, related articles, hyperlinks, comments, reviews. Analysis of this information leads to valuable knowledge about the research article. This knowledge about the context in turn will help in improving the categorization and in the elimination of the uncertainties in them.

## **3 Short Text Categorization Based on Contextual Information**

Using the contextual information in categorization aims to remove the uncertainties in the categorization. Categorization of the research publications based on titles alone leads to unstable results. The titles are either unbiased toward a domain or too short for analysis. This limitation of the titles can be reduced to an extent by using abstracts for the categorization. Abstracts provide a better scope for categorization. However, the term frequencies in abstracts being limited do not provide basis for a strong categorization. The occurrence of each keyword in the abstract is mostly once or twice. To improve the categorization, the terms were weighed using domain-specific ontologies. However, this still had uncertain results as completeness of ontologies for the domains is in itself a major challenge. In this scenario, by bringing in the



**Fig. 1** Short text categorization using contextual information

contextual information of the abstract, the categorization is strengthened. In this proposed system, the gathering of the contextual information is done as a continuous process. The context is extracted from the search key, review and comments, author details, citation and other information that can be retrieved along with the abstract. This surrounding information will be used in decision making whenever there is an uncertainty in the categorization of the abstract.

Contextual information is readily available and can easily be extracted. The context is extracted from the information surrounding the abstract and the title. This includes hyperlink to related articles, comments and reviews about the article. Further, the context provides a better scope to categorize by overcoming the limitations of the abstracts. This categorization is done based on ACM classification of domains as on the year 1998 (<http://www.acm.org/about/class/1998>). Nearly 50 plus pre-classified computer science (or related) domains are used in this. The initial stage (Fig. 1) of the categorization involves NLP methods. The terms after preprocessing are weighted on the basis of their distance in the domain ontology. The contextual information is used to identify the appropriate domain. Once the terms are weighted, they are grouped on their weights. The terms with higher weights are taken forward to the further stages in categorization. This omission of terms with lower weights in the domain makes the input text biased toward categories within the domain.

## 4 Analysis of Semantic Context

The semantic context of the research abstract includes various factors which lead to determining the quality of research abstract and, thereby, the research article. In this context, the article quality of research publications has already been analyzed from

the plagiarism perspective [23]. However, plagiarism analysis of a text document (here, research abstract) will only provide the details about abstract originality. There are various other features to be considered while analyzing the semantic context.

- **Originality** ( $\alpha$ ): Originality is the measure depicting the ‘individuality’ of the abstract. Here comes the plagiarism viewpoint. As research abstracts are meager sections, assessing their originality based on statistical approaches does not suite well. A research abstract should not be a verbatim copy of the any other. Abstract originality can be assessed by means of similarity detection. The originality measure is obtained from fuzzy cognitive maps (FCMs). FCMs are the mind map/mental maps that are obtained from the domain ontological concepts. FCMs thus obtained from the research abstracts are compared to obtain the similarity. This is proposed in order to detect idea/concept wise similarity would be detected.
- **Novelty** ( $\beta$ ): Novelty is highly similar to originality. The ontological concepts relating to each terms present in the abstracts (after preprocessing) are initially identified. Later, the distance between the core terms with other related concepts is measured. The cumulative distance measure so obtained is assigned as a novelty measure.
- **Relevance** ( $\gamma$ ): Relevance is the measure of finding significance of the title with the abstract content. The terms in the title are matched with those in the abstract thereby the relevancy between the title and the abstract is measured.
- **Clarity** ( $\delta$ ): Clarity involves the level of presentation in conveying a fact to the reader. Ambiguity or chaos present in the content delivery need to be removed as readers cannot understand the idea behind the work. Completeness and consistency are to be analyzed while measuring the level of clarity. Also the length of the sentences/sections is accounted in order to find the abstract clarity.
- **Extendibility** ( $\vartheta$ ): This is the factor where a research work needs further enhancement. Generally, the ‘Future work’ section is present in most of the research articles which depicts the requirement of extendibility of the published work. For example, a paper which is published based on architecture of the system can be extended to empirical analysis after performing experiments. In research abstracts the heuristic terms like ‘extend’, ‘improve’.
- **Support** ( $\lambda$ ): In a research article, the support factor would be measured from the various heuristic terms like ‘experiments’, ‘compare’, ‘investigate’. These terms are used to assess that the abstract is providing a supportive description toward the proposed methodology.
- **Conclusion** ( $\omega$ ): This would also be measured from the heuristic terms like ‘conclude’, ‘arrive’, ‘decide’, ‘results’, ‘confirm’. The conclusion that the abstract makes is assessed and a gradient index value is given to this measure.

Using the above contexts, research abstracts are analyzed for quality and are subjected to further categorization. The following section discusses the empirical investigations and the observations obtained.



## 5 Results and Discussion

A total of 500+ titles and abstracts of research publications in the area of computer networks were extracted from journals and conference proceedings. The titles contained 4–20 words and the abstracts from 50 to a few hundred words. Care was taken to ensure that the research articles spanned across various areas of research under computer networks. This was done in order to retain the heterogeneity of the Web. All experiments were done on these 500 titles and abstracts. A number of experiments were done, from using traditional methods on the titles alone to using contextual information to aid in classification on the titles and abstracts together. The context gathered from the surrounding information like the comments, reviews and the search keys is used to ascertain the domain of interest. A static ontology on the terms related to this domain is used to weigh the terms based on their proximity to the domain. These domain ontologies are made available of the system and are selected on the basis of the contextual information. Weights were given in the scale of 1–1.5, the nearest terms given 1.5 and the farther ones 1. This weighing of terms helped to eliminate the uncertainties that were prevailing due to the limited number of terms in the short texts.

### 5.1 *Categorization by Structural Context of Research Abstracts*

Applying traditional feature selection methods on the titles resulted (Table 1) in a high level of uncertainty in the categorization. The titles in general had too little terms for them to be biased toward the domain. Further, few of the titles had very little terms that they were not categorized under any domain. Abstracts, though having limited terms, provided for a better categorization when compared to the titles. The uncertainty remained in the categorization as more than a hundred documents were categorized under more than one domain. Using abstracts and the titles together had very little impact on the categorization as most of the papers had the titles replicated in the abstracts (Table 2).

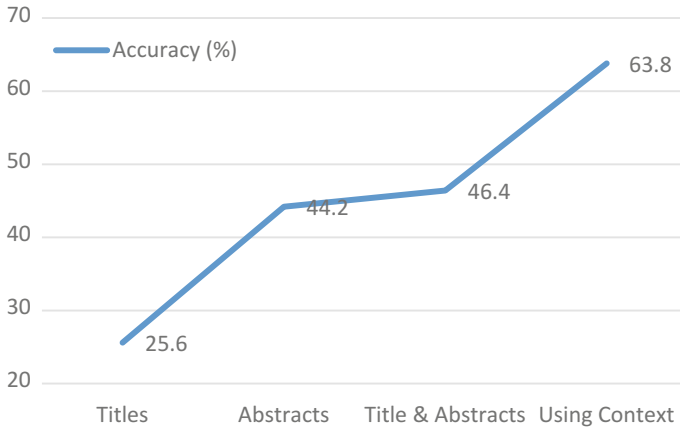
The usage of the contextual information provided a substantial improvement (Fig. 2) in the accuracy of the categorization. The problem of abstracts getting cat-

**Table 1** Traditional feature-based categorization

	Correct	Uncertain	Incorrect	Not categorized
Titles	128	139	251	5
Abstracts	221	28	251	0
Title and abstracts	232	22	246	0
Using contextual information	319	0	181	0

**Table 2** Categorization using contextual information

	Correct	Uncertain	Incorrect	Not categorized
Titles		139	251	5
Abstracts		28	251	0
Title and abstracts		22	246	0
Using contextual information		0	181	0



**Fig. 2** Comparison of short text categorization accuracy

egorized under more than a domain was greatly reduced. However, only 64 percent of the documents were rightly classified under the networks domain. A few straightforward research papers in the networks domain were also categorized incorrectly under the different domains. The paper titled “A Constrained QoS Multicast Routing Algorithm in Adhoc” gets categorized into the domain of numerical analysis. These anomalies occurred due to the limitations of using static domain ontologies. One way to deal with this is to add more weightage to the terms closer to the context’s domain. The accuracy increases with the increase in the weight of the terms in the context. However, increasing the weights may lead to total bias toward the context’s domain and may prove erroneous in the case of misreading the context.

## 5.2 *Categorization by Semantic Context of Research Abstracts*

The 500+ research abstracts are analyzed for above-mentioned semantic features. Table 3 outlines the results obtained for research abstracts across every measure.

The above findings are altered when the abstracts are analyzed for semantic contextual features. For analysis, we have only considered two contextual features: Clar-

**Table 3** Categorization by semantic contextual features

	High	Medium	Low
Clarity	113	327	60
Conclusiveness	276	0	324
Extendibility	168	136	196
Novelty	271	221	8
Originality	368	0	132
Relevance	76	324	100
Support	37	110	353

**Table 4** Categorization accuracy using clarity and relevance as the semantic context

	Relevance	Clarity
Interval 1	145	160
Interval 2	174	159

**Table 5** Categorization accuracy % using clarity and relevance as the semantic context—with finer intervals

	Relevance	Clarity
Interval 1	68	64
Interval 2	62	67
Interval 3	54	54
Interval 4	60	63
Interval 5	61	60

ity and Relevance. The results obtained are convincing (Table 4). Here, we have recorded the improved categorization results for the semantic features: Clarity and Relevance. The 500+ research abstracts are divided into two equal classes. The categorization results with clarity and relevance as the semantic context is shown in Fig. 8. There is not much difference between both the semantic contexts in terms of categorization accuracy.

However, the same test set is divided into five equal intervals and the categorization results are recorded as in Table 5. The average categorization accuracy is 61.6% for clarity and 61% for relevance as the semantic context. It can be seen that the accuracy reduces up to 2% when semantic context is included for text categorization. However, being a valuable measure, including more semantic features as discussed in this paper (Sect. 4) would definitely lead to more accurate and meaningful categorization mechanism which may produce fruitful results in short text mining and related applications.

The objective of this paper is to prove that the short text categorization accuracy improves by including the contextual information, and therefore, we tend to propose more features which are really useful in categorizing the research abstracts more meaningfully as visualized by the minds of a naïve researcher. With this in mind, we

would like to do more things in future: as like including more semantic features, and more minute details about the semantic capital of research abstracts. The implementations are on, and we would like to add more results to the existing approach as we proceed further.

## 6 Conclusion

In this paper, a method for improving the accuracy of categorization of short texts using context information of short text is proposed. This contextual information is used as a means to bias the short texts toward a domain. This biasing is done in order to reduce the uncertainties in the categorization of short texts. The effectiveness of this system can be further enhanced by dynamically updating the proximity and thereby the weights of the terms with respect to the various domains. Our future work in this area would be the development of self-learning ontologies with the aim of improving the completeness and validity of the ontologies. Other improvements in this method can be obtained by concentrating on context collection and analysis of the context. This study will help in improving the effectiveness in the collection of context. The knowledge of the impact of various neighborhood information on the context will in turn improve the categorization in this method.

## References

1. Alexandrov M., Gelbukh A., Rosso P.: An approach to clustering abstracts. In: Proceedings of the 10th International Conference NLDB-05, volume 3513 of Lecture Notes in Computer Science. Springer, pp. 275–285 (2005)
2. Dey, A.K., Jennifer, M.: Designing mediation for context-aware applications. *ACM Trans. Comput. Human Interact. (TOCHI)* (2005)
3. Devaraju, A., Hoh, S., Hartley, M.: A context gathering framework for context-aware mobile solutions. In: *Mobility '07 Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology* (2007)
4. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2010)
5. Fiala, D., Rousselot, F., Ježek, K.: PageRank for bibliographic networks. *Scientometrics* **76**(1), 135–158
6. Fiala, D.: Mining citation information from CiteSeer data. *Scientometrics* **86**(3) (2011)
7. Pinto, David, Benedí, José-miguel: Paolo Rosso. Clustering Narrow-Domain Short Texts by using the Kullback-Leibler Distance, *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science* (2007)
8. Pinto, D., Rosso, P.: KnCr: a short-text narrow-domain sub-corpus of med-line. In: Proceedings of TLH-ENC06, pp. 266–269 (2006)
9. Ingaramo, D., Errecalde, M., Rosso, P.: A general bio-inspired method to improve the short-text clustering task. *Lecture Notes in Computer Science*, 2010, vol. 6008, *Computational Linguistics and Intelligent Text Processing*, pp. 661–672

10. Metzler, D., Dumais, S., Christopher, M.: Similarity measures for short segments of text. In: Proceedings of the 29th European conference on IR research, ECIR'07 (2007)
11. Perez-Tellez, F., Pinto, D., Cardiff, J., Rosso, P.: On the difficulty of clustering company tweets. In: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, October 30, 2010, Toronto, ON, Canada (2010)
12. Castelli, G., Mamei, M., Zambonelli, F.: A Self-organizing approach for building and maintaining knowledge networks. In: Proceedings of Mobile Wireless Middleware, Operating Systems, and Applications. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 48, no. 1, Part 4, pp. 175–188
13. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. KDD* 2(2), 1–25 (2008)
14. Wang, J., Zhou, Y., Li, L., Hu, B., Hu, X.: Improving short text clustering performance with keyword expansion. In: The Sixth International Symposium on Neural Networks Advances in Soft Computing (ISSN 2009), vol. 56, pp. 291–298 (2009)
15. Schneider, K.-M.: Techniques for Improving the Performance of Naive Bayes for Text Classification, Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science, vol. 3406/2005, pp. 682–693 (2005)
16. Abdalgader, K., Skabar, A.: Short-text similarity measurement using word sense disambiguation and synonym expansion. *Lecture Notes in Computer Science*, 2011, vol. 6464, AI 2010: Advances in Artificial Intelligence, pp. 435–444 (2010)
17. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: WWW' 10 Proceedings of the 19th International Conference on World Wide Web (2010)
18. Biryukov, M.: Co-author Network Analysis in DBLP: Classifying Personal Names, Modelling, Computation and Optimization in Information Systems and Management Sciences Communications in Computer and Information Science, vol. 14, Part 1, Part 2, pp. 399–408 (2008)
19. Reuther, P., Walter, B., Ley, M., Weber, A., Klink, S.: Managing the quality of person names in DBLP. In: Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science, vol. 4172/2006, pp. 508–511 (2006)
20. Lee, Pei-Chun, Hsin-Ning, Su, Chan, Te-Yi: Assessment of ontology-based knowledge network formation by vector-space model. *Scientometrics* 85(3), 689–703 (2010)
21. Zelikovitz, S., Hirsh, H.: Improving short-text classification using unlabeled background knowledge to assess document similarity. In: Proceedings of the Seventeenth International Conference on Machine Learning (2000)
22. Klink, S., Reuther, P., Weber, A., Walter, B., Ley, M.: Analysing Social Networks Within Bibliographical Data, Database and Expert Systems Applications. *Lecture Notes in Computer Science*, 2006, vol. 4080/2006, pp. 234–243
23. Deepika, J., Mahalakshmi, G.S.: Towards knowledge based impact metrics for open source research publications. *Int. J. Internet Distrib. Comput. Syst.* (2011)
24. Chen, Q., Yao, L., Yang, J.: Short text classification based on LDA topic model. In: 2016 International Conference on Audio, Language and Image Processing (ICALIP), pp. 749–753. IEEE (2016)
25. Nowak, J., Taspinar, A., Scherer, R.: LSTM recurrent neural networks for short text and sentiment classification. In: International Conference on Artificial Intelligence and Soft Computing, ICAISC 2017: Artificial Intelligence and Soft Computing, vol. 10246, pp. 553–562. LNCS (2017)
26. Li, P., He, L., Hu, X., Zhang, Y., Li, L., Wu, X.: Concept based short text stream classification with topic drifting detection. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, pp. 1009–1014 (2016). <https://doi.org/10.1109/icdm.2016.0128>

# Efficient Classification Technique on Healthcare Data



Rella Usha Rani and Jagadeesh Kakarla

**Abstract** In the process of improvisation of classification accuracy rate, many classifiers are made pass through this framework of classification and prediction. The proposed best classifier is the random forest classifier based on various parameters. The kidney disease diagnosis is done based on the available machine learning classifiers and presented an effective classifier with great accuracy rate of prediction. In the proceedings of ayush to kidney (AtoK) kidney disease diagnose intelligence model, this paper is reporting the best classifier in development of prediction model.

**Keywords** Classifier · Accuracy · Kidney · Machine learning

## 1 Introduction

Through the great more study analysis, the classification is the only way to distinguish the common properties of a class, and similarity index will give the momentum to do categorization by qualitative attributes and termed as classification of qualitative data [1–3]. The criteria selection to do the classification comparison is basically a very complex measure to be undertaken [4]. This shows a clear view of data selection which is the major role in classification analysis. Data preprocessing and analysis from different research report using data mining tool techniques [5]. Taking the learning techniques as fundamental procedure to classify data, a one-dimensional nonlinearity is compared with two-dimensional and continuous action-based systems [6, 7]. The study nature paved the way to incorporate different techniques inclusion into this class of classification techniques. This paper explains how the initial datasets

---

R. U. Rani (✉)

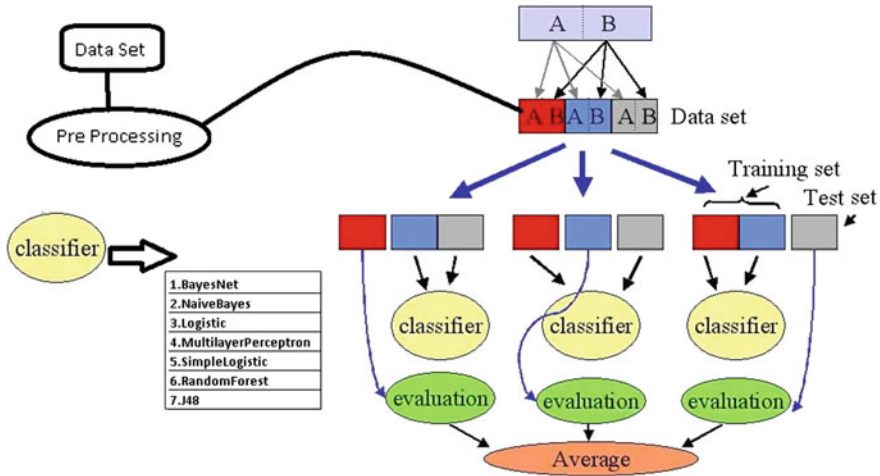
Department of CSE, CVR College of Engineering, Hyderabad, India  
e-mail: usha.shreni@gmail.com

J. Kakarla

Department of CS, Central University of Rajasthan, Ajmer, Rajasthan, India  
e-mail: jagadeesh@curaj.ac.in

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_27](https://doi.org/10.1007/978-981-13-1708-8_27)



**Fig. 1** Classifier model

of kidney disease-affected patients’ health history are cleaned and preprocessed to evaluate the best classification technique that leads to take a decision-making effective in learning aspects.

## 2 Methodology

The dataset is taken from UCI repository as well as collected from hospitals and is cleaned with different data preprocessing techniques, and cross fold technique is used to evaluate the prediction of given dataset by the process of dividing the actual data into training set and testing set, predominantly one as training and all other remaining sets as testing sets [8]. In this k-fold technique of prediction, the k value defined as 10, the model creates a 10 equal size subsample of dataset. The single sample set remains as validation set for testing the data, and all remaining k-1 samples are the datasets that are considered as training data. For effective accuracy, the k value is increased to improve the prediction accuracy rate [9]. The learning classifier system, [6] termed as LCS, handles the function approximation through partitioning the input system for all binary problems. This happens through dividing the input space as hyper-rectangular subspaces through the classifier. Kernel-based classifier is more appropriate for diverse problem condition. Classifier model is the data classification problem framework to classify data (Fig. 1).

Figure 2 is the experimental setup that is carried out through Weka software (a machine learning software for knowledge analysis) [10]. Weka tool is best tool for data mining tasks, data analysis, feature selection, and visualization. To have glance view of weka tool, the screenshot also presented below as Fig. 2 for presenting 400 instances on 25 attributes the preprocessing is shown.



Fig. 2 Data preprocessing

### 2.1 Classifiers Considered

*Bayes net*: “A Bayesian network  $B$  over a set of variables  $U$  is a network structure  $BS$ , which is a directed acyclic graph (DAG) over  $U$  and a set of probability tables  $BP = b(u|ba(u))|u \in U$  where  $ba(u)$  is the set of parents of  $u$  in  $BS$ ”. A Bayesian network represents a probability distributions  $B(U) = \prod_{u \in U} b(u|ba(u))$  [9]. Bayes net evaluate by measuring metrics like divergence score, Bayes score, AIC Score, and entropy values are also being considered. In this proceeding of classification, the inference algorithm calculates the  $\text{argmax } B(y|x)$  through utilizing the distribution  $B(U)$ . It works on Bayes theorem of probability to predict the class of unknown data elements.

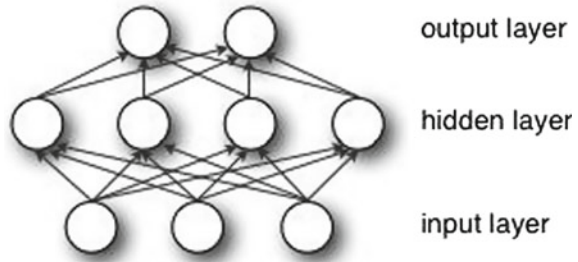
*Simple Logistic*: “Classifier for building linear logistic regression models. Logit-Boost with simple regression functions as base learners is used for fitting the logistic models”. The optimal number of LogitBoost iterations to perform is cross validated, which leads to automatic attribute selection this is the technique for classification.

*Multiplayer Perceptron*: MLP (or artificial neural network—ANN) with a single hidden layer can be represented graphically as follows in Fig. 3.

*Random Forest*: In simple, it can be defined as bagging the randomness occurring in the preferred dataset. It is a classifier algorithm that can perform regression along with classification. This algorithm creates the cross samples or unrelated samples of



**Fig. 3** Layered architecture for MLP



**Table 1** Classification correctness evaluation

Algorithm	Correctly classified	Incorrectly classified	% of correctly	% of incorrect
BayesNet	394	6	98.5	1.5
Naive Bayes	378	22	94.5	5.5
Logistic	391	9	97.75	2.25
Multilayer perceptron	391	9	97.75	2.25
Simple logistic	392	8	98	2
Random forest	396	4	99	1
J48	387	13	96.75	3.25

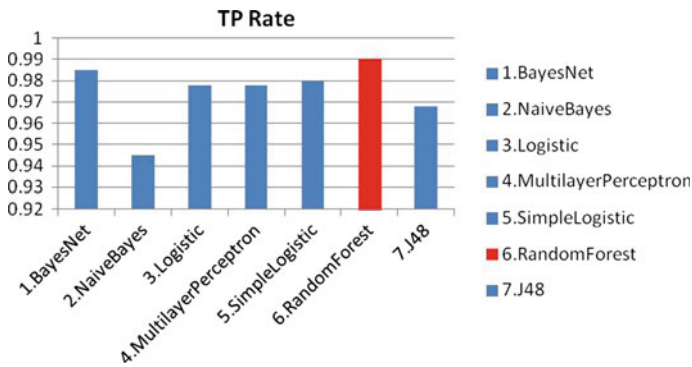
n numbered tree evaluation from the actual data. The changeover taken at this stage of dividing samples is by selecting the best split of data from existing predictors or samples. The grouping or procedure of get-together of these samples framed as a special case of randomness as termed as random forest is formed as trail case of grouping. Finally, this aggregation results with prediction of data with new set of groups called P-tree.

*J48*: Through the different learning, this classifier efficiency is quantized in terms of the parameter checking like confusion matrix, TP rate, FP rate, precision, MCC, ROC area, PRC area. The dividing task of training set and development set as well as a testing test is got initiated from cross folding technique.

Table 1 shows accuracy rate of different classifiers in terms of classified instances. A continuation folding through cross folding technique and after undergoing the repeated classification, the results from different classifiers paved the way of finding which classifier suites best for prediction analysis. The nature of incorrectness in classifier classified instances, in this depiction it was clear that the incorrectness is minute and directly proposes that it is the effective classifier to go ahead with further proceedings of classification on the board of machine learning.

**Table 2** Classification correctness evaluation

Algorithm	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC area	PRC area
Bayes Net	0.985	0.2	0.985	0.985	0.985	0.968	0.999	0.999
Naive Bayes	0.945	0.036	0.951	0.945	0.946	0.891	0.998	0.998
Logistic	0.978	0.016	0.978	0.978	0.978	0.953	0.994	0.992
Multilayer	0.978	0.016	0.978	0.978	0.978	0.953	1	1
Simple logistic	0.98	0.017	0.098	0.98	0.98	0.958	0.999	0.999
Random forest	0.99	0.014	0.99	0.99	0.99	0.979	0.998	0.998
J48	0.968	0.038	0.967	0.968	0.967	0.931	0.976	0.969



**Fig. 4** TP rate on classifiers set

The measuring parameters are TP rate, FP rate, precision, recall, F-Measure, ROC area, PRC area. Each parameter analysis with respect to the different classifiers are graphed below. Table 2 also lists these values for every algorithm which we have considered during experiments. The rating of classifier is the activity performed to present the machine learning effectiveness for the data analytics which is revealed from Figs. 4, 5, 6, 7, 8 and 9.

Each figure presents as one the specific parameter is effective with respective classifier. In taking all this into notice, the effective classifier can be mentioned as the rntandom forest classifier.

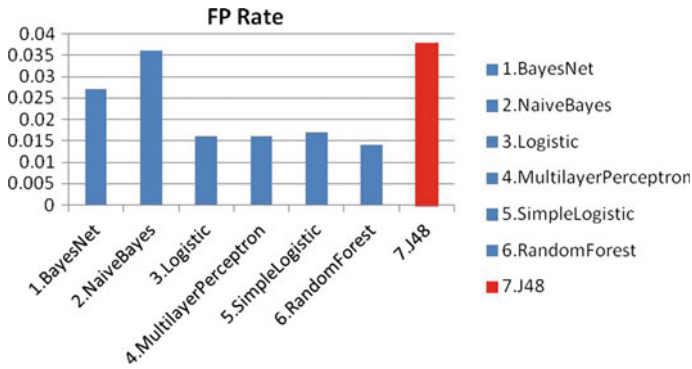


Fig. 5 FP rate on classifiers set

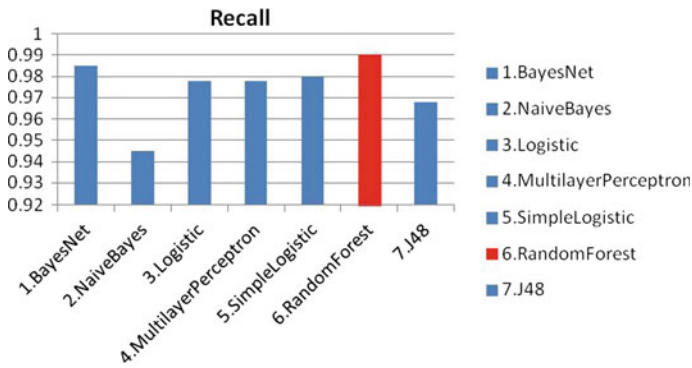


Fig. 6 Recall on classifiers set

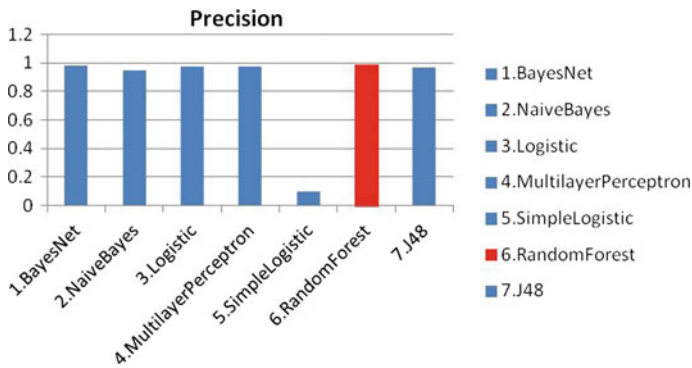


Fig. 7 Precision rate on classifiers set

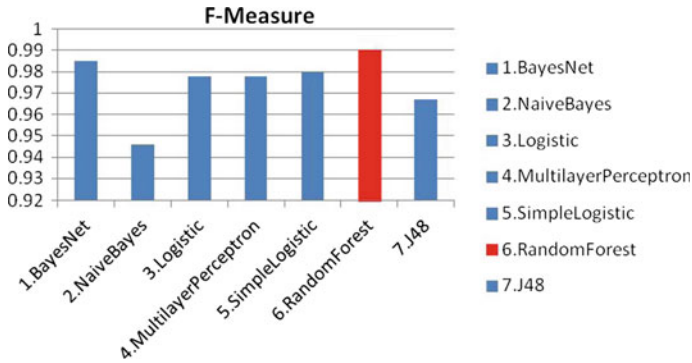


Fig. 8 F-measure on classifiers set

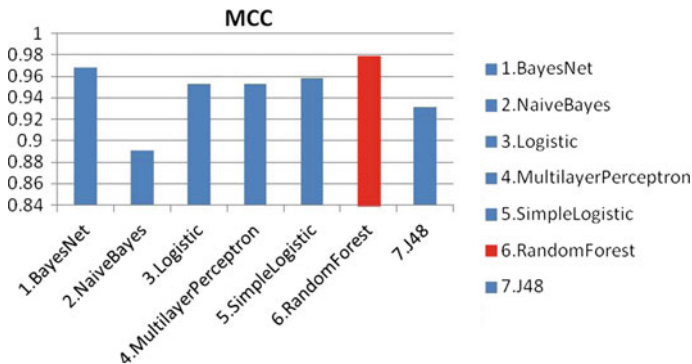


Fig. 9 MCC on classifiers set

### 3 Conclusions

From the classification performed by different techniques, the measurable parameters like TP rate, precision, recall, F-Measure, MCC are showcasing and indicating as the Random Forest classifier is the best one. There is a point that the parameters like FP rate suggest that the J.48 is the best one and according to the parameters like ROC, PRC made us rely that the multilayer perceptron is the best to perform classification under neural network model usage. With the great evaluations, accuracy rate and study process give us a conclusion that the random forest is the best classifier to work with the healthcare data. The future work can be moved toward the other machine learning techniques specifically supervised learning through known features and unknown feature classification to know the classification efficiency for process of prediction.

## References

1. Hyvrinen, L.: 1620 Taxonomy Program (1962)
2. Loomis, R.G., Tanimoto, T.T.: The IBM taxonomy application (ibm 704). In: IBM Application Library (1960)
3. Kupperman, M.: On comparing two observed frequency counts. *Appl. Statist.* 37–42 (1960)
4. Susan, D., Wiley, et al.: *Twenty-First Century Tools For Qualitative Data Analysis* (1996)
5. Krishna Apparao, R.: Statistical and data mining aspects on kidney stones: a systematic review and meta-analysis. *Open Access Scientific Reports* (2012)
6. Iqbal, M., Will, B.N., Zhang, M.: Xcsr with computed continuous action. In: *Australasian Conference on Artificial Intelligence*, pp. 350–361. Springer (2012)
7. Breiman, Leo: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
8. Kanungo, Tapas, Mount, David M., Netanyahu, Nathan S., Piatko, Christine D., Silverman, Ruth, Wu., Angela Y.: An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 881–892 (2002)
9. Butz, M.V.: Kernel-based, ellipsoidal conditions in the real-valued xcs classifier system. In: *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, pp. 1835–1842. ACM (2005)
10. Butz, Martin V: *Rule-Based Evolutionary Online Learning Systems*. Springer (2006)

**Part III**  
**Cryptography and Information Security**

# Two-Phase Validation Scheme for Detection and Prevention of ARP Cache Poisoning



Sweta Singh, Dayashankar Singh and Aanje Mani Tripathi

**Abstract** In data communication, protocols define the set of rules to ensure communication between the hosts over a network. The operation encounters no issue under normal circumstances, but an attacker always seeks for an opportunity to find a loop-hole in the system, to exploit the protocols. ARP cache poisoning is the exploitation of ARP protocol where a malicious attacker aims at binding its hardware address, i.e., MAC with a legitimate entity IP over a LAN. This attempt poisons the cache of the other hosts in the network, causing the traffic diversion to the attacker instead of reaching at genuine host's destination. This paper has proposed a mechanism to validate the new binding received by each host by sending two ICMP probe packets one to the previous binding and other to the new one. New entry of host in the network with no previous entry found in ARP cache is validated using ARP packets to find all the claiming hosts to that IP, used together with ICMP packet to provide a two-phase validation. This scheme being asynchronous in nature also requires no modification in the existing protocol.

**Keywords** Address resolution protocol (ARP) · ARP cache poisoning  
ARP cache spoofing · MITM · ARP vulnerabilities

## 1 Introduction

ARP protocol operating as a transition communication protocol between the network and data link layer aims at mapping the network address of the host with its corresponding data link or MAC address [1]. In LAN, both IP and MAC contribute to

---

S. Singh (✉) · D. Singh · A. M. Tripathi  
Department of Computer Science and Engineering, MMMUT, Gorakhpur, India  
e-mail: swetass22691@gmail.com

D. Singh  
e-mail: dss\_mec@yahoo.co.in

A. M. Tripathi  
e-mail: aanjeymanit09@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_28](https://doi.org/10.1007/978-981-13-1708-8_28)

establish communication between the hosts. A packet is routed over a network using IP address, whereas it is dropped to the correct host using its MAC address. In LAN if a host wants to communicate with other, it should possess both the destination's IP and MAC. If it only knows the IP and not the MAC, calls for ARP. ARP takes the IP of the receiver host and provides its corresponding MAC address. For resolution, Address Resolution Protocol uses two messages—ARP Request and ARP Reply [2]. ARP Request is a broadcast message sent to all available hosts in LAN carrying the IP of destination to identify “who-has the particular destination IP” [3]. The particular IP which matches the destination IP accepts this ARP Request and produces the ARP Reply message defining that the particular IP “is-at” the particular MAC and return with its corresponding MAC. This <IP, MAC> binding is thus maintained in a cache termed ARP cache of the source host for a particular time stamp [2], such that this Request and Reply message need not to be sent again and again. The caching of the binding also relieves network traffic each time for resolution [4, 5]. The entry is stored in host's cache for approx. 20 min and whenever communication is to be initiated is checked each time for the binding. If found in ARP cache of the host, initiates the communication directly and if not again ARP Request is sent for the MAC. In Linux, the ARP table is checked using “arp-a” or “ip neighbour show” command [6].

The functioning of ARP is well operated but comes with two drawbacks—the unauthentic and stateless nature [1]. With its unauthenticated nature, ARP accepts any binding it receives and stores it in cache with no concern if it is coming from a legal end. The stateless nature on the other hand allows the caching of unsolicited ARP reply regardless of whether ARP request was made for that mapping or not. These two loopholes provide an opportunity to the malicious host to attach its MAC and send a fake binding in the ARP reply to the hosts over the network. This binding once cached causes the packet to get rerouted and reach at the attacker computer.

The proposals so far discussed by researchers were classified as per their objective and the adopted approaches. The proposed work aims at detection and prevention with adoption of a non-cryptographic measure of using ICMP probe packet [1, 7, 8]. A table is maintained here to store the MAC-IP binding of each host which persists for a longer duration and is checked each time to validate the binding received. For chance of a new binding, i.e., in case attacker uses an IP of host not in the network an additional validation is adopted by sending ARP request and then next phase is the ICMP-based validation, resulting in 2-phase packet validation. The scheme being distributed do not face issue of centralized failure. An attempt to prevent flooding of spoofed packet is also made in the discussed scheme. The ICMP packets are the echo packets, which are only responded in case of match of both IP and MAC address. The reason of using ARP packets is to analyze the network for malicious host's in the network.



## 1.1 ARP

Address Resolution Protocol (ARP) is a layer 2 protocol functional at the data link layer and associates MAC address with the network (IP) address to discover the nodes to each other. It operates below the network layer in the OSI model as an interface between layer 2 and layer 3, while in TCP/IP protocol suite functioning at layer 2, i.e., the data link layer [1, 3]. Within LAN for communication, it requires both IP and MAC address [2]; IP ensures routing of packet, whereas dropping the packet to correct host is provided using MAC. The sender, for communicating with the destination host, needs to know its MAC. It starts the communication if it retains the <IP, MAC> binding of the destination host but if not found in its local ARP cache sends ARP request to all the hosts in the LAN (broadcast). This ARP request consists of sender IP, MAC and destination IP. The destination IP is matched by all the hosts in the LAN and the host with matching IP only sends a reply whereas other drops the request. The host with intended IP only responds to the requesting host (unicast) providing its MAC. This <IP, MAC> association is held in cache known as ARP table/cache and is updated at a particular interval (approx 20 min). The entries made in the cache can be manual as well as automatic termed as static entries and dynamic entries, respectively. The ARP cache is maintained to avoid sending of ARP request each time resulting in reducing the response time and minimizing the network traffic [4]. Transmission of packet the source host/router knows the IP of the next router except the last router in the path which knows the receiver's IP.

Summarizing the working mechanism, suppose host A (IPA, MACA) wants to start communicating with host B (IPB, MACB) but do not know the MAC of host B, so firstly it will check its cache to find if the <IPB, MACB> binding of B is present in ARP cache. If found A will start communication but if not it will broadcast an ARP request to all the hosts to attain B's MAC address. When B receives this Request sends a response, termed as ARP Reply providing its MAC. With the receipt of ARP reply, host A initiates the communication. This <IPB, MACB> binding will be saved in A's cache for a fixed time duration and is automatically removed after timeout. Both ARP Request and Reply together contribute to attain the binding.

ARP possesses two issues [8, 9]—unauthenticated nature and stateless nature. These limitations of ARP have made it vulnerable to attacks and have gathered the attraction of malicious intention hosts. An attacker could easily send false ARP reply by attaching its MAC with a legitimate host's IP and sending it onto the LAN, which will be accepted by the host without any doubt. Being of unauthenticated nature, there is no way to check if the message is coming from a genuine or a malicious host. Even its stateless nature causes the host to update the cache for an unsolicited reply, i.e., an attacker can easily send an ARP reply irrespective of the Request for it was made or not. Thus by sending fake ARP replies, an attacker can very efficiently poison the cache. The attacker usually aims at sending fake unsolicited ARP replies to attach a spurious <IP, MAC> association. Once successful in updating the cache with fake association, the sender will consider the binding true and will design the packet

to send at the attacker's MAC unknowingly, thus causing all the traffic rerouted at attackers end.

## 1.2 ARP Cache Poisoning

ARP cache poisoning is the mechanism of updating the host's cache with fake <IP, MAC> association where the attacker uses its MAC with a legitimate entity's IP address, causing all the traffic meant for that legitimate host to be dropped or rerouted at attackers end [3]. With this mechanism, the attacker can easily conduct several higher level threats such as posing to be someone else, hiding its identity, interception, etc. [10, 11]. There are several tools available online to perform malicious act. One among them is Ettercap to perform MITM by selecting the target and attaching them to one's MAC. ARP cache poisoning has become one of the most common and easiest ways to perform hacking and exploit the LAN communication. The attacker can easily reside in the network without any trace of its existence.

## 2 Literature Review

Several mechanisms have been previously suggested by the researchers aiming at Detection and Prevention of this poisoning attempt made by the attacker.

A static IP-MAC binding scheme [12] was proposed by S. Puangpronpitag and N. Masusai in which IP-MAC binding was statically held in ARP cache of the host and the packets coming from other hosts whose binding was not present in cache was denied. The drawback incurred in the scheme as it was possible only on a small network and imposed an overhead upon the operating system to maintain the entries.

Nayak and Samaddar [6] coined a continuous monitoring scheme implemented in Linux environment using arping command to retain the binding of gateway or host for larger time duration and continuous checking of ARP table using "arp-a" command to find if there is any duplicate binding.

Nam et al. [13] incorporated a long-term memory along with short-term ARP cache to retain the IP-MAX binding for a longer time period and check the ARP cache for any mismatch case. In case of any new host, it used voting mechanism to find the genuinity of new host and accordingly updated the entry in both tables.

Stateful ARP scheme [14] adopted the concept of analyzing the response time of the packets to discriminate between the genuine and non-genuine hosts. The scheme was based on assumption that there exists a difference between the response time of attacker and the legitimate entity. Secure Unicast Address Resolution Protocol (SUARP) called for modification in the existing network infrastructure where DHCP server was configured with additional parameter. It also maintained the IP-MAC binding of the host. Thus, any host need to enquire about the IP for its MAC,

instead of broadcasting sent a unicast Request to DHCP and was responded with the corresponding MAC.

Cryptographic measures included Secure-ARP (SARP) [15] and Ticket-ARP (TARP) [16]. In SARP, a concept of digital signature was used where a public/private key was attached to each ARP packet allotted by the Authoritative Key Distributor (AKD) to authenticate the host. TARP used a token which served as a digital certificate appended with the packet to authenticate the host. Both the proposals required a central authority to distribute the token/key, respectively, and thus faced a single site failure issue. Cryptographic approach required a large number of packets with every host computer requiring cryptographic implementation [17]. Thus these solutions failed the cost-effectiveness.

Kumar and Tapaswi [3] used a centralized detection scheme in which a central server detected and validated the binding using ARP/ICMP probe packets. Here host and the central server both maintained a long-term cache memory to hold and validate the binding. Jinhua and Kejian [8] also used an ICMP mechanism to validate the binding but maintained a database holding the binding to validate the binding thus faced a single site failure issue. Pandey [4] used ICMP probe packets to validate the new binding sending 2 ping packets; one to the new binding and other to the source host itself with expecting response from a single host only.

Tripathi and Mehtre [1] maintained a secondary table to verify the binding using ICMP packets. They used an entering and existing algorithm to successfully validate the binding. The ICMP echo packet was sent to the previous binding to check the aliveness of the host. The scheme was successful in providing a solution to IP exhaustion problem also.

Arote and Arya [7] involved a voting parameter in centralized detection scheme where the new entering hosts voted to attain the IP-MAC of central server. This central server was responsible in validating the IP-MAC binding using probing mechanism. Certain passive detection tools were proposed such as ARP watch, ARP Guard [2] but were dependent upon attacker's arrival time and were only able to detect attack and raise alarm. They failed in prevention of the attack.

### 3 Proposed Mechanism

This scheme uses a distributed concept, in which each host maintains a secondary table which is stored in form of text file storing the <IP, MAC> bindings of host [1]. The purpose of using the file is to permanently store the binding. Once validated with the record held in secondary table, the primary table is updated along with secondary table. The concept of the scheme is based on concept put forward by Tripathi et al. [1] and Pandey [4]. The detection host first looks into its primary cache, on receiving an ARP Reply to find the binding. If it is obtained without any mismatch in IP or MAC, the entry gets updated again to reside for next timeout session.

The validation phase incurs the use of ICMP and ARP packets. There can be two cases: one the host sends a Request/Reply and other host receives a Request/Reply.

If host wants to send an ARP Request, it simply broadcasts it to all the hosts in the network, comprising of its IP address, MAC address and the IP of the destination to be resolved. Similarly, for sending a Reply, if the Request is intended to it; it sends a unicast message to the host in response to the request proving its MAC else drops the packet. The broadcast ARP Request is answered by only the host who's IP matches the destination IP header whereas others will discard it.

In case a host receives an ARP request, it checks the header destination IP, i.e., if that request is made for it. If it is the intended host, it sends a Unicast ARP reply to the sending host telling its MAC address. For an ARP Reply received, let be <IPX, MACX>, host checks if IPX exists in short-term cache; if found, then the corresponding MAC entry is checked. If the binding is found to be same then the primary cache is updated as well as secondary table. If not, then the secondary table is checked. Now there can be three cases in this. First is that the same binding is found in secondary cache. Second, no binding is found in secondary table and the third that there appears to be a mismatch of IP or MAC in the received binding. Figure 1 represents the flow of mechanism with three case conditions. If the binding is not found in primary cache mean, the entry would have expired after the timeout, thus secondary table is scanned for the entry. If the binding is found and found to be same as stored in secondary cache, both local ARP primary cache and secondary table is updated for next time out.

In case the entry is unavailable in Secondary table too, there is the possibility of the new host. Now the aim of the detection host is to determine all the possible hosts who are claiming to hold the IP. This could uncover all the claimants in the LAN. If there is only a single claiming host, the entry is accepted. The host sends 50–60 ARP Request packets at a time interval of 10–50 ms to obtain reply of any one, preventing any flooding attack, i.e., if the attacker aims at suppressing the response of the legitimate host by sending number of spoofed packet, then it can be failed and the detection host can received at least one Reply from the genuine host. If the reply is received from more than one host, there is the chance that there exists a malicious host who is pretending to be a legitimate host and using the IP of it and the ARP Reply is sent by it to poison the ARP cache. Now the detecting host makes use of 2nd-phase validation to distinguish between the genuine and non-genuine host. For 2nd-phase validation, ICMP ping packets are used. The ICMP echo reply for ping request confirms the identity of the host. If reply is received from more than one host, then send ICMP probe packets to each host from whom reply is received. If the reply is received, accept the binding else discard the entry from local cache. In case ARP reply is received from only one host, the binding is accepted and updated in both primary and secondary table. This scheme provides a two-phase validation for the new host as the attacker cannot falsify both ARP and ICMP echo packet together represented in Fig. 2.

In case of any mismatch found that is mismatch of either IP or MAC when searched in secondary table, an ICMP probe packet is sent to both new and old binding. 2 ICMP probe packets are sent in order to find aliveness of previous host along with finding the legality of new binding received. If only previous host respond ensures the previous host's aliveness in the network then the new binding is removed from primary cache.

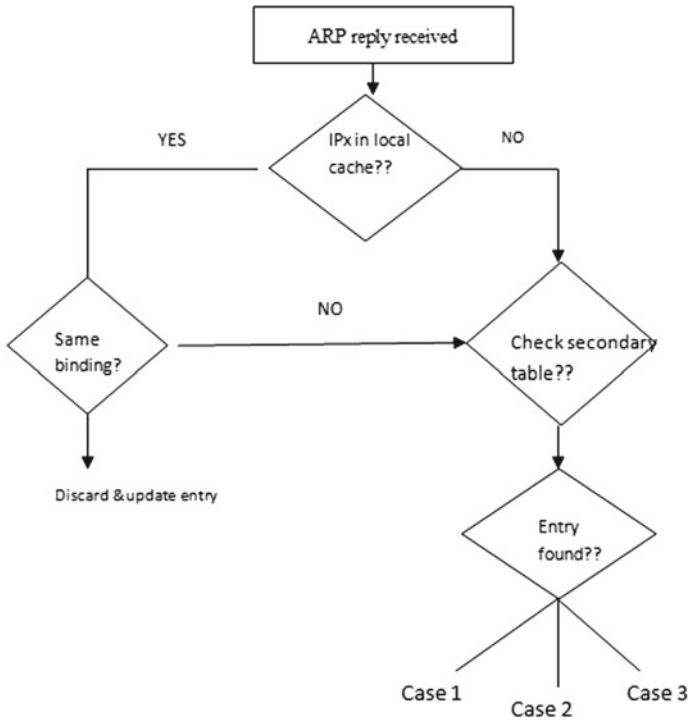


Fig. 1 Binding check in primary table

For no reply from previous binding and only receiving reply from new host means that the IP is allotted to a new host (in case of MAC mismatch) or host is provided a new IP (in case of MAC mismatch). In such case the previous entry is deleted, and new entry is updated in secondary table. For case of response from both hosts, the entry is considered to be spurious and thus removed from the primary cache. Else if no response is received means the previous host has gone offline and the fake binding is sent to the host. Thus the previous binding is removed from secondary table and new entry is removed from primary cache. The mechanism is represented in Fig. 3.

The procedure can be discussed as under:

- If a frame has been received:
- If it is a Request: Check if it is intended to the host
- If it is: Send a Unicast Reply to the querying host
- Else: Drop the packet

If it is a Reply: Check the primary ARP cache:

- If entry found: check if there is a case of mismatch
- If same: accept the binding and update

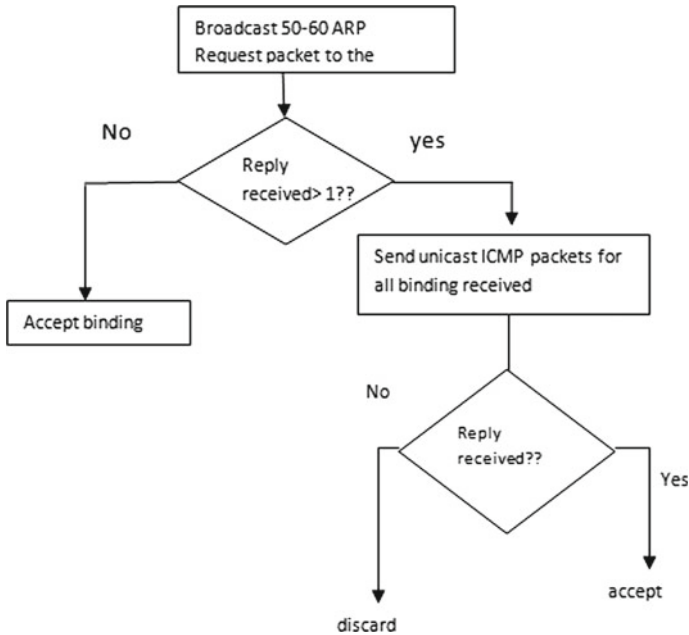


Fig. 2 2-phase validation for new binding

If entry not found in primary cache or there is a case of mismatch in primary ARP table:

If entry found in secondary cache with same <IP, MAC> binding: accept and update primary and secondary table

If entry found but there is a mismatch: Send ICMP ping packet to previous and the new binding

If older binding only responded: drop the new binding and remove the entry from primary ARP table. Raise alarm

If new binding only responded: then remove the associated entry from the secondary table and accept the binding in primary

If none responded: drop the packet

If both responded: drop the new binding and raise the alarm

If entry not found in secondary table: Send 50–60 ARP packets for that binding

If (Reply == 1): accept binding

Else (Reply > 1): send ICMP probe packet to each of the claiming host

If Reply received for the entry and update primary and secondary table accordingly

Else: Drop the binding

Else: discard the binding and remove entry from primary cache

If the frame is to be sent:

If it is a Request: Broadcast to all the hosts in the LAN

If it is a Reply: Send Unicast reply to the host which sent the Request.

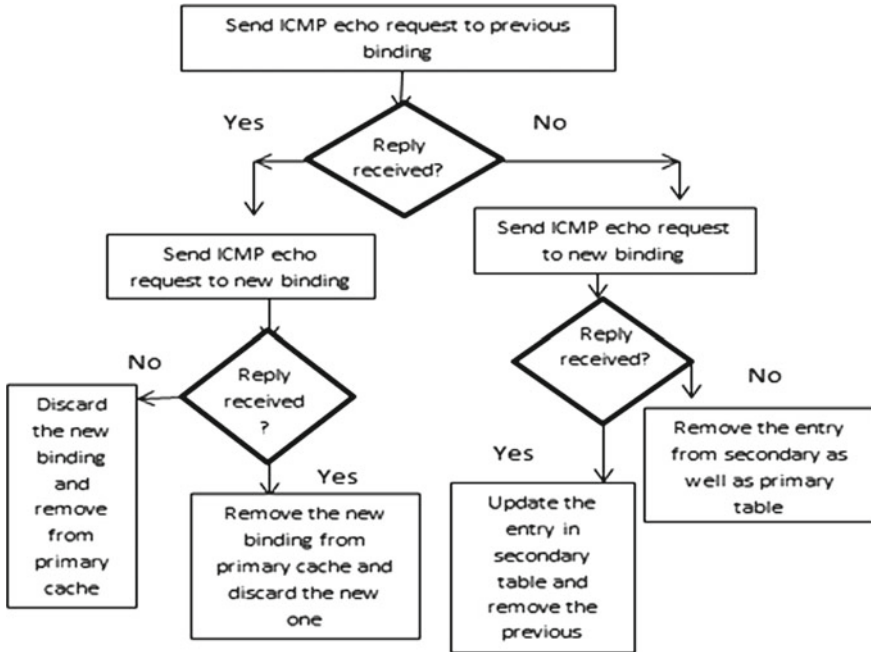


Fig. 3 Case of mismatch in binding

### 4 Simulation and Analysis

The experimentation setup requirement incorporates the use of tools such as Ettercap, PackEth and Wireshark. Ettercap is an easily available tool to perform ARP poisoning attack by selecting the target hosts of the network. Wireshark runs on both the victim and attacker end to analyze the packet that moves in the network. PackEth tool is a packet generating tool that has been used here to generate and send fake ARP Reply message. The entire simulation is carried on “Ubuntu” platform and the system is installed with python.

The simulation of the above-defined procedure is carried out on a smaller scale in python using “scapy”. To poison the cache fake ARP reply is sent using PackEth tool with IP “192.168.10.4”. The attacker has IP “192.168.10.2” which during experimentation was a part of network whose binding was held in secondary table maintained at node. When it claimed to possess the IP “192.168.10.4”, an ICMP echo request was sent to its previous IP as it was held in secondary table. It sent a reply for previous binding and not for the new binding. Hence the fake reply was alarmed using the festival package available in Scapy. The binding pre-maintained is represented in Table 1.

**Table 1** Showing the <IP, MAC> binding pre-stored

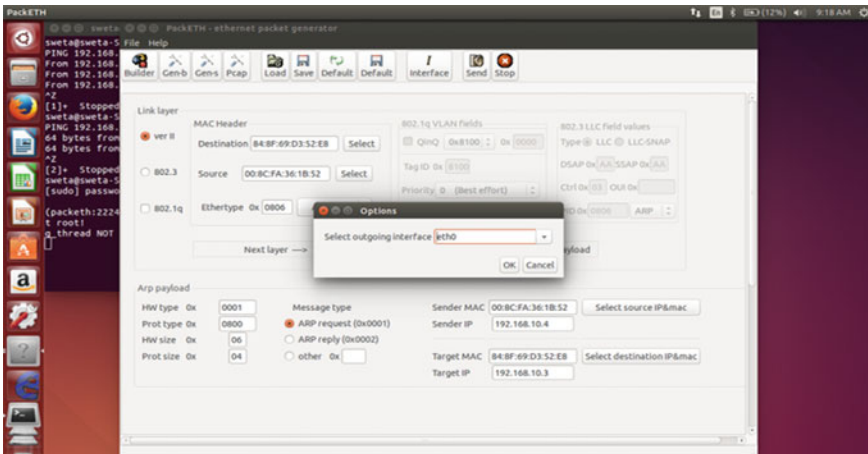
MAC	IP
00:8C:FA:36:1B:52	192.168.10.2
00:1E:68:7C:9A:E2	192.168.10.4
84:8F:69:D3:52:E8	192.168.10.3

An alert together with alarm is generated whenever there is a mismatch case and validation fails. Its performance and experimentation on a larger network is under study and would require much expertise personnel.

The victim is holding IP “192.168.10.3”, and the code is run on victim’s machine. The attacker attaches its MAC “00:8c:fa:36:1b:52” with IP of “192.168.10.4”. “sniff” function is used to “filter” ARP packets which are analyzed to get the packet. Then the packet is read, and its Ethernet address and Network address is obtained.

Next the secondary table is read and for MAC and IP, For MAC corresponding IP is checked and matched, and for IP corresponding MAC is checked. Using packEth tool, a fake ARP packet is sent, in which the host uses IP of another host in the network represented in Fig. 4.

While implementation, the IP of packet received is read and the MAC and IP obtained in packet is printed. Then the corresponding MAC is checked in the secondary table and its corresponding entry obtained is printed. ICMP probe packet is sent to older binding and since is found to be in network has responded but the new binding produced no reply as the IP and MAC belong to two different hosts and ICMP echo packets are only replied by host when its both IP and MAC matches. The code is run the result seen is viewed in Fig. 5.



**Fig. 4** PackEth tool



```

Sent 1 packets, received 1 packets. 100.0% hits.
ARP Reply received
00:8c:fa:36:1b:52
192.168.10.4
MAC entry found in secondary table
192.168.10.2
Sending probe ARP Request for older mapping
RECV 1: Ether / ARP is at 00:8c:fa:36:1b:52 says 192.168.10.2 / Padding

Sent 1 packets, received 1 packets. 100.0% hits.
Received probe ARP reply for older mapping
MAC address 00:8c:fa:36:1b:52 is causing IP Exhaustion Problem
High Performance MPEG 1.0/2.0/2.5 Audio Player for Layers 1, 2 and 3
  version 1.16.0; written and copyright by Michael Hipp and others
  free software (LGPL) without any warranty but with best wishes

Playing MPEG stream 1 of 1: alarm.mp3 ...

MPEG 1.0 layer III, 128 kbit/s, 48000 Hz joint-stereo
^Z
[2]+  Stopped                  python attack.py
root@sweta-Inspiron-5420: /home/sweta/Desktop#

```

Fig. 5 Code executed and result seen

## 5 Conclusion and Future Work

With discussion to the mechanism proposed in this dissertation, an attempt is made to prevent and mitigate ARP poisoning. By preventing attacker from poisoning, the host's cache with fake binding can easily deal with other attacks such as MITM, session hijacking, host impersonation and others. The scheme could be a possible approach for ARP poisoning. Using secondary table, the binding can be maintained permanently and by using two ICMP probe packets send both to previous, and new host can easily fail the attempt of attacker. Even for new bindings, ARP request is sent so that if the host is alive in the network it can respond as well to determine all the hosts which claim to be holding that IP address. Even if it is not alive and attacker sends a spoofed reply to it, a second validation is conducted using ICMP probe packets. This scheme is an asynchronous approach which does not require any periodic monitoring. The implementation has been conducted on a smaller scale to provide the result, conducting at higher level will require more specifications and requirements with expertise personnel.

The aim is at further expanding our work by bringing about some modifications at few sites. At the validation phase or validating the new binding, a voting phase can be adopted such that to accept the new binding only if more than half nodes in the network vote to hold that binding or agree to have that binding. The focus will be at usage of distributed concept in which all the hosts will contribute to authenticate a new host in the network. Voting and probing mechanism can be integrated together contributing to propose a hybrid model. In future, we will work at expanding the

network scope and considering all the scenario of LAN environment together with the possibilities of theft.

## References

1. Tripathi, N., Mehtre, B.M.: An ICMP based secondary cache approach for the detection and prevention of ARP poisoning. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICIC), pp. 1–6. IEEE (2013)
2. Tripathi, N., Mehtre, B.M.: Analysis of various ARP poisoning mitigation techniques: a comparison. In: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 125–132. IEEE (2014)
3. Kumar, S., Tapaswi, S.: A centralized detection and prevention technique against ARP poisoning. In: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), pp. 259–264. IEEE (2012)
4. Pandey, P.: Prevention of ARP spoofing: a probe packet based technique. In: 2013 IEEE 3rd International Advance Computing Conference (IACC), pp. 147–153. IEEE (2013)
5. Jennings, F.: Beware the enemy within. *SC Magazine*. Jul. 2008: Business Source Complete. Web. 25 June. 2011 (2008)
6. Nayak, G.N., Samaddar, S.G.: Different flavours of man-in-the-middle attack, consequences and feasible solutions. In: 2010 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 5, pp. 491–495. IEEE (2013)
7. Arote, P., Arya, K.V.: Detection and prevention against ARP poisoning attack using modified ICMP and voting. In: 2015 International Conference on Computational Intelligence and Networks (CINE), pp. 136–141. IEEE (2015)
8. Jinhua, G., Kejian, X.: ARP spoofing detection algorithm using ICMP protocol. In: 2013 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6. IEEE (2013)
9. Salim, H., Li, Z., Tu, H., Guo, Z.: Preventing ARP spoofing attacks through gratuitous decision packet. In: 2012 11th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), pp. 295–300. IEEE (2012)
10. Tripunitara, M.V., Dutta, P.: A middleware approach to asynchronous and backward compatible detection and prevention of ARP cache poisoning. In: Proceedings of 15th Annual Computer Security Applications Conference (ACSAC 1999), pp. 303–309. IEEE (1999)
11. Abad, C.L., Bonilla, R.I.: An analysis on the schemes for detecting and preventing ARP cache poisoning attacks. In: 27th International Conference on Distributed Computing Systems Workshops, 2007. ICDCSW'07, pp. 60–60. IEEE (2007)
12. Puangpronpitag, S., Masusai, N.: An efficient and feasible solution to ARP Spoof problem. In: 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2009. ECTI-CON 2009, vol. 2, pp. 910–913. IEEE (2009)
13. Nam, S.Y., Kim, D., Kim, J.: Enhanced ARP: preventing ARP poisoning-based man-in-the-middle attacks. *IEEE Commun. Lett.* **14**(2), 187–189 (2010)
14. Wang, Z., Zhou, Y.: Monitoring ARP attack using responding time and state ARP cache. In: The 6th International Symposium on Neural Networks (ISNN 2009), pp. 701–709. Springer, Berlin (2009)
15. Bruschi, D., Ornaghi, A., Rosti, E.: S-ARP: a secure address resolution protocol. In: 2003. Proceedings of 19th Annual Computer Security Applications Conference, pp. 66–74. IEEE (2003)

16. Lootah, W., Enck, W., McDaniel, P.: TARP: Ticket-based address resolution protocol. *Comput. Netw.* **51**(15), 4322–4337 (2007)
17. Goyal, V., Tripathy, R.: An efficient solution to the ARP cache poisoning problem. In: *Information Security and Privacy*, pp. 141–161. Springer, Berlin (2005)

# Software-Defined Networks and Methods to Mitigate Attacks on the Network



Shubham Kumar, Sumit Kumar and Valluri Sarimela

**Abstract** Software-defined network (SDN) is becoming an advance technology. It is not only used to manage IP networks but also manages data centers as well as cloud data and it can be applied in various types of networks. Earlier approaches for IP networks were more complex and IP networks are now a big network; thus, it is very difficult to manage those networks in terms of configuring the network devices, applying policies on the network dynamically and get the knowledge of the faults, load and changes in the network. Software-defined approach made it easy to manage and configure the network. The role of the SDN controller in network devices can be extended with an application that effectively solves a particular problem and provide a flexible management service. One of the protocols used for this technology is OpenFlow. It basically works on southbound interface, i.e., between controller and network devices. Many solutions to utilize the network and exploit as much information possible from the network is one of the aim of researchers and many solutions have been proposed for the same. One of the most important and distinct features is to detect denial-of-service (DoS) attack quickly and precisely. In this paper, we are going to give an introduction about how and why SDN is trending and also analysis of solutions to detect and save a network from DDoS attacks.

**Keywords** Data plane · DoS · Northbound interface · OpenFlow  
Software-defined networks · Southbound interface etc.

---

S. Kumar (✉) · S. Kumar · V. Sarimela  
Central Electronics Laboratory, Bharat Electronics Limited,  
Bengaluru 560013, Karnataka, India  
e-mail: shubhamkumar@bel.co.in

S. Kumar  
e-mail: sumitkumar@bel.co.in

V. Sarimela  
e-mail: vallurisarimela@bel.co.in

## 1 Introduction

SDN is a network architecture in which the control plane is separated from data plane and also managed by the control plane. IP networks are difficult to manage [1] as they are very complex and the structure of the networks change dynamically. SDN decouples the network device from the control plane and the controller becomes external entity, known as SDN controller, NOS, POX, and with various their names. Open Networking Foundation defines SDN as a technology which provides physical separation of the (forwarding) or “Data Plane from the control (intelligent) plane, and where a control plane can control various other devices” [2]. SDN is considered as brain of a network. There are various advantages of this approach; one of them is the ease of programmability and other is the access to more network information. The applications developed can have more network information for the decision making as the global policies; topology and policy decisions are available at the controller. It controls the networks, applies policies on the network, and provides a way of authenticated access into the network. SDN allows open revolution at control plane by providing a programmable infrastructure in the network for dynamic flow table rules. But with more dynamic nature and ease of programmability, new security-related problems and threats are introduced in the network. The applications can take actions from any part of network and there is no need of hard coding a policy or rule, and they can be changed by the controller administrator dynamically depending on the network requirement and capability. Also, the maintenance and integration of the applications and policies will be simple and available at one place. Any new packet coming into the network is first reported to the controller and then decision is taken what to do with the packet.

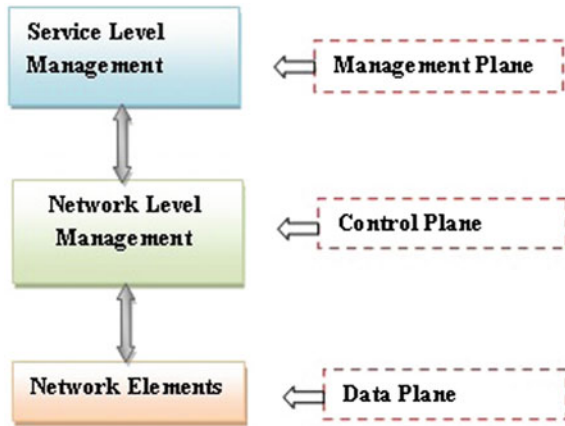
The threat vectors and possible solutions with respect to SDN environment are discussed in [3]. The network devices are connected to the controller and will be informing the controller about the network, and at the same time, if SDN controller wants to get some data from the network, it can get from the devices. If the connection between the controller and switches breaks or somehow controller is unavailable at the instant, then the data plane devices run with the latest configurations/instructions they have received from the controller.

## 2 Terminologies

There are different terms that are defined and will be used in this paper frequently. These basic terms tells the different parts/modules of a basic SDN network [4]. An architecture level diagram is shown in Fig. 1.

1. Data plane is the plane where the network devices work. These devices are also known as data forwarding devices. They just process the packets and inform the control plane about the packets. The data plane receives instructions and policies from the control plane. For different network devices, rules for the

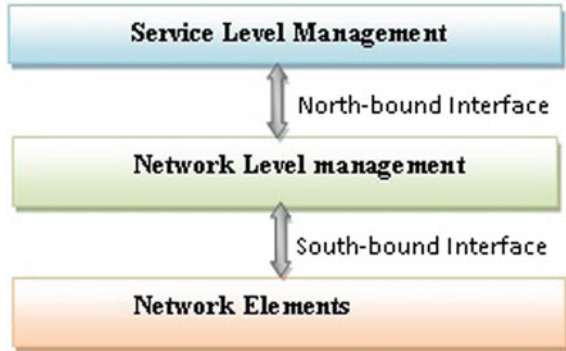
Fig. 1 SDN architecture



devices are different depending on requirement or the manner in which user wants to configure the network. The southbound interface is used for communication between control plane and data plane. One of the standard protocol is OpenFlow, which is discussed in this paper, and is used by many networking companies as well as many open-source SDN controllers which are based on the same protocol.

2. Control plane is the plane which is earlier used to be a part of the network device but now is separated from the devices. It is the brain of the network. The SDN controller sits on network plane, and all the policies and decisions are made at control plane and communicated to data plane through southbound protocol. Also, the network devices send the packets to the control plane if they have no policy of what to do with the packet.
3. *Management plane* is the plane which keeps a number of application set that runs over a computer. The management plane communicates the network policies to the control plane through northbound interface. It is nothing but a set of user interface applications through which user controls the network. The APIs interact with the controller are control plane and control plane applies those policies on the network devices.
4. *Southbound Interface* is the interface which communicates between data plane and control plane through instruction set APIs. OpenFlow protocol is an example of a southbound protocol which communicates between control plane and data plane. There are other protocols that can serve as southbound protocol like LISP, NETCONF, and OVSDB. Depending on the requirement and the level of accessibility and control on the networking devices, any of the above-mentioned protocols can be selected. But OpenFlow is most popular and widely used across globe by leading networking companies.
5. *Northbound Interface* is the interface through which the management plane communicates to the control lane. The APIs present at the management plane pass the network configurations, policies, decisions to the control plane or SDN controller. There is no defined standard protocol for this interface as there is Open-

Fig. 2 SDN interfaces



Flow for southbound interface. The exploration for this and freedom is given to the user. Java is a very helpful tool for a northbound interface development. The user can make applications on Java and through those configurable APIs; the information can be passed to the control plane.

The interfaces mentioned in the SDN architecture are shown in Fig. 2.

### 3 Denial-of-Service Attacks

In general, flooding attacks, denial-of-service (DoS) attacks, and distributed denial-of-service (DDoS) [5] attacks are the main methods to destroy availability of the server or the network to which one is connected to. DoS attacks or DDoS attack is an attempt to make the resource/machine/server/unavailable for service to its intended users. DoS attacks are sent by one system or person where as DDoS attacks are sent by two or more systems or person in order to choke the network. Thus, DoS attacks are one of the sub-types of DDoS attacks. So in this paper, both are indicated under a common name as DDoS attacks.

Most of the attacks are generally TCP packets attacks which are about 95% and rest attacks are by RST packets, ICMP packets, or other packets. SYN-ACK packets, RST packets, and ICMP error messages are communicated back and forth between the victim and attacker indicates that there is an attack. Out of which, SYN-ACK packets clearly indicate that there is an flooding attack but rest packet communication does not clearly indicate if there is an attack or not. Smurf-based attacks are detected, and with packet filtering, it can be prevented with subnet-directed broadcasting addresses. The performance of detection and filtering approach rely on how effective filtering and detection stages are. Its effectiveness can be measured on the basis of two ratios. One is out of total normal packets received, number of packets classified as attack packets that are confirmed as normal packets. This is known as false positive ratio. False negative ratio is just opposite of that. Thus, for an excellent filtering and detection

system, both ratios should be minimum. Now more and more business operation are going online; thus, these attacks are becoming more and more common causing a significant financial loss [6] to business and security. And reports show that these attacks are becoming more and more common these days [7]. Distributed denial of-service (DDoS) attacks makes on-line services unavailable for the users by flooding the packets into the network and thus chocking the network.

Sometimes, this can be a coordinated attack by a number of attackers distributed over the network, and sometimes, it may be an uncoordinated attack. The longer it goes the longer the services will be unavailable and more and more loss to the business. Thus, timely detection and mitigation are very much of importance by analyzing the traffic. With the introduction of SDN technology, it becomes easier to find the network insights into one place. Thus, SDN is an ideal platform for DDoS detection. As soon as the DDoS attacks are detected, counteractions can be taken to save the network or counter policies can be made and introduce into the network. It can be like a proactive method of deadlock detection and prevention method where it predicts that whether the current network conditions can lead to a deadlock, and if yes then what needed to be done. Thus, some of the ways by which attackers attack the network and some of the techniques to detect and mitigate the DDoS attacks are described in the paper.

## 4 OpenFlow Protocol

Southbound interface: OpenFlow protocol is an example of a southbound protocol which communicates between control plane and data plane, hence provides a proper management to deal with DDoS attacks. In the following paragraph, we have describe the OpenFlow protocol.

### 4.1 Introduction to OpenFlow

There are many packet types mentioned in OpenFlow protocol and are increasing in the updated version of the protocol as per the requirement of the network and exploration by researchers. The below part describes the packet description and flow for how a new packet is treated and which protocol packets are involved in creating new flow in the switch table [8]:

**OFPT\_PACKET\_IN** This packet is sent from the switch to controller with the reason specified.

1. TABLE-MISS: If no flow is matched (table-miss flow entry)
2. APPLY-ACTION: if action is output to controller (apply actions)
3. INVALID-TTL: if TTL is invalid in the packet
4. ACTION-SET: if action is output to controller (action set)



5. GROUP: if action is output to controller (group bucket)
6. PACKET-OUT: if action is output to controller (packet out)

The switch sends an OFPT\_PACKET\_IN to the controller when it does not know what to do with the received packet. As per OpenFlow switch 1.5.1 RFC, the above conditions are mentioned.

**OFPT\_FLOW\_MOD** It is send by the controller to the switch, whenever controller needs to do anything with the flow. It sends a command and the switch do the same with that packet. Entry table with a reason is as mentioned below:

1. ADD: a message to create a new flow in the flow table of the switch.
2. MODIFY: to modify all the flows matching.
3. DELETE: to delete matching flows.
4. MODIFY-STRICT: to modify an entry which strictly matches the priority and wildcards.
5. DELETE-STRICT: to delete an entry which strictly matches the priority and wildcards.

So whenever a new packet comes to the switch and it has no flow entry for that packet, then it sends an OFPT\_PACKET\_IN message with OFPR\_TABLE\_MISS reason. In response to that the controller can send an OFPT\_FLOW\_MOD message with OFFFC\_ADD reason indicating to add a flow entry for that packet. In the same way, controller can send message if it want to modify an entry or delete an entry with one of the above reasons specified. The message is sent by the controller to the network device, i.e., switch, router, and they will take the actions on the entry like delete, add, or modify from their flow tables. If no entry is found against a new packet and controller also do not add any entry in the switch, then the packet is dropped. If the switch does not want to allow the packet to pass through, then the controller sends a packet to switch with the reason as drop packet. Then that packet is dropped always. Even if controller takes no decision about that packet, the packet gets dropped at switch.

Every new packet, about which switch has no entry what to do with, is sent to controller. This feature of SDN is sometimes taken advantage by the hackers as they keep on sending the data from different sources at a very fast rate and the switch keeps on sending that data to controller. Thus, the most bandwidth between switch and controller is occupied by unwanted traffic. This is one of the problems that is faced and is like a loophole in the protocol itself. Current DDoS attacks have various forms, e.g., disruption of configuration information, consumption of computational resources. Depending on the types of DDoS attacks, there are different detection and mitigation methods. There are various forms of DDoS attacks and there are various solutions proposed by different researchers. A detailed study of some of those attack is explained below.

1. How the SDN controller is attacked and what happens afterward?
2. DDoS detection in SDN-based environment?

## 5 DDoS in SDN

OpenFlow protocols works between control and forwarding plane. It checks incoming packet and then matches it in its entry into the flow tables, and if a valid entry is found, then it sends and takes the actions on the packet like forwarding, dropping, broadcasting. But if no entry is found, then it sends it to the controller using PACKET\_IN type of packet and with appropriate reason like no entry found, invalid TTL. Thus if such unknown packets are send by attacker and a number of packets are sent continuously, then there is processing delay for the valid packets coming from genuine users and when no entry is made for them, and then those packets will get dropped and the user will not be able to avail the services. Another way is if the attacker sends a lot of packets with different details which appears to be authentic, but after sometimes there will be a lot of entries in flow table and no new valid entries will be filled. Thus, such kind of attacks are considered as DDoS attacks on the SDN controller, and since controller is the decision maker for the data plane devices, they will be unable to serve the purpose to the valid users or traffic. A DDoS attack tolerant system has some essential features associated with redundancy, diversity and independence, these features are easier to implement in SDN based network than traditional networks. Some of the DDoS or attack detection techniques related to SDN are discussed by many researchers. Giotis et al. [9] give anomaly detection using OpenFlow and sFlow on SDN. Other problems like overloading problems due to any internal factors or DDoS are discussed in [10]. In this paper, we are going to describe various approaches and methodologies to detect and mitigate the DDoS attacks.

1. A method of detecting and mitigate DDoS attack using a blocking application.
2. A method of detecting and mitigate DDoS attack using the IP address filtering.
3. A method of mitigate DDoS attack by load balancing.
4. Some other methods and approaches.

### ***5.1 A Method of Detecting and Mitigate DDoS Attack Using a Blocking Application***

A SDN-Oriented DDoS Blocking Scheme for Botnet-Based Attacks: A controller is connected to a number of devices which are distributed in the network. Also, there will be some botnets which are present in the network and will try to block the network. Since the SDN controller has the insights of the network and knows all information about all the devices connected to it, it has the knowledge of the flow tables on each device and the policies implemented on different devices. Thus, one of the solutions to detect and mitigate the effects of the DDoS attacks is the use of a blocking application, which runs on the controller when controller suggests a probable DDoS attack is chocking the network. Whenever a new packets comes to a switch, packet is sent to controller and controller replies the switch to introduce a new

flow in the flow table; thus, the attacker will be sending a large number of packets, and thus, the flow entry table into the switch keeps on increasing. If the controller sees this possible DDoS attack, then the controller notifies the DDoS blocking application for providing a redirected address of server through secure channel.

The blocking application keeps a pool of IP addresses which are used to redirect the traffic to another IP address. Since IPv6 is capable of handling huge IP addresses, then it is easy to find a usable address under the same prefix. The server, when indicated that it is under attack, moves its services logically from the address which is under attack to a new address which is safer to operate. And the genuine users are made to move to the new address for accessing the services. And botnets do not use IP address spoofing so they may be attacking the same IP address and the packets at that port are getting dropped without affecting the network system. Some related work on non-IP spoofing is explained in [11].

## ***5.2 A Method of Detecting and Mitigate DDoS Attack Using the IP Address Filtering***

SDN provides flexibility to network administrator to install flow rules in the controlled switches, if these switches have the capability of ternary content-addressable memory (TCAM) [12]. There are various advantages for using TCAM as it has fast lookup memory and speed, but at the same time TCAM is power hungry and costly.

With the enhancement in technology, the types of DDoS attacks have been changed and there are new improved methods for detection of DDoS have been developed over time [13]. There are different forms of DDoS attacks and different detection methods. When multiple attackers send packets to a particular machine or device and utilize network bandwidth available for the victim and choke its access to network, this approach focuses on detection of those kind of DDoS attacks. The very first aim is to detect the victim quickly and correctly. The idea to identify DDoS attack on a victim is to monitor the flow rate and monitor the flow rate asymmetry. For any potential victim, one needs to monitor the total traffic coming to its IP and total traffic going out from its IP. But TCAM size is very limited, and for a number of IP addresses, one cannot monitor the flow rate for all the IPs. To handle this, divide network in a set of IP ranges and make a rule to monitor the incoming and outgoing traffic for those ranges of IP addresses. When for a particular range if the flow asymmetry increases the threshold or it appeared that it is under DDoS attack, then divide that particular IP range into smaller ranges and try to find out the potential victim. It may be possible that due to limited TCAM size, it is difficult to find out the particular victim but at least one ends up with a small range of IP containing potential victim. Thus, then network admin finds the attacker IP addresses and install rules in OpenFlow switches to drop packets from attackers to the victim. Thus in this way, network resources can be saved and the victim user can avail the network normally. In this way, the decisions or actions for every packet can be taken by matching the flow entries.

Only the controller can introduce or remove a flow entry from flow table, and only controller can associate one or more actions to a packet.

### 5.3 *A Method of Detecting DDoS Attack by Load Balancing*

Load Balancing for Software-Defined Networks against Denial-of-Service attacks: Load balancing is the term used for equal distribution of load and maximum utilization of available resources. Load balancing can be implemented at L4 layer or L7 layer or both. In some load balancing solutions, the algorithm to detect DDoS attacks uses only destination and source IP addresses.

**L7 Load Balancing** This will be defined in the following manner, i.e., statically balancing the load and dynamically balancing the load. Static load balancing is easy to implement but there is possibility that under some conditions, the balancing will be inefficient. In dynamic load balancing, the load is distributed among various servers at the runtime. This is difficult to implement. Under this, the balancers monitor the load on each server, and whenever the load reaches a maximum specified limit, they start applying balancing algorithms [14]. More complex algorithms are compared in [15]. One also includes implementation of Honeybee Foraging Algorithm.

**L4 Balancing** maintains the balance between network devices and the networks equipments or the end systems. Unlike L7 balancing where the load is distributed among various servers, here the load distribution is done among different switches between controllers. Thus, the two load balancing schemes are totally independent. L7 load balancing is about routing the traffic among servers and L4 balancing route packets among various paths in a network. L7 load balancing does independent of the network which lies between the sender and server. Thus through SDN controller, one can balance the loads by changing the flow tables or routes into the network devices at different levels.

### 5.4 *Other DDOS Attacks*

Generally, there are two modes by which we can install flow rules in the switches: proactive mode and reactive mode. In proactive mode, flow rules are installed during the network bootstrap, while in reactive mode, flow rules are installed when switches explicitly request them. However, in reactive modes, the installations of flow rules are prone of vulnerability to denial of services (DOS) for SDN controller and switches. The attacker can send a large number of requests to the controller that can lead to many consequences like:

1. The software components of switches become overloaded.
2. The controllers resource becomes saturated.

All above consequence can affect the performance of a network that can badly affect the controller, switches, or both.

**Switch Software Overloading** Switch sends a request to controller when it found the table\_miss (no entry in flow table). Since switches run on low CPU, they can generate less number of requests of flow per unit time. Thus, switches can be overloaded; as a result, flow from the genuine user may be dropped and delayed.

To mitigate above problem, we can use the soft switches, i.e., Open vSwitch. Since soft switches are dynamic in nature and run on more powerful CPUs, they can handle more number of request per unit time as compare to hardware switches.

1. An entropy-based scheme A beautiful method of detection of DDoS using the entropy is described in [16] which takes into account the total number of packets and the probability for each type of packet header. This method avoids the false negatives and false positives in the network. It is based on the fact that when an attack has happened, the packets from same source IP address will be arriving to a destination, and thus, the entropy will decrease as less number of different packets from other sources will be appearing. In this way when the entropy falls beyond a threshold, it is assumed that the destination host is under attack.

2. The controller's resource becomes saturated

If the attacker sends large number of requests to the controller, they will consume controller's resource (bandwidth, memory, CPU, etc.) for rule installation and computation. If this is not taken care properly, the controller's resource can be flooded by the requests that can lead to network-wide consequences, as the switches connected to the controller will be affected.

To overcome the resource saturation problem, the controller can be share among the switches in a fair manner by introducing the multi-layer fair queue (MLFQ). In MLFQ, the idea is to maintain the queues in controller at multiple layer: queue of group of switches, queue of per switches, queue of per port. Initially, each queue refer to group of switches, but when size of the queue exceeds a threshold, it should dynamically expand into per switch queue. If again the size of queue of per switch exceeds a threshold, it further dynamically expands per port queue. Finally, controller will come up with multi-layer queue and the layer of the queue will depend on the threshold and number of switches.

## 6 Conclusion

This paper gives a brief introduction of SDN along with the most popular southbound interface protocol "OpenFlow." Programmability of the network can be improved by using SDN technology. It provides support for remote controlling of the network along with various network functions. With the improvement in technology, the security challenges also increase. In this paper, we have described about DoS and DDoS attacks and we have primarily focused on different types on DoS and DDoS attacks in SDN environment and methods for identification of those attacks and

how to save a network from those attacks by using a SDN controller. We have presented functionalities of OpenFlow protocol and its use in SDN technology which are effective in preventing DDoS attacks. However, how to make full use of SDN advantages and defend against DDoS attacks, at the same time how to defend the vulnerabilities present in SDN architecture against DDoS attacks are urgent problems for which research is going on.

**Acknowledgements** We would like to express our gratitude to our Principal Scientist, Mr. Manoj Jain, and our Senior Member Research Staff, Mrs. Uma Devi B., who gave us immense opportunity to do this wonderful research on the topic Software-Defined Networking [SDN], which engaged us in doing a lot of researches and helped us in every possible way.

## References

1. Benson, T., Akella, A., Maltz, D.: Unraveling the complexity of network management. In: Proceedings 6th USENIX Symposium Networked Systems Design Implement, pp. 335–348 (2009)
2. Kreutz, B.D., Ramos, F.M.V., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-Defined Networking: a comprehensive survey
3. Mirkovic, J., Reiher, P.: A taxonomy of ddos attack and ddos defense mechanisms. ACM SIGCOMM Comput. Commun. Rev. **34**(2), 39–53 (2004)
4. DDoS Attack Loss. <http://blog.radware.com/security/2013/05/how-much-can-a-ddos-attack-cost-your-business/>
5. Neustar annual ddos attacks and impact report. <https://www.neustar.biz/resources/whitepapers/ddos-protection/2014-annual-ddos-attacks-and-impact-report.pdf>
6. RFC: OpenFlow Switch Specification ver 1.5.1
7. Software Defined networking. <https://www.opennetworking.org/>
8. Kreutz, D., Ramos, F.M.V., Verissimo, P.: Towards secure and dependable software-defined networks. ACM, HotSDN'13, pp. 1–6 (2013)
9. Giotis, K., Argyropoulos, C., Androulidakis, G., Kalogeras, D., Maglaris, V.: Combining OpenFlow and sFlow for an effective and Scalable anomaly detection and mitigation mechanism on SDN environments. J. Comput. Netw. **62**, 122–136 (2014). Elsevier
10. Wang, Y., Zhang, Y., Singh, V., Lumezanu, C., Jiang, G.: NetFuse: short-circuiting traffic surges in the cloud. In: IEEE International Conference on communications, pp. 3514–3518 (2013)
11. Braga, R., Mota, E., Passito, A.: Lightweight DDoS flooding attack detection using NOX/OpenFlow. Proc. IEEE LCN **3**(1), 408–415 (2010)
12. Teamwiki. <http://en.wikipedia.org/wiki/TCAM>
13. Kaushik, V.K., Sharma, H.K., Gopalani, D.: Load balancing in cloud- computing using high level fragmentation of dataset
14. Malik, S.: Dynamic load balancing in a network of workstation, 95.515 Research Report, 19th November, 2000
15. Rajoriya, S.: Load balancing techniques in cloud computing: an overview. Int. J. Sci. Res. **3**(7) (2014)
16. Mousavi, S.M., St-Hilaire, M.: Early detection of DDoS attacks against SDN controllers. In: International Conference on Computing, Networking and Communications, Communications and Information Security Symposium (2015)

# A Fast Image Encryption Technique Using Henon Chaotic Map



Kapil Mishra and Ravi Saharan

**Abstract** Recent advancements in the networking technologies led to increase in network bandwidth, hence allowing transfer of files of large size. A major portion of files being transferred through various networks all over the world consists of multimedia data, particularly digital images. In this project, we propose a new image encryption scheme that may be used for securing the digital images. The scheme uses Henon chaotic map and 128-bit secret key in order to generate the cipher image. Henon chaotic map is a two-dimensional iterated discrete dynamic system that shows chaotic character on specific values of the constants used. Chaotic maps are very sensitive to the initial parameters, i.e., a slight change in the initial conditions drastically changes the overall output generated by the chaotic system. In our scheme, we use Henon chaotic map along with an externally supplied 128-bit secret key is used to encrypt the original image. After encrypting the image, pixel shuffling is performed using a permutation matrix generated using the chaotic map. The algorithm is tested on a standard set of images against various performance metrics like peak signal-to-noise ratio (psnr), entropy, histogram etc. The algorithm was found to be robust against plain-text, statistical attacks, chosen plain-text etc.

**Keywords** Information security · Image encryption · Image security  
Henon chaotic map

## 1 Introduction

Vast improvement had been seen in the field of networking and technologies in the recent years. Such improvements had allowed us to share and transfer large-size media files, high-resolution digital images through various networks all over the

---

K. Mishra · R. Saharan (✉)  
Computer Science and Engineering Department, Central University of Rajasthan, NH-8,  
Bandarsindri, Kishangarh, Ajmer 305817, Rajasthan, India  
e-mail: ravisaharan@curaj.ac.in

K. Mishra  
e-mail: kapilmishra16@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_30](https://doi.org/10.1007/978-981-13-1708-8_30)

world. Digital images being transferred through such networks may be of personal or business nature which would be sensitive enough to require security from the attackers and illegitimate users through the network as they may try to extract the information which can be then leaked or used for wrong purposes. An increase had been seen for such incidents about leaking of personal photos. So digital image security is an increasing concern for the researchers all over the world and hence in recent years, a lot of interest among researchers had been seen in the field of digital image encryption [1–4].

We already have techniques for encrypting data being transferred through the network. Various techniques including DES, IDEA, AES etc. are used for data encryption through a secure network. A general information security model include a sender side encryption module, network for data transfer and a decryption module at receiver's end. The information security is ensured by encrypting and decrypting the data being transferred through the network. Conventional encryption techniques discussed above provide proven security to the data. But due to the distinct characteristics of digital image data as compared to conventional text data, we need to design specialized algorithms in order to encrypt digital images [5, 6]. The distinct characteristics of image data can be named as shown below:

- High Redundancy
- High Correlation
- High error tolerance or less sensitive to error

High redundancy and error tolerance exist in digital images due to the large number of pixels presented in the image and inability of human eye to detect small change in pixel values of the image [7, 8]. High correlation in digital images exist due to 8 immediate neighbors of each pixel while in conventional text data, there exist only two immediate neighbors.

To adapt according to the different characteristics of digital image, we need to develop dedicated encryption algorithm for images. Any such technique should be efficient to deal with peculiar characteristics of images and simple enough to make it easy to implement [9, 10]. Here, it is worth mention that information security is considered to be formed by three pillars that are confidentiality, integrity and availability. Image encryption techniques deal with ensuring the confidentiality of the digital image data by utilizing cryptographic principles and techniques. Cryptography is the branch of science that deals with the methods for secret communication in the presence of third party (which may include other users, attackers etc. in our context).

Cryptographic methods are used to design encryption or decryption algorithms. The main goal of all such methods is to ensure security of data in the presence of untrusted users. Cryptographic methods for encryption are broadly classified on the basis of key distribution policies as shown below:

1. Private key or symmetric key cryptography, and
2. Public key, also known as asymmetric cryptography.

In private key or symmetric key cryptography, same key is used to encrypt and decrypt the data at senders and receivers side respectively (Fig. 3). In such methods,



secure exchange of keys is essential. These kinds of methods are computationally low cost and require less resources. While in public key or asymmetric cryptography, different keys, one for encryption at sender side while other for decryption at receiver side are used. These methods are of computationally high cost and require greater resources. Apart from above-mentioned division, image encryption algorithms are classified based upon nature of techniques used as chaotic techniques that use chaotic maps and non-chaotic techniques that do not use chaos at all.

## 2 Literature Survey

Prior to our work, we performed an exhaustive literature survey; some of the previous works are described briefly in this section. In [1], the proposed method is based upon SCAN language, which generates a large set of unique patterns based upon a small set of predefined patterns. The method shuffles the pixels of the original image based upon the encryption keys generated using the SCAN language. Encryption keys are nothing but patterns generated by a SCAN word. Encryption keys are used in such a way that no pixel is accessed more than once. The SCAN word is a combination of two patterns:

- SCAN pattern
- Partition pattern

A SCAN pattern is further dependent upon four general patterns, each one of which contains eight transformations numbered from 0 to 7. These are: Continuous Raster C, Continuous Diagonal D, Continuous Orthogonal O, and Continuous Spiral S. Each of the partition patterns is dependent on 3 general patterns B, Z and X. Each one of these also depends upon eight transformation patterns. The input image is divided into four subregions. The partition pattern decide which sub region is traversed first by scanning path. Scanning is done for each subregion in a separate manner. Huge no. of possible patterns makes the method resistible against the brute-force attack.

In [2], various chaos-based methods are described. Chaotic maps are very sensitive to the initial conditions and hence are extensively used in cryptographic methods. Various image encryption techniques use chaotic maps. A general approach of a chaotic technique consists of the following phases:

- Pixel Shuffling (Confusion) phase which involves changing location of pixels and hence decorrelation of pixels. After this phase, statistical information like histogram of the image does not change.
- Pixel modification (Diffusion) phase involves modification of pixel values.

Based upon above techniques, chaotic methods are classified into three categories: pixel permutation or transposition techniques, pixel modification only techniques, i.e., only changing values of the pixels, and visual transformation technique that involves both transposition as well as pixel modification operations.

In [3], a survey paper was presented that reviews major chaotic encryption techniques proposed in recent years. Different techniques are discussed along with their respective problems and strength factors with probable application areas of the technique. A hybrid image encryption and authentication technique using hashing and digital signature technique is discussed. Another technique using error correcting codes was also presented. Different algorithms were presented and concluded that each one of them is suitable for different applications. It is also concluded that if the algorithm was not designed properly then the image may be insecure and can be forged.

In [4], authors described an image encryption method based upon Henon chaotic map and w7 cipher using 128-bit external secret key. The method consists of following phases:

- At the first stage, the original image is shuffled using a permutation map generated by the Henon chaotic map.
- At second stage, XOR operation is applied between the earlier generated shuffled image and the cipher image generated using w7 cipher.

Permutation matrix is a matrix that consists of single one in each row and column. It is used for shuffling because it avoids computation cost at the decryption side as its inverse is nothing but its transpose, so it prevents heavy computation which otherwise would be needed at the decryption side. As the algorithm uses both phases viz. pixel shuffling and modification in order to generate the final encrypted image, the algorithm is secure enough to thwart various types of attacks. But since it uses w7 cipher for encryption which is a stream cipher, it needs to generate a huge cipher stream of size  $m * n * 8$  (for an image of size  $m * n$ ) and then reshape operation needs to be performed, the execution time of the algorithm is increased. The schematic model is shown in Fig. 1.

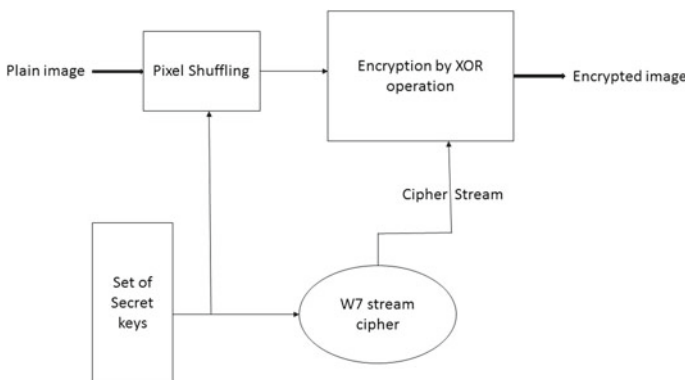


Fig. 1 Image encryption by w7 cipher and Henon chaotic map

### 3 Proposed Algorithm

After performing an exhaustive literature survey of various image encryption technique, we realize a need of an efficient algorithm which is secure as well as is less complex and fast. It is concluded that the algorithm should perform well on various security parameters so that it may sustain various kinds of attacks. In our work, we designed an approach based upon the Henon chaotic map and externally supplied 128-bit secret key.

#### 3.1 Henon Map

Henon map may be stated as a two-dimensional iterated discrete-time dynamical system with a chaotic attractor as proposed by Henon in 1976 [11]. It can be stated by following pair of equations:

$$X_{n+1} = 1 + y_n - \alpha x_n^2 . \tag{1}$$

$$Y_{n+1} = \beta x_n . \tag{2}$$

With  $x_0, y_0$  as initial point,  $(x, y)$  denote the present state of the system. Henon showed that if  $S$  is the area bounded by four points  $(-1.33, 0.42), (1.32, 0.133), (1.245, -0.14)$  and  $(-1.06, -0.5)$ , and if the initial point lies in the area  $S$ , then the subsequent points— $(x_i, y_i)$  for  $i \geq 1$ , also lie in  $S$  [12].

The proposed work generates permutation matrix for shuffling of pixels of the image (confusion phase) and cipher image for encryption of the shuffled image (diffusion phase) using Henon chaotic map and the 128 bit externally supplied secret key. As it is a private key algorithm, we assume the same key to be available at both sender and receiver ends. The architecture of the proposed scheme is shown with the help of a schematic diagram in Fig. 2. The algorithm is described below:

#### 3.2 Encryption Algorithm

To encrypt a given image, following is the algorithm:

1. Take source image of size  $m*n$  as input.
2. Generate random variables from Henon chaotic map by following steps:
  - (a)  $X, Y = \text{Henon}(m*n)$  in order to generate random variables using Henon chaotic map of size equal to no. of pixels.
  - (b)  $X = \text{abs}(\text{floor}(X(1:m*n)*1000000))$ ;
  - (c)  $Y = \text{abs}(\text{floor}(Y(1:m*n)*1000000))$ ;
  - (d)  $X = \text{reshape}(X,m,n)$ ;  $Y = \text{reshape}(Y,m,n)$ ;

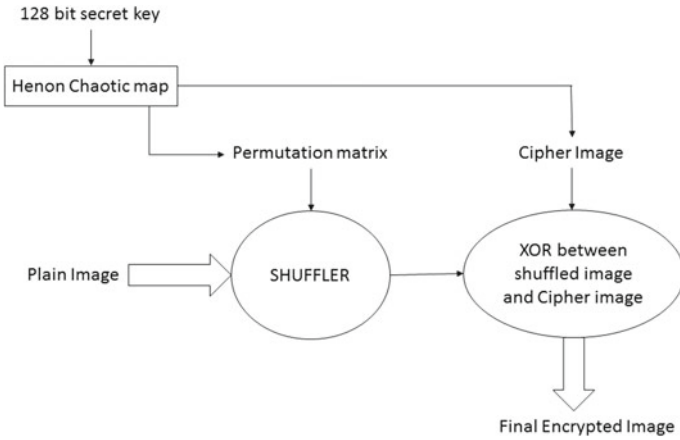


Fig. 2 Architecture of proposed scheme

- (e)  $D = X * Y;$
- (f)  $D = D * \text{sum}(\text{key})$  sum of the digits of 128 bit key.
- 3. Generate permutation matrix  $P(m * n)$  by calculating position for each row  $i$  as
  - (a)  $\text{pos} = \text{mod}(D, n) + 1$
  - (b)  $P(i, \text{pos}) = 1$
  - (c) Other entries being zero for the row  $i$
- 4. Perform combined shuffling operation by first performing vertical shuffling and then horizontal shuffling as shown below: For each  $i, j$  from 1 to  $n$ 
  - (a)  $vI(1:n, j) = P * I(1:n, j)$
  - (b)  $cI(j, 1:n) = vI(j, 1:n) * P$
- 5. For  $i = 1:m$ 
  - (a)  $j = \text{mod}(i, 16) + 1$
  - (b)  $o D(i, 1:m) = \text{bitxor}(D(i, 1:m), okey(1, j))$
- 6.  $D = \text{mod}(D, 255) + 1$ . cipher image
- 7. Generate final encrypted image by applying XOR operation between the shuffled image and the cipher image.

For decrypting, we follow the same algorithm in reverse order as of encryption process just replacing the permutation matrix by its inverse which is nothing but its transpose. All other steps remaining the same make the algorithm very simple to implement it on encryption side as well as decryption side.

## 4 Results and Analysis

### 4.1 Experimental Setup

Proposed algorithm is tested and implemented on machine with following configuration:

Operating system	Windows 7 ultimate 64 bit
Processor	Intel core i5
Memory	2 GB
Software used	Matlab 2015a
Input image	Standard gray scale images of size 512 * 512 are used.

### 4.2 Key Space Analysis

Large key-space is required for an efficient digital image encryption algorithm in order to resist brute-force attack. In our proposed algorithm, we use 128-bit external secret key making the key space  $2^{128}$  and furthermore if we include two seed points of the Henon map as part of secret key, then the key space becomes even larger. If the floating point precision of the machine is  $10^{-14}$ , it makes the key space of the algorithm as large as  $2^{128} \times 10^{14} * 2$  which is enough to resist brute-force attacks.

### 4.3 Key Sensitivity Analysis

An efficient digital image encryption algorithm needs to be highly key sensitive. The algorithm must give a totally different output even after a slight change of one bit in the security key. In the proposed algorithm, Henon chaotic map is used which due to its chaotic character, is highly sensitive to initial conditions. Also, we are using 128-bit external key for image encryption, which is highly sensitive as well.

### 4.4 Histogram Analysis

Histogram of an image provides information about the frequency distribution of its pixels and regarding density estimation. A cipher image should have a uniform

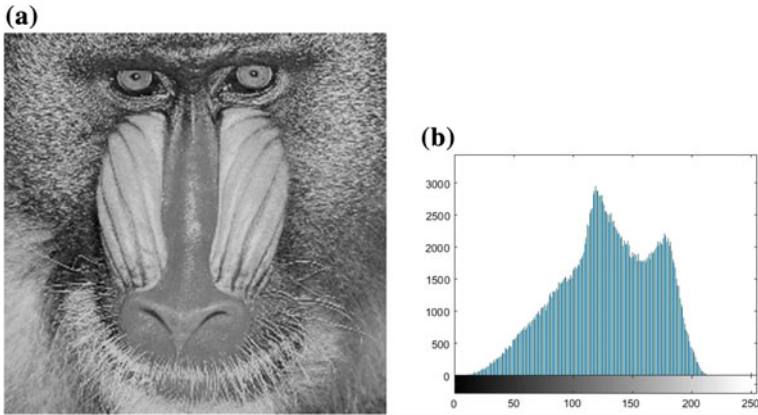


Fig. 3 Histogram analysis of plain baboon image

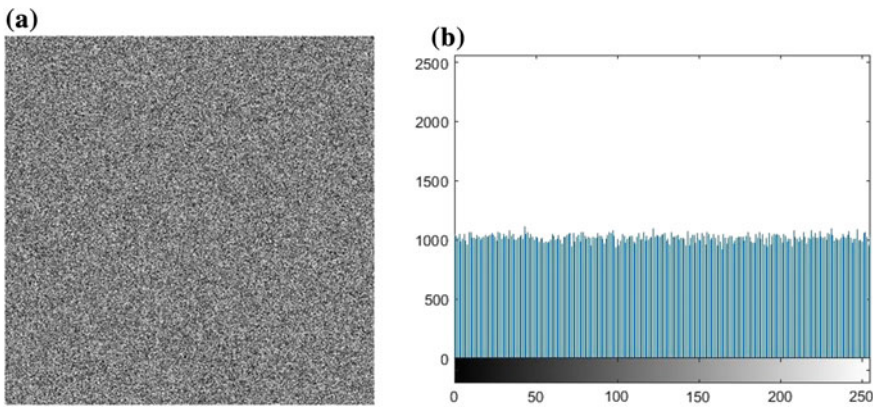


Fig. 4 Histogram analysis of encrypted baboon image

histogram to be secure from known plain-text attack. Figure 3 depicts the histogram of the original baboon image and Fig. 4 shows the histogram of the encrypted image. Since the histogram of the encrypted image is nearly uniform, the proposed algorithm is expected to prevent known plain-text attack. As the histogram of the encrypted image is uniform, it can be concluded that the proposed algorithm is highly resistant to statistical attacks.

### 4.5 Information Entropy

Information entropy is a measurement of uncertainty or randomness in a signal or image. A good encryption technique must incorporate randomness property and follow uniform distribution [5]. It is calculated by the following formula:

$$H(m) = - \sum_{i=0}^{2^N-1} P(m_i) \log_2 [P(m_i)] . \tag{3}$$

where,

P (mi) = Probability of a pixel, and

N = Bit-depth of each pixel

### 4.6 Correlation Analysis

High correlation is one of the most important characteristics of data belonging to the class of digital images. Each pixel is strongly correlated with its neighboring pixels which may be horizontal, vertical or diagonal in position. Scatter plots are shown in Figs. 5, 6, and 7 for depicting the correlation between randomly selected 5000 pixel pairs of each of the horizontal, vertical and diagonal locations for both the original plain image and the generated cipher image. The standard 512 \* 512 sized gray scale cameraman image was used for the correlation testing. Correlation coefficients are calculated using the Eq. 1 shown. The correlation coefficients of the various plain images and that of respective encrypted images are shown in Fig. 8. Generally, for a normal image, pixels are highly correlated and the coefficients are very close to 1, while for the encrypted image, the coefficients are close to 0.

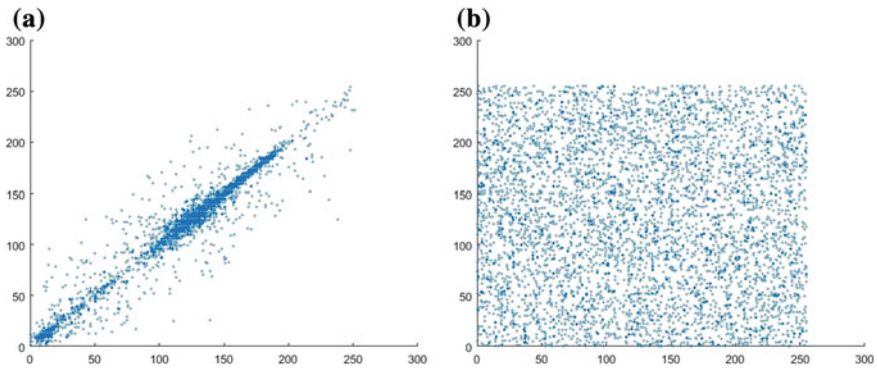
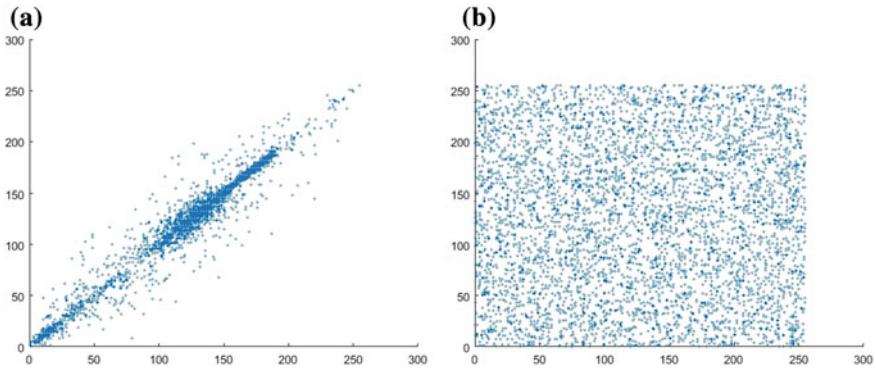
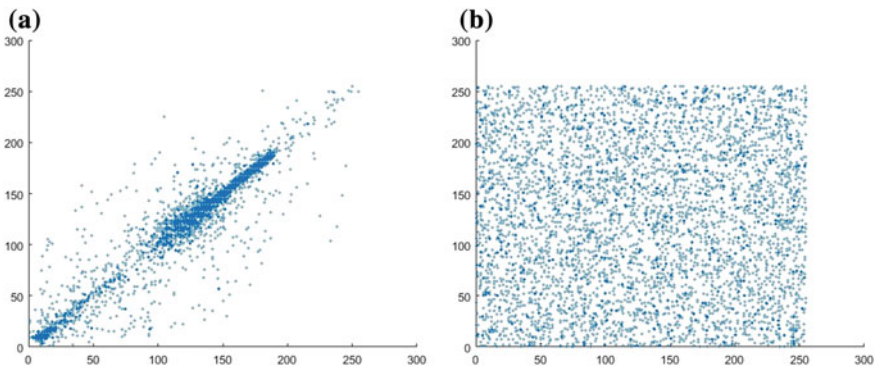


Fig. 5 Correlation analysis of horizontal pixel pairs



**Fig. 6** Correlation analysis of vertical pixel pairs



**Fig. 7** Correlation analysis of diagonal pixel pairs

## 5 Conclusion

In this paper, a new method for image encryption is proposed. The technique is based upon using chaotic properties of Henon map as pseudo-random number generator along with 128 bit secret key to obtain permutation matrix for shuffling of the original image and a cipher image that is used to finally encrypt the shuffled image. The method is vigorously tested on standard test images based upon various security parameters of digital image encryption. The focus is kept on keeping the mechanism simple enough, making it easy to implement in practical applications.

The future scope of the work may constitute the optimization of the algorithms for applications in sensor nodes and military applications where the processing ability of the nodes is extremely low. As, any algorithm that is costly in terms of computational cost, that can not be implemented in the discussed scenario.



Image	Plain entropy	Cipher entropy	Plain image correlation			Cipher image correlation		
			HC	VC	DC	HC	VC	DC
Baboon (512*512)	7.2925	7.9993	0.9337	0.9123	0.8669	0.0027	0.0013	-5.4562e-04
Cameraman (512*512)	7.0480	7.9994	0.9831	0.9900	0.9733	-0.0022	2.4476e-04	-0.0035
Lena (512*512)	7.4451	7.9992	0.9719	0.9850	0.9593	-7.4248e-04	-7.5330e-04	9.1292e-04
Peppers (512*512)	7.5925	7.9992	0.9767	0.9791	0.9638	-9.4483e-06	0.0014	0.0020
Woman_blonde (512*512)	7.0390	7.9992	0.9386	0.9595	0.9153	0.0023	0.0020	-4.2859e-04

Fig. 8 Security analysis of the proposed algorithm

## References

- Chen, C.-S., Chen, R.-J.: Image encryption and decryption using SCAN methodology. In: 7th International Conference on Parallel and Distributed Computing, Applications and Technologies, 2006. PDCAT06, pp. 61–66. IEEE (2006)
- Sankpal, P.R., Vijaya, P.A.: Image encryption using chaotic maps: a survey. In: 2014 Fifth International Conference on Signal and Image Processing (ICSIP), pp. 102–107. IEEE (2014)
- Rajput, A.S., Mishra, N., Sharma, S.: Towards the growth of image encryption and authentication schemes. In: 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 454–459. IEEE (2013)
- Jolfaei, A., Mirghadri, A.: An image encryption approach using chaos and stream cipher. *J. Theor. Appl. Inf. Technol.* **19**(2), 117–125 (2010)
- Kumar, M., Aggarwal, A., Garg, A.: A review on various digital image encryption techniques and security criteria. *Int. J. Comput. Appl.* **96**(13) (2014)
- Wei-bin, C., Xin, Z.: Image encryption algorithm based on Henon chaotic system. In: 2009 International Conference on Image Analysis and Signal Processing. IEEE (2009)
- Ping, P., Mao, Y., Lv, X., Xu, F., Xu, G.: An image scrambling algorithm using discrete Henon map. In: 2015 IEEE International Conference on Information and Automation, pp. 429–432. IEEE (2015)
- Nithin, N., Bongale, A.M., Hegde, G.P.: Image encryption based on FEAL algorithm. *Int. J. Adv. Comput. Sci. Technol.* (2013)
- Hamad, S., Khalifa, A., Elhadad, A., Rida, S.Z.: A modified playfair cipher for encrypting digital images. *Mod. Sci.* (2013)
- Soleymani, A., Nordin, M.J., Sundararajan, E.: A chaotic cryptosystem for images based on Henon and Arnold cat map. *Sci. World J.* (2014)
- Forre, R.: The Hnon attractor as a keystream generator. In: *Advances in Cryptology-EuroCrypt*, vol. 91, pp. 76–81 (1991)
- Hnon, M.: A two-dimensional mapping with a strange attractor. *Commun. Math. Phys.* **50**(1), 69–77 (1976)

# A New Approach to Provide Authentication Using Acknowledgment



Vijay Paul Singh, Naveen Aggarwal, Muzzammil Hussain  
and Charanjeet Kour Raina

**Abstract** Whenever there is communication between two nodes, there is a possibility of vulnerability. When node A communicates with node B, a lot of problems arise between them, out of which security breach is one. If data are sensitive it is essential that the data should be safely delivered to the authorized node. To handle security issues, cryptography technique is used, through which security attacks are not possible. For secure communication, there are a number of security protocols that protect from attackers. One of the cryptosystems is the public-key encryption. If attackers obtain the private key, they can easily read encrypted messages and masquerade it or perform some unknown activity which can be highly adverse for the communication. To overcome this problem, we introduce acknowledgment-based authentication in which even if the attacker knows the private key he/she could not communicate with authorized parties continuously.

**Keywords** Authentication · Security · Encryption · Keys · Acknowledgment  
Decryption

---

V. P. Singh (✉) · N. Aggarwal  
Department of CS, UIET, Panjab University, Chandigarh, India  
e-mail: vpaul9678@gmail.com

N. Aggarwal  
e-mail: navagg@gmail.com

M. Hussain  
Central University Rajasthan, Kishangarh, India  
e-mail: mhussain@curaj.ac.in

C. K. Raina  
PTU, Kapurthala, India  
e-mail: ckraina23@gmail.com

## 1 Introduction

Security is an essential issue during communication, either communication between two nodes or multiple nodes. Whenever we talk about security we focus on loopholes, i.e., possible security attacks. That is why computer security requires confidentiality, integrity, authentication, authorization and availability [1]. Every security protocol tries to fulfill these requirements so that there is less possibility of security breach. As we know there are two types of attacks: active attack and passive attack. Active attack changes the system resources, such as masquerades, replay, DoS, modification of messages, etc., whereas passive attack uses the information between nodes or we can say eavesdropping is possible by it. Passive attack is very silent; sender and receiver cannot determine this type of attack. To handle such types of attack cryptography technique is used. There are number of protocols to handle such types of attack. Even in the current scenario, to protect from this attack, sender and receiver generally use encryption and decryption techniques. Through encryption and decryption plaintext is converted into ciphertext in such a way that no masquerader can read that ciphertext and only authorized node can understand that ciphertext. To perform encryption and decryption, generally there are two techniques. It is done by using either symmetric or asymmetric keys. In symmetric technique, encryption and decryption are done by the same key and only authorized parties know about the keys, whereas in asymmetric technique, a pair of keys, i.e., public key and private key, is used. Public-key cryptography is one of the common and more secure methods, but in reality security of any encryption and decryption depends upon the length of key; the higher the size of the key, the higher will be the security of the encrypted message [2, 3]. In public-key cryptography, encryption of the message is done with the help of public key of the authorized receiver and decryption of that message is done by the private key of the authorized receiver [4]. In public-key cryptography everyone knows the public key of receiver, but no one knows the private key of receiver. So, only authorized receiver can decrypt that encrypted message with his own private key. In this way sender and receiver transmit data securely. But at every stage of the communication both sender and receiver cannot re-verify each other. And if somehow an attacker gets private key he/she can easily decrypt message at any stage and understand the encrypted message. To consider this type of situation we introduce acknowledgment-based authentication, through which the authorized user can re-verify that he/she communicates with genuine user and even they can check at any time. Through acknowledgment-based authentication authorized user can check at any time or at any step during communication. In this way attacker cannot understand the encrypted message and the communication between authorized users is secure. In this paper we elaborate acknowledgment-based protocol and implement it. The following sections are in this paper: Sect. 2 presents the related work; Sect. 3 presents the proposed work; and then Sect. 4 presents the implementation. Eventually conclusion and future work are given in Sects. 5 and 6, respectively.

## 2 Related Work

Cryptography is a technique that hides sensitive information in such a way that masquerader attack cannot possible. Cryptography is mainly based on mathematical relations that fulfill the security requirements.

There are a number of security protocols which protect from intruder, hacker, adversary, etc. Every protocol uses cryptographic technique through which no attacker can succeed in breaching the securities. There are few security protocols which are little bit related to our proposed algorithm; such common protocols are Needham and Schroeder, Kerberos, IPSec, Point-to-Point Protocol, Internet Key Exchange, Transport Layer Security, handshake protocol, etc. [5].

Needham–Schroeder protocol is a shared-key authentication protocol. In this protocol, whenever node A communicates with node B they use secure symmetric keys which are provided by the trusted key server S and use nonce during message transfer for freshness [6]. Also one more paper regards sharing of session key securely between nodes following Needham–Schroeder protocol [7]. Kerberos authentication protocol is based on ticket-granting ticket service provided by ticket-granting server; there are six steps required for performing Kerberos authentication protocol [8]. Handshake protocol was divided into four phases such as establish security capabilities between client and server then later they authenticate with the help of key exchange after that change Cipher Spec and finished [9]. One more paper used acknowledgment for verification of the authorized user [10].

On the other side of Acknowledgment, it is one of the common processes used during communication. There are many protocols which are acknowledgment based; one of them is TCP. Acknowledgment is generally used for confirmation; whenever two nodes communicate they send acknowledgment to each other after the successful reception of the message. In most cases the size of acknowledgment is 32 bits [11]. It contains information about the source and the sender.

## 3 Proposed Work

We know the sender and the receiver both want secure communication. To perform secure communication they use encryption and decryption techniques. In this way unauthorized user/attacker cannot understand the encrypted message; generally, RSA, Diffie–Hellman, etc., are used for encryption and decryption. Through acknowledgment-based authentication secure communication is possible between two authorized users and users can cross-check at any time. In this way, the man-in-the-middle attack is not possible to authorized users (Table 1).

Step 1: A sends M to B in encrypted form using the public key of B.

$$A \rightarrow B : Y = E [PU_B, M]$$

**Table 1** Symbols

Name	Symbol
Alice	A
Bob	B
Plaintext	M
Encryption	E
Decryption	D
Ciphertext	Y, Z
Public key of Bob	PU <sub>B</sub>
Private key of Bob	PR <sub>B</sub>
Public key of Alice	PU <sub>A</sub>
Private key of Alice	PR <sub>A</sub>
Random number	R

Step 2: B receives A's message and decrypts it with private key PR<sub>B</sub>.

$$B \rightarrow A : M = D [PR_B, Y]$$

Step 3: B replies to A and encrypts the M using the public key of A.

$$B \rightarrow A : Z = E [PU_A, M]$$

Note: In this way, they communicate with each other using this technique; no intruder can understand their encrypted message. If adversary "T" got the private key they could easily decrypt the message and read it. To handle such type of situation we introduce this proposed algorithm.

### 3.1 Proposed Algorithm

Step 1: A sends M to B in encrypted form using the public key of B.

$$A \rightarrow B : Y = E [PU_B, M]$$

Step 2: B receives A's message and decrypts it with private key PR<sub>B</sub>.

$$B \rightarrow A : M = D [PR_B, Y]$$

Step 3: B re-verifies that A is genuine or not and sends random number R to it encrypted with public key PU<sub>A</sub>.

$$B \rightarrow A : Z = E [PU_A, R]$$

Here, the value of R is based on acknowledgment table. If somehow an attacker decrypts this, still he/she cannot reply to B because he/she has no acknowledgment table. R lies between a number of acknowledgments in acknowledgment table maintained by B.

Step 4: A receives that message and decrypts it using  $PR_A$  and uses that random number as a serial number of secure acknowledgment table and fetches value from that table ACK and encrypts the message with that ACK and sends it to B.

$$A \rightarrow B : Q = E [ACK, M_2]$$

Step 5: B receives the encrypted message and decrypts it with the same ACK value and finds whether A is genuine or not.

$$ACK = ACK$$

In this proposed algorithm both maintain the table of acknowledgments. Whenever they communicate with each other, eventually B sends an acknowledgment to the A and that acknowledgment is maintained by both for further verification. With the help of these acknowledgments, they can verify each other at any moment. Figure 1 shows the data transfer mechanism between client and server, where the client communicates with server and server verifies it.

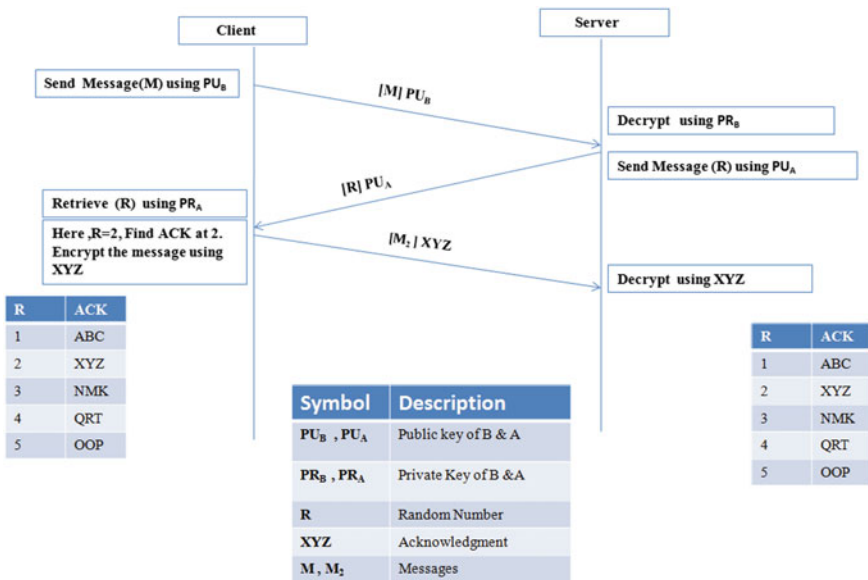


Fig. 1 Data transfer mechanism

### 4 Implementation

To simulate this we used Java programming and with the help of socket programming, we perform the communication. To maintain acknowledgment table on both sides we use MYSQL, and to perform it we use NetBeans and implement in Windows 7.

Here are two cases: In the first case, when their communication takes place for the first time, there is no need of acknowledgment verification. In the second case, whenever they communicate with each other they use the above-proposed algorithm and they ensure that they communicate with the authorized user. The figures show case 2 when they re-verify each other; whenever A communicates with B, B rechecks that A is genuine or not using Step 3 and the aftermath by receiving message from A in Step 7, where B matches Acknowledgment with the stored actual acknowledgment. In this way, they protect from man-in-the-middle attack.

Figure 2 shows Steps 1, 4 and 5 of the proposed algorithm at A.

Figure 3 shows Steps 2, 3, 6 and 7 of the proposed algorithm at B.

```
run:
-----Step 1-----
A send message to B:- Hello B I am A
Encrypted Form : 3165962034131166545789524621494147022074755768374
-----Step 4-----
A received Message by B:-1028071702528
Decrypted Form : 4
-----Step 5-----
A send message to B:- AASADA289GTHYJU7
Encrypted Form : 6320346566283129167066262652691626888463373084026
```

Fig. 2 Node A

```
run:
-----Step 2-----
B received A message:- 3165962034131166545789524621494147022074755768374
Decrypted Form : Hello B I am A
-----Step 3-----
B re-verified that A is genuine or not send Random Number:- 4
Encrypted Form : 1028071702528
-----Step 6-----
B received A message:- 6320346566283129167066262652691626888463373084026
Decrypted Form : AASADA289GTHYJU7
-----Step 7-----
A is Genuine
```

Fig. 3 Node B

## 5 Conclusion

Security is an extremely important aspect whenever sensitive information is transferred between two nodes. There are a number of security protocols which protect from attackers. The above-proposed algorithm verifies that users securely communicate with each other. There are a number of circumstances where this proposed algorithm provides security and mitigates different attacks, and this algorithm secures the communication at every level and protects it from attacks by an attacker.

## 6 Future Work

In this paper, we focus on security issues between A and B nodes; in future work, we will introduce machine learning techniques through which they can check at any time whenever they feel misbehavior feature of A.

## References

1. Stallings, W.: *Cryptography and Network Security: Principles and Practices*. Pearson Education India (2006)
2. Callegati, F., Cerroni, W., Ramilli, M.: Man-in-the-middle attack to the HTTPS protocol. *IEEE Secur. Priv.* **7**(1), 78–81 (2009)
3. Jonsson, J., Kaliski, B.: *Public-key Cryptography Standards (PKCS)# 1: RSA Cryptography Specifications Version 2.1* (2003)
4. Forouzan, B.A., Mukhopadhyay, D.: *Cryptography and Network Security (Sie)*. McGraw-Hill Education (2011)
5. Salomaa, A.: *Public-key Cryptography*. Springer Science & Business Media (2013)
6. Lowe, G.: An attack on the Needham-Schroeder public-key authentication protocol. *Inf. Process. Lett.* **56**(3), 131–133 (1995)
7. Gupta, A., Hussain, M.: Secure session key sharing using public key cryptography. In: *Proceedings of the Third International Symposium on Women in Computing and Informatics*. ACM (2015)
8. Neuman, B.C., Ts'o, T.: Kerberos: an authentication service for computer networks. *IEEE Commun. Mag.* **32**(9), 33–38 (1994)
9. Wager, D., Schneier, B.: Analysis of the SSL 3.0 protocol. In: *The Second USENIX Workshop on Electronic Commerce Proceedings*, vol. 1. no. 1 (1996)
10. Singh, V.P., Hussain, M., Raina, C.K.: Authentication of base station by HDFS using trust based model in WSN. In: *International Conference on Communication and Electronics Systems (ICCES)*. IEEE (2016)
11. Mathis, M., et al.: TCP selective acknowledgment options. No. RFC 2018 (1996)



# Prevention of Replay Attack Using Intrusion Detection System Framework



Mamata Rath and Binod Kumar Pattanayak

**Abstract** In current mobile technology, mobile ad hoc networks feature numerous challenges to sustain constant connectivity when there are adverse network situations such as link failure, routing attacks and security threats. In such cases, the network exhibits reactive approach using detection mechanism to prevent attacks. At the same time this network also demonstrates proactive approach by making attempts to avoid an attacker from doing attacking activity. It uses various cryptographic methodologies as the prevention mechanism for doing so. The proposed article presents an intrusion detection mechanism-based framework (IDMBF) for secured routing in a very specialized and challenging mobile ad hoc network. Simulation results show that the proposed system exhibited reduced packet loss rate and comparatively reduced end-to-end delay during data transmission when compared to other similar approaches.

**Keywords** MANET · IDS · Mobile agent · DoS attack  
Security protocol

## 1 Introduction

For sustaining security in network, intrusion detection system [1] plays a vital role in MANET. According to the increasing trends in technology, there is a critical requirement for development of robust secured systems using combined technology of different domains such as cross-layer communication or neural network implementation or game theory-based strategic solutions, etc. MANETs do not have any

---

M. Rath (✉)

C. V. Raman College of Engineering, Bhubaneswar, Odisha, India  
e-mail: mamata.rath200@gmail.com

B. K. Pattanayak

Department of Computer Science and Engineering, SOA University,  
Bhubaneswar, Odisha, India  
e-mail: binodpattanayak@soauniversity.ac.in

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_32](https://doi.org/10.1007/978-981-13-1708-8_32)

fixed infrastructure [2], and the architecture based on various topologies frequently changes due to the fact that numerous mobile devices continuously are attached to a network and go away from the network range frequently; therefore, keeping track of regular security aspects is another primary challenge [3]. There are some specific purposes. MANETs are designed for military and defence purposes that also face problems during search and rescue operations. In many applications like natural calamity or in battle fields, the infrastructure may not be there, or if there, then it has been damaged, so it requires formation of a new MANET within short period of time. Design and implementation of an efficient and attack-preventive secured IDS [4] is highly desired in such wireless technology-based specialized network.

## 2 Related Work

Targeting the issue of delay in intercommunication, paper [5] discusses about the nodes in a MANET, which are swarm robots with good control of sensing and transportation. An innovative approach has been used here for better interaction among robots in a MANET using spring force laws based on attraction and repulsion laws. An Extended Virtual Spring Mesh Framework has been presented here with better performance with adaptive control parameters. To prevent the network from various attacks, it is more important to develop secured systems. In this motive, article [6] presents an analysis and detailed study on various IDS structures, especially the way they have been transformed from normal IDS systems of the primitive wireless ad hoc networks to the ambient intelligent scenario of the computing systems. To investigate the DDOS (distributed denial-of-service) attack, using non-address spoofing flood in MANET, paper [7] proposes an innovative approach. Here, the detected features based on statistical analysis of IDS log files are proposed. Various NASF-based attacks and their patterns are simulated and tested. In paper [8] a new security framework has been designed with an intention to perform quick adaptation to dynamic link conditions, minimum processing overhead and very slow utilization of the network. Another new concept of powerful security system has been proposed in [9], named as EAACK, and it uses hybrid cryptography and bouncing theory for minimizing network overhead. Paper [10] gives a method for reorganization of malicious node in the network by comparing many parameters such as nodal energy and the level of reliability. Another paper [11] addresses the security aspects of the MANET with the help of some combined technology used in an IDS. It improves the network security by using a secondary network sensor and an improved location algorithm. This technique was simulated using a military tactical scenario. In another approach [12], a hybrid technique is used to lessen the network overhead due to digital signature by combining the principles of RSA and AES algorithms effectively. The proposed system in [13] detects the anomaly in the network with the help of ANN (artificial neural network) theory to prevent network attacks. In [14]

the authors have proposed a new intrusion detection system, called EAACK, which is specially intended for mobile ad hoc networks. This strategy exhibits higher rate of detecting malicious behaviour among the nodes without disturbing the network performance to a greater extent. In [15] an efficient scheme has been proposed for MANET, which analyses and optimizes the period for which the IDS needs to remain active. This is carried out by making good cooperation among IDS and neighbour nodes to minimize their personal activation time. Otherwise, an IDS has to always remain active to monitor the suspicious activities in the network.

### 3 Design of the Secured Framework

This section provides illustrations of detailed design of the proposed framework for MANET that is based on a robust energetic protocol called PDO-AODV (Power and Delay Optimized AODV protocol), which is our significant previous research work [2]. Originally, we had developed a network framework in TCP/IP suite at the network layer that performs the routing among the mobile nodes of MANET with energy efficiency, managed delay and load balancing [3]. Now as an extension of this work we are proposing the same energy-efficient technique in our proposed IDS framework. As efficient communication among the workstations in mobile ad hoc networks is a challenging task, so to increase the network scalability and to improve the network lifetime along with maintaining the correct level of quality of service [16] and to prolong the communication period, a secured IDS framework called intrusion detection mechanism-based framework (IDMBF) for secured routing has been presented here.

This system is specifically meant for MANET scenario. As per this technique, a group of mobile stations in a MANET, which are also participating in a real-time application, undergo registration process with a dynamic mobile agent. Thereafter, the mobile agent assigns unique registration identification numbers to these stations. Those members with a particular mobile agent identification number are identified in the network as authorized entities to get the mobile agent (MA) service under the IDS framework [17] to get prevented from replay attack. As depicted in the flow chart in Fig. 1 the MA calls a check authentication function in which it directly communicates with the original source node with its initial IP address to validate. It uses control messages to confirm a positive reply from the source. Figure 2 depicts the functionality of the function. After receiving an authenticated message the MA [16] allows the receiver to further send the route reply message; otherwise, if the validation is not successful, it reverts back to the receiver with an alert message not to send route reply or any other sensitive information.

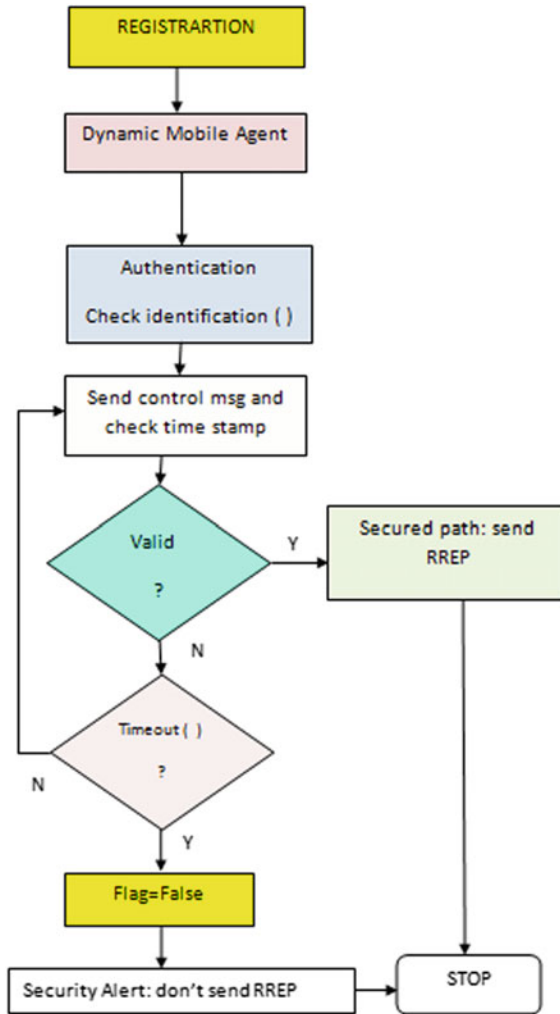


Fig. 1 Flow chart of the security module in the proposed approach

### 4 Simulation and Results

The proposed work has been simulated using a well-known simulation tool called NETSIM. It has very powerful features for performance measurement. It has rich set of library files for most of the wired and wireless protocols, availability of C source code, animated features during simulation, easy debugging and coding facility. It supports the functions of advanced wireless networks such as mobile ad hoc networks and WiMAX. Most of the MANET protocols such as DSR, AODV, IETF RFC 4728,

**Fig. 2** The proposed algorithm used by the mobile agent

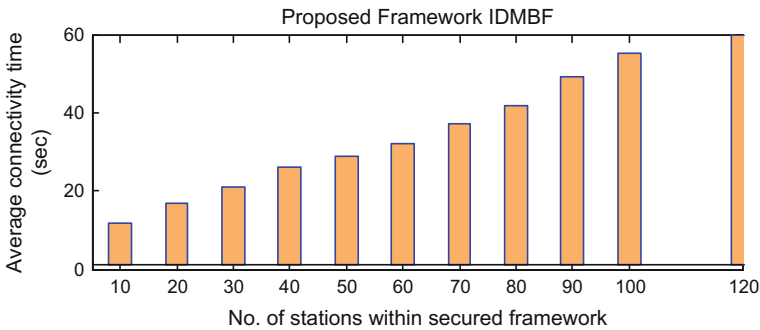
```
Float check_identification (source_id, time )
{
    Flag=True
    Send Hello message to source ip-address;
    While not getting a response
    {
        Send Hello message;
    }
    If time_out( )
    {Flag=False; break ;}
    If (Flag == True)
        Command: not a False route : send RREP
    Else
        {Alert: Stale Route: Dont send RREP}
```

IEEE 802.16 D are configured in it for simulation purpose. The network parameters used during simulation are indicated in Table 1 [18].

Figure 3 describes the simulated connectivity time under IDS framework, and Fig. 4 shows the Simulated Data Drop result in IDS framework IDMBF. The proposed framework can be applied in very drastic situation of network where MANET is the only preferable solution for network formation and data communication. We have performed the comparative analysis of this protocol with other prominent protocols designed for such extreme scenes by considering some other critical issues. In this section we present the detailed mechanism of the similar security approaches using IDS by other researchers so that we can compare our proposed work with these valuable contributions. Intrusion detection systems with multilayer detection technology have rapidly evolved recently to prevent various attacks that prevent authorized users from accessing network resources. In this direction an innovative cross-layer intrusion detection architecture has been designed by the authors in [19] to find out the malicious nodes and prevent the network from various types of DoS attacks.

**Table 1** Network parameters [18]

Parameter name	Parameter value
Channel type	Wireless channel
Radio model	Two-ray ground
Network interface type	Wireless PHY
Type of traffic	VBR
Simulation time	5 min
MAC type	Mac/802_11
Max. speed	50 m/s
Network size	1600 × 1600
Mobile nodes	120
Packet size	512 Kb
Interface queue type	Queue/Drop tail
Simulator	NETSIM

**Fig. 3** Simulated connectivity time under IDS framework

Cooperative anomaly intrusion detection with data mining (CAIDD) technique has been used here to improve the proposed system. A fixed width clustering algorithm has been implemented in the said article for detecting the problems more effectively in MANET traffic. The simulation results by the OPNET simulator indicate better results in terms of stable network stability and improved network lifetime when compared to other similar types of approaches. Paper [20] proposes the application of a combination of techniques for the design of a collaborative MANET intrusion detection system (CMIDS) that improves the network security using a predictive location algorithm by sharing the protocol functionality. The objective of designing a secured network model with a strong IDS based on unpredictable malicious behaviour is to eliminate the higher rate of network vulnerability. In this context, paper [21] presents ZIDS (zone-based intrusion detection system) which applies the game theory concept to mine the uncertain strategies of the malicious nodes. Simulation of this process performs better with respect to correct detection rate.

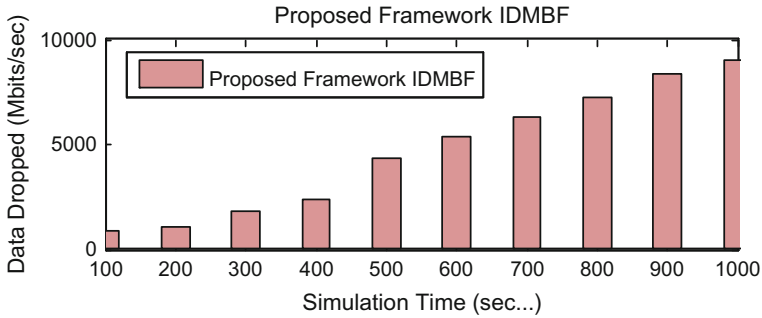


Fig. 4 Simulated data drop result in IDS framework IDMBF

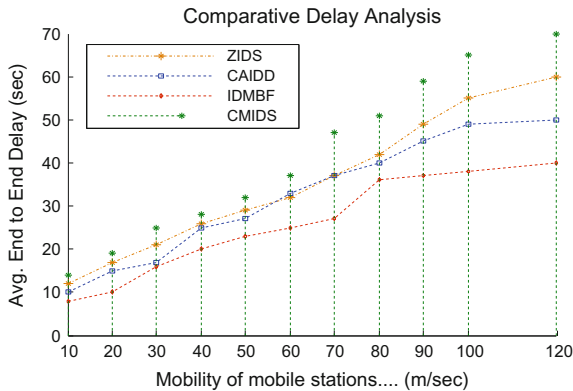


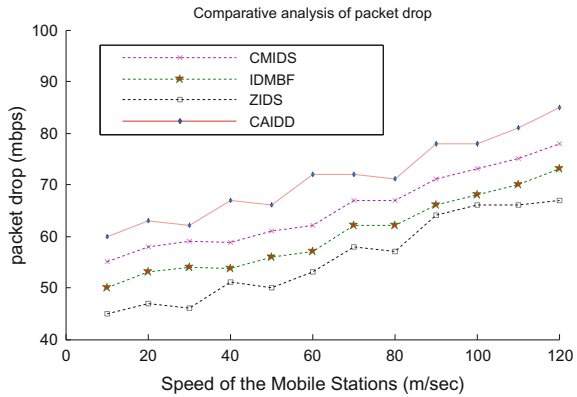
Fig. 5 Comparative delay analysis

In the above section comparative analysis of end-to-end delay and packet drop rate has been performed to compare the simulation results with similar approaches. As depicted in Fig. 5, it can be analysed that the proposed approach exhibits reduced end-to-end delay in comparison with other similar approaches, and similarly, it can be concluded from Fig. 6 that the proposed secured approach IDMBF eliminates the packet loss problem to a greater extent.

## 5 Conclusion

Security issues in mobile ad hoc networks are dynamic challenges. There is an effort in the said article to provide a secured base for the prevention of an important cyberattack called replay attack during the transmission of sensitive information over the wireless network. An improved security logic has been used here to save the victims of this attack by registering the mobile stations with an intrusion detection

**Fig. 6** Comparative analysis of packet drop



mechanism. This mechanism is controlled by a dynamic mobile agent that regularly checks the authenticity of the source node and alerts the member nodes before any attack is imposed on them. Simulation results show that the proposed approach performs better in terms of reduced packet loss and enhanced network lifetime.

## References

1. Butun, I., Morgera, S.D., Sankar, R.: A survey of intrusion detection systems in wireless sensor networks. *IEEE Commun. Surv. Tutor.* **16**(1), 266–282, First Quarter (2014)
2. Rath, M., Pattanayak, B.K., Pati, B.: Energy efficient MANET protocol using cross layer design for military applications. *Def. Sci. J.* **66**(2) (2016)
3. Rath, M., Pattanayak, B.K.: Energy competent routing protocol design in MANET with real time application provision. *Int. J. Bus. Data Commun. Netw.* **11**(1), 50–60 (2015)
4. Nadeem, A., Howarth, M.P.: A survey of MANET intrusion detection & prevention approaches for network layer attacks. *IEEE Commun. Surv. Tutor.* **15**(4), 2027–2045, Fourth Quarter (2013). <https://doi.org/10.1109/surv.2013.030713.00201>
5. Derr, K., Manic, M.: Adaptive control parameters for dispersal of multi-agent mobile ad hoc network (MANET) swarms. *IEEE Trans. Industr. Inf.* **9**(4), 1900–1911 (2013)
6. Chaki, R.: Intrusion detection: Ad-hoc networks to ambient intelligence framework. In: 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krackow, pp. 7–12 (2010)
7. Guo, Y., Lee, I.: Forensic analysis of DoS attack traffic in MANET. In: 2010 Fourth International Conference on Network and System Security, Melbourne, VIC, pp. 293–298 (2010)
8. Shrestha, R., Sung, J.Y., Lee, S.D., Sik-Yun, P., Choi, D.Y., Han, S.J.: A secure intrusion detection system with authentication in mobile ad hoc network. In: 2009 Pacific-Asia Conference on Circuits, Communications and Systems, Chengdu, pp. 759–762 (2009)
9. Awatade, S., Joshi, S.: Improved EAACK: develop secure intrusion detection system for MANETs using hybrid cryptography. In: 2016 International Conference on Computing Communication Control and automation, Pune, pp. 1–4 (2016)
10. Nanaware, P.M., Babar, S.D.: Trust system based intrusion detection in mobile ad-hoc network (MANET). In: 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, pp. 1–4 (2016)



11. Carvalho, J.M.A., Costa, P.C.G.: CMIDS: collaborative MANET intrusion detection system. In: 2016 International Conference on Cyber Conflict (CyCon U.S.), Washington, DC, pp. 1–5 (2016)
12. Patil, T., Joshi, B.: Improved acknowledgement intrusion detection system in MANETs using hybrid cryptographic technique. In: 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, pp. 636–641 (2015)
13. Indira, N.: Establishing a secure routing in MANET using a hybrid intrusion detection system. In: 2014 Sixth International Conference on Advanced Computing (ICoAC), Chennai, pp. 260–263 (2014)
14. Shakshuki, E.M., Kang, N., Sheltami, T.R.: EAACK—a secure intrusion-detection system for MANETs. *IEEE Trans. Industr. Electron.* **60**(3), 1089–1098 (2013)
15. Marchang, N., Datta, R., Das, S.K.: A novel approach for efficient usage of intrusion detection system in mobile ad hoc networks. *IEEE Trans. Veh. Technol.* **66**(2), 1684–1695 (2017)
16. Rath, M., Pattanayak, B.K.: MAQ: a mobile agent based quality of service platform for MANETs. *Int. J. Bus. Data Commun. Netw.* **13**(1) (2017)
17. Pattanayak, B., Rath, M.: A Mobile agent based intrusion detection system architecture for mobile ad hoc networks. *J. Comput. Sci.* **10**, 970–975 (2014)
18. Rath, M., Pattanayak, B.K., Pati, B.: Energetic routing protocol design for real-time transmission in mobile ad hoc network. In: *Computing and Network Sustainability, Lecture Notes in Networks and Systems*, vol 12. Springer, Singapore (2017)
19. Shrestha, R., Han, K.H., Choi, D.Y., Han, S.J.: A novel cross layer intrusion detection system in MANET. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications, Perth, WA, pp. 647–654 (2010)
20. Carvalho, J.M.A., Costa, P.C.G.: Collaborative approach for a MANET intrusion detection system using multilateration. In: 2016 11th International Conference on Computer Engineering & Systems (ICCES), Cairo, pp. 59–65 (2016)
21. Sangeetha, V., Kumar, S.S.: ZIDS: zonal-based intrusion detection system for studying the malicious node behaviour in MANET. In: 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, pp. 276–281 (2015)

# Appending Photoplethysmograph as a Security Key for Encryption of Medical Images Using Watermarking



M. J. Vidya and K. V. Padmaja

**Abstract** In the current scenario, once a patient has been diagnosed with a disease, an expert physician's opinion is sought forthwith, in accordance with advent of technology, and techniques do come into a key role in medical diagnosis. Most of the patients and physicians prefer to get a viewpoint before proceeding further with procedural treatment plans. Henceforth, securing and transmitting of data plays a vibrant role in terms of accuracy, security and other parameters. Cyber criminals involved in hacking medical data, look at it as an opportunity to hawk these sensitive data, leading to the hour of concern. The augment is to allow the patient information to govern and share, to end parties with at most level of seizure; so that information cannot be leaked. Because Government, International and National Medical Associations are looking at medical data security as a priority, it is very important to have an efficient algorithm or a method. The novelty in the proposed work lies in using the patient data as a security protocol and appending three stages of security: bundle encryption generated based on patient ID and age as the first stage, augmentation index derived from bioelectric signal source—photoplethysmograph (PPG)—as a pivotal opener and hybrid discrete wavelet transform–discrete cosine transform (DWT-DCT) watermarking in the second stage, last level of de-watermarking of embedded data from facial photograph of the patient.

**Keywords** Watermarking · Photoplethysmograph · Electronic Patient Record

## 1 Introduction

With the advancement of communication, accessing healthcare system and distributing patient information has also advanced. Together with this, the difficulties engaged with information security have additionally expanded. Secrecy (approved clients can

---

M. J. Vidya (✉) · K. V. Padmaja

Department of Electronics & Instrumentation Engineering, R. V. College of Engineering, R. V. Vidyaniketan Post, Mysuru Road, Bengaluru 560059, India  
e-mail: vidyamj@rvce.edu.in

K. V. Padmaja

e-mail: padmajakv@rvce.edu.in

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_33](https://doi.org/10.1007/978-981-13-1708-8_33)

just approach the information), integrity (data ought not be adjusted) and authenticity (accessed by approved staff) are the real qualities required to satisfy the prerequisite of information security [1, 2]. As per the 2017 Global Health Care Outlook Report published by Deloitte [3], cyber-theft and cyber espionage continue to endanger patient privacy and the use of sensitive patient information in the medical industry. Annually close to \$5.6 billion is spent by medical industry for medical data security.

As indicated by Thales Healthcare Report [4], 81% of US healthcare organizations and 76% of worldwide medicinal services associations will invest more on security of medical data in 2017. Numerous researchers have proposed different computation methods which depend on spatial and frequency domain.

Gran et al. [5] recommended that the performance of LZW algorithm resulted in better lossless compression watermarking for ultrasound medical images. The outcomes demonstrate that there is a reduction in effective lossless recovery and effective tamper detection.

Saman Iftikhar et al. [6] proposed a reversible watermarking technique for used to empower highest information security utilizing Z-notation based formal specification.

Frequency domain technique based on DCT was proposed by Chao et al. [7] where an Electronic Patient Record (EPR) data are embedded in DCT of watermarked images. Ritu Agarwal [8] proposed a therapeutic imaging watermarking strategy which is strongly utilizing M-ary modulation amid the DCT band for two cerebrum imaging modalities with high subtlety. Additionally DCT-based scheme where ECG information is inserted into medical images was proposed by Acharya et al. [9].

Maity et al. [10] utilized a technique for contrast mapping to reverse watermark and subsequently making the framework powerful.

Together with the watermarking techniques, programmed human recognizable proof utilizing biometric framework has increased huge significance in healing and treatment centers and enterprises. Certain highlights of human practices or qualities of the body can be utilized as methods for human distinguishing proof. Some of the cases of biometrics being utilized are face recognition [11], voice recognition [12], electroencephalograph (EEG) mapping [13], fingerprint identification [14] and electrocardiograph (ECG) mapping [15].

In the current circumstances, numerous applications utilize any of these or a blend of them to give human biometric distinguishing proof. The conventional strategies for human confirmation accessible have many drawbacks: confronts can be tricked by approach-measured photographs, voices can be imitated, fingerprints can be reproduced in latex, and EEG or ECG is lumbering to some degree with many electrodes attached from the measuring instrument to the subject. Contrasted with the conventional biometric approaches, photoplethysmograph (PPG) system has many focal points, for example, simple to use with no confounded method, low advancement cost and accessible from various locales from the human body like ear lobe, arm, midline of the forehead, forefingertip and wrist [16].

A photoplethysmogram is the representation of optically obtained volumetric changes of blood in the vascular bed in the thumb or index finger by

photoplethysmograph. The physiological parameters, for example, heartbeat and saturated oxygen content in blood, can be obtained by the optical properties of PPG.

In all the proposed frameworks, watermarking strategy is predominantly used to conceal information, making it an effective technique for information stowing away; however, all papers have tended to information covering up biomedical signal, image and electronic patient record. Along with the information concealing system, the biomedical signal of PPG can be utilized as a simple biometric in view of its lower complexity in acquisition and processing techniques.

So the need is to develop a framework which can be summed up for an extensive variety of restorative data to be watermarked, utilization of an easy to acquire biometric and furthermore address the issue of information misfortune with data security.

## 2 Methodology

The proposed method work can be split into 3 as shown in following figures. Figure 1 describes Module 1, Fig. 2 and Fig. 4 describes Module 2 and Module 3, respectively.

### 2.1 Module 1: Generating an Envelope Key

Module 1 involves taking the required data (age, gender, contact address and phone number) of the patient at the doctor’s registration desk. At that point of registration, a unique identification number (UID) is assigned to the patient. Utilizing the contact number, patient UID and their age, the information is converted to binary digits with equal length. In the event of unequal length, zeros are annexed. A secret key is generated, this password is used to protect individual person’s data in the framework,

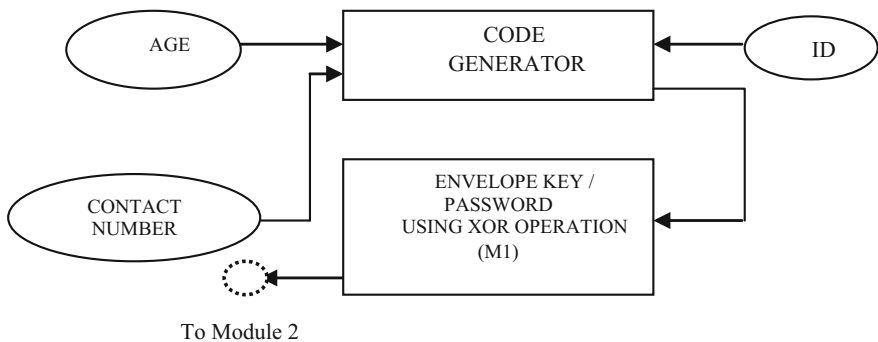
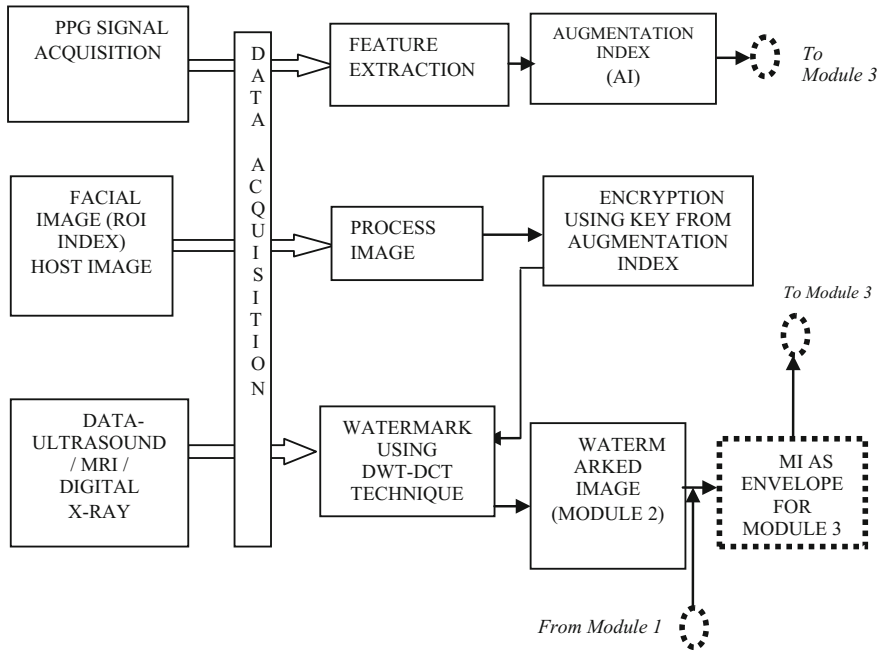


Fig. 1 Envelope key generator—first stage of security



**Fig. 2** Watermarking of data

and furthermore, same password is utilized like a level of security for transmission. The next module depicts usage of PPG as a key for encryption and Least Significant Bit (LSB) technique to watermark patient information with the host image. Matlab ® has been utilized to build up the entire framework. A self-explanatory flow for Module 2 is shown in Fig. 2.

### 2.2 Module 2: Watermarking of Data

The data acquisition system is the main block of the watermarking module as shown in Fig. 2.

The three main inputs to this block are the PPG signal, the facial image of the subject as the host image and the information (ultrasound or MRI images or digital X-ray image) to be watermarked. The diastolic and systolic peaks are extracted from PPG signal to compute Augmentation Index (AI), and the flowchart is shown in Fig. 3. This key gives another level of secure correspondence between the sender and receiver. The inserting of the information as watermark is done by a novel hybrid discrete wavelet transform–discrete cosine transform watermarking algorithm as depicted in Fig. 4. The embedding of watermark is done using the Eq. (1).

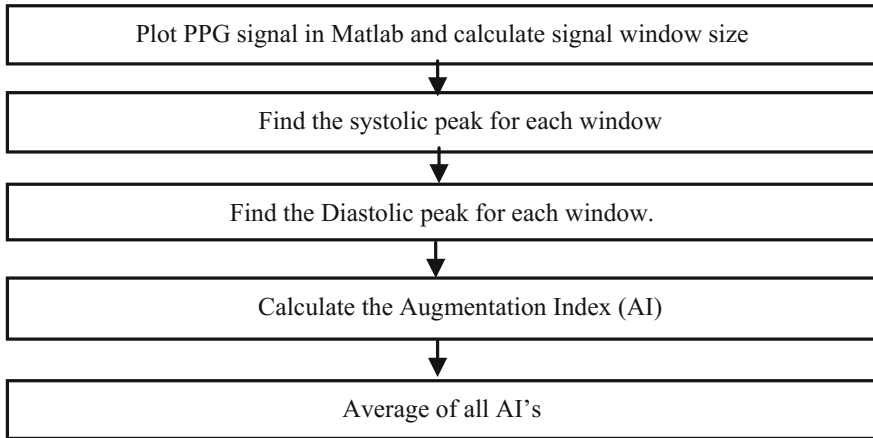


Fig. 3 Computation of augmentation index

$$C = Y + \alpha W \tag{1}$$

where Y is host image, W is watermark, the parameter  $\alpha$  is called embedding intensity and C is the watermarked image.

### 2.3 Module 3: De-Watermarking Using Inverse Hybrid DWT-DCT Watermarking Technique

De-watermarking begins with unfastening the watermarked file with the password as shown in Fig. 5. This follows Eq. 2.

$$W = (C - Y)/\alpha \tag{2}$$

The next step is by using a decoder which will be the decryption key, where the key is drawn from Module 3. As the decryption key matches with the encryption key, the watermarked image is decrypted to give the host facial image and the hidden data.

## 3 Results

This section describes the results obtained by watermarking the electronic patient record on the facial image of the subject for a large dataset of 100 subjects. The distortion caused by watermarking is assessed by using peak signal-to-noise ratio

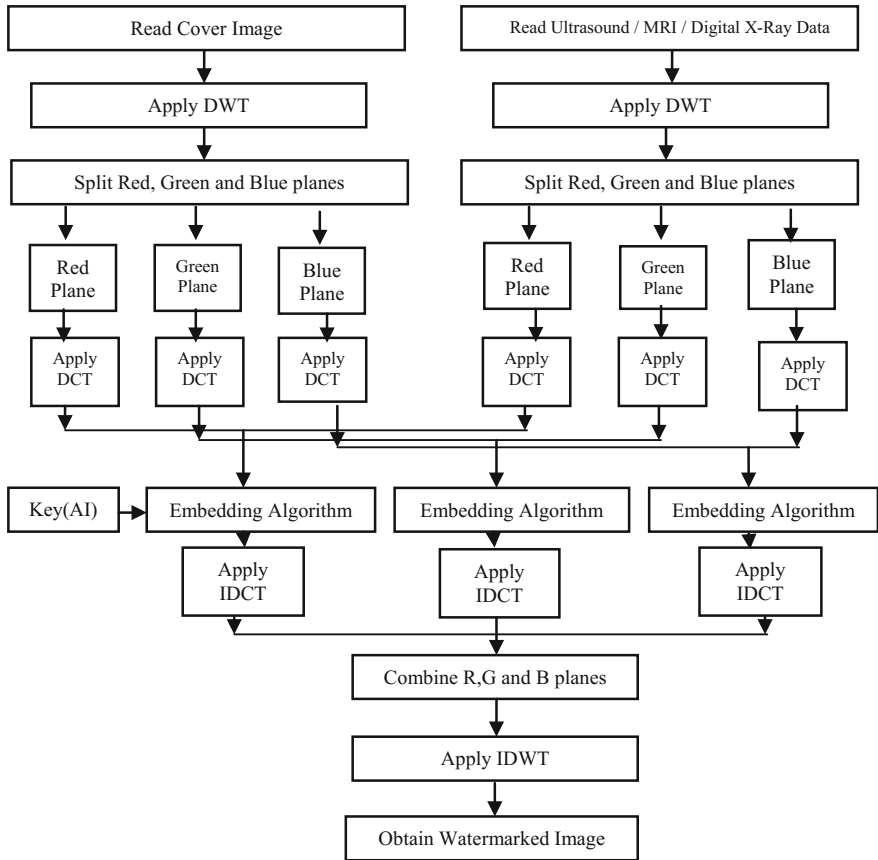


Fig. 4 Hybrid DWT-DCT watermarking technique

(PSNR), Structural Similarity Index Measure (SSIM), Normalized Cross-Correlation (NCR) and Mean Difference (MD). These performance criteria also measure the amount of imperceptibility of watermarking technique (Eq. 2).

Here, Leena.jpeg is presented as the host image instead of subject’s facial image in order to maintain the privacy of the subject. Figures 6 and 7 show the host image and subject’s dental OPG.jpeg image, respectively. A graphical user interface (GUI) has been created for loading the host image and information for the watermarking and de-watermarking process as shown in Fig. 8 and Fig. 9, respectively. Figure 10 shows the watermarked image after applying hybrid DWT-DCT technique. Figures 11 and 12 depict the de-watermarked image and the extracted opg.jpeg data at the receivers end.

Here are the results obtained for different OPG.jpeg images embedded inside Leena.jpeg host image as shown in Table 1.

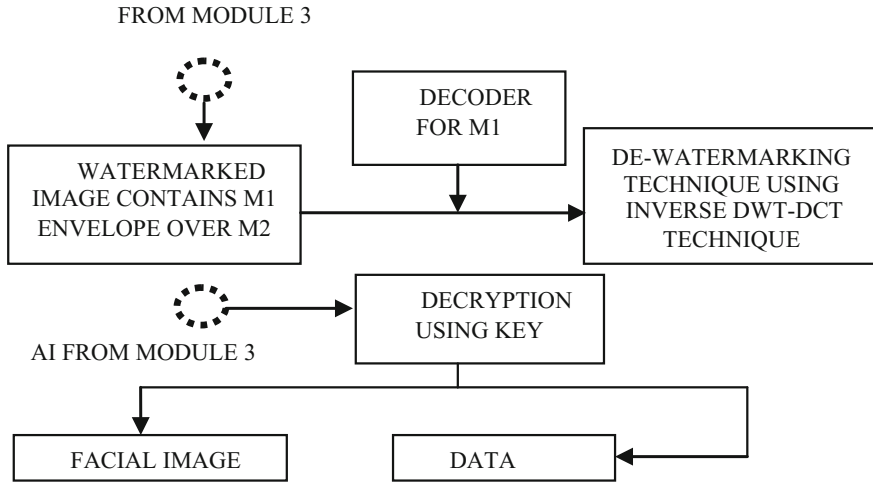


Fig. 5 De-watermarking using hybrid DWT-DCT technique



Fig. 6 Host image: Leena.jpeg

The entire work is conducted on a large dataset of 100 images, and it is found that the hybrid watermarking technique is an efficient method of watermarking an EPR on to the subject’s facial image.

From the results, the hybrid technique has high peak signal-to-noise ratios, which means that the hidden data are more imperceptible. The hybrid technique of watermarking gives a better normalized cross-correlations which relates to better robustness of watermark. The PPG acquisition is feasible, easily accessible and more efficient than any other biometric acquisition techniques.



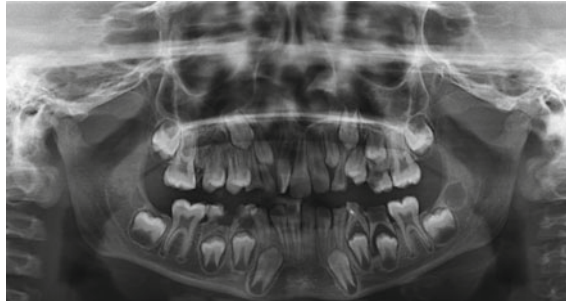


Fig. 7 Digital OPG.jpeg

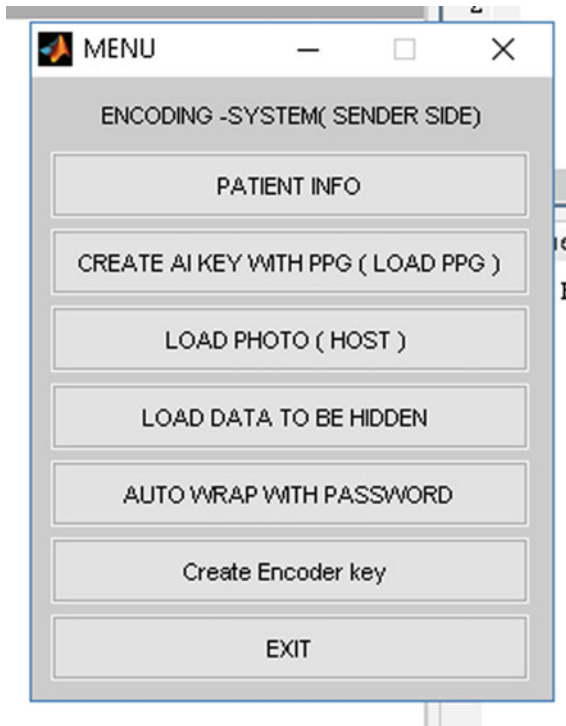


Fig. 8 GUI for sender

## 4 Conclusion

This paper presents frequency domain image watermarking using a hybrid DWT-DCT technique along with the usage of subject's facial image and PPG signal as an authentication key. In the future work, embedding of multiple images with data should be performed with higher imperceptibility and robustness.

Fig. 9 GUI for receiver

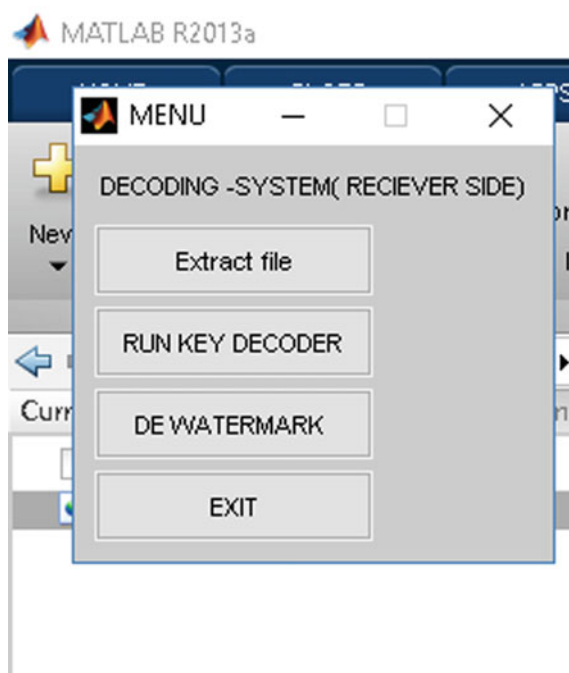
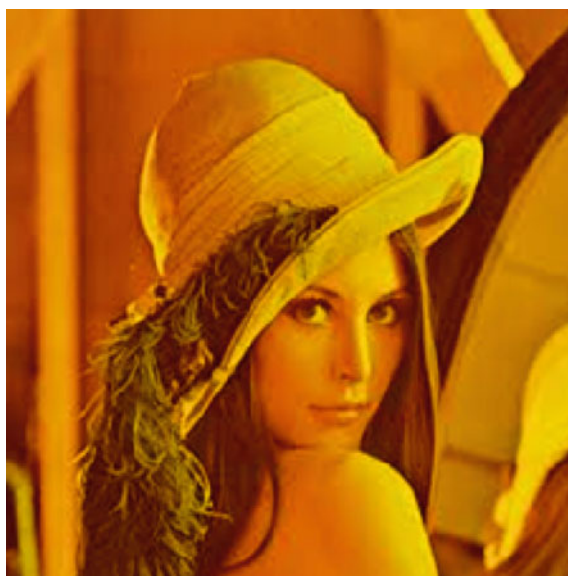


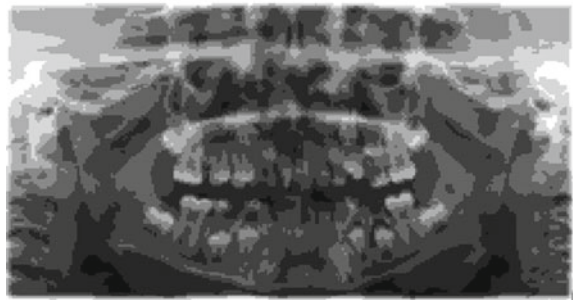
Fig. 10 Watermarked image



**Fig. 11** De-watermarked image



**Fig. 12** Extracted data



**Table 1** Comparison of performance analysis of watermarking technique for  $\alpha = 0.045$

Parameters	OPG1.jpeg	OPG1.jpeg	OPG1.jpeg
MSE	0.0018	0.0017	0.0017
PSNR	63.92	62.92	64.72
MD	11	11	11
SC	0.38	0.36	0.33
NCR	1.46	1.42	1.13

## References

1. Pradeepkumar, G., Usha, S.: Effective watermarking algorithm to protect Electronic patient record using image transform. In: 40th Annual International Conference of the EMBS Minneapolis. IEEE (2014)
2. Organization for Economic Co-operation and Development: Strengthening Health Information Infrastructure For Healthcare Quality Governance: Good practices, New opportunities and Data Privacy Protection Challenges. OECD Publishers (2015)
3. Deloitte: A report on global health care outlook, common goals, Competing Priorities (2017)
4. Thales Data Threat Report: New technologies and old habits driving data breaches and risk in global healthcare. Healthcare Edition (2017)
5. Gran, B., et al.: Watermark compression in medical image watermarking using Lempel-Ziv-Welch (LZW) lossless compression technique. *J. Digit. Imaging PubMedC. Web*, 216–225 (2017)
6. Iftikhar, S., et al.: A reversible watermarking technique for social network data sets for enabling data trust in cyber, physical, and social computing. *IEEE Syst. J.* **11**, 197–206 (2016)
7. Chao, H.M., Hsu, C.M., Miaou, S.G.: A data hiding technique with authentication, integration and confidentiality for electronic patient records. *IEEE TransInf. Technol. Biomed.* **6**, 46–52 (2002)
8. RituAgrawal, M.S.: Medical image watermarking technique in the application of E-diagnosis using M-Ary modulation. *Proced. Comput. Sci.* **85**, 648–655 (2016)
9. Acharya, U.R., Niranjana, U.C., Iyengar, S.S., Kannathal, N., Min, L.C.: Simultaneous storage of patient information with medical images in the frequency domain. *Comput. Methods Program. Biomed.* **76**, 13–19 (2004)
10. Maity, H.K., Maity, S.P.: Joint robust and reversible watermarking for medical images. *Proced. Technol.* **6**, 275–282 (2012) ISSN 2212-0173
11. Brunelli, R., Poggio, T.: Face recognition: features versus templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(10), 1042–1052 (2003)
12. Li, F.F.: Sound-based multimodal person identification from signature and voice. In: 5TH International Conference on Internet Monitoring and Protection(ICIMP) (2010)
13. Nakanishi, I., Baba, S., Miyamoto, C.: EEG based biometric authentication using new spectral features. In: International Symposium on Intelligent Signal Processing and Communication Systems (2009)
14. Kumar, D., Ryu, Y.: A brief introduction of biometrics and fingerprint payment technology. *Int. J. Adv. Sci. Technol.* **4** (2009)
15. Biel, L., Pettersson, O., Philipson, L., Wide, P.: ECG analysis: a new approach in human identification. *IEEE Trans. Instrum. Meas.* **50**(3), 808–812 (2001)
16. Elgendi, M.: On the analysis of fingertip Photoplethysmogram signals. *Curr. Cardiol. Rev.* **8**, 14–25 (2012)

# Hierarchical Autoconfiguration Scheme for IPv6-Based MANETs



T. R. Reshmi

**Abstract** Assigning address to the nodes is a challenging task in MANETs due to the lack of infrastructure and dynamic topology. Even though the existing autoconfiguration schemes ensure unique identities or addresses to the nodes, these schemes fail to guarantee minimal protocol overhead and address acquisition delay. This paper proposes a hierarchical addressing scheme for IPv6-based MANETs that ensures unique addresses with minimal overhead and address acquisition delay. The scheme also uses an address reclamation mechanism that ensures the availability of free addresses to the newly entering nodes. The proposed scheme is implemented in NS-2 and compared with an existing scheme. The results conclude that the scheme outperforms the existing scheme with less address acquisition delay, protocol overhead and packet losses.

**Keywords** Autoconfiguration · Hierarchical addressing · IPv6 · IP address MANETs

## 1 Introduction

The mobile ad hoc networks (MANETs) are self-organizing infrastructure-less networks which forward packets and communicate with other nodes by multi-hop communication. The intermediate nodes between the communicating nodes act as routers to forward the communication packets to destination nodes. The nodes in MANETs require a unique IP address for identification and proper routing of the packets. The nodes are not preconfigured with the IP addresses as the network often requires reconfiguration due to merging and partitioning. To assign the unique addresses, MANETs require an autoconfiguration protocol. The autoconfiguration protocols are broadly divided into two types: stateless and stateful autoconfiguration protocols. The stateless autoconfiguration protocols allow nodes to self-generate its IP

---

T. R. Reshmi (✉)  
VIT University, Chennai, India  
e-mail: reshmi.tr@vit.ac.in

addresses and detect duplicates using the duplicate address detection (DAD) method in the network. But as these protocols do not maintain any database of the already allocated IP addresses, it uses a flooding mechanism to check whether the node's chosen address is already existing or not. If any of the nodes in the network is already assigned with the same address, duplication intimation will be acknowledged to the node. So in case of duplication detected, the node selects another address and repeats the DAD process. The stateful autoconfiguration protocol maintains a database of the allocated addresses and maintains a list of free addresses that can be assigned to newly entering nodes in the MANET. These protocols use dynamic distributed host configuration protocol (DDHCP) servers, as MANETs characterize a dynamically changing topology and hence it is impossible to assign a single node as the centralized DHCP server.

There are two versions of Internet protocols used in current networks: Internet Protocol Version 4 (IPv4) and Internet Protocol Version 6 (IPv6). IPv4 addresses are 32-bit addresses, and  $2^{32}$  addresses are available in its address space. IPv6 addresses are 128-bit addresses, and  $2^{128}$  addresses are available in its address space. Deployment of IPv6 was started as the IPv4 address space was not enough to meet the growing needs of unique IP addresses. The proposed work concentrates on IPv6 protocol, as an initiative toward the next-generation Internet and Internet of Things (IoT). There are different types of flat and hierarchical addressing styles deployed in networks. The flat addressing structure reduces the flexibility of extending the scalability of the protocols. The hierarchical addressing structure eases the deployment and management of IP addresses. In large-scale MANETs, the use of hierarchical addressing structure is beneficial for address assignment and reclamations. The general characteristic features of the autoconfiguration protocols are listed below.

- Unique IP addresses: Each node in the network should be configured with a unique IP address.
- Scalable address space: The address space allocated for distribution to the nodes based on request must be scalable with the increase in nodes.
- Non-integrated routing protocols: The autoconfiguration protocols must be independent of the routing protocols and should not be integrated with routing protocols.
- Reusability: The addresses of the departed nodes must be recovered and reused by the protocol to allocate to the new nodes.
- Reliability of the service: The protocol must overcome the network failure issues.

The address assigning plan can affect the network security and management. The addressing structure of a network defines the fundamental organization and function of a network. The numbering plan, hierarchical addressing to support security segmentation, security implications of EUI-64 addresses, address management and privacy extension in the IPv6 addressing plan reduce the threats to security and management. A hierarchical addressing structure for stateful autoconfiguration in MANETs is the proposal of the paper. The paper highlights the benefit of using a quad-tree structure with four different levels of nodes for address assignment. The proposed protocol ensures the uniqueness of IP addresses based on the stateful infor-

mation stored by its agent node. The protocol overcomes the necessity of duplicate detection mechanism and extra overhead messages. The nodes entering the network self-generate interface address based on its hierarchical level and Sector ID. The address space synchronization which is considered as the tedious mechanism in stateful autoconfiguration is addressed with a lightweight scheme in the proposal. Hence, the computational and communication complexity of the stateful autoconfiguration protocol is reduced and thereby eases the management. The autoconfiguration protocols [1–20] have been evolving from the date of MANET standardization, and the researches are still in progress. All these schemes target MANET characteristics and ensure characteristic requirements of autoconfiguration. The schemes highlight the problem of the limited address space and encourage the use of IPv6 addresses. The paper is organized into five sections. Section 2 describes the algorithm and functioning of the hierarchical autoconfiguration Protocol. Section 3 discusses the performance of the protocol. Section 4 discusses the conclusion on the analysis of the proposed work.

## 2 Proposed Work

A hierarchical autoconfiguration scheme for the IPv6-based MANET has been proposed in the paper. The scheme uses an address structure with two parts: 64-bit Network ID and 64-bit Host ID. The Host ID of the node has two parts: the Sector ID and the Node ID. Each node possesses a disjoint block of addresses which are used to allocate to other nodes. So any neighbor node can act as an address agent and new nodes. The address-requesting node is called as ‘requestor,’ and the node distributing the address is called as ‘allocator’. Each node in the network maintains a table called ‘address information base’ (AIB) containing the hierarchical information about its requestor nodes and the allocator nodes. This AIB is updated during address allocation, network merging and partitioning. Hence, this hierarchical address distribution makes the address management easy and scalable.

### 2.1 System Design

The hierarchical scheme uses Global Positioning System (GPS) to obtain a node’s position such as altitude, latitude and longitude. The scheme uses a geographic forwarding scheme to route packets among nodes. The network topology is divided into many hierarchical grid structures with increasing sizes of squares. The grid structures are represented in terms of sectors. The random combination of lower sector squares cannot construct any higher sector square. Sector-1 square is the smallest unit of the grid. Four Sector-1 squares form Sector-2 square, again four Sector-2 squares form Sector-3 square, and so on. The nodes in the network are categorized into different levels as shown in Fig. 1. The Level 1 node is the root

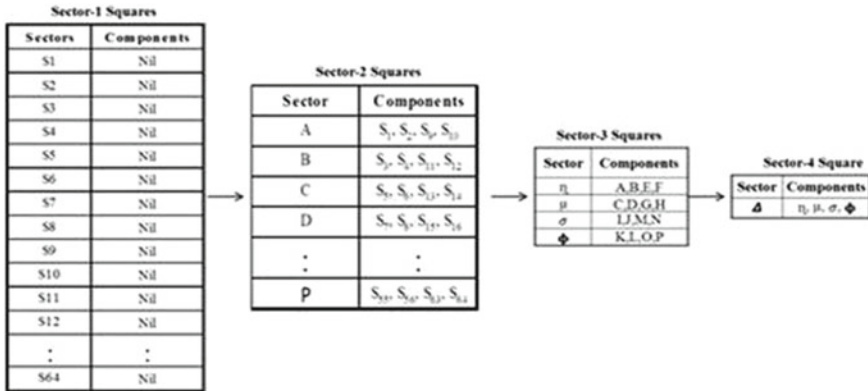


Fig. 1 Hierarchical sectors and components

node of the network and is assigned with the first host address. The Level 2 nodes are the aggregator nodes of the Sector-3 squares. So for each Sector-3 square there exists a Level 2 node which is selected based on the first address of the allocated address space of its geographic location. Likewise, the nodes selected in each level communicate with the hierarchical level nodes for address information exchanges and updates. The IPv6 addresses are identifiers of 128-bit length, and each has the first 64 bits used for network identification and the second 64 bits used for the level and host identification. The Level 1 node also called the root node of the network generates a MANET ID and 64-bit address suffix. The first 2 bits of address suffix represent the Level ID, and usually, the rest of the bits are generated to create the address as the first address in address space.

For the Level 2 nodes, the first 2 bits represent its level, the next 2 bits represent its Sector ID ( $\eta=00, \mu=01, \text{etc.}$ ), and the rest of the bits are randomly generated. For the Level 3 nodes, the first 2 bits represent its level, the next 2 bits represent the next hierarchical Sector ID ( $\eta=00, \mu=01, \text{etc.}$ ), the next 4 bits represent the Sector ID (A = 0000, B = 0001, etc.), and the rest of the bits are randomly generated. For the Level 4 nodes, the first 2 bits represent its level, the next 2 bits represent the second hierarchical Sector ID ( $\eta=00, \mu=01, \text{etc.}$ ), the next 4 bits represent the first hierarchal Sector ID (A = 0000, B = 0001, etc.), and the next 6 bits represent the Sector ID (1 = 000000, 2 = 000001, 3 = 000010, etc.) padded with randomly generated 48 bits.

## 2.2 Address Information Base

Every node maintains a table called ‘address information base’ (AIB) containing the hierarchical information about its requestor nodes and the allocator nodes. The fields in AIB are listed in Table 1.



**Table 1** Address information base (AIB)

Fields	Description
Level_id	Hierarchical level of the node
Free_addr	The number of free address in the block
Req_addr	The number of allocated addresses
Seq_no	The sequence number of the message
Life_time	Life time of the IP address set
Lost_addr	The number of requestor nodes lost from the network
Hierarchical_jumps	The number of hierarchical shifts due to its lost allocators

### 2.3 Operations

Root node configuration: When a node arrives at the MANET, it sends ‘Router Solicitation’ multi-cast messages and waits for a specific period of time. If the node does not receive any ‘Router Advertisement’ messages, the node concludes that there are no preconfigured nodes in the network and declares it as the root node. The root node starts initializing the address by generating the MANET ID, Level ID and Host ID. The root node initiates its AIB parameters as given below.

$$\text{Level\_ID} = 1, \text{ Free\_addr} = 262 - 1, \text{ Req\_addr} = 1, \text{ Seq} = 1, \text{ Life\_time} = 1500 \text{ s}, \\ \text{Lost\_addr} = 0, \text{ Hierarchical\_jumps} = 0$$

Node arrival: The root node configured in the network acts as the allocator to the newly entering nodes in the MANET. If the requestor node receives more than one message from allocators nodes with Free\_addr value > 0 in AIB, then the requestor node selects the node with smallest Level ID as allocator and sends ‘IP\_assign’ message by generating an IP. Table 1 shows the peak signal-to-noise ratio of performance of our proposed method of watermarked image and original image with various watermark images, where our watermarked images’ peak signal-to-noise ratio has a better performance than others. The allocator checks for duplication and acknowledges the conflict. If the requestor does not receive any message until the timer expires, an ‘IP\_conf’ message is sent to the allocator to confirm the address configuration in its interface. The parameters in the AIB of node configured by the Level-1 allocator node are listed below.

$$\text{Level\_ID} = 2, \text{ Free\_addr} = 260 - 1, \text{ Req\_addr} = 1, \text{ Seq} = 1, \text{ Life\_time} = 1500 \text{ s}, \\ \text{Lost\_addr} = 0, \text{ Hierarchical\_jumps} = 0$$

Support for merging: When the nodes are configured with an IP address, the node starts updating their aliveness to the allocator using an ‘Update’ message at regular intervals. So when a node merges with another MANET, it will not receive response from nodes of its own network and detect a MANET ID mismatch with

other nodes. In that case, the node understands that it has merged with another MANET and releases its IP address by sending an 'IP\_release' message and starts reconfiguring its interface. If it is a group of nodes merging in the MANET, the prime allocator (allocator with the lowest Level ID when compared to other nodes) detects the merging as it cannot communicate with its allocator. The prime allocator of the merged nodes checks for its logical hierarchy and compares it with the prime allocator of the other MANET. The allocator with low Level ID is believed to have complex addressing pattern as it manages more number of nodes. So the prime allocator with high Level ID starts initiating an 'IP\_drop' among its requestor nodes and triggers their interfaces for a reconfiguration.

**Support for partitioning:** Nodes leave the network with or without any prior notifications. If the nodes have thorough knowledge about their mobility pattern, then these nodes will be aware when they leaves a network. So these nodes send an 'IP\_release' message and flush the configuration in its interfaces and AIBs. The allocator of the moved node also updates its Lost\_addr field in the AIB to record the departure of one of its requestor nodes. When a partitioning occurs, the nodes detect the cleavage in the network as they are unable to communicate with its allocator. So the allocator of the network with fewer levels initiates the 'IP\_drop' among its nodes and triggers for reconfiguration of its interfaces. In other case, if the higher level agent node departs the network, the IP address is released using an 'IP\_release' message and the next low level node with the smallest IP address is elected as allocator. The Hierarchical\_jump field in its AIB will be incremented by 1 to denote it as the new allocator to the next level nodes.

**Address recovery:** The scheme recovers the addresses of the left-out nodes by explicit or implicit mechanism coordinated by the allocator nodes. When the nodes are configured with an IP address, the node starts updating their aliveness to the allocator using an 'Update' message at regular intervals. So if the allocator does not receive messages from its next level node, it reclaims the address and increments the number of addresses in the Free\_addr of AIB.

### 3 Performance Evaluation

The performance of the proposed autoconfiguration scheme is validated and compared with DDCP scheme [13] using NS 2 [21]. DDCP scheme is an appreciable autoconfiguration scheme implemented in a flat topology and selects the allocator based on available address space. The messages formats used for the simulation are structured as directed in the 'Generalized MANET Packet/Message Format' [9]. The additional information (node's performance parameters, allocated address, key assigned) is included in the type-length-value (TLV) block of the messages. The simulation parameters used for the simulation are summarized in Table 2.

**Table 2** Simulation parameters

Number of nodes	50, 100, 150, 200, 250
Simulation area	1250 × 1250 m <sup>2</sup>
MAC protocol	IEEE 802.11
Mobility model	Random waypoint
Transmission range	250 m
Node mobility	10–50 m/s
Simulation duration	50 s
Traffic source	CBR
Packet size	512 bytes
Routing protocol	AODV

The following metrics are used to evaluate the schemes.

- Protocol overhead: The total number of control and maintenance packet exchanges in the network.
- Address acquisition delay: The time taken by the scheme to configure a node interface with an allocated IP address.
- Packet loss: The total numbers of control and maintenance packets dropped during the packet exchanges of the scheme.
- Address reclamations: The number of addresses reclaimed during the autoconfiguration scheme.

The nodes are deployed in a coverage area of 1250 × 1250 m<sup>2</sup>, and the size of a Sector-1 square is assumed to be 156.25 × 156.25 m<sup>2</sup>. The simulation results of the schemes were plotted with an average of 25 runs. Figure 2 shows the protocol overhead of the autoconfiguration schemes at varying node populations. The results conclude that the hierarchal message exchanges of the proposed scheme reduce the overhead of control and maintenance messages when compared to the flat message exchanges in DDCP [13] scheme. Figure 3 shows the delay for acquiring the addresses in the interface of the MANET nodes during autoconfiguration. The average delay for the address acquisition at varying node population is plotted. The results conclude that the proposed scheme can ensure less delay with minimal packet exchanges and procedure. The guaranteed uniqueness also reduces the time delay in duplicate address detection. The proposed scheme outperforms the existing scheme with the hierarchical message exchanges.

Figure 4 shows the packet losses during the autoconfiguration. The hierarchical scheme has proven to exchange less messages, and hence, the packet losses of the scheme are comparatively low compared to the DDCP [13] scheme. Figure 5 shows the address recovered during the autoconfiguration. The DDCP [13] scheme uses a

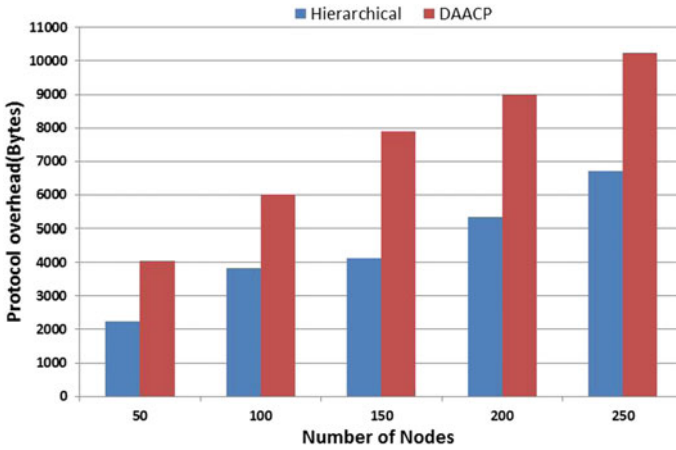


Fig. 2 Protocol overhead of the autoconfiguration schemes

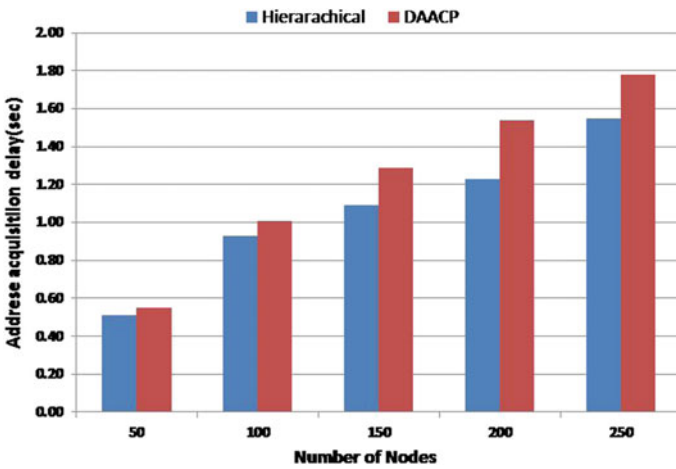


Fig. 3 Address acquisition delay of the autoconfiguration schemes

threshold level for the address space. So when the available free space is reduced below a threshold, it uses a flooded recovery mechanism to reclaim lost and free addresses. The hierarchical scheme uses a time-bounded address reclamation process which avoids flooding at instances. It also ensures high availability of addresses compared to DDCP [13] scheme.

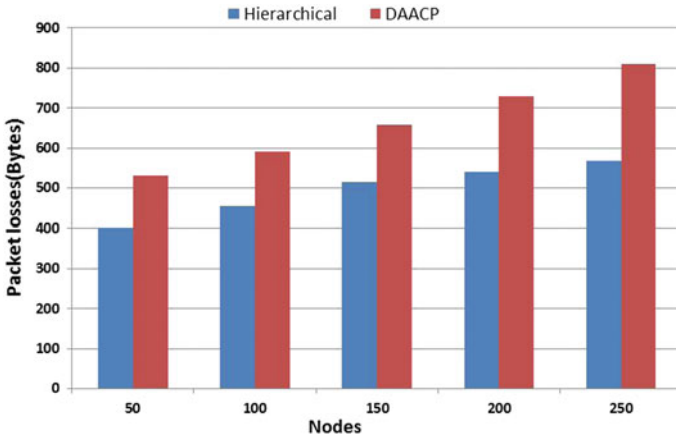


Fig. 4 Packet losses of the autoconfiguration schemes

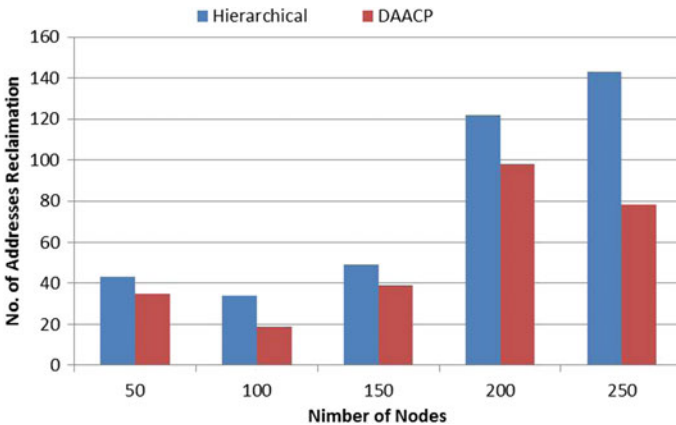


Fig. 5 Address reclamations of the autoconfiguration schemes

## 4 Conclusion

The paper proposed a hierarchical addressing protocol for IPv6-based MANETs. It uses the fixed-size address pool for the address allocation. The number of messages exchanged and the latency are less in the scheme as the messages are not flooded. The messages are initially limited to one-hop and incremented to more hierarchical levels, if the trials are failed. The messages are limited to one-hop neighbors in successful trials. So the messages exchanged are reduced during the working of scheme. The address space management is well defined in hierarchical addressing schemes. The protocol has proven to perform well in small- and large-size networks.

## References

1. Thomson, S., Narten, T.: IPv6 Stateless Address Autoconfiguration, RFC 2462, Dec 1998
2. Clausen, T., Dearlove, C., Dean, J.: Mobile Ad-Hoc Network (MANET) Neighborhood Discovery Protocol (NHDP), Internet Draft, Apr 2011
3. Droms, R., Bound, J., Volz, B., Lemon, T., Perkins, C., Carney, M.: Dynamic Host Configuration Protocol for IPv6 (DHCPv6), RFC 3315, July 2003
4. Perkins, C., Malinen, J., Wakikawa, R., Royer, E., Sun, Y.: IP Address Autoconfiguration for Ad Hoc Networks, Internet Draft, Nov 2001
5. Mohsin, M., Prakash, R.: IP address assignment in a mobile ad hoc network. In: MILCOMM: Military Communication Conference Proceedings, pp. 1–6 (2002)
6. Park, I., Kang, N., Song, H.Y.: Address autoconfiguration for hybrid mobile ad hoc networks, Internet-Draft, MANET Autoconfiguration (AUTOCONF) (2007)
7. Thoppian, M.R., Prakash, R.: A distributed protocol for dynamic address assignment in mobile ad hoc address auto-configuration protocol lha in ad-hoc networks. *IEEE Trans. Mob. Comput.* **5**(1)
8. Perkins, C., Clausen, T.: MANET address autoconfiguration for legacy address pool. Internet Draft, Jan 2011
9. Clausen, T., Dearlove, C., Dean, J., Adjih, C.: Generalized mobile adhoc network (MANET) packet/message format. IETF, RFC 5444, Feb. 2009 hosts, Internet Drafts, Apr 2011
10. Kim, S., Lee, J., Yeom, I.: Modeling and performance analysis of address allocation schemes for mobile ad-hoc networks. *IEEE Trans. Veh. Technol.* (2008)
11. <http://www.theiet.org> (2006). Last accessed April 2006
12. Villanueva, M.J., Calafate, C.T., Torres, A., Cano, J., Cano, J.C., Manzoni, P.: Seamless MANET autoconfiguration through enhanced 802.11 beaconing. *Mob. Inf. Syst.* **9**(1), 19–35 (2013)
13. Gammar, S.M., Amine, E., Kamoun, F.: Distributed address auto configuration protocol for Manets. *Telecommun. Syst.* **44**(1–2) (2010)
14. García Villalba, L.J., Matesanz, J.G., Sandoval Orozco, A.L., Márquez Díaz, J.D.: Distributed dynamic host configuration protocol (D2HCP). *Sensors* (2011)
15. García Villalba, L.J., Matesanz, J.G., Kim, T.H.: E-D2HCP: enhanced distributed dynamic host configuration protocol. *Computing* (2013)
16. Fernandes, N.C., Moreira, M.D.D., Duart, O.C.M.B.: An efficient and robust addressing protocol for node autoconfiguration in ad hoc networks. *IEEE/ACM Trans. Netw.* (2013)
17. Reshmi, T.R., Murugan, K.: Filter-based address autoconfiguration protocol (FAACP) for duplicate address detection and recovery in MANETs. *Computing* 1–23 (2014)
18. Bouk, S.H., Sasase, I.: IPv6 autoconfiguration for hierarchical MANETs with efficient leader election algorithm. *J. Commun. Netw.* (2009)
19. Nazeeruddin, M., Parr, G., Scotney, B.: DHAPM: a new host au-to-configuration protocol for highly dynamic MANETs. *J. Netw. Syst. Manag.* **14**(3) (2006)
20. Bouk, S.H., Sasase, I.: IPv6 autoconfiguration for hierarchical MANETs with efficient leader election algorithm. *J. Commun. Netw.* **11**(3), 248–260 (2009)
21. Network Simulator: <http://www.isi.edu/nsnam/ns>

# Improved $(k, n)$ Visual Secret Sharing Based on Random Grids



Pritam Kumari and Rajneesh Rani

**Abstract**  $(k, n)$  Visual secret sharing (VSS) is a cryptographic procedure in which secret information in the form of images is encoded into ‘ $n$ ’ shares (or shadow images) such that secret is recovered only if ‘ $k$ ’ or more shares are superimposed, whereas superimposing not as much as ‘ $k$ ’ shares provides no information regarding secret. In this paper,  $(k, n)$  VSS based on random grids is proposed which requires no design for codebook and does not involve pixel expansion. The proposed scheme produces better results in terms of improved contrast which decides visual quality. Many experiments are executed to assess the efficacy and security of the proposed approach. The proposed scheme is compared with existing schemes to show its benefits.

**Keywords** Pixel expansion · Random grid · Visual secret sharing · Contrast

## 1 Introduction

Since the Internet applications have started growing, more digital information is accessed and distributed via Internet. But security of such digital information is still a threat. To protect digital information from unauthorized access, various techniques like cryptography, steganography and watermarking are employed. But these involve much computational cost in the decryption phase. So, optimal solution for the protection of digital images is visual secret sharing (VSS). This scheme does not require complex computations to decode the secret.

Naor and Shamir [1] presented  $(k, n)$  VSS technique. In  $(k, n)$  VSS, the secret visual information is encoded into ‘ $n$ ’ random and innocent shares (also called

---

P. Kumari (✉) · R. Rani  
Department of Computer Science and Engineering, Dr. B. R. Ambedkar National  
Institute of Technology, Jalandhar 144011, Punjab, India  
e-mail: pritamkumari07@gmail.com

R. Rani  
e-mail: ranir@nitj.ac.in

shadow images), with each share given to one participant. Shares are xeroxed onto transparencies which separately do not disclose any useful data regarding the secret. Human visual system (HVS) can easily recover the secret when any 'k' or more transparencies are superimposed together, whereas stacking transparencies less than 'k' provides no data about the secret. The visual quality increases as the number of shares to be stacked increases. Unfortunately, this scheme requires codebook design, also expands pixel and gives the poor visual quality of the revealed secret.

Kafri and Keren [2] investigated the idea of sharing images (binary) securely using a novel approach of random grids (RG) in 1987: encrypting secret into two innocent looking random grids (shadow images) which are as large as the original secret image. The decoding phase recovers secret by stacking two random grids together. This RG-based VSS requires no codebook designing and eliminates the pixel expansion problem.

The remaining portion of the proposed paper is structured as follows: related work and short explanation of traditional RG-based VSS is given in Sect. 2. Section 3 introduces the proposed  $(k, n)$  VSS based on random grids, explains some definitions and states assumptions. Results based on experiments and comparisons with existing schemes are presented in Sect. 4, while Sect. 5 gives the conclusion of paper.

## 2 Related Work

VSS schemes based on general access structures were presented [3] to give more adaptable sharing system. Many researchers tried to extend Naor and Shamir's work to improve the visual quality [4, 5] and reduce pixel expansion [6, 7]. Scheme [7] eliminates the pixel expansion problem completely. Conventional VSS makes use of random shares which attract the hacker's attention. So, to solve this problem, extended visual secret sharing [8, 9] was developed which uses some innocent looking images as shares rather than random shares. It also tackles the issue which arises during managing non-meaningful shares. Many researchers [10, 11] introduced Halftone visual cryptography (HVC) which creates halftone shares consisting of significant information.

Shyu [12] broadened Kafri and Karen's approach to produce VSS based on random grids which encodes color images and grayscale images using halftoning. Schemes [2, 13, 14] reduce pixel expansion but are not meant for generalized  $(k, n)$  VSS. Shyu [14] presented  $(n, n)$  visual cryptography sharing based on random grids. Chen and Tsao [13] used the concept of random grids to demonstrate  $(2, n)$  and  $(n, n)$  VSS with its ability to encode binary as well as color images. Chen and Tsao [15] gave a new method to formulate  $(k, n)$  threshold VCS schemes by RG. Guo et al. [16] presented  $(k, n)$  VSS using concept of RG which helps increase the visual quality as long as  $k$  is less than or equal to  $n/2$ . The proposed paper introduces  $(k, n)$  random grid-based VSS scheme. The original secret can be revealed with improved contrast by stacking (Boolean OR) sufficient number of shares together by HVS.



Traditional (2, 2) RG-based VSS [15], which is used in our proposed scheme, is explained below. A random grid [2] is characterized as a rectangular matrix of pixels in which every element is pixel and can have one of two possible values, i.e., ‘0’ meaning white or transparent pixel and ‘1’ meaning black or opaque pixel. Each pixel has an equal probability of being transparent or opaque.  $\otimes$  and  $\oplus$  denote Boolean OR and XOR operations, respectively.

*Traditional (2, 2) RG-based VSS*

*Input:* Binary secret image  $SI$  with size  $r \times c$

*Output:* Two random grids  $RS_1$  and  $RS_2$

1. Create the first random grid  $RS_1$  by randomly choosing each pixel to be either ‘0’ or ‘1.’
2. Compute the second random grid  $RS_2$  as given in Eq. (1) for each secret pixel  $SI(p, q)$

$$RS_2(p, q) = \left\{ \begin{array}{l} RS_1(p, q), \text{ if } SI(p, q) = 0 \\ \overline{RS_1(p, q)} \text{ if } SI(p, q) = 1 \end{array} \right\} \tag{1}$$

In recovery phase, recovered secret image  $RI$  is obtained by stacking (Boolean OR) of both shares  $RS_1$  and  $RS_2$  as given in Eq. (2).

$$\begin{aligned} RI(p, q) &= RS_1(p, q) \otimes RS_2(m, n) \\ &= \left\{ \begin{array}{l} RS_1(p, q) \otimes RS_1(p, q) = RS_1(p, q), \text{ if } SI(p, q) = 0 \\ RS_1(p, q) \otimes \overline{RS_1(p, q)} = 1, \text{ if } SI(p, q) = 1 \end{array} \right. \end{aligned} \tag{2}$$

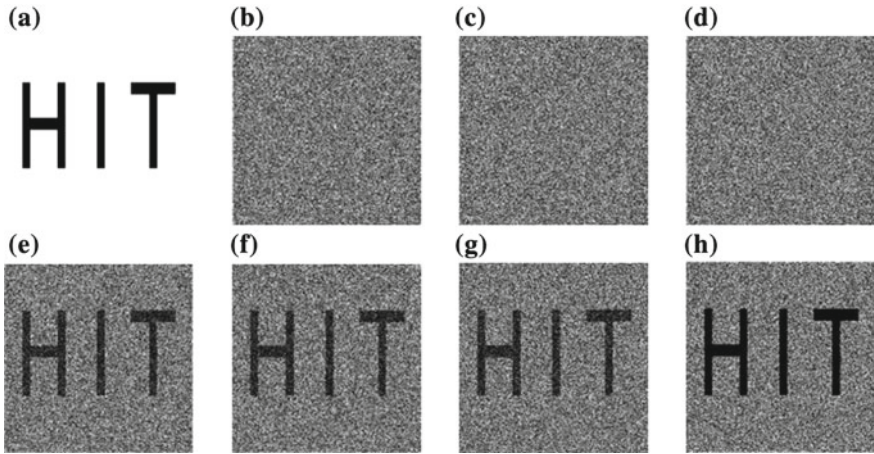
If secret pixel is black, i.e.,  $SI(m, n) = 1$ , then the recovered image bit is always black, whereas the recovered bit has half probability to be white or black if  $SI(m, n) = 0$  because  $RS_1$  is generated randomly.

### 3 The Proposed Scheme

This section introduces (k, n) RG-VSS ( $2 \leq n \leq 5$  and  $2 \leq k \leq n$ ) which uses the concept of RG and stacking operation and also explains some definitions which are used in the proposed work.

In order to analyze VSS based on RG, few definitions have been borrowed from [12, 13] which are explained below (Fig. 1):

**Definition 1** (Average light transmission [12, 13]) Let  $L(i)$  refer to the light transmission of a pixel ‘i’ in binary secret image  $SI$  having size  $h \times w$  which is ‘1’ for



**Fig. 1** Implementation results of the proposed (2, 3) case. **a** Binary image *Text*, **b** Share  $RS_1$ , **c** Share  $RS_2$ , **d** Share  $RS_3$ , **e**  $RS_1 \otimes RS_2$ , **f**  $RS_1 \otimes RS_3$ , **g**  $RS_2 \otimes RS_3$ , **h**  $RS_1 \otimes RS_2 \otimes RS_3$

transparent pixel and ‘0’ for opaque pixel. The average light transmission of *SI* is computed as

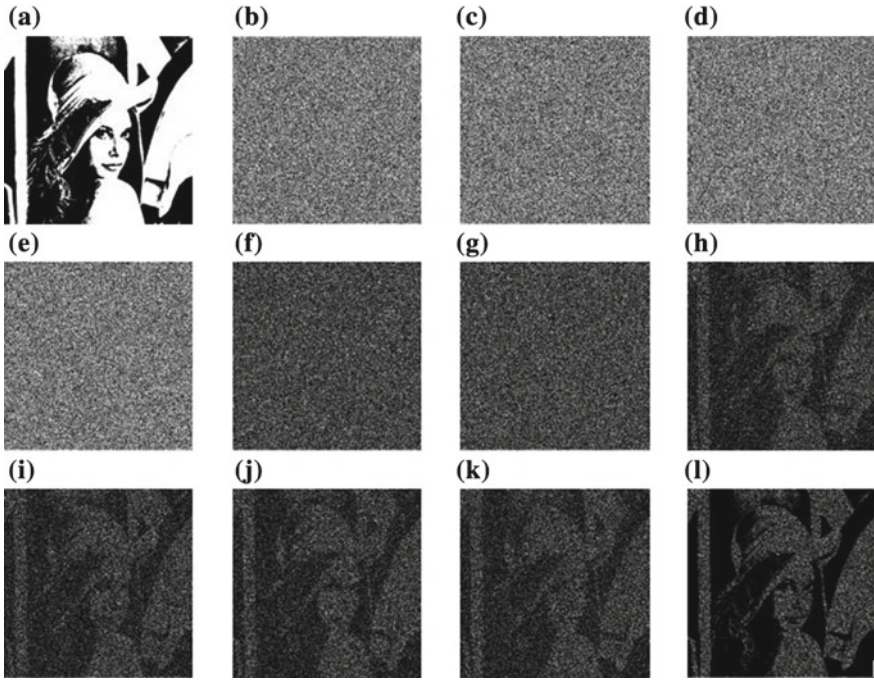
$$L(SI) = \frac{\sum_{i=1}^{i=h} \sum_{j=1}^{j=w} L(SI(i, j))}{h \times w} \tag{3}$$

**Definition 2** (*Contrast* [12]) Given the original secret image *SI*, the superimposed image *RI* has contrast ( $\alpha$ ) given in Eq. (4), as follows

$$\alpha = \frac{L(RI[SI(0)]) - L(RI[SI(1)])}{1 + L(RI[SI(1)])} \tag{4}$$

where  $RI[SI(1)]$  (respectively  $RI[SI(0)]$ ) is that portion in the recovered image *RI* that corresponds to all the black (respectively white) pixels of *SI*. Contrast is one of important metrics which evaluates the image quality of recovered image. Large contrast is desirable due to fact that recovered secret will be easily recognized by HVS if the stacked result has large  $\alpha$ .

**Definition 3** (*Visually recognizable* [13]) The original image *SI* and the recovered image *RI* are more similar if the contrast of recovered image is greater than 0. The condition  $L(RI[SI(0)]) > L(RI[SI(1)])$  implies that *RI* can be recognized as *SI* by HVS.



**Fig. 2** Implementation results of the proposed (3, 4) case. **a** Binary secret image *Lena*, **b** Share  $RS_1$ , **c** Share  $RS_2$ , **d** Share  $RS_3$ , **e** Share  $RS_4$ , **f**  $RS_1 \otimes RS_2$ , **g**  $RS_3 \otimes RS_4$ , **h**  $RS_1 \otimes RS_2 \otimes RS_3$ , **i**  $RS_1 \otimes RS_2 \otimes RS_4$ , **j**  $RS_1 \otimes RS_3 \otimes RS_4$ , **k**  $RS_2 \otimes RS_3 \otimes RS_4$ , **l**  $RS_1 \otimes RS_2 \otimes RS_3 \otimes RS_4$ .

### 3.1 Shares Generation Phase and Secret Recovery Phase

Shares generation phase is shown in Fig. 2 which generates ‘n’ shares  $RS_1, RS_2, \dots, RS_n$ . The steps of shares generation phase of the proposed scheme are described below:

#### Shares Generation Algorithm

*Input:* Binary secret image  $SI$  of size  $r \times c$

*Output:* ‘n’ shadow images  $RS_1, RS_2, \dots, RS_n$ .

Repeat steps 1–3 for every secret pixel  $SI(m, n)$ ,

1. Utilize traditional (2, 2) RG-based VSS to encrypt a secret pixel  $SI(m, n)$  to generate two bits  $b_1$  and  $b'_2$ . Encode  $b'_2$  in the similar way as two bits  $b_2$  and  $b'_3$  are generated, similarly encode  $b'_3$  into  $b_3$  and  $b'_4$ . Repeat this operation until  $b_1, b_2, b_3, \dots, b_{k-1}, b'_k$  are produced (the last bit  $b'_k$  is same as  $b_k$ ).
2. Generate remaining ‘n – k’ bits, i.e.,  $b_{k+1}, b_{k+2}, \dots, b_n$  by assigning each of them to  $b_k$ .

3. Randomly arrange the 'n' bits  $b_1, b_2, b_3, \dots, b_n$  and assign the rearranged bits to shares  $RS_1(m, n), RS_2(m, n), \dots, RS_n(m, n)$ .
4. Output 'n' shadow images  $RS_1, RS_2, \dots, RS_n$ .

In step 1 of the above algorithm, the traditional (2, 2) RG-VSS is applied repeatedly 'k' times to generate 'k' bits. Remaining 'n - k' bits are obtained by equating them to kth bit which is shown in step 2. In step 3, these 'n' bits are randomly rearranged to show that all shares are of equal importance. In the proposed approach, we construct (k, n) RG-VSS from step 1 given in the above algorithm.

### 3.2 Extending Grayscale Images

The proposed algorithm encrypts grayscale images by applying halftoning like dithering, error diffusion [13, 15] which takes grayscale image and produces binary halftoned image. After that, the proposed algorithm is implemented on halftoned binary image.

### 3.3 Assumptions

The assumption [17] given below must be met for showing that the proposed VSS is valid construction.

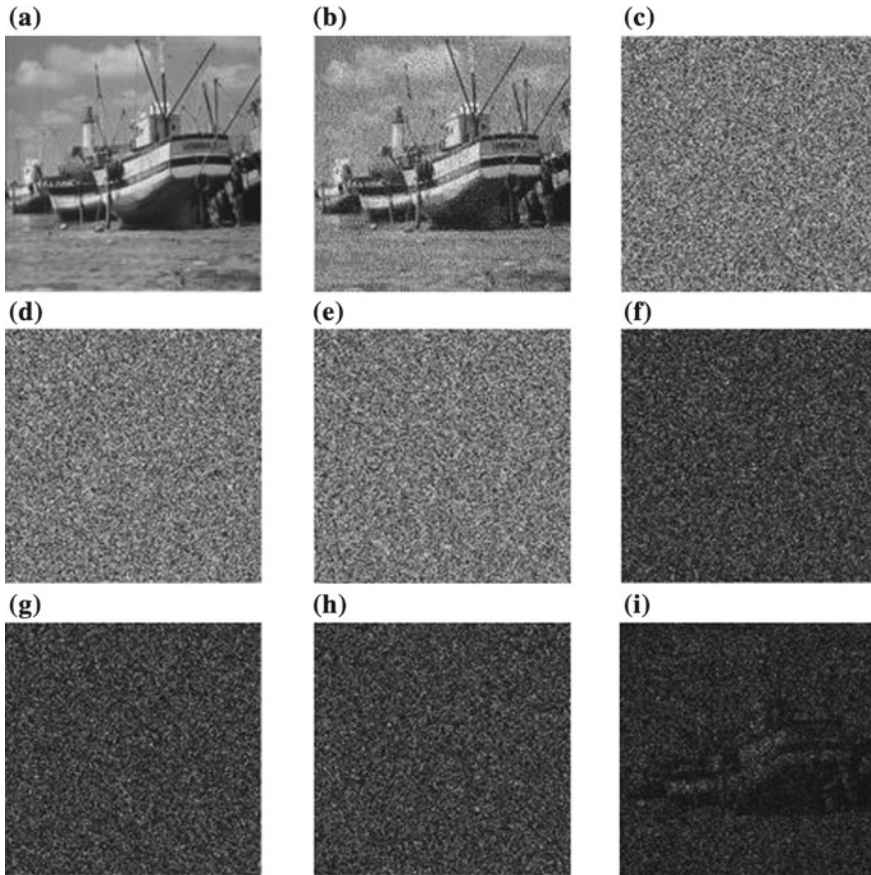
*Assumption:* The following three conditions are fulfilled to claim the RG-based VSS as a valid construction:

1. Each shadow image is a random grid and does not produce any data about the secret image:  $L(R_t[SI(0)]) = L(R_t[SI(1)])$  where  $2 \leq t \leq n$ .
2. If  $t < k$ , the stacked result  $R_{1 \otimes 2 \otimes 3 \dots \otimes t} = R_1 \otimes R_2 \otimes \dots \otimes R_t$  provides no hint about the secret:  $L(R_{1 \otimes 2 \otimes 3 \dots \otimes t}[SI(0)]) = L(R_{1 \otimes 2 \otimes 3 \dots \otimes t}[SI(1)])$ .
3. If  $t \geq k$ , the stacked result  $R_{1 \otimes 2 \otimes 3 \dots \otimes t} = R_1 \otimes R_2 \otimes \dots \otimes R_t$  reveals the secret image:  $L(R_{1 \otimes 2 \otimes 3 \dots \otimes t}[SI(0)]) > L(R_{1 \otimes 2 \otimes 3 \dots \otimes t}[SI(1)])$ .

The first condition ensures that the individual share is not capable of delivering any information related to secret image. The second condition claims that stacking inadequate number of shares reveals no hint about the secret information. The third condition says that only stacking sufficient number of shadow images reconstructs the secret image. These assumptions are proved by conducting experiments in the next section.

### 4 Experimental Results and Analyses

We have presented results for the experiments we conducted on images in this section. In the proposed (k, n) random grid-based VSS scheme, experiments are implemented for  $2 \leq n \leq 5$  and  $2 \leq k \leq n$ . Several images of size  $512 \times 512$  like binary secret image *Text* given in Fig. 1a, image *Lena* in Fig. 2a, grayscale secret image *Boat* given in Fig. 3a are taken as input to do experiments to analyze the performance of the proposed scheme.



**Fig. 3** Implementation results of proposed (3, 3) case. **a** Grayscale secret image *Boat*, **b** Halftoned binary image, **c** Share  $RS_1$ , **d** Share  $RS_2$ , **e** Share  $RS_3$ , **f**  $RS_1 \otimes RS_2$ , **g**  $RS_1 \otimes RS_3$ , **h**  $RS_2 \otimes RS_3$ , **i**  $RS_1 \otimes RS_2 \otimes RS_3$

## 4.1 Image Illustration

In Fig. 1, experiment results carried out for (2, 3) (i.e.,  $k=2$ ,  $n=3$ ) VSS scheme are given. Figure 1a shows the original binary secret image *Text*. Figure 1b–d shows the three shares generated. These shares individually cannot disclose anything about the secret image. Reconstructed secret is shown in Fig. 1e–g which is generated by stacking any two shares. Stacked result when all shadows are collected is presented in Fig. 1h.

In Fig. 2, experiment results conducted for (3, 4) (i.e.,  $k=3$ ,  $n=4$ ) VSS scheme are given. In Fig. 2a, the secret image *Lena* is shown, and Fig. 2b–e shows the four shares created. Figure 2f–g shows the two of stacked results when any two shares are stacked together. Stacking when any three shares are collected is displayed in Fig. 2h–k. Figure 2l shows the stacked result when all four shares are collected.

In our experiments, results for (3, 3) (i.e.,  $k=3$ ,  $n=3$ ) VSS scheme conducted on grayscale image *Boat* are shown in Fig. 3a. Halftoned binary image is given in Fig. 3b which is obtained by applying halftoning on the grayscale image *Boat*. Three shadow images which are generated are demonstrated in Fig. 3c–e. Stacked results when any two shadow images are stacked are presented in Fig. 3f–h. Figure 3i shows the stacking on all three shares.

The following observations can be made from experiments conducted for (k, n) VSS:

- Every share is noise-like random image.
- If  $t \geq k$ , recovered secret image is visually recognizable, where ‘t’ is the number of stacking shares.
- When  $t < k$ , recovered secret gives no information about the original secret image.
- Contrast decreases when we increase the value of ‘n.’ Therefore, we assume the maximum value for ‘n’ is 5.
- The proposed scheme is also applicable for grayscale images.

## 4.2 Comparison with Related Schemes

This section shows the comparison of the proposed scheme with related schemes to show the advantages of proposed approach. Like [18, 15, 19, 20], the proposed RG-based VSS also benefits from no codebook designing and removes the problem of pixel expansion.

### 4.2.1 Contrast Comparison

Contrast in definition 2 evaluates the image quality for reconstructed secret. We have compared the proposed algorithm with scheme [20] in terms of average contrast.



**Table 1** Average contrast of the proposed approach and [20] for binary images *Text* and *Lena*

(k, n)	Average contrast of [20]				Average contrast of proposed scheme			
	t = 2	t = 3	t = 4	t = 5	t = 2	t = 3	t = 4	t = 5
(2, 2)	0.50041				0.50190			
(2, 3)	0.12578	0.50104			0.28631	0.49990		
(3, 3)		0.24950				0.25040		
(2, 4)	0.06594	0.11084	0.25024		0.19918	0.33299	0.49953	
(3, 4)		0.05220	0.24974			0.11097	0.25002	
(4, 4)			0.12508				0.12535	
(2, 5)	0.04067	0.06800	0.07117	0.12465	0.15456	0.25074	0.36389	0.49980
(3, 5)		0.02243	0.04682	0.12477		0.06214	0.13541	0.25025
(4, 5)			0.02278	0.12463			0.04620	0.12488
(5, 5)				0.06186				0.06290

**Table 2** Comparison of features of previous related schemes with the proposed scheme

Schemes	Type of VSS	Pixel expansion Problem	Involves codebook Design	Recovering measure
Ref. [1]	(k, n)	✓	✓	Stacking
Ref. [15]	(k, n)	×	×	Stacking
Ref. [10]	(k, n)	✓	✓	Stacking
Ref. [19]	(k, n)	×	×	Boolean
Ref. [18]	(n, n), (2, n)	×	×	Boolean
Ref. [20]	(k, n)	×	×	Boolean
Ours	(k, n)	×	×	Stacking

Table 1 demonstrates the results of contrast of the decoded secret for the scheme [20] and the proposed scheme. The results show that:

- Contrast comes out to be greater than 0 as t increases than k. Here, ‘t’ represents the number of shares to be stacked.
- The average contrast of the proposed algorithm exceeds the average contrast given by scheme [20].

### 4.2.2 Features Comparison

We have compared the proposed schemes with related schemes with respect to some features. It is listed in Table 2. Benefits of the proposed RG-based VSS over related schemes are displayed in Table 2.

## 5 Conclusion

This paper presented  $(k, n)$  random grid-based VSS which could be adopted for encryption of binary or grayscale images. The shadow images are copied onto transparencies. Secret image could be well identified by HVS when 'k' or more transparencies are superimposed together. One cannot get any useful data about the secret on stacking not as much as 'k' shares. Compared with [20], the proposed approach benefits from the greater visual quality in terms of the large contrast. The proposed scheme has advantages of no codebook designing, avoiding pixel expansion, having same importance for all shares. In the future, there may be further improvement in the visual quality of the revealed secret. This scheme could be stretched to encrypt color images. Also, meaningful shares can be generated instead of random shares. There is still scope for improvement in security of shadow images.

## References

1. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
2. Kafri, O., Keren, E.: Encryption of pictures and shapes by random grids. *Opt. Lett.* **12**(6), 377–379 (1987)
3. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Visual cryptography for general access structures. *Inf. Comput.* **129**(2), 86–106 (1996)
4. Blundo, C., Bonis, A.D., Santis, A.D.: Improved schemes for visual cryptography. *Des. Codes Cryptogr.* **24**(3), 255–278 (2001)
5. Blundo, C., Arco, P.D., Stinson, D.R.: Contrast optimal threshold visual cryptography schemes. *SIAM J. Discret. Math.* **16**(2), 224–261 (2003)
6. Ito, R., Kuwakado, H., Tanaka, H.: Image size invariant visual cryptography. *IEICE Trans. Fundam.* **E82-A**(10), 2172–2177 (1999)
7. Yang, C.N.: New visual secret sharing schemes using probabilistic method. *Pattern Recogn. Lett.* **25**, 481–494 (2004)
8. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Extended capabilities for visual cryptography. *Theor. Comput. Sci.* **250**(1), 143–161 (2001)
9. Liu, F., Wu, C.: Embedded extended visual cryptography schemes. *IEEE Trans. Inf. Forensics Secur.* **6**, 07–322 (2011)
10. Zhou, Z., Arce, G.R., Di Crescenzo, G.: Halftone visual cryptography. *IEEE Trans. Image Process.* **15**(8), 2441–2453 (2006)
11. Wang, Z., Arce, G.R., Di Crescenzo, G.: Halftone visual cryptography via error diffusion. *IEEE Trans. Inf. Forensics Secur.* **4**(3), 383–396 (2009)
12. Shyu, S.: Image encryption by random grids. *Pattern Recognit.* **40**(3), 1014–1031 (2007)
13. Chen, T., Tsao, K.: Visual secret sharing by random grids revisited. *Pattern Recognit.* **42**(9), 2203–2217 (2009)
14. Shyu, S.: Image encryption by multiple random grids. *Pattern Recognit.* **42**(7), 1582–1596 (2009)
15. Chen, T., Tsao, K.: Threshold visual secret sharing by random grids. *J. Syst. Softw.* **84**(7), 1197–1208 (2011)
16. Guo, T., Liu, F., Wu, C.: Threshold visual secret sharing by random grids with improved contrast. *J. Syst. Softw.* (2013). (in press)



17. Wu, X., Sun, W.: Improving the visual quality of random grid-based visual secret sharing. *Signal Process.* **93**(5), 977–995 (2013)
18. Wang, D., Zhang, L., Ma, N., Li, X.: Two secret sharing schemes based on Boolean operations. *Pattern Recognit.* **40**(10), 2776–2785 (2007)
19. Wu, X., Sun, W.: Random grid-based visual secret sharing with abilities of OR and XOR decryptions. *J. Vis. Commun. Image Represent.* **24**(1), 48–62 (2013)
20. Yan, X., Wang, S., Niu, X., Yang, C.N.: Random grid-based visual secret sharing with multiple decryptions. *J. Vis. Commun. Image R.* **26**, 94–104 (2015)

# Efficient Motion Encoding Technique for Activity Analysis at ATM Premises



Prateek Bajaj, Monika Pandey, Vikas Tripathi and Vishal Sanserwal

**Abstract** Automated teller machines (ATMs) have become the predominant banking channel for the majority of customer transactions. However, despite the multitudinous advantages of ATM, it lacks in providing security measures against ATM frauds. Video surveillance is one of the prominent measures against ATM frauds. In this paper, we present an approach that can be used for activity recognition in small premises such as ATM rooms by encoding the motion in images. We have used gradient-based descriptor (HOG) to extract features from image sequences. The features obtained are classified using random forest classifier. Our employed method is successful in determining abnormal and normal human activities both in case of single and multiple personnel with an average accuracy of 97%.

## 1 Introduction

The goal of computer vision is to facilitate the machine to interpret the world through the process of digital signal [1]. Various technologies such as motion detection and facial recognition are based on computer vision. Automating the video surveillance with the help of computer vision to detect any suspicious activity or personnel is an effective way to the cover up some flaws in the security. Video surveillance detects moving object through a sequence of images [2, 3]. ATM surveillance is a sub-domain of video surveillance. ATM crime has become one of the most prominent

---

P. Bajaj (✉) · M. Pandey · V. Tripathi · V. Sanserwal  
Department of Computer Science and Engineering, Graphic Era University,  
Dehradun 248002, Uttarakhand, India  
e-mail: prateekbajaj552@gmail.com

M. Pandey  
e-mail: monikapandey234@gmail.com

V. Tripathi  
e-mail: vikastripathi.be@gmail.com

V. Sanserwal  
e-mail: vishuchaudhary28@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_36](https://doi.org/10.1007/978-981-13-1708-8_36)

issues nationwide [4] as they are at the public places and vulnerable to thefts. The usual security measure in an ATM system is CCTV which is not automated and requires authority to monitor them. The slow response time of a CCTV is a reason for its under-efficiency and adds to the vulnerability of security. Automated surveillance system detects any unusual activity in their frame view and automatically takes the desired actions [5]. In a recent survey based on video surveillance, Cucchiara [6] reported that there are various problems than hinder motion detection in non-ideal conditions. Various techniques that have been used for motion analysis using automated systems are based on the framework of temporal templates and spatiotemporal templates optical flow, background subtraction, silhouettes, histograms and several others [7–10]. In this paper, we have further extended [11] by introducing a motion encoding technique called motion identifying image (MII). In MII, we have incorporated root-mean-square of thresholded images. We have analyzed four categories of human actions which are classified from a single camera view. They are single, single abnormal, multiple and multiple abnormal. The paper is further organized in the following manner: Sect. 2 reviews the recent work; Sect. 3 describes the methodology we have used; Sect. 4 gives results and analysis; and Sect. 5 concludes the paper.

## 2 Literature Review

Video surveillance has contributed to the enhancement of security and protection in every possible field [12]. There are various ways to detect an activity in computer vision. In this section, we present the previous work conducted to improve video surveillance. Several approaches have been presented to recognize human actions. Davis and Bobick [13] have used temporal templates using motion history image (MHI) and motion energy image (MEI) for recognizing human activity. The temporal approaches utilize vector images where each vector points motion in the image [14]. Directional motion history image (DMHI) is an extension of MHI introduced by Ahad et al. [15, 16]. Poppe [17] has presented a detailed overview of human motion analysis using MHI and its variants. Al-Berry et al. [18], motivated by MHI, introduced a stationary wavelet-based action representation, which has been used to classify variant actions. There are various descriptors such as spatiotemporal interest feature points (STIPs), histograms of oriented flow (HOF) and histograms of oriented gradients (HOGs) which are used to compute and represent actions. Space–time interest point (STIP) detectors are extensions of 2D interest point detectors that incorporate temporal information. HOG is a window-based descriptor which is used to compute interest points. Further, the window is divided into a grid of ( $n * n$ ). Frequency histogram is generated from each cell of the grid to show edge orientation in the cell [19], whereas the descriptor HOF gives information using optical flow [20]. Another descriptor named Hu moments extracts interest points based on shape, independent of position, size and orientation of the image [21], and since it is a shape descriptor, it requires comparatively less computation [22–24]. Zernike moments

descriptor is another shape descriptor which is more efficient than Hu moments [21]. Sanserwal et al. [25] in their paper have proposed algorithm in which they have used HOG descriptor, Hu moments and Zernike moments descriptor for activity detection from a single viewpoint [26] Vikas et al. proposed an approach that makes use of motion history image and Hu moments to extract features from a video. Rashwan et al. [27] proposed optical flow model with new robust data obtained from histogram of oriented gradients (HOGs) computed between two consecutive frames. But the approaches such as HOG can be highly computational [28]. Huang and Huang [29] in his paper uses look-up table along with the method of integral image to speed up HOG. Uijlings et al. [30] proposed a framework that can increase the efficiency of densely sampled HOG, HOF and MBH (motion boundary histograms) descriptors. Ryan Kennedy and Camillo J. Taylor used a method in which optical flow is calculated over triangulated images [31]. In our approach, we have used three consecutive frames to encode motion into image which is then provided to gradient-based descriptor HOG. We have described that our framework can effectively recognize ATM events.

### 3 Methodology

The proposed methodology makes use of computer vision-based framework to detect normal and abnormal activities in indoor premises such as ATM room. Figure 1 represents working of our framework. It shows that the method consists of the camera feed in the form of video, which is converted into threshold images. Our framework consists of two parts, conversion of an image into encoded motion using MII and conversion of encoded image into features using a descriptor. MII involves preprocessing the thresholded images using root-mean-square formula. The features thus obtained are classified using random forest classifier. The algorithmic representation of our framework is shown in Fig. 2.

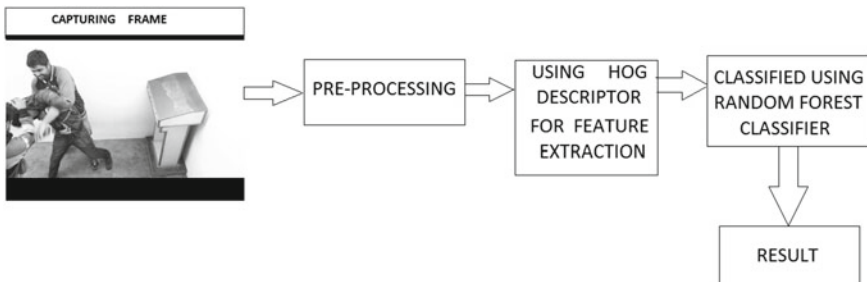


Fig. 1 Architecture for the proposed method

<ol style="list-style-type: none"> <li>1. Compute Thresholdimages</li> <li>2. Initialize z=0</li> <li>3. While z &lt; frames do</li> <li>4. Initialize x=1</li> <li>5. no=M(z)2+M(z+x)2</li> <li>6. n=M(z)2+ M(z+x+1)2</li> <li>7. <math display="block">Y = \frac{\sqrt{n} + \sqrt{no}}{3}</math></li> <li>8. <math display="block">Y = \sqrt{Y}</math></li> <li>9. z = z+1</li> <li>10. Compute HOG</li> </ol>	<ol style="list-style-type: none"> <li>1. Computing thresholdimages</li> <li>2. Initialize z</li> <li>3. Frames=no. Of frames</li> <li>4. Initialize x</li> <li>5. no=sum of square of 1st and 2nd frame</li> <li>6. n=sum of square of 1st and 3rd frame</li> <li>7. Calculating Y</li> <li>8. Calculating square root of Y</li> <li>9. Increment z</li> <li>10. Computing histogram of gradient</li> </ol>
--	--

Fig. 2 Generation of descriptor

### 3.1 Preprocessing

In this section, we abstract three consecutive frames and convert them into thresholded images. The method we employed for converting the frames into thresholded images is adaptive thresholding. In adaptive thresholding, we calculate different threshold values for different regions of same image. Now threshold values can be calculated using the mean of neighborhood areas or using the weighted sum of neighbor values where weights are a Gaussian window. Later, a constant is subtracted from the calculated threshold value. If the value of pixel is less than the threshold value, it is assigned to zero; otherwise, it is assigned to the desired maximum value. In our method, we have calculated threshold values for each region using mean with the block size (size of neighborhood area which is used to calculate threshold value) of eleven and the constant (which is subtracted) two. The value of constant may vary for some other set of videos. Let  $T(x, y)$  is a pixel after thresholding,  $t$  be the thresholded value,  $m$  be the maximum value that can be assigned to the pixel and  $I(x, y)$  is a pixel of a frame. The equation for adaptive thresholding is given in Eq. (1).

$$T(x, y) = \begin{cases} m & \text{if } I(x, y) \geq t \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Further, we compute squares of each pixel in first, second and third frames we get after thresholding as shown in Eqs. (2), (3) and (4). Then, we calculate two values A and B, by adding the values in Eqs. (2) and (3), (2) and (4), respectively, as shown in Eq. (5) and Eq. (6). The square root of these A and B is calculated and is then

divided by the number of frames which in our case is three as depicted in Eq. (7). The operation of square root is again applied to the achieved result C, Eq. (8).

$$F1 = IMG1^2 \tag{2}$$

$$F2 = IMG2^2 \tag{3}$$

$$F3 = IMG3^2 \tag{4}$$

$$A = F1 + F2 \tag{5}$$

$$B = F1 + F3 \tag{6}$$

$$C = \frac{\sqrt{A} + \sqrt{B}}{3} \tag{7}$$

$$R = \sqrt{C} \tag{8}$$

Figure 3 shows the complete diagrammatic representation of preprocessing. After preprocessing, we obtain motion identifying image which is then fed to our descriptor HOG for feature extraction.

### 3.2 Descriptor

We have used histogram of orientation gradient (HOG) to compute features of motion identifying image. HOG describes the appearance of a local object within an image by distribution of intensity gradient or edge directions. The image that we give as an input to the descriptor is divided into small regions, which are called cells. These cells are connected. Histogram of gradient directions is calculated for each pixel within these cells. HOG computes the derivative of image (M) with respect to x and y as shown in Eqs. 9 and 10.

$$M_x = M * DX \text{ where } DX = [-1 \ 0 \ -1] \tag{9}$$

$$M_y = M * DY \text{ where } DY = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \tag{10}$$

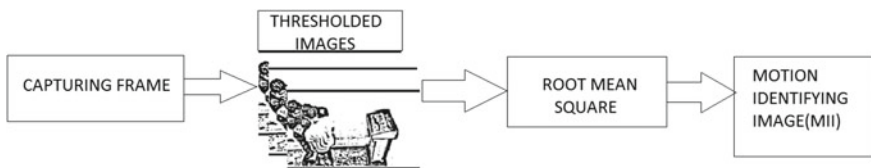


Fig. 3 Preprocessing

Further, we calculate magnitude and gradient of  $M$  in Eqs. 11 and 12.

$$|\text{Gr}| = \sqrt{M_x^2 + M_y^2} \quad (11)$$

$$\theta = \arctan\left(\frac{M_x}{M_y}\right) \quad (12)$$

Finally, cell histograms are created and then normalized using L2 normalization as shown in Eq. 13.

$$\mathcal{F} = \frac{n}{\sqrt{n_2^2 + \vartheta}} \quad (13)$$

Here,  $n$  represents vector without normalization containing all histograms of current block and  $\vartheta$  represent small constant.

We have used random forest classifier that works by creating multiple decision trees during training. In our case, the model had been trained using random forest classifier which creates 100 trees.

## 4 Results and Analysis

The proposed framework has been tested and trained, using Python 3.0 and OpenCV on computer having the specifications Intel i5 clocked at 2.4 GHz processor with the RAM of 16 GB, on videos for calculating various shape descriptors. The videos analyzed by the presented algorithms have a minimum resolution of  $320 \times 240$ . These videos are recorded in indoor premises such as ATM room. We have analyzed four categories of video captured as shown in Fig. 4: (i) single: when normal activities are being performed by a single person in a single camera view; (ii) single abnormal: when abnormal activities are being performed by a single person in a single camera view; (iii) multiple: when normal activities are being performed by a multiple person in a single camera view; (iv) multiple abnormal: when abnormal activities are being performed by a multiple person. There are a total of 49 videos in all the four classes (10 single, 10 single abnormal, 20 multiple and 9 multiple abnormal). In India, it is common for multiple personnel to enter the ATM room together. So for this sole activity we have taken a class of videos multiple. The framework is trained using these videos for extracting features from image sequences. The framework uses different videos for both testing and training purposes. The algorithm is tested for three frames, and its comparison against various other algorithms is shown in Table 1. Table 2 shows the value of W, X, Y and Z, the four classes that we have used in our dataset.

In Table 1, we have given comparative analysis with two other methods for motion encoding, which produces the best accuracy when an input of ten frames is given to the descriptor. First method uses (a) motion history image (MHI) as a descriptor;



**Fig. 4** Four classes of videos

**Table 1** Comparison with other descriptor (in percentage %)

Algorithm used	Result (%)
1. Combination of MHI and HU Moments	95.73
2. Combination of HOG and Zernike moments	95.02
3. Motion identifying image (MII) on thresholded images	97.24

**Table 2** Confusion matrix of MII on thresholded images

	W	X	Y	Z
W = Single	571	0	0	0
X = Single Abnormal	11	179	0	2
Y = Multiple	6	0	710	26
Z = Multiple Abnormal	0	0	6	332

second method uses (b) the fusion of histogram of gradient (HOG) and Zernike moments. In general, the more frames we give to the descriptor, the more accuracy we get, as temporal information increases but even after using ten frames as an input to the descriptors used in other two algorithms, their result is comparatively less than what we acquired using MII of three frames. Hence, from the figure it is clear that our descriptor MII is better in detecting motion than MHI and fusion of HOG and Zernike.



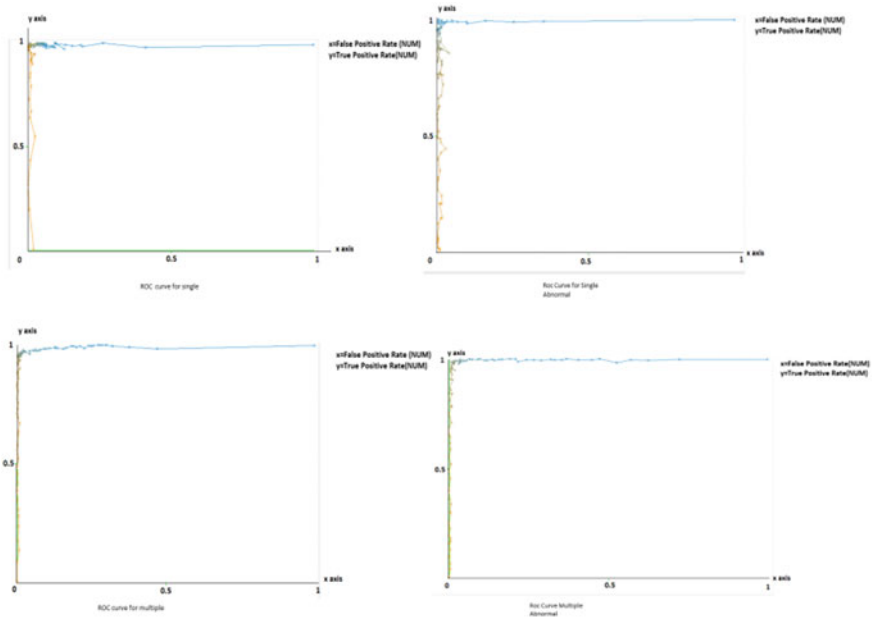


Fig. 5 ROC curve

Figure 5 shows the corresponding ROC graphs for all the four classes that is single, single abnormal, multiple and multiple abnormal for the testing dataset. In all the four graphs, the x-axis represents false-positive rate and the y-axis shows true-positive rate.

## 5 Conclusion

In this paper, we have proposed an algorithm that makes use of motion encoding technique called motion identifying image (MII) and a gradient-based descriptor HOG to recognize motion. It can be used in enhancing the security of ATM surveillance as well as in any other similar areas. The algorithms are tested for both normal and abnormal events with single as well as multiple personnel that can occur in ATM. It can contribute to the security of ATM as there is a tremendous increase in ATM frauds. In our method, the accuracy is about 97% when used with three frames. In the future, this motion encoding technique can be combined with any other descriptor to obtain higher accuracy. Also, an advanced and better classifier can be used for better recognition.

## References

1. Wang, C., Komodakis, N., Paragios, N.: Markov random field modeling, inference & learning in computer vision & image understanding. A survey. *Comput. Vis. Image Underst.* **117**(11), 1610–1627 (2013)
2. Chen, P., Chen, X., Jin, B., Zhu, X.: Online EM algorithm for background subtraction. *Procedia Eng.* **29**, 164–169 (2012)
3. Blanco Adán, C.R., Jaureguizar, F., García, N.: Bayesian visual surveillance: a model for detecting and tracking a variable number of moving objects. In: 18th IEEE International Conference on IEEE Image Processing (ICIP), pp. 1437–1440 (2011)
4. Boateng, R.: Developing e-banking capabilities in a Ghanaian Bank. Preliminary lessons. *J. Internet Bank. Commer.* 213–234 (2006)
5. Kumar, P., Mittal, A., Kumar, P.: Study of robust and intelligent surveillance in visible and multi-modal framework. *Informatica (Slovenia)* **32**(1), 63–77 (2008)
6. Cucchiara, R.: Multimedia surveillance systems. In: Proceedings of the Third ACM International Workshop on Video Surveillance & Sensor Networks, pp. 3–10. ACM (2005)
7. Babu, R.V., Ramakrishnan, K.R.: Compressed domain human motion recognition using motion history information. In: 2003 International Conference on Image Processing, vol. 3, pp. 321–324. IEEE (2003)
8. Gupta, R., Jain, A., Rana, S.: A novel method to represent repetitive and overwriting activities in motion history images. In: 2013 International Conference on Communications and Signal Processing (ICCSP), pp. 556–560. IEEE (2013)
9. Zhou, F., De la Torre, F., Hodgins, J.K.: Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(3), 582–596 (2013)
10. Bradski, G.R., Davis, J.: Motion segmentation and pose recognition with motion history gradients. *Mach. Vis. Appl.* **13**(3), 174–184 (2002)
11. Pandey, M., Sanserwal, V., Tripathi, V.: Intelligent vision based surveillance framework for ATM premises (2016)
12. Sujith, B.: Crime detection and avoidance in ATM. *Int. J. Comput. Sci. Inf. Technol.* 6068–6071 (2014)
13. Davis, J.W., Bobick, A.F.: The representation and recognition of human movement using temporal templates. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 928–934 (1997)
14. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit.* **47**(6), 2280–2292 (2016)
15. Ahad, M.A.R., Ogata, T., Tan, J.K., Kim, H.S., Ishikawa, S.: Directional motion history templates for low resolution motion recognition. In: 34th Annual Conference of IEEE, pp. 1875–1880 (2008)
16. Ahad, M.A.R., Ogata, T., Tan, J.K., Kim, H.S., Ishikawa, S.: Template-based human motion recognition for complex activities. *IEEE International Conference*, pp. 673–678 (2008)
17. Poppe, R.: A survey on vision-based human action recognition. *Image Vis. Comput.* **28**(6), 976–990 (2010)
18. Al-Berry, M.N., et al.: Action recognition using stationary wavelet-based motion images. *Intelligent Systems*, pp. 743–753 (2014). Springer International Publishing (2015)
19. Hu, R., Collomosse, J.: A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Comput. Vis. Image Underst.* **117**(7), 790–806 (2013)
20. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
21. Hu, M.K.: Visual pattern recognition by moment invariants. *IEEE Trans. Inf. Theory* **8**(2), 179–187 (1962)
22. Amato, A., Lecce, V.D.: Semantic classification of human behaviors in video surveillance systems. *J. WSEAS Trans. Comput.* **10**, 343–352 (2011)

23. Chen, Q., Wu, R., Ni, Y., Huan, R., Wang, Z.: Research on human abnormal behavior detection and recognition in intelligent video surveillance. *J. Comput. Inf. Syst.* **9**(1), 289–296 (2011)
24. Srestasathiern, P., Yilmaz, A.: Planar shape representation and matching under projective transformation. *Comput. Vis. Image Underst.* **115**(11), 1525–1535 (2011)
25. Sanserwal, V., Pandey, M., Tripathi, V., Chan, Z.: Comparative analysis of various feature descriptors for efficient ATM surveillance framework (2017)
26. Tripathi, V., et al.: Robust abnormal event recognition via motion and shape analysis at ATM installations. *J. Electr. Comput. Eng.* (2015)
27. Rashwan, H.A., et al.: Illumination robust optical flow model based on histogram of oriented gradients. In: *German Conference on Pattern Recognition*, pp. 354–363. Springer, Berlin, Heidelberg (2013)
28. Hirabayashi, M., et al.: GPU implementations of object detection using HOG features and deformable models. In: *IEEE 1st International Conference on IEEE Cyber-Physical Systems, Networks, and Applications (CPSNA)*, pp. 106–111 (2013)
29. Huang, C., Huang, J.: A fast HOG descriptor using lookup table and integral image (2017). [arXiv:1703.06256](https://arxiv.org/abs/1703.06256)
30. Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off. *Int. J. Multimed. Inf. Retr.* **4**(1), 33–44 (2015)
31. Kennedy, R., Taylor, C.J.: Optical flow with geometric occlusion estimation and fusion of multiple frames. In: *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 364–377. Springer International Publishing (2015)

# EKRV: Ensemble of kNN and Random Committee Using Voting for Efficient Classification of Phishing



A. Niranjan, D. K. Haripriya, R. Pooja, S. Sarah, P. Deepa Shenoy and K. R. Venugopal

**Abstract** Any efficient anti-phishing tool must be able to classify phishing activity as ‘phishing’ with utmost accuracy. The key factor that influences the accuracy of an anti-phishing tool is the selection of a classification algorithm whose prediction accuracy is the maximum with nil or least false-positive rate. This paper proposes the implementation of a hybrid approach involving random committee that is a type of Ensemble classification technique and k-nearest neighbor (kNN) algorithm which is available as IBK (instance-based with k neighbors) on WEKA, resulting in most encouraging prediction accuracy values. The proposed scheme is followed after the preprocessing phase that involves feature extraction using Consistency Subset Eval algorithm with the Greedy Stepwise search technique.

**Keywords** Random committee · kNN · Phishing · Voting · Ensemble classifiers

## 1 Introduction

Any attempt of obtaining sensitive credentials of a person such as username, password, OTP, PIN and other details, for malicious reasons, by mimicking a trustworthy entity is referred to as *phishing* [1]. Phishing attack is launched mostly through an email that has a link to a fake website, with almost identical look and feel as that of the legitimate one. The objective of phishers is to make users believe that they are interacting with trusted online sites. An efficient method of identifying and classifying phishing websites is required in order to protect users’ sensitive data.

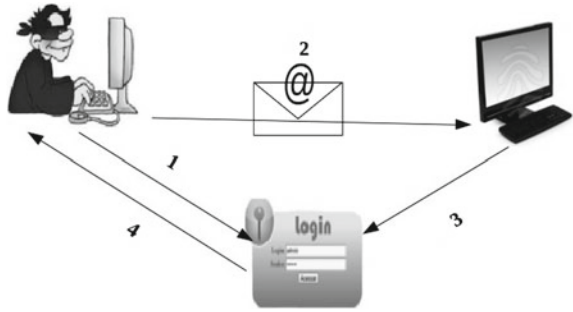
Presently available browser and other application programs are designed keeping the relative ease of use by users with no or very less concern for security. The main reason behind the users falling prey for the phishing attacks is the lack of

---

A. Niranjan (✉) · D. K. Haripriya · R. Pooja · S. Sarah · P. Deepa Shenoy · K. R. Venugopal  
Department of Computer Science and Engineering, University Visvesvaraya  
College of Engineering, Bangalore University, Bengaluru 560001,  
Karnataka, India  
e-mail: a.niranjansharma@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_37](https://doi.org/10.1007/978-981-13-1708-8_37)

403

**Fig. 1** Phishing life cycle

skills required to distinguish between a genuine and phishing websites. The phishing websites though appear almost identical to the genuine ones, often leave some clues. Experts in the domain look for such clues to identify the site as a fraudulent site. When the amount of requested information is too high or the URL of the site is too small or too big or it contains too many foreign anchors (Links to other URLs) or the presence of an IP Address instead of a domain name in the URL or the presence of more than five dots or slashes in the URL, the site that is being visited could possibly be a phishing site.

### ***1.1 Phishing Life Cycle***

The phishing life cycle typically involves the following steps as depicted in Fig. 1.

- Step 1: The phisher creates a phishing site which is almost similar to a genuine site to lure the customers and to make them believe that the visited site is a legitimate site.
- Step 2: The phisher now sends an email to all the targetted audience and prompts them to click on the link in the mail to the phishing site.
- Step 3: The user is directed to the phishing site where all his personal credentials are collected by the phisher.
- Step 4: This information of the user is now misused by the phisher to launch the phishing attacks.

### ***1.2 Types of Phishing Attacks***

Some of the most commonly found types of phishing attacks are:

- Deceptive phishing
- Malware-based phishing
- DNS-based phishing

- Content-injection phishing
- Man-in-the-middle phishing and
- Search engine phishing

Deceptive phishing involves the collection of user credentials by first sending a broadcast mail to a number of clients and asking them to take action against severe system failure by reentering the login information or urging them to login to their account to check fictitious charges to their bank accounts by clicking on the links provided. Malware-based phishing on the other hand involves the introduction of malware programs as attachments to an email or by exploiting the system vulnerabilities to getaway with the essential user information to launch a phishing Attack. The attackers can direct the user requests to a fake site either by configuring the Domain Name System or by tampering the host files. Such an attack is referred to as DNS-based phishing or pharming. When a small portion of the content of a legitimate site is replaced with malicious code by a phisher for the collection of user credentials; then, it is called Content-Injection phishing. A man-in-the-middle phishing involves a phisher positioning himself conveniently in between the user and the legitimate website to record the information being entered without affecting the ongoing transaction. In search engine phishing, an attacker creates websites with too attractive luring offers and have them indexed with search engines in a very legitimate way. Customers come across these sites in the normal course of searching for products or services and are fooled into giving up their information.

It is therefore necessary for the anti-phisher application to recognize the attack as phishing on its onset. The anti-phisher can make use of machine learning techniques to classify the ongoing activity as phishing or as legitimate [1]. The performance of an anti-phisher depends mainly on the prediction accuracy of the chosen classification algorithm.

AIM: The current work aims at providing a high prediction accuracy with a significant reduction in FPR and absolute error rate. This is achieved by selecting various classifying machine learning algorithms with better individual performances and their combinations.

This work was taken up to achieve the following objectives:

- Application of different classifier algorithms to determine the most efficient classifiers in terms of prediction accuracy, false-positive rate and absolute error.
- Extraction of the most relevant features from the UCI machine learning repository phishing data set for dimensionality reduction using the best Feature Selection Algorithm.
- To use a hybrid model of the better classification algorithms to further enhance the performance of classification in terms of prediction accuracy, false-positive rate and absolute error.

The rest of the paper is organized as follows: Sect. 2 summarizes the related work carried out in the current domain, while Sect. 3 presents the proposed EKRK model. Experimental results are discussed in Sect. 4, and conclusion forms Sect. 5 of the paper.

## 2 Related Work

Lakshmi and Vijaya [2] in their paper have experimented with multilayer perceptron, Decision tree induction and Naïve Bayes classification techniques of machine learning techniques. Multilayer perceptron is a type of neural network classifier with a number of models organized into multiple layers. The decision tree induction on the other hand involves the construction of a decision tree model on the train set, and the test data is classified based on this model. The Naïve Bayes technique involves two steps for classification with the estimation of parameters for probability distribution as the first step and computation of posterior probability and performing classification based on the largest posterior probability value of the test sample. Their results indicate that decision tree classification involving J48 algorithm shows better classification accuracy compared to the other two techniques. However, the train set that they have used is created using URLs of only 200 sites.

The authors of [3] in their paper have proposed an associative classification algorithm called FACA for the efficient detection of phishing sites. Classification technique on one hand involves assigning or predicting test instances to their pre-defined classes, while association rule mining involves determining relationships between attributes in a large database and forming rules based on these relationships. Associative classification (AC) is a hybrid approach that combines classification with association rule mining techniques. AC aims at classifying unseen test instances based on association rules. The FACA, however, has a prediction accuracy of just over 92%.

The paper proposed by the authors of [4] discusses on hybrid approach that makes use of sequential minimal optimization (SMO) and the genetic algorithm (GA) for phishing website detection. SMO is basically used to solve the quadratic problem that arises while training a model using support vector machines (SVMs). The Genetic algorithm in this hybrid model is used to optimize the parameters. The prediction accuracy of this model is, however, just over 96%.

Chowdhury et al. have proposed multilayer hybrid strategy (MHS) involving ten layers of processing [5]. One of the layers encompasses a hybrid approach for feature extraction. Another layer involves a hybrid approach for pruning. For classification, they have proposed an ensemble of random forests. The prediction accuracy of this complex model is only around 95%.

A hybrid approach involving fuzzy-based and associative classification techniques is discussed in paper [6]. This method, however, has a prediction accuracy of 92%.

## 3 Proposed EKRV Model

The UCI machine learning data set on phishing has 30 features with 6157 legitimate and 4898 phishing instances out of a total of 11,055 instances distributed in the ratio of 55.7 and 44.3%. The model has to be trained before carrying out classification.

The data set was therefore divided into train and the test sets with the train set getting a share of 9087 instances, while the test set getting 1968 instances. The same ratio that existed between the legitimate and phishing samples were maintained even in the train and the test sets. Thus, the train set has 5061 legitimate and 4026 phishing samples, while the test set has 1096 legitimate and 872 phishing samples. The data set has a total of 30 attributes with an additional class attribute that indicates whether the sample in question is phishy or legitimate or suspicious. A sample is labeled as  $-1$  to indicate that it is phishy,  $1$  as legitimate and  $0$  as suspicious. Preprocessing is an essential operation in data mining. This could involve feature extraction, normalization, cleaning and other operations. Feature extraction often aims at dimensionality reduction resulting in faster execution of an application by removing the unnecessary or unwanted features. Hence, it was decided to use most suitable feature extraction algorithm that results in least number of features. It was found that the Consistency Subset Eval Feature Extraction Algorithm along with Greedy Stepwise Search method produces least number of features totaling 23 features with an additional class label. The next step of our approach is to cross check whether the chosen features are optimal or not. Delta values of the ranks are determined, and the min, max and the mean of the delta values are computed. The features numbered 1–20 are found to have a delta value of 0. The features around feature 20 are 19 and 21. So feature 21 is also considered to be optimal. Delta value is 0 once again at 22. Feature around this number is 23. Hence, the 23<sup>rd</sup> feature is considered. The chosen 23 features thus are proven to be optimal.

The extracted features are then subjected to an ensemble of random committee with random tree as the base classifier and kNN available as IBK on WEKA with cover tree as the base classifier using voting for classifying the incoming sample as phishy or legitimate. The algorithm of our proposed EKRK is listed in Algorithm 1.

**Algorithm 1 EKRK:**

- 1: **while** True **do**
- 2:     Load the UCI Machine Learning Phishing Data Set
- 3:     Extract features using Consistency\_Subset\_Eval Feature Extraction Algorithm with Greedy Stepwise Search method
- 4:     **if** the extracted features are optimal **then**
- 5:         Choose Random Committee with Random Tree as Base Classifier and KNN/IBK with Cover Tree as the search algorithm using Voting for the classification of the samples. Choose Product of probabilities as the combination Rule while voting.
- 6:     **end if**
- 7: **end while**



### 3.1 System Model of the Proposed EKRK

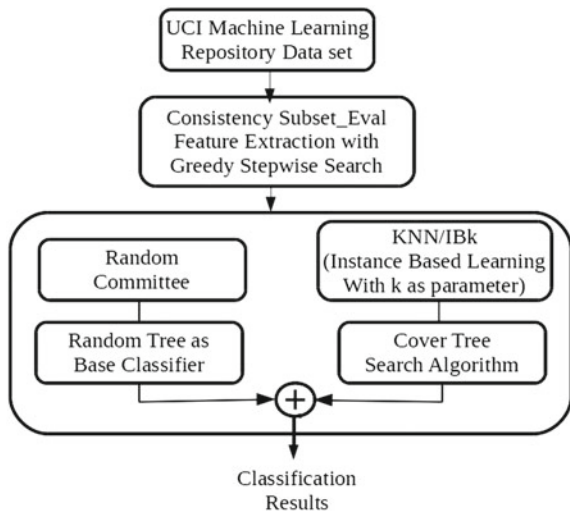
EKRK is basically a hybrid technique that involves a combination of K-nearest neighbor (KNN) and random committee techniques. It is possible to combine more than 1 classifier either through voting or stacking. When the two classifiers were fused using voting, the prediction accuracy was found to be comparatively higher than stacking as listed in Table 1. For this reason, voting was chosen as the fusion Scheme. The system model of the proposed scheme is illustrated in Fig. 2. As discussed earlier, EKRK involves two phases: preprocessing and classification. The preprocessing phase involves extraction of the features from the UCI machine learning phishing data set, using Consistency Subset Eval Algorithm with the default Greedy Stepwise Search. This step results in 23 features.

The test set is now subjected to the classification phase that involves a hybrid algorithm that combines random committee and kNN algorithms using voting. Voting is a process of creating a number of sub-models and combining the predictions of all sub-models for the final prediction. Random committee is considered to be one of the ensemble classifiers. An ensemble classifier utilizes a base classification algorithm to differently permuted training sets. The size of all these training sets would, however, be the same. An unlabeled data sample is assigned a particular class based on the highest number of votes received among the individual classifiers' predictions. In random committee, a number of case classifiers are built using a

**Table 1** Voting versus stacking

Ensemble approach	Prediction accuracy	False-positive rate
Stacking	97.1	0.031
Voting	97.4	0.028

**Fig. 2** System model of the proposed EKRK



different random number seed and the average of the predictions generated by the individual base classifiers forms the final prediction. kNN is a type of instance-based learning, or lazy learning algorithm, in which computation is deferred until classification. A test sample is classified based on what is the prediction of the majority of its neighbors, with the sample being assigned to the class most common among its k-nearest neighbors (k is a positive integer, typically small). If k = 1, then the sample is simply assigned to the class of that single nearest neighbor. The system model for the proposed EKRV is illustrated in Fig. 2.

### 4 Experimental Results and Discussion

As a part of our research, it was initially decided to run all available classification algorithms on all the 30 features of the data set and to determine the best classification algorithm(s). Table 2 lists the best classification algorithms under each classifier category available on WEKA, their prediction accuracy, false-positive rates and the build time. Prediction accuracy of a classification algorithm is in turn related to true-positive rate (TPR), false-positive rate (FPR), true-negative rate (TNR) and false-negative rate (FNR) [7]. Let  $N_{Leg}$  and  $N_{Phish}$  be the total number of legitimate and phishing samples, respectively. The number of legitimate samples classified as legitimate can be denoted as  $N_{Leg \rightarrow Leg}$  and the number of phishing samples classified as phishing as  $N_{Phish \rightarrow Phish}$ . Let  $N_{Leg \rightarrow Phish}$  denote the number of legitimate samples misclassified as phishing and  $N_{Phish \rightarrow Leg}$  denote the number of phishing samples misclassified as legitimate.

The true-positive rate (TPR) is the rate of phishing samples classified as phishing out of the total phishing samples.

$$TPR = N_{Phish \rightarrow Phish} / N_{Phish} \times 100 \tag{1}$$

False-positive rate (FPR) is the rate of phishing websites misclassified as legitimate out of the total phishing websites.

**Table 2** Performance of various classifiers

Classification algorithm	Prediction accuracy	False-positive rate	Build time
Random committee	97.3	0.029	0.27
kNN/IBK	97.2	0.03	0
PART	96.8	0.035	1.06
Logistic	94.0	0.064	1.36
HNB	93.7	0.069	0.09
WAODE	93.7	0.065	0.11
VFI	92.7	0.077	0.03

$$\text{FPR} = N_{\text{Phish} \rightarrow \text{Leg}} / N_{\text{Phish}} \times 100 \quad (2)$$

False-negative rate (FNR) is the rate of legitimate samples misclassified as phishing out of the total legitimate samples.

$$\text{FNR} = N_{\text{Leg} \rightarrow \text{Phish}} / N_{\text{Leg}} \times 100 \quad (3)$$

True-negative rate (TNR) is the rate of legitimate samples classified as legitimate out of the total legitimate samples.

$$\text{TNR} = N_{\text{Leg} \rightarrow \text{Leg}} / N_{\text{Leg}} \times 100 \quad (4)$$

Prediction accuracy (PA) is the total number of phishing and legitimate samples that are identified correctly with respect to the total of all the samples.

$$\text{PA} = (N_{\text{Leg} \rightarrow \text{Leg}} + N_{\text{Phish} \rightarrow \text{Phish}}) / (N_{\text{Leg}} + N_{\text{Phish}}) \times 100 \quad (5)$$

It may be noticed from Table 2 that only random committee and IBK (kNN) algorithms exhibit maximum prediction accuracy with least FPR. Build time of some of the classifiers are smaller, but their accuracy levels are also low. Hence, it was decided to use a hybrid approach involving an ensemble of random committee and kNN (IBK) classification algorithms. To further reduce the dimensionality of the data set and the computation time, a preprocessing stage involving feature extraction was introduced. To determine the best feature extraction Algorithm out of a number of feature extraction algorithms available on WEKA, prediction accuracy and false-positive rates of both random Committee and kNN for different feature extraction algorithms were computed as listed in Table 3. It may be noticed that Consistency Subset Eval Algorithm results in 23 features [8] without much degradation in the prediction accuracy levels.

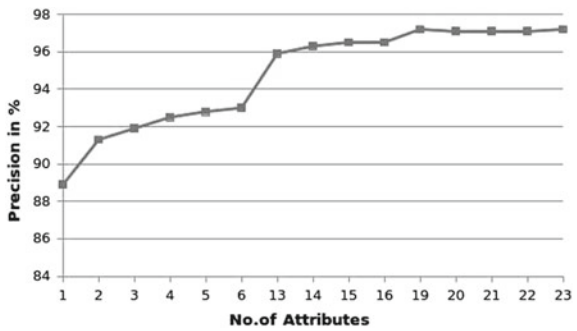
The next important step in the proposed EKRV model is to check whether the chosen 23 features by the Consistency Subset Eval are indeed optimal or not. For this purpose, we decided to compute the delta values for the generated ranks and to determine min, max and the median values of the delta values. They were found to be 0, 0.024 and 0.003483, respectively. The feature present at serial number 20 (F4) is found to have a delta value of 0. Features found at serial numbers 19 and 21 that is F19 and F30 are also considered to be optimal, and hence, the next feature present at serial number 22 which is F3 is also included in the chosen feature set. As the other features after serial number 23 have a delta value of 0, they are discarded and only the feature present at 23 that is F22 is considered to be optimal. Table 4 lists all the features, their ranking values and the delta values. To further prove that the chosen 23 features are optimal, the prediction accuracy of random committee is plotted against the number of attributes as illustrated in Fig. 3 and that of the kNN in Fig. 4.

It was decided to fuse random committee and kNN techniques through voting for the classification of the test samples as the combined prediction accuracy of

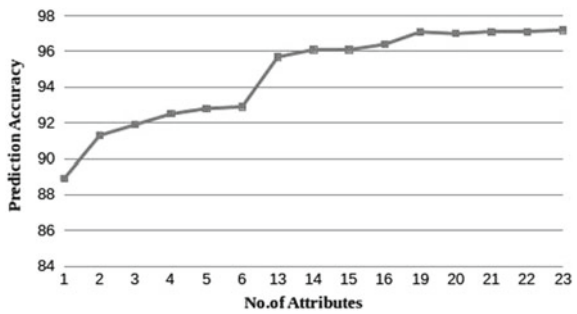
**Table 3** Performance of random committee and kNN with different feature extraction algorithms

Feature extraction algorithm	Extracted features	Random committee	kNN
CFSSubsetEval	F6, F7, F8, F13, F14, F15, F16, F26, F28	94.7	94.5
ChiSquaredAttributeEval	F1–F30	97.3	97.2
ClassifierSubsetEval	Nil	0.31	0.31
ConsistencySubsetEval	F1–F4, F6–F9, F12–F17, F19, F22, F24–F30	97.2	97.1
FilteredAttributeEval	F1–F30	97.3	97.2
FilteredSubsetEval	F8, F14	91.3	91.3
InfoGainAttributeEval	F1–F30	97.3	97.2
OneRAttributeEval	F1–F30	97.3	97.2
ReliefFAttributeEval	F1–F30	97.3	97.2
Symmetrical UncertAttributeEval	F1–F30	97.3	97.2
WrapperSubsetEval	Nil	0.31	0.31

**Fig. 3** Accuracy in random committee



**Fig. 4** Accuracy in kNN/IBK



these schemes resulted in better prediction accuracy rate of about 97.4 and lesser false-positive rate as shown in Fig. 5.

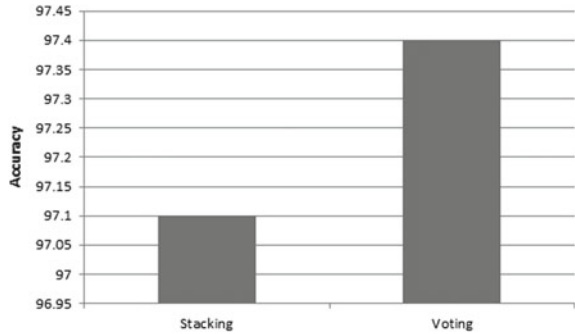
**Table 4** Ranking values generated by Consistency Subset Eval Algorithm for all 30 features along with the difference between the features (Delta)

Serial number	Ranking value	Feature number	Delta
1	0.889	F8	–
2	0.913	F14	0.024
3	0.919	F15	0.006
4	0.925	F6	0.006
5	0.928	F29	0.003
6	0.934	F26	0.006
7	0.946	F7	0.012
8	0.952	F24	0.006
9	0.957	F2	0.005
10	0.962	F25	0.005
11	0.967	F1	0.005
12	0.972	F13	0.005
13	0.977	F27	0.005
14	0.98	F28	0.003
15	0.983	F9	0.003
16	0.985	F17	0.002
17	0.987	F16	0.002
18	0.988	F12	0.001
19	0.989	F19	0.001
20	0.989	F4	0
21	0.99	F30	0.001
22	0.99	F3	0
23	0.99	F22	0
24	0.99	F5	0
25	0.99	F10	0
26	0.99	F11	0
27	0.99	F18	0
28	0.99	F20	0
29	0.99	F21	0
30	0.99	F23	0

## 5 Conclusion

Feature extraction is applied initially on the UCI machine learning phishing data set to reduce dimensionality and to remove all the redundant and irrelevant features. Out of the 30 features that are present in it, only 23 features are considered. For the elimination of the redundant features, Consistency Subset Eval Algorithm was employed and only after ensuring that the chosen features are indeed optimal, our EKRv model was used. Voting feature on WEKA was utilized for combining the

**Fig. 5** Stacking versus voting



**Table 5** Performance of the proposed EKRV

Mechanism	Prediction accuracy	False-positive rate	Absolute error
Using entire set as test set (11,055)	99	0.01	0.0117
Tenfold cross-validation	97.4	0.028	0.0275
Training set (9087)	99.1	0.01	0.0105
Test set (1968)	94	0.063	0.0653
Tenfold cross-validation	97.4	0.028	0.0272

random committee and kNN classifiers. A tenfold cross-validation too was applied for the proposed classification model. The experimental results as listed in Table 5 prove that prediction accuracy, TPR and FPR rates are better compared to the existing models. The proposed model is required to be tested on other data sets as well, and total build time required to carry out classification has to be further reduced.

## References

1. Niranjana, A., Nitish, A., Deepa Shenoy, P., Venugopal, K.R.: Security in data mining—a comprehensive survey. *Global J. Comput. Sci. Technol.* **16**(5), 52–73 (2017)
2. Lakshmi, V.S., Vijaya, M.S.: Efficient prediction of phishing websites using supervised learning algorithms. *Procedia Eng.* **30**, 798–805 (2012)
3. Hadi, W., Aburub, F., Alhawari, S.: A new fast associative classification algorithm for detecting phishing websites. *Appl. Soft Comput.* **48**, 729–734 (2016)
4. Yan, Z.: A genetic algorithm based model for Chinese phishing E-commerce websites detection. In: *International Conference on HCI in Business, Government and Organizations*, pp. 270–279 (2016)
5. Chowdhury, M.U., Abawajy, J.H., Kelarev, A.V., Hochin, T.: Multilayer hybrid strategy for phishing email zero-day filtering. In: *Concurrency and Computation: Practice and Experience* (2016)
6. Shah, R.K., Hossain, M.A., Khan, A.: Intelligent phishing possibility detector. *Int. J. Comput. Appl.* **148**(7), pp. 1–8 (2016)

7. Moghimi, M., Varjani, A.Y.: New rule-based phishing detection method. *Expert Syst. Appl.* **53**, 231–242 (2016)
8. Mohammad, R.M., Thabtah, F., McCluskey, L.: Phishing Websites Features (2015). [http://eprints.hud.ac.uk/24330/6/RamiPhishing\\_Websites\\_Features.pdf](http://eprints.hud.ac.uk/24330/6/RamiPhishing_Websites_Features.pdf)

# Enhanced Digital Video Watermarking Technique Using 2-Level DWT



Rashmi Jakhmola and Rajneesh Rani

**Abstract** “Watermarking” is the strategy for camouflage digital information in an exceptionally carrier signal; the shrouded information should, however does not need to be constrained for holding a reference to the carrier signal. They can also be used to verify or to recognize the house owner’s actual real identity or then again to confirm the validity or respectability of the carrier signal. It is prominently utilized for following copyright encroachments and for paper cash verification. This paper proposes an improved system of video watermarking which maximizes the normalized correlation coefficient, making the system tougher, so that they can bear common attacks. The suggested algorithm applies 2-level DWT for embedding watermark in the video, frame by frame along with a Gaussian filter, which further improves the features of suggested algorithm’s watermarked content with less deterioration of visual quality of the video. After applying the algorithm, the result shows that the suggested algorithm is sturdier against most of the common attacks with improved PSNR and imperceptibility level.

**Keywords** Digital watermarking · DWT · PSNR · NCC  
Image enhancement filters

## 1 Introduction

Steganographic techniques are basically used in two of the category of data hiding, i.e., steganography and digital watermarking in order to embed the secret information within the noisy or carrier signal. A digital watermark can also be taken as passive protection tool, since original and watermarked copy appears to be same even though

---

R. Jakhmola (✉) · R. Rani  
Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar 144011, Punjab, India  
e-mail: nitu.csdept@gmail.com

R. Rani  
e-mail: ranir@nitj.ac.in

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_38](https://doi.org/10.1007/978-981-13-1708-8_38)



after the watermarking procedure. Steganography mainly targets the physical property observed by human senses to strengthen itself, whereas digital watermarking's main target is to improve and make itself efficient. It essentially stamps the information; however it does not corrupt it or administration access to the information. One application of digital watermarking is supply or source following. A watermark is embedded into a digital signal at every purpose of distribution. If a replica of the work is found later, then the watermark could also be retrieved from the copy and also the supply of the distribution is understood. This method reportedly has been accustomed to observe the supply of lawlessly derived movies. According to [1], there are so many types of watermark algorithm based on type of document, human perception, information for detection, robustness and embedding process. In the proposed algorithm, transform domain is used as it provides more robustness and imperceptibility. Watermark may degrade the quality of image or video, so based on [2], filters are used in order to lower the degradation of image, i.e., deblurring filters are used like unsharp filter, Gaussian filter and blind deconvolution filter, while the proposed algorithm uses Gaussian filter with standard deviation of 0.5 for further improving the peak signal-to-noise ratio (PSNR) and normalized correlation coefficient (NCC) value as in [3].

This paper is distributed in the following sections: Sect. 1 provides a brief description about digital video watermarking and Sect. 2 discusses about research work which has been already done on video watermarking. Section 3 consists of the detailed description of the suggested algorithm which is being implemented, which is accompanied by the detailed analysis and then implementation strategies is further overviewed. Section 3.1 carries the final results which we get after implementing proposed algorithm, and later on, Sect. 4 presents the final results and brief discussions related to that. Section 5 revolves around possible conclusion and related future work of the work done in paper.

## 2 Related Work

In the area of digital video watermarking, a wide amount of work has been done which are reviewed here. Various algorithms and strategies are applied to make it more efficient. According to [4], 2D-discrete cosine transform (DCT) is used for embedding watermark in V-component of each frame considering human visual system (HVS). The algorithmic rule is fairly strong against attacks like image cropping and noise like Gaussian noise and removal attack. In [5] highly smooth or textured blocks are recognized, then watermark is placed in them, and through this procedure, this formula of watermarking human sensory system is exploited. Every video has the textured blocks which possess very vital information of the video as there is a relation of this with the vital half in an exceeding video. In [3] the main system's motive is to make sure about the: (i) the image's confidentiality by applying the encryption, (ii) the image's reliability by embedding a reliability proof before the encryption process and (iii) content final user by tracing back through the means of watermarking.

In [6] watermark is embedded within the non-moving a part of every color in close all 3 RGB channels. Initially, some frames are chosen. After that, the video moving and non-moving components are separated from every frame square. In [7] a blind video watermarking technique is proposed which is resistant to most of rotation attacks and is DCT based in nature. Rotation changelessness property of the complicated Zernike moments is employed to attain the goal, whereas to create the theme strong against collusion, style of the theme is completed in such a way that the embedding blocks can vary for the sequential frames of the video. In [8] they slightly changed the pixel value ordering (PVO) predictor. Extended pixel-based PVO (PPVO) was incontestable to possess a bigger embedding capability and better marked image quality than PPVO, PVO, improved PVO and PVO-K. The experiments additionally showed that this technique mistreatment changed PPVO outperformed four different progressive strategies with moderate payloads, while in [9] a secure watermarking algorithm is granted which supported DCT, Schur transform and firefly algorithm in order to fulfill the augmented demand. These demands target the improvement in a secure algorithm which also maintains the basic properties of the algorithm which are: capability to hold enough payload, distinguished level of physical property and strength.

While [10] presents a unique digital image watermarking-supported ripple rework in HSI color area [11] proposed two new changes in the concept of watermarking scheme which is singular value decomposition-based SVD. These suggested changes will improve the physical property and capability once the watermark is planted into the U and V elements of the taken picture, in place of planting watermark into the U part planned earlier, from the theoretical analysis. In [12] first, they tend to scramble the slices of watermark by the mistreatment of Arnold rework key, after that it moves toward placing the disorganized slices or chunks of the watermark bits in place of the singular values associated with the DWT sub-base on the top of non-motion frames of CT domain. This algorithm has an upper edge due to the restoring nature of its embedding algorithm which further improves physical property. The reconciling nature of the embedding issue improves more the extent of physical property.

From the on top of analysis, it is clearly visible that attaining more hardiness in case of geometric attacks whereas continuing the physical property and hardiness in case of alternative usual attacks constitute one among the foremost difficult options for any video watermarking formula. Nearly all plans did not contemplate geometric attacks in comparison with alternatives failed to continuing the physical property within the instance of other regular attacks. Nearly all those projected solutions mainly focus on the hardiness of the formula but a lesser concern on the security concern, which itself is of high importance.

### 3 Proposed Methodology

The proposed algorithmic rule uses 2-level DWT with 2D-Gaussian filter that makes it a more efficient and strong against common attacks because it improves the PSNR and NCC values for constant dataset utilized in [13].

The procedure of video watermarking basically consists of 3 major steps, i.e., (i) Generation and Embedding, (ii) Distribution and potential attacks, and (iii) detection. And based on domain of embedding process, the technique of watermarking is divided into following categories and discussed as follows:

- (i) **Spatial Domain:** Here the watermark bits are directly embedded into the pixel value of image. So this domain is basically the image plane itself. Under this domain, the pixel value intensity is modified at a minor level. This technique cannot stand up to low-pass filtering and other customary image processing attacks as it possesses least quality with high payload.
- (ii) **Transform Domain:** Under this domain has impalpability and in addition powerful. Through this domain, we have the capability to estimate the frequency by adjusting the point original value, that is why it can be also called as frequency domain.

While using this approach low-frequency segment of picture information ought to be altered in as indicated by the watermarked information heartily through the change area methods.

In frequency domain methods, after altering the modification in the coefficients of the casings of the video grouping, the watermark is implanted. Discrete Fourier transform (DFT), discrete cosine transform (DCT) and the discrete wavelet transform (DWT) are most of the popularly used transforms. For the most part, the primary downside of change space strategies is their higher computational prerequisite.

- (i) **DFT Video Watermarking Technique:** This approach initially removes the splendor of the watermarked outline, figuring its full-outline DFT consuming most of the significance of the coefficients. The watermark is basically a product of two alphanumeric strings. In the procedure the DFT coefficients are adjusted, followed by then IDFT. In this procedure primary frame is processed for watermarking only, which was made out of twelve frames, as a result alternate ones are left uncorrupted. It is of great robustness to the standard picture preparing as nonlinear/linear filtering, JPEG compression, oppose to geometric changes as scaling, turn and trimming and at last sharpening. The watermark outline and the watermark inclusion methodology do not include any changes. Straightforward methods like expansion or swap are utilized for the blend of watermark. DFT-based watermarking plan with layout coordinating can oppose various assaults, including pixel expulsion, revolution and shearing. The reason for the layout is to empower resynchronization of the watermark payload spreading succession.
- (ii) **DCT Video Watermarking Technique:** Discrete cosine transform (DCT) is an imperative strategy for video watermarking. A considerable measure of com-

puterized video watermarking calculations implants the watermark into this space. The ease of use of this change is on account of that a large portion of the video pressure measures depend on DCT and some other related changes. In this space some DCT coefficients of the video are chosen and isolated into gatherings, and after that the watermark bits are implanted by doing modification in each gathering.

- (iii) DWT Video Watermarking Technique: The disseminations of the frequency is changed in each progression of DWT, where subscript behind them deals with the quantity of layers of changes, L deals with low recurrence and H deals with high recurrence. Sub-chart LL speaks to the lower-resolution estimation of the first video, while high-frequency and mid-frequency subtle elements sub-diagram LH, HL and HH speak to vertical edge, level edge and slanting edge points of interest. The procedure can be rehashed to process the different scale wavelet deterioration.

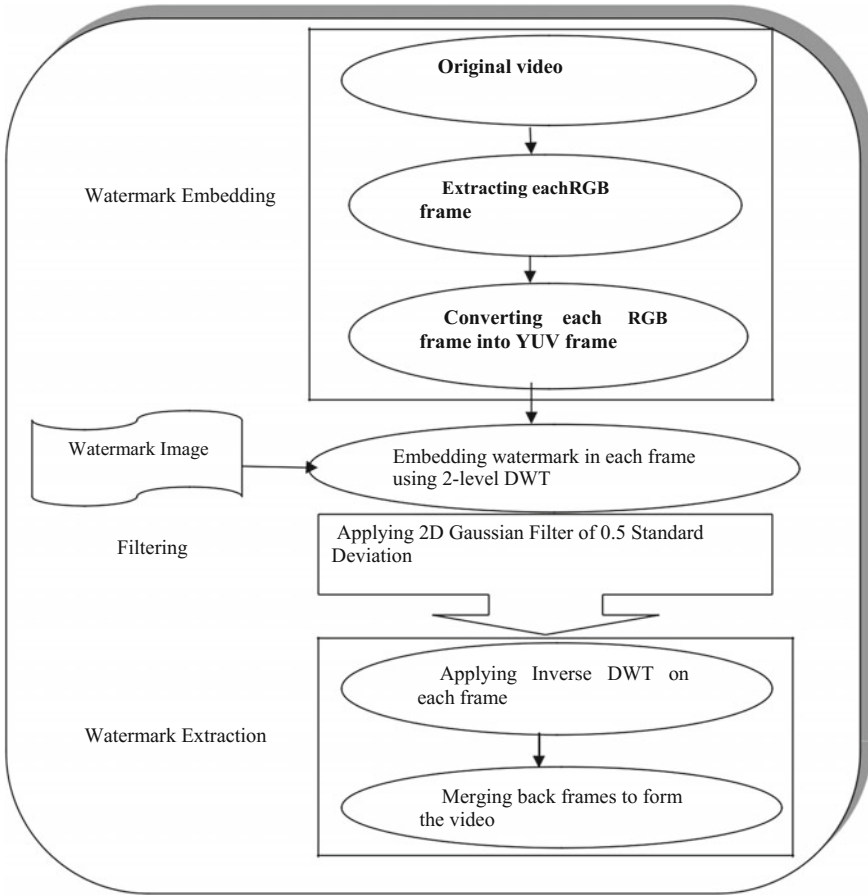
In the wake of watermarking, the frames of video may degrade. So, the corruption of the picture is diminished by utilizing image enhancement filters like unsharp filter, blind deconvolution filter and Gaussian filter. Their brief depiction is given as:

- (a) Unsharp Filter: This type of filters expand the high-frequency regions and can be straight or non-direct channel in nature. Negative Laplacian is most of the well-known method to deal with execute unsharp filter, it concentrate the high-frequency segments in a picture, which is then superimposed on the first picture, another possible ways are also available.
- (b) Blind Deconvolution: This is a deconvolution method that is utilized to upgrade obscured picture from a solitary or an arrangement of “obscured.” It depends on the suspicion that, because of the flaws of imaging (scanner, camera and magnifying instrument), the picture is convolved with the psf. psf is the scientific function that portrays the way mutilation of a hypothetical point wellspring of light through the imaging framework.
- (c) Gaussian filter: Mathematically, applying a Gaussian blur to a picture is the same as convolving the image with a Gaussian perform. This is often additionally called a two-dimensional Weierstrass transform. In the proposed algorithm, this filter is used with standard deviation of 0.5.

Figure 1 defines the basic steps of procedure followed in the proposed framework. It demonstrates how the three basic phases of watermarking happen. In spite of the fact that there are three separate stages, the filtering phase can only be executed before extraction phase. Figure 1 shows the basic flow of stages, that how the taken video is watermarked, filtered and finally how the watermark is extracted (Table 1).



### 3.1 Proposed Watermarking Embedding System Procedure

Video is partitioned into casings. RGB edges are changed over to YUV outlines. 2-DWT is connected on it. RGB watermark picture is changed over into a vector




**Fig. 1** Block diagram of proposed digital video watermarking algorithm

**Table 1** Sample video frame of dataset “Suzie.mpg” and sample watermark image





Sample video frame	Video size	Sample watermark	Watermark size
	352 × 288		256 × 198

$P = \{p_1, p_2, \dots, p_{32 \times 32}\}$  of ones. This vector  $P$  is again separated into  $n$  parts. At that point each part is implanted into each of the comparing LL and HH sub-groups. The watermark pixels are inserted with quality  $x$  into the most extreme coefficient  $M_i$  of every principal component piece  $Y_i$ , where  $x$  is the watermark installing quality.

**Table 2** Sampled watermarked video frame with PSNR value

Sampled watermarked video frame	PSNR value
	85.0278

**Table 3** Result value of NCC for sampled video frame without any attack

Sampled video frame	Sample watermark	Sample watermarked video frame	Extracted watermark	Calculated NCC
				1

Converse DWT is connected to acquire the watermarked luminance segment of the casing. At last watermarked casing is remade and watermarked video is acquired (Table 2).

### 3.2 Filtering

2D Gaussian deblurring channel is connected on each frame having standard deviation of 0.5. It must be connected before watermark extraction prepare.

### 3.3 Proposed Watermarking Extraction System Procedure

The means utilized for watermark extraction is the same as the means in the installing, however in the invert course. As takes after Watermarked video is changed over into frames. Each RGB frame is changed over to YUV portrayal. DWT is connected. LL and HH sub-groups separated into  $n \times n$  non-covering blocks. The removed watermark is contrasted, and the first watermark utilizes NCC, where NCC is the

**Table 4** Result of robustness against various attacks

Different attacks	NCC value	
	PRVWA Algo. [13]	Proposed Algo.
Without attack	0.9973	1.0
Salt and pepper at 0.03	0.9972	0.9985
Salt and pepper at 0.01	0.9973	0.9978
Gaussian at 0.1	0.9968	0.9979
Gaussian at 0.01	0.9974	0.9982
Poisson	0.9974	0.9993
Median filtering	0.9974	0.9986
Contrast adjustment	0.9975	0.9995
Histogram attack	0.9976	0.9983

normalized correlation coefficient. NCC esteem is 1 when the watermark and the extricated watermark are indistinguishable and zero if the two are not the same as each other. Calculated NCC of extracted watermark gives 1.

## 4 Experimental Results and Discussion

Above algorithm is connected to an example video succession *suzie.mpg* utilizing watermark logo “secret.png.” The first examined frame and its relating watermarked frame are presented in Table 4, and watermarked outline shows up outwardly indistinguishable to the first. The performance of algorithm can be measured as far as its intangibility and heartiness against the conceivable attacks. Watermarked frame is subjected to an assortment of attacks, for example, salt and pepper, Gaussian, Poisson, median filtering, contrast adjustment and histogram attack. To evaluate the performance of any watermarking framework, PSNR and NCC are utilized as a general measure of the visual nature of the watermarking framework (Figs. 2, 3 and Table 3).

## 5 Conclusion

At last we can conclude that, in the today’s advancing and innovative scenario, the requirement for advanced video watermarking is high and improvement of vigorous systems is a need as far as copyright security and validation. For video watermarking purposes, a noteworthy number of calculations have been acquainted all together with given the most ideal elements required for this concept. Our proposed framework

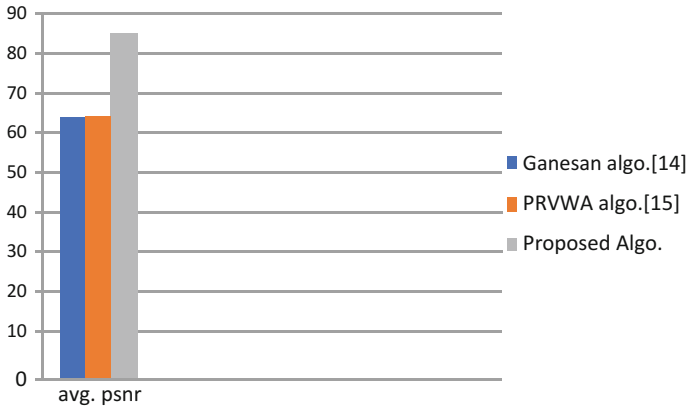
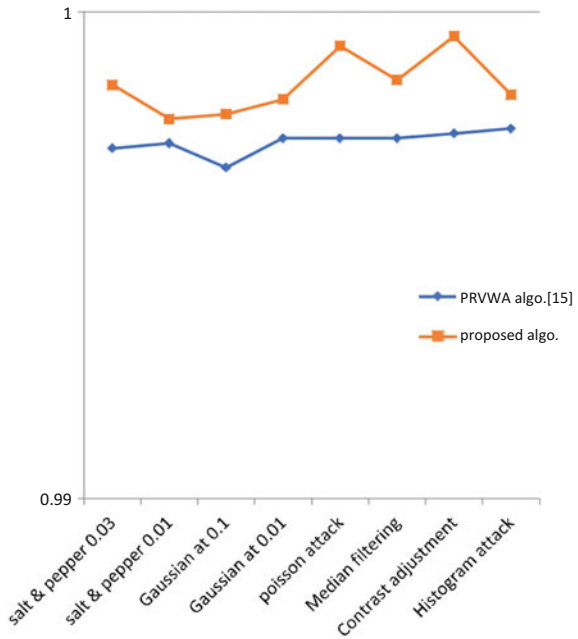


Fig. 2 Comparison of average PSNR value between proposed and previous algorithms

Fig. 3 Comparison of current and proposed algorithm w.r.t NCC values under different attacks



improves some parameters based on the present calculation given by [13]. Without influencing computational time, the proposed framework has accomplished this.

Results are improved even though the frame size is  $352 \times 288$  (which is  $176 \times 144$  in [13]) and the watermark size is  $256 \times 198$  (which is  $80 \times 80$  in [13]), which is quite larger.



## References

1. Rewani, R., Kumar, M., Pundir, A.K.S.: Digital image watermarking: a survey. *Int. J. Eng. Res. Appl.* **3**, 1750–1753 (2013)
2. Poljičak, A., Mandić, L., Kurečić, M.S.: The influence of image enhancement filters on a watermark detection rate. *Influ. Image Enhanc. Acta Graphica* **22**(3–4), 53–60 (2011)
3. Bouslimi, D., Coatrieux, G.: A crypto-watermarking system for ensuring reliability control and traceability of medical images. *Signal Process. Image Commun.* **47**, 160–169 (2016)
4. Azeem, N., Ahmad, I., Jan, S.R., Tahir, M., Ullah, F., Khan, F.: A new robust video watermarking technique using H.264/AAC codec Luma components based on DCT. In: *Int. J. Adv. Res. Innov. Ideas Educ.* **2**, 2395–4396 (2016)
5. Buhari, A.M., Ling, H.C., Baskaran, V.M., Wong, K.S.: Fast watermarking scheme for real-time spatial scalable video coding. *Signal Process. Image Commun.* **47**, 86–95 (2016)
6. Rasti, P., Samiei, S., Agoyi, M., Escalera, S., Anbarjafari, G.: Robust non-blind color video watermarking using QR decomposition and entropy analysis. *J. Vis. Commun. Image R.* **38**, 838–847 (2016)
7. Karmakar, A., Phadikar, A., Phadikar, B.S., Maity, G.K.: A blind video watermarking scheme resistant to rotation and collusion attacks. *J. King Saud Univ. Comput. Inf. Sci.* **28**, 199–210 (2016)
8. Thomasa, N., Thomasb, S.: A low distortion reversible data hiding technique using improved PPVO predictor. In: *International Conference on Emerging Trends in Engineering, Science and Technology. Procedia Technol.* **24**, 1317–1324 (2016)
9. Swaraja, K., Madhavelatha, Y., Reddy, V.S.K.: Robust video watermarking by amalgamation of image transforms and optimized firefly algorithm. *Int. J. Appl. Eng. Res.* **11**, 216–225 (2016)
10. Haribabu, M., Bindu, C.H., Swamy, K.V.: A secure and invisible image watermarking scheme based on wavelet transform in HSI color space. In: *International Conference on Advances in Computing and Communications*, pp. 6–8 (2016)
11. Chung, K.L., Yang, W.N., Huang, Y.H., Wu, S.T., Hsu, Y.C.: On SVD-based watermarking algorithm. *Appl. Math. Comput.* **188**, 54–57 (2007)
12. Loganathan, A., Kaliyaperumal, G.: A robust color video watermarking scheme based on hybrid embedding techniques. *Multimed. Tools Appl.* · (2015)
13. Maharjan, R., Alsadoon, A., Prasad, P.W.C., Rahma, A.M.S., Elchouemi, A., Senanayake, S.A.: A proposed robust video watermarking algorithm: enhanced extraction from geometric attacks. In: *International Conference on Multimedia and Image Processing* (2016)
14. Huang, H.C., Chang, F.C., Chen, Y.H., Chu, S.C.: Survey of bio-inspired computing for information hiding. *J. Inf. Hid. Multimed. Signal Process.* **6** (2015)
15. Mohammad, A.A., Alhaj, A., Shaltaf, S.: An improved SVD-based watermarking scheme for protecting rightful ownership. *Signal Process.* **88**, 2158–2180 (2008)

# Cryptanalysis on Digital Image Watermarking Based on Feature Extraction and Visual Cryptography



Neha Shashni, Ranvijay and Mainejar Yadav

**Abstract** This paper presents the cryptanalysis on a method that uses the visual cryptography and feature extraction for watermarking in images. Many of the existing digital watermarking techniques use a method which does not involve embedding of the watermark into the vulnerable images, but rather it only uses the features of image as the security key for generating watermark. This is assisted with the visual cryptography which makes it possible to perform watermarking without embedding any data onto it. Although they perform very well with many different attacks like scaling, rotating, compression, different kind of noises, etc., it still fails to prove the copyrights of image when the attack as illustrated in the paper is applied to the watermarked image. This method allows the attacker to use his own valid watermark to be introduced to the image, which will result in failure of the owner to prove its copyright.

**Keywords** Cryptanalysis · Digital watermarking · Feature extraction  
Visual cryptography

## 1 Introduction

Cryptanalysis refers to the study of ciphers in a way that finds out the weaknesses in the systems that claim to be a security system that means the system that involves a secret coding. It can be either weaknesses or breaking the system. Breaking the system is in the sense that the attack method is capable of revealing the secret from the cipher code without knowing the key or algorithm of the cryptosystem. It may either use the brute force or some advancement in it to detect the fault in the method. Weaknesses can also refer to finding the fault or such property in the design or implementation of cryptosystem that can reduce the complexity of the brute force.

---

N. Shashni (✉) · Ranvijay · M. Yadav  
Computer Science Engineering Department, MNNIT Allahabad, Allahabad 211004,  
Uttar Pradesh, India  
e-mail: shashnineha123@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_39](https://doi.org/10.1007/978-981-13-1708-8_39)

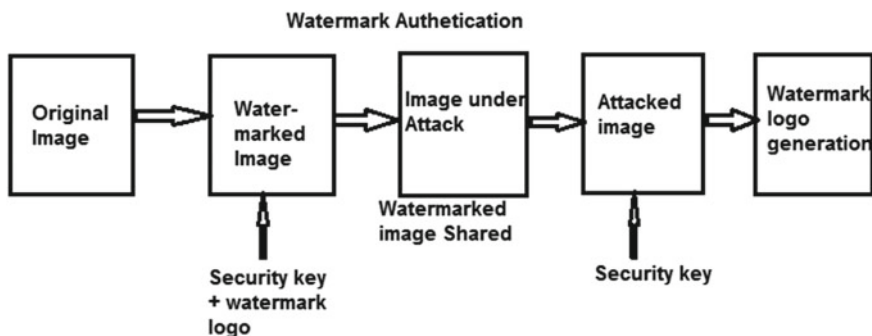
Cryptanalysis in the digital watermarking can be performed in order to analyze the security of watermarking method and to see how efficient it is in providing the basic properties of a method. It is also done to see whether there is some possibility that watermarking method is breakable that is not sufficiently capable of proving the real owner of the multimedia.

There are number of attacks to perform cryptanalysis. Some of them are as follows:

1. Known-plaintext analysis: In this technique, the cryptanalyst has the knowledge of some of the plaintext that is a portion of secret from the ciphertext using which he will attempt to deduce the keys to break the system.
2. Chosen plaintext analysis: This is when the cryptanalyst tries to deduce the key by comparing the plain text and ciphertext, thereby breaking the system.
3. Ciphertext-only analysis: This means that the cryptanalysis will only use the ciphertext to break the system. In this method, an accurate guesswork is needed.
4. Man in the middle: In this technique, no text is analyzed but rather it is a forced method to reveal the secret somehow by either forcing a member or some other way.

In this paperwork, the cryptanalysis of the digital watermarking method falls under the known-plaintext analysis but not completely. Since the cryptanalysis does not depend on the owners algorithm of watermarking, that is the algorithm is not known nor the plaintext, but it is still able to break the watermarking method (Fig. 1).

The original image is made to undergo the watermarking process by any watermarking method that owner has chosen. Owner uses a secret image or logo called watermark which is used with security keys that owner keeps with him and generates the watermarked image. This image is then transmitted to the media or anyplace where owner wants to use it. So, now it is shared openly. When an attacker finds the image and wants to use it illegally, he will either make it go through some attacks so that owner will not be able to authenticate the image or use it as it is for his own purpose. Now, when the owner gets to know about the illegal use of his image, the owner will claim that this image is his and not the one who is using it illegally. To prove this, the owner will try to generate the watermark on the image that is



**Fig. 1** General authentication technique of digital watermarking system

being shared by the attacker and the attacker will have to justify it. This is called the authentication, and if owner is successfully able to prove his authenticity over the image using some watermarking method, then the watermarking method is supposed to pass the cryptanalysis successfully.

As we know that the Internet has led to a profound development in the era of computer science generation, but it has also led to many copyright issues of data. The digital media has took a pace over than the other data, but its vulnerability is a risk, because in the world of Internet it is easy to copy the image without any loss of quality, also there is no limit on the number of times it can be shared. There have been many researches proposed for the protection of the copyright of digital media in the last two decade, which can be categorized into groups mainly spatial and frequency domain. But not all of them are able to fulfill all the requirements for the copyright protection. For an example, some methods may provide good quality of the image, but image watermarking may not be able to work under different kinds of attacks, while some may provide good security against the attacks but fail to maintain the quality of the image because of watermarking embedding. So, the motive was to develop a method that would provide a great deal of security plus maintaining the quality of the image, whether or not the watermark is visible in it.

Watermarking techniques complement encryption by embedding a secret imperceptible signal, a watermark, directly into the original data in such a way that it always exists here. The following purposes are used for such type of watermarks.

1. Owner identification of digital multimedia: Identifying the owner of specific digital work like photography, music albums or videos, etc., can be difficult task.
2. Digital art work: There are so many kinds of digital art works nowadays, for example a cartoon character. It can be made sure that this is not used by any unauthorized party by using the watermarking.
3. Medical Application: To identify a patient in his medical document like CT scan report, medical images by embedding the name or information related to patient. It is very crucial that his critical information is not risked while embedding the data.
4. Broadcast: While broadcasting, for example, an advertisement so that owner can be sure that his advertisement is aired on everywhere safely. The monitoring results of this can be used for royalty or copyright protection purposes.
5. Transaction tracking: It can play a very important role in maintaining the track of every transaction history by embedding the critical information as watermark onto the receipt.
6. Medical application: Embedding the date and the patient's name in the medical images could.
7. Internet: The movie producer can know which recipient was the source of leak of movie.
8. Identity card: The biometric information can be embedded onto the identity information of any person as a watermark.

Watermarking scheme must be robust and reliable because it is used for security purposes. Robustness of the scheme against the attacks depends on the algorithm

which is being used for embedding; watermark should be hidden such that no one should be able to find the watermark other than the authenticated users. One more aspect is also there which is equally important for the scheme which is quality of the multimedia content that is watermarking technique should not so much affect the quality of the image.

Cryptanalysis over the digital watermarking method is of a great importance. This shows us how much secure is the watermarking method and of what potential it is so that the system is unbreakable. It gets the idea whether it is safe to allow the method to be applied practically or not.

## 2 Related Work

In the digital watermarking area, there has been a lot of work already done because of its versatility of use and effectiveness. Numerous methods are also developed for the same purpose which comes under the space domain or frequency domain, i.e. the encryption methodology. But many of the methods also employ the visual cryptography that may use any of these two domains, which depends on the method. Visual cryptography is a technique of hiding the secret image which involves generating the shares out of the secret image and sharing them with different members so when a minimum required number of shares from different members are overlapped, only the secret information can be revealed. These shares are unrecognizable from the naked eyes and seem almost meaningless random pixel values. This methodology is also used for watermarking purpose by using some of its security features in different ways. It can be seen in numerous methods.

Naor and Shamir [1] proposed the concept of visual cryptography. This concept is being used by many authors for different types of applications by adding something extra and constantly improving the method. Actually, visual cryptography is described as a secret sharing scheme extended of digital images, and this will be discussed in detail in the next section.

Some researchers propose detection-based watermarking techniques such as one based on visual cryptography (VC) that does not alter the original image in order to preserve the visual quality of the image, but generates two shares known as the ownership share, which is registered to a certified authority (CA) and is kept with the owner so that it can be used later for verification and identification share, which is generated from the suspected copyrighted document, to be used with the ownership share [2]. Hwang proposed a scheme based on VC that uses the (MSB) for comparison with the global mean of the intensity of the image in generation of the shares [3]. Hwang proposed a method [4] for the watermarking which uses the VC in combination with DCT of the blocks and compared it with mean coefficients.

Method proposed by Pushpa Devi et al. [5] presented a new method for watermarking of images based on visual cryptography technique. The two shares generated using VC are used: One is ownership share that is generated in the watermark registration phase, and the other one is master share that is used with this ownership share

to retrieve the watermark during identification phase. These shares do not reveal anything individually and generate watermark when both are stacked together. The ownership share is generated by comparing the pixel values against the mean of the pixel value of scrambled image in a block. It used Cat Map for scrambling of image. This method does not really involve the embedding part of watermark and can efficiently work with the most of the attacks. This watermark fails to prove the original copyright when the attacker also introduces his own valid watermark to the watermarked image. Since the image shared was not made to undergo any changes, this favors the attacker by sharing the original image as it is.

Fridrich et al. [6] in her thesis work presented a cryptanalysis on a watermarking technique where she has shown that the method is vulnerable in different types of substitution and impersonation attacks. In this paper, the author has used the same security key that is insertion function which means that the attacker is aware of the watermarking method applied onto the image. But in the proposed work, the attacker need not to be aware of the watermarking method the owner has applied and still be able to break the system that is any kind of watermarking system.

In paper [7], Daniel used two attacks on Teng et al.'s fragile watermarking algorithm and both of them allow the attacker to apply valid watermarks to the watermarking method.

In papers [8–13], some methods are proposed which are related in the sense that they require no embedding in order to do the watermarking.

### 3 Proposed Work

The concept of attacking by simply embedding the new watermark in the already watermarked image as discussed earlier is proved in this paper on the method in [5].

The algorithm is explained in detail here. The author has used the concept of visual cryptography as a method of watermarking, by generating two shares using original image and watermark. It tells that one of the shares is kept secret key with the copyright holder, while the other share can be again extracted during watermark extraction. So, according to VC when these two shares will be stacked together, the secret information that is the watermark should be visible. Also the Cat Map is used on the watermark and the luminance channel of image before using it to generate share for the additional security purpose. The luminance channel is partitioned into blocks where the mean and middle values of block are compared. Using the codebook, the ownership share is generated using the obtained values in block and the Cat Mapped watermark. Here  $a$ ,  $b$ ,  $P1$ ,  $P2$  and ownership share are the security keys that are kept with the owner and are used while the extraction process. Also a thing to be noted here is the original image does not undergo through any changes and is shared as it is.

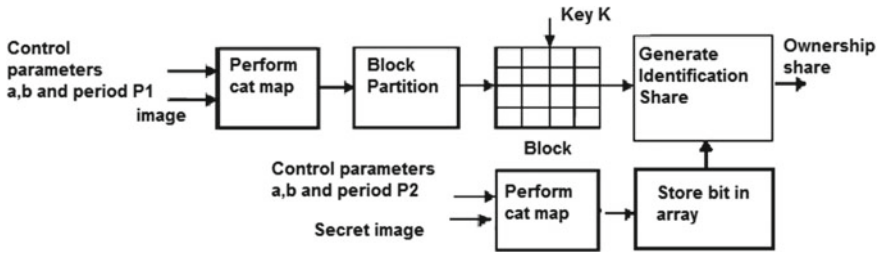


Fig. 2 Watermark embedding phase

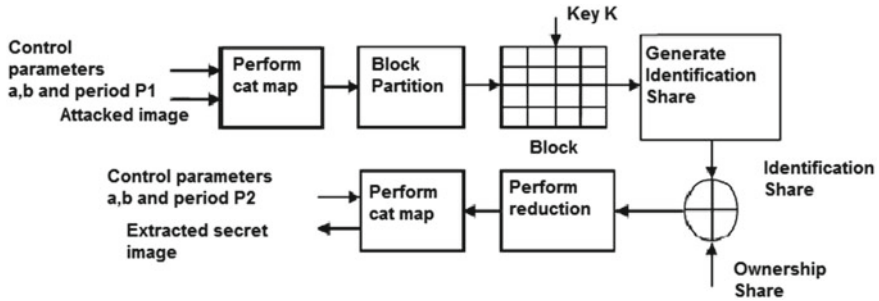


Fig. 3 Watermark extraction phase

During the watermark authentication process, the share is again generated which is called the identification share by undergoing through the same process as in the ownership share generation but using the second codebook. It will be again Cat Mapped for the rest of the period when the watermark is obtained by stacking the ownership and identification shares to finally make the watermark readable. The watermark embedding procedure can be seen in Figs. 2, and 3 represents the watermark extraction phase.

The codebook in Tables 1 and 2 is used in order to generate the ownership share and identification share, respectively, using the watermark values and feature values of image using visual cryptography.

This watermarking method is used as an attack in this paper. This favors the attacker because of its feature extraction properties, which makes it difficult to detect and to remove; moreover, the image is not modified with this method. The next section shows the experiment results when this method is used by attacker.

**Table 1** Codebook C1 for generation of ownership share

Feature	$B_i(3,3) < \mu_i$						$B_i(3,3) \geq \mu_i$											
	0			1			0			1								
scrambled bit	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
$\text{mod}(i,6) =$	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
Ownership share																		



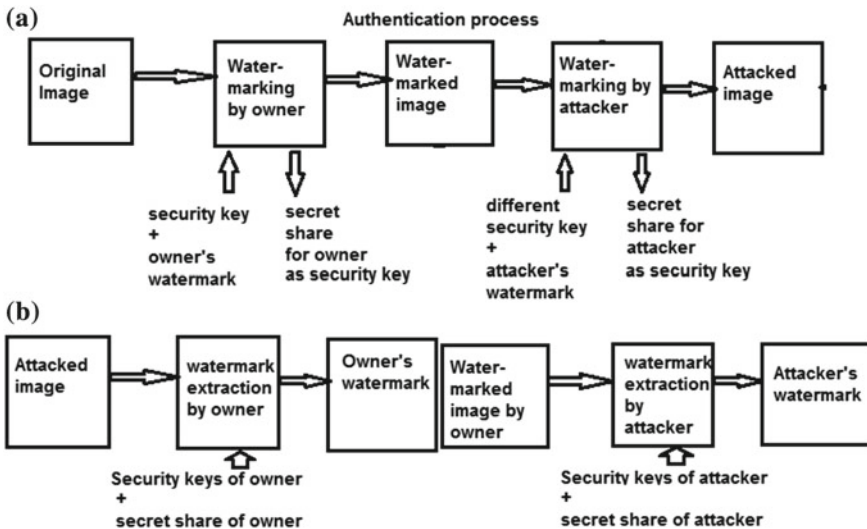
**Table 2** Codebook C2 for generation of identification share

Feature mod (i,6) =	$B_i(3,3) < \mu_i$						$B_i(3,3) \geq \mu_i$					
	0	1	2	3	4	5	0	1	2	3	4	5
Identificat ion share												

### 4 Experiment Results

In this experiment, the watermarking technique is also the same as the one explained above. That means both the methods by owner and attacker are same. The attacker is using another valid binary watermark which is not related to the one used by owner in anyway. Here, the attacker does not know about the security keys of owner and is still able to authenticate itself. Authentication process is shown in the figure. Figure 4a shows the process of watermarking by owner and communicated, which is then attacked by the attacker. Figure 4b shows the authentication process of the watermark of owner and the attacker. Both will have to use their keys and prove their copyright onto the image. Now if both the attacker and owner are successfully able to generate their watermarks, this will clearly result in the failure of the owners watermarking method to protect the copyright of the image.

In this experiment, we have used the ‘Lena’ image on which the watermarking is applied. In Fig. 5, we can see the image (Fig. 5c) and the watermark (Fig. 5a) used by the owner. In this method, the image size is required to be more than the 8 times

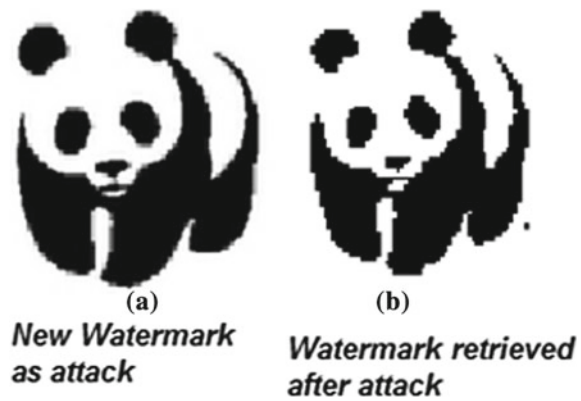


**Fig. 4** Attacking model

**Fig. 5** Watermark used by the owner, watermark retrieved when authenticated and the image used in the experiment



**Fig. 6** Watermark used by attacker and the watermark retrieved when authenticating attacker's watermark



multiple of the binary watermark chosen. Or it can be said that the size of watermark logo should be less than or equal to the 1/8th times of dimensions of image.

The watermark image used by the attacker is in Fig. 6a. Here in Fig. 6, the owner uses its own key ( $a = 2$  and  $b = 3$  and the ownership share 1 generated by the owner) to retrieve the watermark (Fig. 6b) and share the image as it is, because this method does not involve any change to the image. Now to that same image but with a different watermark and different key ( $a = 3$ ,  $b = 4$  and the ownership share 2 generated by the attacker), the attacker is able to retrieve its watermark (Fig. 6b) with a good quality as well. Now, the watermarks retrieved by both the attacker and owner are shown in Fig. 6 and Fig. 7, respectively, since there is no change in the original image. Both of the owner's and attacker's watermarks are matching with the ones that are retrieved from the image. So, in the conclusion, both the owner and attacker are able to generate their watermark using their secret shares and secret keys with an equal visibility and clarity. This does not prove the original owner of the image, and owner fails to prove the claim (Table 3).

**Table 3** Experiment result in PSNR

Image	Retrieved owner's watermark	Retrieved attacker's watermark
PSNR (dB)	99	99
KEY	a = 2, b = 3	a = 3, b = 4

## 5 Conclusion

This paper presents the cryptanalysis done on the watermarking method that uses the feature extraction and visual cryptography. The attacker is able to introduce new and valid watermark to the images shared by owner after watermarking. This can be a big threat to watermarking method because of its property that it is able to use only the features and needs no embedding. Since there is almost no embedding done to image, no one other than owner will be able to figure out what features are used to generate the secret share. This property was considered as robust for the method, but the same property leads the attacker to prove some other watermark onto the image which is a big threat. This method of attacking can be used on other watermarking methods also because this method will work for all the identical images as far as the features that are used by the attacker to introduce his watermark are still there. This means that the owner will fail to claim his copyright onto the image. The decision cannot be made in favor of any party as both are successfully able to generate their own watermark due to which copyright dispute is still left unresolved.

**Acknowledgements** This work is supported by CSED MNNIT Allahabad, Allahabad.

## References

1. Naor, M., Shamir, A.: Visual cryptography. In: Proceedings of the Advances Cryptology. EUROCRYPT94. LNCS, vol. 950, pp. 1–12. Springer-Verlag, Berlin (1995)
2. Hsu, C.S., Hou, Y.C.: Copyright protection scheme for digital images using visual cryptography and sampling methods. *Opt. Eng.* **44**(7), 077003–077010 (2005). <https://doi.org/10.1117/1.1951647>
3. Hwang, R.J.: A digital image copyright protection scheme based on visual cryptography. *Tamkang J. Sci. Eng.* **3**(2), 97–106 (2000)
4. Rawat, S., Raman, B.: A publicly verifiable lossless watermarking scheme for copyright protection and ownership assertion. *Elsevier AEU* **66**(11), 955–962 (2012). <https://doi.org/10.1016/j.aeu.2012.04.004>
5. Pushpa Devi, B., et al.: A watermarking scheme for digital images based on visual cryptography. *Contemp. Eng. Sci.* **8**(32), 1517–1528 (2015)
6. Fridrich, J., et al.: Cryptanalysis of the Yeung-Mintzer Fragile Watermarking Technique. SUNY Binghamton, Binghamton, NY
7. Caragat, D., et al.: Cryptanalysis of an improved fragile watermarking scheme. *AEU-Int. J. Electron. Commun.* (2016)
8. Abusitta, A.H.: A visual cryptography based digital image copyright protection. *J. Inf. Secur.* **3**, 96–104 (2012)

9. Hassan, M.A., Khalili, M.A.: Self watermarking based on visual cryptography. *World Acad. Sci. Eng. Technol.* **8** (2005)
10. Surekha, B.: A multiple watermarking technique for images based on visual cryptography. *Int. J. Comput. Appl.* (0975–8887) **1**(11) (2010)
11. Shao, Z.: Robust watermarking scheme for color image based on quaternion-type moment invariants and visual cryptography. *Signal Process. Image Commun.* **48**, 12–21 (2016)
12. Datta, S., Nath, A.: Data authentication using digital watermarking. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2**(12) (2014)
13. Saturwar, J., Chaudhari, D.N.: Digital watermarking scheme for secret images using visual cryptography. *Int. J. Sci. Eng. Res.* **4**(6) (2013). ISSN 2229-5518

# A Spoofing Security Approach for Facial Biometric Data Authentication in Unconstrained Environment



Naresh Kumar and Aditi Sharma

**Abstract** Security is ever a challenging issue of research at national and international level due to the privacy standards of object detection and recognition in unconstrained environment. Almost real-life activities are performed under the unconstrained conditions in which exact determination of any activity fails due to the lack of complete information of facial parts, hands and relevant objects. However, from social interaction views, face detection is a quite saturated research in normal conditions but, due to highly sensitive and easy availability of facial data, encourages the researchers to work by cryptographic aspects in unconstrained environment. In this work, we focus on security issues for automated facial biometric data authentication by local features extraction. By keeping space and time intricacy, we ensure the fusion of Gabor, center-symmetric LBP and discriminative robust LBP features to improve the performance. The feature matching is performed by majority of vote in which difference of Gaussian and robust local ternary pattern is used. Since we choose one sample to match with stored faces, the reduction in space complexity improves the performance up to 89% of matching accuracy.

**Keywords** Automated face recognition (AFR) · Local binary pattern (LBP) · Center-symmetric local binary pattern (CS-LBP) · Robust local ternary pattern (RLTP) · Discriminative rotational local binary pattern (DR-LBP)

---

N. Kumar (✉)

Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee 247667, Uttarakhand, India  
e-mail: atrindma@iitr.ac.in

A. Sharma

Department of Computer Science & Engineering, MBM Engineering College, Jai Narain Vyas University, Jodhpur 342011, Rajasthan, India  
e-mail: aditi11121986@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713, [https://doi.org/10.1007/978-981-13-1708-8\\_40](https://doi.org/10.1007/978-981-13-1708-8_40)

# 1 Introduction

Security issues due to human being are commonly resolved by biometrics [1, 2] which include hand geometry, fingerprints, signature, iris and gait recognition. This is noticeable that publically available facial information can create the biased issues in the research which can be considered either as rich source of information (as almost the human perspective can be predicated by facial information being very rich source of data) or as a high risk of theft of the personal information. The progress in smart sensor based technology ensures that surveillance of personal facts can be disclosed easily [3, 4]. Photographs are used for any government or private registration to verify the face shape, structure and global information about all facial parts which is integrated with residential information. The still image of frontal face [5] can give sufficient information, but it happens in a very rare context of real-time automated secure applications.

## 1.1 Automated Face Recognition System

Recognition with automation is demanded in every domain of organization. Takeo Kanade [6] developed first automated system to recognize psychology of facial features which gave poor results to deal big dataset in the context of processing aspects. The problem [6] was resolved by Kirby and Sirovich [7] in which Karhunen-Loeve transform is used to drive features representation. The sounding work by Turk and Pentland on Eigen's face [8] established benchmark for automated face recognition algorithms.

The final stage of face recognition [9] is divided into verification and identification which is simply comparing the test face image with known face image of dataset. Highly controlled imaging conditions [1], such as passport photographs, are resulted high error rate. Training of dataset in less controlled imaging conditions [5] like illumination, camera movement and variant pose, can be considered a hot research problem. The common limitation of face biometric identity algorithms is to process fairly restrictive and labor-intensive training data. Highly distinct configuration of human face is contributed by internal and external facial components from which seamlessly geometrical and vision based is expected for unconstrained face detection and recognition.

## 1.2 Motivation of Face Recognition System

Human vision is highly focused on facial parts for any operations from psychovisual prediction to highly secured defense mission or health issues. Photographs are widely augmented with the manual directory as a common biometric authentication

like driving license, passport and academic institutions. The frontal or lateral face image captured in natural conditions can be verified to find missing children, electoral registration, banking and searching for police booking, etc. Facial marks and its activities in augmentation can highlight the international most wanted terrorist. Difference of Gaussian (DoG) and robust local ternary pattern (RLTP) are applied as a local texture-based matching method for face recognition. In the first phase, facial features are extracted to provide edge information at various illuminations using DoG at face images. The discrimination for texture and edge is created by RLTP which is a useful feature to distinguish the samples accurately. The novelty of the work is produced at the reduction of space complexity due to single sample which is sufficient for improving performance in our results.

## 2 Related Work

Face detection from spoofing security aspects is highly sounding in vision and cryptographic [2, 10] approaches. Automatic face recognition by Lenc and Král [11] is proposed in unconstrained environment to support SIFT-based Kepenekci approach that is verified highly scoring on a large dataset. Confidence measure was highlighted as a future issue on large dataset. Human biometric [12] opens all about traits and behavior by computing features from DNA, iris, palm prints, fingerprints retina and other facial components. Facial landmark and texture features [13] are computed to use facial expression as a facial biometric authentication. The face recognition is simply face verification which is obtained by robust invariant pose alignment methods proposed by Li et al. [9] where it is ensured that high-resolution data based on pore scale facial features, PPCASIFT, outperforms better than PSIFT by reducing computational time. Further, an efficient automatic face recognition is proposed by Aly et al. [1] based on SIFT algorithms for computing invariant local features. In the same domain, Bay et al. [14] introduced speed-up robust (SURF) features to compute key points that condensed highly sensitive features. The face image highly invariant to various modes of facial expressions and lighting direction is represented by Belhumeur et al. [15], ensuring that Fisher's face has higher vote in comparison with Eigen's face for computing error rate. Face detection technique which is invariant to illumination and noise is proposed by Faraji and Qi [16] which extended LDP to eight LDP (ELDP) for illumination insensitive and robustly noise-free face representation. Andries et al. [17] focused on basics of computer of computer vision problems, detection, tracking and recognition and ensured that the techniques can show better results for multimodal object tracking. Elastic Bunch Graph Matching (EBGM) algorithm by Kepenekci [18] which outperforms the classical algorithms for face detection and Gabor filter is used to get dynamic face landmark profile.

The novelty of the work proposed by Lu et al. [19] is discriminative manifold and subspace learning for single face per person (SSPP) in which dimensionality issues are resolved by single training sample. Face normalization is a basic step of training phase which is performed by eye-centric features by Dutta et al. [20] for

improving the performance of facial representations. Variation in face image [21] is controlled by edge map, intensity derivative and convolution by Gabor filter of face image. The complexity of human face recognition is introduced [22] in the context of various picture-processing aspects. Deng et al. [23] proposed locally matching pursuit methods outperform traditional ICA, Fisher's face and Eigen's face for high-dimensional facial space. Eye detection of mobile iris and gaze estimation [24] is closely related to face biometric which is proposed to work in unconstrained conditions.

The cryptographic aspect in [25] as a hash biometric of face bit extraction is proposed which compromises the dimension of big data of varied samples. The recent technology for face recognition in highly varied big data is introduced [26] which proposed rank-order clustering for big data of unlabeled face image and clustered them according to their individual identity. Rank-order clustering uses the concepts of deep learning and gives better results than traditional clustering as k-means and spectral clustering. Deep convolution network (DCN) was integrated with linear SVM [27] to authenticate one-to-many as a generalized face biometric parameter. This model archives state of the art for DCN-based deep learning biometric authentication. Overall literature on face recognition in constrained environment is still immature in security aspects.

### 3 Proposed Methodology

To deal with the security issues in visual media, texture based local features and its variations like DR-LBP, R-LTP are used to achieve the sounding performance. In this section, we present the proposed work for facial data authentication.

#### 3.1 Facial Gabor Template

Gabor filters have high range of applications including OCR, fingerprints and iris recognition. In this work, the objective of preprocessing phases includes noise removal and accuracy of face recognition for which we choose Gabor filter to detect linear edge in face image. The orientation or rotation by an angle to a face image ranges from 0 to 270°, and optimal joint localization is computed in spatial and frequency domain. The combination of real and imaginary parts of Gabor filter in orthogonal direction is given in (3.1), represented by  $x' = x \cos \theta + y \sin \theta$  and  $y' = x \sin \theta + y$ .

$$s(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\left(\frac{x'^2 + y'^2 \lambda^2}{2\sigma^2}\right) \exp\left(i\left(2\pi \frac{x'}{y} + \sigma\right)\right)\right) \quad (3.1)$$



In (3.1),  $\lambda$  is used as wavelength parameter of 5, the envelope of Gaussian is  $\sigma$ , and  $\gamma$  is considered as a spatial aspect ratio with range (0.5). Again,  $\psi$  is the phase offset range from 0 to  $\pi/2$  and  $\theta$  is the orientation of angle ranges from 0 to  $360^\circ$ . Gabor filter  $\psi_{(f,\theta)}(x, y)$  in 2D is given in (3.2) which is a complex sinusoidal signal for modulated Gaussian kernel.

$$\psi_{(f,\theta)}(x, y) = \exp\left[-\frac{1}{2}\left\{\frac{x^2\theta_n}{\sigma^2_x} + \frac{y^2\theta_n}{\sigma^2_y}\right\}\right] \exp(2\pi fx\theta_n) \tag{3.2}$$

In (3.2),  $\theta$  denotes the orientation of image and  $f$  is central frequency of Gabor function in 2D sinusoidal plane. The convolution is performed by Gabor filter on face image to obtain facial Gabor template. This operation is shown (3.3) at  $x$ - $y$  coordinate system for grayscale  $I(x, y)$  face image, where  $\theta$  is convolutional operator. The magnitude of real and imaginary part of kernelled facial image used as a mask of facial template presented in Eq. (3.4).

$$g_{(f,\theta)}(x, y) = I(x, y)\psi_{(f,\theta)}(x, y) \tag{3.3}$$

$$|g_{(f,\theta)}(x, y)| = \sqrt{R^2g_{(f,\theta)}(x, y) + I^2g_{(f,\theta)}(x, y)} \tag{3.4}$$

### 3.2 Local Texture and Its Variant Features

Local binary pattern (LBP) is basic texture feature in Fig. 1, which is obtained by dividing the image in blocks of  $3 \times 3$  pixels. The binary-coded string of the block is computed by decimal sum of the values which are greater than the central pixel. Rest of values in neighboring pixels reduces to zero. The complete procedure for LBP computation is explained in Fig. 2. Finally, histogram is computed from decimal values for all the blocks. So, LBP (3.5) given in Fig. 1 is summing up 100 as  $0+2+4+8+0+32+64$ , which is equivalent to binary code 01110110.

$$LBP_{p,r} = \sum_{p=0}^{p=-1} S(g_p - g_c)2^p \tag{3.5}$$

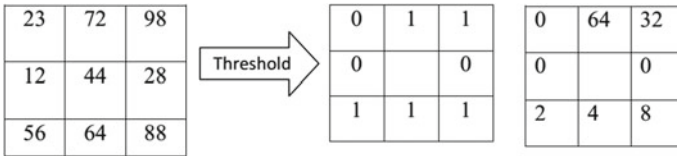


Fig. 1 Example of  $3 \times 3$  LBP

$$S(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \textit{otherwise} \end{cases} \tag{3.6}$$

In the literature, several versions of LBP are available like center-symmetric LBP (CS-LBP), discriminative rotate LBP (DR-LBP), extended LBP (ELBP) and elliptical LBP (EPLBP). Out of these we introduced in this work CS-LBP and DR-LBP only. The local gray-level structure of facial texture is summarized by Ojala [28] in the form LBP. Center-symmetric local binary patterns (CS-LBPs) given in (3.7) were developed for effective regional information. CS-LBP was designed to have greater support in the lower image regions.

$$CS_{LBP_{p,r,t}}(x, y) = \sum_{p=1}^{p=(\frac{r}{2}-1)} S\left(g_p - g_{p+(\frac{r}{2})} - t\right)2^p \tag{3.7}$$

$$S(z) = \begin{cases} 1, & z \geq 0 \\ 0, & \textit{otherwise} \end{cases} \tag{3.8}$$

The computational procedure of CS-LBP values is illustrated in Fig. 2 where 44 is taken as a central value and decimal values of the binary string are computed, which is generated by assigning 1 and 0 to these differences in clockwise fashion. Furthermore,

DR-LBP is introduced by Ojala [28] for gray-level texture in (3.9). If unsupervised texture segmentation highly outperforms, DR-LBP operator is jointly a good option with local contrast measure. The values discrimination for 0 and 1 in (3.9) follows the same as (3.7), and  $i_c$  and  $i_n$  are grayscale values of central pixels  $c$  and  $n$ .

$$DR_{LBP(x_c-y_c)} = \sum_{n=0}^{n=7} 2S(i_n - i_c) \tag{3.9}$$

In case of DR-LBP, a uniform histogram is obtained by applying histogram equalization on the regular grids of cell. The threshold value is chosen for computing the difference against the neighboring values. The positive difference is assigned 1, and negative difference is 0. DR-LBP performance is better due to high dimensional

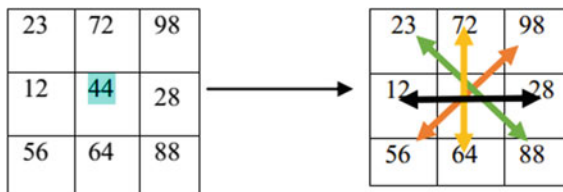


Fig. 2 CS-LBP 3 × 3 matrix

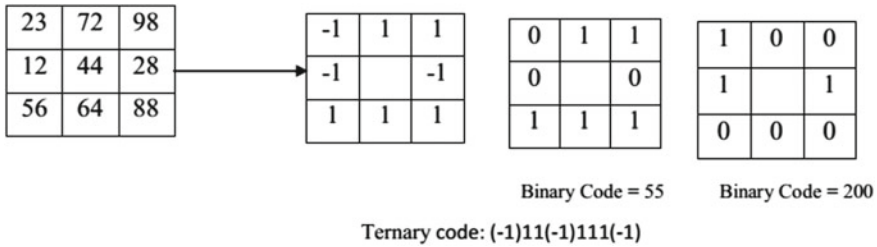


Fig. 3 Splitting of RLTP into positive and negative LBP

reduction. One more extension of LBP is robust local ternary pattern (RLTP) which uses threshold constant for three values instead of binary 0 and 1. These texture features are invariant lighting effects and monotonic gray-level transformation as proved highly discriminative for texture classification. By keeping these views, LTP is introduced (3.10). The complete procedure to compute the values by setting some threshold is explained. In this example, we have chosen threshold value five. The ternary code in the example is (-1)11(-1)111(-1).

$$S(u, ic, t) = \begin{cases} 1, & u \geq ic + t \\ 0, & |u - ic| < t \\ -1, & u < ic - t \end{cases} \tag{3.10}$$

The computational procedure for RLTP is explained in Fig. 3 which represents the splitting of RLTP into positive and negative LBP codes. Two different channels of texture features LBP and LTP have used to proceed our experimental work. Robust local tetra pattern is highly resistive to noise.

### 3.3 Classification of Locally Fused Texture Features

Distance measure-based classification techniques are simple to proceed. The most general form of Euclidean distance equation is represented in (3.11) for facial feature classification.

$$d(x, y) = \sqrt{\sum_{s,i} (x_{s,i} - y_{s,i})^2} \tag{3.11}$$

If we have sample data from specific population, Chi-square test is goodness of fit to apply for classification given in (3.12).

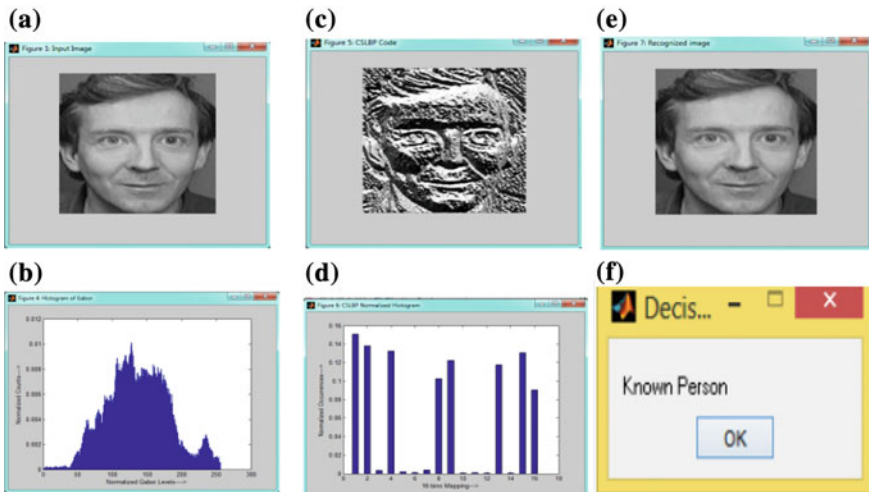
$$d(x, y) = \sum_{i=1}^n \frac{(x_i - y_i)^2}{(x_i + y_i)} \tag{3.12}$$

### 4 Results and Evaluation

The input to process the images is taken from the ORL face database. The statistics of the dataset is described by containing 40 subjects with 10 images for each subject. The experimental work is carried out by randomly choosing 20 images of 3 different subjects, and training set contains 20 images.

#### 4.1 CS-LBP and DR-LBP Results

Histogram of the Gabor face image, Fig. 4a, is shown as sum in Fig. 4b. Similarly, CS-LBP is computed, and the corresponding results are shown in Fig. 4d against the CS-LBP image, Fig. 4c. Further decision is taken to classify the person which is justified in Fig. 4e, f consequently. By the justification as Fig. 4f, we can use feature fusion to make a feature set from whole dataset. The experimental work is performed on database created the same for CS-LBP experiment. Here image size is  $92 \times 112$  and sample pictures are chosen 10. The comparative performance results show that DR-LBP is better for dimensionality aspects and uses less threshold than LBP which is shown in Fig. 5a–f. The numerical data in the graphical representation by histogram range from 0 to 255.



**Fig. 4** a Input image, b histogram of input image, c CS-LBP input images, d histogram of CS-LBP images, e recognized image, f decision feature

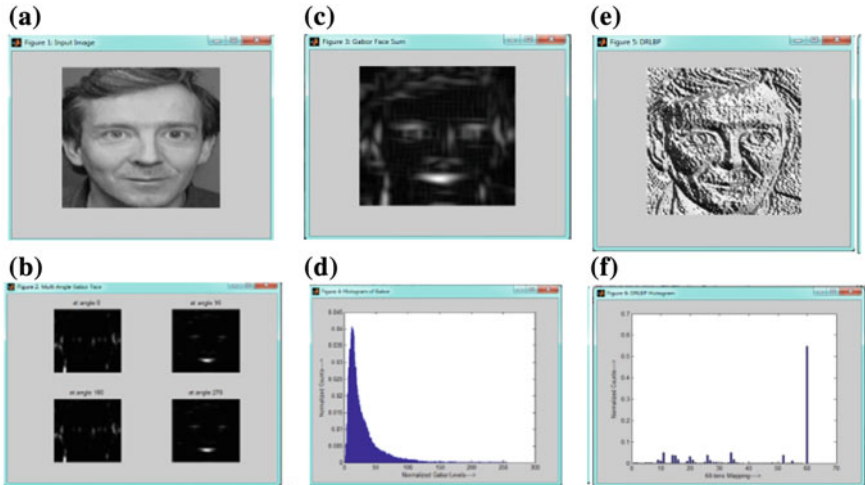


Fig. 5 a Input image, b multi-angle image, c Gabor sum image, d histogram of Gabor image, e DR-LBP image, f histogram of DR-LBP image

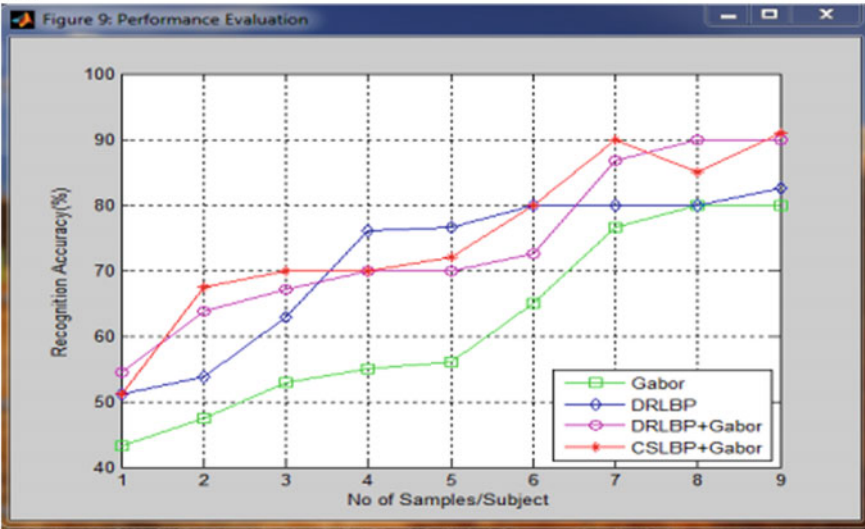
### 4.2 Recognition Rate and Performance Evaluation

In the literature, holistic matching methods are used with ANN, PCA and LDA by employing multiple samples per class which increases training samples space. The novelty of this work is highlighted as employing single sample per class to reduce training space of facial images. It ensures the performance of the system being less sensitive to noise in which space and time complexity is reduced. This happens due to fusion of Gabor filter with CS-LBP and DR-LTP shown in Table 1. The experimental work is processed using standard ORL with the resolution of  $128 \times 128$  images. The comparison of our results is shown in Fig. 6a, b against existing LBP which becomes considerable as number of CS-LBP and DR-LBP images. The novelty consideration of this work is its computational efficiency.

Table 1 Recognition rate comparison

Recognition algorithms	Maximum recognition rate (%)
Gabor	76
DR-LBP	79
Gabor+DR-LBP	84
Gabor+CS-LBP	89
DoG+SIFT	80

(a)



(b)

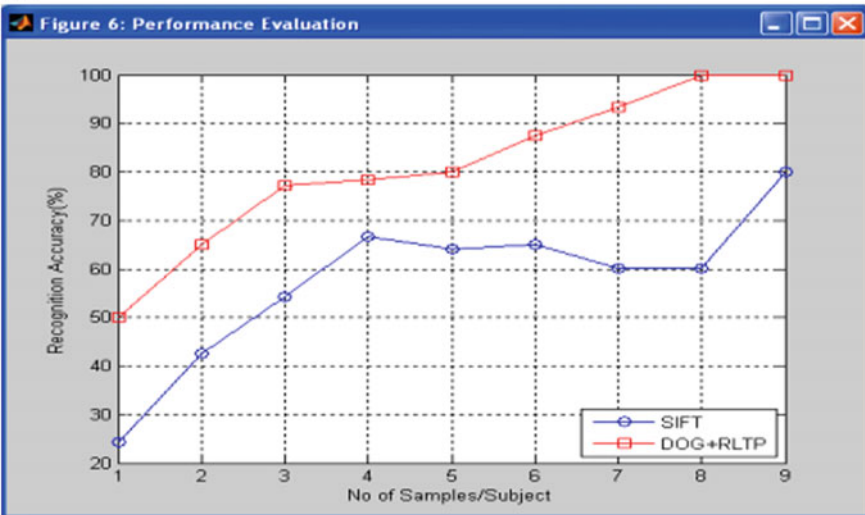


Fig. 6 a Performance graph of LBP methods, b performance graph of LTP methods

### 5 Conclusion and Future Scope

This work demonstrates the securities issues in unconstraint by facial biometric authentication. In the processing context, we used Gabor filter with local texture features with its variation. The face template obtained from textual features is converted into binary template by Gabor filter from which CS-LBP, DR-LBP and LTP

features are computed. We used single sample per class for training which results the enhancement in space and time complexity. The matching by features fusion of Gabor, CS-LBP and DR-LBP images is achieved by applying majority of vote to authenticate the person. Difference of Gaussian (DoG) and RLTP is used to achieve the matching from training samples. The future scope of this work is quite challenging as we think about deep unconstraint features due to various facial actions and fully unconstraint environmental conditions. Furthermore, this research is supposed to achieve keen interest because of easier availability of facial data and visual cryptographic issues related to daily life problems.

## References

1. Phillips, P.J., Grother, P., Micheals, R.J., Blackburn, D.M., Tabassi, E., Bone, M., Face, R.V.T.: Evaluation report. Facial Recognition. Vendor Test 2002, (2003)
2. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. *IEEE Trans. Inf. Forensics Secur.* **1**(2), 125–143 (2006)
3. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Technol.* **14**(1), 4–20 (2004)
4. Ross, A., Othman, A.: Visual cryptography for biometric privacy. *IEEE Trans. Inf. Forensics Secur.* **6**(1), 70–81 (2011)
5. Grother, P.J., Quinn, G.W., Phillips, P.J.: Report on the evaluation of 2D still-image face recognition algorithms. NIST interagency report, 7709, 106, (2010)
6. Wayman, J., Jain, A., Maltoni, D., Maio, D.: *An Introduction to Biometric Authentication Systems*, pp. 1–20. Springer, London (2005)
7. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(1), 103–108 (1990)
8. Turk, M., Pentland, A.: Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**(1), 71–86 (1991)
9. Li, D., Zhou, H., Lam, K.M.: High-resolution face verification using pore-scale facial features. *IEEE Trans. Image Process.* **24**(8), 2317–2327 (2015)
10. Hao, F., Anderson, R., Daugman, J.: Combining crypto with biometrics effectively. *IEEE Trans. Comput.* **55**(9), 1081–1088 (2006)
11. Lenc, L., Král, P.: Automatic face recognition system based on the SIFT features. *Comput. Electr. Eng.* **46**, 256–272 (2015)
12. Sun, Y., Zhang, M., Sun, Z., Tan, T.: Demographic analysis from biometric data: achievements, challenges, and new frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.* (2017)
13. Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., Freeman, W.T.: Face Synthesis from Facial Identity Features (2017). [arXiv:1701.04851](https://arxiv.org/abs/1701.04851)
14. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008)
15. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 711–720 (1997)
16. Faraji, M.R., Qi, X.: Face recognition under illumination variations based on eight local directional patterns. *IET Biom.* **4**(1), 10–17 (2015)
17. Andries, M., Simonin, O., Charpillat, F.: Localization of humans, objects, and robots interacting on load-sensing floors. *IEEE Sens. J.* **16**(4), 1026–1037 (2016)
18. Kepenekci, B.: Face recognition using Gabor wavelet transform. Doctoral dissertation, Middle East Technical University (2001)
19. Lu, J., Tan, Y.P., Wang, G.: Discriminative multimaniifold analysis for face recognition from a single training sample per person. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 39–51 (2013)

20. Dutta, A., Günther, M., El Shafey, L., Marcel, S., Veldhuis, R., Spreeuwens, L.: Impact of eye detection error on face recognition performance. *IET Biom.* **4**(3), 137–150 (2015)
21. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(7), 721–732 (1997)
22. Kanade, T.: Picture processing system by computer complex and recognition of human faces. Doctoral dissertation, Kyoto University, 3952, pp. 83–97 (1973)
23. Deng, W., Liu, Y., Hu, J., Guo, J.: The small sample size problem of ICA: a comparative study and analysis. *Pattern Recogn.* **45**(12), 4438–4450 (2012)
24. Jung, Y., Kim, D., Son, B., Kim, J.: An eye detection method robust to eyeglasses for mobile iris recognition. *Expert Syst. Appl.* **67**, 178–188 (2017)
25. Ngo, D.C., Teoh, A.B., Goh, A.: Biometric hash: high-confidence face recognition. *IEEE Trans. Circuits Syst. Video Technol.* **16**(6), 771–775 (2006)
26. Otto, C., Wang, D., Jain, A.K.: Clustering millions of faces by identity (2016). [arXiv:1604.00989](https://arxiv.org/abs/1604.00989)
27. Crosswhite, N., Byrne, J., Parkhi, O. M., Stauffer, C., Cao, Q., Zisserman, A.: Template adaptation for face verification and identification (2016). [arXiv:1603.03958](https://arxiv.org/abs/1603.03958)
28. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern. Recogn.* **29**(1), 51–59 (1996)



**Part IV**  
**Optical and Wireless Networks**

# Design and Implementation of OFDM Transceiver Using Different Modulation Technique over CDMA



Shikha Bharti, Hemant Rathore, Arun Kumar and Manish Kumar Singh

**Abstract** In this work, OFDM (Orthogonal Frequency Division Multiplexing) is studied for different transmission systems on the basis of BER (Bit Error Rate), transmitting signal, received signal, constellation diagram and the best transmitting scheme is determined for OFDM. Simulation results are utilized to determine: how the Bit Error Ratio (BER) of a transmission varies when signal to Noise Ratio and Multi-Propagation effect are altered for 16-QAM (Quadrature Amplitude Modulation) and QPSK (Quadrature Phase Shift Keying) modulation.

**Keywords** OFDM · QAM · QPSK · ISI · BER

## 1 Introduction

The use of OFDM modulation is being extensively used in wired and wireless communication. The reward of this technique is its increasing usage. The implementation of OFDM is simple; its bandwidth utilization is OFDM being simple; its bandwidth usage is simple and can furnish a high data rate with enough robustness to the channel [1]. The numbers of sub channel and sub-carrier are orthogonal to each other. The sub carriers are huge in number; the class of sub carrier is kept small as possible. The use of OFDM is more because of its ability to over-come the fading effect caused by the multipath interference at the receiver side [2]. Inter Symbol Interference (ISI)

---

S. Bharti (✉) · H. Rathore · A. Kumar · M. K. Singh  
Department of Electronics & Communication, JECRC University,  
Jaipur 303905, Rajasthan, India  
e-mail: shikhabharti712@gmail.com

H. Rathore  
e-mail: rathore1994hemant@gmail.com

A. Kumar  
e-mail: arun.kumar@jecrcu.edu.in

M. K. Singh  
e-mail: mmanishsingh009@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_41](https://doi.org/10.1007/978-981-13-1708-8_41)

and Frequency Selective Fading are two main effects of OFDM. Due to flat channel characteristics of subcarrier, the fading is a simple, handle by using a simple equalizer, but with higher order modulation, more complex equalizer is needed and its hardware implementation become so difficult. The use of OFDM signal is correct if input signal consist of an accurate compensation of carrier frequency offset. A frequency offset as small as few percent is enough for enhancing or predicting the performance of OFDM Receiver [3]. The carrying out of the cognitive radio concept is simple in OFDM since it provides a proven, scalable adaptive technology for wireless engineering. Instead of using a wideband carrier a large act of narrow band parallel sub-carrier are used to convey the data [4]. In [5] OFDM transmitter and receiver is implemented by using a QAM modulation technique and simulation results reveal that PAPR can be reduced by using a technique like clipping and peak cancellation. In [6] OFDM Transceiver for IEEE 802.11 is designed by using an FPGA Kit where area efficiency and low power dissipation model operated at 20 MHz In [7] OFDM with wavelet de-noising technique is designed and simulated a wavelet. The main focus of the work was to replace FFT with wavelet technology. In [8] MIMO-OFDM with different modulation scheme is presented and output result concludes that using V-Blast technique spectral efficiency can be improved [8]. In [9] the performance of an OFDM-CDMA waveform using a different modulation scheme on a high frequency multipath channel is analyzed. In [10] OFDM is designed by using higher radix FFT FPGA where speed and accuracy is optimized and enhanced. In [11] OFDM acoustic is designed whose data rate is 3.2 Kbps for QPSK and 6.4 kbps with 16-QAM modulation at signal bandwidth of 6 kHz. In [12] studied the effect of the transmitter and Receiver IQ Imbalance under the carrier off set in OFDM system. The main purpose of the present work is to evaluate the performance of OFDM for QPSK and QAM modulation scheme.

## 2 Proposed Methodology

A block of symbol is encoded by mean of FFT and transmitted parallel over a number of sub-channels. Cyclic prefix is interleaved in every block of data and it is modulated. After the transmitter antenna, the signals go through all the anomaly and hostility of the wireless channel. Channel estimation is performed using a demodulated pilot. Received data are obtained by using the estimation. At this time channel decoding and de-interleaving is performed to recover the original signal free from any obstacle. The block diagram of OFDM is shown in Fig. 1.

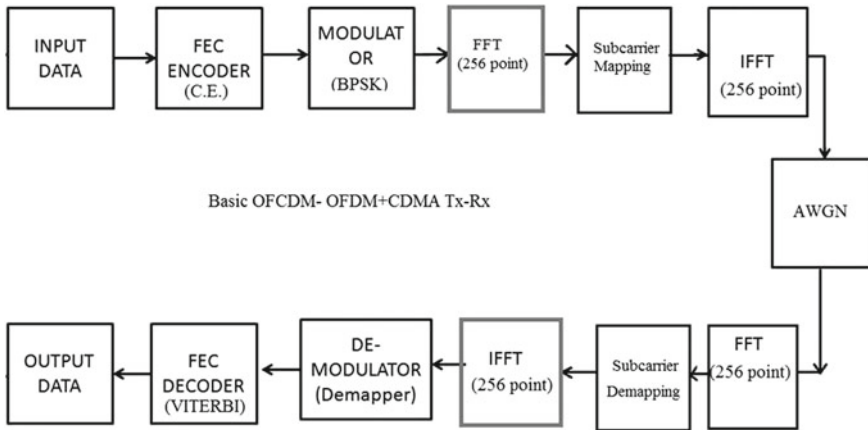


Fig. 1 Block diagram of OFDM

### 3 Result and Discussion

The performance of the proposed work has been studied by using a Matlab.

#### 3.1 Simulated Result for QAM

Figure 2, shows the input signal which is generated by using a random generator. The input signal is sampled and modulated in AWGN (Additive White Gaussian Noise) channel which adds a noise indicated by Fig. 3. The effect of PAPR (Peak Average Power ratio) is introduced in modulated signal results due to multi-path interference of a signal is shown in Fig. 4. At the receiver, the demodulated signal is shown by Fig. 5. Figures 6 and 7 illustrates the fading effect which results in ISI as the signal are very close to each other and even some are overlapping with each other. The scatter plot provides analysis of ISI shown by Fig. 8, it can be conclude that performance of an amplifier is reduced by high PAPR which is considered to be one of the major challenges in OFDM. Figure 9, shows the simulated and theoretical plot of BER versus  $E_b/N_0$ . It indicates that proposed system BER is better than conventional system. The throughput of the system is also analysed by plotting BER for different SNR (Fig. 10).

#### 3.2 Simulated Result for QAM

Figures 11 and 12, shows the input signal generated by random generator. The transmitted data are modulated through an AWGN channel which simply adds a white

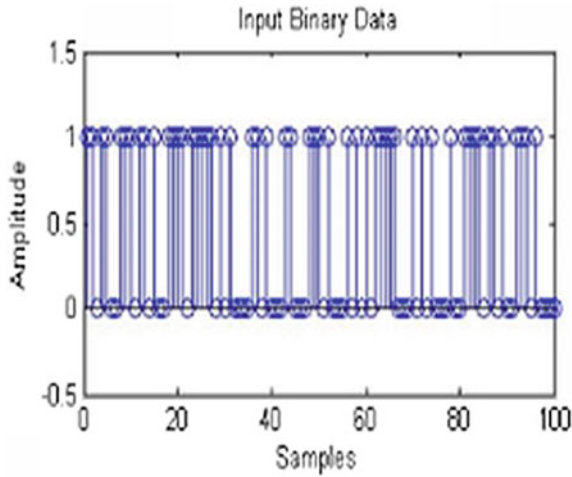
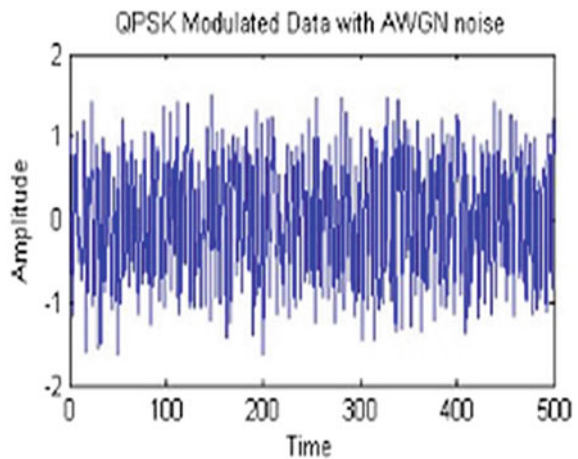


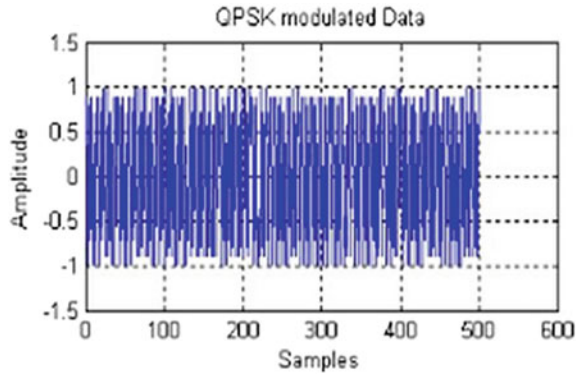
Fig. 2 QPSK binary input data

Fig. 3 QPSK modulated data with AWGN noise

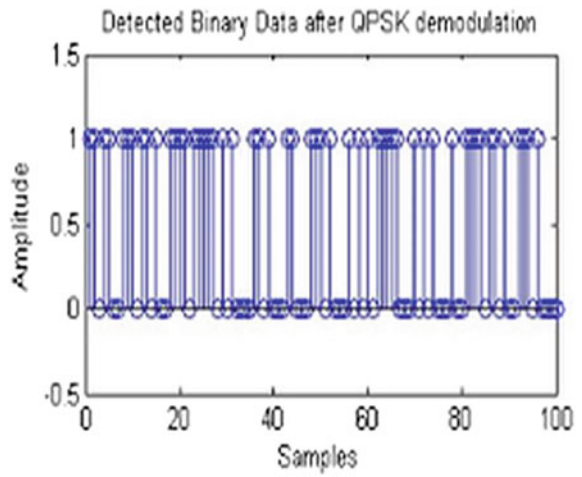


noise to it as shown by Fig. 13. Figures 14 and 15, shows the OFDM signal and the effect of PAPR which is due to multi-path interference. The envelope of a signal continues to display a larger amplitude variation. The variation placed high requirement of power amplifier which degrade its performance indicated by Fig. 16. Figure 17 shows how multi-path interference (fading) has created ISI between the signals. As compared to QPSK, here we see that ISI effect is reduced in QAM. BER versus  $E_b/N_0$  is shown by Fig. 18. Figure 19, represent BER for different value of SNR for QAM-OFDM.

**Fig. 4** QPSK modulated data



**Fig. 5** QPSK demodulated data



**Fig. 6** QPSK OFDM signal

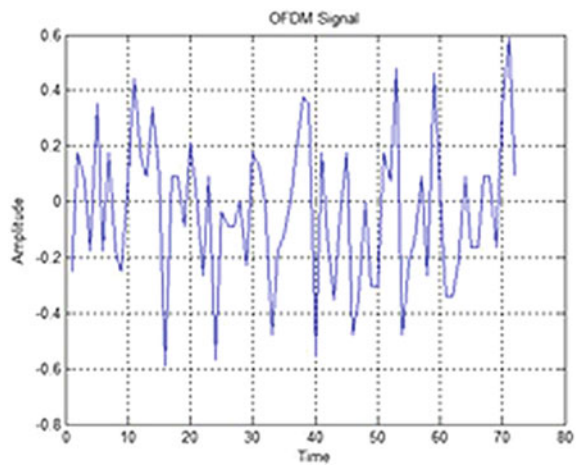


Fig. 7 Receive QPSK data

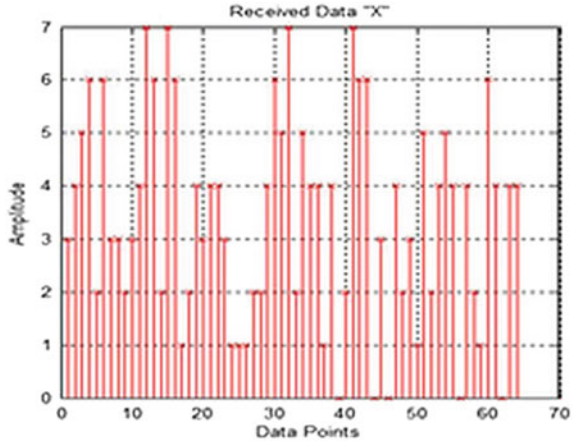


Fig. 8 Diagram of QPSK receive data in phase

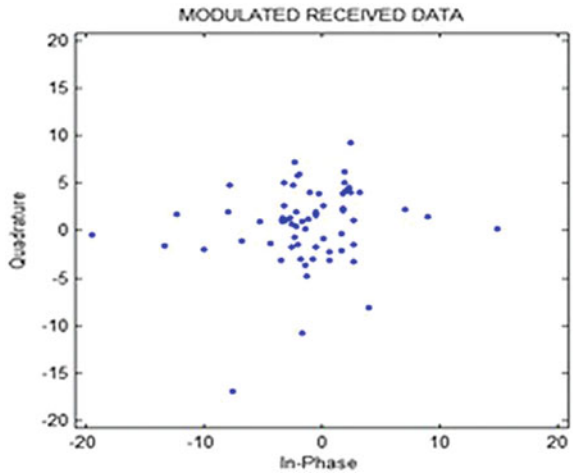
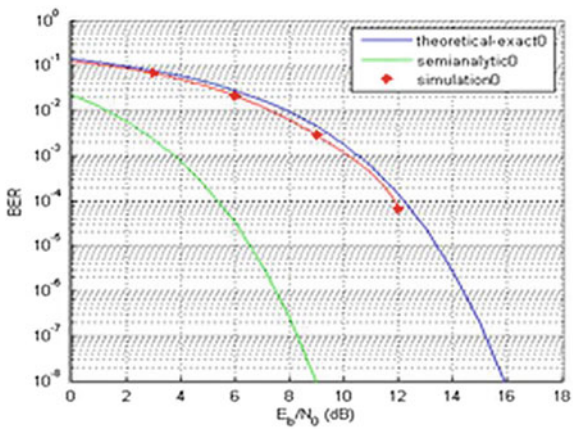


Fig. 9 BER versus Eb/No for QPSK (theoretical)



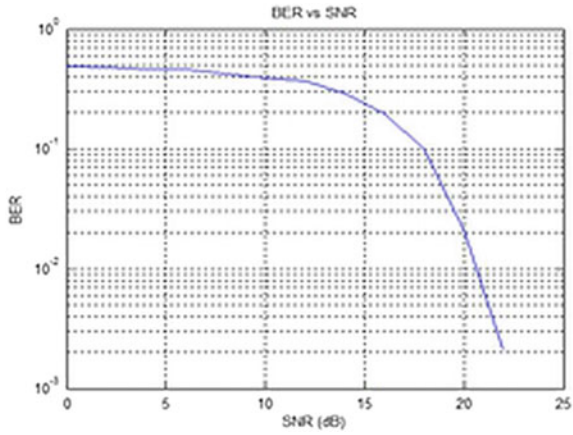


Fig. 10 BER versus SNR for QPSK modulation scheme

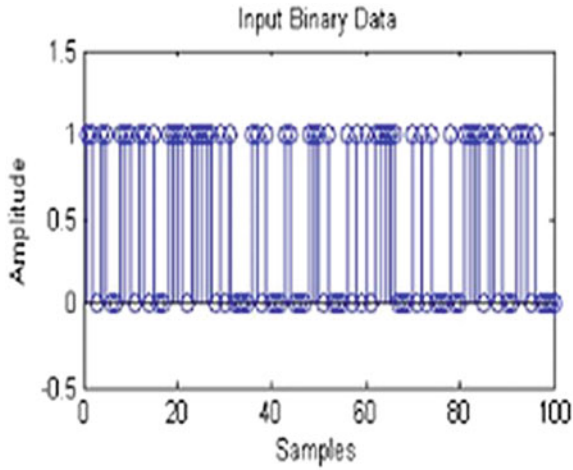


Fig. 11 Binary input data for QAM



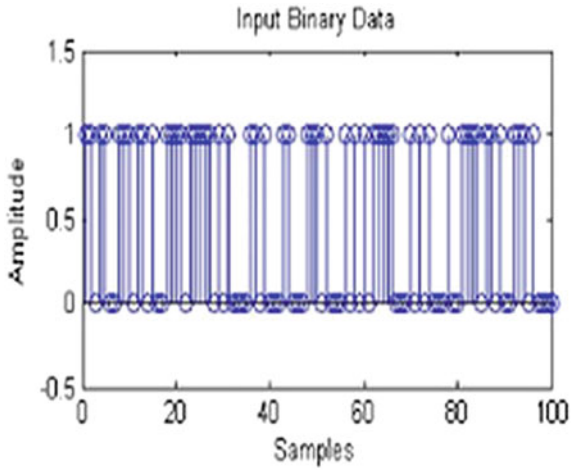


Fig. 12 Transmitted data for QAM

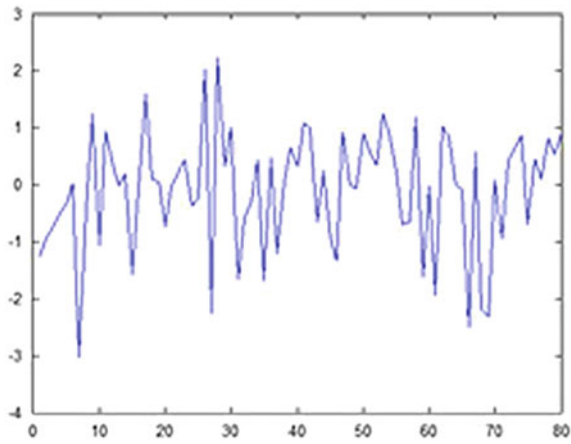


Fig. 13 AWGN noise in QAM

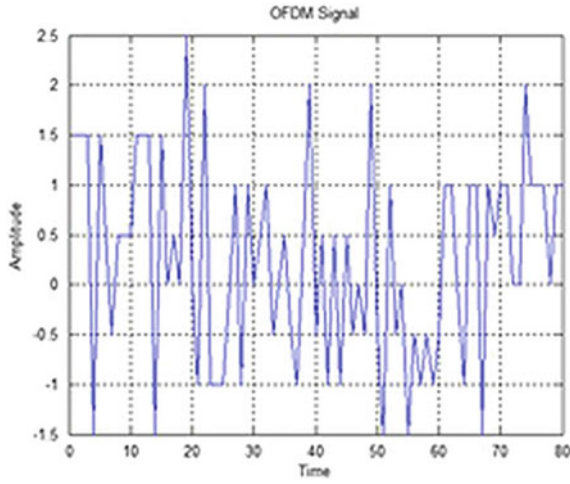


Fig. 14 QAM OFDM signal

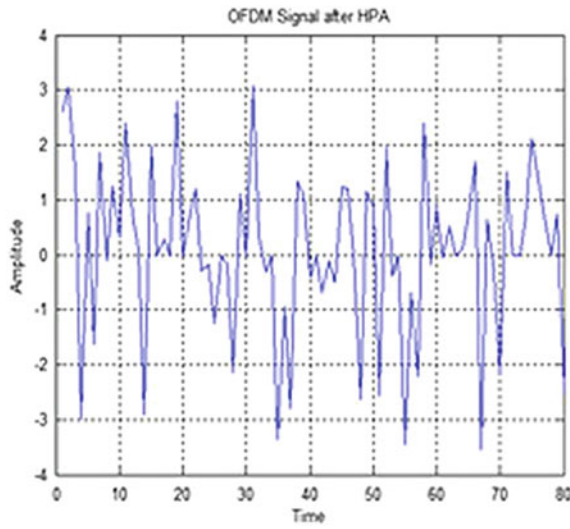


Fig. 15 QAM OFDM signal after HPA

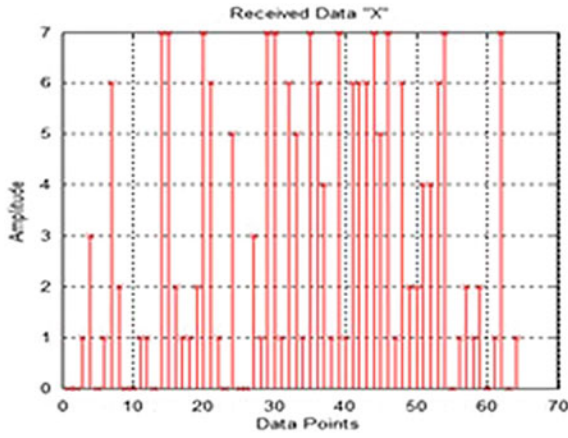


Fig. 16 Receive data in “X” channel

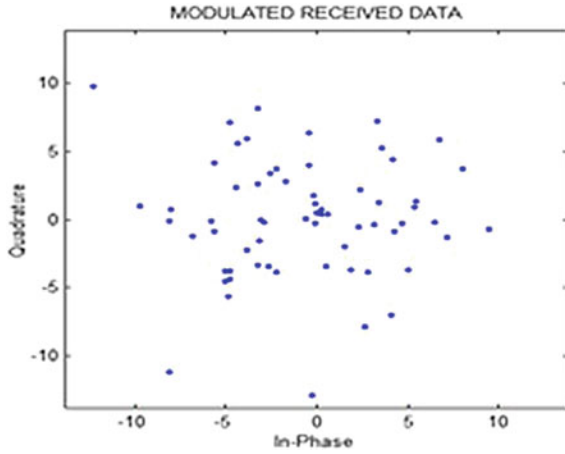


Fig. 17 Modulated receive data for QAM

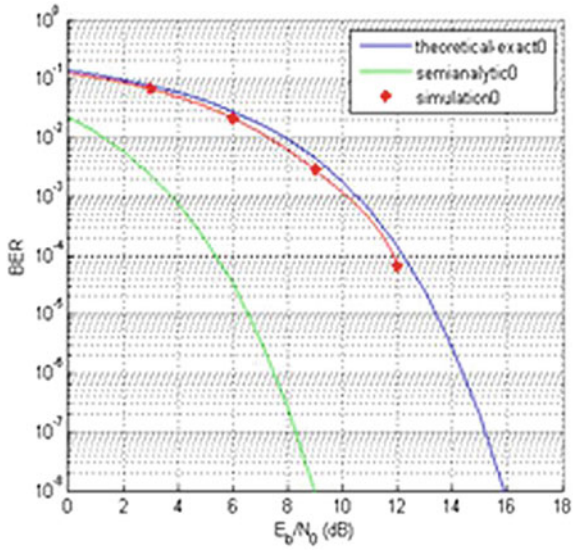


Fig. 18 Plot of BER versus  $E_b/N_0$  for QAM

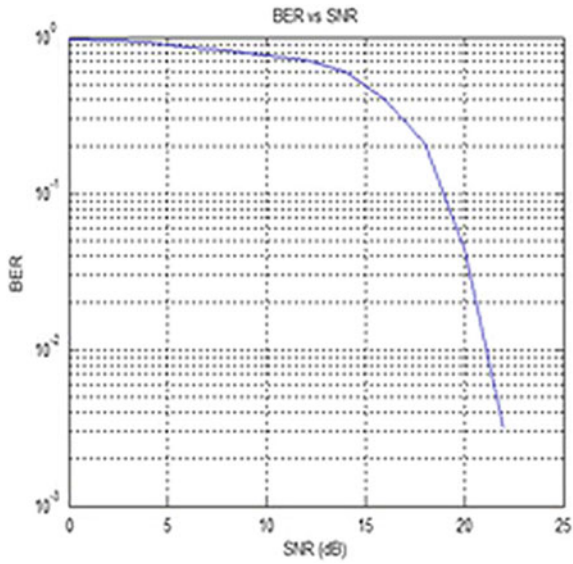


Fig. 19 BER versus SNR for QAM

## 4 Conclusion

OFDM has several interesting properties that suit its use over wireless channel and hence many standards have started to use OFDM for transmission and multiple accesses. In this work, OFDM framework for QAM and QPSK is analyzed and studied. Experimental result suggests that, performance of OFDM-QAM is better than QPSK-OFDM. It is also concluded that, the effect of ISI will be more and more for high order modulation schemes and the insertion Cyclic prefix results in loss of bandwidth as same information is repeated. Hence, this scheme may not be suitable for next generation communication system.

**Acknowledgements** We would like to thanks Electronics & Communication Department of JECRC University for providing the support and help. We would also like to thanks to Prof. (Dr.) Ram Rattan, Dean Engineering, Dr. Dinesh Sethi, HOD ECE, Prof. (Dr.) Manisha Gupta, HOD Physics of JECRC University.

## References

1. Kumar, A., Gupta, M.: Key technologies and problem in deployment of 5G mobile communication systems. *Commun. Appl. Electron.* **3**, 4–7 (2015)
2. Kumar, A., Gupta, M.: Design of OFDM and PAPR reduction using clipping method. *Artificial Intelligence and Network Security 2015*, vol. 1. pp. 221–229. DRDO Delhi, Desi-doc (2015)
3. Kumar, A., Gupta, M.: A review on OFDM and PAPR reduction techniques. *Am. J. Eng. Appl. Sci.* **8**, 202–209 (2015)
4. Kumar, A., Gupta, M.: Design of 4:8 MIMO OFDM with MSE equalizer for different modulation technique. *J. Wirel. Pers. Commun.* **95**, 4535–4560 (2017)
5. Ghorpade, S., Sankpal, S.: Behaviour of OFDM system using Matlab simulation. *Int. J. Adv. Comput. Res.* **3**, 67–71 (2013)
6. Murtuza, W., Srivastava, N.: Implementation of OFDM based transceiver for IEEE802.11a on FPGA. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **4**, 24–28 (2014)
7. Raju, G., Uma, B.: Design and simulation of wavelet OFDM with wavelet denoising on AWGN channel. *Int. J. Adv. Res. Comput. Commun. Eng.* **2**, 3015–3018 (2013)
8. VaniDivyatha, M., SivaReddy, B.: Design and BER performance of MIMO-OFDM for wireless broad band communication. *Int. J. Modern Eng. Res.* **3**, 1382–1385 (2013)
9. Nieto, J.: An investigation of OFDM-CDMA using different modulation schemes on HF multipath fading channel. In: *Proceedings of the SPIE Digital Wireless Communication*, vol. 55, pp 300–308 (2014)
10. Jober, M.A., Massicotke, D., Achouri, Y.: A higher radix FFT FPGA implementation suitable for OFDM system. In: *IEEE International Conference on Electronic Circuit & System*, vol. 1, pp. 744–747 (2011)
11. Yan, H., Wan, L., Zhou, S., Shi, Z., Huang, J.: DSP based receiver implementation for OFDM acoustic modem. *Phys. Commun.* **5**, 22–32 (2012)
12. Tandu, D., Moonen M.: Joint adaptive compensation of transmitter and receiver under carrier frequency offset in OFDM based system. *IEEE Trans. Signal Process.* **55**, 5246–5252 (2007)

# Energy Harvesting-Based Two-Hop Clustering for Wireless Mesh Network



Sudeep Tanwar, Shivangi Verma and Sudhanshu Tyagi

**Abstract** The deployment of wireless mesh networks (WMNs) in numerous applications ranging from civilian to military and much more has acquired attention among researchers in recent years. A mesh is formed from tiny sensor nodes (SNs), deployed in some predefined manner which can perform task as per the need of an application. Effective utilization of energy of each SN in any network is one of the challenging task, as replacement of battery of these nodes degrades the performance of the network. Considering this optimal energy utilization, in this paper, we have proposed a two-hop optimized clustering-based approach in WMN using energy harvesting (EH), which increases the network lifetime and stability. Sensors those are nearest to the collection point (CP) can directly communicate with the base station (BS); however, SNs those are far away from CP used the cluster-based approach for data communication. Result shows that there has been significant improvement on lifetime enhancement, and stability as compared with other state-of-the-art protocols which exist under the same category.

**Keywords** Wireless mesh network (WMN) · Collection point (CP)  
Energy harvesting (EH) · Clustering

---

S. Tanwar (✉)

Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, India

e-mail: sudeep149@rediffmail.com

S. Verma

Department of IT, Chandra Vati Tiwari Girls Degree College, Kashipur, India

e-mail: shivangi0007@gmail.com

S. Tyagi

Department of ECE, Thapar Institute of Engineering and Technology, Deemed to be University, Patiala, Punjab, India

e-mail: sudhanshutyagi123@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713, [https://doi.org/10.1007/978-981-13-1708-8\\_42](https://doi.org/10.1007/978-981-13-1708-8_42)

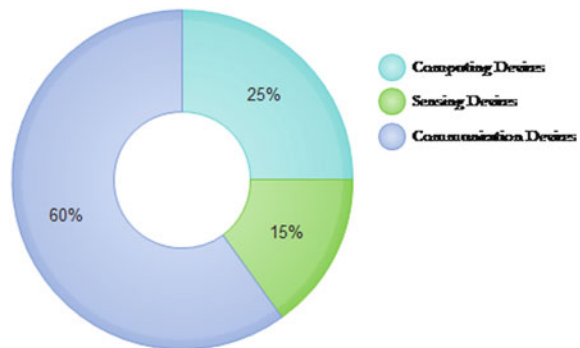
## 1 Introduction

Over a couple of years, wireless mesh network (WMN) is gaining popularity and has a potential to deliver on Internet broadband, wireless local area network, and connective network for operators or customers at low costs. Since, the wireless technologies are growing at rapid rate, selection of appropriate route between sender and receiver is always crucial. Present mesh routers have the bridge functionality to connect WMNs with preexist wireless networks, such as mobile and wireless sensor networks [1]. Mesh topology provides many alternate paths to transmit the sensed/observed data between sources and destination. Router must be intelligent, especially during the data transfer operation, in the aspect that during a node failure path could be reconfigured through the alive SNs.

SNs are very tiny and have the combinations of microcontroller, memory space for storage purpose, transceiver for data communication, and importantly small battery. Sensing, processing or computation, and communication are the three major components, where energy of a SN will be consumed; out of the these three components, communication will consume the major portion as shown in Fig. 1. Yole development survey predicts a growing market for EH modules. According to this survey, combination of automation in infrastructure and industrial sector will capture the adequate percentage of the digital market by the end of 2017. It further forecasts that EH in some sectors like mechanical and thermal will have 39% and 15% of the market share, respectively. Automation in infrastructure and industrial sector will have the largest segment of EH market possibly in the year 2017 as shown in Fig. 2. In this paper, we have proposed a two-hop communication in WMN using EH (THCEH) technique which includes the two-layer-based architecture for sensing and transmission. Sensed data are delivered to the CP passing through cluster head (CH) and nearest EH node. Inclusion of EH node in-between the route has the purpose to save the energy.

Rest of the paper is organized as follows: Previous work done by researcher in this domain is briefly discussed in Sect. 2. Section 3 highlights our proposed work. Section 4 discuss the energy model of the proposed approach. Section 5 discusses

**Fig. 1** Distribution of energy consumption



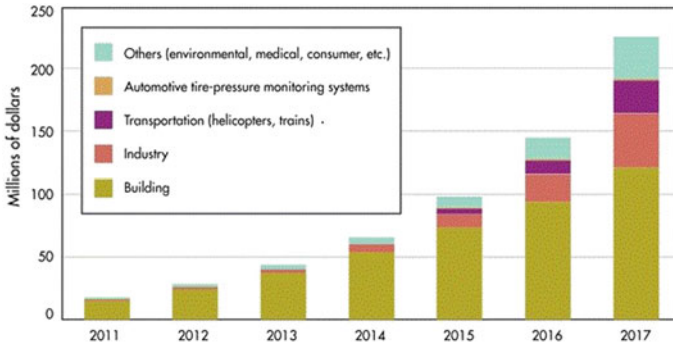


Fig. 2 Growth of energy harvesting in different applications

the results obtained after extensive simulations, and Sect. 6 concludes our work with future work.

## 2 Related Work

In this section, various state-of-the-art protocols which exist in this category are briefly described. Lacuesta et al. [2] proposed two protocols named as weak and strong protocol for secure and energy-saving spontaneous ad hoc mechanism under WMCN. Trust between users and mesh routers are the major concern in these protocols. These protocols were performed the operations such as: random checking, verification division, incorrect packets elimination, and node authentication. It does not considers the energy consumption for all nodes when spontaneous network nodes used the Internet access. Zhao et al. [3] proposed location management solution for Internet-based WMN. It provides support to the mobile users and also minimizes the overhead in terms of location update (LU) over mesh backbone. It provides better scalability and power consumption. It saves power and adaptively supports any type of network. In video delivery aspect, better quality of video could not be maintained for a longtime period. This problem was taken by Chen et al. [4] by proposing an energy-efficient cross-layer solution for video delivery in WMN. E-mesh technology provides a high-multimedia QoS and saves energy. Liliana et al. [5] proposed the clustering-based approach in WSN, where SNs send data to the BS through CH.

Sehgal et al. [6] suggested residual energy-efficient heterogeneous (REEH) to extend the network lifetime. This was based on heterogeneous environment, and section of CHs were depends on the residual energy of the SNs. CH gets data from neighboring nodes and sends to the CP after compression. Mamidisetty et al. [7] considered a mesh topology which arises by an embedding node in 2D-base grid. Embedded function is given in [7], where each node is placed on the grid within the fixed transmission range. Function equation gives the mesh topology, where



each node has  $q$  neighbors, which gives the count of neighbors of each node and  $q_k$  denotes the count of nodes in a cluster.

Mamidisetty et al. [8] considered the regular mesh topology for  $q=3, 4, 6, 8$  and also provide mechanism to select the CH depending upon  $k$ -dominating sets. Zhang et al. [9] proposed the clustering algorithm for improvement in lifetime of WSN with EH sensors. This increased the network lifetime by introducing EH nodes, CH received, and transmitting data directly to BS or via servers as relay nodes with EH. In this paper, we have extended the idea presented by [8, 9] and placed the EH nodes between the CH and CP. Three parameters for the validity of cluster were used which depend on the following points:

- Node communicates with its immediate neighbor
- Required energy to communicate with CPs by CH
- Calculate how much energy is required to send data to CP by CH.

$K$ -dominating CH approach is possible when

$$k < \frac{3E}{2e} - \frac{1}{2}. \quad (1)$$

In final outcome, the need of systematic approach for the selection of CHs in networked sensing systems. Major problems in the existing literature are degradation of network lifetime and energy efficiency. Main objectives of proposed approach are to increase the network lifetime and energy efficiency in WMN through the THCEH protocol to address the aforementioned issues.

### 3 Proposed Approach

To increase the lifetime and also optimize the energy efficiency of the network, we proposed THCEH mesh architecture as shown in Fig. 3, where we have considered the mesh topology for  $q=4$  and  $k=1$  or  $2$  hop. In proposed approach, a cluster depends on the following parameters:

- We assume that for inner layer required energy for a node to communicate with its neighbor is  $e_1$  and for outer layer is  $e_2$ , where CH is same for both clustering levels.
- CH collects the data and sends to EH, and at EH energy is obtained from various external sources (like solar, thermal, wind, kinetic) and stored for further use in devices used in WSN.
- EH compressed the data and sends to the CP.

The next section highlights the energy model used in our proposed approach.

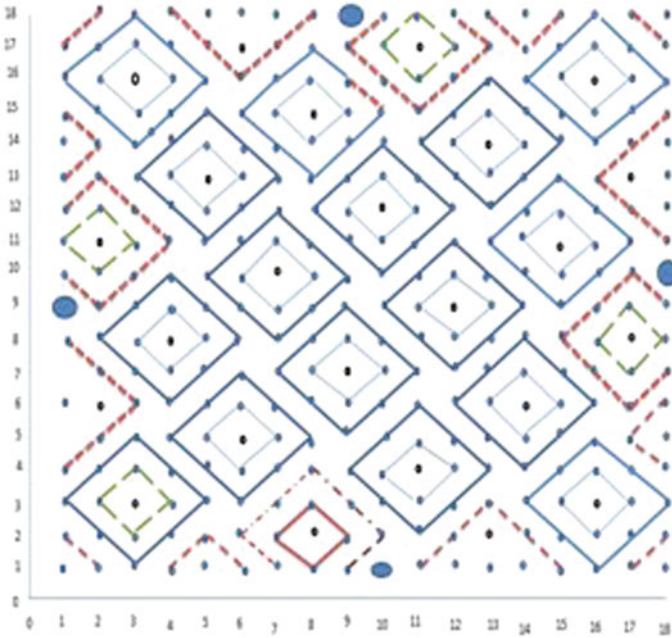


Fig. 3 Square and semi-square in E4 with k = 1 or k = 2 hop

### 4 Energy Model

Let  $e_1$  and  $e_2$  be the enhanced energy of neighboring nodes. A node  $n_i \in N$  sends the data to CH over a path of length  $k$ ,  $k(e_1$  and  $e_2)$ , the unit of energy expended for each message, i.e., sends to  $n_i$  to  $c_i$ . Let  $E$  represent the energy expended by CHs to EH and  $E$  represent the CH sends the data message to CP. The worst case of expended energy is:

$$E + k(e_1 + e_2)p(t_k^q - 1), \tag{2}$$

where  $p$  is the number of messages produced at each node in which CH receives the data from each node that belongs to the same cluster and sends it to EH. Under no CH condition, there are  $t_k^q - 1$  nodes in the cluster. Expended the above model by  $pEt_k^q - 1$  amount of energy to send the data to EH, CH receives the message from cluster and sends to the EH. EH compressed the data and sends to the CP. Compressing data are used for reducing the network traffic. Let  $C$  be compressed ratio used to reduce the traffic when CH sends the data to EH, i.e.,

$$(1 - C)pEt_k^q - 1 + k(e_1 + e_2)p(t_k^q - 1). \tag{3}$$

### 4.1 Least Set of Cluster Head

$E_4$  mesh topology,  $n_{ij}$  is adjacent to the nodes in first-level clustering is

$$n_{i+1,j} \ n_{i-1,j} \ n_{ij+1} \ n_{ij-1} \tag{4}$$

and intersecting adjacent to the nodes in second-level clustering is  $n_{i+2j}, n_{i-2j}, n_{ij+2}, n_{ij-2}, n_{i-1,j+1}, n_{i+1,j-1}, n_{i-1,j-1}, n_{i+1,j+1}$ . Figure 3 shows an example of clusters in  $E_4$ , where  $R=C=18$  and  $k=1, k=2$ . Each square in the center as one CH, i.e., is same for both cluster levels  $k=1$  or  $2$ . Some of the squares at the boundary line are shown with semi-square. Dark nodes along the boundaries are called EH nodes.

### 4.2 Best Possible of Cluster Head

We are now presenting the results to establish the best solution for CH selection. In the mesh topology,  $E_q$  is:

$$t_k q = qk(k + 1)/2 + 1. \tag{5}$$

Consider a fixed  $q$  and variate the value of  $k$ , node  $n$  and  $q$  neighbors that are involved in the process. Thus, as a basic for induction, when  $k=1, t_1 q=(q+1)$ . Recall that  $e_1$  and  $e_2$  are the energy expended from the  $k=1$  and  $k=2$  hop. CH received the message from the  $e_1$  and  $e_2$  nodes,  $E$  (cluster head) and  $E$  (energy harvesting) is the energy expended when a node sends the message to CP. Energy required for  $A$  data aggregation rounds, using CH in  $k=1$  or  $k=2$  hop is:

$$AE + (e_1)n_k q \tag{6}$$

$$E' + A \left[ E + \sum_{i=1}^k q_i (ie_1) + \sum_{i=1}^k q_i (ie_2) \right] \tag{7}$$

$$E' + A \left[ E + \sum_{i=1}^k i^2 (qe_1) + \sum_{i=1}^k i^2 (ie_2) \right]$$

$$E' + A \left[ E + q(e_1 + e_2) \frac{k(k + 1)(2k + 1)}{6} \right] \tag{8}$$

Let  $n_k^q = \frac{k(k+1)(2k+1)}{6}$ , and then the above equation is expended as:

$$A[E + (e_1 + e_2)n_k^q] + E^0 \tag{9}$$

Some nodes directly communicates with EH nodes; it means there is no CH involved. Hence, there are  $t_k^q = q \left( \left( \frac{k(k+1)}{2} \right) \right)$  nodes. Therefore, energy required for A aggregation rounds for the nodes is:

$$AE(t_k^q) \quad (10)$$

$c$  is the compressed ratio, and then the required energy to send data from CH to EH is

$$A((1-c)Et_k^q + (e_1 + e_2)n_k^q) \quad (11)$$

To save the mesh topology  $E_4$ , Eq. 11 must be less than Eq. 12; since  $q=4$ , we need

$$E(t_k^4) - ((1-C)Et_k^4 + (e_1 + e_2)n_k^4) > 0 \quad (12)$$

$$E(t_k^4) - ((1-C)Et_k^4 + (e_1 + 2e_1)n_k^4) > 0 \quad (13)$$

Now, on substituting the values of  $t_k^4$  and  $n_k^4$  from above equations, we get

$$E \left[ q \frac{k(k+1)}{2} + 1 \right] \quad (14)$$

$$\left[ (1-C)E_q \frac{k(k+1)}{2} + 3qe_1 \frac{k(k+1)(2k+1)}{6} \right] > 0 \quad (15)$$

### 4.3 Algorithm of THCEH

Algorithm-1 of the proposed approach starts with initializing the network size,  $n_{ij}$ , which is at the intersection of  $i$ th and  $j$ th column and is adjacent to the nodes in first-level clustering that is  $n_{i+1j}$ ,  $n_{i-1j}$ ,  $n_{ij+1}$ ,  $n_{ij-1}$  and the intersecting adjacent to the nodes in second-level clustering is  $n_{ij+2j}$ ,  $n_{i-2}$ ,  $n_{ij+2}$ ,  $n_{ij-2}$ ,  $n_{ij+1,j+1}$ ,  $n_{ij+1,j-1}$ ,  $n_{i-1,j-1}$ ,  $n_{ij+1,j+1}$ . Each square in the center as one CH, i.e., is same for both cluster levels  $k=1$  or  $2$ . CHs receive the data from the  $k=1$  and  $k=2$  hop, and all CHs transmit the data for their neighboring EHs. This received data is transmitted to CP.

---

**Algorithm 1** Pseudo code for TECH
 

---

- 1: **Input:** initializing the network elements
  - 2: **Output:** Communication of CH to EH
  - 3: Initialization network size  $A = (x_1, y_1), (x_2, y_2), \dots, (x_N, y_N), q = 4$
  - 4: Intersection of  $i^{th}$  Row and  $j^{th}$  Column is adjacent to the nodes in 1<sup>st</sup> level clustering.  
 $n_{i+1j}, n_{i-1j}, n_{ij+1}, n_{ij-1}$
  - 5: similarly for the nodes in 2<sup>nd</sup> level clustering.  $n_{i+2j}, n_{i-2j}, n_{ij+2}, n_{ij-2}, n$  6: Calculate CH:  

$$X = i(2k + 1) \bmod t^{q_k} \text{ And } Y = i$$
 Constraint  

$$\forall i(2k + 1) \bmod t^{q_k} \leq R, i \leq C$$
  - 7: Calculate the energy at 1<sup>st</sup> level and 2<sup>nd</sup> level required energy in order to communicate with neighbor of node,  $e_1$  and  $e_2$ .  

$$A[E + (e_1)n^{q_k}]$$
 And  

$$A[E + (e_1)n^{q_k}] + E^0$$
  - 8: CH takes the energy from 1<sup>st</sup> level and 2<sup>nd</sup> level  

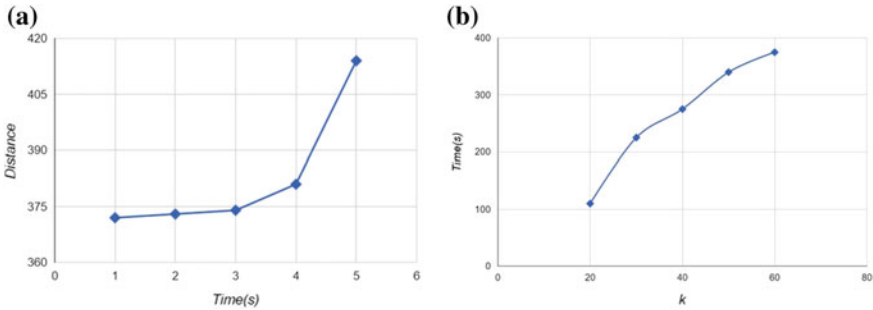
$$E = CHCP \text{ and } E = CH \rightarrow EH$$
  - 9: Which node have no CH, they can direct communicate to the EH  $AEt^{q_k} + 1$
- 

## 5 Discussion on Result

To validate the proposed approach, we performed simulations on MATLAB. The underlying 2D-base grid had  $R=18$  rows and  $C=18$  columns. Using embedding functions ( $E_q, q=4$ ), nodes are embedded on this 2D-base grid. Every node always transmitted 5 messages out of hold buffer of 50 messages. CHs received the message from  $k=1$  or  $k=2$  hop nodes and transferred it to the EH nodes. EH nodes take the data and transmit to CP after aggregation. For the mesh topology  $E_q$ , we have selected a  $q+1$  frame size. The total number of nodes  $N=324$  is distributed to the squared shape network field. All mesh nodes and the location of CP are fixed. For data packet transmission, size of message is set to 5 bytes and parameters values are same as LEACH [10]. System reliability, lifetime enhancement, and energy efficiency are the parameters for the performance evaluation of THCEH. Table 1 shows the simulation parameters. We have established the THCEH to data aggregation, given a fixed  $k=1$  and  $k=2$ , a fixed compression ratio  $C$ , required inner layer energy for communication of node with neighbors is  $e_1$  and required outer layer energy for communication of node with neighbors is  $e_2$ , where CH is same for both clustering levels. CH received the data and sends to EH, E (energy harvesting) and EH compressed the data sends to the CP (collection point), E. For each transmission, the energy expended is calculated as in [10] which includes two major segments: (a) the electronic circuits and (b) the

**Table 1** Simulation parameters

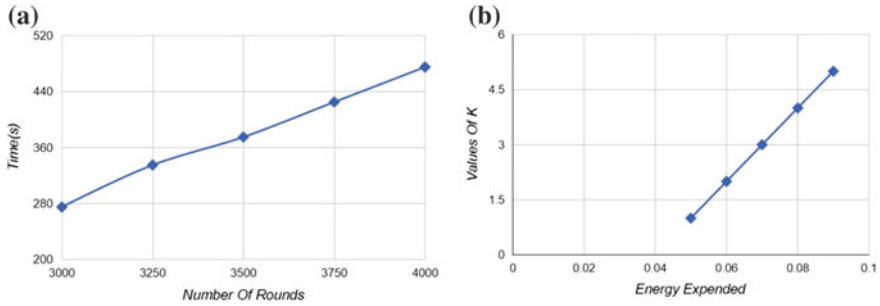
Network size	18 * 18, 324 fixed nodes
Transmitter electronic	50 nJ/bit
$R_x$ electronic	50 nJ/bit
$T_x$ amplifier	100 pJ/bit
Size of message for data packet	5 bytes
$K$	40 $\sqrt{\text{pJ}}$
$d1, d2, d3$	1, $2d_1, 3$
Simulation duration	210 s



**Fig. 4** **a** Latency observed in  $E_4$  on 18 \* 18 versus distance, **b**  $K$  versus time for E4

energy amplifier. Using the radio model [10], we have considered the required energy by the electronic circuits to obtain and develop a bit 50 nJ. The value of compression ratio is  $C=0.9$ . The required energy by an amplifier to transmit one bit over a distance of 1 m is 100  $\sqrt{\text{pJ}}$ . The separation among the two adjacent grids was 1 m and diagonally was 3.

Nowadays, regular topologies are used in many systems such as streetlight monitoring systems, traffic monitoring, energy support system and electronic water meter monitoring. In Fig. 4a, distance between nodes will change in one hop and then time will increase. If we increase the value of  $k$  changes, then the time will increase as shown in Fig. 4b; if we increase the number of rounds, then the execution time will increase. Figure 5a shows the relation between energy and time, and Fig. 5b indicates that the throughput at CP unchanged irrespective of time. The number of messages sent by an EH to CP depends on the number of messages received from the clusters. When  $k$  increases, energy will increase.



**Fig. 5** **a** Estimated time when all nodes die in  $E_4$ , **b** energy expended in  $E_4$  when  $k < 5$

## 6 Conclusions with Future Scope

In this paper, we propose a two-hop communication and energy harvesting-based routing protocol for WMN, where EH nodes performed the data aggregation based on dominating sets of graphs. Focusing on embedding nodes in 2D-base grid, proposed approach selects a CH in these topologies. Without EH, CHs directly send the message to CP and consume more energy. In proposed approach, for inner layer the required energy to communicate by a node with its neighbor is  $e_1$  and for outer layer is  $e_2$ , where CH is same for both clustering levels. CHs receive the data and send to EH node; finally, EH compressed the data and sends to the CP. Results show that there has been a significant improvement on lifetime enhancement, energy and latency. In future, we implement this approach with different mesh topologies for  $q = 6, 8$ . THCEH can be designed to detect the collision issues and can also be extended with heterogeneous nodes.

## References

1. Akyildiz, F., Wang, X., Wang, W.: Wireless mesh networks: a survey. *Comput. Netw.* **47**(4), 445487 (2005)
2. Lacuesta, R., Lloret, J., Garcia, M., Pealver, L.: Two secure and energy-saving spontaneous ad-hoc protocol for wireless mesh client. *J. Netw. Comput. Appl.* **34**(2), 492505 (2011)
3. Zhao, W., Xie, J.: Domain: a novel dynamic location management solution for internet-based infrastructure wireless mesh networks. *IEEE Trans. Parallel Distrib. Syst.* **24**(8) (2013)
4. Chen, S., Muntean, G.-M., E-Mesh: an energy-efficient cross layer solution for video delivery in wireless mesh networks. In: This work was supported in part by the Irish Research Council for Science, Engineering and Technology Enterprise. Partnership with Everseen Ltd. (2011)
5. Liliana, M., Abrboleda, C., Nasser, N.: Comparison of clustering algorithms and protocols for wireless sensor networks. In: Proceedings of IEEE CCECE and CCGEI (2006)
6. Sehgal, L., Choudhary, V.: REEH: residual energy efficient heterogeneous clustered hierarchy protocol for wireless sensor networks. *Int. J. Sci. Eng. Res.* **2**(12), 1–5 (2011)

7. Mamidisetty, K., Ghamande, M., Sastry, S., Ferrara, M.: A domination approach to clustering nodes for data aggregation. In: Proceeding of IEEE Global Telecommunications Conferences, pp. 1–5 (2008)
8. Mamidisetty, K., Ferrara, M., Sastry, S.: Systematic selection of cluster heads for data collection. *J. Netw. Comput. Appl.* 1548–1558 (2012)
9. Zhang, F.P., Xiao, G., Tan, P.H.: Clustering algorithm for maximizing the lifetime of wireless sensor network with energy-harvesting sensors. *Comput. Netw.* (2013)
10. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the IEEE Hawaii International Conference on System Sciences, p. 110 (2000)



# A Compact and High Selective Microstrip Dual-Band Bandpass Filter



Dwijjoy Sarkar and Tamasi Moyra

**Abstract** This paper presents a compact and novel dual-band bandpass filter (DBBPF) using a microstrip dual-mode resonator, folded stepped impedance resonator (SIR) and an etched ground structure (EGS). The first and second passbands are generated by the dual-mode and single-mode resonators. The proposed bandpass filter (BPF) produces passbands centered at 2.4 GHz and 3.5 GHz, respectively. The passband performance of the DBBPF is enhanced by etching a pattern in the ground plane which provides less than 1-dB insertion loss in both of the passbands with more than 20-dB isolation in the stopband. The BPF is analyzed using classical odd–even-mode technique. The simulated full-wave electromagnetic (EM) results obtained using IE3D are in well agreement with the theoretical results. The proposed DBBPF can be employed in WLAN and WiMAX applications.

**Keywords** BPF · DBBPF · SIR · Dual mode · EGS · WLAN · WiMAX

## 1 Introduction

Dual-band RF and microwave bandpass filters (DBBPFs) are subject of interest in recent days due to the growing demand for dual-band appliances. A DBBPF is an essential component in such systems. A conventional DBBPF can be designed by cascading two different bandpass filters (BPFs) having separate passbands, by embedding a different resonator into the primary resonator of the BPF and by the separation of various resonant modes generated by a multimode resonator. The first method is very straightforward and easy to implement, but such kind of DBBPF circuits is greater in size compared to others, and therefore, they are very rarely used

---

D. Sarkar (✉) · T. Moyra  
Department of ECE, National Institute of Technology Agartala, Agartala 799046,  
Tripura, India  
e-mail: dwijjoysarkar@gmail.com

T. Moyra  
e-mail: tamasi\_moyra@yahoo.co.in

nowadays. The second and third methods are prevalent for DBBPF designing. Some recent research articles related to different DBBPFs are shown in [1–9], which are implemented using [1] parallel-coupled stepped impedance resonators (SIRs), [2] microstrip line resonator loaded with an open-loop ring in one end which acts as an embedded resonator, [3] multimode E-shaped resonator, [4–6] dual-mode resonators, [7, 8] spiral resonators and [9] end-coupled resonators loaded with metallic via holes, respectively.

Although undoubtedly enormous research is already conducted on DBBPFs, further research is still required for the implementation of cost-effective and compact DBBPFs with good dual passband performance. In this paper, a novel and compact DBBPF is proposed using a dual-mode resonator, folded single-mode resonator and an etched ground structure (EGS). The dual-mode resonator generates the first passband near 2.4 GHz, and the folded microstrip line resonator produces the second passband near 3.5 GHz. The insertion loss in both of the passband is reduced by incorporating an EGS in the ground plane, which improves the cross-coupling between the resonators and feedline. All the resonators are analyzed using classical odd–even-mode technique. The proposed DBBPF is simulated in IE3D with RO4003C substrate ( $\epsilon_r = 3.38$ ,  $h = 0.508$  mm,  $\tan \theta = 0.0018$ ), which provides less than 1-dB insertion loss in both of the passbands and more than 20-dB rejection level up to twice the midband frequency.

## 2 Design and Analysis of the Proposed BPF

The proposed DBBPF is depicted in Fig. 1. All the labeled dimensions of the BPF are listed in Table 1. The BPF only consumes  $18.75 \times 9.55$  mm<sup>2</sup> core layout area. The coupling architecture of the DBBPF is illustrated in Fig. 2. The source and load

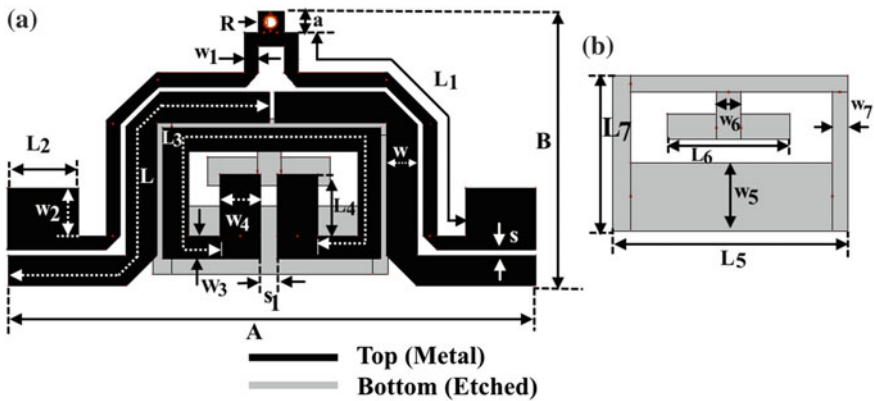
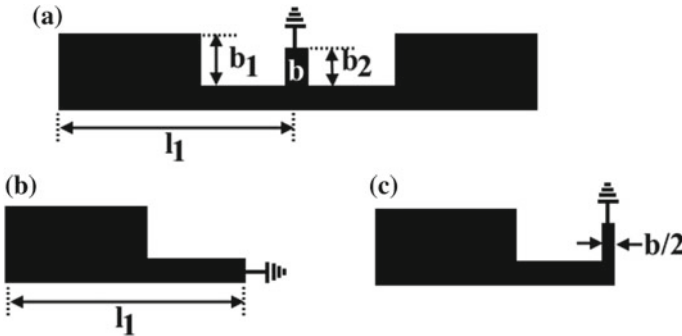
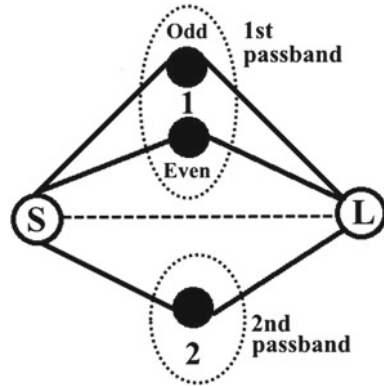


Fig. 1 a Proposed DBBPF; b EGS (bottom)

**Table 1** DBBPF dimensions

Dimensions	Values (mm)
$L, L_1, L_2, L_3, L_4, L_5, L_6, L_7$	14.215, 12.11, 2.6, 14.35, 2, 8.32, 4.35, 5.5
$w, w_1, w_2, w_3, w_4, w_5, w_6, w_7$	1.1, 0.5, 1.7, 0.85, 1.45, 2.4, 0.9, 0.55
$a, R, s, s_1$	0.8, 0.25, 0.2, 0.55
$A, B$	18.75, 9.55

**Fig. 2** Coupling architecture of the proposed BPF



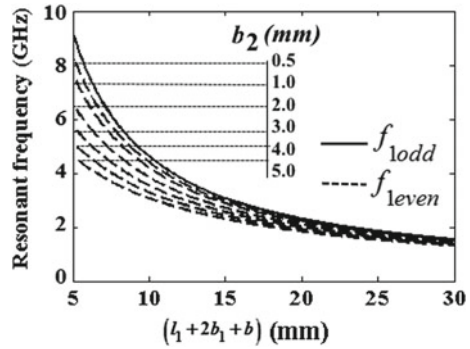
**Fig. 3** a Dual-mode resonator; b odd-mode circuit; c even-mode circuit

are directly coupled to each other, whereas they are in cross-coupling mode with the dual-mode resonator (resonator 1) and folded SIR (resonator 2).

The proposed DBBPF is constructed using a fundamental dual-mode resonator of Fig. 3 and a microstrip line SIR of Fig. 4. Using classical transmission line and odd-even-mode techniques [10], the resonant frequencies of the dual-mode resonator in Fig. 3 are approximately calculated from its odd-even-mode equivalent circuits, which are shown in Eqs. (1) and (2), respectively.

$$f_{1odd} = \frac{c}{4(l_1 + 2b_1 + b)\sqrt{\epsilon_e}} \tag{1}$$

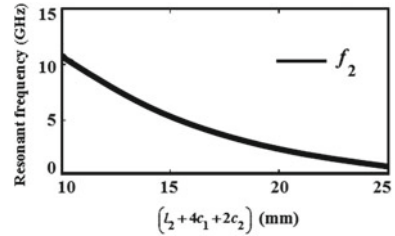
**Fig. 4** Odd- and even-mode resonant characteristics



**Fig. 5** Stepped impedance resonator (SIR)



**Fig. 6** Resonant characteristic of the SIR



$$f_{1even} = \frac{c}{4(l_1 + 2b_1 + b + b_2)\sqrt{\epsilon_e}} \tag{2}$$

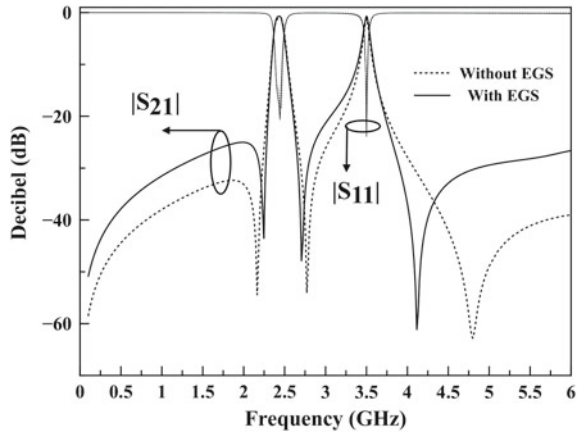
[where ‘c’ is the velocity of light, ‘ε<sub>e</sub>’ is the effective dielectric constant, and (l<sub>1</sub> + 2b<sub>1</sub> + b) and (l<sub>1</sub> + 2b<sub>1</sub> + b + b<sub>2</sub>) are average lengths of the odd- and even-mode resonators.]

From Eqs. (1) and (2), the odd- and even-mode resonant characteristics are extracted by considering RO4003C microstrip transmission line, which are depicted in Fig. 4. It can be found that the resonant frequencies  $f_{1odd}$  and  $f_{1even}$  are closely spaced, and they can be separated by increasing the length of the dimension ‘b<sub>2</sub>.’

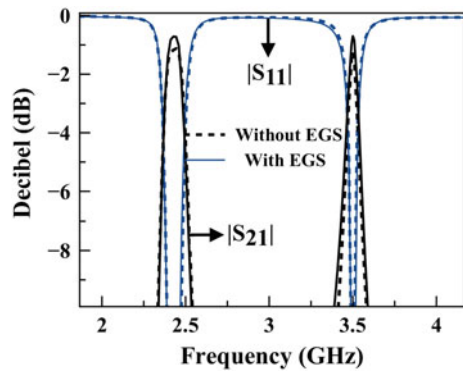
The second resonator of Fig. 1 is the folded version of a basic SIR shown in Fig. 5. The resonance frequency of the SIR is approximately calculated by using the transmission line techniques mentioned in [10], which is shown in Eq. (3). The tunability characteristic of the resonance frequency is depicted in Fig. 6, which indicates a monotonical decay of the resonance frequency as the average resonator length is increased.

$$f_2 = \frac{c}{2(l_2 + 4c_1 + 2c_2)\sqrt{\epsilon_e}} \tag{3}$$

**Fig. 7** Simulated scattering parameters



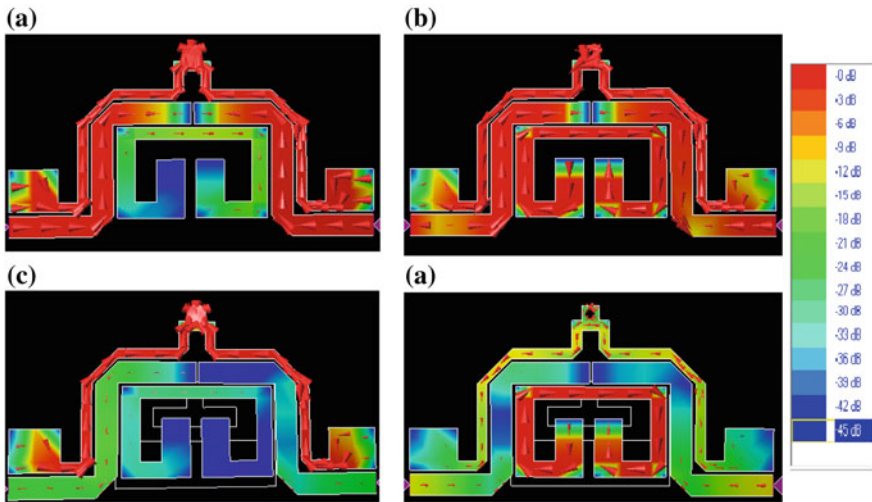
**Fig. 8** Simulated scattering parameters zoomed view



### 3 Results and Discussion

The simulated scattering response of the proposed DBBPF (Fig. 1) in the presence and absence of EGS is depicted in Fig. 7. The initial dimensions of the DBBPF are extracted using Eqs. (1)–(3), and thereafter, some manual optimization is performed to get better results. At first, the resonant modes  $f_{1odd}$  and  $f_{1even}$  are set near to 2.4 GHz by adjusting the dimensions  $L_1, L_2, w_2, a$  and the radius  $R$  of the dual-mode resonator of Fig. 1. The center frequency of the second passband is set near to 3.5 GHz by adjusting the dimensions  $L_3, L_4$  and  $w_4$ , respectively.

Figure 8 depicts the zoomed view of the DBBPF scattering response in the presence and absence of EGS. Although it can be found that the transmission zero locations are somewhat changed due to the inclusion of EGS in the ground plane, its effect in the passband can be more accurately observed in Fig. 8. The insertion loss in both of the passband is reduced near to 0.7 dB, which was around 1.2 dB for the previous case. The fractional bandwidth (FBW) in the passbands is 4.2% and 2%, respectively, which ensures achievement of high selectivity in both of the passbands.



**Fig. 9** Current distribution: **a** at 2.4 GHz without EGS; **b** at 3.5 GHz without EGS; **c** at 2.4 GHz with EGS; **d** at 3.5 GHz with EGS

This improvement in the passband performance can be explained by the current distribution patterns illustrated in Fig. 8. It can be found that the unused resonator in each of the frequency bands is properly deactivated due to the inclusion of EGS and therefore more power is transmitted in that specified passband. The folded SIR at 2.4 GHz is better deactivated in Fig. 9c compared to Fig. 9a. On the other hand at 3.5 GHz, the dual-mode resonator in Fig. 9d is better deactivated compared to Fig. 9b; thereby, more power transfer in each of the passband is obtained.

## 4 Conclusion

In this paper, a compact and novel DBBPF based on dual-mode resonator, folded SIR and EGS is proposed. The simulated DBBPF provides high selectivity in the both of the passbands with 20-dB outside-band isolation and less than 1-dB insertion loss in the operating frequencies (2.4, 3.5 GHz). The proposed BPF is suitable for WLAN (2.4 GHz) and WiMAX (3.5 GHz) applications.

**Acknowledgements** This work is supported by National Institute of Technology Agartala.

## References

1. Zhang, Y., Sun, M.: Dual-band microstrip bandpass filter using stepped-impedance resonators with new coupling schemes. *IEEE Trans. Microwave Theory Tech.* **54**, 3779–3785 (2006)
2. Hsu, C.-Y., Chen, C.-Y., Chuang, H.-R.: A miniaturized dual-band bandpass filter using embedded resonators. *IEEE Microwave Wirel. Compon. Lett.* **21**, 658–660 (2011)
3. Kuo, Y.-T., Chang, C.-Y.: Analytical design of two-mode dual-band filters using E-shaped resonators. *IEEE Trans. Microwave Theory Tech.* **60**, 250–260 (2012)
4. Sun, S.: A dual-band bandpass filter using a single dual-mode ring resonator. *IEEE Microwave Wirel. Compon. Lett.* **21**, 298–300 (2011)
5. Lerdwanittip, R., Namsang, A., Jantree, P.: Dual-band bandpass filter using stubs to controllable passband. *Procedia Comput. Sci.* **86**, 11–14 (2016)
6. Xu, J., Zhu, C.-M.: Compact dual-band bandpass filter using uniform-impedance resonators and dual-mode resonator. *Microwave Opt. Technol. Lett.* **58**, 1537–1540 (2016)
7. Xu, Z., Wei, B., Cao, B., Guo, X., Zhang, X., Heng, Y., Jiang, L., Zheng, T., Wang, J.: A compact dual-band bandpass superconducting filter using microstrip/CPW spiral resonators. *IEEE Microwave Wirel. Compon. Lett.* **23**, 584–586 (2013)
8. Xiao, M., Sun, G., Xu, F.: Compact dual-band bandpass filters based on a novel defected ground spiral resonator. *Microwave Opt. Technol. Lett.* **57**, 1636–1640 (2015)
9. Reja, A.H., Khader, A.A.-H., Ahmad, S.N., Salih, A.A.A.: Dual-band band-pass filters based on metallic via holes. *Procedia Comput. Sci.* **58**, 748–754 (2015)
10. Hong, J.-S.: *Microstrip filters for RF/microwave applications*. Wiley, Hoboken, NJ (2011)

# A Reliable Routing Protocol for EH-WSAN



Jagadeesh Kakarla

**Abstract** In one of our previous work, we have proposed a “delay- and energy-aware reliable coordination mechanism for sensor and actor networks.” In that protocol, the sensor acts as a backup cluster head in the absence of actor, which leads to higher energy consumption of the sensors and also causes funneling effect (the sensor which is nearer to the cluster head consumes more energy and loses a large number of packets as compared to the other sensors in the cluster) in the network. In this paper, energy harvesting sensors are introduced in the sensor and actor networks area (EH-WSAN). Further, we have used mid-point K-mean technique for the placement of actors. In this work, we have deployed two types of sensors (energy harvesting sensors and normal sensors) and actors. “The actor acts as a primary cluster head, and an energy harvesting sensor acts as a backup cluster head in the absence of actor.” The proposed energy-aware coordination mechanism is simulated, and results are analyzed with its competitive protocols.

**Keywords** Sensor · Energy · Cluster · Actor · Harvesting

## 1 Introduction

“In wireless sensor and actor network (WSAN), sensors detect the changes in the environment and actors perform actions based on the sensors data.” A sensor normally consists of five different components such as sensing unit, analog-to-digital converter (ADC), processor and storage, transceiver, and power unit as shown in Fig. 1. A sensor generates an analog signal by sensing the physical area, which is reformed into a digital signal using ADC. The digital signal is transmitted to a processor, which in turn consists of microcontroller that performs computing operations. The transceiver is used to send information to the destination sensor. The

---

J. Kakarla (✉)

Department of CSE, Indian Institute of Information Technology Design & Manufacturing  
Kancheepuram, Chennai, Tamil Nadu, India  
e-mail: jagadeeshk@iiitdm.ac.in



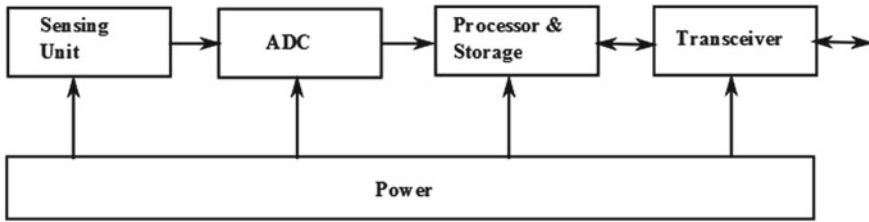


Fig. 1 Sensor architecture

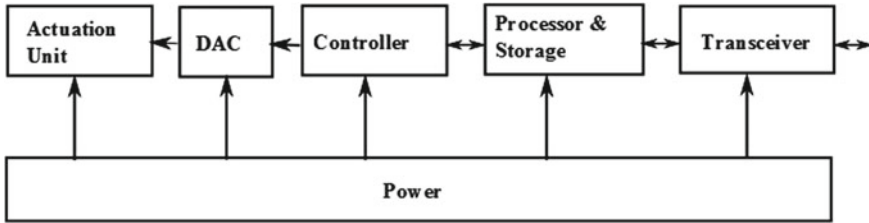


Fig. 2 Actor architecture

power unit supplies power to all the components in a sensor node. The actors are resource-rich (battery, transmission range, number of radios, etc.) as compared to sensors [1]. Figure 2 describes the architecture of an actor. WSN has important role in different practical applications like health monitoring, industrial safety, fire-hazard monitoring, and many other applications.

The network’s lifetime of WSN mainly depends on battery lifetime of a sensor. To improve the network lifetime of sensor networks, various researchers have proposed energy-efficient protocols. Another group of researchers have worked in the direction of energy harvesting from environment (solar and wind power), to ensure an unlimited energy supply to the sensor. “This category of sensor networks is called as energy harvesting wireless sensor networks (EH-WSN).” In our work, we have introduced energy harvesting sensor for wireless sensor and actor networks called energy harvesting sensor and actor networks (EH-WSAN). In these sensors, energy harvesting unit and buffer unit are added to the components of normal sensors.

## 2 Related Work

In this section, we have analyzed the existing routing protocols for energy harvesting sensor networks. Cao et al. proposed a hybrid routing metric which considers residual energy and energy harvesting rate (EHR) [2]. By using the above-mentioned method, authors have selected 1-hop node to forward its data to the sink. However, the results indicate the lifetime is less as compared to optimal approach. Bouachir et al. have

considered layer concept to design opportunistic routing and data dissemination protocol for real-time applications [3]. Li et al. have designed a distributed cluster-based routing protocol [4]. It considers sensor residual energy and harvesting capability for selecting cluster heads. It achieves 30% more throughput as compared to standard leach protocol. Yi et al. have proposed energy ability algorithm and greedy perimeter stateless routing mechanism for energy harvesting sensor networks [5]. The authors have considered sensor location, residual energy, and harvesting capability in the next hop selection to forward data for the sink.

Gong et al. have modified the existing AODV accordingly to meet the challenges of energy harvesting sensor networks. The results of the authors' proposed work outperform existing AODV protocol [6]. Cui et al. have identified that considering only harvesting capability of a sensor does not lead to better performance [7]. Hence, they have included bit error rate with harvesting capability in the routing process. Kawashima et al. have identified the drawbacks of existing relay traffic-based transmission power control [8]. "To overcome the drawback, the authors have proposed a routing protocol based on power generation pattern of sensor networks." Further, they have proved that their mechanism produces higher packet delivery ratio compared to the existing mechanisms.

Yin et al. have proposed a genetic algorithm-based routing protocol for EH-WSN [9]. "It consists of unequal clustering algorithm and adaptive routing protocol." Further, the sink initiates clustering process and also selects cluster heads. Jian et al. have proposed energy harvesting-aware routing protocol [10]. It improves the network's lifetime and throughput by 15–19% as compared to leach. Meng et al. have proposed an adaptive routing mechanism for EH-WSN [11]. The primary objective for their work is to achieve higher throughput, and they have achieved it with the help of regioning scheme. Further, the authors have implemented their work on a test bed consists of 20 energy harvesting sensors. It performed very well as compared to the existing mechanisms. Dong et al. have proposed a distance- and energy-aware routing protocol for EH-WSN (DEARER) [12]. They have analyzed their work using theoretical analysis and experiments. The analysis proved that their work performed well for EH-WSN.

### 3 Energy-Aware Coordination Mechanism (EACM)

"In one of our previous works, we have proposed a delay- and energy-aware coordination mechanism for WSAN" [13]. In the backup cluster head (BCH) scenario, the sensor which is one-hop way from actor consists of higher residual energy, and node degree is elected as BCH. It gathers information from the cluster members till the actor comes back to its original position. However, in this mechanism as sensors act as a BCH and the sensors which are one-hop away from primary cluster head, BCH has to forward the data from the cluster members to the head. Hence, a lot of energy will be consumed from these sensors and die early as compared to the remaining

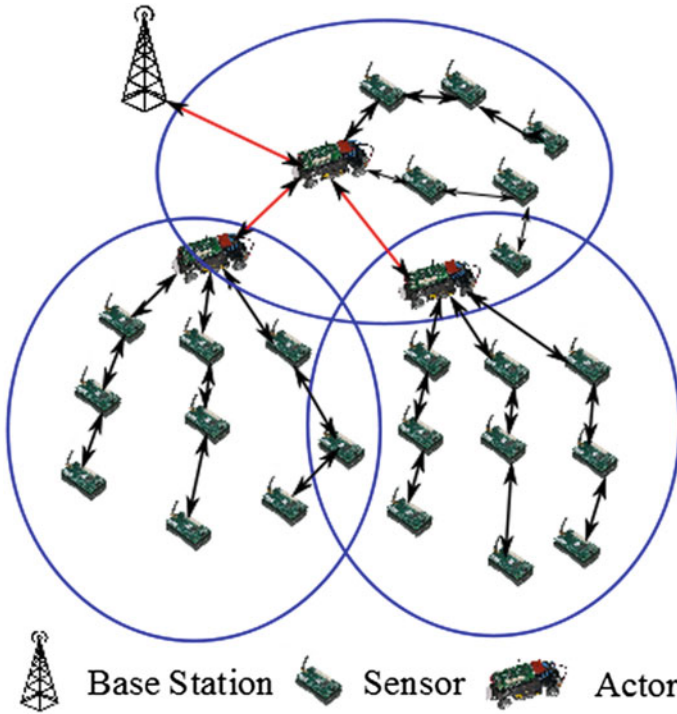


Fig. 3 Proposed network architecture

sensors in the network, which leads to network partition. This problem is stated as funneling effect in sensor networks.

“To reduce funneling effect, we have proposed an energy-efficient coordination mechanism for EH-WSAN.” In this work, we have deployed a set of actors, energy harvesting sensors, and large of normal sensors. The actor takes the duty as primary cluster head, energy harvesting sensor takes the role of BCH and the sensors behaves as cluster members. “In this work, sensors are organized into clusters in the first level, where the resource-rich actor acts as a cluster head (CH). In every cluster, sensors send their event information to the corresponding cluster head (actor).” The nodes’ organization of the proposed work is shown in Fig. 3.

In our proposed clustering mechanism, we have used the mid-point based K-mean clustering technique to compute the location of actors. The K defines the optimal number of actors, and it is computed using the below formula,

$$K = \left\lceil \sqrt{\frac{NS}{2\pi}} \sqrt{\frac{E_{fs}}{E_{mp}}} \frac{NA}{di^2_{toBS}} \right\rceil \tag{1}$$

where  $NS$  is sensors' quantity,  $NA$  is simulation area,  $E_{fs}$ ,  $E_{mp}$  are amplifier energy of the free space, multi-path model, respectively.  $d_{toBS}$  defines the average distance from actor to base station, and it is measured as

$$d_{toBS} = 0.765 \frac{M}{2} \quad (2)$$

The working principle of the algorithm is explained below.

---

**Algorithm 1** Mid-point-based K-mean clustering

---

1. Find the distance between each point and origin.
  2. Sort the obtained distances in the ascending order.
  3. Segregate the distances into K equal sets.
  4. The middle point of every set is considered as the initial centroid.
- 

The above algorithm segregates the deployed sensors into K number of clusters. In each cluster, actor is placed in the middle point and acts as a cluster head. In the above algorithm, the network area is formed into K number of clusters and each actor has same workload. The actor (cluster head) broadcasts a cluster setup packet composed of its identity and location. Once the sensor receives the cluster setup message from an actor, it transfers acknowledgment to the actor composed of its identity and residual energy.

When the actor leaves the cluster, then the sensor which has highest energy harvesting capability and residual energy is selected as the backup cluster head. The main feature of this work is that energy harvesting sensors save some portion of energy during the non-cluster head mode. The saved energy is utilized when it acts as a BCH.

The mean energy arrival rate and energy required to transfer a packet to the nearest actor is considered to find the relation between energy harvesting capability and residual energy of a sensor. The energy harvesting sensor priority to act as a backup cluster head is computed using the following equation

$$P_n = \frac{\frac{Ar_n}{reqe_n}}{\sum_{k=1}^N \frac{Ar_n}{reqe_n}} \quad (3)$$

where  $Ar_n$  defines mean energy arrival rate and  $reqe_n$  denotes energy required to transfer a packet to actor.

With the help of above equation, we calculate the priority of each energy harvesting sensors then the sensor which has highest priority in the cluster is selected as the BCH. The elected BCH informs this information to all the cluster members by sending its identity to all the members of a cluster. It collects the information from all of its members and transfers the aggregated data to a nearest actor. The BCH process has been selected to reduce packet loss ratio and reduce energy consumption for cluster re-establishment process.

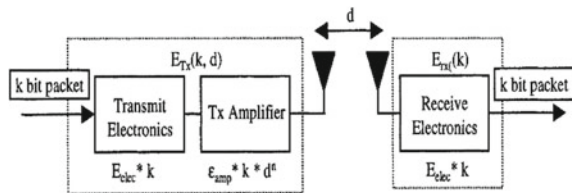
**Table 1** Simulation parameters

Parameters	Values
Duration	400 s
Data flow mechanism	Constant bit rate
Mobility model	Random waypoint
Initial energy of sensor	2J
Packet size	64 B
ATIM window size	20 ms
Beacon interval	100 ms
Transfer rate	20–60 pkt/s
Number of data channels	3–4

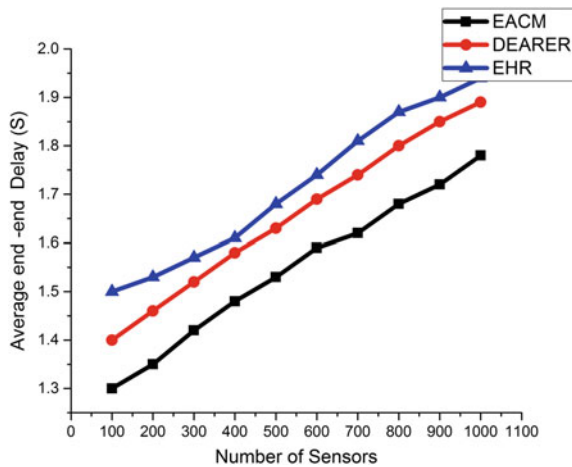
### 4 Analysis of Results

“In this section, we have analyzed the performance of the proposed coordination mechanism with its competitive mechanisms like ECH [2] and DEARER [12] by using various performance metrics such as packet delivery ratio, average residual energy, and average end-to-end delay.” The simulation of the proposed mechanism and existing mechanism is done in NS2 simulator, and the parameters are listed in Table 1).

**Fig. 4** Radio model



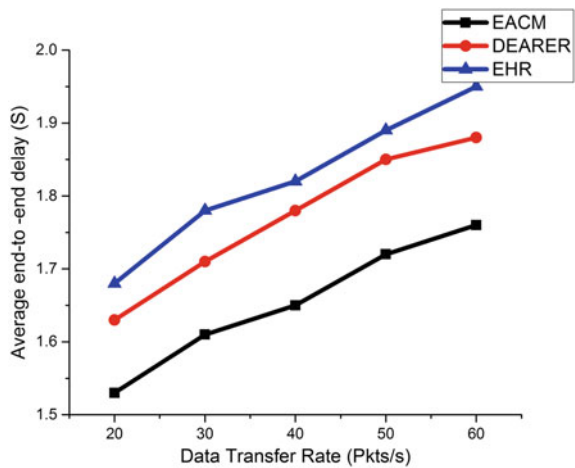
**Fig. 5** Delay versus sensors



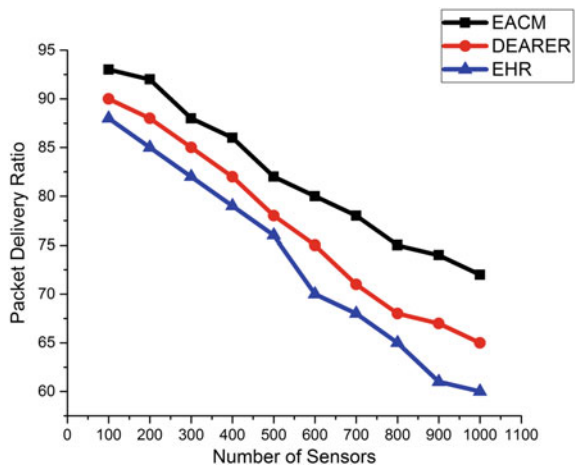
The sensors are resource-constrained nodes; hence, energy is an important metric which is need to be considered while designing any MAC protocol for WSAN. Hence, in the simulation we have used first-order radio model (Fig. 4) which is used by several researchers to evaluate the energy utilization of the protocols. The radio model is used to compute how much energy is utilized in the network for a certain amount of time.

“End-to-end delay refers to the time taken for a packet to be transmitted across a network from source to destination.” Figure 5 shows the performance of all the three protocols for average end-to-end delay by varying the number of sensors. Similar results are shown in Fig. 6 by varying the data transfer rate. Both the figures indicate that the proposed mechanism delivers data with less delay as compared to the existing mechanisms.

**Fig. 6** Delay versus data transfer rate



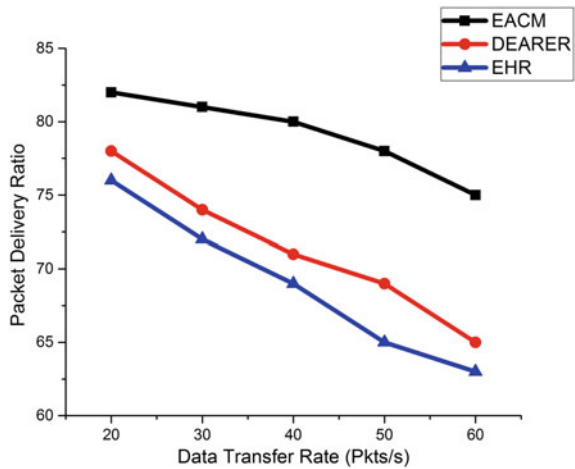
**Fig. 7** Packet delivery ratio versus sensors



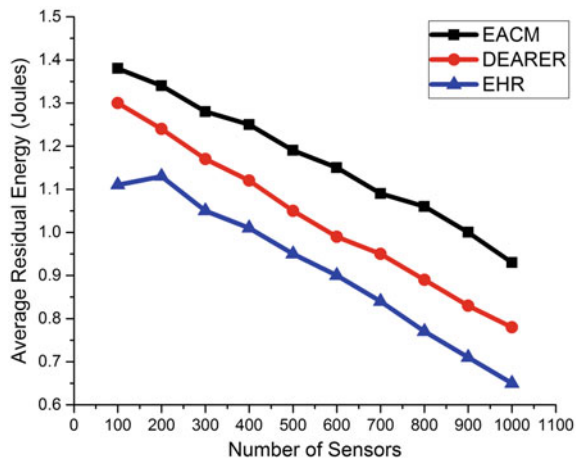
“The packet delivery ratio is defined as the ratio of packets that are successfully delivered to a destination compared to the number of packets that have been sent out by the sender.” Figures 7 and 8 show the packet delivery ratio vs number of sensors and data transfer rate, respectively. In both the cases, the proposed mechanism (EACM) outperformed the existing mechanisms like EHR [2] and DEARER [12].

“The average residual energy of sensor is denoted as sensor’s remaining energy in the network after the simulation running time.” Figures 9 and 10 depict the performance of all the three mechanisms for average residual energy in the network. The results indicate that the proposed mechanism (EACM) consumes lesser energy among three mechanisms.

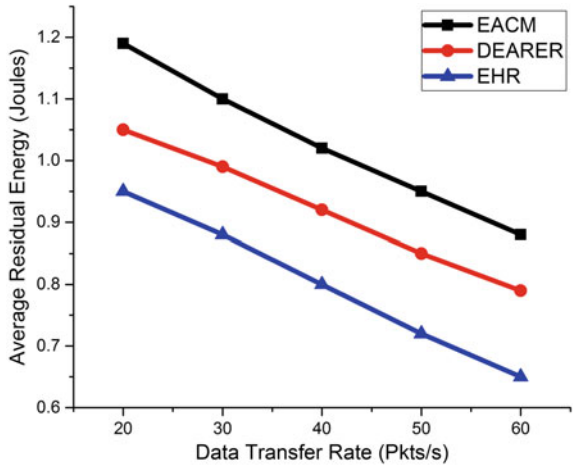
**Fig. 8** Packet delivery ratio versus data transfer rate



**Fig. 9** Average residual energy versus sensors



**Fig. 10** Average residual energy versus data transfer rate



## 5 Conclusions

“A delay- and energy-aware reliable coordination mechanism for sensor and actor networks has been proposed in our previous work.” In that work, the sensor works as a cluster head in the absence of actor. It consumes a lot of energy from sensors and also causes funneling effect (the sensors which are nearer to the cluster head consume more energy and loose a large number of packets as compared to the other sensors in the cluster) in the network. To overcome this problem, we introduced energy harvesting sensors in the WSAN. In this work, we deployed two types of sensors (energy harvesting sensors and normal sensors) and actors. The actor performs as CH, and an energy harvesting sensor acts as a BCH in the absence of actor. “The proposed protocol is simulated in NS2, and results are compared with its competitive protocols.”

## References

1. Sabri, N., Aljunid, S., Ahmad, R., Malik, M., Yahya, A., Kamaruddin, R., Salim, M.: Wireless sensor actor networks. In: IEEE Symposium on Wireless Technology and Applications, pp. 90–95. IEEE (2011)
2. Cao, Y., Liu, X-Y., Kong, L., Wu, M-Y., Khan, M.K.: EHR: routing protocol for energy harvesting wireless sensor networks. In: 2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS), pp. 56–63. IEEE (2016)
3. Bouachir, O., Mnaouer, A.B., Touati, F., Crescini, D.: Opportunistic routing and data dissemination protocol for energy harvesting wireless sensor networks. In: 2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2016)
4. Li, J., Liu, D.: An energy aware distributed clustering routing protocol for energy harvesting wireless sensor networks. In: 2016 IEEE/CIC International Conference on Communications in China (ICCC), pp. 1–6. IEEE (2016)



5. Yi, S., Huang, X., Wang, C.: EA-GPSR, a routing protocol for energy harvesting wireless sensor networks. In: 2015 4th International Conference on Computer Science and Network Technology (ICCSNT), vol. 1, pp. 1029–1032. IEEE (2015)
6. Gong, P., Xu, Q., Chen, T.M.: Energy harvesting aware routing protocol for wireless sensor networks. In: 2014 9th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), pp. 171–176. IEEE (2014)
7. Cui, R., Qu, Z., Yin, S.: Energy-efficient routing protocol for energy harvesting wireless sensor network. In: 2013 15th IEEE International Conference on Communication Technology (ICCT), pp. 500–504. IEEE (2013)
8. Kawashima, K., Sato, F.: A routing protocol based on power generation pattern of sensor node in energy harvesting wireless sensor networks. In: 2013 16th International Conference on Network-Based Information Systems, pp. 470–475, Sept 2013
9. Yin, W., Wenbo, L.: Routing protocol based on genetic algorithm for energy harvesting-wireless sensor networks. *IET Wirel. Sens. Syst.* **3**(2), 112–118 (2013)
10. Meng, J., Zhang, X., Dong, Y., Lin, X.: Adaptive energy-harvesting aware clustering routing protocol for wireless sensor networks. In: 2012 7th International ICST Conference on Communications and Networking in China (CHINACOM), pp. 742–747. IEEE (2012)
11. Eu, Z.A., Tan, H.P.: Adaptive opportunistic routing protocol for energy harvesting wireless sensor networks. In: 2012 IEEE International Conference on Communications (ICC), pp. 318–322. IEEE (2012)
12. Dong, Y., Wang, J., Shim, B., Kim, D.I.: Dearer: a distance-and-energy-aware routing with energy reservation for energy harvesting wireless sensor networks. *IEEE J. Sel. Areas Commun.* **34**(12), 3798–3813 (2016)
13. Kakarla, J., Majhi, B., Battula, R.: A delay and energy aware coordination mechanism for WSN. *Int. J. Commun. Syst* (2016)

# A Simulation Study: LMI Based Sliding Mode Control with Attractive Ellipsoids for Sensorless Induction Motor



Deepika, Shiv Narayan and Sandeep Kaur

**Abstract** This paper proposes an LMI based robust sliding mode control for an uncertain induction motor with attractive ellipsoid approach. The non-linear control strategy is based on minimization of an ellipsoid's size for convergence of motor dynamics onto an optimal sliding manifold, in minimum time. The controller gains as well as sliding manifold are found by solving matrix inequalities (LMI), using YALMIP and SEDUMI toolboxes. To demonstrate the efficacy of the methodology, a detailed analysis is done for this motor with time varying mismatched perturbations or external disturbances using MATLAB software. Simulations have proved the robustness of the proposed controller with high performance time domain responses, in presence of parametric and load variations.

**Keywords** Induction motor · Linear matrix inequalities (LMI)  
Attractive ellipsoid · Quasi-Lipschitz functions · Sliding mode control (SMC)  
Lyapunov function

## 1 Introduction

Induction motor drives are widely used in industrial applications [1]. Various control strategies have been discovered in past for the induction motor drive [2–9]. Vector control method is a well known technique which considers stator current as a set of two vectors- flux and torque [2, 3]. But, this method needs pre-requisite information about magnetic field lines of forces, for which sensors have to be implanted in closed loop system. Some schemes require the sensing of the magnetic fields, and some

---

Deepika (✉) · S. Narayan · S. Kaur  
Department of Electrical Engineering, PEC University of Technology, Chandigarh, India  
e-mail: sharmadeepika504@gmail.com

S. Narayan  
e-mail: shivnarayan@pec.ac.in

S. Kaur  
e-mail: sandipsaroa@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_45](https://doi.org/10.1007/978-981-13-1708-8_45)

are based on estimation of the flux distribution methods. But, both the techniques increase either complexity of the system or control problem.

Furthermore, robust control designs such as Youla parameterized two degree of freedom controllers [4],  $H_\infty$  [5], LQR (Linear Quadratic Regulators) [6], Adaptive control [7] etc. have also been devised for speed as well as position control of induction motor. But, most of these methods require optimization of weighted sensitivity and complementary sensitivity functions [4–6], which needs tuning of weighting functions, making a tedious process for complex non-linear systems.

Another inherent robust technique is sliding mode control (SMC), which belongs to the class of variable structure control theory. System dynamics are enforced onto the prescribed sliding surface in finite time, with a proper non-linear control law. The sliding phase has always proved to be ineffective towards the external disturbances as well as matched uncertainties, which enter into input matrix of the system [8, 10]. But, the problem lies in the mismatched uncertainties, which affect sliding dynamics. This challenge has largely attracted the control research community. Abera et al. [9] shows the multi-rate output feedback sliding mode control scheme utilized to control sensorless induction motor, with an integral action.

The Attractive Ellipsoid method is also known for ameliorating the effects of the mismatched perturbations on the system response. It is based on minimizing the dimensions of the ellipsoids, which contain all the state trajectories. These trajectories will be enforced near origin (equilibrium), present inside ellipsoid [11–13]. In [11], sliding mode controller is well presented by a same approach for mitigating mismatched disturbances, with a numerical example. In [12], a robust linear feedback controller is derived for a spacecraft using invariant (attractive) ellipsoid method, by solving matrix inequalities [13, 14].

To the best of author's knowledge, such a technique is not yet applied to any of the electrical systems. Therefore, this paper proposes the new kind of LMI based sliding mode control strategy for the control of induction motor parameters such as error in actual speed and currents from desired ones. The efficacy of new technique is shown using detailed analysis of various steady state and transient performances and convergence of solutions to the domain in vicinity of equilibrium is demonstrated, in presence of the time varying mismatched uncertainties, load and parametric variations.

Organization of this paper is as follows: Sect. 2 describes the mathematical model of induction motor. Next, Sect. 3 shows the sliding mode control methodology with LMI constraints. Section 4 gives the simulations and results obtained using YALMIP and SEDUMI toolbox in MATLAB software. Section 5 states the conclusion.

## 2 Dynamical Model of an Induction Motor

The mathematical model for a dynamical system for a three phase induction motor is derived in this section. The non-linear equations under d-q reference frame are given by [1, 2, 9]:

$$vds = p\phi_{ds} + R_s i_{ds} - w_s \phi_{qs} \quad (1)$$

$$vqs = p\phi_{qs} + R_s i_{qs} - w_s \phi_{ds} \quad (2)$$

$$vdr = p\phi_{dr} + R_r i_{dr} - (w_s - w_m)\phi_{dr} \quad (3)$$

$$vqr = p\phi_{qr} + R_r i_{qr} - (w_s - w_m)\phi_{qr} \quad (4)$$

$$T_m = 1.5P(\phi_{ds}i_{qs} - \phi_{qs}i_{ds}) \quad (5)$$

$$pw_m = (T_m - T_l - vw_m)/J \quad (6)$$

where  $vds, vqs$  are stator voltages and  $vdr, vqr$  are rotor voltages, in d-q reference frame.  $T_m, T_l$  are electromagnetic developed torque and load torque, respectively.  $w_s, w_m$  are synchronous and motor speed, respectively.  $p = \frac{d(\cdot)}{dt}$ ,  $J$  is moment of inertia of motor.  $\nu$  denotes viscous coefficient and  $P$  gives number of poles.  $R_s, R_r, L_s, L_r$  are respective stator and rotor resistances and inductances.  $\phi_{dr}, \phi_{qr}, \phi_{ds}, \phi_{qs}$  are flux linkages of stator and rotor windings, respectively, in d-q frame. Also,  $i_{ds}, i_{qs}, i_{dr}, i_{qr}$  are the stator and rotor currents, respectively. Here,  $L_m$  is mutual inductance. Solving (1)–(6), we get:

$$pi_{dr} = \sigma R_s i_{ds} - i_{qs} L_m w_s \vartheta - \sigma v_{ds} + i_{qr} w_s - \vartheta R_r i_{dr} + i_{qr} (-\chi L_s w_m)$$

$$pi_{ds} = \chi R_s i_{ds} + i_{qs} L_m w_m \sigma + \chi v_{ds} + i_{qs} w_s + \sigma R_r i_{dr} + i_{qr} \sigma L_r w_m$$

$$pi_{qr} = \sigma L_s w_m i_{ds} + i_{qs} \sigma R_s - \sigma v_{qs} - i_{dr} w_s + \chi L_s w_m i_{dr} - i_{qr} R_r \vartheta w_m \quad (7)$$

$$pi_{qs} = -w_s i_{ds} - i_{ds} \sigma L_m w_m + \chi v_{qs} - i_{dr} \chi L_m w_m + \sigma R_r i_{qr} - i_{qs} \chi R_s \quad (8)$$

where  $\sigma = \frac{L_m}{\chi}$ ,  $\chi = -L_m^2 + L_r L_s$ ,  $\eta = \frac{L_r}{\chi}$ ,  $\vartheta = \frac{L_s}{\chi}$ .

Above system of equations are non-linear in  $\dot{X}$ , affine in input  $u$  and can be written in the form:

$$\dot{X} = f(X) + G(X)u \quad (9)$$

where, state vectors are chosen as:  $X = [i_{ds} \ i_{qs} \ w_m \ i_{dr} \ i_{qr}]'$ , input vector is given as:

$u = [v_{ds} \ v_{qs} \ T_l]'$ , output vector:  $y = [i_{ds} \ i_{qs}]'$ . Further, (8) can be linearized about an operating point, to obtain the continuous time system of the form:

$$\begin{aligned} \Delta \dot{X}(t) &= A \Delta X(t) + Bu(t) + Dg(t, X) \\ Y(t) &= C \Delta X(t) \end{aligned} \quad (10)$$

where,  $\Delta X = X - X_e$  is error in system dynamics,  $Y \in R^k$ ,  $u \in R^m$ ,  $A \in R^{n \times n}$ ,  $B \in R^{n \times m}$ ,  $C^{k \times n}$ ,  $g$  represents bounded Quasi-Lipschitz non-linear external disturbances,

gains of the perturbations/disturbances are defined by matrix  $D$ . Now, system dynamics are defined by:

$$\begin{aligned}\Delta X &= [\Delta i_{ds} \Delta i_{qs} \Delta w_m \Delta i_{dr} \Delta i_{qr}] \\ Y &= [\Delta i_{ds} \Delta i_{qs}] \\ u &= [\Delta v_{ds} \Delta v_{qs} \Delta T_l]\end{aligned}\tag{11}$$

### 3 SMC Based on Attractive Ellipsoids Scheme

#### 3.1 Control Objective

The main control aim is to obtain a robust control law which ensures convergence of all the error states to zero, even in presence of mismatched uncertainties in finite time.

$$\Delta X = [\Delta i_{ds} \Delta i_{qs} \Delta w_m \Delta i_{dr} \Delta i_{qr}] \rightarrow 0\tag{12}$$

**Assumption 1** The disturbances are assumed to be bounded and bounds are given as:

$$g'Q_g g \leq g_0 + \Delta X'Q_x \Delta X\tag{13}$$

where  $Q_g \in R^{2 \times 2}$ ,  $Q_x \in R^{5 \times 5}$  are positive definite matrices and  $g_0$  is a positive constant.

#### 3.2 Problem Statement

To achieve the prescribed control objective, firstly, decompose matrix  $B$  into  $B_1 \in R^{2 \times 3}$ ,  $B_2 \in R^{3 \times 3}$  and then, modify the system dynamics (10) with transformation matrix  $T$ :  $\Omega = T \Delta X$

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad T = \begin{bmatrix} I_{n-m} & -B_1 B_2^{-1} \\ 0 & B_2^{-1} \end{bmatrix} \quad n = 5, m = 3, |B_2| \neq 0$$

$$\dot{\Omega}_1 = A_{11} \Omega_1 + A_{12} \Omega_2 + D_1 g(t, x)$$

$$\dot{\Omega}_2 = A_{21} \Omega_1 + A_{22} \Omega_2 + u(t) + D_2 g(t, X)$$

$$\Omega = \begin{bmatrix} \Omega_1 \\ \Omega_2 \end{bmatrix}, TAT^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, TD = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \tag{14}$$

$$\Omega_1 \in R^2, \Omega \in R^3, A_{11} \in R^{2 \times 2}, A_{22} \in R^{3 \times 3}, A_{12} \in R^{2 \times 3}, A_{21} \in R^{3 \times 2}, D_1 \in R^{2 \times 2}, D_2 \in R^{3 \times 2}$$

Sliding manifold is chosen according to regulation objectives:

$$C_1\Omega = C\Omega_1 + \Omega_2 = 0 \tag{15}$$

where,  $C_1 = \begin{bmatrix} C & 1 \end{bmatrix}$  will be found from convex optimization problem, illustrated in next section. In (16), the first term denotes the continuous equivalent control law for sliding mode and second term shows the discontinuous control for reaching mode.

$$u(t) = -(C_1B)^{-1}C_1A\Omega(t) - K(\Omega)sign(C_1\Omega(t)) \tag{16}$$

where,  $|C_1B| \neq 0$  and constant gain is  $K(\Omega) = \sqrt{a + \Omega^T R \Omega}$  with positive constant  $a$ , and matrix  $R \in R^{5 \times 5}$  such that:

$$0 \leq a_{\min} \leq a \leq a_{\max}, 0 \leq \|R\| \leq b \tag{17}$$

### 3.3 Controller Design

**Definition** The ellipsoid given by:

$$\varepsilon(P) = \Omega^T P \Omega < 1, P > 0, \Omega \in R^5 \tag{18}$$

is invariant when any system state trajectory starting in ellipsoid remains inside it for  $t > 0$ . Further, it is attractive when system state trajectory starts outside this ellipsoid and converges to it. Hence, we take optimization problem as:  $\max trace(P) = \min trace(P^{-1})$  for minimizing the sum of squares of ellipsoid's semiaxis.

The Attractive ellipsoid scheme depends on minimization of an ellipsoid's size, subjected to certain linear matrix inequalities (LMIs) [13]. This linear optimization involves the linear objective function as well as constraints, which makes it convex optimization. As the stability of the sliding mode dynamics has always been proved by using Lyapunov's concept of stability. Therefore, the proofs of the results described below, are also explained by Lyapunov's theory [12, 13]. To achieve convergence of system states onto the smallest invariant ellipsoid, certain LMIs have to be solved using certain lemmas with objective function as trace (Z). Let  $Y = CP$ .

**Lemma [13]:** *LMIs obtained by S-procedure for fixed  $\tau_1, \tau_2, \delta$ , give the solutions  $(a, V, Z, Y, P, R)$ :*

$$\begin{aligned}
 & \begin{bmatrix} b^2 I_5 & R \\ R & I_5 \end{bmatrix} > 0, R - \frac{a}{g_0} Q_x \geq 0, \begin{bmatrix} P & [P \ -Y^T] \\ [P \ -Y] & TZT^T \end{bmatrix} \geq 0 \\
 & \begin{bmatrix} \frac{a}{g_0} Q_g & D^T T^T \\ TD & \begin{bmatrix} \delta P & 0 \\ 0 & V \end{bmatrix} \end{bmatrix} > 0, \begin{bmatrix} \frac{P}{\delta} & Y^T \\ Y & I_3 - V \end{bmatrix} \geq 0, \begin{bmatrix} PA_{11}^T + A_{11}P - Y^T A_{12}^T & [P \ -Y^T] \\ -A_{12}Y + \tau P + \tau_2 D_1 R_g^{-1} D_1^T & \\ & \begin{bmatrix} P \\ -Y \end{bmatrix} \\ & & -\tau_2 Q_x^{-1} \end{bmatrix} \leq 0 \\
 & \tau \geq 0, P > 0, V > 0, \lambda > 0 \tag{19}
 \end{aligned}$$

Here,  $a, \tau, \lambda$  are positive numbers,  $P$  denotes  $2 \times 2$  positive definite matrix and  $Y$  is  $5 \times 5$  positive definite matrix. This gives the quasiminimal attractive (invariant) ellipsoid  $\varepsilon(Z^{-1})$  for the system. The linear sliding surface is given by:

$$C_1 = [YP^{-1} \ I_3]T \tag{20}$$

Hence, optimal sliding motion is given by:

$$\dot{\Omega}_1(t) = (A_{11} - A_{12}C)\Omega_1(t) + D_1g(t, x), \ \Omega_2(t) = -C\Omega_1(t) \tag{21}$$

### 4 Simulation Results

The various parameters of induction motor are given in Table 1. The uncertain matrices for the linearized system are as follows:

$$g = \begin{bmatrix} 0.0025 \cos(0.4t) - 0.007 \sin(0.4t) \\ 0.05 \cos(0.4t) + 0.023 \sin(0.4t) \end{bmatrix}, \ D = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ -1 & 1 \\ 0 & -0.002 \\ 0 & 0.025 \end{bmatrix} \tag{22}$$

Simulations are carried out through MATLAB programming with YALMIP and SEDUMI toolboxes. The differential equations are solved through ode23 solver with the tolerance of  $1e-2$ . The LMI results are obtained as:

$$P = \begin{bmatrix} 178.6645 & 16.0159 \\ 16.0159 & 316.5438 \end{bmatrix} \tag{23}$$

**Table 1** Induction motor rating

Parameters	Values
Line voltage	460 V
Power	3 HP
Phases	3
Frequency	60 Hz
Full load speed	1750 rpm
Full load efficiency	88.5%
Power factor	80%
Poles	4
Full-load current	4 A
Xm	139.0 Ω
Rs	1.77 Ω
Rr	1.34 Ω
Xs	5.25 Ω
Xr	4.57 Ω

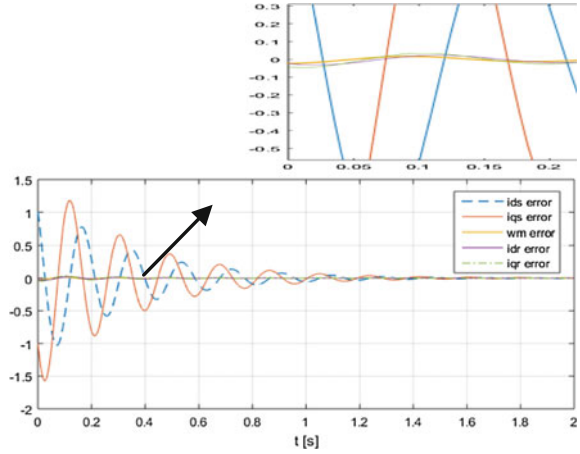
Controller parameters are presented in Table 2. Also, Fig. 1 shows the convergence of all error states in 1 s to zero. Error in speed is regulated to zero in 0.1 s with maximum peak overshoots 0.01%. Figure 2 describes the two dimensional plots between error in stator current and speed. This shows that the speed error is converged to domain in vicinity of origin, corresponding to finite stator current error. Figure 3 and Fig. 4 gives three dimensional plots between various stator currents, rotor currents and speed errors, respectively. Figure 5 plots the convergence of sliding manifold to zero in 1 s. Figure 6 describes chatter free control effort versus time. Figure 7 shows

**Table 2** Controller parameters

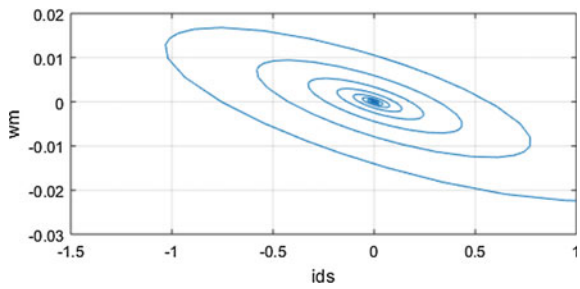
Parameters	Values
$\tau_1$	1.526
$\tau_2$	1
$\delta$	0.1
$Q_x$	$\begin{bmatrix} 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0.0001 & 0 & 0 & 0 \\ 0 & 0 & 0.0001 & 0 & 0 \\ 0 & 0 & 0 & 0.00003 & 0 \\ 0 & 0 & 0 & 0 & 0.00003 \end{bmatrix}$
$Q_g$	$\begin{bmatrix} 130 & 15 \\ 15 & 400 \end{bmatrix}$
$g_0$	1
$\beta$	1e8



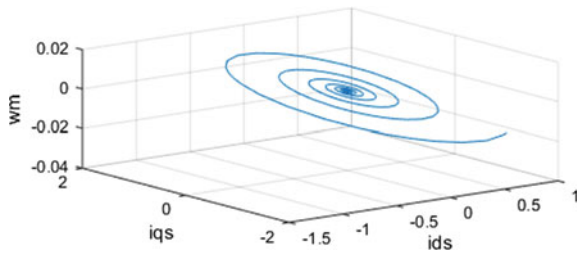
**Fig. 1** Error states versus time



**Fig. 2** Speed error versus current error

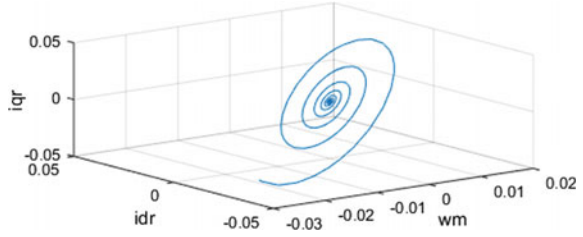


**Fig. 3** 3D plots b/w speed and stator currents

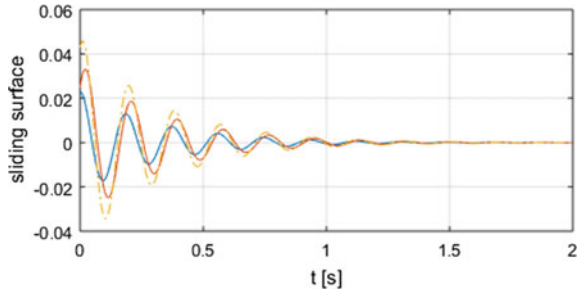


that the closed loop system is found to be robust in presence of parametric variations in winding resistances. Figure 8 shows that actual speed of motor tracks the varying reference properly, without any oscillations. Figure 9 shows that the system is robust to 10% load disturbance, added from 1 to 2 s.

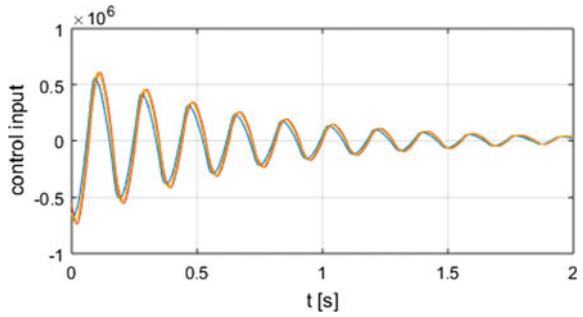
**Fig. 4** 3D plots b/w speed and rotor currents



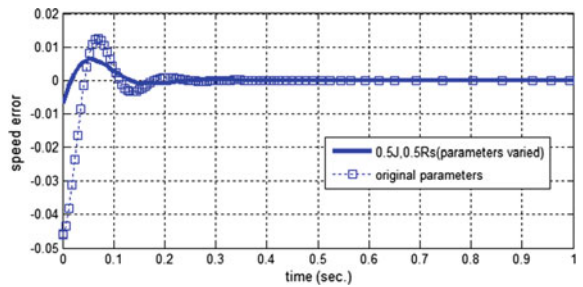
**Fig. 5** Convergence of sliding surface



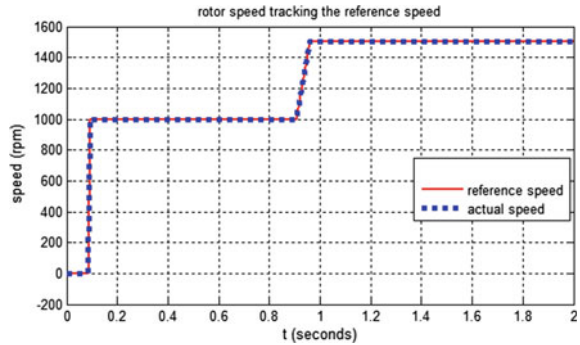
**Fig. 6** Control effort versus time



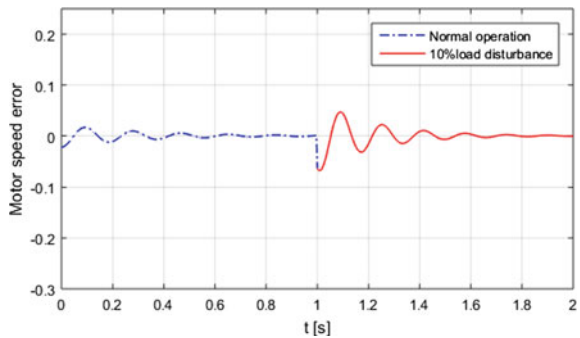
**Fig. 7** Speed error with parametric variations



**Fig. 8** Tracking of reference speed



**Fig. 9** Convergence of speed error with 10% load disturbance from 1 to 2 s



## 5 Conclusion

In this work, an optimal sliding mode controller is designed with LMI constraints, for an induction motor, using attractive ellipsoids approach. YALMIP and SEDUMI toolboxes are used to solve the optimization problem with MATLAB software. The conclusions that can be derived from the presented simulations and results are as follows:

1. The convergence of error states to the quasiminimal region of equilibrium has been achieved with the proposed controller for an induction motor.
2. The various responses of the controlled system with the proposed controller provided very less settling time (maximum 1.2 s) and peak overshoots.
3. Robust performances are also achieved even with the non-linear disturbances, load torque disturbances and variation in parameters.

## References

1. Krause, P.C.: Analysis of Electric Machinery, McGraw-Hill (1986)
2. Paice, D.A.: Induction motor speed control by stator voltage control. *IEEE Trans. Power Appar. Syst.* **585**–590 (1968)
3. Abdel-Salam, M., Abou-Shadi, S., Sayed, Y.: Speed sensorless vector control of induction motor as influenced by core-loss. *Electr. Mach. Power Syst. J.* **27**, 921–939 (1998)
4. Sayed, Y., Abdelfatah, M., Abdel-Salam, M., Abou-Shadi, S.: Design of robust controller for vector controlled induction motor based on Q-parameterization theory. *Electr. Power Compon. Syst.* **30**, 981–999 (1998)
5. Kao, Y.T., Liu, C.H.: Analysis and design of microprocessor-based vector controlled induction motor drives. *IEEE Trans. Ind. Electron.* **39**, 46–54 (1992)
6. Yassine, B., Fatiha, Z., Chrifi-Alaoui, L.: LQR-PI controller dedicated to the indirect vector control without speed sensor for an asynchronous motor. In: 2015 16th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), Monastir, pp. 634–640 (2015)
7. Chang, C., Liu, C.H.: Adaptive speed sensorless induction motor drive for very low speed and zero stator frequency operation. *Elect. Power Compon. Syst.* **38**, 804–819 (2010)
8. Utkin, V.I.: Sliding Modes and Their Applications in Variable Structure Systems. Nauka, Moscow (1974)
9. Abera, G.E., Agarwal, V., Bandyopadhyay, B., Janardhanan, S.: Multirate output feedback sliding mode controller for sensorless induction motor. In: 2005 IEEE International Conference on Industrial Technology, pp. 877–881 (2005)
10. Utkin, V.I.: Sliding mode control design principles and applications to electric drive. *IEEE Trans. Ind. Electron.* **40**, 23–36 (1992)
11. Polyakov, A., Poznyak, A.: Invariant ellipsoid method for minimization of unmatched disturbances effects in sliding mode control. *Automatica* **47**, 1450–1454 (2011)
12. Gonzalez-Garcia, S., Polyakov, A., Poznyak, A.: Using the method of invariant ellipsoids for linear robust output stabilization of spacecraft. *Autom. Remote Control* **72**, 540–555 (2011)
13. Polyakov, A., Poznyak, A., Vadim, A.: Attractive Ellipsoids in Robust Control. Springer (2014)
14. Boyd, S., Ghaoui, E., Feron, E., Balakrishnan, V.: Linear Matrix Inequalities in System and Control Theory. SIAM, Philadelphia (1994)

# A Study of Environmental Impact Assessment on the Performance of Solar Photovoltaic Module



Sanhita Mishra, S. C. Swain, P. C. Panda and Ritesh Dash

**Abstract** In this paper, few experiments are conducted in a stand-alone PV system and various current versus voltage and power versus voltage are drawn to study different characteristics and one of them is verified through MATLAB/SIMULINK. The performance of the system is calculated using battery connected load under both AC and DC System. In this experimental work battery charging, discharging property is also studied. Effect of different colour spectrum on solar PV panel is investigated. How the partial shading and tilting of solar panel with some angle affect the output power is clearly visible in the experimental work.

**Keywords** PV panel · Colour spectrum · Partial shading · Tilt angle · Batteries

## 1 Introduction

Solar irradiation without emitting greenhouse gases is good remedies for today's energy and environmental issue. As the fossil fuels are exhaustible and degrading the environment, renewable source has become rightful nowadays [1, 2]. In this paper, various tests are conducted for a stand-alone PV system set-up which exist in energy system laboratory of KIIT University. A stand-alone PV system consists of an adjustable PV panel, regulated lamps and a main controller with batteries. The standard specification of the panel is  $P_{MPP} = 37$  W and total cell = 36,  $V_{OC} = 21.8$  V,  $I_{SC} = 2.4$  A,  $V_{MPP} = 17.2$  V and  $I_{MPP} = 2.20$  A maximum irradiation is  $1000$  W/m<sup>2</sup> and standard temperature is 25°. Every time PV module absorbs a wide range of radiation and its characteristic changes from morning to evening. So experiments are conducted on solar panel to analyse the performance and its properties can be studied briefly. As different colours have different wavelength and solar light is combination of all colour visible spectrum, it is proved experimentally

---

S. Mishra (✉) · S. C. Swain · P. C. Panda · R. Dash  
School of Electrical Engineering, KIIT University, Bhubaneswar, India  
e-mail: sanhita.mishra@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_46](https://doi.org/10.1007/978-981-13-1708-8_46)

that red colour has maximum power in comparison with purple colour. When the panel is tilted with some angle away from the sun, then automatically maximum power will be reduced which is verified in the experiment.

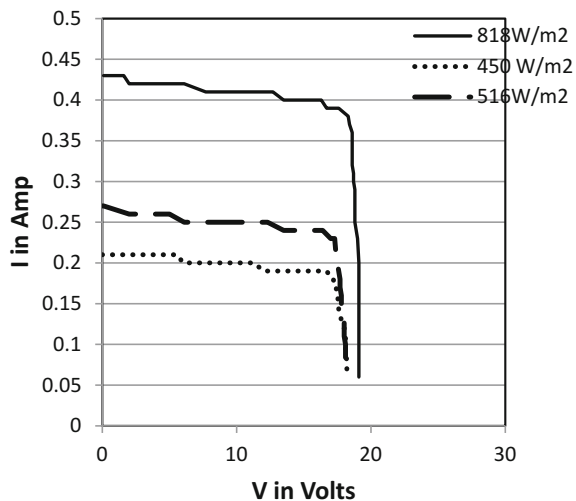
## 2 Different Types of Experiment on Solar Panel

**2.1.** A variable resistor, ammeter and a voltmeter are incorporated with the PV module and irradiance has been measured using solar power meter and then the curve was drawn to see the  $P_{MPP}$  with different irradiance. Using MATLAB/Simulink, the P-V and I-V property is verified for a single irradiance [3, 4]. The set-up used for conducting the experiments is given below (Figs. 1, 2, 3 and 4).

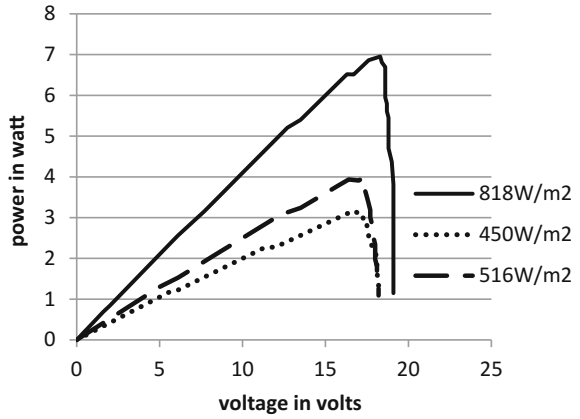
**2.2.** A case study is conducted to verify effect of different colour filters in a solar module. Considering two colour such as red and purple and also without any colour, the properties are being studied. Normal sunlight consists of all colour of visible spectrum [5]. As we know, energy of photon is  $E=hc/\lambda$  [6] where  $h$  is Plank's constant and  $\lambda$  is the wavelength of light. Wavelength of different colours greatly affects photovoltaic module. The performance of solar PV panel is experimented by exposing the panel to different color spectrum and their corresponding wavelength. So it is seen from the graph that red colour has maximum current in comparison with other but with normal sunlight the current is maximum (Figs. 5 and 6).

**2.3.** When the sunlight is incident on the solar panel, the output power is being affected by the proper position of the panel. In this paper, experiment is being done by considering different tilt angles. The panel is being tilted with  $15^\circ$ ,  $5^\circ$  and  $20^\circ$ , and it is investigated that more is the tilt angle of the panel away from the light less

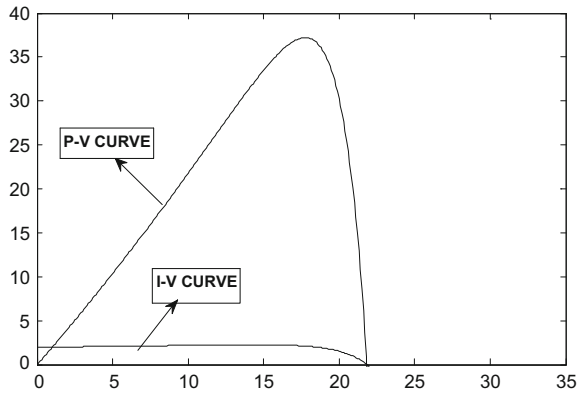
**Fig. 1** I-V characteristics with different irradiance



**Fig. 2** P-V characteristics with different irradiance



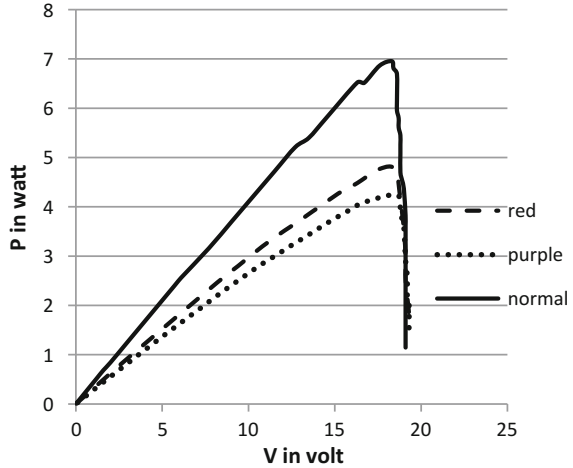
**Fig. 3** I-V and P-V curves from simulation



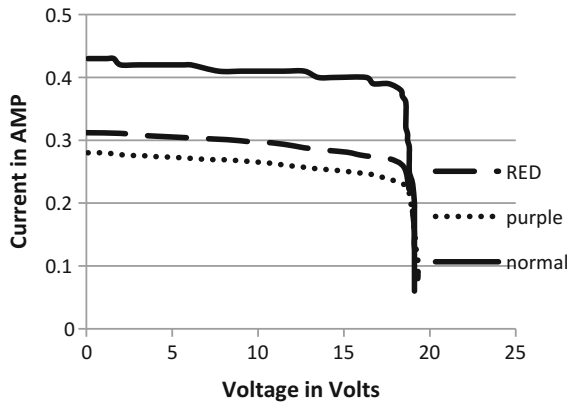
**Fig. 4** Set-up to conduct the experiments



**Fig. 5** P-V curve with different colour spectrum



**Fig. 6** P-V curve with different colour spectrum



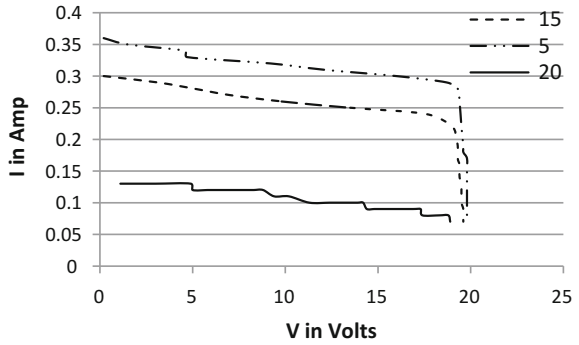
current will be generated. Positioning a solar module is very difficult because earth daily rotates round the sun. A solar module also could be mounted on a sun tracker to follow the sun as it travels through the sky. So through this experiment it can be concluded that tilt angle plays vital role for generating power in solar panel (Figs. 7 and 8).

**2.4.** Shading of solar panel affects the efficiency of solar panel because exposure of sun towards solar panel decreases [7]. When shading is done in a solar cell, there is a drop occurs in pumping electrons from one side to others. So fitting of bypass diode helps in improving the efficiency. Here partial shading is done by making one of the panels completely covered from two panels connected in series. Then, the characteristic is studied and shown in the Fig. 9.

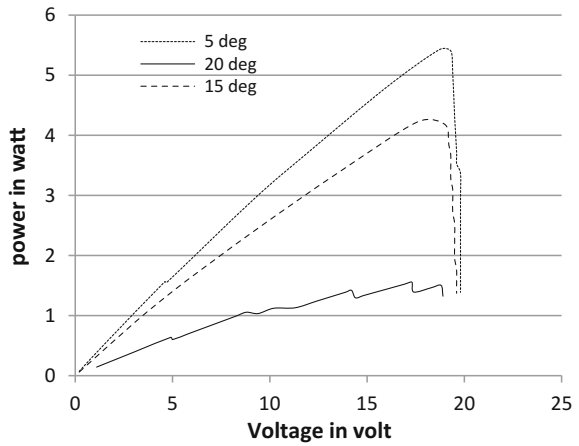
**2.5.** Stand-alone PV contains DC load and battery for storage purpose. Mainly charge controller helps in controlling the module voltage. As test is conducted by considering DC load of 420 Mega ohm, it is also concluded that array power = load



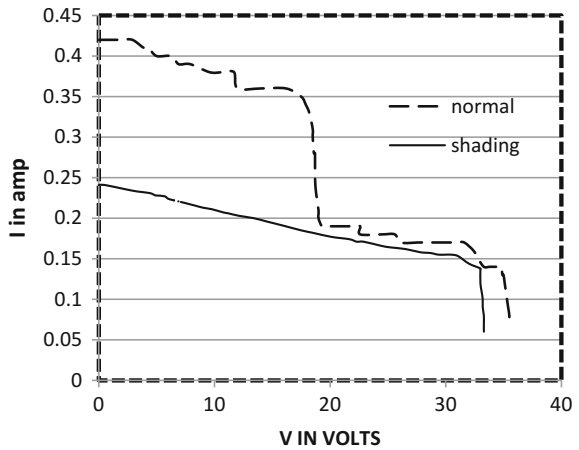
**Fig. 7** I-V characteristics with different tilt angles



**Fig. 8** P-V characteristics with different tilt angles



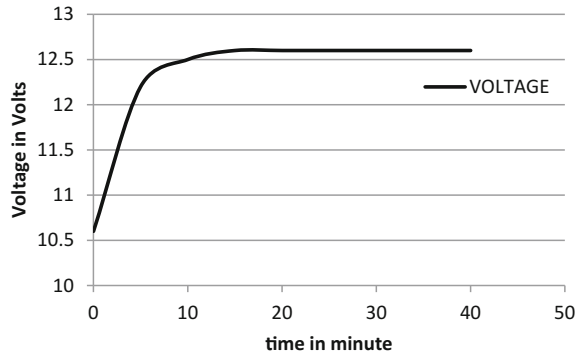
**Fig. 9** I-V characteristic for shading



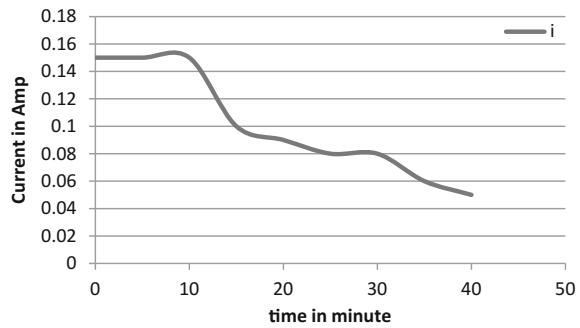
**Table 1** P-V panel operated with no load

No. of module	Module current	Module voltage	Module power	Battery I/P current	Battery I/P voltage	Battery I/P power	DC load current	DC load voltage	DC load power
2	0.16 A	28	4.48	7 A	0.5 V	3.5	0.02	7.4	0.148

**Fig. 10** Charging voltage for storage device



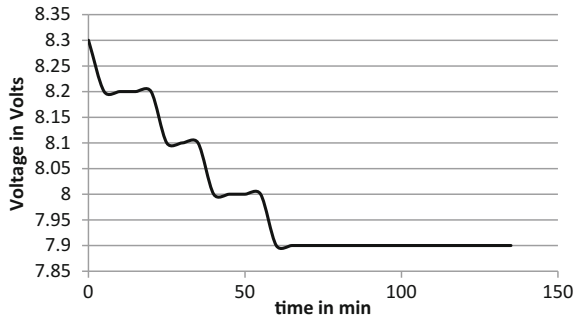
**Fig. 11** Charging current for storage device



power + battery power + power loss by charge controller. If the battery power is coming as negative, that means battery is discharging through load [8] (Table 1).

2.6. As battery is very important for solar model, it is also vital to improve the performance of the battery and this battery performance depends on charging of battery and the case study has been done to charge a 12 V battery and 4 methods of battery charging method trickle charging, constant current charging, constant voltage charging and float charging [9]. So in this experiment also the battery charging and discharging property are observed. Figures 10 and 11 show the charging voltage and current for a storage device. From Fig. 10, it can be found that charger takes 14 min to boost the voltage up to 12.2 V, which is quite satisfactory in this case and that of charging current is decreasing gradually from peak value towards zero magnitude showing successful operation of charger battery system. Here the battery has 8.3 V and it starts to discharge after 150 min continuous reading and it is shown that the battery voltage is decreasing (Fig. 12).

**Fig. 12** Discharging voltage of storage device



### 3 Conclusion

Aim of this study is to find the effect of environmental parameters on the performance of solar PV module. Being a promising energy in the world, it is affected by many factors and hence affecting the reliability and efficiency of module. It is usually affected by many factors such as dust, colour, irradiance and shading. In this experiment, a storage device is connected at the input side for increasing the performance under harsh environmental condition.

**Acknowledgements** Authors would like to thank the school of electrical engineering and KIIT University for providing necessary laboratory facility and support to carry out the experiment.

### References

1. Kim, J.Y., Kim, J., Amsden, J.J., Roh, J., Park, I., Yoon, D.Y., Kim, H., Lee, C.H.: Temperature dependence and impedance characteristics of hybrid solar cells based on poly(phenylene vinylene): ZnO nanoparticles with added surfactants. *IEEE J. Photovolt.* **7**(4) (2017)
2. Omran, W.: Performance analysis of grid-connected photovoltaic systems. Ph. D. thesis, Waterloo, Ontario, Canada (2010)
3. Pavan Kumar, A.V., Parimi, A.M., Vma Rao, K.: Performance analysis of a two-diode model of PV cell for PV based generation in MATLAB. In: ICACCCT. IEEE (2014)
4. Swain, S.C., Ali, S.M., Dash, R., Mohanta, A.K.: Performance evaluation of photovoltaic system based on solar cell modelling. In: ICCPCT. IEEE (2015)
5. Bhol, R., Pradhan, A., Dash, R., Ali, S.M.: Environmental effect assessment on performance of solar PV panel. In: ICCPCT. IEEE (2015)
6. Sudhakar, K., Jain, N., Bagga, S.: Effect of color filter on the performance of photovoltaic module (2013). 978-1-4673-6030-2
7. Ishaque, K., Salam, Z., Syafaruddin: A comprehensive MATLAB Simulink PV system simulator with partial shading capability based on two-diode model. *Elsevier Solar Energy* **85**, 2217–2227 (2011)

8. Jain, A., Tharani, K., Dhall, H., Jain, A., Tharani, K., Dhall, H., Singh, N.K., Bhatia, S.: Solar home lighting system with AC and DC loads. IOSR J. Electr. Electron. Eng. (IOSR-JEEE) **12**(3) Ver. II (2017)
9. Chauhan, A.: MPPT control PV charging system for lead acid battery. M.Tech. thesis, NIT, Rourkela (2014)

# Design of a Compact Ultra-wideband Bandpass Filter Employing Defected Ground Structure and Short-Circuited Stubs



Sarbani Sen and Tamasi Moyra

**Abstract** A compact-sized ultra-wideband (UWB) bandpass filter (BPF) comprising short-circuited stubs and defected ground structures (DGS) is presented in this paper. The BPF structure is obtained by cascading the DGS-based lowpass filter (LPF) and the optimum distributed highpass filter (HPF). The optimum BPF design provides comparatively low insertion loss and a wide attenuation band. Since the structure is compact, it improves the selectivity. The BPF is designed using the RO5880 substrate of thickness 1.57 mm, and the dielectric constant of it is  $\epsilon_r = 2.2$ . The average group delay calculated for the BPF is approximately 0.5 ns. The frequency response of UWB composite BPF supports the UWB range (3.1–10.6 GHz), which is suited for Bluetooth and other wireless applications.

**Keywords** Ultra-wideband (UWB) bandpass filter · Optimum distributed HPF Defected ground structure (DGS)

## 1 Introduction

Nowadays, the ultra-wideband technology has become very popular due to its enormous applications. The Federal Communications Commission (FCC) has granted the application of ultra-wideband (ranges from 3.1 to 10.6 GHz) for commercial purposes in 2002 [1]. Numerous devices like antennas, power combiner/divider, and RF amplifiers support this technology [2–4]. Also, the microwave filters using this technology have been drawing more attention of researchers, working in the microwave field. There are several methods available to implement the UWB bandpass filter structure [5–8]. It is often seen that lumped-element filters are difficult to implement

---

S. Sen (✉) · T. Moyra  
Electronics and Communication Department, National Institute of Technology  
Agartala, Agartala 799046, Tripura, India  
e-mail: sensarbani77@gmail.com

T. Moyra  
e-mail: tamasi\_moyra@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_47](https://doi.org/10.1007/978-981-13-1708-8_47)

at microwave frequencies [9]. Hence, conventional microstrip filters are preferred. Besides this, a microstrip bandpass filter of compact size and wide attenuation band is highly demanded [10]. If a UWB filter is designed using lossy substrate, it may absorb high-frequency signals [11].

This paper proposes a simple, compact-sized structure of UWB bandpass filter. Separately, an optimum distributed highpass filter (HPF) of 3.5 GHz cutoff frequency and a lowpass filter (LPF) of 11.11 GHz cutoff frequency are designed. Later on, both the structures have been cascaded to provide a bandpass response. The highpass filter is constructed using such elements which are known as commensurate transmission line elements [12]. It is formed by cascading six short-circuited stubs having electrical length  $\theta_c$  at its cutoff frequency, and these stubs are separated by some connecting lines having an electrical length of  $2\theta_c$ . The filter that consists of 'n' stubs has an insertion function of  $(2n - 1)$  so that the highpass response has  $(2n - 1)$  ripples. These 'n' ripples compare with the n-stub bandpass filter. The electrical length can be calculated as

$$\left(\frac{\pi}{\theta_c} - 1\right)f_c = f, \quad (1)$$

where  $f_c$  indicates the cutoff frequency and ' $f$ ' specifies the extended passband frequency. For the practical design of HPF, two to six stubs and a passband ripple of 0.1 dB for  $\theta_c = 25^\circ, 30^\circ, 35^\circ$  are needed.

Similarly, the lowpass filter (LPF) is designed in the ground plane using the technique called defected ground structure (DGS). The structure is obtained in such a way that it does not disturb the signal plane properties. The combined structure [13] provides the ultra-wideband response.

## 2 Design Theory and Simulated Results

For designing this BPF, the primary requirement is to design a highpass filter of wide passband at least up to 10.6 GHz. Many methods are there to implement an HPF, but in this paper, the optimum distributed structure has been chosen to provide the ultra-wideband response.

### 2.1 HPF Structure and Its Circuit Model

The highpass filter has been designed using six stubs. Each of the stubs is connected to the ground via hole. The short-circuited stub has the electrical length,  $\theta_c = 25.23^\circ$ , and the connected lines have the electrical length of,  $2\theta_c = 50.46$ . The filter has a primary passband that ranges from  $\theta_c$  to  $(\pi - \theta_c)$ , where ' $\theta_c$ ' is referred as the cutoff frequency. It is said that the smaller the electrical length in a particular cutoff

frequency, the wider the passband. That is why the optimum value of ‘ $\theta_c$ ’ has been chosen as  $25^\circ$ . By making use of the element values tabulated in [10] for optimum distributed highpass filter (HPF) with 0.1 dB passband ripple, the impedance values of short-circuited stubs and the connecting lines can be found out. The formulas are given below:

$$Z_i = Z_0/y_i, \tag{2}$$

$$Z_{i,i+1} = Z_0/y_{i,i+1}, \tag{3}$$

where  $Z_0$  is said to be the characteristic impedance and ‘ $i$ ’ is the variable.  $Z_i$  calculates the impedance values of short-circuited stubs, and  $Z_{i,i+1}$  calculates the impedance values of the connecting lines. For an example, the element value can be calculated as

$$y_1 = 0.25038 + \frac{(0.35346 - 0.25038)}{5} \times 9.5 \tag{4}$$

Similarly, all the ‘ $y$ ’ values can be calculated by using the table mentioned in [12]. The impedance values are calculated using these ‘ $y$ ’ values. The impedance values for the respective stubs are calculated and shown in the next section (Table 1).

The highpass filter is a Chebyshev model which is realized using microstrip transmission line, and the software used for designing purpose is MOM-based IE3D. Later on, an equivalent circuit is proposed which shows a good agreement with the simulated one. The circuit models are designed using the ADS software. Using these impedance values and associated electrical lengths, the length and width of each stub for the microstrip structure are obtained. The values are shown in Table 2.

**Table 1** Impedance value for the stubs

Elements	Value ( $\Omega$ )
Z1 = Z6	112.05
Z2 = Z5	80.69
Z3 = Z4	71.78
Z1,2 = Z5,6	48.27
Z2,3 = Z4,5	49.75
Z3,4	50.14

**Table 2** Design parameters for microstrip

Terms	Length (mm)	Width (mm)
Z1 = Z6	5.15	1.04
Z2 = Z5	5.065	2.14
Z3 = Z4	5.04	2.66
Z1,2 = Z5,6	9.88	4.8
Z2,3 = Z4,5	9.89	4.8
Z3,4	9.9	4.8

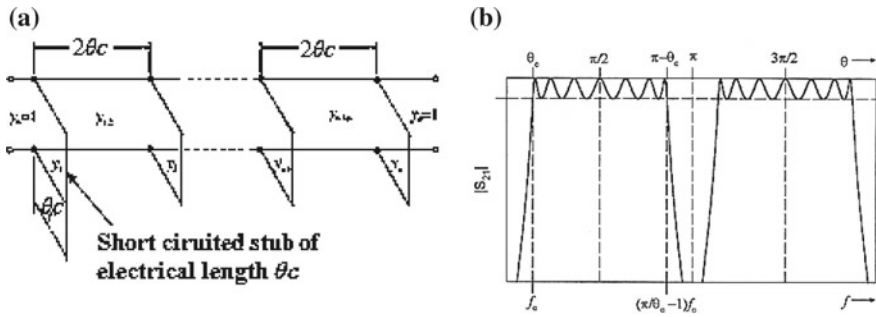


Fig. 1 a Optimum distributed HPF and b filtering characteristics of the HPF

The conventional circuit representation of the proposed highpass filter (HPF) is shown in Fig. 1.

Using the parameter values from Table 2, the optimum HPF structure has been designed in microstrip. The diameter of the holes which are attached to the stubs is 3 mm. The microstrip structure and the simulated response are shown, and also the group delay of the passband is calculated (Fig. 2).

The frequency response shows both S11 and S21 values in dB. The cutoff frequency obtained from the designed structure is 3.5 GHz.

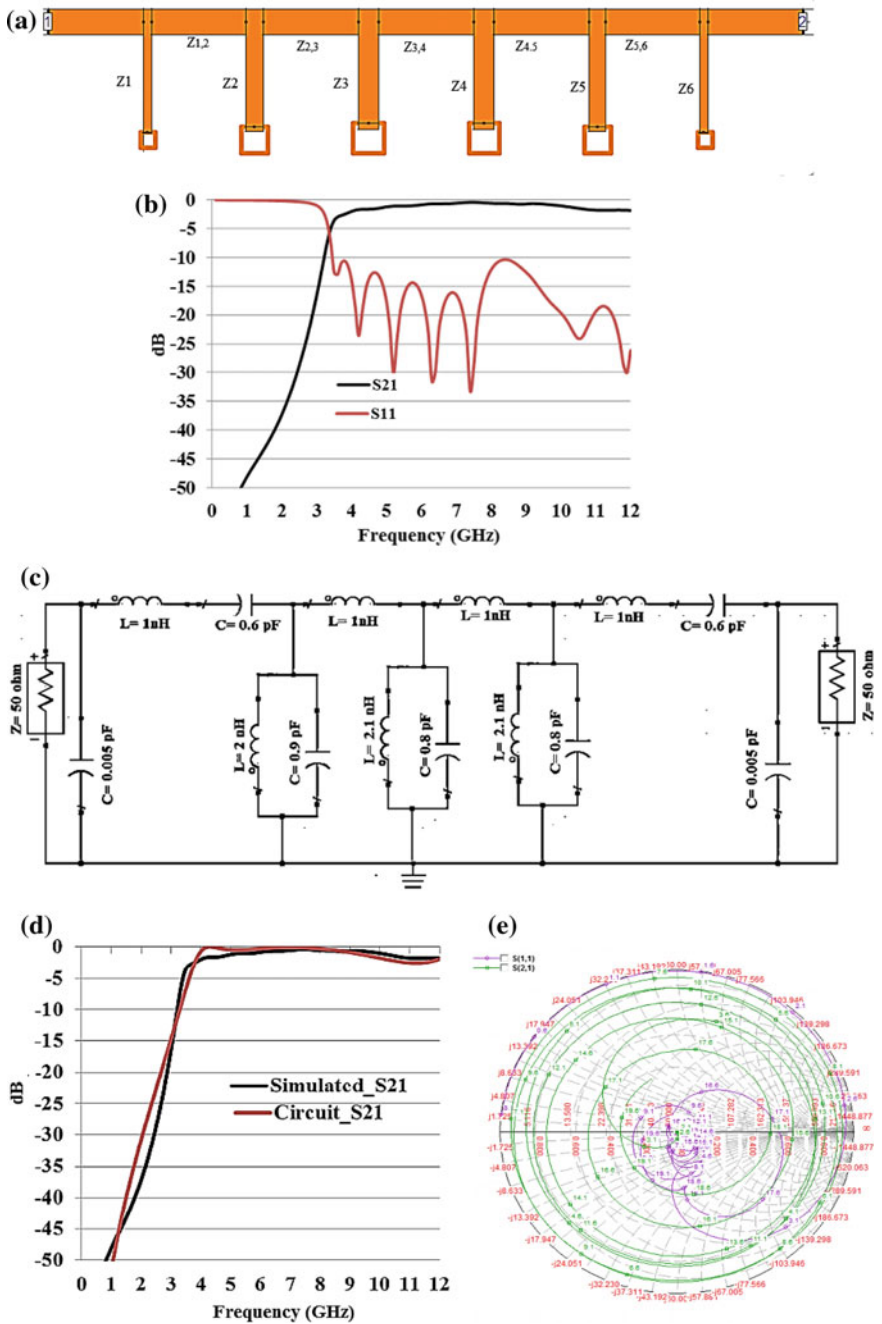
### 2.2 Overview of Defected Ground Structure (DGS)

As the name suggests, DGS is a ‘defect’ in the ground plane. The ground plane of any transmission line (stripline/microstrip/CPW) is modified to produce a better response. Nowadays, defected ground structures are preferred in many structures to enhance the performance [13, 14]. The elements designed in the ground plane are kept directly under the transmission line. It is also adjusted for efficient coupling. There are different shapes of defected ground structures available. Each of them produces a different response. According to the requirement, the structures are applied. The equivalent circuit diagram is shown in Fig. 3.

### 2.3 Lowpass Filter Using DGS Technique

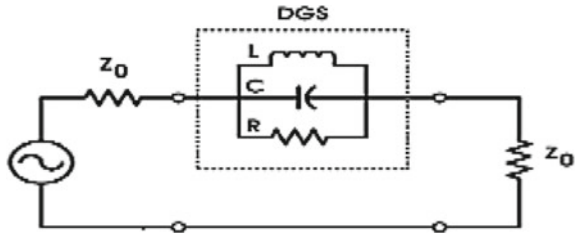
As we know, single-element DGS provides a gradual fall in 3 dB. So, an array of elements is preferred for sharpness and better selectivity. Among the various numbers of shapes, ‘H’ shape is preferred in this work as it meets the desired requirements. The structure and the response are shown in Fig. 4.





**Fig. 2** a Optimum distributed highpass filter, b simulated output response, c equivalent circuit of the HPF, d response of simulated  $S_{21}$  versus circuit  $S_{21}$ , and e frequency response in Smith chart of HPF

**Fig. 3** An equivalent DGS element circuit



It is clearly seen that the LPF is designed in the ground plane without disturbing the signal plane. Hence, the desired structure becomes easier to implement. The LPF response has the cutoff frequency of 11.2 GHz. The roll-off factor obtained after parametric analysis is 25 dB/GHz for  $S = 4$ . The Smith chart shows the location of the zeros and poles ( $S_{21}$  and  $S_{11}$ ).

### 3 Realization of Ultra-wideband Bandpass Filter

The conventional BPF filter structure is obtained by cascading the optimum distributed highpass filter and the DGS-based lowpass filter. To satisfy the specifications of FCC, the proposed filter is designed to achieve better selectivity at its edges of passband. The roll-off obtained at the upper cutoff is 20 dB/GHz, and the lower cutoff frequency is 29 dB/GHz approximately. Since the connecting lines of the optimum HPF have similar impedance values, the microstrip structure is slightly modified to have the same physical width and it is typically 3 mm for the FR4 substrate. The simple and compact BPF structure with its response is shown in Fig. 5.

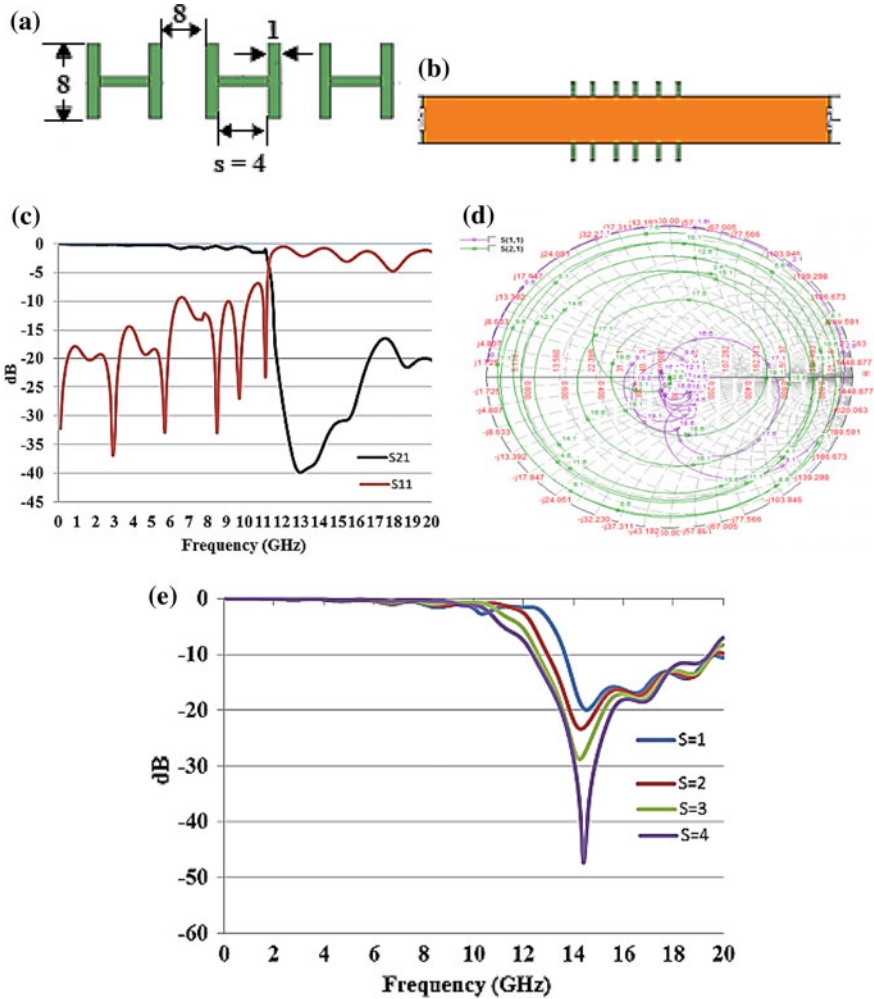
The 3D structure has been shown to demonstrate the clear connection between the signal plane and the ground plane.

#### 3.1 Group Delay

Group delay is the time delay of the passband envelopes. It is almost proportional to the order of the filter. A filter that produces flattest group delay in the passband is more acceptable. Since Chebyshev bandpass filter produces equal ripple in the passband, group delay is not very flat. The group delay can be calculated as

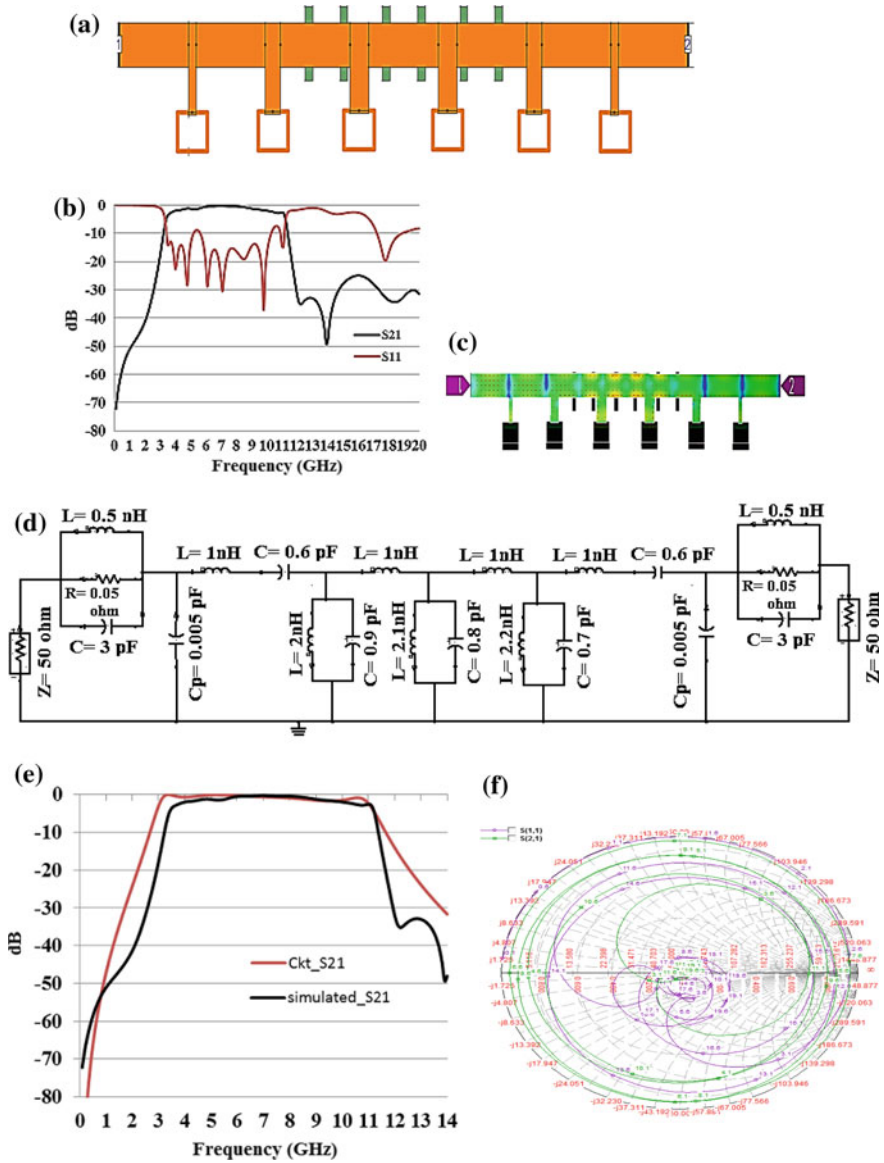
$$\text{Group delay} = -\frac{\Delta\phi}{\Delta\omega} \quad \text{and} \quad \omega = 2\pi f,$$

where ‘ $\phi$ ’ is phase angle of  $S_{21}$  and ‘ $\omega$ ’ is the angular frequency whose unit is rad/s. The graphs representing group delay of the each filter (HPF, LPF, and BPF) are shown in Fig. 6.



**Fig. 4** **a** ‘H’-shaped DGS, **b** LPF using the array of DGS, **c** simulated response of the lowpass filter, **d** frequency response in Smith chart of a microstrip LPF, and **e** variation of length of ‘S’ in the DGS array

If the group delay in passband of each filter is almost flat, it makes the filters more accurate. The calculated average group delay of HPF is 0.44 ns, and for LPF, it is 0.5 ns.



**Fig. 5** a Cascaded microstrip-based bandpass filter, b frequency response of the BPF, c current distribution shown in the 3D structure, d lumped equivalent circuit of the BPF, e response graph of circuit S21 versus simulated S21, and f frequency response in Smith chart of the BPF

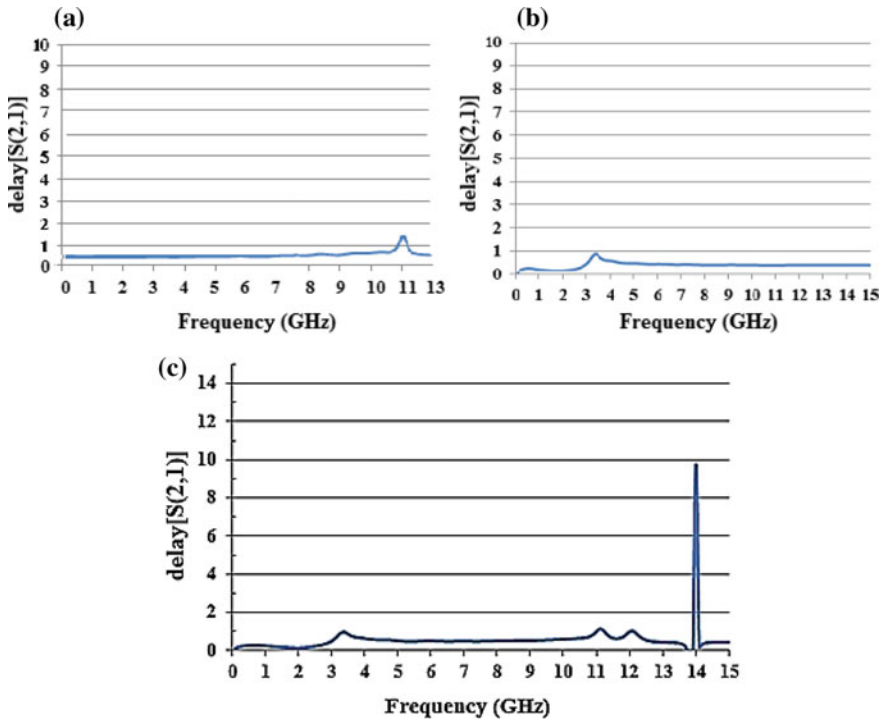


Fig. 6 Group delay of a DGS-based LPF, b highpass filter, and c bandpass filter

### 3.2 Obtained Parameters

This section includes the key parameters that have been obtained from the proposed filter. The performance characteristics are summarized in Table 3. Since  $\frac{f_2}{f_1} > 1$ , center frequency is calculated using  $\sqrt{f_1 * f_2}$  formula, where ‘ $f_1$ ’ is referred to the lower cutoff and ‘ $f_2$ ’ to the upper cutoff frequency.

Table 3 Key parameters of the microstrip bandpass filter

Sl. no.	Parameters	Values
1	Fractional bandwidth	123%
2	Passband insertion loss	Less than -2 dB
3	Average rejection level	-25 dB
4	Group delay [S(2,1)]	0.57 ns
5	Center frequency	6.2 GHz
6	Shape factor	4.7
7	Roll-off factor	20, 29 dB/GHz

## 4 Conclusion

A highly selective, microstrip-based ultra-wideband bandpass filter is proposed in this paper. This design employs six short-circuited stubs and an array of ‘H’-shaped defected ground structure. The cascaded structure provides the desired bandpass response that ranges from 3.5 to 11.11 GHz with FBW of 104%. The simulated result shows good roll-off factor, wide attenuation band, and almost flattest group delay. In addition to that, the equivalent circuits of each filter have been proposed. Because of its simplified structure and compact size, it can be widely used in microwave and wireless communication systems.

## References

1. Revision of Part 15 of the Commission’s Rules Regarding Ultra-Wide-band Transmission System. ET-Docket 98-153, First note and order, Federal Communication Commission, 14 Feb 2002
2. Kumar, M., Basu, A., Koul, S.K.: Active UWB antenna. In: Proceedings of URSI International Symposium on Electromagnetic Theory (EMTS), pp. 497–500. IEEE (2010). 978-1-4244-5153-1/10/\$26.00
3. Vanci, J., Sokol, V., Cerny, P., Skvo, Z.: The UWB amplifier 3.1–10.6 GHz. In: Proceedings of the 14th Conference on Microwave Techniques (COMITE), Prague, April 2008
4. Safarian, A., Zhou, L., Heydari, P.: CMOS distributed active power combiners and splitters for multi-antenna UWB beamforming transceivers. *IEEE J. Solid-State Circuits* **42**(7), 1481–1491 (2007)
5. Ishida, H., Araki, K.: Design and analysis of UWB band pass filter with ring filter. *IEEE MTT-S International Microwave Symposium Digest*, 0-7803-8331-1/04/\$20.00, pp. 1307–1310, June 2004
6. Chin, K., Lin, L., Kuo, J.: New formulas for synthesizing microstrip bandpass filters with relatively wide bandwidths. *IEEE Microw. Guided Wave Lett.* **14**(5), 231–233 (2004)
7. Li, K., Kurita, D., Matsui, T.: An ultra-wideband bandpass filter using broadside-coupled microstrip-coplanar waveguide structure. *IEEE MTT-S International Microwave Symposium Digest*, 0-7803-8846-1/05/\$20.00, pp. 675–678, June 2005
8. Srivastava, R., Kumar Pandey, A., Chauhan, R.K.: Design of Ultra-Wideband Filter with Reconfigurable Notches. *IEEE* (2016). <https://doi.org/10.1109/etct.7882999>
9. Pozar, D.M.: *Microwave Engineering*, 2nd edn. Wiley, New York (1998)
10. Orellana, M., Selga, J., Sans, M., Rodríguez, A., Boria, V.E.: Design of capacitively loaded coupled-line bandpass filters with compact size and spurious suppression. *IEEE Trans. Microw. Theory Tech.* **65**(4) (2017)
11. Saito, H.H., Nishikata, A.: Development of band pass filter for ultra wideband (UWB) communication systems. In: Proceedings of the IEEE Conference on Ultra Wideband Systems and Technology, 0-7803-8187-4/03/\$17.00, pp. 76–80 (2003)
12. Hong, J.-S., Lancaster, M.J.: *Microstrip Filters for RF/Microwave Applications*. Wiley (2001). ISBNs: 0-471-38877-7 (Hardback); 0-471-22161-9 (Electronic)
13. Weng, L.H., Guo, Y.C., Shi, X.W., Chen, X.Q.: An overview on defected ground structure. *Prog. Electromagn. Res. B* **7**, 173–189 (2008)
14. Jiayuan, L., Wang, J., Hui, G.: Design of compact balanced ultra-wideband bandpass filter with half mode dumbbell DGS. *Electron. Lett.* **52**(9), 731–732 (2016)

# Performance Evaluation of Wireless Sensor Network in the Presence of Wormhole Attack



Manish Patel, Akshai Aggarwal and Nirbhay Chaubey

**Abstract** Sensor nodes are densely deployed in the sensor field and are remotely monitored. The communication between nodes is also unreliable. An attacker can easily launch an attack. Possible attacks on sensor networks include selective forwarding, wormhole, black hole, sinkhole, denial of service, jellyfish. Wormhole is the gateway of all these attacks. After launching the wormhole attack, an attacker can launch any of these attacks. We have measured the impact of wormhole attack in AODV protocol in wireless sensor network. Also we have discussed some countermeasures against wormhole attacks and future research issues.

**Keywords** Wormhole · Security · Hostile · Countermeasure · AODV

## 1 Introduction

Wireless sensor network consists of densely deployed sensor nodes. Major components of sensor nodes are battery power, processor, analog-to-digital converter and transceiver. Major applications of sensor network include military applications, environment applications and health-related applications.

Security is very crucial issue for wireless sensor networks because of their fundamental characteristics. Sensor nodes are vulnerable to security attacks such as selective forwarding, wormhole, sinkhole, black hole, denial of service. Among all attacks, detecting wormhole attack is very hard [1–4]. After launching the wormhole, an attacker can launch many more attacks. In this paper, wormhole attack is simulated and performance of wireless sensor network is measured in the presence

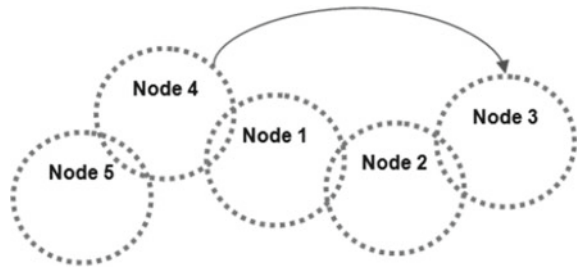
---

M. Patel (✉) · A. Aggarwal · N. Chaubey  
Smt. S. R. Patel Engineering College, Gujarat Technological University, Ahmedabad, India  
e-mail: it43manish@gmail.com

A. Aggarwal  
e-mail: akshai.aggarwal@gmail.com

N. Chaubey  
e-mail: nirbhay@ieee.org

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_48](https://doi.org/10.1007/978-981-13-1708-8_48)

**Fig. 1** AODV protocol

of an attacker. We have simulated the effect of wormhole attack in wireless sensor network. Our simulation results show that in the presence of multiple wormhole attackers, packet delivery ratio and throughput sharply decrease.

In Sect. 2, we have discussed the functionality of AODV (ad hoc on demand distance vector routing) protocol. Section 3 presents the description of wormhole attack. Section 4 presents simulation results. Section 5 presents conclusions and future research issues.

## 2 AODV Protocol

In AODV protocol path is established on demand. To establish the path, a source node broadcasts the route request packet. All the neighbor nodes forward the route request packet to their neighbors and finally RREQ packet is received by destination node. After that destination node sends RREP (route reply) packet on the reverse path and the path is established from source to the destination. All the data packets are transferred from source to the destination via this path.

As shown in Fig. 1, node 4 is the source node and node 3 is the destination node. Node 4 broadcasts RREQ packet. It is received by node 1 and node 5. There is no path from node 5, so the packet is dropped. Node 1 broadcasts to node 2, and packet reaches to node 3. When node 3 receives the RREQ packet, it sends RREP packet to node 4. After establishing the path, the data transfer occurs between source and destination.

## 3 Wormhole Attack Mechanism

To launch the wormhole attack, attacker needs two malicious nodes and both the nodes are located in different areas. These malicious nodes are hidden. They are connected through high-speed low-latency tunnel. One malicious node captures the packets from one area and tunnels to another malicious node in another area. Nodes located in one area falsely conclude that the nodes located in another area are their



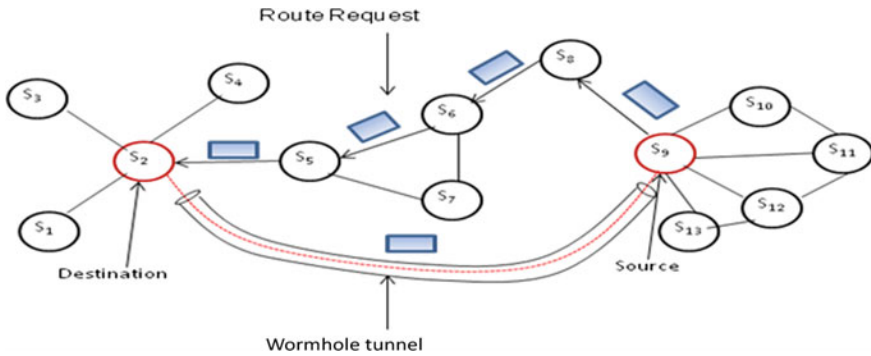


Fig. 2 Wormhole attack against on demand routing protocol

one hop neighbors and vice versa. Detecting wormhole attack is very hard. To launch the attack, an attacker does not need to know the preloaded secret material in the sensor node.

As shown in Fig. 2, node s9 broadcasts RREQ packet to establish path to node s2. The route request packet is captured by one malicious node and forwards to another malicious node at the other end of the network. Due to the tunnel, node s9 and s2 are one hop neighbors. The normal route request follows the path from s9-s8-s6-s5-s2. Due to the tunnel, RREQ packet reaches faster compared to the normal route. Node s2 sends RREP (route reply) packet on the reverse path, and the path is established between node s9 and s2. All the data packets are transferred from node s9 to s2 following the path through the tunnel. Malicious node can analyze the traffic, drop, delay and modify the packets.

### 4 Simulation Results

For simulation of wormhole attack, we have used NS2. Our results are based on the simulation of 10, 20 and 30 sensor nodes. The nodes are deployed in 500 \* 500 m area. We have measured packet delivery ratio and throughput for varying number of wormhole links. Packet delivery ratio is the ratio between the number of packets received and the number of the packets sent. Per unit of time the number of data packets sent from source to destination is throughput. The results after creating one and two wormhole links are given in Table 1.

**Table 1** Result with one and two wormhole links

No. of nodes	Without attack		With one wormhole link		With two wormhole links	
	PDF	Throughput	PDF	Throughput	PDF	Throughput
10	99.65	83.80	76.20	69.40	61.15	56.70
20	99.65	83.80	76.25	69.45	61.15	56.75
30	99.70	83.90	76.25	69.45	61.20	56.75

## 5 Countermeasures Against Wormhole Attack

Distance-based techniques are presented in [1–5]. Wormhole path contains less number of hop counts. To detect wormhole attack, location-based approaches [6–10] are used in wireless sensor network. It requires GPS or directional antenna which increases the network cost. Some techniques [11, 12] are based on maintaining neighbor information to detect wormhole attack. For dense network, it requires additional storage and processing power for storing and analyzing neighbor information. Wormhole route has greater than average time per hop compared to a normal route [13–15].

## 6 Conclusion

Wormhole attack is a serious threat against wormhole attack. We have analyzed the effect of wormhole attack in wireless sensor network. Network performance sharply decreases in the presence of multiple wormhole attackers. We have also discussed some existing countermeasures against wormhole attack. One challenging research area is detecting multiple wormhole attacks simultaneously. Another challenging research area is detecting wormhole attack in mobile wireless sensor network.

**Acknowledgements** The authors are highly thankful to Smt. S. R. Patel Engineering College, Gujarat Technological University, for providing the opportunity to conduct this research work.

## References

1. Honglong, C., Lou, W., Sun, X., Wang, Z.: (2010) A secure localization approach against wormhole attacks using distance consistency. *EURASIP J. Wirel. Commun. Netw.* **2010**, 11
2. Shokri, R., Poturlski, M.: A practical secure neighbor verification protocol for wireless sensor networks. In: *WiSec'09*, 16–18 Mar 2009. ACM, Zurich, Switzerland (2009)
3. Xu, Y., Ouyang, Y., Le, Z., Ford, J., Makedon, F.: Analysis of range-free anchor-free localization in a WSN under wormhole attack. In: *MSWiM'07*, 22–26 Oct 2007. ACM, Chaina, Greece
4. Sookhak, M., Akhundzada, A., Sookhak, A., Eslaminejad, M., Gani, A., Khan, M.K., Li, X., Wang, X.: Geographic wormhole detection in wireless sensor networks. *J. PLOS ONE*, 20 Jan 2015. <https://doi.org/10.1371/journal.pone.0115324>

5. Lai, G.H.: Detection of wormhole attacks on IPv6 mobility-based wireless sensor network. *EURASIP J. Wirel. Commun. Netw.* (2016)
6. Hu, Y.-C., Perrig, A., Johnson, D.B.: Wormhole attacks in wireless networks. *IEEE J. Sel. Areas Commun.* **24**(2), 370–380 (2006)
7. Poovendran, R., Lazos, L.: A graph theoretic framework for preventing the wormhole attack in wireless ad hoc networks. *Wirel. Netw.* **13**, 27–59. Springer (2007)
8. Khalil, I., Bagchi, S., Shroff, N.B.: MOBIWORP: mitigation of the wormhole attack in mobile multihop wireless networks. *J. Ad Hoc Netw.* **6**, 344–362. Elsevier (2008)
9. Hu, L., Evans, D.: Using directional antennas to prevent wormhole attacks. In: *Network and Distributed System Security Symposium (NDSS)*, pp. 131–141 (2004)
10. Lu, X., Dong, D., Liao, X.: MDS-based wormhole detection using local topology in wireless sensor networks. *Int. J. Distrib. Sens. Netw.* **2012**, 9, Article ID. 145702
11. Zhang, Y., Liu, W., Lou, W., Fang, Y.: Location-based compromise—tolerant security mechanisms for wireless sensor networks. *IEEE J. Sel. Areas Commun.* **24**(2) (2006)
12. Singh, R., Singh, J., Singh, R.: WRHT: a hybrid technique for detection of wormhole attack in wireless sensor networks. *J. Mob. Inf. Syst.* **2016**, 13. Hindawi Publishing Corporation, Article ID 8354930
13. Khabbaziyan, M., Mercier, H., Bhargava, V.K.: Severity analysis and countermeasure for the wormhole attack in wireless ad hoc networks. *IEEE Trans. Wirel. Commun.* **8**(2), 736–745 (2009)
14. Qazi, S., Raad, R., Mu, Y., Susilo, W.: Securing DSR against wormhole attacks in multirate ad hoc networks. *J. Netw. Comput. Appl.* 582–593 (2013)
15. Mukherjee, S., Chattopadhyay, M., Chattopadhyay, S., Kar, P.: Wormhole detection based on ordinal MDS using RTT in wireless sensor network. *J. Comput. Netw. Commun.* **2016**, 15, Article ID 3405264
16. Buttyan, L., Dora, L., Vajda, I.: Statistical wormhole detection in sensor networks. In: *SAS*, pp. 128–141. Springer (2005)

**Part V**  
**Social Networks**  
**and Sentiment Analysis**

# Fused Sentiments from Social Media and Its Relationship with Consumer Demand



Pushkal Agarwal, Shubham Upadhyaya, Aditya Kesharwani  
and Kannan Balaji

**Abstract** Social media is an ideal platform for influencers to share their experiences and product sentiments, as consumers frequently trust the recommendations of their peers. Consumer-created reviews and ratings are the preferred source of information about product and service value, price, and product quality. Social media mining is the process of storing, analyzing, and extracting useful patterns from social media data. Mining social information helps businesses in understanding the demand of their products like cars, movies, fashion goods, electronics, and so on. In this research, we present an analytical model which quantifies and engenders the insight of fused social information, gathered from multiple data sources. Subsequently, we discuss the results obtained by the proposed model for some movie use cases like “Angry Birds” by sourcing data from Twitter, Rotten Tomatoes, and the Internet Movie Database (IMDB).

**Keywords** Social media analytics · Sentiment analysis · Fusion

---

P. Agarwal · S. Upadhyaya (✉)  
The Nielsen Company, SLN Terminus, Gachibowli, Hyderabad 500032, Telangana, India  
e-mail: shubham.upadhyaya@nielsen.com

P. Agarwal  
e-mail: pushkalagarwal@gmail.com

A. Kesharwani  
Goldman Sachs Services Pvt. Ltd., Crystal Downs, Embassy Golf Links  
Business Park, Bengaluru 560071, Karnataka, India  
e-mail: aditya.kesharwani@ny.email.gs.com

K. Balaji  
The Nielsen Company, 501 Brooker Creek Blvd, Oldsmar, FL 34677, USA  
e-mail: balaji.kannan@nielsen.com

# 1 Introduction

According to Merriam-Webster dictionary [1], social media can be defined as forms of electronic communication through which users create online communities to share ideas, personal messages, information, and other contents. Some also consider it as the use of technology combined with social interaction to create or co-create values [2]. The recent growth of social media networks like Facebook, YouTube, and Twitter has significantly changed the nature of communication from unidirectional to bidirectional, not only between the firms and the organizations, but also among the consumers [3]. Thus, it is conspicuous that we are surrounded by social media from all the sides.

Social media does affect the sales of various products like cars, movies, fashion goods, electronic devices, etc. [4]. Hence, increase in the social media usage has massively increased the amount of data accessible for research and analyzing consumer behavior which we refer as social media mining [2] and can also be described as the process of representing, analyzing, and extracting useful patterns from social media [5]. In this paper, we will compare the reviews for movie from different social media platforms with its chronological sales and will be talking in detail about the improvement led by the fusion of reviews from these platforms. Thereafter, the final results will demonstrate the leading source of movie reviews for a time slice (here on daily basis) with respect to the actual sales observed.

Taking the movie data into consideration, firstly, we gathered the publicly available tweets data with respect to “#AngryBirdsMovie” using the “twitteR” [6] library defined for R. R is a GNU project specifically optimized for statistics and interactive graphics [7]. Secondly, to aid further, we took data from two of the most sought-after movie review platforms—Rotten Tomatoes and IMDB (Internet Movie Database) [8] through scrapping. For the aforementioned task, we used the Selector Gadget tool, which helps in picking up the appropriate CSS selector [9] of a Web page content, say, reviews, rating, and time stamp. The R script enabled the usage of this tool by taking the whole text and comparing with passed CSS selector [9].

The Sentiment analysis [10, 11] approach that is used to extract and quantify the reviews’ text was also implemented in R via sentiment library [12]. Positive and negative scores are generated using “Naive Bayes Classifier” [13] and used for further formulations that will be discussed in later sections.

Finally, the tool which we used for visualization is “Shiny by RStudio” [14], applications of which can be deployed on Shiny Cloud [15] easily. Shiny also combines the computational power of R with the interactivity of the modern Web. Some other advantages of Shiny are seamless compatibility with R, support from RStudio community, and reactive input–output [16] to improve the performance of Shiny applications. Subsequently, we created some word clouds and relevant plots to gain more insights in what and where people are actually talking while giving reviews for movies.

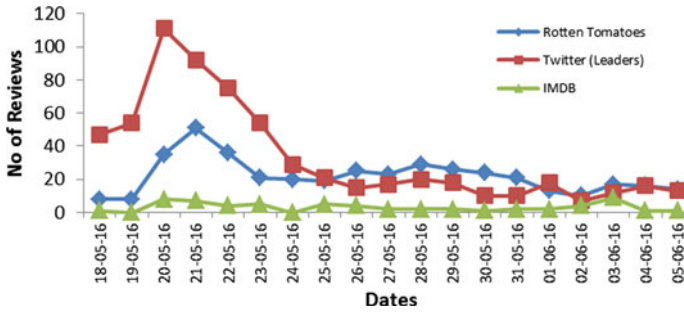


Fig. 1 Daily reviews count from sources

## 2 Proposed Approach

The following subsections provide the details of the steps needed to be followed for the fusion-based comparison, between social information and consumer demand. The subject of the research is “The Angry Birds Movie” released under the banner of “Sony Pictures” [17]. The consumer demand, i.e., the US box office revenue, has been taken from “The Numbers” [18]. Let us get a better insight of the same.

### 2.1 Data Gathering

As claimed by Economics Bulletin report [19], Twitter can be a contributor to movie success; therefore, one of the data sources was Twitter. The tweets were retrieved for US region (longitude: 37°05’24.0”N, latitude: 95°42’00.0”W, radius: 2000 miles), looking the map and coverage. After setting up the Twitter API [6], we searched the tweets related to #AngryBirdsMovie for the aforementioned market which have attributes like text, screenName, id, created, isRetweet, retweetCount, etc. [6]. As per Nielsen’s 2013 American Moviegoing report [20], 41% of millennials said, they refer the critic ratings from platforms such as Rotten Tomatoes and IMDB, before going to a movie. Figure 1 shows the number of reviews from Rotten Tomatoes and IMDB for the same movie. Data were scrapped from these two Web sites using Selector Gadget [9] tool. Reviews, creation dates, and ratings given by the users were extracted using their CSS selectors chosen by the Selector Gadget [9] tool. The resultant data were segregated on the basis of time slice, which was considered to be “one-day” (Fig. 1).

**Table 1** Source schema

Source	Attribute			
	A1	A2	A3	A4
Twitter	Positive scores	Negative scores	Retweet count score	Leaders tweets count
Rotten Tomatoes	Positive scores	Negative scores	Ratings sum	Reviews count
IMDB	Positive scores	Negative scores	Ratings Sum	Reviews count

## 2.2 Data Preprocessing

The Economics Bulletin report [19] claims that the diffused information from a popular personality (say star) has an immediate impact on people choices. So we introduce “Leaders” approach in the selecting Twitter data for generating results. Out of all the tweets, we selected a subset that is original (`IsRetweet == True`) and diffused (`RetweetCount > 0`). The following algorithm tells the selection of leaders’ tweets, given the extracted tweets as parameter:

```

Initialize leaders
Initialize leaderTweets
for each tweet in tweets
    if (tweet[isRetweet]=F AND tweet[retweetCount]>0)
        then
            append (leaders, tweet[screenName])
            append (leaderTweets, tweet[text])
        end
    end
end

```

Based on the above algorithm, leaders’ tweets are selected. The daily leaders’ tweets count is shown in Fig. 1. From 9,104 tweets spanned over 19 days, we got 639 only as leaders’ tweets. Once data are gathered and divided, they are preprocessed before they are analyzed further. Preprocessing includes the removal of numbers, punctuations, stopwords, links and other unwanted text, lowering case, and finally stemming of words. The corpus is now ready for sentiment analysis. Sentiment analysis [10] is the process of quantifying the text provided as input, in terms of positive and negative scores. These scores are used afterward to judge the overall sentiment of the writer of the text excerpt. The package which was used for the aforementioned purpose was “sentiment” [12], which is an R package with tools for sentiment analysis using Naïve Bayes Classifiers [13] for positivity/negativity and emotion classification. After data preprocessing, the data from all the sources were stored according to the schema in Table 1. These attributes were summed up for the considered time slice, i.e., one day.



### 2.3 Factor Formulation

We now try to establish a formula that incorporates the obtained positive and negative scores while taking total number of tweets also into consideration. Now, we establish a formula (Eqs. 1–3) for these three sources of data that best explains the coherence of the social media discussions with the sales of the movie on a daily basis. After the regular refinement of these data and doing the sentiment analysis, we obtained positive and negative scores. Now these scores, along with the number of reviews on a daily basis were put together combining the various aspects pertaining to each source of data. To underscore the importance of compatibility, a formula was developed for these three sources that best explained the coherence value with the sales data, trying out various possible combinations of the scores obtained. These coherence values are nothing but the Pearson correlation coefficient [21] of the sales and each of the individual forms of the data, and out of various combinations, the one that gave the maximum correlation has been considered the formula for that source of data. The established formula which gave factor value for these sources is given below:

- Rotten Tomatoes:

$$\text{Score}_{RT} = \left[ \frac{\text{Positive Score}}{\text{Negative Score}} * \text{Reviews Count} * \text{Ratings Sum} \right] * 10^{-3} \quad (1)$$

- Twitter (Leaders):

$$\text{Score}_{TW} = [(\text{Positive Score} - \text{Negative Score})] * \frac{10^{-3}}{14} \quad (2)$$

- IMDB:

$$\text{Score}_{IMDB} = \left[ \frac{(\text{Positive Score} - \text{Negative Score}) * \text{Reviews Count}}{* \text{Ratings Sum}} \right] * 10^{-3} \quad (3)$$

Here, RT stands for Rotten Tomatoes, TW for Twitter, IMDB for the Internet Movie Database. After obtaining the perfect formula that best explains the sales for these three sources, we scaled these scores with the sales value by multiplying with numerical constants to generate a plot to explain coherence visually.

## 2.4 Fusion

After defining factor formulae in the previous section, we further explored to fuse the contributions of these three sources. The fusion was carried out using the following two factors.

1. Correlation between the sales and scores for each source.
2. Fraction of daily number of reviews or tweets from each source.

Combination of these two factors gave rise to the following formulae:

$$\alpha_t = \left[ \text{Correlation}(\text{Score}_{RT}, \text{Sales}) * \frac{\text{reviewCount}_{RT}(t)}{(\text{reviewCount}_{RT+TW+IMDB}(t))} \right] \quad (4)$$

$$\beta_t = \left[ \text{Correlation}(\text{Score}_{TW}, \text{Sales}) * \frac{\text{reviewCount}_{TW}(t)}{(\text{reviewCount}_{RT+TW+IMDB}(t))} \right] \quad (5)$$

$$\gamma_t = \left[ \text{Correlation}(\text{Score}_{IM}, \text{Sales}) * \frac{\text{reviewCount}_{IMDB}(t)}{(\text{reviewCount}_{RT+TW+IMDB}(t))} \right] \quad (6)$$

Combining these, we formulated:

$$\text{Fusion\_score}_t = \alpha_t * \text{Score}_{RT} + \beta_t * \text{Score}_{TW} + \gamma_t * \text{Score}_{IMDB} \quad (7)$$

where  $\alpha_t$  is weight assigned to Rotten Tomatoes score, similarly  $\beta_t$  for Twitter score and  $\gamma_t$  for IMDB score and  $t$  for time slice, i.e., 1 day, where  $t$  varies from 1 to the last possible day.

Subsequently, we calculated  $\alpha$ ,  $\beta$ , and  $\gamma$  for each day using Eqs. (4)–(6).

These values were plugged into Eq. (7) to calculate ‘‘Fusion\_score.’’ Results of the same are discussed in the next section.

## 3 Discussion

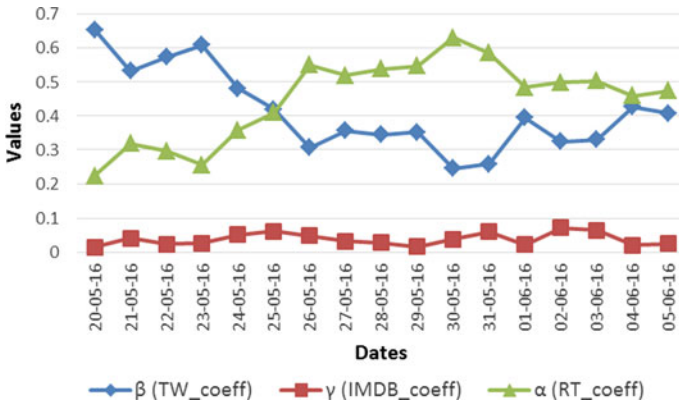
As discussed in Sect. 2.4, we calculated  $\alpha$ (RT\_coeff),  $\beta$ (TW\_coeff), and  $\gamma$ (IMDB\_coeff) for each day. For computation of these coefficients, correlation values of each data source with the sales of the movie were considered as shown in Table 2. One can easily infer from Table 2 that Rotten Tomatoes is the best data source when it comes to the reviews of the movie.

An interesting observation can be inferred from Fig. 2 that  $\alpha$  is initially lower than  $\beta$  due to high discussion for movie on Twitter in first weekend as compared to Rotten Tomatoes but as the time passes by,  $\alpha$  surpasses  $\beta$ . Hence, in the long run, Rotten Tomatoes shows better retention properties as compared to Twitter. Here,  $\gamma$  does not play any significant role in fusion score due to less number of reviews.

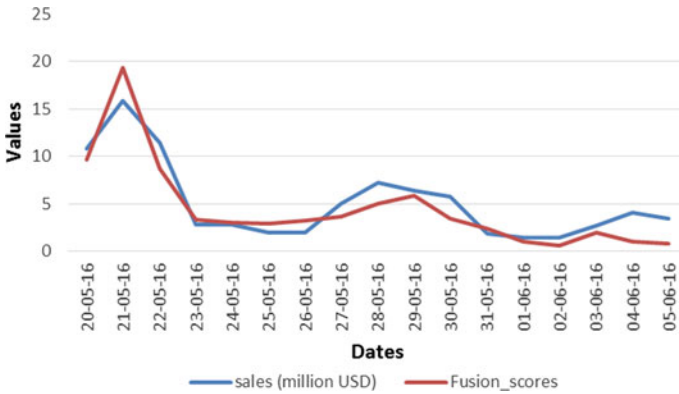
Finally, results which we got after merging scores of all data sources by plugging them into Eq. (7) along with the respective coefficients are shown in Fig. 3.

**Table 2** Correlation values of data source with the sales of movie

Data source	Correlation value
Twitter	0.88
Rotten Tomatoes	0.95
IMDB	0.68
Fusion (using Eq. (7))	0.94



**Fig. 2** Daily trend for  $\alpha$ ,  $\beta$ , and  $\gamma$



**Fig. 3** Daily trend for Fusion\_scores

From Fig. 3, it is clearly evident that fusion incorporates characteristics of all the data sources in the sense that Twitter results were good only in the first week and for Rotten Tomatoes results were better after the first week, which can be inferred from Fig. 2, whereas after fusion, results are coherent for all three weeks. Moreover, the aforementioned improvement by fusion is evident by the correlation values as shown in Table 2.

## 4 Conclusion

Social networks are gaining popularity to understand the interest of customers and target them based on their interest. Sentiments of customers have a high correlation with the sales of products. In this paper, we proposed a model to understand the sale of products as the sentiments of customers vary. The model is developed by combining data from multiple data sources. We analyzed the developed model using “Angry Birds Movie” data. In the analysis, a clear interdependence between movie sales and social media discussions about the movie was observed when the three sources of data were combined together, while with individual source, it explained some of the parts. The publicly available data were retrieved from Twitter, IMDB, and Rotten Tomatoes, which were analyzed with the sales of the “Angry Birds Movie” in the US market on a span of three weeks. The discussion about the movie starts before the release of the movie because of the excitement among the people. This movie was popular because of a mobile gaming application launched several years ago. So, with this analysis, we were able to relate the sentiments of the people with the sales of the movie.

## References

1. Dictionary, M.W.: Definition of Social Media by Merriam-Webster. <https://www.merriam-webster.com/dictionary/social%20media>
2. Buettner, R.: Predicting user behavior in electronic markets based on personality-mining in large online social networks. *Electr. Markets* 1–19 (2016)
3. Hennig-Thurau, T., Malhotra, E.C., Frieger, C., Gensler, S., Lobschat, L., Rangaswamy, A., Skiera, B.: The impact of new media on customer relationships. *J. Serv. Res.* **13**(3), 311–330 (2010)
4. Chevalier, J.A., Mayzlin, D.: The effect of word of mouth on sales: online book reviews. *J. Mark. Res.* **43**(3), 345–354 (2006)
5. Bali, A., Agarwal, P., Poddar, G., Harsole, D., Zaman, N.M.: Consumer’s sentiment analysis of popular phone brands and operating system preference. *Int. J. Comput. Appl.* **155**(4) (2016)
6. Gentry, J.: Repository for TwitteR Package in R. <https://CRAN.R-project.org/package=twitter> (2015)
7. R Development Core Team.: R: A Language and Environment for Statistical Computing. <http://www.R-project.org/> (2011)
8. Yang, Y.: Article on Movie Reviews on Freenuts. <http://freenuts.com/movie-reviews> (2010)
9. Wickham, H.: Repository for Selectorgadget in R. <https://cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html> (2016)
10. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: *Proceedings of the 2nd International Conference on Knowledge Capture*, pp. 70–77. ACM (2003)
11. Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A., Benevenuto, F.: Sentibench—a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci.* **5**(1), 1–29 (2016)
12. Jurka, T.P.: Repository for Sentiment Package in R. <https://cran.r-project.org/src/contrib/Archive/sentiment> (2012)
13. Rish, I.: An empirical study of the naive Bayes classifier. In: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, no. 22, pp. 41–46 (2001)

14. Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J.: Shiny: Web Application Framework for R. <http://CRAN.R-project.org/package=shiny> (2015)
15. Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J.: Shiny Cloud Official Website. <https://www.shinyapps.io/> (2015)
16. Server, R.S.: RStudio: Official Website. <http://shiny.rstudio.com/> (2015)
17. Pictures, S.: Angry Birds Movie. <http://www.sonypictures.com/movies/theangrybirdsmovie> (2016)
18. Numbers, T.: Angry Birds Collections from Numbers. <http://www.the-numbers.com/movie/Angry-Birds-Movie-The#tab=box-office> (2016)
19. Treme, J., VanDerPloeg, Z.: The twitter effect: social media usage as a contributor to movie success. *Econ. Bull.* **34**(2), 793–809 (2014)
20. Company, N.: #Twothumbsup: Moviegoing at a Theater Near You. <http://www.nielsen.com/us/en/insights/news/2014/twothumbsup-moviegoing-at-a-theater-near-you.html> (2014)
21. Rodgers J.L., Nicewander, W.A.: Thirteen ways to look at the correlation coefficient. *Am. Stat.* **42**(1), 59–66 (1988)

# A Prototype for Semantic Knowledge Retrieval from Educational Ontology Using RDF and SPARQL



S. Mahaboob Hussain, Prathyusha Kanakam and D. Suryanarayana

**Abstract** Nowadays, there is a huge amount of data available on the current Web and they are still growing, which raises a serious problem to obtain accurate search results since the Web offers unstructured textual data. The process of understanding the natural language query (NLQ) by a machine is a challenging task to retrieve the user query from the database, but in this scenario, users always presume that the process of information retrieval is a simple task. To really make it simple, semantic Web is introduced to certainly make the retrieval process simple with SPARQL from ontologies and Resource Description Framework (RDF). In fact, users fail to understand the syntax of the SPARQL to retrieve information from the semantic Web. Hence, this paper explores the process of transforming SPARQL to natural language to apply on semantic Web and illustrates the internal working of the system which reads the user-entered NLQ and the process of tokenization, pos tagging to the sentences by checking the grammar. Accordingly, technologies and data storage possibilities are analyzed and evaluated to retrieve accurate results. Overall, this paper illustrates semantic information retrieval using SPARQL from RDF knowledge base which helps to furnish appropriate information to the user.

**Keywords** Semantic web · Information retrieval · SPARQL  
RDF · OWL · NLP

---

S. Mahaboob Hussain (✉) · D. Suryanarayana  
Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India  
e-mail: mahaboobhussain.smh@gmail.com

D. Suryanarayana  
e-mail: suryanarayanadasika@gmail.com

P. Kanakam  
MVGR College of Engineering, Vizianagaram, Andhra Pradesh, India  
e-mail: prathyusha.kanakam@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_50](https://doi.org/10.1007/978-981-13-1708-8_50)

# 1 Introduction

Every system needs to answer the user query which will be in any logical or syntactical natural language. In practical, this system should understand the input and must convert in its own machine language to generate the reliable output. The major difficulty of the systems at the 1960s is responding the simple information retrieval-based queries and logical knowledge-based semantic information [1]. In most of the IR-based systems, information is acquired from the Web which is a rich source of documents from where the text is driven by the user-posed query. So, these type of text-based retrieval systems can be noted as a question-answering system.

Semantic Web implements this relational semantic information which can be efficiently recognized and read by the crawlers [2]. The functionality of the semantic Web comprises presenting accurate information, storing hierarchical databases in the distinct format and a query processing system can extract the triplets which are understandable by machines. Each user-posed query can be lemmatized to discover the root words in that query, and these certain words support a question-answering system to exact the answers for users' queries. The obtained answers are one-of-a-kind entities that consists of person, location and time details. Utmost of the intelligent machines can efficiently understand the theme of the query with the words in the sentence after ignoring the stop words and anticipate the answer [3].

This result will be placed in the same query by removing the interrogative words. Only the definition of the question is interpreted by some of the systems. For example, for the question: Who is the CEO of Google and who is his PA? The query processing will produce the results. *Answer Type*: CEO, PA of CEO. *Query*: CEO, PA, Google. The entire query process system is shown in Fig. 1.

Search engines and the crawlers play an essential role in providing accurate information, but it will be a critical task when the world of Web data increases timely.

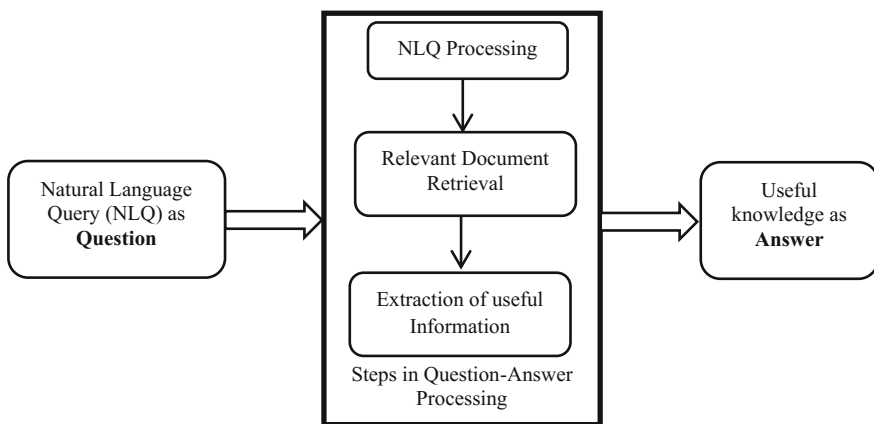


Fig. 1 Query processing system

Most of the users depend upon the Web for the information, and it will be searched and retrieved by the Google, Yahoo, Bing and other famous search engines. There is a concern about the performance and accuracy of these crawlers and no guarantee that these search engines will satisfy the users with their outcomes from large interconnected clusters throughout the world data centers [4].

Most of the organizations, demographic agencies, industries and Government sectors announce and publish their information on the Web for the easy access to their databases for the common people through the search engines. All the time, search engines provide results from the Web but may not be the accurate and predicted results. Hence, it's a Deep Web with the huge intensity of database where search engines sometimes unable to fetch the required data. However, this approach, even with search, can hardly be automated. Representing the data and creating query forms is a simple task using the HTML, but the background process, i.e., searching and retrieving the information from the Web. Moreover, predicting the intention of the user query is a difficult task.

## 2 Preliminary and Related Work

This paper proposes the key concept of understanding the cognition of a human by the machine by converting natural language into a machine understandable format, i.e., in SPARQL [5]. It will be processed on the ontology knowledge base to retrieve the information. Here, career ontology is used to explain the semantic Web process model as in Fig. 2. Every concept in the ontology is mapped by the keywords in the ontology, and the graph is used to retrieve available relations between these concepts. In the user-posted natural language query, certain relations are not stated expressly by the user and these are most likely named as the gaps in the query. The user knows the complexity, for example, searching for a postgraduation course and duration of engineering, the obvious relation would be “post graduation course”. To avoid inaccurate results in search time, RDF is used in the representation of triples such as subject, predicate, and object [6].

### 2.1 Resource Description Framework (RDF)

RDF is a data model which is used to represent the Web with the well-defined knowledge that is used to associate for RDF-based languages and specifications [7]. RDF graphs (subject–predicate–object) are triples and the elements are URIs and blank nodes which are the two descriptions of an RDF model.

For example, consider a sentence *Eamcet is the entrance exam for engineering*, and it should convert into triplets as shown in Fig. 3. From those triplets, SPARQL query is generated to give accurate results with the help of ontology and RDF knowl-



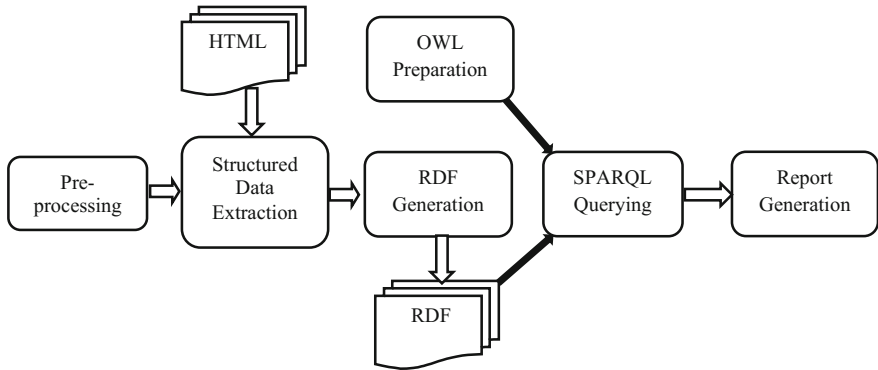


Fig. 2 Process of semantic web information retrieval

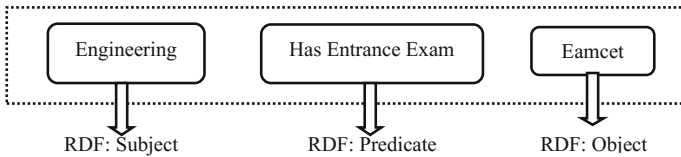


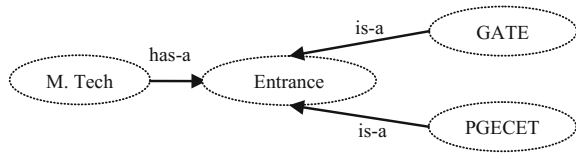
Fig. 3 Resource Description Framework representation with triplets

edge base. The complete set of a graph is to represent an RDF and consists of two nodes *subject* and *object* and a connecting node *predicate*.

## 2.2 Web Ontology Language (OWL)

Ontologies are the consistent and relational knowledge for representing the data with a Web ontology language. Ontologies are the tremendous repository of the hierarchical relation data which illustrates the relationships among the classes [8]. This collection of data defines the knowledge for various domains. Here, the nouns and verbs represent the classes and the relations between the objects as shown in Fig. 4. Web ontology language (OWL) represents a domain with classes and its properties which further incorporate the generous description of the objects and the properties of Web resources described by these ontologies.

**Fig. 4** Ontological relationships for a sentence in OWL representation



### 2.3 SPARQL

SPARQL is pronounced as sparkle and its acronym stands for SPARQL Protocol and it is a special language to perform queries on the databases to manage the data which are stored in RDF format [5]. In this paper, SPARQL is used to retrieve the information which is available in the RDF format from the semantic Web that provides specific information to the user. Authors illustrated the SPARQL with an instance that shows an SPARQL query to obtain all the courses after engineering from the information in the given RDF graph. The SPARQL query consists several keywords in which SELECT clause and the WHERE clause are applied. By using a SELECT and WHERE clause in the SPARQL query, it results a triplet pattern from a class.

## 3 Working Methodology

Authors worked on a strategy for parsing the sentences that extract all methods' signatures and tag each of the words that are given. Then, identification of each word individually in every method is required. Parsing method technique divided into two main phases: POS of the each word is tagged with an identifier and the second one is using content to choose a particular part-of-speech and the identifier will be pieced into its lexical segments. A natural language query of the user is admitted to the proposed search system, and then tagging will be done to the words available in the question and it will be parsed. Then, the stop word algorithm is applied; thus, it will generate triplets as subject, object and predicate from the natural language question. An SPARQL query will be generated from the obtained results of triplets. Once the process is performed, then the semantics of the input question is achieved; therefore, it can be transformed into an appropriate query language (SPARQL) and then executed against the knowledge base. The result of querying the knowledge base is a sub-graph of the ontology RDF graph. This means that the required information is stored in the triplet form, and it should be transformed into a more human-readable output. However, this step is system specific and depends on the desired usability of a particular system. The answer can vary from simple triplet representation to a full-sentence answer. Consider an experimental query on the well-known search engine. *What are the courses and jobs after BTech?* Figure 5 shows the experimental results of the proposed system.

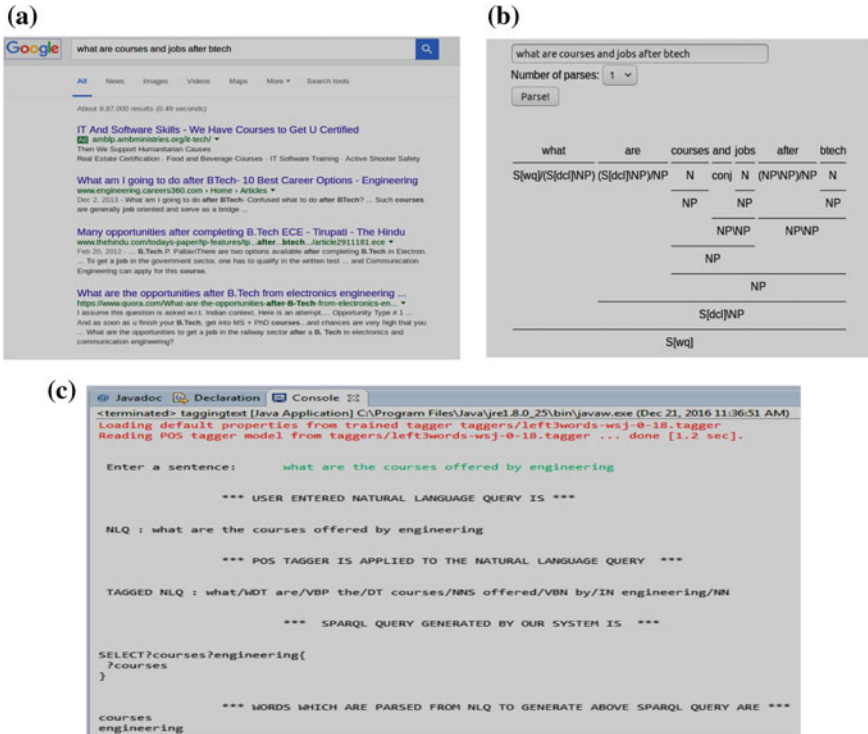


Fig. 5 a Result from Google search engine for a natural language query, b process of semantic Web information retrieval, c conversion of NL query to SPARQL

A natural language query of the user is admitted to the proposed search system and then tagging will be done to the words available in the question and it will be parsed. Further, the stop word algorithm is applied; thus, it will generate triplets as subject, object and predicate from the natural language question. An SPARQL query will be generated from the obtained results of triplets. Once the process is performed, then the semantics of the input question is achieved; therefore, it can be transformed into an appropriate query language (SPARQL) and then executed against the knowledge base. The result of querying the knowledge base is a sub-graph of the ontology RDF graph. This means that the required information is stored in the triplet form, and it should be transformed to a more human-readable output. However, this step is system specific and depends on the desired usability of a particular system. The answer can vary from simple triplet representation to a full-sentence answer. Consider this query on the well-known search engine. *What are the courses and jobs after BTech?* The results we are getting from the Google search engine are shown in Fig. 5.

The given natural language query undergone several processes and obtain SPARQL query which is generated by using triplets identified by using POS tagging

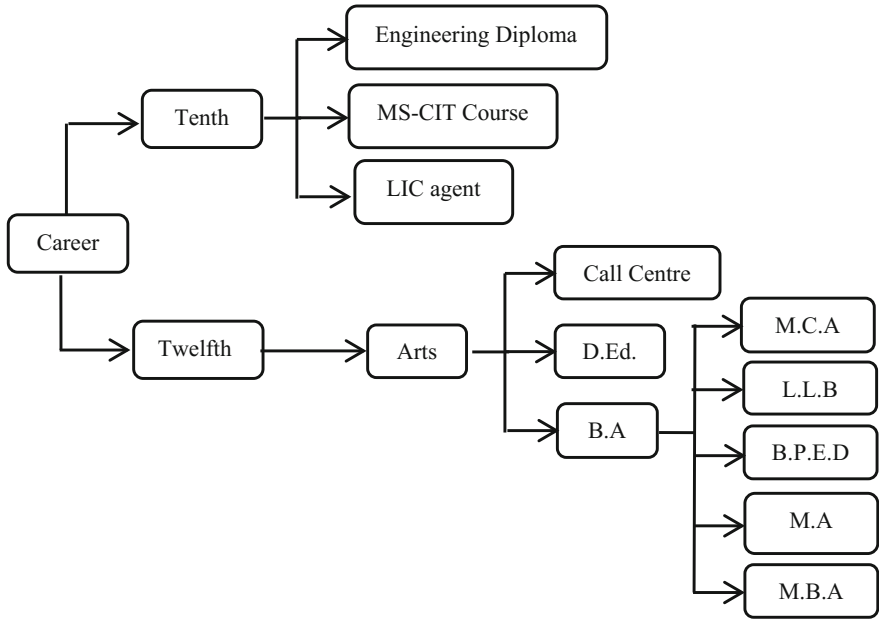
[9]. Ontology is the main aspect of semantic web and it is the relational database to retrieve semantic results. Preferably than retrieving the information ordinarily a semantic information retrieval is worthy for accurate and predicted data. Therefore, the optimal solution is to develop an ontology with RDF which consists of relational semantic triples as subject, object, and predicate written in RDF/XML. To retrieve the information from this constructed semantic database, an SPARQL query language is used. For instance, an RDF model is represented.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
<!ENTITY dbp "http://dbpedia.org/property/" >
<!ENTITY dbpedia "http://dbpedia.org/resource/" >
<!ENTITY dbo "http://dbpedia.org/ontology/" >
<!ENTITY owl "http://www.w3.org/2002/07/owl#" >
<!ENTITY tto "http://example.org/tuto/ontology#" >
<!ENTITY ttr "http://example.org/tuto/resource#" >
<!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
]>
<owl:Class rdf:about="&ont;B.A">
  <rdfs:subClassOf rdf:resource="&ont;arts"/>
  <tto:duration
rdf:datatype="&xsd;decimal">3.0</tto:duration>

<ont:postgraduationcourse>B.P.E.D</ont:postgraduationcourse>
  <tto:PGcourses>B.P.Ed.</tto:PGcourses>
  <tto:PGcourses rdf:resource="&ont;M.A."/>
  <tto:PGcourses rdf:resource="&ont;M.C.M"/>
  <tto:job>Digital Marketing</tto:job>
</owl:Class>
<owl:Class rdf:about="&ont;L.L.BFoundation">
  <rdfs:subClassOf rdf:resource="&ont;arts"/>
  <tto:duration
rdf:datatype="&xsd;decimal">2.0</tto:duration>
  <tto:duration
rdf:datatype="&xsd;decimal">5.0</tto:duration>
  <tto:PGcourses>L.L.M.</tto:PGcourses>
  <tto:job>Lawyer, Judiciary Editing Law
books</tto:job>
</owl:Class>
<owl:Class>
```

Figure 6 represents the ontology for the RDF written above; hence, it shows the hierarchical architecture for the student career after tenth and twelfth.

Now, the converted natural language query (looking for courses, duration, jobs) into SPARQL will be applied on the ontology [10] as shown below and the results after the process of retrieval are framed as shown in Table 1.



**Fig. 6** Hierarchical representation of ontology created using RDF

**Table 1** Resultant information for the SPARQL query in triplet form

Subject	Predicate	Object
Chemical engineering	Jobs	Scientist, professor, trainee software, sales manager
Civil engineering	Jobs	Site engineer, project engineer
Computer science and engineering/CSE	Jobs	Software engineer, system analyst, assistant professor, software tester, bank jobs
Electronics communication engineering	Jobs	Hardware networking engineer, telecom engineer graphic designer
Electrical electronics engineering	Jobs	Development engineer, embedded trainer
Mechanical engineering	Jobs	Space instrument engineer, machine design engineer

```

Prefix dbo: <http://dbpedia.org/ontology/>
Prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix tto: <http://example.org/tuto/ontology#>
Prefix ttr: <http://example.org/tuto/resource#>
SELECT ?courses ?Jobs where {
    ?thing rdf:type dbo:B.Tech.
    ?thing ont:courses ?courses.
    ?thing tto:Jobs ?Jobs.
}

```

## 4 Conclusion

This paper focuses on how the future Web 3.0 might look like and the interaction between the human and WWW to give accurate results for any natural language query. Transforming from the current to future Web, definitely it provides the knowledge rather than results for the users which consists of structured data and knowledge representation. However, the most important role of the information retrieval system is to retrieve the ontology from various domains to obtain the user predicted results out of the semantic Web. Here, with example ontology, the work was determined and results were analyzed. The process of removing the unnecessary words from the query and formulating to estimate the related set of triplets make the search efficient. What next? is the main concern of the students to forward in their education path for their future according to their interest. With this semantic search facility, it will be more illuminating for the students in education domain by getting the accurate information rather than getting more links and unrelated information. This paper shows the transformation of natural language query to SPARQL to provide accurate results for given user query, and the results will be satisfactory needs for users. Therefore, the semantic search engine will produce more precise results compared to the traditional search engine.

## References

1. Lewandowski, D.: Evaluating the retrieval effectiveness of web search engines using a representative query sample. *J. Assoc. Inf. Sci. Technol.* **66**, 1763–1775 (2015)
2. AbdulNabi, S.: Effective performance of information retrieval on web by using web crawling. *Int. J. Web Semant. Technol.* **3**, 63–72 (2012)
3. Suryanarayana, D., Hussain, S., Kanakam, P., Gupta, S.: Stepping towards a semantic web search engine for accurate outcomes in favor of user queries: Using RDF and ontology technologies. In: 2015 IEEE international conference on computational intelligence and computing research, pp. 1–6 (2015)
4. Mahaboob Hussain, S., Surya Narayana, D., Kanakam, P., Gupta, S.: Palazzo matrix model: an approach to simulate the efficient semantic results in search engines. In: 2015 IEEE international conference on electrical, computer and communication technologies (2015)

5. Zhang, Z., Yang, T.: SPARQL ontology query based on natural language understanding. *J. Comput. Appl.* **30**, 3397–3400 (2011)
6. Neumann, T., Weikum, G.: RDF-Stores und RDF-Query-Engines. *Datenbank-Spektrum* **11**, 63–66 (2011)
7. Tripathi, A.: Resource description framework (RDF) for organised searching on internet. *DESI-DOC Bull. Inf. Technol.* **21**, 3–7 (2001)
8. Antoniou, G., Van Harmelen, F.: *Web Ontology Language: Owl. Handbook on Ontologies*, pp. 67–92. Springer, Berlin Heidelberg (2004)
9. Shaik, S., Kanakam, P., Hussain, S., Narayana, D.: Transforming natural language query to SPARQL for semantic information retrieval. *Int. J. Eng. Trends Technol.* **41**, 347–350 (2016)
10. Suryanarayana, D., Kanakam, P., Hussain, S.M., Gupta, S.: High-performance linguistics scheme for cognitive information processing. In: *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, pp. 369–378. Springer, Singapore (2018)

# Social Trust Analysis: How Your Behavior on the Web Determines Reliability of the Information You Generate?



Rhea Sanjay Sukthanker and K. Saravanakumar

**Abstract** Can I trust a review? A very common question for someone accustomed to online shopping. The Internet hosts a large number of reviews. Many e-commerce Web sites like Amazon, eBay, Flipkart ask their customers for their reviews once the product is bought. There is an important aspect of trust in an online context. Often reviews diverge widely on their star ratings from 1–5 which clearly show bias for a brand or product. What actually guarantees the reliability of a review? Some of the effective ways to ensure the trustworthiness of a review are to use the reviewer profile information and his previous reviews. Opinions of others have a greater impact on consumers rather than verified information provided by the product’s producer, thus, ensuring that misleading reviews do not creep in is a necessity. The goal of this work is to develop a trustworthy reviews model by taking into consideration all the factors which make a review reliable.

**Keywords** Social networks · Feature extraction · Unsupervised clustering  
Trustworthiness · E-commerce

## 1 Introduction

With the evolution of Internet technology at a very fast pace, information about any product we wish to buy or any movie we are planning to watch is right at our fingertips. Social media platforms like Facebook have more traffic than search engines like Google. Thus, it is evident that we are moving toward a world which revolves around social interactions and information exchange. But what guarantees that this information is correct, unbiased and trustworthy? The economist Kenneth Arrow defines trust as a “lubricant of the social system.” Social trust is the essence of

---

R. S. Sukthanker (✉) · K. Saravanakumar  
VIT University, Vellore, India  
e-mail: sukthankerrhea.sanjay2014@vit.ac.in

K. Saravanakumar  
e-mail: ksaravanakumar@vit.ac.in



efficient functioning of communities which facilitates collaborative growth and self-enhancement. The field of trust analysis in social networking sites like Facebook, Twitter has been studied extensively in [1]. But as we evolve and let online vendors replace local stores, what actually matters now is the trustworthiness of product reviews. Online reviewing works like a breeding ground for public opinion and can influence business trends in a disguised manner [2]. We tend to trust the “word of mouth” phenomenon and the product reviews more than the information provided by vendors. After the age of bloggers [3], we have reached the age of reviewers; thus, people easily trust reviews of widely acknowledgeable reviewers. What further makes this trust issue more convoluted are the myriad factors which play a role in determining the trust. Some approaches to tackle this problem use graph-based approaches [4] to detect review spam. Another significant work in this area which uses the correlation between reviewer honesty and trustworthiness is [5]. Another factor which often goes unnoticed is that the text of the review also plays a critical role in the review helpfulness and trustworthiness. This paper will attempt to cluster similar reviewers by extrapolating from reviewer history, reviewer buying/reviewing behavior and count of useful words used in their reviews.

## 2 Related Work

The intricacies involved in the field of social trust analysis surfaced less than a decade ago when the web became a strong platform for opinion expression. There have been many significant approaches to tackle this issue which vary widely depending on the task or context in which they are used. One such example is given in [6] which extract social trust relationships between users on Twitter using factors such as influence, cohesion and the valence (sentiment) of the user in an unsupervised manner. Recently there has been an upsurge in the use of personalization in user recommendation which closely relates to our work. One of the pioneers in the area of social-context aware trust influence [7] considers both the participant’s personal characteristics as well as mutual relations for improved recommendation. Some other significant works are [8–10] which further establish that a person prefers recommendations from trusted friends. Sinha and Swearingen [11] and Bedi et al. [12] also demonstrated that given a choice between the recommendation from trusted friends and a recommendation system, the former is more preferred. These provide a firm background to establish the necessity and need for trust-based review recommendation.

One of the dataset widely used to mine trust relationships is the Epinions [13] web of trust dataset which is crawled from [14] and is available in many data repositories. Many researchers [15] use this dataset for mining trust relationships. Epinions dataset has been widely exploited using Bayesian analysis in [16] for trust aware recommendation and for rating prediction in [17]. The task of trustworthiness prediction becomes even more complex when a dataset like Amazon reviews [5] is used which is meant to function as a staging ground for businesses rather than opinions. So broadly speaking there are two types of datasets which are widely exploited in

this regards one which uses user–user and user–product rating like Epinions or Slash Dot and others like Amazon datasets which uses only user–product ratings in trust prediction. Both types of datasets have their own idiosyncrasies. Prediction accuracy in the former type of datasets can be tested, and the negative or positive links can be predicted [17]. The major issue faced in evaluating the trust in the latter is that the accuracy of predictions cannot be adequately supported. Some examples of earlier work in trust and helpfulness detection are discussed in [18].

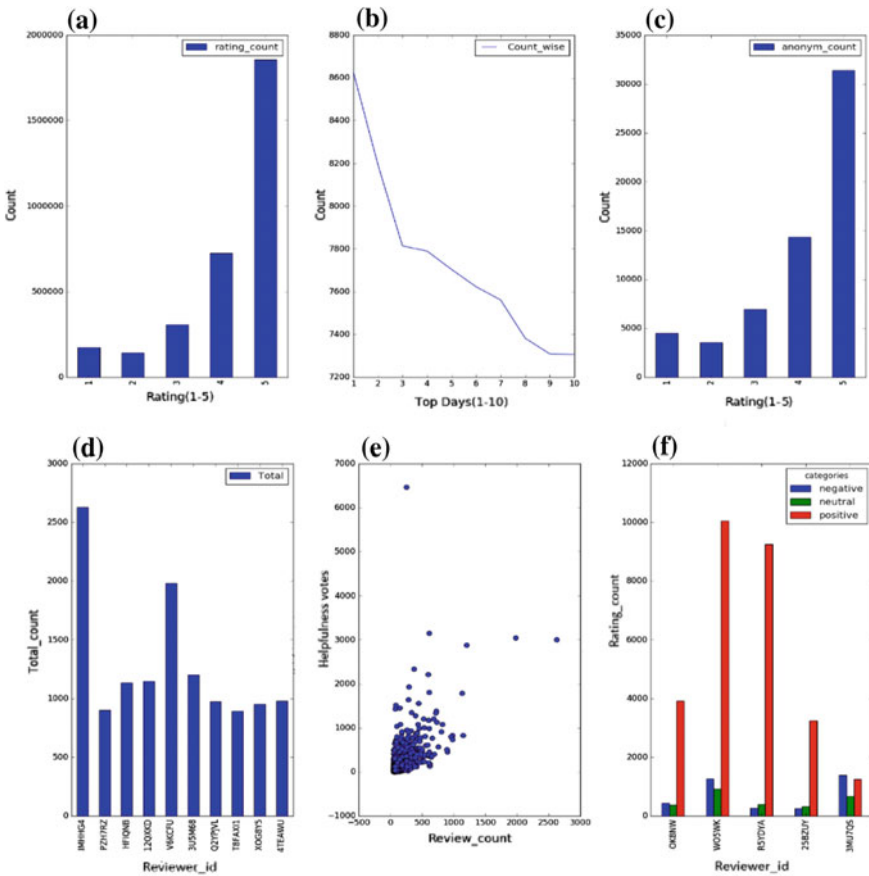
The idea for this project emerged from the analysis and discrimination of the earlier research work conducted in this area the most significant ones being [2] which explore how online review forums provide customers with powerful platforms to express opinions and influence business trends. This paper contributed toward creating a trustworthy co-created recommendation model. The foremost pioneers of this idea were Prahalad and Ramaswamy [18] who defined the concept of co-creation in customer communities. The first functionality of the review is derived from the aforementioned paper. Here the reviewer’s profile is unboxed to evaluate the trust metrics. The parameters used will be reviewer profile history, past helpful votes, reviewer rank, percentage of posts made, account activity, etc. Another paper which significantly inspired our work was the recent work of Wu et al. [19] on how credibility of an advisor is actually perceived in a marketplace. This motivated the second functionality of our project—the advisor segmentation phase.

### 3 Problem Statement

E-commerce Web sites have bought the world closer and have enabled businesses to make huge strides in their profit. What we often tend to ignore is another and perhaps the most significant influence of e-commerce Web sites. It has developed trust relationships among the customers and between the buyer and seller. E-commerce Web sites have heralded the era of mutual trust. Now the question which arises next is whom shall we trust and who can be trusted? Does the reviewer rank alone suffice to trust his/her review? The answer is no. Though reviewer’s rank does play a role in review helpfulness, it does not guarantee a frank and unbiased review. A number of attempts have been made toward the goal of developing a trusted e-commerce network. Most customers prefer to buy expensive electronics and delicate objects in person rather than buying them through e-commerce Web sites. This is because a certain factor of trust does influence buying decision making. Most of us do not trust the social web enough to risk critical buying decisions. This can be changed if we are able to identify and eliminate spammers and provide the vulnerable customer with only trusted reviews.

### 4 Exploratory Analysis on the SNAP Amazon Kindle Dataset: Getting to Know the Data

Figure 1 shows the exploratory analysis of the dataset. Each graph and the insights derived from it are discussed below. The plot (a) displays the number of reviews grouped by their rating. As clearly shown in Table 1, the web is overall positive. A customer does not give a negative review unless he has had a very bad experience with the product. Also the distribution is highly skewed. This means that the number of positives is much greater than the number of negatives. Notice that more than 50% of the reviews are extremely positive. This can mean either they are extremely



**Fig. 1** a Ratings from 1 to 5 and the count in the Kindle dataset. b The top 10 days when the maximum number of reviews were recorded. c Ratings with the anonymity count. d A subset of reviewers and their total review count. e The reviewer's review count and his total helpfulness. f Product id and their positive, negative and neutral review counts

**Table 1** Distribution of ratings in the dataset

Rating	Nature	Rating percentage
1	Extremely negative	5.3974
2	Negative	4.468179
3	Neutral	9.573706
4	Positive	22.647745
5	Extremely positive	57.912966

biased or the data crawled is for books that were highly acknowledgeable. Another insight is the number of reviews per day in Fig. 1b which shows a sudden hike in the number of reviewers given on the Web site during specific days. In the plot below, only the top 10 days with the highest number of reviews are considered. A huge hike in number of reviews on September 6, 2012 is noticed, i.e., 6619 recorded reviews could be possibly attributed to the release of Kindle Fire on the same day in Europe.

Another thing which is quickly revealed on exploring the dataset is that there are a significant proportion of reviews which are anonymous. As can be seen clearly from Fig. 1c, the numbers of anonymous reviewers giving a rating of 5 are large in number. Can anonymous reviews be trusted? The answer is mainly a no as no background information about the reviewer can be obtained. It can be possible that these anonymous reviewers are extremely biased friends of the book’s author or publisher. For the sake of the project, we will not be considering anonymous reviewers though they can be said to show unique patterns and deviations.

The graph in Fig. 1e shows the count of the number of votes per reviewer and the total helpfulness votes that particular reviewer received. There are some visible outliers and also a majority of reviewers with no helpfulness votes. On one hand, there exist some reviewers with relatively few reviews and abnormally large number of helpfulness votes and on the other hand there exist reviewers with a large number of reviews but very little helpfulness votes.

The graph in Fig. 1f shows the number of positives, neutral and negative reviews (y-axis) for a product (x-axis), and the following conditions are considered. A small subset of the products is displayed in the graph. The rating is discretized to categories negative, if ratings are 1 and 2, neutral, if rating is 3, and positive, if ratings are 4 and 5. Another factor considered for dataset exploration is the review count per user. Only some users have counts above 1200, and only two of them show deviations from others in the category showing exorbitantly high number of review count. These could be either spammers or extremely popular and active reviewers.

## 5 The Proposed Model

### 5.1 A Brief Overview

Figure 2 shows the concise model of the proposed approach. Each and every phase is described in detail below.

**Phase 1: Data Collection.** The data sources used are available at SNAP [20] repository, which is the data repository for many large social network datasets. We use the data crawled by jmcauley [21] from the Amazon Web site, which consists of review dataset of a number of product categories. As shown in Fig. 2b, we have considered Amazon Kindle dataset as our major dataset and others like baby products, pets, sports, beauty, health and cell phones as our assistive datasets. The assistive datasets will be majorly used for comparison of reviewer helpfulness across myriad categories. These categories are used to measure the competence of the reviewer and his expertise in a particular category. The main dataset is the Kindle dataset on which trustworthiness is to be predicted.

**Phase 2: Handling Massive Dataset.** The dataset of kindle reviews is split into multiple parts and stored in a compressed format called pickles in python for faster processing. The JSON files are converted to Python pandas dataframe for the analysis purpose.

**Phase 3: Data Cleaning and Transformation.** The next phase is the phase of data preprocessing. In this phase, the following two major subtasks are performed:

- (a) *Data Cleaning.* In this phase, the attributes irrelevant to the study like the reviewTime, reviewerName are eliminated from the dataframe. The dataframe is checked to ensure that any of the attributes considered for the study are not missing, and any noisy attributes are eliminated. The review text is cleaned

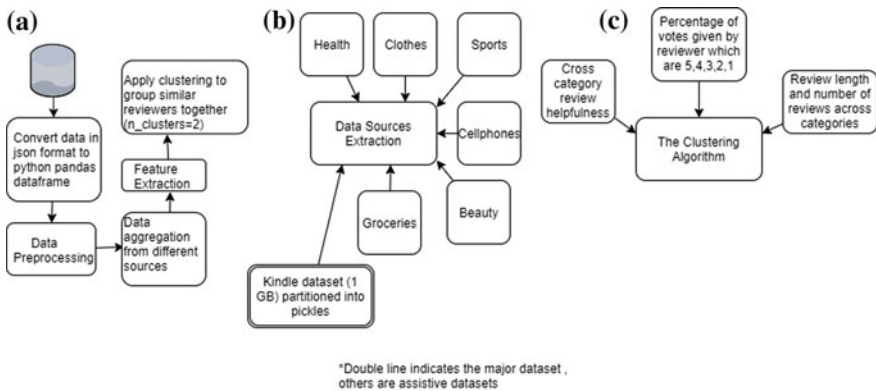


Fig. 2 a Overall architecture. b Data aggregation phase. c The clustering phase

by tokenizing it, eliminating punctuations and eliminating stopwords. Only the feature of interest, i.e., the word count of the review is retained.

- (b) *Data Transformation*. In this stage, a number of transformations are applied on the dataframe attributes. Data transformation stage is crucial for extracting useful features from the dataframe. Details of the feature extraction phase are given in Sect. 5.2.

**Phase 4: Clustering**. In this phase, the reviewers with similar online behavior and reputation are grouped into clusters using K-means algorithm. The algorithm is discussed further in Sect. 5.3.

## 5.2 Feature Extraction

Real-world datasets like the one used in our study are composed of redundant attributes, all of which may not be particularly useful toward achieving the goal. Feature selection is a phase in which the attributes providing obsolete or redundant information are eliminated from the dataset. Feature extraction on the other hand is related to dimensionality reduction; it starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant. The Amazon review dataset has a very high dimensionality. Thus, the use of feature extraction to extract most relevant features is both inevitable and necessary.

**Feature 1: Weighted Helpfulness in All Categories**. The formula below calculates the weighted helpfulness of reviewer over different categories. It is to assign weight depending on the total number of helpfulness votes.

$$WH_i = \alpha * \sum_{j=1}^n \frac{H_j}{T_j} \quad (1)$$

where  $WH_i$  is the weighted helpfulness for reviewer  $i$  and  $H_j$  is the total helpfulness votes for review  $j$  and  $T_j$  is the total views for review  $j$  including both instances when it was voted and not voted.  $\alpha$  is the weight associated with the reviewer. This scheme of weighting is used to ensure that uninformative data, e.g., 1/1 helpfulness which corresponds to 100% helpfulness is handled appropriately (Table 2).

**Feature 2: Count of Useful Words Used**. Average of the review word count of all the reviews for a particular reviewer is taken and is exploited as a useful feature. Tokenize function tokenizes the text. Eliminate punctuations removes the punctuation marks. Eliminate stopwords removes stopwords like and, or, to.

$$AWC_i = \sum_{j=1}^n wci / length(RTi) \quad (2)$$

**Table 2** Deciding value of  $\alpha$  for reviewer  $i$

$T_j$	$\alpha$
$T_j > 1000$	1
$500 < T_j \leq 1000$	0.75
$100 < T_j \leq 500$	0.5
$50 < T_j \leq 100$	0.25
$T_j > 10$	0.1

where  $AWC_i$  is the average word count of reviewer  $i$ ,  $w_c$  is the total number of useful words in review  $j$  and  $RT_i$  is the total number of reviews by reviewer  $i$ .

**Feature 3: Percentage of Ratings.** Percentage of ratings 1 through 5 for each reviewer’s comments is calculated as follows:

$$PR_{ij} = \frac{RAT_i}{T_j} \tag{3}$$

where  $PR_{ij}$  is the percentage of rating  $i$  for reviewer  $j$ ,  $RAT_i$  is the total number of ratings for rating  $i$  ( $i = 1-5$ ), and  $T_j$  is the total number of reviews for reviewer  $j$ .

We classify the type of reviewers in the social e-commerce network as follows:

- (1) *The Popular ones (feature: Popularity).* These are the reviewers who are highly appreciated by other customers. Their typical characteristics are high number of helpful votes in a particular category as compared to other categories. Popularity is calculated as follows:

$$Popularity_i = \sum_i^{Num\_cat} \alpha * WH_i / Num\_cat \tag{4}$$

where  $WH_i$  is weighted helpfulness in category  $i$  (whose weighted helpfulness is non zero) and  $Num\_cat$  is the number of categories in which  $WH_i$  is not zero,  $\alpha$  the weight metric is 1 when the category is books and  $\alpha$  is 0.5 otherwise.

- (2) *The Unbiased (fair) ones (feature: Fairness).*

$$Fairness_i = -0.75 * P5_i - 0.5 * P4_i + 0.5 * P3_i - 0.5 * P2_i - 0.5 * P1_i \tag{5}$$

where  $P5_i, P4_i, P3_i, P2_i, P1_i$  are percentage ratings for reviewer  $i$  calculated earlier. This weighting scheme is used to assign a highly negative value to reviewers with a large percentage of highly negative or positive rating. Here we use modified version of the method proposed in [2] for universal applicability. Above equation assigns a high negative weight to percentage of rating 5 as reviewers with high percentage of highly positive reviews are not fair and a slightly lower negative weight to percentage of rating 4. Percentage of neutral

is added unchanged, and weighing scheme similar to 5 and 4 is followed for percentage of 1 and 2 ratings, respectively.

Five rating is extremely positive; thus, reviewers with a very high percentage of reviews as positive are not trustworthy. Reviewers with a high percentage of 4 ratings are moderately trustworthy, while those with a high percentage of neutral reviews are considerably trustworthy. Also extremely negative reviewers with percentage of negative ratings very high are highly not trustworthy, and similar behavior corresponds to reviewers with high number of 2 star ratings.

- (3) *The Experts (feature: Expertise)*. These are typically the domain experts of the category (e.g., voracious readers in the kindle books categories). They typically write long reviews and are devoted to reviewing a particular category only. This particular feature uses cross-category comparison. According to the Merriam dictionary [21], an expert is a person having, involving or displaying special skill or knowledge derived from training or experience. Thus, experience of a reviewer reflects how much he is trusted in the community. Some reviewers are as the saying goes “jacks of many and masters of none.” Thus, it is necessary to extract the expertise as a feature (Table 3).

$$E_i = [(bc_i) - [(bbc_i) + (byc_i) + (spc_i) + (cpc_i) + (htc_i) + (ptc_i)]] * (awc_i) \quad (6)$$

where  $E_i$  specifies expertise of particular reviewer. In the above equation, the difference between the total review count of book category and sum of total review count of other categories is taken mainly to detect the distribution of the reviewer’s reviews over different amazon categories. For an expert reviewer of the Kindle category, this difference will be significantly high. This value is later multiplied with the average word count of the reviewer. Thus, for an expert reviewer, this equation yields a very high value. All the features above are normalized for better visualization of clusters during clustering.

**Table 3** Abbreviations used

Feature	Abbreviation used
Book category review count	bc
Baby category review count	bbc
Beauty category review count	byc
Sports category review count	spc
Health category review count	htc
Pets category review count	ptc
Cell phones category review count	cpc
Average word count	awc



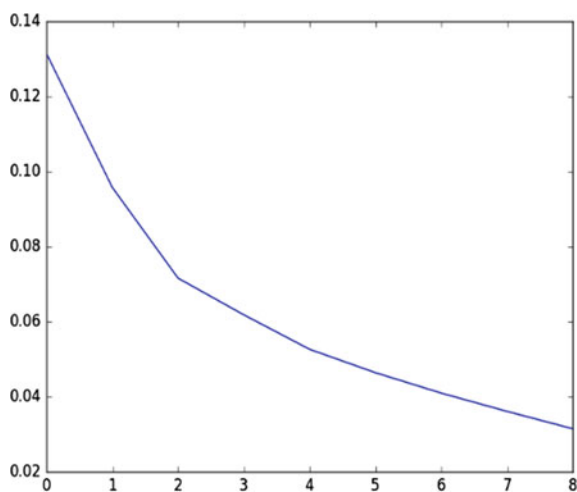
### 5.3 Clustering Phase and Results

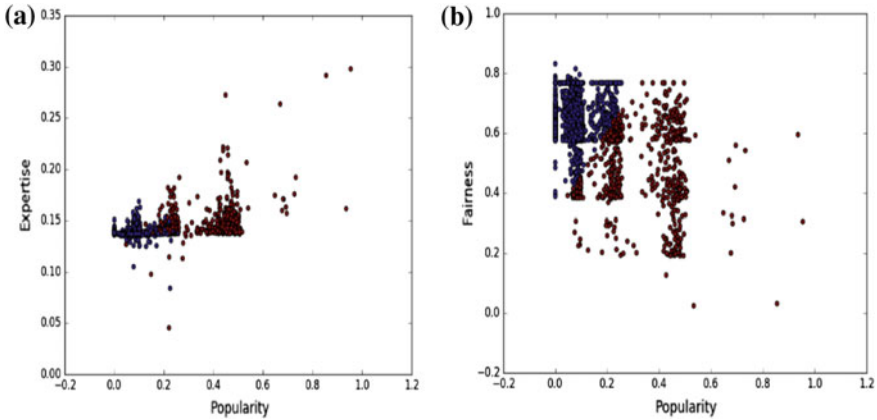
As it is obvious that the detection of trustworthy reviews on an e-commerce platform like Amazon wherein trust relations are not explicitly known, is an unsupervised type of problem. K-Means clustering is the type of clustering in which the items in the dataset which are closely related or whose behavior is similar are said to belong to a particular cluster. K-medoids is also widely used as a clustering metric but here we refrain from using K-medoids mainly because the time complexity is  $O(n^2 * k * i)$ , whereas K-Means runs in  $O(n * k * i)$ . Our dataset spans over a large number of reviewers; thus, in this work, we use K-Means [22] to find the reviewers whose behavior closely resembles each other. The type of clustering performed is 3-D clustering based on the three types of personality traits, i.e., “Popularity”, “Fairness”, “Expertise”, while deciding the number of clusters, the elbow method is used as shown in the graph below. Therefore, the number of clusters to be selected is 2 (Figs. 3, 4).

In K-means, the labels of the clusters are not known. Thus, explicit segregation of trustworthy and not trustworthy reviewers is not possible. This is an issue faced by majority of the unsupervised learning algorithms. Domain-specific knowledge can be incorporated in this study to make the identification of trustworthy reviewers easier (Table 4).

After applying domain knowledge to cluster, the reviewers based on the 3 parameters are categorized into two clusters: the cluster with label as 1 is found to be the group of trustworthy reviewers. The ranks of the labeled trustworthy reviewers were used to determine the prediction accuracy of the algorithm. The ranks were obtained by manually inspecting the rank the reviewers in the sample are assigned by Amazon. A small sample of 67 reviewers shows the results as tabulated above. Thus, analyzing the results obtained the following insights into reviewer trustworthiness:

**Fig. 3** Elbow plot to decide optimum number of clusters





**Fig. 4** **a** Clustering on basis of popularity and fairness. **b** Clustering on basis of popularity and fairness

**Table 4** Percentage of trustworthy reviewers per category in a given sample

Reviewer ranks (sample)	% of trustworthy sample they compose (%)
Rank < 1000	17.91
1000 ≤ Rank < 2000	25.37
2000 ≤ Rank < 3000	11.94
3000 ≤ Rank < 4000	14.92
4000 ≤ Rank < 5000	7.46
5000 ≤ Rank < 6000	7.46
6000 ≤ Rank < 7000	1.49
7000 ≤ Rank < 8000	1.49
8000 ≤ Rank < 9000	2.98
9000 ≤ Rank ≤ 10,000	1.49
Rank > 10,000	7.46

- (1) The majority of the trustworthy reviewers are neither the ones ranked at the top nor the unranked ones but those who lie somewhere between these 2 extremities.
- (2) Though the distribution of trustworthy reviewers does not show a very even pattern in general, trust decreases as the reviewer rank increases, i.e., the probability that a reviewer with rank  $x$  greater than  $y$  is generally less trustworthy.

## 6 Future Work and Conclusion

An in-depth study in the field of social trust analysis defines several possible future extensions to our work. This study does not take into consideration many of the crucial factors which can to a large extent determine the trustworthiness of the reviewer. The

first area of work can be in studying the reviewing behavior of the reviewer. A customer who reviews related products is much more reliable than a customer who reviews products in random categories and significantly unrelated products. Another potential work can be a more detailed analysis of the review text, which tells us a lot about reviewer's personality traits and reliability. First phase of this particular work can be to extract the category relevant product aspect extraction [23], e.g., in books category story, characters, suspense, motivation. The polarity of a review also need to be evaluated carefully to test whether a review is positive or negative [24]. Further the synset may need to be generated to identify similar aspects, e.g., story, plot. Thus, if a customer's review speaks about large number of aspects of the product, the reviewer is considered more trustworthy. Also there are other domains like loyalty in social media [25] which provide scope for a comparative analysis of trust and loyalty.

Thus, this work makes an attempt to identify the trustworthy reviewers in large social networks like Amazon, where users do not explicitly have any trust relations. This is a typical problem of unsupervised pattern recognition. From studying a sample of reviewers clustered on basis of similarity based on their trustworthiness level, it is evident that though highly ranked reviewers are generally trustworthy, this assumption may not always be true. Thus, discovering more accurate and better approaches for social trust analysis is definitely the need of the hour.

## References

1. Chen, L., Tsoi, H.K.: Privacy concern and trust in using social network sites: a comparison between french and chinese users. In: IFIP Conference on Human-Computer Interaction, pp. 234–241. Springer, Berlin, Heidelberg (2011)
2. Li, S.T., Pham, T.T., Chuang, H.C., Wang, Z.W.: Does reliable information matter? towards a trustworthy co-created recommendation model by mining unboxing reviews. *ISEB* **14**(1), 71–99 (2016)
3. Sunny, S., Divya, M., Rachna, M., Soorea, L., Revathi, C., Saravanakumar, K.: Recommendation of blogs in E-learning. *Int. J. Eng. Technol.* **5**(3), 2515–2518
4. Wang, G., Xie, S., Liu, B., Philip, S.Y.: Review graph based online store review spammer detection. In: 2011 IEEE 11th International Conference on Data Mining (icdm), pp. 1242–1247. IEEE (2011)
5. Shinzaki, D., Stuckman, K., Yates, R.: Trust and Helpfulness in Amazon Reviews: Final Report (2013)
6. Vedula, N., Parthasarathy, S., Shalin, V.L.: Predicting Trust Relations Among Users in a Social Network: On the Role of Influence, Cohesion and Valence
7. Wang, Y., Li, L., Liu, G.: Social context-aware trust inference for trust enhancement in social network based recommendations on service providers. *World Wide Web* **18**(1), 159–184 (2015)
8. Berscheid, E., Reis, H.T.: Attraction and Close Relationships (1998)
9. Fiske, S.T.: Social Beings: Core Motives in Social Psychology. Wiley (2009)
10. Yaniv, I.: Receiving other people's advice: influence and benefit. *Organ. Behav. Hum. Decis. Process.* **93**(1), 1–13 (2004)
11. Sinha, R.R., Swearingen, K.: Comparing recommendations made by online systems and friends. In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, vol. 106 (2001)
12. Bedi, P., Kaur, H., Marwaha, S.: Trust based recommender system for semantic web. *IJCAI* **7**, 2677–2682 (2007)

13. Unbiased reviews by real people. [www.Epinions.com](http://www.Epinions.com)
14. Zhang, Y., Yu, T.: Mining trust relationships from online social networks. *J. Comput. Sci. Technol.* **27**(3), 492–505 (2012)
15. Guo, L., Ma, J., Jiang, H.R., Chen, Z.M., Xing, C.M.: Social trust aware item recommendation for implicit feedback. *J. Comput. Sci. Technol.* **30**(5), 1039–1053 (2015)
16. Yuan, W., Guan, D., Lee, Y.K.: The small-world trust network. *Appl. Intell.* **35**(3), 399–410 (2010)
17. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: *Proceedings of the 19th International Conference on World Wide Web*, pp. 641–650. ACM (2010)
18. Prahalad, C.K., Ramaswamy, V.: Co-creating unique value with customers. *Strateg. Leadersh.* **32**(3), 4–9 (2004)
19. Wu, K., Noorian, Z., Vassileva, J., Adaji, I.: How buyers perceive the credibility of advisors in online marketplace: review balance, review count and misattribution. *J. Trust Manag.* **2**(1), 2 (2015)
20. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection, 2011. <http://snap.stanford.edu/data/index.html> (2014)
21. He, R., McAuley, J.: Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: *Proceedings of the 25th International Conference on World Wide Web*, pp. 507–517. IW3C Steering Committee (2016)
22. Kanungo, T., et al.: An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 881–892 (2002)
23. Poria, S., Cambria, E., Ku, L.W., Gui, C., Gelbukh, A.: A rule-based approach to aspect extraction from product reviews. In: *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pp. 28–37 (2014)
24. Banan, T., Sekar, S., Mohan, J.N., Shanthakumar, P., Saravanakumar, K.: Analysis of student feedback by ranking the polarities. In: *Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 203–214 (2016)
25. Hamilton, W., Zhang, J., Danescu-Niculescu-Mizil, C., Jurafsky, D., Leskovec, J.: Loyalty in online communities. In: *AAAI International Conference on Weblogs and Social Media (ICWSM)* (2017)

# Automatic Emotion Classifier



Hakak Nida, Kirmani Mahira, Mohd. Mudasar, Muttoo Mudasar Ahmed and Mohd. Mohsin

**Abstract** Nowadays various social networking sites are popularly used all across the world. People keep on updating their status, thoughts, opinions, suggestions, etc., across the continent which often includes their emotions and sentiments. Thus these sites could provide a very vast and diverse amount of emotion data coming from all cultures and traditions over the globe. Our research would be using the data from one such site, twitter, as its emotion corpus for analysis. Using this data, we will be creating an automatic emotion classifier which can then be used as an efficient emotion classifier for any future data.

**Keywords** Emotion classifier · WordNetAffect · Features

## 1 Introduction

A fundamental demarking feature between humans and rest of the animals in this world, and even in robotic machines, which have replaced humans at many places, are the emotions. Emotions form a very important aspect of our lives. All our writings, sayings, acts, thinking are affected by our emotions. Thus by analyzing these emotions, we are analyzing the humans and their behavior. We can express our emo-

---

H. Nida (✉) · K. Mahira

Department of CSE, Maharshi Dayanand University, Rohtak, Haryana, India

e-mail: hakaknida04@gmail.com

K. Mahira

e-mail: mahira.kirmani@yahoo.com

Mohd. Mudasar

Department of CS, Kashmir University, Srinagar, J&K, India

M. Mudasar Ahmed

Department of CSE, SSM College, Pattan, J&K, India

Mohd. Mohsin

Department of CSE, Kurukshetra University, Kurukshetra, Haryana, India

© Springer Nature Singapore Pte Ltd. 2019

B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713, [https://doi.org/10.1007/978-981-13-1708-8\\_52](https://doi.org/10.1007/978-981-13-1708-8_52)

tions in various ways, like text, voice, expressions, etc. For our survey, we will be analyzing the emotions through textual words obtained from the data on twitter [1]. There has been much work done in this regard [2], where text has been divided into various classes. However, my approach would be to perform sentence-level classification based on Ekman's six emotion classes plus a neutral class [3]. The emotion classification has many applications in real world [4], some of them being in e-learning environment [5], in suicide prevention [6], etc.

In this paper, we will be creating a classifier based on the emotion words which are taken from twitter. The corpus taken is an annotated corpus, annotation being done manually by group of judges and setting up some level of agreement between them for giving a class to a particular sentence. The process consists of three steps: (1) creation of the feature dictionary based on emotion words of the WordNetAffect, (2) creation of the vectors for the words in the feature dictionary and the classes, (3) training the classifier with these vectors, and finally, (4) calculating the precision of our classifier.

The paper is organized as follows. Section 2 describes the related work in the field, Sect. 3 describes how the data is obtained, Sect. 4 describes the methodology, Sect. 5 represents experiments, evaluation and results, and finally, Sect. 5 gives the conclusion of our research.

## 2 Related Work

The primary thing that we need for our classifier is an emotion corpora which is a dataset labeled with emotion classes. Previous works done in the field the dataset have been mostly tagged manually where some 3 or 4 or more judges sit and annotate the data based on the emotion words they find in the corpus. It then involves reaching some level of agreement between the judges as the annotation is purely subjective in nature thus making manual annotation a time-consuming and tedious process. Recently, some works have been done by employing automatically annotating the corpus [7], thus eliminating the need for judges, agreements between them and all. Once we have annotated corpora, we then employ these corpora to train our classifier using supervised learning approach [8] which maps it to a new inferring function for further classification.

Related works in the field of manual annotation using Ekman's classes [3, 9] include annotation of: children stories [10, 11], spoken dialogue [12–14] annotated with emotion categories, web-logs [15] annotated with emotion categories and intensity, news headlines [16] annotated with emotion categories and valence. Besides Ekman's classes, some other manual annotated corpora include: [17] annotated with 14 emotion categories, [18] annotated with 28 emotion categories, [19] includes annotation with 15 emotion categories and at three different levels of document, sentence and element.

Regarding the automatic annotation, several research works have exploited the use of hash tags, emoticons and other emotion indicators, to use them for automatically

annotating the data [20–22]. For example, [23] annotates the data based on emotion hashtags as keywords [24, 25] annotates data based on emoticons identifying six primary emotions of Ekman’s classification based on facial expressions indicated by the emoticon.

In our work, we have employed three different emotion corpora which are annotated manually. We use this data and fully use it to train our classifier using a supervised learning approach to infer a model that can then classify all future examples with higher values of accuracy.

### 3 Methodology

#### 3.1 Process

The overall process involved in our automatic emotion classifier is depicted as under (see Fig. 1).

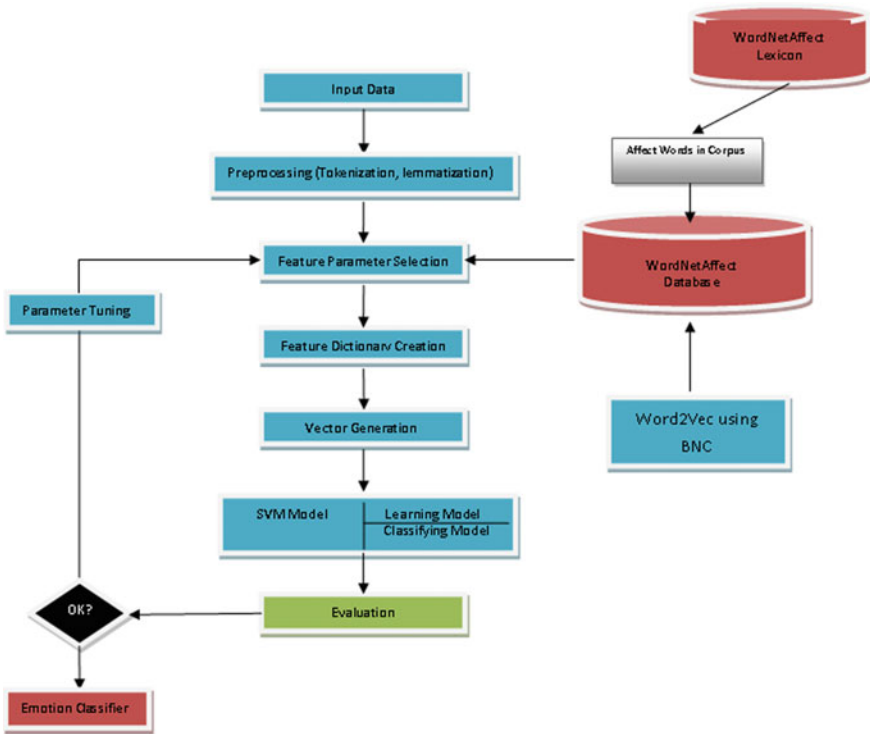


Fig. 1 Process model of automatic emotion classifier

### 3.2 *DataSet*

For the purpose of our research, we have employed three labeled corpora which we have been named as corpus1, corpus2, corpus3; the first one is a dataset of around 1000 sentences obtained from the work of [26] who have collected it from the most popular newspaper headlines. The data has been annotated manually by selecting and setting different degrees of emotional load with each emotion word; the second dataset we have used is an annotated corpus obtained from the work of [27]. They in turn have used the dataset developed by [28], wherein the English tweets are manually labeled by Plutchik's eight emotion classes [29]. By employing this vast dataset, we have collected two separate subsets of data named as corpus2 and corpus3 for our research purpose; corpus2 is a sample of around 30,000 sentences containing many emotion elements like words, emoji, emoticons, etc. But for the experimentation purpose of the classifier, we have manually removed all emojis and emoticons from this subset as we wanted to evaluate our classifier first, as purely based only on the emotion words. Further, instead of Plutchik's eight classes we use only 6 emotion classes from Ekman's basic classification for emotions [29]. Further, corpus3 has been created as another sample from the same corpus as the one used in corpus2 but with different text lines, and most importantly, this time all the emoticons as well as emojis have been included, to evaluate their effect on the accuracy of our classifier. These emojis represent the facial expressions which further augment my WordNetAffect database. This dataset consists of around 40,000 sentences of the corpus.

### 3.3 *Feature Selection*

The whole process is implemented using java. The features are extracted and stored in a feature dictionary in the format;

F: W

Where F is the feature number and W is the weight associated with the respective feature.

The seed word extraction is done by Stanford CoreNlp package. For emotion word labeling purpose, the synonyms for each emotion word are generated using the publically available WordNetAffect [30]. Thus associated with each of the six basic emotions is a group of various affect words generated by WordNetAffect model, creating a large lists of emotion words. Further to expand the affect words more, we have used a Word2Vec model (W2V) on the British National Corpus (BNC). This results in further expansion of our WordNetAffect database.

To obtain an emotion word in context of its left and right emotion words, we have created a file named configuration file, thus by using the file we can obtain the desired context of any emotion word to improve its possibility for getting more appropriate class labeling.



**Table 1** Feature set

Feature	Description	Extracted and obtained through
Personal pronoun	Consists of all personal pronouns present in a sentence	Automatically by Stanford Penn-Bank POS-tagger
Adjectives	Consists of all the adjectives present in a sentence	Automatically by Stanford Penn-Bank POS-tagger
Unigram	Consists of single meaningful words present in a sentence	Generated by Stanford Corenlp
Bigrams	Consists of different groups of paired words present in a sentence	Generated by combination of unigrams
POS	Consists of POS tagging of each word in a sentence	Automatically by Stanford Penn-Bank POS-tagger
POS bigrams	Consists of POS tagging of bigrams of a sentence	Automatically by Stanford Penn-Bank POS-tagger
WordNetAffect lexicon	Consists of list of emotion words associated with each basic emotion word	WordNetAffect emotion list created as defined in this section above
WordNetAffect lexicon with context	Consists of each emotion word associated with its left and right context	Configuration file created as defined in this section above
WordNetAffect POS	Consists of POS tagging of WordNetAffect	Combination of WordNetAffect emotion lexicon and Stanford POS-tagger
Dependency parsing	Consists of dependency parsing for each sentence	Stanford dependency parser

To obtain the dependency of the sentences, we are using the dependency associated with the personal pronouns with a verb feature through Stanford dependency parser [31]. We have also employed porter stemmer [32] for stemming process and a stopwords file that lists the words to be removed from our dataset as being stopwords.

Thus to summarize, the features set includes (Table 1).

### 3.4 Training Model

The model that we have employed for our emotion classifier is an SVM (support vector machine). First of all a feature dictionary is created based on the seed words present in the corpus. In the next step, these features and their values are converted to vectors so that the SVM can work on them and learn from them, since an SVM can operate only on the numeric data. The SVM first learns from the corpus, i.e., trains itself using a loo (leave one out model) model, and finally predicts the class for

each sentence in the same corpus during classification process. Thus by comparing the actual original class label with the one predicted by the SVM, we can measure the accuracy of our system. The formulation used by SVM is described in [33].

## 4 Experiments, Evaluation and Results

Each of the sentences in our corpus is represented by a vector consisting of all the features extracted above, following which is the classification of the sentences to some emotion class. Further this emotion class is also given a numeric label. To create an inferring function for future examples, we use the SVM [32] model, which involves two steps in its classification model generation. First it learns from the training data using the principle of leave one out cross-validation, generating a model file that will map all the future examples and then secondly it predicts and thus classifies the sentences present in the same given corpus. Thus SVM generates the training dataset and the testing dataset from the same annotated corpus provided to it. All the predictions generated are recorded in a report file in the format:

Sentence, original label, predicted label

Finally, we compare the two labels to find if a match exists or not. In this way, the precision of our system is calculated using the formula

$$\text{Precision} = (\text{match}/\text{count}) * 100$$

where,

match represents the total matched original class label and the predicted labels and count is the total number of lines of text in corpora

The results calculated for the three corpora during the evaluation process are represented by a graph as follows (Fig. 2) (Table 2).

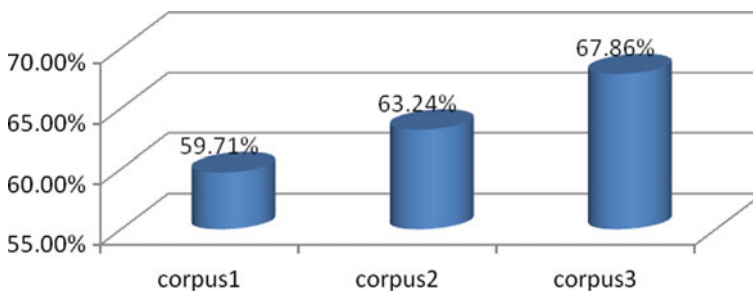


Fig. 2 System accuracy on three corpora

**Table 2** Accuracy of emotion classification

Corpus	Accuracy (%)
Corpus1	59.71
Corpus2	63.24
Corpus3	67.86

## 5 Conclusion

Our classification involves the data that had been labeled manually. Using this annotated data, we have extracted various features from it by using the affect words present in the sentences of corpora. We have expanded these affect words further by using word2vec model. The features generate a classifier by employing SVM whose accuracy is better than many previous works done on the same idea.

## References

1. Information from Social networking site. [www.twitter.com](http://www.twitter.com)
2. Balabantaray, R.C., Mohammad, M., Sharma, N.: Multi-class twitter emotion classification: a new approach. *Int. J. Appl. Inf. Syst.* **4**(1) (2012)
3. Anusha, V., Sandhya, B.: A learning based emotion classifier with semantic text processing. In: *Advances in Intelligent Informatics*, pp. 371–382. Springer International Publishing (2015)
4. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. *Comput. Intell.* **29**(3) (2013)
5. Rodríguez, P., Ortigosa, A., Carro, R.M.: Extracting emotions from texts in e-learning environments. In: Barolli, L., Xhafa, F., Vitabile, S., Uehara, M. (eds.) *Complex Intelligent and Software Intensive Systems (CISIS)*. IEEE Computer Society, pp. 887–892 (2012)
6. Desmet, B., Hoste, V.: Emotion detection in suicide notes. *Expert Syst. Appl.* **40**(16), 6351–6358 (2013)
7. Canales, L., Strapparava, C., Boldrini, E., Martínez-Barco, P.: Exploiting a bootstrapping approach for automatic annotation of emotions in texts. In: *2016 IEEE International Conference on Data Science and Advanced Analytics*
8. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, July 2012
9. Ekman, P.: An argument for basic emotions. *Cogn. Emot.* **6**, 169–200 (1992)
10. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for textbased emotion prediction. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 579–586. Association for Computational Linguistics, Vancouver, British Columbia, Canada, Oct 2005
11. Mohammad, S.: From once upon a time to happily ever after: tracking emotions in novels and fairy tales. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. pp. 105–114. Association for Computational Linguistics, Portland, OR, USA, June 2011
12. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. *J. Am. Soc. Inf. Sci. IEEE Trans. Speech Audio Process.* **13**(2), 293–303 (2005)
13. Chuang, Z.J., Wu, C.H.: Multi-modal emotion recognition from speech and text. *Int. J. Comput. Linguist. Chin. Lang. Process.* **9**(2), 45–62 (2004)

14. Danisman, T., Alpkocak, A.: Emotion classification of audio signals using ensemble of support vector machines. In: PIT, pp. 205–216 (2008)
15. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* 293–303 (2005)
16. Mishne, G.: Experiments with mood classification in blog posts. In: Proceedings of the Style 2005: The 1st Workshop on Stylistic Analysis of Text for Information Access, SIGIR 2005, Salvador, Brazil, 15–19 Aug 2005
17. Jung, Y., Park, H., Myaeng, S.H.: A hybrid mood classification approach for blog text. In: Lecture Notes in Computer Science, 4099, pp. 1099–1103 (2006)
18. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: Proceedings of the 2008 ACM Symposium on Applied Computing, ser. SAC’08. ACM, New York, NY, USA, pp. 1556–1560 (2008)
19. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Recognition of affect, judgment, and appreciation in text. In: Proceedings of the 23rd International Conference on Computational Linguistics, ser. COLING’10. Stroudsburg. Association for Computational Linguistics, PA, USA, pp. 806–814 (2010)
20. Liew, J.S.Y., Turtle, H.R., Liddy, E.D.: EmoTweet-28: a fine-grained emotion corpus for sentiment analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (2016)
21. Boldrini, E., Martínez-Barco, P.: EMOTIBLOG: a model to learn subjective information detection in the new textual genres of the web 2.0-multilingual and multi-genre approach. Ph.D. Dissertation, University of Alicante (2012)
22. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *Processing*, pp. 1–6 (2009)
23. Mohammad, S.: #Emotional tweets. In: \*SEM 2012: The First Joint Conference on Lexical and Computational Semantics—Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pp. 246–255. Association for Computational Linguistics, Montreal, Canada, 7–8 June 2012
24. Purver, M., Battersby, S.: Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 482–491. Association for Computational Linguistics, Avignon, France, Apr 2012
25. Mohammad, S.M., Kiritchenko, S.: Using hashtags to capture fine emotion categories from tweets. *Comput. Intell.* **31**(2), 301–326 (2015)
26. Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: affective text. In: SemEval-Proceedings of the 4th International Workshop on Semantic Evaluation, pp. 70–74 (2007)
27. Meo, R., Sulis, E.: Processing affect in social media: a comparison of methods to distinguish emotions. *Tweets ACM Trans. Internet Technol.* **17**(1), 7 (2017)
28. Suttle, J., Ide, N.: Distant supervision for emotion classification with discrete binary values. *Comput. Linguist. Intell. Text Process.* **7817**, 121–136. Springer
29. Plutchik, R.: Emotion: theory, research and experience. In: *Theories of Emotion*, vol. 11, no. 01, p. 399. Academic Press (1980)
30. Strapparava, C., Valitutti, A.: Wordnet-affect: an affective extension of wordnet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon (2004)
31. Information gathered from: <http://nlp.stanford.edu/software/stanforddependencies.shtml>
32. Information gathered from: <http://tartarus.org/martin/PorterStemmer>
33. Information gathered from: [http://svmlight.joachims.org/svm\\_multiclass.html](http://svmlight.joachims.org/svm_multiclass.html)

# Sentiment Analysis of Tweets Through Data Mining Technique



Taranpreet Singh Ruprah and Nitin Trivedi

**Abstract** Sentiment analysis extracts the mood of a speaker or an author with respect to some subject or the overall contextual polarity of a document. In this paper, an algorithm is proposed for sentiment analysis of tweets extracted from social networking site, i.e., twitter. The comments of users are analyzed and are divided into two parts: positive and the negative sentiments. Algorithms used are keyword spotting and lexical affinity. Tweets are extracted from Twitter through REST API and real time. After that, filtration processes such as stemming, elimination of stop wordsetc are performed and then algorithms are applied on the remaining dataset. The correctness of those algorithms is checked using the results from the standard ALCHEMY API, and the accuracy of those algorithms is calculated individually. A new algorithm is proposed which is the hybrid of both keyword spotting and lexical affinity. The results of that algorithm are generated, and the accuracy is calculated and compared with rest of the algorithms. Machine learning is implemented using NLTK (natural language toolkit).

**Keywords** Hybrid algorithm · ALCHEMY API · Lexical affinity  
Keyword spotting

## 1 Introduction

Sentiment analysis uses the computational linguistics which is used to identify and extract subjective information from source materials. It aims to find out the behavior of a speaker or an author with respect to some statement or the overall contextual polarity of a document. The behavior may be his or her judgment or evaluation, emotional condition, or may be some specific communication. The first step in sentiment

---

T. S. Ruprah (✉) · N. Trivedi  
ADCET-ASHTA, Ashta, India  
e-mail: tpr\_cse@adcet.in

N. Trivedi  
e-mail: trivedini@gmail.com

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_53](https://doi.org/10.1007/978-981-13-1708-8_53)

analysis is to pass the text document to classifying the polarity of a given text at the document, sentence, or feature/aspect level—whether the expressed opinion in a document, a sentence, or an entity feature/aspect is positive, negative, or neutral. Advanced, “beyond polarity” sentiment categorized looks, for instance, at emotional states such as “angry”, “sad”, and “happy”. From the last decade, the sentiment analysis automatically extracts the various expressions like positive or negative expression from the message.

There are mainly two approaches for the sentimental analysis: bag of words (BOW) and feature-based sentiment (FBS) [1]. In Bag of Words technique, all files of data are like a collection of words. The bag of words is not used where the opinion about the product and the feature is examined. In such cases, it is required to extract features. Feature-based sentiment has developed as an approach for examining the assumptions of items and their components. The consequences of slant grouping are introduced in different configurations in various areas: positive/negative, similar to/loathe suggested/not prescribed, great/terrible, purchase/don't purchase [2].

This paper aims at understanding the sentiment by analyzing the reviews, for example, managing the box office revenues of the movie by understanding the sentiment of the movie reviews posted in various social networking sites such as Twitter. The data are extracted from [www.twitter.com](http://www.twitter.com) using REST API and real time and are filtered by using filtration techniques such as stemming, stop-words elimination. The data finally left are the filtered data. The accuracy of the algorithms is checked by ALCHEMY API [3]. Algorithms such as keyword spotting and lexical analysis are applied on the filtered tweet. Dataset is maintained, and in keyword spotting algorithm, the filtered tweet is divided into two words. Each word is checked in the database; if the word in sentence matches the word in database, then the corresponding score is marked. At the end of the sentence, the score of individual words is added and based on that score the sentiment of the sentence that is tweet is analyzed as positive, negative, or neutral. The accuracy of each algorithm is calculated and compared. In this research, a new algorithm is proposed which is the combination of keyword spotting and lexical analyzer. The accuracy of the proposed algorithm is also measured using the same algorithm, and the accuracies of all the three algorithms are compared.

## 2 Literature Survey

Past some work has been done in the area of sentiment analysis. This problem is addressed. Wide range of researches have been done on this field. The following research has been previously done in this field.

1. The basic algorithm used to analyze the sentiment of the tweets is keyword spotting [4].
2. Some of the problems structured with this algorithm are stop-word elimination.

3. This is a shallow algorithm. There is no notation of sarcastic reviews or expressions.
4. This algorithm is totally dependent on database.
5. Nasukawa and Yi [5] used Markov model in natural language processing and statics-based technique to identify the sentiments in the subject.
6. Yi et al. [6] calculate the relation between the topic and the sentiment with the help of mixture model. They calculate the sentiment of particular subject not the whole subject.
7. Godbole et al. [7] analyze the sentiment with the help of seed list, by introducing synonyms and antonym in the positive and negative polarity.
8. Yang et al. [8] use SVM and CRF to calculate the sentiment at per sentence and then calculate the sentiment of the whole document.
9. Naamen et al. apply human coding and analysis of comments to understand the user activities.
10. Go et al. [9] use the model using SVM (Support Vector Machine, MaxEnt, Bayes for those tweets which are positive which end with “:)” “:-)” and the negative tweets with end like “:(” “:-(”).
11. Gamon [10] analyzes the sentiment, based on the data from feedback through global support survey. They analyze the feature like POS tags. They analyze extensive feature and feature selection and calculate the classifier accuracy.

If size of database is small, then it is not effective.

Another algorithm used is lexical affinity [11].

Flaws of this algorithm are summarized as follows:

- (1) It is purely statistical and probability based.
- (2) There is no natural language processing.
- (3) Pictorial representations are unrecognizable.
- (4) Negations are not detected, and hence this affects the accuracy of this algorithm.

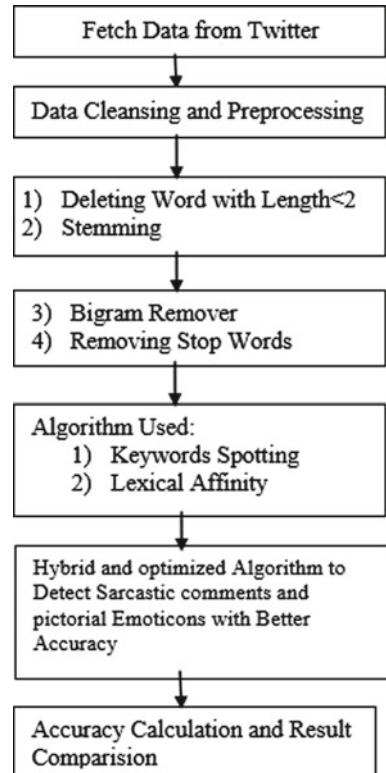
These were the predefined algorithms with their drawbacks. To overcome some of these drawbacks, a new algorithm has been proposed which is the hybrid of both the basic algorithms used previously. The accuracy of hybrid algorithm is much better than the rest of the algorithms.

### 3 Proposed Algorithm

#### 3.1 Keyword Spotting Algorithm

In keyword spotting, the sentence formed after applying filtration techniques is split into words. A file is maintained in the database which contains all the words extracted from thesaurus dictionary. The file contains the words and along with them the score of each word based on the sentiment, for example positive one for positive word and negative one for negative word. Each word from the sentence is checked in that

**Fig. 1** Graphical representation of flow of the proposed system



file containing list of words, and its score is marked. After each and every word of the sentence is checked, the total score is calculated. According to the total score, the sentence or tweet is split as positive or negative words. The accuracy of that algorithm is calculated.

This algorithm lags in many ways.

It is a shallow algorithm; for example, it cannot recognize sarcastic reviews and expressions. It cannot be recognizing when the sentence has negotiation, and it depends on the surface feature [12]. It is dependent on the size of database. It is not effective if the size of database is small [12] (Fig. 1).

### 3.2 *Lexical Affinity*

Lexical affinity is a statistical approach. It is based on probability. After that, two files are maintained in database, namely “pos” and “neg” containing positive and negative words, respectively. Each word from the sentence is checked in both files containing the list of words, and after each word has been checked, the probability of



each sentence is calculated. After calculating the probability, the sentence is analyzed as positive or negative. The accuracy of this algorithm is also calculated. Machine learning is implemented using NLTK (natural language tool kit). This algorithm lags in many ways.

It operates individual word level, so it is easily tricked with the sentence [12].

The probability approach is depending on the text of particular genre, so it is difficult to develop a reusable, domain-independent model [12].

### **3.3 Hybrid Algorithm**

The new proposed algorithm is the hybrid of keyword spotting and lexical affinity algorithms. After analyzing the drawbacks of both the predefined algorithms, we have worked upon making a new optimized algorithm. Checker has been added for detecting pictorial emoticons, for example ☹, ☺, :’ (etc. We detected sarcastic reviews using keyword spotting and extracted the result of lexical affinity.

Thus, the new hybrid algorithm is used for analyzing the sentiment of the tweets from Twitter, which is even more accurate than the rest of the algorithms. The drawbacks of both the previous algorithms are solved in the combined algorithm, which is probabilistic as well as uses bag of words model also.

In the new algorithm, the database which is created for keyword spotting using SQLite is used to mark and check all the deciding adjectives. This is how the positive points of bag of words model and probabilistic model have been combined. As keyword spotting works only on the surface, it is shallow. Therefore, we negated the results in case of sarcasm and those results were by default same as given by lexical affinity algorithm.

### **3.4 Calculating the Agency**

Accuracy of all the algorithms is calculated by the following method. Suppose in a dataset four sentences or tweets are analyzed as positive and six as negative, then we will import the variables by using python and then we check it using ALCHEMY API. Suppose ALCHEMY API finds five sentences as positive and as negative, then accuracy of the algorithm is calculated by finding the difference between the ALCHEMY and keyword spotting algorithm and calculates the probability by dividing it by total number of reviews. This is how the accuracy of keyword spotting is calculated. In the same way, the accuracy of lexical affinity is calculated.

## 4 Experiment and Results

### 4.1 Dataset Description

In this work, dataset is collected by parsing tweets from Twitter using real time. First, 20 real-time tweets are parsed from Twitter using Twitter API, access token, and access key. It gave us 1000s of tweets at a time. We have taken 1000 positive, 1000 negative tweets, and IMDB reviews also. The database for both keyword spotting and lexical affinity is created using SQLite. Words and their synonyms are extracted from thesaurus dictionary and are saved in the database.

### 4.2 Data Preprocessing

Preprocessing includes many steps like deleting words whose length is less than 2. Next step in preprocessing is stemming. Stemming [12] is done using porter stemmer. Other steps such as bigram and stop-words removal are also done.

Stemming is the term utilized as a part of data retrieval to depict the procedure for diminishing arched words to their pledge stem. Stemming programs are regularly alluded to as stemming calculations or stemmers.

Others stemmers such as Lovins stemmer have less accuracy compared to porter stemmer.

## 5 Results

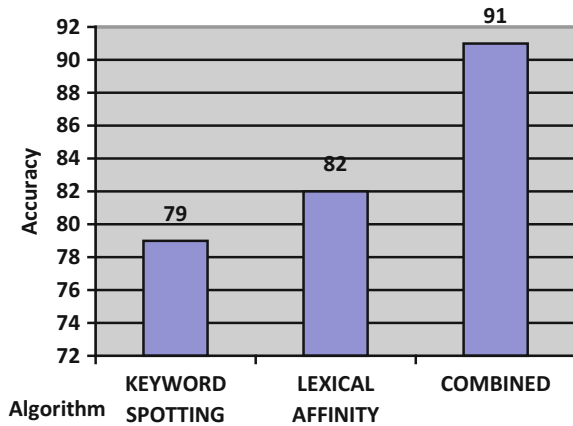
### 5.1 Graphical Representation

See Fig. 2.

### 5.2 Statistical Approach

The results of all the three algorithms (keyword spotting, lexical affinity, and hybrid) are calculated, and their accuracy is calculated using the results of ALCHEMY API is the standard and computerized algorithm for analyzing the sentiment. The results of all the three algorithms are compared with the results from ALCHEMY API, and accuracy is calculated. The accuracies of the entire algorithm used are represented graphically as well as statistically. Comparing the accuracies, we find that the “Hybrid algorithm” is better than the rest of the two algorithms and thus it should be used for analyzing sentiments (Table 1).

**Fig. 2** Graphical representation of accuracy of all three algorithms



**Table 1** Tabular representation of accuracy of all three algorithms

Algorithm	Accuracy (%)
Keyword spotting	79
Lexical affinity	82
Hybrid algorithm	<b>91</b>

## 6 Conclusion

The detailed study of sentiment analysis and its algorithms has been done. We started with extraction of tweets from Twitter by using Twitter REST API and real time. Various filtration techniques such as stemming, stop-words elimination etc have been applied on the tweets. The filtered sentence is split into words, and individually two data mining algorithms such as keyword spotting and lexical affinity have been applied. According to the analysis, the reviews were classified as positive or negative. The analysis was cross-checked using ALCHEMY API. Based on the above analysis, accuracy is calculated. After that, a new algorithm is proposed which is the hybrid of both algorithms applied earlier. The new algorithm was more optimized, and we calculated the accuracy of the proposed algorithm. The accuracies were compared, and it is found that the new algorithm is more accurate than the other two on the same dataset. Algorithms were implemented, and complexity analysis of every algorithm was done successfully along with implementation of a comparison table for the different algorithms.

## References

1. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing (2010)
2. Binali, H., Potdar, V., Wu, C.: A state of the art opinion mining and its application domains. In: IEEE International Conference on Industrial Technology, pp. 1–6, Feb 2009
3. Alchemy Api.: Available from: [www.alchemyapi.com](http://www.alchemyapi.com). Last visited Mar 2012
4. Batool, R., Khattak, A.M., Maqbool, J., Lee, S.: Precise Tweet Classification and Sentiment Analysis. IEEE (2013)
5. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture, pp. 70–77. ACM (2003)
6. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: International Conference on Data Mining, ICDM 2003, Third IEEE, pp. 427–434. IEEE (2003)
7. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), pp. 219–222 (2007)
8. Yang, C., Lin, K.H.Y., Chen, H.H.: Emotion classification using web blog corpora. In: WI'07 Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 275–278. IEEE Computer Society, Washington, DC, USA (2007)
9. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical Report, Stanford (2009)
10. Gamon, M.: Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th International Conference on Computational Linguistics (2004)
11. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining, 2010 IEEE (2013)
12. Cambria, E.: An Introduction to Concept-Level Sentiment Analysis. IEEE (2013)

# UPLBSN: User Profiling in Location-Based Social Networking



G. U. Vasanthakumar, G. R. Ashwini, K. N. Srilekha, S. Swathi,  
Ankita Acharya, P. Deepa Shenoy and K. R. Venugopal

**Abstract** Online social networks serve various purposes and help mankind in many ways. The amount of information in social networks is increasing everyday, making it huge data source for its users. All the data available in social networks may not be trustworthy. In this work, we present an intelligent, crowd-powered information collection system that identifies the set of trusted experts topic-wise in Twitter social network. The proposed UPLBSN algorithm presented in this work identifies trusted experts by finding the relationship between content of tweets and the tweet location. The topic(s) of user posts are clustered by extracting the keywords and are stored in the database. Profiled profound users are presented to the business users based on the topic searched by them. The proposed UPLBSN algorithm is evaluated by conducting experiments on Twitter data set to demonstrate its adequacy.

**Keywords** Data mining · Geolocation · Online social networking  
Topic of interest · User profiling

## 1 Introduction

The use of online social networking sites has tremendously increased since the last decade. Different kinds of social networking sites are available which provide various exposure, feeling, and sharing mechanisms to their users. Few sites allow users to post text data, few others allow only posts related to professional life, whereas other sites allow its users to post text, multimedia, and many more. All sites provide platform for its users to share their interest, knowledge, and other useful information. Such sites also provide ways to connect to their peers who might be geographically located long distance apart. Young generation are more attracted by these social networking sites

---

G. U. Vasanthakumar (✉) · G. R. Ashwini · K. N. Srilekha · S. Swathi · A. Acharya  
P. Deepa Shenoy · K. R. Venugopal  
Department of Computer Science and Engineering, University Visvesvaraya  
College of Engineering, Bangalore University, Bengaluru 560001, Karnataka, India  
e-mail: vasanthakumar.gu.in@ieee.org

© Springer Nature Singapore Pte Ltd. 2019  
B. Pati et al. (eds.), *Progress in Advanced Computing and Intelligent  
Engineering*, Advances in Intelligent Systems and Computing 713,  
[https://doi.org/10.1007/978-981-13-1708-8\\_54](https://doi.org/10.1007/978-981-13-1708-8_54)

in comparison with elderly people, and most of them use at least one of the social networking sites like Facebook, Twitter, Forum, Google+, LinkedIn. In general, users may have profiles in more than one social networking site, and one may have many profiles in single site with different names.

The mechanisms available are easy to reach and use social networking sites, and hence, people around the world use these sites everyday and almost all the time, especially youth. In many cases, these social networking sites help in getting tied up with some unknown people around the world who share same interest and important information. For those seeking some information regarding unknown places or things, these social networking sites may be helpful since they include people around the globe. Queries raised in these sites may be answered by many, and one can pick the required information from the available ones. Other advantages include being in touch with family or friends even though they are geographically separated, forming friends or groups who share same interest, knowledge sharing through images or videos or texts, also to know current affairs from all over the world and connect with any person in the world.

In spite of having multiple advantages, there are some loopholes in social networking like trust while making friends or sharing information, safety for the posted images, and correctness of the information available. Since there are high risks of malicious usage of photos being shared, wrong information may lead to wrong conclusions; friendship with unknown person may create multiple problems and so on. The need for mechanisms to control or avoid such disadvantages is required in every social media. Even though there are some safety measures available in some social networking sites, yet there needs more prominent mechanisms to avoid all possible problems.

Today, many available social networking sites are competing with each other in providing vast and variety of facilities to their users. In addition, these sites are adopting emerging trends and changing needs of users for providing facilities accordingly. Location tagging is one such facility provided by social networking sites where the users can tag the current location from where they are accessing the site. Apart from this, user can also tag some location to their post on which they are commenting. There is no mechanism provided by online social networking sites to check the correctness of location tags with respect to the posted content. Depending upon the interest, the users can post their thoughts, arguments, photos, media, videos, and so on, either with or without tagging the location.

Sometimes the location tagging of users in social networking sites help in various purposes, like analyzing the user activities and identifying user pattern of movement, which contributes in various research topics. As study clarifies that many users provide wrong information regarding location while creating their profiles, such users can be identified by analyzing their location-tagged information for particular period of time [1]. In other words if a user requires some help from others in social network groups/friends, then user can share his/her location which would be helpful in identifying the seeker easily. From many angles, location sharing or tagging by the user is helpful for many purposes.

In many of social networking sites like Twitter, Facebook, Google+, the facility of location tagging has been provided but is not mandatory to share the location for users. Hence, few may utilize these features all the time, whereas few others may not use it at all. Location tagging in social networks is achieved using a well know mechanism called global positioning system (GPS). GPS is a radio navigation, i.e., global navigation satellite system which provides the time and geolocation anywhere on earth, in all climatic conditions. By considering the entire earth along with latitude and longitude, map projection GPS satellite will identify the accurate location.

User activities, pattern of actions, kind of posts, timings, etc., of users contribute in analyzing user behavior, which in turn add up in profiling online social network users [2]. For the same purpose, the newly added attribute is location tagging. From posts, comments, photos, or any media shared by the user along with location tagging provides the user movement pattern and interest. Apart from that, we can also combine user post data along with location-tagged information in order to verify the purpose of post as well as the correctness of the post with respect to location. In many cases, user's posts cannot be blindly believed since the user who posted something may either be true or false. In other cases, user may post some spams or hoax information with respect to some person, location, or event. Hence, there needs a mechanism for filtering and identifying such posts from users which has positive relationship between the posted data, and the location shared is huge. This mechanism combines the content of posted data and location tag and provides it to further processing if true relationship exists. The data is then analyzed to extract keywords which are processed using NER approach and WordNet approach to form unique clusters. This way we make sure that the data we present to those who require is correct and true. Several data mining algorithms allow us to analyze the collected data dynamically [3].

**Motivation:** With more than 200 million accounts on Twitter in diverse geographical locations, the short messages or tweets form a huge data set that can be analyzed to extract geographical information of users. The information obtained can be used to provide users with personalized services such as local news, local advertisements, application sharing. If user needs any idea or help regarding either personal, business, or professional, or any other subject-/topic-oriented information, then those users who are proficient in relevant topic(s) with their interest may provide various ideas and suggestions.

**Contributions:** To get questions answered or to get clarification/help by profound users, proficient in relevant topic(s) of interest, users of Twitter are profiled with respective topic(s) of their interest based on their tweets content with respective locations and are made available to be used by other users in the network.

The remainder of the paper is organized as follows: Sect. 2 gives a glimpse of literature work carried out. The definition of the problem is described in Sect. 3, whereas the proposed system is discussed in detail in Sect. 4. User Profiling in Location-based Social Networking (UPLBSN) algorithm is presented in Sect. 5. Simulation and result analysis are discussed in Sect. 6, whereas various applications of proposed algorithm are presented in Sect. 7. The entire work is summarized with conclusions in Sect. 8.

## 2 Literature Survey

From the new phenomenon developed by Zhang and Li [4], by defining all the attributes of interactions in social networks, they constructed a transmission graph which clearly depicts that the interactions and the duration of interactions are directly depended on the number of individuals. The kind of data that is available in social networks is huge, and many a times the data is not trustworthy. Kefalas et al. [5] have provided complete details of location-based social networks (LBSN) by conducting survey on all related algorithms. By considering the entities like users, groups, their activities, location, and quality, they examined the strength and weakness of LBSN with respect to three perspectives like time awareness, user's privacy issue, and recommendations.

Based on Bayes' rule, by combining both temporal and spatial perspectives, Song et al. [6] have proposed a probability model which provides solution for problems encountered during location prediction for friendship recommendation system in LBSN. With same intention, to provide related and connected recommendations for users based on preferences, a personalized recommender has been proposed by Berjani and Strufe [7] which reads location information of users to understand their preferences. From the results of simulations obtained on Gowalla data set, it is proved that the inclusion of geolocation attribute has made the recommendation algorithm in providing better results for identifying the user behavior and also in providing preferences. To characterize the group of friends formed in social networks and to provide related friend recommendations to users, Silva et al. [8] have developed a friend recommendation system.

Human behavior with respect to social ties has been analyzed by Cho et al. [9]. They considered both the location data available from cell phone and location-based social network for the analysis. From recursive analysis, they proved that human travels which are short and periodic are irrelevant with respect to social ties, whereas the long distance travels are directly dependent on social ties in the network. To identify users' interest requires collecting and analyzing few things from social networks [10]. Debnath et al. [11] have analyzed such data by applying the collaborative filtering on user's data to obtain user interest. They also presented a recommendation system by combining content and collaborative filtering. This combined information utilized by Wan et al. [12] led in proposing another friend recommendation system. Results obtained by experiments on Digg data proved that the method outperforms other systems.

Influential users in the network spread information wider and others in the network follow the trends in the network [13]. Social computing system [14] which analyzes social data to identify the sequential user behavior, i.e., user's behavioral changes with respect to the data shared by one user. As data spreads, it makes influence on others, and other users tend to change accordingly. Decision-making systems help nodes in the network to make decision before accepting friend request or in making decision about the neighboring nodes. Chung-Kai et al. [15] presented a decision-



making system on distributed information sharing system, and Wang and Djuric [16] provided a similar method for the same purpose.

A recommendation algorithm along with prediction model using crowdsourcing “qCrowd” is presented by Mahmud et al. [17] which selects strangers as connection point to get response for the query from others in network. It automatically selects the node in Twitter to get response. For better recommendation for users who has to make decision on accepting friend’s request and to form bond as social ties, IntRank a trustworthy system which can be used in securing social network users by providing better recommendation is proposed by Zhang et al. [18]. This is developed based on repeated analysis on social ties, ranking, and interaction pattern. Using symmetric key encryption mechanism [19], users can access social networks easily and securely.

Decentralized mobile networks are searched when a query is raised. Carbanar et al. [20] have presented a framework PROFIL-R for location centric profiles (LCPs) which is constructed over profiles of users who have visited discrete locations. This mechanism provides much privacy for users and also provides correct data. The method is efficient and has been proved by experimenting on resource-constrained mobile devices. The trends in LBSNs have changed in greater extent to tag the location for each post. With the support of location, i.e., places of visit and user attributes in profile, Wang et al. [21] have proposed a novel framework for co-clustering with multiattribute feature to discover the hierarchical and overlapping communities in LBSNs. Experimental results obtained on Foursquare data set revealed that the proposed framework efficiently identifies the overlapping communities with various perspectives.

Yin et al. [22] have presented a location-aware probabilistic generative model LALDA for recommendation system. Experimental results prove that this method outperforms both cold start problem as well as top-k recommendations in user profiling effectively and efficiently. Mohamed and Abdelmoty [23] have demonstrated how different dimensions of data when combined with location-based networks can help in profiling users. This model applies similarity measures and co-occurrence method along with semantic analysis as tags. Place properties and annotation of place by user are combined to semantically analyze the user interest to profile them.

### 3 Problem Definition

Challenge in present social networks is to find trustworthy people. All the tweets posted by the users may not be trusted as one can tweet anything on any matter without any co-relevance of the content with the location from where they tweet. Our intuition is that a conversation between users can be related to a set of topics such as weather and sports including certain location-specific topics, such as an event related to a city, or a reference to a specific place or an entity in a city.

Given a set of tweets posted by the user, an attempt is made to extract trustworthy tweets by finding relevance if any between the tweeted post and its location and profile the twitter users accordingly with their topic(s) of interest using location-

based social networking. We assume that the users would have enabled their location services while tweeting.

## 4 Proposed System

In this paper, we concentrate on profiling the user, based on their activities in the social media like Twitter. Here, the information of the user is extracted from their tweets which are location tagged and the trustworthiness of the user is found by relating the keywords from the tweets and their location and are presented to the business user.

### 4.1 *Named Entity Recognition [NER] and WordNet*

NER also known as entity identification is a subtask of information extraction that seeks to locate and classify named entities in tweets into predefined categories such as the person, location, organization, money, percent, date, and time. Other keywords are classified using synsets of WordNet dictionary. The keywords are generated and classified using NER and WordNet approach.

For example, if a tweet is posted like “Ankitha is attending an air show in Bangalore on February 18”, then NER takes an unannotated block of text and produces an annotated block that highlights the names of the entities like, Ankitha [person], air show [event], Bangalore [location], and February-18 [date].

### 4.2 *Reverse Geocoding*

Reverse geocoding is the process of back coding of a point location (latitude and longitude) to a readable address or place name understandable by the end user. This permits the identification of nearby street address, places and/or areal subdivisions such as neighborhoods, state, or country.

For example, if a tweet appears like, “The crowd is very much excited when Indian cricket team enters the stadium”, then through reverse geocoding the location will be “Chinnaswamy stadium, Bangalore”. Whereas, NER categorizes Indian, stadium [name], and cricket [event]. Here, since the user has tweeted from Chinnaswamy stadium about the cricket, and hence, this tweet is trusted.

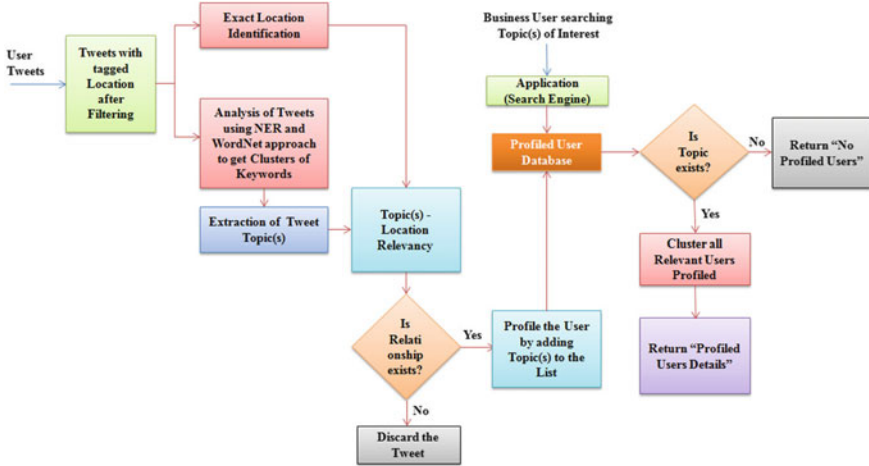


Fig. 1 System architecture diagram

### 4.3 System Architecture

The location-tagged tweets are collected and are preprocessed. They are then fed to the classifier which is a combination of NER and WordNet approaches to classify the keywords into respective categories of topics. Later, the relevance between the topic(s) and the location extracted from where the person has tweeted are established using reverse geocoding. Relevant topic(s) will be listed under each individual user, and they are profiled with those topic(s) and respective trustworthy tweets and are stored in a database as shown in Fig. 1.

When an application user searches for a particular topic, then all those profiled users w.r.t their relevant topic(s) are displayed to the business user where he can explicitly query that trusted profiled user for the help regarding those topic(s) of interest.

## 5 Algorithm

The proposed UPLBSN Algorithm is as shown in Algorithm 1, using which the profound twitter users are profiled with their relevant topic(s) of interest and stored in the database for further use by the business users. As we cannot trust the user by considering a single tweet while profiling, a threshold is fixed so that if the number of trustworthy tweets are greater than or equal to the threshold, only then the person is trusted for the topic before storing his details in the database under trusted persons/users.

---

**Algorithm 1** *User Profiling in Location Based Social Networking (UPLBSN) Algorithm*


---

```

1: while True do
2:   Initialize trustedperson[topic] = 0
3:   for (Every Twitter User) do
4:     Collect and Filter to retain only Location tagged Tweets
5:     for (Every location tagged tweet of the user) do
6:       Initialize trustworthiness of all the topics for the user as equal to zero
         i.e, trustworthy[userid][topic] = 0
7:       for (topic = 1 to n) do
8:         Extract Keywords from tweets using NER and WordNet to Identify
           the topic(s) of Interest
9:       end for
10:      From longitude and latitude values, get exact location using Reverse
        Geo-coding
11:      Check for relevance between location and topic
12:      if (Relevance exist) then
13:        trustworthy[userid][topic]++
14:      else
15:        Discard
16:      end if
17:    end for
18:    for (topic = 1 to n) do
19:      if (trustworthy[userid][topic] >= threshold) then
20:        trustedperson[topic]+=userid
21:      end if
22:    end for
23:  end for
24:  if (Searched topic of Application User exist in Database) then
25:    Display the userids of trustedperson[topic]
26:  end if
27: end while

```

---

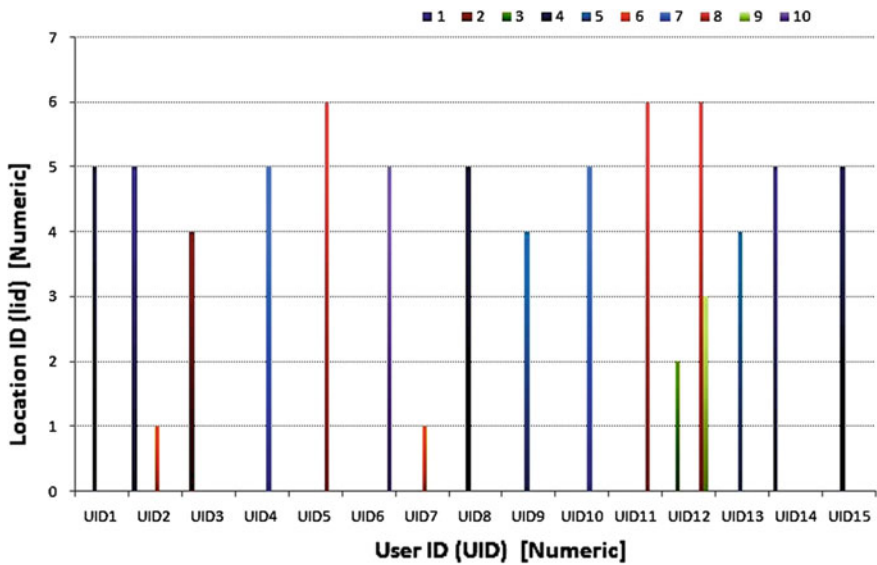
## 6 Simulation and Result Analysis

The proposed algorithm is implemented in MySQL and Java language. A system with Intel Pentium i7 having 4 GB RAM and Windows 8 platform is used for the purpose of simulation.

Twitter APIs, NER Classifiers, WordNet 2.1 are integrated. The data is collected from Twitter for a period of 1 month compounding to 550 tweets from 75 users. The data set consists of user ID, tweet ID, tweet content, location, and timestamp which are preprocessed. Analyzing the data set with the proposed UPLBSN algorithm, profound users are profiled and stored in database. The following search topics searched by the business users along with the profiled users' IDs in respective location are as shown in Table 1. The users are profiled location-wise with respect to relevant

**Table 1** Profiled users with location upon respective topics

Search topic	Topic ID	Location	Location ID	Profiled users
Cricket	tid-1	Bangalore	lid-5	UID-2, UID-14
Shopping malls	tid-2	Chennai	lid-4	UID-3
Cooking tips	tid-3	Kolkata	lid-2	UID-12
Stress management	tid-4	Bangalore	lid-5	UID-1, UID-8, UID-15
Health	tid-5	Chennai	lid-4	UID-9, UID-13
Demonetization	tid-6	Delhi	lid-1	UID-2, UID-7
Colleges	tid-7	Bangalore	lid-5	UID-4, UID-10
Restaurants	tid-8	Hyderabad	lid-6	UID-5, UID-11, UID-12
Food streets	tid-9	Mumbai	lid-3	UID-12
Air show	tid-10	Bangalore	lid-5	UID-6



**Fig. 2** Location-wise topics of profiled users

topic(s) of their interest and are displayed to the business users when they search for information on some specific topic and are shown in Fig. 2. It can be seen from Fig. 2 that location id-4 consists of three profiled users for two different search topics, whereas location id-5 consists of eight profiled users for four different search topics.

## 7 Applications and Discussion

The proposed algorithm is useful in real-time applications to query the information about vehicle traffic, queue in the hospital, crowd in the stadium, queue near ATMs and Banks, and also for new visitors of places to know the information about hotels, colleges, restaurants, etc. It can also be used in applications like “Quora” to find the experts who can answer the query on specific topic.

Twitter users are increasing day by day, and they not only get updates from those whom they elect to follow but also by explicitly choosing to consume new information by searching for a topic in addition to following accounts. Twitter has also added features like “trending topics” to make its users updated, and even major search engines have also been included to search public social streams. With all these potential information and news being spread on Twitter, users do need assurance of reliable information. In order to solve such problems to certain extent, our proposed algorithm is one such method to profile the experts for further dissemination of information.

## 8 Conclusions

To access and connect with the right user for getting further information in social networks, our proposed UPLBSN algorithm may be used efficiently. The algorithm presented in this paper identifies trusted experts by finding relationship between the content of tweets and their locations. Using NER and WordNet approach, the keywords along with synonyms are extracted to form respective clusters. Users are profiled based on the topics to which they belong to with respect to location. This provides business users with the relevant profiled trusted users and their details w.r.t the query requested so that they get connected to them to gather further information of their interest. The simulation results depict the adequacy of proposed UPLBSN algorithm.

## References

1. Cranshaw, J., Toch, E., Hong, J., Kittur, A., Sadeh, N.: Bridging the gap between physical location and online social networks. In: ACM 12th International Conference on Ubiquitous Computing, pp. 119–128, Apr 2010
2. Vasanthakumar, G.U., Sunithamma, K., Shenoy, P.D. Venugopal, K.R.: An overview on user profiling in online social networks. *Int. J. Appl. Inf. Syst.* **11**(8), 25–42. Foundation of Computer Science (FCS), NY, USA, Jan 2017. ISSN 2249–0868
3. Shenoy, P.D., Srinivasa, K.G., Venugopal, K.R., Patnaik, L.M.: Evolutionary approach for mining association rules on dynamic databases. In: *Advances in Knowledge Discovery and Data Mining*, pp. 325–336, Apr 2003
4. Zhang, Y.Q., Li, X.: Temporal dynamics and impact of event interactions in cyber-social populations. *Chaos: An Interdisc. J. Nonlinear Sci.* **23**(1) (2013)

5. Kefalas, P., Symeonidis, P., Manolopoulos, Y.: New perspectives for recommendations in location-based social networks: time, privacy and explainability. In: Fifth ACM International Conference on Management of Emergent Digital Eco Systems, pp. 1–8 (2013)
6. Song, Y., Hu, Z., Leng, X., Tian, H., Yang, K., Ke, X.: Friendship influence on mobile behavior of location based social network users. *J. Commun. Netw.* **17**(2), 126–132 (2015)
7. Berjani, B., Strufe, T.: A recommendation system for spots in location-based online social networks. In: ACM 4th Workshop on Social Network Systems, p. 4 (2011)
8. Silva, N.B., Tsang, R., Cavalcanti, G.D., Tsang, J.: A graph-based friend recommendation system using genetic algorithm. *IEEE Congr. Evol. Comput.* 1–7 (2010)
9. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: ACM 17th SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1082–1090 (2011)
10. Bhat, V.H., Malkani, V.R., Shenoy, P.D., Venugopal, K.R., Patnaik, L.M.: Classification of email using BeaKS: behavior and keyword stemming. In: IEEE TENCON (2011)
11. Debnath, S., Ganguly, N., Mitra, P.: Feature weighting in content based recommendation system using social network analysis. In: 17th ACM International Conference on World Wide Web, pp. 1041–1042 (2008)
12. Wan, S., Lan, Y., Guo, J., Fan, C., Cheng, X.: Informational friend recommendation in social media. In: 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1045–1048 (2013)
13. Vasanthakumar, G.U., Priyanka, R., Vanitha Raj, K.C., Bhavani, S., Asha Rani, B.R., Shenoy, P.D., Venugopal, K.R.: PTMIBSS: profiling top most influential blogger using synonym substitution approach. *Int. J. Soft Comput.* **7**(2). ICTACT, India, Jan 2017. ISSN 2229-6956
14. Gao, Y., Chen, Y., Liu, K.J.R.: Understanding sequential user behavior in social computing: to answer or to vote? *IEEE Trans. Netw. Sci. Eng.* 112–126 (2015)
15. Chung-Kai, Y., van der Schaar, M., Sayed, A.H.: Information-sharing over adaptive networks with self-interested agents. *IEEE Trans. Signal Inf. Process. Over Netw.* **1**(1), 2–19 (2015)
16. Wang, Y., Djuric, P.M.: Social learning with bayesian agents and random decision making. *IEEE Trans. Signal Process.* **63**(12) (2015)
17. Mahmud, J., Zhou, M.X., Megiddo, N., Nichols, J., Drews, C.: Recommending targeted strangers from whom to solicit information on social media. In: ACM International Conference on Intelligent User Interfaces, pp. 37–48 (2013)
18. Zhang, L., Fang, H., Ng, W.K., Zhang, J.: IntRank: interaction ranking-based trustworthy friend recommendation. In: IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, pp. 266–273 (2011)
19. Zhang, L., Li, X.Y., Liu, K., Jung, T., Liu, Y.: Message in a sealed bottle: privacy preserving friending in mobile social networks. *IEEE Trans. Mob. Comput.* **14**(9) (2013)
20. Carbanar, B., Rahman, M., Ballesteros, J., Rische, N., Vasilakos, A.V.: PROFILR: toward preserving privacy and functionality in geosocial networks. *IEEE Trans. Inf. Forensics Secur.* **9**(4) (2014)
21. Wang, Z., Zhang, D., Zhou, X., Yang, D., Yu, Z., Yu, Z.: Discovering and profiling overlapping communities in location-based social networks. *IEEE Trans. Syst. Man Cybern. Syst.* **44**(4) (2014)
22. Yin, H., Cui, B., Chen, L., Hu, Z., Zhang, C.: Modeling location-based user rating profiles for personalized recommendation. *ACM Trans. Knowl. Discov. Data* **9**(3) (2015)
23. Mohamed, S., Abdelmoty, A.: Uncovering user profiles in location-based social networks. In: International Conference on Advanced Geographic Information Systems, Applications, and Services (2016)

# Performing Interest Mining on Tweets of Twitter Users for Recommending Other Users with Similar Interests



Richa Sharma, Shashank Uniyal and Vaishali Gera

**Abstract** With an upsurge in the popularity of microblogging sites, Twitter has emerged as a huge source of assorted information. People often use Twitter to post their ideas and beliefs about the prevailing issues, feedbacks about products they use and opinions on the topics which appeal to them. Therefore, Twitter is considered to be one of the most appropriate virtual environments for information retrieval through data extraction as well as for analysis and drawing out inferences. This paper proposes a system that maintains a database of the Twitter users, fetches their areas of interests and accordingly recommends them the lists of other users with similar interests whom they may like to follow. The prototype of the system is developed in R and has been evaluated on various datasets. The results are promising and portray decent levels of accuracy, i.e., the proposed system is able to discover the correct area of interests of the users and accordingly make appropriate recommendations.

**Keywords** Twitter · Tweets · Information retrieval · Interest mining · R

## 1 Introduction

Due to increased use of social media applications, interest mining is gaining prominence as a hot topic of research. Interest mining involves techniques to extract information regarding the interests of people from texts, images, music playlists, etc. Users post their views on products and services used by them, opinions about political and religious issues or simply present some factual information related to their

---

R. Sharma · S. Uniyal · V. Gera (✉)  
Department of Computer Science, Keshav Mahavidyalaya, University of Delhi,  
New Delhi, Delhi, India  
e-mail: vaishaligera95@gmail.com

R. Sharma  
e-mail: rsharma@cs.du.ac.in

S. Uniyal  
e-mail: uniyalshashank94@gmail.com



interests. Such social media applications include blogs, bookmarks, communities, files, forums, microblogs, profile tags, wikis and so forth. Out of these, microblogs and profile tags most accurately reflect a person's area of expertise or interest [4].

Twitter is one such microblogging site with more than 313 million monthly active users from around the world [3]. This volume of Twitter users tweeting regularly on varied topics makes a rich repository of data available on Twitter for analysis and research purposes. Since Twitter data are abundant and freely available, researchers see it as a valuable source of input for their research in various subfields of data mining, [14] for example, sentiment analysis [7] and text mining [19]. Besides being popular among users, Twitter restricts the users to frame meaningful tweets within the limit of 140 characters, making the tweets easier to parse.

The aim of this research is to determine the areas of interest of a Twitter user on the basis of what the user posts frequently and accordingly suggest him people with similar interests he can follow. Through this, we bring together people with similar interests. The idea is to generate a list containing people sharing similar interests; this list is self-evolving such that as soon as the system finds the areas of interest of a user it makes a new entry for it.

The mentioned approach matches root words present in a particular tweet with a predefined list, and based on the number of matches the genre of the tweet is determined. For example, consider the following tweet by the cricket expert Harsha Bhogle:

*So enjoyed watching @ImZaheer bowl. That first inswinger to Rahane was a classic. Wonder if there is another IPL left in him...*

Here, the terms—*ImZaheer*, *bowl*, *inswinger*, *Rahane*, *IPL*—are associated with cricket and hence can be categorized to be related to cricket and if more such tweets are found in his account, then it can be inferred that Harsha Bhogle is a cricket enthusiast.

The work done under this research can be divided into 3 sections:

- (1) Applying parsing techniques on extracted tweets of Twitter users to find their interest areas and store this information in a database.
- (2) Use the database having the information about interests of previous users to suggest every next user the list of people he can follow.
- (3) Examining the accuracy of the algorithm.

Among the different software packages that can be used to analyze Twitter, R offers a wide variety of libraries and packages that meet the requirements of this research. R is open source and provides a large integrated collection of tools for data analysis. R is designed to interface well with other technologies that included programming languages and databases [5, 13].

For this research, Twitter API was used to collect a corpus of text posts from 16 Twitter users to users in accordance with the genre of their tweets mainly into two categories—(1) politics and (2) cricket (these two being the areas of interest catered in this research). For this, it was required to create two dictionaries containing the terminologies related to these fields.



Fig. 1 Interest mining framework

Figure 1 illustrates the framework of the followed approach. First the tweets corresponding to a particular Twitter user are fetched and stored in a.csv file. Then, using various R libraries every tweet is split into individual words and these words are then compared with the predefined dictionaries to categorize the people on the basis of their interests and this information is stored in a MySQL database. Finally, in accordance with areas of interest of a user, a list of people is suggested whom he can follow. An entry of the current user is also made in the database such that he could also be recommended to other people making this system self-evolving.

The organization of the paper is as follows: The next section presents the related research work. Section 3 puts forth the proposed methodology. Section 4 discusses the results followed by challenges faced in doing this research and the work to be carried in future in Sect. 5. Section 6 concludes the paper.

## 2 Related Work

Traditionally, Twitter feeds have been used as a corpus for sentiment analysis and opinion mining as Twitter is used by people to express opinion about different topics. Twitter contains huge number of text posts which come from celebrities, company representatives, world leaders and general people. Thus, the data of Twitter become valuable for marketing and social studies. Prior work done in this field is related to classification of tweets as positive, negative or neutral. In [10], the author used “TreeTagger” for POS tagging and observed the patterns in distributions among positive, negative and neutral sets and concluded that emoticons and facts are stated by the use of syntactic structures. Read in [12] used emoticons and formed a training set for sentiment classification. For this purpose, the author used “usenet” newsgroups to

get emoticons from texts. The dataset was divided into “positive” (happy emoticons) and “negative” (sad or angry emoticons) samples for application of machine learning techniques.

Twitter data has also proved its worth for evaluation of performance of different machine learning algorithms. In [1], the authors used emoticons as noisy labels and showed that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) with certain preprocessing steps have 80% accuracy when trained with emoticon data.

In [2], authors have used Twitter to measure the popularity of a user and his influence on Twitter using 3 measures of influence—in degree, retweets and mentions. Through their research they found that popularity is not gained spontaneously but through continuous efforts.

Recently, mining the interests and areas of expertise of a Twitter user has gained prominence. Research scholars have used Twitter data (text, photographs) to extract the areas of interest of a person. In [4], the evaluation done by the authors compared the usefulness of eight different social media applications for mining expertise and interests. The results suggest that socialization sources such as people’s tag and blogs are more accurate for extracting the areas of interest/expertise in comparison to the collaborating sources, such as files and wikis. In [11], Qiu and Cho through their research tried to observe patterns in users’ past search histories to know their interests. Wang et al. [16], Wen and Lin [17] projected to deduce user interests from users’ social connections and interactions. Li et al. [8] used the information about places visited by people to mine their interests. In [6], Kim et al. categorized user interests by reading level and topic distributions. In [18], the authors studied the problem of interest mining from personal photographs. They proposed an approach of user image latent space model to model the user’s interest and image content. In [9], the authors suggest an approach named “*twopics*”, which characterizes users’ topics of interest, by recognizing the entities that appear very frequently in a tweet. The tweet is parsed for its entities which are disambiguated first and are then discovered (power becomes power play). The discovered entities are then used to determine the topics of interest. Their system was able to achieve 52.33% accuracy.

This paper proposes a system to find areas of interest of twitter users from what they post on Twitter and accordingly suggest them other Twitter users with similar interests whom they can follow. The application implements preprocessing of tweets of users to find their interest and then recommends to them other users with similar interests.

### 3 Proposed Methodology

The proposed methodology involves the stages as shown in Fig. 2.

The input of the Application is user Tweets. User Tweets are being fetched through Twitter API and *twitterR* library of R. The rest of the process is described in the following phases.

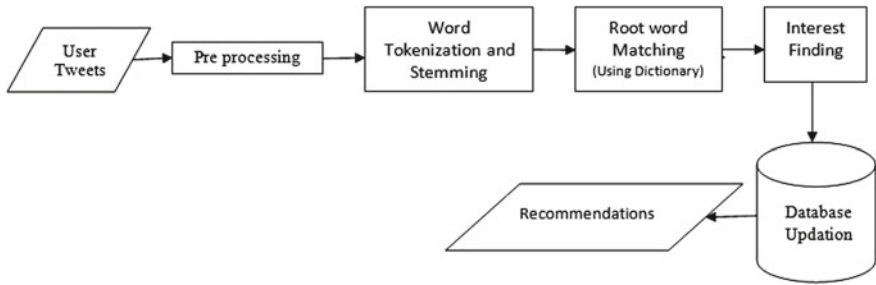


Fig. 2 Work flow diagram

A. *Preprocessing*: Preprocessing phase has to be done on tweets to clean and prepare them for classification with better accuracy. This cleanup is done by R’s regex-driven global substitute, `gsub()`. Preprocessing is done by performing following operations:

1. **Stop words removal**: Since stop words (a, about, further, every, also, is) do not possess any relevant information, therefore to make searching process easy, they must be removed.
2. **Punctuation removal**: Punctuation characters (! “ # \$ % & ‘ () \* +, - . / : ; < = > ? @ [ \ ] ^ \_ ‘ { } ~) are removed from each individual tweet.
3. **Control words removal**: Since content of control words (n, r) determine action rather than meaning, they must be removed.
4. **Digits removal**: Digits are also removed to make a tweet concise having valuable information.

For example, in the following tweet:

*No School Bag Day Will Break Rote Learning & Foster All Round Development. Thumbs Up to the Idea @myogiadityanath.,*

“no, will, &, ;, all, ., up, to, the, @” will be removed; hence, the output after the preprocessing steps will be:

*school bag day break rote learning amp foster round development thumbs idea myogiadityanath.*

B. *Word tokenization and stemming*: For further processing, each Tweet is broken into individual words. These individual words are replaced with their root words using Porter’s algorithm for matching with the dictionary [15]. The output of the above preprocessed tweet will be:

*“school” “bag” “day” “break” “rote” “learn” “amp” “foster” “round” “develop” “thumb” “idea” “yogiadityanath”.*

C. *Root word matching*: The root words obtained from the previous step are compared with the dictionary containing the politics and cricketing terms. This

gives a count for number of matches which is prerequisite for finding the areas of interest.

- D. *Interest finding*: The number of tweets which do have some words that match to the terms in the dictionary are calculated and checked if it exceeds a threshold value to infer the areas of interest.
- E. *Database updating*: After having known the interest of a user, an entry of the current user and his interest is made into the database.

Finally, on the basis of a person's interest a list of people who share the same interest is recommended to him.

## 4 Results

Unlike some other microblogging services, the data posted by different twitter users is freely available. This data can be used for creating standardized datasets for various purposes. For this research, Twitter feeds from 16 different Twitter users were used as the corpus. Among the 16 users, 10 were experts in cricket, while 6 held interest in politics. The algorithm was applied on this data to evaluate the accuracy measures.

Tables 1 and 2 depict statistical data for politics and cricket as areas of interest. The value for expected count was calculated manually. "Expected" count is the actual number of tweets belonging to either area of interest from the total fetched tweets, whereas the "measured" count is the number of tweets belonging to either interest as detected by the application. These two values are further used to measure the accuracy of the system.

The formula used to measure the accuracy of the classification process where the person's interests are being fetched is given as:

$$\text{Accuracy} = \frac{\text{Measured value}}{\text{Expected Value}} \times 100 \quad (1)$$

Table 1 depicts the accuracy measure for cricket. The results are encouraging and went up to 80% in many cases but at the same time dipped to 60% in some other cases.

From Table 2, it can be observed that the numbers for accuracy achieved for politics are better than in case of cricket. The accuracy was more than 90%, in fact it even went up to 100% in many cases.

The rows which do not have any values for accuracy are the cases where measured value exceeds the expected value, i.e., the number of tweets calculated by our application for that particular interest is more than the actual number of tweets for that interest.

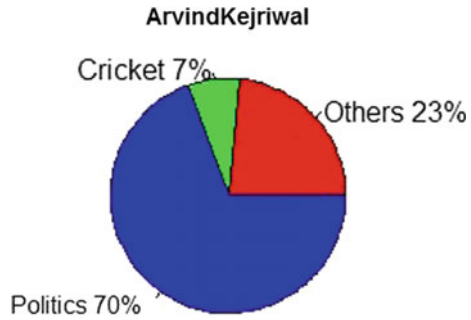
**Table 1** Statistical data for cricket-based tweets

User name	Expected count	Measured count	Accuracy for cricket (%)
Bhogleharsha	164	117	71.3
Cricketwallah	114	77	67.5
ArvindKejriwal	2	9	–
ShashiTharoor	9	8	88.8
Sanjaymanjrekar	19	16	84.2
Rgcricket	44	38	86.3
RajatSharmaLive	0	0	100
VijayGoelBJP	2	4	–
Sardesai Rajdeep	5	13	–
SeerviBharath	88	76	86.3
Mohanstatsman	64	37	57.8
Virendersehwag	17	20	–
Kp24	7	6,8	85.7
Narendramodi	0	0	100
Gauravkapur	35	32	91.4
Kartikmurli	4	4	100

**Table 2** Statistical data for politics-based tweets

User name	Expected count	Measured count	Accuracy for politics (%)
Bhogleharsha	0	9	–
Cricketwallah	11	27	–
ArvindKejriwal	113	89	78.7
ShashiTharoor	75	72	96
Sanjaymanjrekar	0	2	–
Rgcricket	1	3	–
RajatSharmaLive	30	29	96.6
VijayGoelBJP	75	77	–
Sardesai Rajdeep	46	43	93.4
SeerviBharath	0	0	100
Mohanstatsman	3	4	–
Virendersehwag	2	6	–
Kp24	2	2	100
Narendramodi	30	27	90
Gauravkapur	0	5	–
Kartikmurli	0	0	100

**Fig. 3** Distribution of *ArvindKejriwal* tweets



For example, in case of *sardesairajdeep*, the value for accuracy for cricket is calculated as:

$$\frac{13}{5} \times 100 = 260$$

The value exceeds 100% and is thus ambiguous because there are certain words which find place in both the dictionaries (for example, the word “power” as “power—play” in cricket and simply “power” in politics) as a result such words affect the count for measured value.

Figure 3 illustrates the distribution of tweets for *ArvindKejriwal* which clearly shows his interest in politics.

Recall, Precision and F-score were used as metrics for evaluation of the application.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

$$F\text{-score} = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

Calculation of these is done using confusion matrix. In the confusion matrix, the “correct” cells are:

**True Negative (TN):** Case was negative and predicted negative, i.e., a user who did not have interest in a particular area was correctly identified as not having interest in that area.

**True Positive (TP):** Case was positive and predicted positive, i.e., a user who did have interest in a particular area was correctly identified as having interest in that area.

And the “error” cells are:

**Table 3** Confusion matrix for cricket-based tweets

	Interested	Not interested
Interested	6(TP)	3(FN)
Not interested	1(FP)	6(TN)

**Table 4** Confusion matrix for politics-based tweets

	Interested	Not interested
Interested	4(TP)	2(FN)
Not interested	0(FP)	10(TN)

**False Negative (FN):** Case was positive but predicted negative, i.e., a user who did have interest in a particular area was incorrectly identified as not having interest in that area.

**False Positive (FP):** Case was negative but predicted positive, i.e., a user who did not have interest in a particular area was incorrectly identified as having interest in that area.

Tables 3 and 4 show the confusion matrices for both areas of interest. They were calculated separately to know the levels of accuracy for both cases.

$$\text{Recall} = \frac{6}{6 + 3} = 0.66$$

$$\text{Precision} = \frac{6}{6 + 1} = 0.85$$

$$\text{F-score} = \frac{2 * 0.66 * 0.85}{0.66 + 0.85} = 0.74$$

$$\text{Recall} = \frac{4}{4 + 2} = 0.66$$

$$\text{Precision} = \frac{4}{4 + 0} = 1$$

$$\text{F-score} = \frac{2 * 0.66 * 1}{0.66 + 1} = 0.79$$

On evaluation the application gave better results in case of politics than cricket. This can be attributed to the fact that linguistics for politics is much less diverse than in case of cricket. There is a specific trend that can be observed in most politics genre tweets. For example, users mentioned names of political leaders and political parties either by twitter handle name or by hash tag. However, in cricket no such trend could be observed as every expert had his own creative way of expressing his views.



The cases that resulted in ambiguous values for accuracy were more in politics than cricket. This is because of the intersecting words in dictionaries of both areas of interest which resulted in ambiguity. For example, root word “*power*” is used as “*power play*” in cricket but only “*power*” in politics.

## 5 Challenges and Future Work

The application is a basic prototype, yet it generates a lot of encouraging results. It accurately retrieves the areas of interest of users and makes appropriate recommendations accordingly.

However, there are certain challenges associated with the application. The predefined dictionaries maintained for every interest (or field) need to be updated frequently and should be made as specific as possible. Even after listing almost all relevant terms specific to a particular field, there still remain words that can be found in more than two dictionaries leading to conflicting results. Since different people have different style of writing; we cannot define our dictionary in accordance with the choice of every single person. For example, Harsha Bhogle refers to the cricket team Mumbai Indians as *#mipaltan*, while Aakash Chopra refers to them as *#mi*. Therefore, automatic analysis of such diverse and ambiguous tweets poses a challenge [10].

Apart from this, there were certain limitations while performing Twitter analysis using R—firstly, the number of retrieved tweets was less than the number of requested tweets; secondly, the older tweets could not be retrieved.

In future, we plan to expand the system to improve upon the results by incorporating self-updating dictionaries, disambiguation via context, inclusion of third-party tools for better processing and integration with machine learning techniques.

## 6 Conclusion

Through this paper, we researched on how Twitter may prove to be a powerful source of data that can be analyzed to give out purposeful information. Each user on Twitter wants to follow people having similar areas of interest to stay updated on any information regarding the common area of interest, to formalize opinion and to maintain better social relationships.

This paper summarized the results of our application, whose objective is to recommend a Twitter user people he can follow according to his interest. After having fetched the areas of interest of different Twitter users from their tweets, we store their details in a database, thereby making suggestions to every user regarding people he can follow according to his interest, thus clustering together the people with similar interests.

## References

1. Alec, G., Lei, H., Bhayani, R.: Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group (2009)
2. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: the million follower fallacy. In: International AAAI Conference on Weblogs and Social Media (2010)
3. Company>About (Twitter). Retrieved from <https://about.twitter.com/company> 26 May 2017
4. Guy, I., Avraham, U., Carmel, D., Ur, S., Jacovi, M., Ronen, I.: Mining expertise and interests from social media. In: International World Wide Web Conference, Rio de Janeiro, Brazil (2013)
5. Hennessy, A.: Sentiment Analysis of Twitter Using Knowledge Based and Machine Learning Techniques (2014)
6. Kim, J.Y., Collins-Thompson, K., Bennett, P.N., Dumais, S.T.: Characterizing web content, user interests, and search behavior. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, pp. 213–222. ACM, Seattle (2012)
7. Kiruthika, M., Woonna, S., Giri, P.: Sentiment analysis of twitter data. *Int. J. Innov. Eng. Technol.* **6**(4), 264–273 (2016)
8. Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.-Y.: Mining user similarity based on routine activities. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information System, p. 34. ACM, Irvine (2008)
9. Matthew, M., Macskassy, S.A.: Discovering users' topics of interest. In: Proceedings of Workshop of Analytics of Noisy and Unstructured Text Data(AND). ACM, Toronto, Ontario, Canada (2010)
10. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC (2010)
11. Qiu, F., Cho, J.: Automatic identification of user interest. In: Proceedings of the 15th International Conference on World Wide Web, pp. 727–736. ACM, Edinburg (2006)
12. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *The Association for Computer Linguistics* (2005)
13. Seefeld, K., Linder, E.: *Statistical Using R with Biological Examples*. University of New Hampshire, Durham (2007)
14. Tarlekar, A.K.P.K.: Sentiment analysis of twitter data from political domain using machine learning techniques. *Int. J. Innov. Res. Comput. Commun. Eng.* **3**(6), 5590–5597 (2015)
15. The Porter Stemming Algorithm. Retrieved 26 May 2017, from Tartarus: <https://tartarus.org/martin/PorterStemmer/> Jan 2006
16. Wang, T., Liu, H., He, J., Du, X.: Mining user interests from information sharing behaviors in social media. In: *Advances in Knowledge Discovery and Data Mining*. Springer, Berlin, pp. 85–98 (2013)
17. Wen, Z., Lin, C.-Y.: On the quality of inferring interests from social neighbors. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Washington, DC, pp. 373–382 (2010)
18. Xie, P., Pei, Y., Xing, Y.X.: Mining User Interests from Personal Photos. *Association for the Advancement of Artificial*, Pittsburgh (2015)
19. Zhao, Y.: *Text Mining with R—Twitter Data Analysis*, Melbourne (2014)

# Author Index

## A

Aanhey Mani Tripathi, 303  
Aditi Sharma, 437  
Aditya Kesharwani, 531  
Akshai Aggarwal, 523  
Aman Jatain, 13  
Anand Jatti, 3, 25  
Ankita Acharya, 581  
Arun Kumar, 451  
Ashish Kumar, 75  
Ashwini, G. R., 581

## B

Barsha Abhisheka, 167  
Bekri, M. A., 157  
Benhlina, S., 157  
Bennis, L., 157  
Bhavna Gupta, 119  
Binod Kumar Pattanayak, 349

## C

Charanjeet Kour Raina, 341

## D

Dayashankar Singh, 303  
Deepa Shenoy, P., 403, 581  
Deepika, S., 493  
Deepu Krishna, S., 25  
Dhvani Shah, 109  
Divya Meena, 49  
Dwijjoy Sarkar, 475

## F

Fatima Gulnashin, 149

## G

Geeitha, S., 139

## H

Hakak Nida, 565  
HariPriya, D. K., 403  
Harish Sharma, 149  
Hemant Rathore, 451

## I

Ishita Dutta, 85  
Iti Sharma, 149

## J

James Raj, A. M., 257  
Jayanthi Muthuswamy, 269  
Julian Benadit, P., 257

## K

Kakarla, J., 293, 483  
Kakoty, N. M., 61  
Kannan Balaji, 531  
Kapil Mishra, 329  
Kathirvalavakumar Thangairulappan, 97  
Kirmani Mahira, 565  
Krishnappa Veena Divya, 25  
Kumar, V., 225

## L

Lida Barba, 177

## M

Mahaboob Hussain, S., 541  
Mahalakshmi, G. S., 281

Mainejar Yadav, 425  
 Mamata Rath, 349  
 Manimegalai Rajkumar, 203  
 Manish Kumar Singh, 451  
 Manish Patel, 523  
 Mohd. Mohsin, 565  
 Mohd. Mudasir, 565  
 Monika Pandey, 393  
 Mukhopadhyay, S., 37  
 Muttoo Mudasir Ahmed, 565  
 Muzzammil Hussain, 341

**N**

Naresh Kumar, 437  
 Naveen Aggarwal, 341  
 Neha Shashni, 425  
 Nibaldo Rodriguez, 177  
 Nikita Khanna, 119  
 Niranjan, A., 403  
 Nirbhay Chaubey, 523  
 Nitin Trivedi, 573

**P**

Padmaja, K. V., 359  
 Panda, P. C., 505  
 Pandiselvam Pandiyarajan, 97  
 Pankaj Kumar Jadwal, 189  
 Pati, R., 225  
 Phukan, N., 61  
 Pooja Kherwa, 237  
 Pooja Sharma, 85  
 Pooja, R., 403  
 Poonam Bansal, 237  
 Prashant Khanna, 189  
 Prateek Bajaj, 393  
 Prathyusha Kanakam, 541  
 Preeti Joon, 13  
 Pritam Kumari, 381  
 Pujari, A. K., 225  
 Pushkal Agarwal, 531

**R**

Rafi, M., 37  
 Rahul Shrivastava, 129  
 Rajasree, P. M., 3  
 Rajeev Chatterjee, 167  
 Rajesh Kambattan Kovarasan, 203  
 Rajneesh Rani, 381, 415  
 Rani, R. U., 293  
 Ranvijay, 425  
 Rashmi Jakhmola, 415  
 Ravi Saharan, 49, 329

Reshmi, T. R., 371  
 Revan Joshi, P., 25  
 Rhea Sanjay Sukthanker, 551  
 Richa Sharma, 593  
 Ritesh Dash, 505

**S**

Sagayaraj Fancis, F., 257  
 Sai Prasad Potharaju, 215  
 Saksham., 119  
 Sandeep Kaur, 493  
 Sanhita Mishra, 505  
 Sarah, S., 403  
 Saravanakumar, K., 551  
 Sarbani Sen, 513  
 Sarimela, V., 317  
 Sarthak Kanodia, 119  
 Sendhilkumar, S., 281  
 Shalini Bhaskar Bajaj, 13  
 Shanmugavadvu, P., 75  
 Shashank Uniyal, 593  
 Shikha Bharti, 451  
 Shiv Narayan, 493  
 Shivam Singhal, 247  
 Shivangi Verma, 463  
 Shubam Kumar, 317  
 Shubham Upadhyaya, 531  
 Sonal Jain, 189  
 Sreedevi, M., 215  
 Srilekha, K. N., 581  
 Sudeep Tanwar, 463  
 Sudhakar Tripathi, 129  
 Sudhanshu Tyagi, 463  
 Sumit Kumar, 317  
 Suryanarayana, D., 541  
 Swain, S. C., 505  
 Swathi, S., 581  
 Sweta Singh, 303

**T**

Tamasi Moyra, 475, 513  
 Taranpreet Singh Ruprah, 573  
 Thangamani, M., 139  
 Thekkekara Joel Philip, 109

**U**

Umesh Gupta, 189

**V**

Vaishali Dabral, 85  
 Vaishali Gera, 593  
 Vasantha Kumar, V., 281

Vasanthakumar, G. U., [581](#)  
Venugopal, K. R., [403](#), [581](#)  
Vidya, M. J., [359](#)

Vijay Paul Singh, [341](#)  
Vikas Tripathi, [85](#), [247](#), [393](#)  
Vishal Sanserwal, [393](#)