

Bidyut B. Chaudhuri  
Masaki Nakagawa  
Pritee Khanna  
Sanjeev Kumar *Editors*

# Proceedings of 3rd International Conference on Computer Vision and Image Processing

CVIP 2018, Volume 2

# **Advances in Intelligent Systems and Computing**

## **Volume 1024**

### **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

### **Advisory Editors**

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,  
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering,  
University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,  
Gyor, Hungary

Vladik Kreinovich, Department of Computer Science, University of Texas  
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao  
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,  
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute  
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,  
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen, Faculty of Computer Science and Management,  
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,  
The Chinese University of Hong Kong, Shatin, Hong Kong

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.

**\*\* Indexing: The books of this series are submitted to ISI Proceedings, EI-Compendex, DBLP, SCOPUS, Google Scholar and Springerlink \*\***

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

Bidyut B. Chaudhuri · Masaki Nakagawa ·  
Pritee Khanna · Sanjeev Kumar  
Editors

# Proceedings of 3rd International Conference on Computer Vision and Image Processing

CVIP 2018, Volume 2



Springer

*Editors*

Bidyut B. Chaudhuri  
Techno India University  
Kolkata, India

Pritee Khanna  
Department of Computer Science  
Indian Institute of Information Technology,  
Design and Manufacturing  
Jabalpur, Madhya Pradesh, India

Masaki Nakagawa  
Division of Advanced Information  
Technology and Computer Science  
Tokyo University of Agriculture  
and Technology  
Koganei-shi, Tokyo, Japan

Sanjeev Kumar  
Department of Mathematics  
Indian Institute of Technology Roorkee  
Roorkee, Uttarakhand, India

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-32-9290-1

ISBN 978-981-32-9291-8 (eBook)

<https://doi.org/10.1007/978-981-32-9291-8>

© Springer Nature Singapore Pte Ltd. 2020

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, expressed 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

The Third International Conference on Computer Vision and Image Processing (CVIP 2018) was organized at PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur (IIITDMJ), during September 29–October 1, 2018. The conference was endorsed by the International Association of Pattern Recognition (IAPR) and co-sponsored by BrahMos, Council of Scientific and Industrial Research (CSIR), Defence Research and Development Organisation (DRDO), Indian Space Research Organisation (ISRO), MathWorks, and Science and Engineering Research Board (SERB), India.

The general theme of the conference being Computer Vision and Image Processing, papers on Image Registration, Reconstruction, and Retrieval; Object Detection, Recognition, Classification, and Clustering; Biometrics, Security, Surveillance, and Tracking; and Deep Learning Methods and Applications, as well as Medical Image Processing and Remote Sensing, were submitted from India and abroad. Out of the 206 submitted papers, 81 were accepted for presentation after peer review, making the acceptance rate of 39%. Among those, 47 were oral and 34 posters. Out of 81 papers, 76 papers were presented at the conference. These papers were distributed over eight oral and four poster sessions.

In addition to the contributory paper sessions, the conference included plenary talks of Prof. Masaki Nakagawa (Tokyo Institute of Agriculture and Technology, Japan), Prof. Venu Govindaraju (SUNY, Buffalo, USA), and Prof. Hironobu Fujiyoshi (Chubu University, Japan). Besides, Dr. R. Venkatesh Babu (IISc, Bangalore) and Ms. Ramya Hebbalaguppe (TCS Innovation Labs, New Delhi, India) delivered the invited talks. The talks ranged from historical epochs of Artificial Intelligence Research to theoretical- and application-based works on Deep Learning. The talks enriched and expanded the knowledge horizon of the researchers who attended the conference.

Besides these, a coding challenge involving supervised classification of bird species from a set of bird images was organized, and the works of the respondents were judged by a committee. A pre-conference workshop was organized by Mathworks on Deep Learning for Computer Vision Applications. Like the previous year, the best oral and poster papers were selected by a panel of experts. The last

session of the conference unfolded the Best Paper Award, the Best Student Paper Award, and the Best Poster Award along with the declaration of the names of the Coding Challenge winners and the venue and organizing institute name of the next CVIP conference (2019).

Overall, the conference was a grand success. Young researchers were immensely benefitted by interacting with academic and industry experts. Like previous years, the proceedings of this conference are also compiled in this edited volume brought out by Springer Nature under their series of *Advances in Intelligent Systems and Computing*.

The success of such an event was due to harmonious contributions of various stakeholders including the members of the international advisory committee, the technical program committee, the plenary and invited speakers, the local organizing committee, the sponsors, the endorser and the researchers who attended the conference. Our sincere thanks to all of them! Last but not least, warm thanks are due to Springer for printing and publishing these proceedings in such a beautiful form.

Kolkata, India

Koganei-shi, Japan

Jabalpur, India

Roorkee, India

Bidyut B. Chaudhuri

Masaki Nakagawa

Pratee Khanna

Sanjeev Kumar

# Contents

<b>Descriptor-Length Reduction Using Low-Variance Filter for Visual Odometry .....</b>	1
Shrijay S. Kalambe, Elizabeth Rufus, Vinod Karar and Shashi Poddar	
<b>Remote Multimodal Biometric Authentication using Visual Cryptography .....</b>	13
Harkeerat Kaur and Pritee Khanna	
<b>Caption-Based Region Extraction in Images .....</b>	27
Palash Agrawal, Rahul Yadav, Vikas Yadav, Kanjar De and Partha Pratim Roy	
<b>Facial Expression Recognition Using Improved Adaptive Local Ternary Pattern .....</b>	39
Sumeet Saurav, Sanjay Singh, Ravi Saini and Madhulika Yadav	
<b>Cell Extraction and Horizontal-Scale Correction in Structured Documents .....</b>	53
Divya Srivastava and Gaurav Harit	
<b>DeepAttent: Saliency Prediction with Deep Multi-scale Residual Network .....</b>	65
Kshitij Dwivedi, Nitin Singh, Sabari R. Shanmugham and Manoj Kumar	
<b>Copy-Move Image Forgery Detection Using Gray-Tones with Texture Description .....</b>	75
Anuja Dixit and Soumen Bag	
<b>Object Labeling in 3D from Multi-view Scenes Using Gaussian-Hermite Moment-Based Depth Map .....</b>	87
Sadman Sakib Enan, S. M. Mahbubur Rahman, Samiul Haque, Tamanna Howlader and Dimitrios Hatzinakos	

<b>Zernike Moment and Mutual Information Based Methods for Multimodal Image Registration .....</b>	101
Suraj Kumar Kashyap, Dinesh Jat, M. K. Bhuyan, Amit Vishwakarma and Prathik Gadde	
<b>A Novel Deep Learning Approach for the Removal of Speckle Noise from Optical Coherence Tomography Images Using Gated Convolution–Deconvolution Structure .....</b>	115
Sandeep N. Menon, V. B. Vineeth Reddy, A. Yeshwanth, B. N. Anoop and Jeny Rajan	
<b>D<sup>2</sup>ehazing: Real-Time Dehazing in Traffic Video Analytics by Fast Dynamic Bilateral Filtering .....</b>	127
Apurba Das, Shashidhar Pai, Vinayak S. Shenoy, Tanush Vinay and S. S. Shylaja	
<b>Joint Bit Allocation for 3D Video with Nonlinear Depth Distortion—An SSIM-Based Approach .....</b>	139
Y. Harshalatha and Prabir Kumar Biswas	
<b>A Modified FCM-Based Brain Lesion Segmentation Scheme for Medical Images .....</b>	149
Anjali Gautam, Debanjan Sadhya and Balasubramanian Raman	
<b>Word Spotting in Cluttered Environment .....</b>	161
Divya Srivastava and Gaurav Harit	
<b>Physical Intrusion Detection System Using Stereo Video Analytics .....</b>	173
G. Aravamuthan, P. Rajasekhar, R. K. Verma, S. V. Shrikhande, S. Kar and Suresh Babu	
<b>Identification of Fraudulent Alteration by Similar Pen Ink in Handwritten Bank Cheque .....</b>	183
Priyanka Roy and Soumen Bag	
<b>Faster RCNN-CNN-Based Joint Model for Bird Part Localization in Images .....</b>	197
Arjun Pankajakshan and Arnav Bhavsar	
<b>Structural Analysis of Offline Handwritten Mathematical Expressions .....</b>	213
Ridhi Aggarwal, Gaurav Harit and Anil Kumar Tiwari	
<b>L1-Regulated Feature Selection and Classification of Microarray Cancer Data Using Deep Learning .....</b>	227
B. H. Shekar and Guesh Dagnew	
<b>Image Embedding for Detecting Irregularity .....</b>	243
M. K. Sharma, D. Sheet and Prabir Kumar Biswas	

<b>Optimal Number of Seed Point Selection Algorithm of Unknown Dataset . . . . .</b>	257
Kunal Chowdhury, Debasis Chaudhuri and Arup Kumar Pal	
<b>TexFusionNet: An Ensemble of Deep CNN Feature for Texture Classification . . . . .</b>	271
Swalpa Kumar Roy, Shiv Ram Dubey, Bhabatosh Chanda, Bidyut B. Chaudhuri and Dipak Kumar Ghosh	
<b>Person Identification Using Footprint Minutiae . . . . .</b>	285
Riti Kushwaha and Neeta Nain	
<b>Target Tracking Based Upon Dominant Orientation Template and Kalman Filter . . . . .</b>	301
Nikhil Kumar, Puran Dhakrey and Neeta Kandpal	
<b>Robust Single Image Super Resolution Employing ADMM with Plug-and-Play Prior . . . . .</b>	313
V. Abdu Rahiman and Sudhish N. George	
<b>Indoor–Outdoor Scene Classification with Residual Convolutional Neural Network . . . . .</b>	325
Seema Kumari, Ranjeet Ranjan Jha, Arnav Bhavsar and Aditya Nigam	
<b>Comparison Between LGBP and DCLBP for Non-frontal Emotion Recognition . . . . .</b>	339
Hardik Dosi, Rahul Keshri, Pravin Srivastav and Anupam Agrawal	
<b>Co-Detection in Images Using Saliency and Siamese Networks . . . . .</b>	351
Milan Zinzuvadiya, Vatsalkumar Dhameliya, Sanjay Vaghela, Sahil Patki, Nirali Nanavati and Arnav Bhavsar	
<b>Hand Down, Face Up: Innovative Mobile Attendance System Using Face Recognition Deep Learning . . . . .</b>	363
Aditi Agrawal, Mahak Garg, Surya Prakash, Piyush Joshi and Akhilesh M. Srivastava	
<b>Trajectory Classification Using Feature Selection by Genetic Algorithm . . . . .</b>	377
Rajkumar Saini, Pradeep Kumar, Partha Pratim Roy and Umapada Pal	
<b>Action Recognition from Egocentric Videos Using Random Walks . . . . .</b>	389
Abhimanyu Sahu, Rajit Bhattacharya, Pallabh Bhura and Ananda S. Chowdhury	
<b>A Bag of Constrained Visual Words Model for Image Representation . . . . .</b>	403
Anindita Mukherjee, Jaya Sil and Ananda S. Chowdhury	

<b>Activity Recognition for Indoor Fall Detection in 360-Degree Videos Using Deep Learning Techniques . . . . .</b>	417
Dhiraj, Raunak Manekar, Sumeet Saurav, Somsukla Maiti, Sanjay Singh, Santanu Chaudhury, Neeraj, Ravi Kumar and Kamal Chaudhary	
<b>A Robust Watermarking Scheme for Copyright Protection . . . . .</b>	431
Satendra Pal Singh and Gaurav Bhatnagar	
<b>Multi-scale Image Fusion Scheme Based on Gray-Level Edge Maps and Adaptive Weight Maps . . . . .</b>	445
Jitesh Pradhan, Ankesh Raj, Arup Kumar Pal and Haider Banka	
<b>Comparison of Reconstruction Methods for Multi-compartmental Model in Diffusion Tensor Imaging . . . . .</b>	461
Snehlata Shakya and Sanjeev Kumar	
<b>Design of Finite Impulse Response Filter with Controlled Ripple Using Cuckoo Search Algorithm . . . . .</b>	471
Anil Kumar, N. Agrawal and I. Sharma	
<b>Robustly Clipped Sub-equalized Histogram-Based Cosine-Transformed Energy-Redistributed Gamma Correction for Satellite Image Enhancement . . . . .</b>	483
Himanshu Singh, Anil Kumar and L. K. Balyan	
<b>Author Index . . . . .</b>	497

# Editors and Contributors

## About the Editors

**Bidyut B. Chaudhuri** is currently Pro Vice Chancellor of Techno India University, Salt Lake, Calcutta, India. Previously he was INAE Distinguished Professor at Indian Statistical Institute, Calcutta. He received his B.Sc. (Hons), B.Tech, and M.Tech. degrees from Calcutta University, India, and his Ph.D. from the Indian Institute of Technology Kanpur, in 1980. He did his postdoc work as Leverhulme fellow at Queen's University, UK and acted as a visiting faculty at the Technical University, Hannover, Germany. His main research interests are in Pattern Recognition, Image Processing, Language processing and Machine learning in which he has published 450 research Papers and five books. He is a life fellow of IEEE, IAPR, TWAS as well as fellow of Indian Academies like INAE, INS, INASc. He has received many awards for his research work. Prof. Chaudhuri is now an Associate Editor of the International Journal of Document Analysis and Recognition (IJDAR), International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI). In the past he worked in such capacity in several other international journals.

**Masaki Nakagawa** is a Professor of Media Interaction at the Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Japan. He graduated from the University of Tokyo in March 1977 and pursued an M.Sc. course in Computer Studies at Essex University, England, sponsored by the Japanese Government. In March 1979, he graduated from the University of Tokyo with an M.Sc. in Physics. In July 1979, he completed his M.Sc. in Computer Studies at Essex University in England, and in December 1988, he completed his Ph.D. at the University of Tokyo. His work chiefly focuses on handwriting recognition and pen-based user interfaces and applications, especially educational applications. Prof. Nakagawa has over 300 publications to his credit.

**Pratee Khanna** is an Associate Professor and Head of the Computer Science & Engineering Discipline, PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur. Her main areas of interest include Biometrics, Biomedical Image Processing, Image Retrieval and Indexing, Dynamic Gesture Recognition, and Computer Aided Product Design. She is a recipient of UGC Fellowship, India and Long Term JSPS Fellowship, Japan. She is a senior member of the IEEE Computer Society and a life member of IAENG. She has published more than 90 papers in journals and conference proceedings.

**Sanjeev Kumar** is an Associate Professor of Mathematics at the IIT Roorkee, India. His areas of interest include Computer Vision & Mathematical Imaging, Inverse Problems, and Machine Learning. He completed his Ph.D. in Mathematics at the IIT Roorkee in 2008. He is a member of the IEEE Computer Society and International Association of Pattern Recognition, and a life member of the ACEEE and IACSIT. He has published over 40 papers in journals and conference proceedings.

## Contributors

**V. Abdu Rahiman** Government College of Engineering Kannur, Kannur, India

**Ridhi Aggarwal** Indian Institute of Technology Jodhpur, Jodhpur, India

**Aditi Agrawal** Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India

**Anupam Agrawal** Interactive Technologies and Multimedia Research Lab, Indian Institute of Information Technology, Allahabad, India

**N. Agrawal** PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Madhya Pradesh, India

**Palash Agrawal** Indian Institute of Technology Roorkee, Roorkee, India

**B. N. Anoop** Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

**G. Aravamuthan** Homi Bhabha National Institute (HBNI), Mumbai, India

**Suresh Babu** EISD, Bhabha Atomic Research Centre, Mumbai, India

**Soumen Bag** Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand, India

**L. K. Balyan** Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India

**Haider Banka** Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, India

**Gaurav Bhatnagar** Department of Mathematics, Indian Institute of Technology Karwar, Karwar, India

**Rajit Bhattacharya** Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

**Arnav Bhavsar** School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India;

MAS Lab, School of Computing and Electrical Engineering, IIT Mandi, Mandi, India

**Pallabh Bhura** Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

**M. K. Bhuyan** Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Prabir Kumar Biswas** Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India

**Bhabatosh Chanda** Indian Statistical Institute, Kolkata, India

**Kamal Chaudhary** Samsung Research India, New Delhi, India

**Bidyut B. Chaudhuri** Indian Statistical Institute, Kolkata, India

**Debasis Chaudhuri** DRDO Integration Centre, Panagarh, West Bengal, India

**Santanu Chaudhury** CSIR-Central Electronics Engineering Research Institute, Pilani, India

**Ananda S. Chowdhury** Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

**Kuntal Chowdhury** Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), IIT(ISM), Dhanbad, Jharkhand, India

**Guesh Dagnew** Department of Computer Science, Mangalore University, Mangalore, India

**Apurba Das** PES University, Bengaluru, India

**Kanjar De** Indian Institute of Technology Roorkee, Roorkee, India

**Puran Dhakrey** Instruments Research and Development Establishment, Defence Research and Development Organization, Dehradun, India

**Vatsalkumar Dhameliya** Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

**Dhiraj** CSIR-Central Electronics Engineering Research Institute, Pilani, India

**Anuja Dixit** Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, India

**Hardik Dosi** Interactive Technologies and Multimedia Research Lab, Indian Institute of Information Technology, Allahabad, India

**Shiv Ram Dubey** Indian Institute of Information Technology, Sri City, Andhra Pradesh, India

**Kshitij Dwivedi** Singapore University of Technology and Design, Singapore, Singapore

**Sadman Sakib Enan** Department of EEE, BUET, Dhaka, Bangladesh

**Prathik Gadde** School of Informatics and Computing, Indiana University, Purdue University, Indianapolis, IN, USA

**Mahak Garg** Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India

**Anjali Gautam** Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

**Sudhish N. George** National Institute of Technology Calicut, Kozhikode, India

**Dipak Kumar Ghosh** Adamas University, Kolkata, India

**Samiul Haque** Department of ECE, North Carolina State University, Raleigh, NC, USA

**Gaurav Harit** Indian Institute of Technology Jodhpur, Jodhpur, India

**Y. Harshalatha** Indian Institute of Technology Kharagpur, Kharagpur, India

**Dimitrios Hatzinakos** Department of ECE, University of Toronto, Toronto, ON, Canada

**Tamanna Howlader** ISRT, University of Dhaka, Dhaka, Bangladesh

**Dinesh Jat** Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Ranjeet Ranjan Jha** MAS Lab, School of Computing and Electrical Engineering, IIT Mandi, Mandi, India

**Piyush Joshi** Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India

**Shrijay S. Kalambé** Vellore Institute of Technology, Vellore, Tamil Nadu, India

**Neeta Kandpal** Instruments Research and Development Establishment, Defence Research and Development Organization, Dehradun, India

**S. Kar** Homi Bhabha National Institute (HBNI), Mumbai, India

**Vinod Karar** CSIR-Central Scientific Instruments Organization, Chandigarh, India

**Suraj Kumar Kashyap** Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Harkeerat Kaur** PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Madhya Pradesh, India

**Rahul Keshri** Interactive Technologies & Multimedia Research Lab, Indian Institute of Information Technology, Allahabad, India

**Pritee Khanna** PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Madhya Pradesh, India

**Anil Kumar** Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India

**Manoj Kumar** Samsung Research Institute, Bengaluru, India

**Nikhil Kumar** Instruments Research and Development Establishment, Defence Research and Development Organization, Dehradun, India

**Pradeep Kumar** IIT Roorkee, Roorkee, India

**Ravi Kumar** Samsung Research India, New Delhi, India

**Sanjeev Kumar** Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

**Seema Kumari** MAS Lab, School of Computing and Electrical Engineering, IIT Mandi, Mandi, India

**Riti Kushwaha** Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India

**S. M. Mahbubur Rahman** Department of EEE, BUET, Dhaka, Bangladesh

**Somsukla Maiti** CSIR-Central Electronics Engineering Research Institute, Pilani, Pilani, India

**Raunak Manekar** CSIR-Central Electronics Engineering Research Institute, Pilani, Pilani, India

**Sandeep N. Menon** Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

**Anindita Mukherjee** Dream Institute of Technology, Kolkata, India

**Neeta Nain** Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India

**Nirali Nanavati** Indian Institute of Technology Mandi, Mandi, India

**Neeraj** Samsung Research India, New Delhi, India

**Aditya Nigam** MAS Lab, School of Computing and Electrical Engineering, IIT Mandi, Mandi, India

**Shashidhar Pai** PESIT, Bengaluru, India

**Arup Kumar Pal** Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), IIT(ISM), Dhanbad, Jharkhand, India

**Umapada Pal** ISI Kolkata, Kolkata, India

**Arjun Pankajakshan** School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India

**Sahil Patki** Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

**Shashi Poddar** CSIR-Central Scientific Instruments Organization, Chandigarh, India

**Jitesh Pradhan** Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, India

**Surya Prakash** Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India

**Partha Pratim Roy** Indian Institute of Technology Roorkee, Roorkee, India

**Ankesh Raj** Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, India

**Jeny Rajan** Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

**P. Rajasekhar** EISD, Bhabha Atomic Research Centre, Mumbai, India

**Balasubramanian Raman** Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

**Partha Pratim Roy** IIT Roorkee, Roorkee, India

**Priyanka Roy** Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand, India

**Swalpa Kumar Roy** Jalpaiguri Government Engineering College, Jalpaiguri, India

**Elizabeth Rufus** Vellore Institute of Technology, Vellore, Tamil Nadu, India

**Debanjan Sadhya** Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

**Abhimanyu Sahu** Department of Electronics and Telecommunication Engineering, Jadavpur University, Kolkata, India

**Rajkumar Saini** IIT Roorkee, Roorkee, India

**Ravi Saini** Academy of Scientific and Innovative Research (AcSIR), Chennai, India;

CSIR-Central Electronics Engineering Research Institute, Pilani, India

**Sumeet Saurav** Academy of Scientific and Innovative Research (AcSIR), Chennai, India;

CSIR-Central Electronics Engineering Research Institute, Pilani, Pilani, India

**Snehlata Shakya** Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India

**Sabari R. Shanmugham** DataRobot, Singapore, Singapore

**I. Sharma** PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Madhya Pradesh, India

**M. K. Sharma** Advanced Technology Development Centre, IIT Kharagpur, Kharagpur, India

**D. Sheet** Electrical Engineering, IIT Kharagpur, Kharagpur, India

**B. H. Shekar** Department of Computer Science, Mangalore University, Mangalore, India

**Vinayak S. Shenoy** PESIT, Bengaluru, India

**S. V. Shrikhande** Homi Bhabha National Institute (HBNI), Mumbai, India

**S. S. Shylaja** PES University, Bengaluru, India

**Jaya Sil** IIEST Sibpur, Howrah, India

**Himanshu Singh** Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India

**Nitin Singh** Samsung Research Institute, Bengaluru, India

**Sanjay Singh** Academy of Scientific and Innovative Research (AcSIR), Chennai, India;

CSIR-Central Electronics Engineering Research Institute, Pilani, Pilani, India

**Satendra Pal Singh** Department of Mathematics, Indian Institute of Technology Karwar, Karwar, India

**Akhilesh M. Srivastava** Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, Indore, India

**Divya Srivastava** Indian Institute of Technology Jodhpur, Jodhpur, India

**Pravin Srivastav** Interactive Technologies and Multimedia Research Lab, Indian Institute of Information Technology, Allahabad, India

**Anil Kumar Tiwari** Indian Institute of Technology Jodhpur, Jodhpur, India

**Sanjay Vaghela** Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

**R. K. Verma** EISD, Bhabha Atomic Research Centre, Mumbai, India

**Tanush Vinay** PESIT, Bengaluru, India

**V. B. Vineeth Reddy** Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

**Amit Vishwakarma** Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

**Madhulika Yadav** Department of Electronics, Banasthali Vidyapith, Vidyapith, Rajasthan, India

**Rahul Yadav** Indian Institute of Technology Roorkee, Roorkee, India

**Vikas Yadav** Indian Institute of Technology Roorkee, Roorkee, India

**A. Yeshwanth** Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, India

**Milan Zinzuvadiya** Sarvajanik College of Engineering and Technology, Surat, Gujarat, India

# Descriptor-Length Reduction Using Low-Variance Filter for Visual Odometry



Shrijay S. Kalambe, Elizabeth Rufus, Vinod Karar and Shashi Poddar

**Abstract** Visual odometry is a popular technique used to estimate motion in GPS-challenged environment, whose accuracy depends on the features extracted from the images. In past attempts to improved feature distinctiveness, these features have become complex and lengthier, requiring more storage space and computational power for matching. In this paper, an attempt is made toward reducing the length of these feature descriptors while maintaining a similar accuracy in pose estimation. Elimination of feature indices based on variance analysis on feature column sets is proposed and experimented in this paper. The features with reduced descriptor length are applied over the 3D-2D visual odometry pipeline and experimented on KITTI dataset for evaluating its efficacy. The proposed scheme of variance-based descriptor length reduction is found to reduce the overall time taken by the motion estimation framework while estimating the transformation with similar accuracy as that with full-length feature vector.

## 1 Introduction

Visual Odometry (VO) is the science of finding the current location of an agent with the help of images captured by the camera mounted on the agent. Recently, VO has drawn attention of researchers all over the globe with the rising investment of huge capital in such navigation systems by automobile manufacturers. These conglomerates are trying to implement automatic driver assistance system (ADAS) which will provide autonomy and increase the seating capacity in the vehicles. Although GPS is a well-developed technology, their signals are very weak in densely populated urban structure, under dense trees, inside buildings, underwater, etc. The growing bias errors in inertial sensors necessitates external aiding to these systems for which

---

S. S. Kalambe · E. Rufus  
Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India

V. Karar · S. Poddar (✉)  
CSIR-Central Scientific Instruments Organization, Sector 30, Chandigarh, India  
e-mail: [shashipoddar@csio.res.in](mailto:shashipoddar@csio.res.in)

GPS is used as an external corrector in a periodic manner. However, with GPS outages for long duration or its unavailability leads to the requirement of alternate correction mechanism. Vision-based navigation or visual odometry (VO) is thus a very suitable choice for aiding the inertial navigation or to be used independently for estimating motion. The history of visual odometry dates back to 1980 when Moravec first tested the VO technology on planetary rovers [15]. The term visual odometry was coined by Nister in his landmark paper wherein he mentioned the advantages of VO over traditional wheel odometry [18]. The basic pipeline of the visual Odometry requires capturing image sequence, detecting features and matching them between two consequent image frames, running outlier rejection sub-routine and estimating transformation between these corresponding feature points.

The image sequence fed to the VO scheme can be either from a monocular or a stereo camera. In this work, the stereo camera setup is considered because of its ability to have appropriate scale value. Motion estimation is one of the critical steps in VO pipeline which can either be an absolute orientation or Perspective from n point scheme. It can then be followed by a local optimization scheme such as bundle adjustment to refine. The vector-based descriptors such as SIFT and SURF are very accurate because of their lengthy descriptor dimension and invariance properties. However, these descriptors requires lots of storage space and computation time in the matching process. It is thus required to have compact descriptors which can maintain similar accuracy with reduced feature descriptor length.

Different feature-length reduction techniques have been proposed in the literature such as the usage of principal component analysis (PCA) [9], random projections [21], linear discriminant analysis [23], partial least squares [14]. These schemes are compared in [8] and cites the advantage of feature-length reduction technique in reduced memory requirement, computational time and even improved accuracy at times. Zhou et al. proposed a multidimensional scaling based scheme to reduce the feature descriptor length similar to PCA [24]. An attempt is also made to reduce the dimension of scale invariant feature transform (SIFT) vector with the help of neural network structure called autoencoder [10]. Chandrasekhar et al. proposed a transform coding approach to compress descriptors from floating-point representation to 2 bits/dimension [3]. Dubey et al. presented a dimensionality reduction technique for intensity order based descriptor by considering neighbors as sets of interleaved neighbor, constraining the dimensionality length with increasing sample points [4]. In this paper, the length of speeded up robust feature (SURF) descriptor is reduced with the help of variance-based analysis on feature columns and is explained in detail in the following section.

Several other improvements have been carried out over the traditional VO scheme such as bucketing [11], feature selection, new constraints that can reduce outliers [12], pose refinement strategies, improved outlier rejection schemes [16, 20], specific feature detectors, machine learning techniques [22], etc. [19]. However, to the best of our knowledge, the usage of this kind of reduced feature-length descriptors for vision based motion estimation task does not exist and is thus attempted here in this work. The feature descriptor length is reduced by pruning indices on the basis of variance analysis on feature columns. The complete paper is organized as follows: Sect. 2

discusses the theoretical background, Sect. 3 presents the proposed methodology of pruning descriptor based on variance information. Section 4 analyzes the reduced feature descriptor for visual Odometry by experimenting on KITTI dataset and finally Sect. 5 concludes the paper.

## 2 Theoretical Background

### 2.1 Feature Detection and Matching

Features are those interest points in an image which can be matched distinctively between two image frames with the help of their properties. These features can be either a corner descriptor, an edge detector, blob detector or a region detector. Among these four classes of detectors, blob descriptors are more popularly used owing to the rotation and scale-invariant properties induced in some of them. Although corner detectors have been used for the computer vision applications, they are not as efficient as the recent detectors like SIFT, SURF, BRIEF, BRISK, etc., which are invariant to isometric transformations in an image. Scale-invariant feature transform (SIFT) is one of the landmarks technique that paved the way toward different scale-invariant techniques. These schemes create a descriptor vector for each of the feature point that describes the region around it and provides it the distinctiveness required to match them uniquely between two images.

In this paper, an improved variant of SIFT, that is, speeded up robust feature (SURF) [2] is used for experimenting with its reduced length for feature matching purpose. The SURF feature descriptor uses a window of 20 pixel points around the key-point. This region is subdivided into  $4 \times 4$  subregions, in which the Haar wavelet response in  $X$  and  $Y$  direction is computed to create a 4-dimensional vector for each subregion, building a overall 64 length descriptor. These feature descriptors can be matched through brute force or nearest neighbor search technique. The brute force matcher computes the distance between each feature vector to every other feature vector and the nearest neighbor technique follows a tree-based search to reach the feature point, which is nearer to the query feature. For features with large dimensions, the nearest neighbor technique requires several backtracking scans degrading the linear scan performance. The distance measure used for vector-based descriptors like SIFT and SURF are Manhattan distance or Euclidean distance and Hamming distance for binary descriptors. In this paper, the vector-based SURF descriptor is used and the squared sum of difference (SSD) between two vectors is calculated as matching criteria.

## 2.2 Pose Estimation

Pose estimation is the process of obtaining transformation between corresponding point clouds. The outlier rejection scheme removes the inappropriate matches and improves the motion estimation accuracy. The feature points detected in stereo image pair for previous and current time instant are matched in a circular manner such that the new feature subset is contained in all the four images. The motion estimation can be carried out in 3D–3D, or 3D–2D, or 2D–2D framework. In the 3D–3D framework, the corresponding matched image points are triangulated to yield corresponding 3D point clusters between previous and current time instant,  $X_k$  and  $X_{k-1}$ , respectively. These 3D points are used to obtain transformation,  $T_k$  by minimizing the equation:

$$\text{Error} = \arg \min_{T_k} \sum_i \|X_k^i - T_k(X_{k-1}^i)\| \quad (1)$$

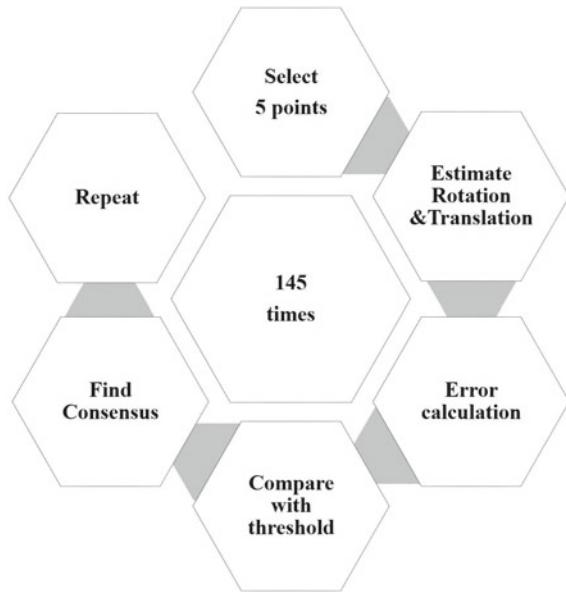
Here,  $k$  denotes the time instant, and superscript  $i$  denotes the  $i$ th feature point. The solution proposed by Arun et al. [1], which computes translation as the difference of the 3D point cluster centroid and rotation using SVD, is very commonly used. The 3D–2D motion estimation framework obtains transformation by minimizing the image re-projection error rather than 3D–3D position error and is more accurate [17]. Consider the re-projection of 3D point in previous frame  $X_k$  to the current frame  $I_k$  by a transformation  $T_k$  as  $p_{k-1}$  and the corresponding image point in current frame as  $p_k$ , the transformation can be obtained by minimizing the image re-projection error given as

$$\text{Re-projection Error} = \sum_{i=0}^N \|p_{k-1}^i - \pi_1(P^i)\|^2 + \|p_k^i - \pi_2(P^i, R, T)\|^2 \quad (2)$$

This problem is also called as PnP problem, which is perspective from n point problem (PnP) [17]. Although several solutions were proposed for obtaining this transformation, the direct linear transformation algorithm is the simplest and most commonly used [6]. The pose estimated using either of the schemes is then concatenated at each time instant to yield the current pose with respect to the starting frame of reference.

## 2.3 RANSAC

The corresponding feature points obtained by applying feature detection and matching step is then used for estimating transformation between two time instants. However, owing to the inaccuracies in the feature matching and noises in the image, some outliers creep into the process which needs to be removed. Random Sampling and Consensus (RANSAC) [5] is one of the most popular techniques used for rejecting outliers and is used here. Inliers are the points which hold key information about the

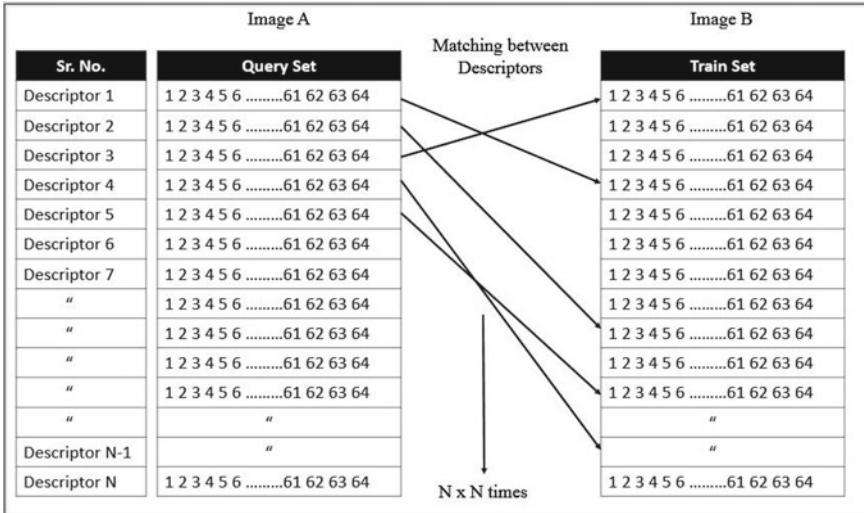
**Fig. 1** RANSAC scheme

model whereas the points, which are vague and don't hold any information are known as outliers. Figure 1 depicts the key steps of 5-point RANSAC scheme incorporated in this paper for estimation motion.

### 3 Proposed Method

Visual odometry is one of the core navigation frameworks that help in estimating motion using image frames. Feature detection and matching is a very critical step that determines the accuracy and computational complexity of the software architecture. Although corner and edge detectors have shown good performance in the past, they are invariant to scale and rotation invariance. In 2004, David Lowe [13] proposed a scale-invariant feature transform (SIFT) technique to detect and describe a feature which can be used to track a key point over several image frames accurately. Over time, several other feature detectors and descriptors were proposed of which SURF, BRIEF, ORB, and BRISK are some of the popular techniques. The feature descriptors used to describe the local patch around a keypoint help in providing uniqueness for its accurate matching in a transformed image. These floating-point lengthy descriptors require large storage space and requires relatively larger time in matching one feature to the other. In this paper, an attempt has been made towards reducing the length of these feature vectors with the help of variance analysis between these features.

Figure 2 depicts the brute force matching process in which each feature from image 1 is checked with all the other features from image 2. As seen, this process takes

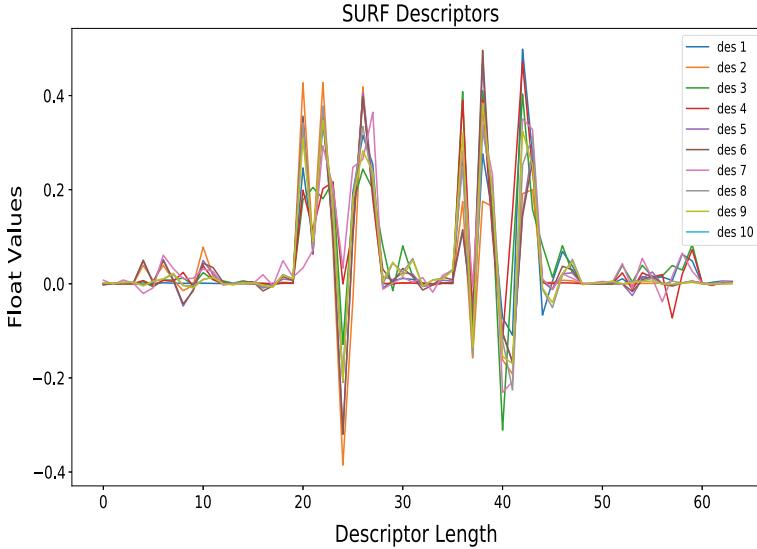


**Fig. 2** Matching between SURF descriptors

a large time in comparing each feature vector of  $N$  length with the other features of  $N$  length. Although several faster matching processes have been proposed in the literature, the vector length of these feature descriptors is a major bottleneck in reducing the feature matching time. In an empirical study on different feature vectors from different images, it was found that all vector points do not yield enough information. For example, the SURF feature descriptor plot shown in Fig. 3 for 10 different features clearly depicts low variance at around 2, 15, and 50th vector point than other locations. These descriptors are chosen randomly and do not belong to one specific area of the image and only 10 features have been selected for clarity purpose. As known, entropy is a direct measure of information in any data. This principle is used here to reduce the length of the feature descriptors by considering only those vector points that have high variance and removing the ones which are near to zero. The variance is calculated across each of the vector columns for all the features detected in an image. The vectors columns which yield more variance are retained and the ones with lower variance are removed using a threshold basis chosen empirically.

This scheme reduces the feature length drastically resulting in a faster matching process. Since the attempt is made toward discarding low entropy data, maximum information of a feature is retained with lower feature vector length. This process of removing low variance vector points reduces the feature length while retaining similar feature matching accuracy which will be shown in the following section.

The threshold value chosen empirically for variance is 0.0025 which approximately reduces the feature length from 64 to 15. The variance calculation is not done on individual vectors as that lead to different indexes for different feature points and varying feature lengths. The index calculated for image 1 are used in image 2



**Fig. 3** Few descriptors

to discard both query and train descriptor vector points on a one to one basis. The variance of the feature column sets is calculated as

$$\text{variance} = \frac{\sum_{i=0}^N [x_i - \bar{x}]^2}{N - 1} \quad (3)$$

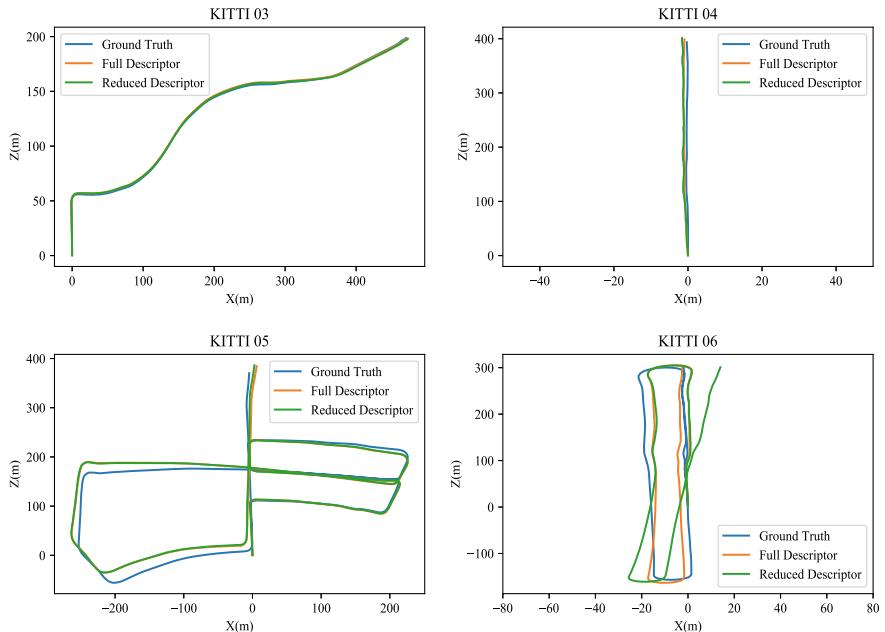
Here,  $x_i$  denotes  $i$ th data element and  $\bar{x}$  is a mean and  $N$  is total no. of elements. The overall algorithm for feature-length reduction usage in visual odometry pipeline is presented in following steps:

- Step 1:** Detect and compute key points, descriptors respectively in stereo image pairs.
- Step 2:** Compute column-wise variance of the feature descriptor set for left images.
- Step 3:** Discard those columns from query set whose variance is below chosen threshold.
- Step 4:** Use same index for discarding the columns from descriptors of right image.
- Step 5:** Match features using reduced feature length.
- Step 6:** Triangulation is performed for transforming matched 2D points into 3D.
- Step 7:** Motion estimation is done by reducing the re-projection error between 2D and corresponding 3D points from previous frame.
- Step 8:** Pose formation and concatenation.

## 4 Results and Discussion

Features detection and matching is one of the most critical step in the visual odometry pipeline that govern the accuracy of the estimated motion. With increasing demands for motion estimation in varied robotics industry, it is not always possible to have adequate resources to store and process lengthy vector-based feature descriptors. Although binary descriptors have been proposed in the literature, they are not always a good choice in low texture surroundings. In order to reduce the storage space and the computational time required in matching vector-based features, this article proposed random rejection and low variance vector indices from the full-length feature set. This process reduces the feature length while retains the matching accuracy required to estimate pose between two instants.

In this paper, the KITTI visual odometry dataset [7] consisting of grayscale stereo image pairs for different paths taken by vehicle in Karlsruhe city is used to compare the performance of full-length feature descriptor based VO to that of the reduced-length descriptor based pose estimation. The experimentation is carried out over 11 KITTI data sequences 0–10 for analysis purpose. Figure 4 depicts the performance of variance-based reduced feature-length descriptor usage in 3D to 2D pose estimation framework. Four different KITTI sequences 03–06 is considered for depicting the variation in performance of estimated trajectory with reduced feature-length as



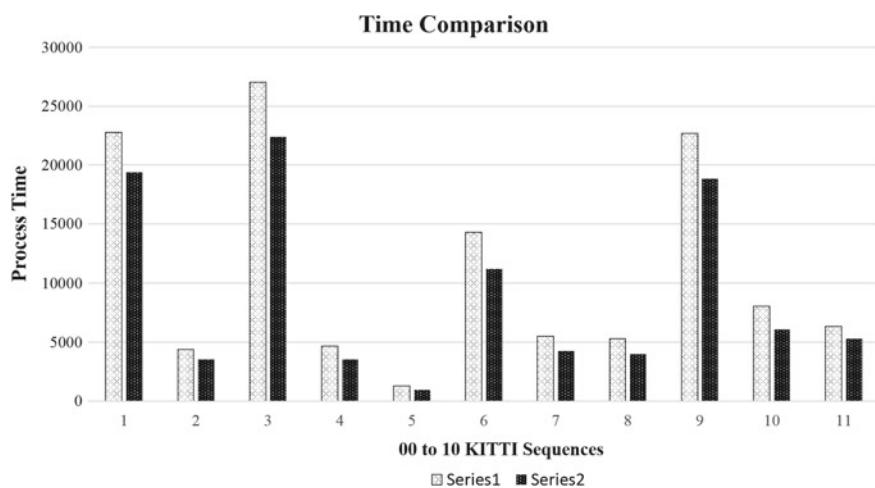
**Fig. 4** Comparison between full and reduced descriptor

compared to full-length descriptor. As seen, the estimated path using reduced-length descriptor is nearly the same as that with the full-length descriptor.

Numerical comparison of VO based on full-length and reduced-length descriptor is provided in Table 1. In this table, rotation and translation error for random index based feature-length reduction of 64-point SURF feature to 32 and 16-points vector is provided. As seen, the random performance of variance based reduced length descriptor performs similar to the normal VO. In order to study the effect in computational time, it is found that the total time taken in VO pipeline reduces by approximately 15%. Figure 5 depicts the bar chart for time taken by all the 11 datasets for VO with

**Table 1** Rotation and translation error value for full descriptor and variance based reduction

Seq No.	Descriptor-64		Variance based	
	$\tilde{t}$ (%)	$\tilde{r}$ ( $^{\circ}$ /m)	$\tilde{t}$ (%)	$\tilde{r}$ ( $^{\circ}$ /m)
0	3.6295	0.0197685	3.655	0.020628
1	56.1868	0.005667	57.1423	0.006824
2	7.2534	0.0405684	8.2087	0.0265299
3	4.3908	0.0194247	4.3647	0.0195393
4	2.1134	0.021774	3.2129	0.0231492
5	3.3129	0.0175911	2.6236	0.0144969
6	2.3383	0.0127206	2.9477	0.0410841
7	2.4682	0.0225762	2.8916	0.0225
8	3.6771	0.019482	5.0813	0.023493
9	4.7388	0.0185652	4.1935	0.0140958
10	3.1522	0.0202269	2.3777	0.0143823



**Fig. 5** Time comparison between the full descriptor and reduced descriptor algorithm

full-length descriptor as compared to variance filter based reduced-length descriptor. The series 1 here refers to the time taken by VO algorithm with full length SURF descriptor and the series 2 with reduced descriptor length.

## 5 Conclusion

Features are a very important aspect of a visual odometry pipeline which determines motion estimation accuracy. In this work, an attempt is made toward reducing the length of feature descriptor vector while maintaining its matching accuracy. Two different approaches have been proposed here, one in which random vector indices are removed from the descriptor and the other in which these vector indices are selected on the basis of feature column set variance. The experimental analysis on KITTI dataset proves the claim of reduced descriptor vector length in its ability to estimate motion with similar accuracy and reduced computational time. In the future, the authors would like to incorporate machine learning techniques that can help in obtaining certain patterns which leads to furthermore reduction in description length while retaining similar matching accuracy.

**Acknowledgements** This research has been supported by DRDO—Aeronautical Research & Development Board through grant-in-aid project on design and development of visual odometry System.

## References

1. Arun, K.S., Huang, T.S., Blostein, S.D.: Least-squares fitting of two 3-D point sets. *IEEE Trans. Pattern Anal. Mach. Intell.* **5**, 698–700 (1987)
2. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: speeded up robust features. In: European Conference on Computer Vision, pp. 404–417. Springer (2006)
3. Chandrasekhar, V., Takacs, G., Chen, D., Tsai, S.S., Singh, J., Girod, B.: Transform coding of image feature descriptors. In: Visual Communications and Image Processing 2009, vol. 7257, p. 725710. International Society for Optics and Photonics (2009)
4. Dubey, S.R., Singh, S.K., Singh, R.K.: Rotation and illumination invariant interleaved intensity order-based local descriptor. *IEEE Trans. Image Process.* **23**(12), 5323–5333 (2014)
5. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. In: Readings in Computer Vision, pp. 726–740. Elsevier (1987)
6. Fraundorfer, F., Scaramuzza, D.: Visual odometry: Part I: The first 30 years and fundamentals. *IEEE Robot. Autom. Mag.* **18**(4), 80–92 (2011)
7. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. *Int. J. Robot. Res.* **32**(11), 1231–1237 (2013)
8. González Valenzuela, R.E., et al.: Linear dimensionality reduction applied to SIFT and SURF feature descriptors (2014)
9. Ke, Y., Sukthankar, R.: PCA-SIFT: a more distinctive representation for local image descriptors. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 2, pp. II–II. IEEE (2004)

10. Keser, R.K., Ergün, E., Töreyin, B.U.: Vehicle logo recognition with reduced dimension SIFT vectors using autoencoders. In: Multidisciplinary Digital Publishing Institute Proceedings, vol. 2, p. 92 (2018)
11. Kitt, B., Geiger, A., Lategahn, H.: Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In: 2010 IEEE Intelligent Vehicles Symposium (IV), pp. 486–492. IEEE (2010)
12. Kottath, R., Yalamandala, D.P., Poddar, S., Bhondekar, A.P., Karar, V.: Inertia constrained visual odometry for navigational applications. In: 2017 Fourth International Conference on Image Information Processing (ICIIP), pp. 1–4. IEEE (2017)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
14. Maitra, S., Yan, J.: Principle component analysis and partial least squares: two dimension reduction techniques for regression. In: Applying Multivariate Statistical Models, vol. 79, pp. 79–90 (2008)
15. Moravec, H.P.: Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, Stanford Univ CA Dept of Computer Science (1980)
16. More, R., Kottath, R., Jegadeeshwaran, R., Kumar, V., Karar, V., Poddar, S.: Improved pose estimation by inlier refinement for visual odometry. In: 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), pp. 224–228. IEEE (2017)
17. Nistér, D.: An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(6), 756–770 (2004)
18. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004, vol. 1, pp. I–I. IEEE (2004)
19. Poddar, S., Kottath, R., Karar, V.: Evolution of visual odometry techniques (2018). [arXiv:1804.11142](https://arxiv.org/abs/1804.11142)
20. Raguram, R., Chum, O., Pollefeys, M., Matas, J., Frahm, J.M.: USAC: a universal framework for random sample consensus. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 2022–2038 (2013)
21. Sulic, V., Perš, J., Kristan, M., Kovacic, S.: Efficient dimensionality reduction using random projection
22. Wang, S., Clark, R., Wen, H., Trigoni, N.: DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 2043–2050. IEEE (2017)
23. Ye, J., Ji, S.: Discriminant analysis for dimensionality reduction: an overview of recent developments. In: *Biometrics: Theory, Methods, and Applications*. Wiley-IEEE Press, New York (2010)
24. Zhou, Z., Cheng, S., Li, Z.: MDS-SIFT: an improved SIFT matching algorithm based on MDS dimensionality reduction. In: 2016 3rd International Conference on Systems and Informatics (ICSAI), pp. 896–900. IEEE (2016)

# Remote Multimodal Biometric Authentication using Visual Cryptography



Harkeerat Kaur and Pritee Khanna

**Abstract** This work proposes an architecture for multimodal biometric recognition systems where user, recognition system, and template database are remotely located over a network. As the number of biometrics are limited and once lost they are compromised forever, it becomes imperative to design systems that optimize recognition rates and also address security and privacy issues for biometric-enabled authentication schemes. The proposed architecture provides revocability to multimodal biometric templates and secures their storage and transmission over a remote network with the help of visual cryptography technique. The proposed architecture gives a good matching performance and also fulfills four template protection criteria, i.e., security, diversity, revocability, and performance. Various attack scenarios such as phishing, replay, database, man-in-middle, and attack via record multiplicity are also addressed.

**Keywords** Multimodal biometric security · Remote authentication · Revocability · Visual cryptography

## 1 Introduction

Biometric authentication systems are one of the most reliable means for providing secure access. Most of the biometric information is stored as digital entities, which are at the risk of theft due to hacking and other malicious activities. Storage and transmission are two important aspects to be considered while securing the template. Unsupervised, remote, and web-based applications provide attackers ample time and opportunities to intercept the system. Cryptography provides a good mechanism for securing biometric templates, but it usually involves complex computations for encryption/decryption of templates. Visual cryptography (VC) techniques are

---

H. Kaur · P. Khanna (✉)

PDPM Indian Institute of Information Technology, Design and Manufacturing,  
Jabalpur, Madhya Pradesh, India  
e-mail: [pkhanna@iitdmj.ac.in](mailto:pkhanna@iitdmj.ac.in)

intuitive for data protection. As a result, some VC-based schemes are proposed for protection of biometric data. These techniques provide security to biometric templates by first decomposing these into meaningless secret shares and then storing them over distributed database servers. This prevents illegal usage of biometric data by hackers or internal administrator. Yet, there does not exist any scheme that offers security to multiple biometric modalities and at the same time fights with attacks common to network-based biometric authentication systems. This work proposes a novel architecture for biometric authentication through VC technique that offers security, privacy, and revocability to security-critical applications. It addresses a remote authentication environment for multiple biometric modalities, where ensuring secure transmission and storage of templates is important. Various attack scenarios are also addressed. The protocols use lightweight computations to support real-time operations and implemented with the two most common modalities, face and fingerprint.

The work is organized as follows. Section 2 gives a basic introduction to VC and existing approaches for biometric privacy. The proposed architecture is discussed in Sect. 3. Experimental results are given in Sect. 4. Security and privacy issues are discussed in Sect. 5. The work is concluded in Sect. 6.

## 2 Visual Cryptography for Biometric Privacy

Naor and Shamir (1994) proposed the basic model of Visual Secret Sharing (VSS) schemes [1]. A  $k$ -out of- $n$  VSS encrypts a secret image  $S$  into  $n$  meaningless shares such that the secret can be revealed by any combination of  $k$  or more shares only. Any VSS scheme is characterized by two parameters, namely, *pixel expansion* and *contrast*. Sharing a pixel into  $n$  shares may require its encoding into more than one sub-pixel leading to increase in size. Contrast refers to the quality of reconstructed image after decryption. Biometric template protection is one of the practical applications of VSS schemes. Monoth and Babu (2010) utilized a simple XOR-based VSS technique to secure storage, transmission, and tampering of biometric image templates in the network [2]. Although simple and easy to implement, the scheme causes loss in contrast of the reconstructed biometric image that results in the loss of discriminative information. Other techniques like random scrambling and block permutations can be applied before encryption to further enhance the security [3]. A significant contribution in biometric template protection is made by Ross and Othman (2011) with the use of Gray-level Extended Visual Cryptography Scheme (GEVCS) [4]. Instead of having a noisy and meaningless appearance shares produced by GEVCS look like natural images. This mitigates the concern of the adversary for the presence of some secret information. In case of a compromise, the existing shares are revoked and new shares can be easily created by changing host images. The technique successfully fulfills biometric template protection criteria of *revocability*, *diversity*, *security*, and *performance* [5].

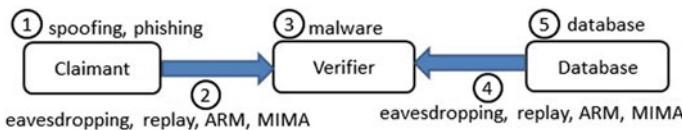
Recursive Visual Cryptography (RVC), another variant of VC techniques, recursively creates and embeds secret image into shares. Takur et al. (2012) used

multiple resolution RVC schemes which allow smaller secrets to be hidden within one large share in recursive manner to encrypt biometric images [6]. However, creating shares using RVC requires complex computations as small shares are recursively encoded into large shares. Patil et al. (2015) proposed a multi-secret sharing method for two fingerprints belonging to the same user using a codebook [7]. The scheme involves pixel expansion, which increases the size of shares and degrades visual quality on reconstruction. Nandakumar et al. (2017) proposed solutions for multi-party transactions using one-time biometric tokens generated with Shamir's secret sharing algorithm and blockchain technology [8].

It can be observed that a limited contribution of VC techniques exists in securing biometric applications. Ross and Othman (2010) are able to provide revocability and diversity to biometric templates using GEVCS technique, but it involves complex computations unsuitable for real-time applications. Also, it suffers from the problem of pixel expansion that results in the loss of biometric information on reconstruction. The proposed research is motivated toward development of an effective framework that fulfills all the important template protection requirements, secures the system at database level, and provides protocols for secure transmission of biometric templates such that intermediate attacks can be overcome. As compared to the previous schemes, the technique used in this work is simple, does not require any codebook, and also allows lossless recovery of secret without any pixel expansion.

### 3 The Proposed Architecture

Remote authentication systems are unattended and possess greater security risks as it is difficult to supervise and practice controls over the claimant and the verifier. Figure 1 shows links between claimant, verifier, and databases are susceptible to various attacks. The adversary manipulates the claimant by communicating as a valid verifier and obtains biometric data fraudulently in phishing attack. The attacker records and resubmits the signal to gain access in replay attack. In attack via record multiplicity (ARM), the attacker obtains multiple copies of template to get user information. It is possible for an adversary to insert himself into communication link and impersonate both parties to gain information in man-in-middle attack (MIMA). Also, the database can be hacked and analyzed to extract personal information for sharing and cross-matching.



**Fig. 1** Vulnerabilities involved in the remote authentication systems

### 3.1 Enrollment Phase

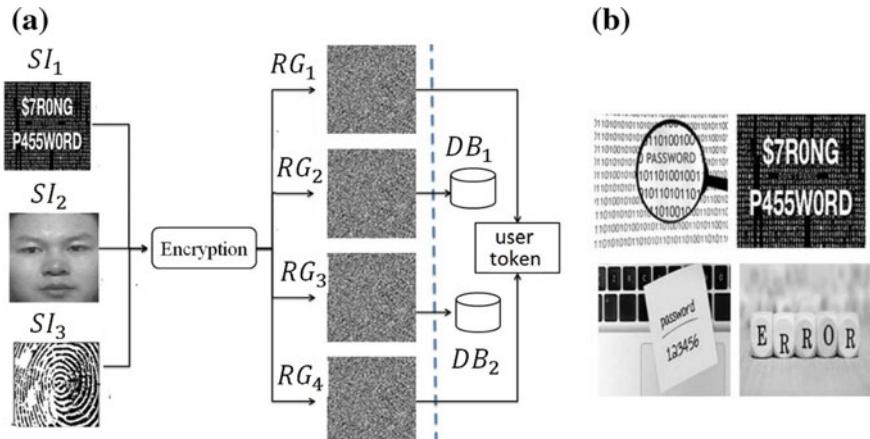
Enrollment module is illustrated in Fig. 2a. A unique identification number  $ID$  is assigned to each enrolled user (known as claimant). For each user, two secret biometric image samples, here face [9] and fingerprint [10] and a random cover image [11] are subjected to VC algorithm in order to generate four meaningless shares. Two of the shares are provided to the user in a tokenized format and the remaining are stored over distributed databases. The acquired samples are processed and encrypted as described below.

**Preprocessing Samples:** As the encryption algorithm requires input images in binary format, grayscale face image samples are subjected to error diffusion halftoning algorithm [12]. Halftoning also gives an illusion of continuous tone to binary images. Fingerprint images are binarized by thresholding.

**Selecting Cover Image:** A cover image is randomly chosen from a public database created with a number of text or password images (Fig. 2b).

**Encryption:** Proposed encryption technique extends shifted Random Grid (RG) technique for three images [13]. A random grid is used to encrypt two images into two shares such that the first secret image is recovered by directly stacking two shares, while the second secret image is recovered by shifting the second share horizontally by a certain width and then stacking it on the first share. The width of the shift can be determined by the user. The second recovered image suffers from visual distortion proportional to the width of shift. The encryption algorithm uses three definitions given below [14].

**Definition 1**  $f_{ran}(\cdot)$  creates a binary cypher grid defined by randomly assigning 0 or 1 with equal probabilities



**Fig. 2** a The proposed architecture—Enrollment phase and b Example cover images

$$f_{ran}(.) = \begin{cases} 0 & \text{with the probability } 1/2 \\ 1 & \text{with the probability } 1/2 \end{cases} \quad (1)$$

For example, let  $X$  be  $2 \times 2$  random grid generated by randomly assigning 0/1 to each of its pixel position as  $X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ .

**Definition 2**  $F_{RG} : Z \leftarrow F_{RG}(X, Y)$  generates second cypher grid pixel  $Z$  from the first cypher grid pixel  $X$  and a secret image  $Y$  pixel, given as

$$F_{RG}(X, Y) = \begin{cases} X, & \text{if } Y = 0 \\ 1 - X, & \text{if } Y = 1 \end{cases} \quad (2)$$

Let  $X$  be the first cypher grid generated above using  $f_{ran}(.)$  and  $Y$  be the secret image  $Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$  then, using  $F_{RG}$  the second cypher grid  $Z$  is  $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ .

**Definition 3**  $f_{RG} : (X, Z) \leftarrow f_{RG}(Y)$  generates two cypher grid pixels  $X$  and  $Z$  for a given secret image pixel  $Y$  as

$$X = f_{ran}(.) \text{ and } Z = F_{RG}(X, Y) \quad (3)$$

For a given secret image  $Y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ , the first cypher grid  $X$  generated using  $f_{ran}(.)$  is  $X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$  and second cypher grid  $Z$  using  $F_{RG}(X, Y)$  is  $Z = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ . It is evident that the secret  $Y$  can be recovered from the two cypher grids as  $Y = X \oplus Z$ , where  $\oplus$  represents XOR operation.

## The Proposed Algorithm

**Input:**  $SI_1$ (cover),  $SI_2$ (face), and  $SI_3$  (fingerprint) images each of size  $m \times n$ .

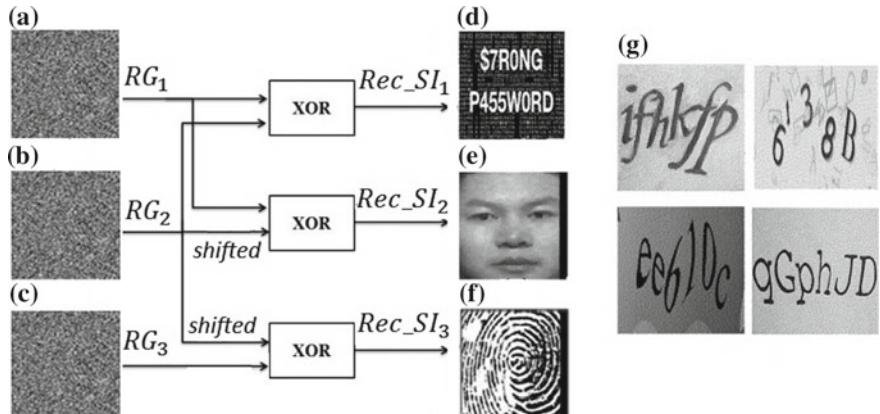
**Output:** Shares  $RG_1$ ,  $RG_2$ ,  $RG_3$ ,  $RG_4$  each of size  $m \times n$ .

- Step 1. Repeat steps 2–7 for position  $i = 0$  to  $m - 1$  and  $j = 0$  to  $n/p - 1$  to generate encrypted shares  $RG_1$ ,  $RG_2$ , and  $RG_3$  with  $shift = 1/p$ .
- Step 2. Select a binary pixel  $SI_1(i, j)$  from the first secret image  $SI_1$  and set a counter variable  $c = 0$ .
- Step 3. Generate binary pixels for shares  $RG_1$  and  $RG_2$  at positions  $(i, j + f.c)$ , where  $f = \frac{n}{p}$ , using binary pixels of first secret image  $SI_1$  as  $RG_1(i, j + f.c)$ ,  $RG_2(i, j + f.c) = f_{RG}(SI_1(i, j + f.c))$ .
- Step 4. Generate binary pixel for share  $RG_2$  at position  $(i, j + f.(c + 1))$  using pixel  $RG_1(i, j + f.c)$  and second secret image pixel  $SI_2(i, j + f.c)$  as  $RG_2(i, j + f.(c + 1)) = F_{RG}(RG_1(i, j + f.c), SI_2(i, j + f.c))$ .
- Step 5. Generate binary pixel for share  $RG_3$  at position  $(i, j + f.c)$  using pixel  $RG_2(i, j + f.(c + 1))$  and third secret image pixel  $SI_3(i, j + f.c)$  as  $RG_3(i, j + f.c) = F_{RG}(RG_2(i, j + f.(c + 1)), SI_3(i, j + f.c))$ .

- Step 6. Generate binary pixel for share  $RG_1$  at position  $(i, j + f.(c + 1))$  using pixel  $RG_2(i, j + f.(c + 1))$  and first secret image pixel  $SI_1(i, j + f.(c + 1))$  as  $RG_1(i, j + f.(c + 1)) = F_{RG}(RG_2(i, j + f.(c + 1)), SI_1(i, j + f.(c + 1)))$ .
- Step 7. If  $c < p - 1$ , increment counter  $c = c + 1$  and repeat steps 3 to 7.
- Step 8. Create another random grid  $RG_4(i, j) = f_{ran}()$  for  $1 \leq i \leq m, 1 \leq j \leq n$ .

As shown in Fig. 3, each individual share ( $RG_1$ ,  $RG_2$ ,  $RG_3$ , and  $RG_4$ ) is meaningless and has noise like appearance. These shares reveal secret images when overlapped and XORed at appropriate positions. The first secret image  $Rec\_SI_1$  (i.e., cover) is recovered when  $RG_1$  and  $RG_2$  are directly overlapped and XORed. To recover the second secret image  $Rec\_SI_2$  (i.e., face [9]),  $RG_2$  is shifted horizontally by  $1/p$  of its width and then overlapped and XORed with  $RG_1$ . Similarly to recover the third secret image  $Rec\_SI_3$  (i.e., fingerprint [10]),  $RG_3$  is shifted horizontally by  $1/p$  of its width and XORed with  $RG_1$ . Recovered secret images for  $p = 16$  are shown in Fig. 3d–f. It is visible that the first secret image is fully recovered. Visual area proportional to the shifting quantity is lost for both second and third recovered images, but without affecting the contrast quality. Here  $p$  is equal to 16 and the size of image is  $256 \times 256$  pixels. Thus, visual area lost on reconstruction is  $256 \times 16$  pixels which is  $1/16$  of the width. The area lost can be reduced by decreasing the shifting factor, say  $shift = \{1/32, 1/64 \dots\}$ , but at the cost of extra computation. In case of compromise, the encrypted shares can be revoked and the new shares can be generated by changing the cover image.

**Distribution of shares:** Original biometric images are destroyed after creating the shares. For a user having identity  $ID$ , shares  $RG_1$  and  $RG_4$  are provided to him in a tokenized format and the other two shares  $RG_2$  and  $RG_3$  are stored on distributed



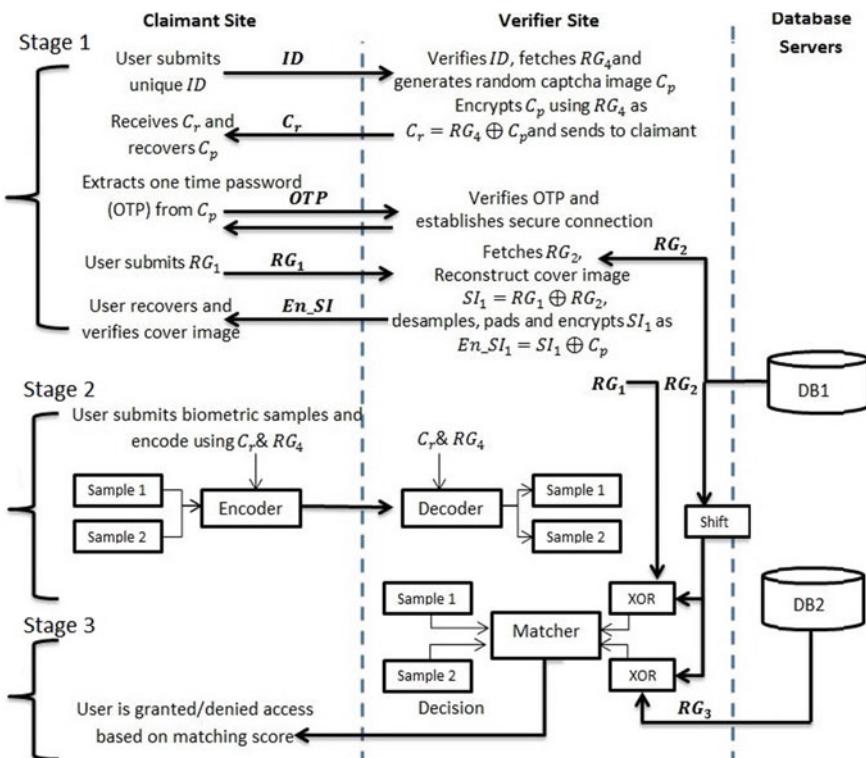
**Fig. 3** **a**  $RG_1$ , **b**  $RG_2$ , **c**  $RG_3$ , **d** recovered cover image, **e** recovered face image, **f** recovered fingerprint image, and **g** sample captcha images

servers. The random grid and associated identity pair ( $RG_4, ID$ ) is maintained at the verifying site.

### 3.2 Authentication Phase

A multistage verification approach is followed as illustrated in Fig. 4. A one-time password based security is applied at **stage 1**. The claimant checks authenticity of the verifier at **stage 2** before providing biometric samples for encryption process. The templates are recovered from the encrypted shares and matched to grant or deny access at **stage 3**.

**Stage 1:** One-time password (OTP) using captcha images is applied and subsequently used for encryption in later stages. The claimant submits his unique  $ID$  to the verifier site. Verifier fetches the associated share  $RG_4$  from its local storage for  $ID$ . Also, it generates a random binary captcha image  $C_p$  of size compatible to random shares.



**Fig. 4** Authentication phase for the proposed architecture

Figure 3g shows sample binary captcha images. Verifier encrypts captcha image  $C_p$  using share  $RG_4$  as  $C_r = RG_4 \oplus C_p$ .

This generates another random share  $C_r$  which is sent to the claimant and asked to verify the OTP. At claimant site user with correct share recovers captcha image as  $C_p = RG_4 \oplus C_r$ , and submits the revealed text as OTP. This establishes a secure connection between claimant and verifier successfully, and user is asked to submit specific token data ( $RG_1$ ). Now verifier fetches the random share  $RG_2$  from the database server and XORs it with input token  $RG_1$  to reconstruct cover image  $SI_1$ . The cover image is desampled by 50% and randomly padded with 0/1's to restore its original size. It is again encrypted using one-time image  $C_p$  and sent to claimant as  $En\_SI_1$ . The user at claimant site recovers modified cover image as  $Rec\_SI_1 = En\_SI_1 \oplus RG_4 \oplus C_r$  and determines if it is the same password image as generated during enrollment. If so, then it is a trusted site, otherwise it is a phishing/fake site. As much of the information is lost while desampling, this step prevents an attacker having stolen token ( $RG_1$ ) on claimant site to obtain encrypted share  $RG_2$  from the knowledge of recovered cover image. Encryption using one-time image  $C_p$  prevents an attacker to obtain any information about  $RG_2$  or cover image.

**Stage 2:** The claimant's face and fingerprint image samples ( $Sample_1$  and  $Sample_2$ ) are acquired from the sensor after confirming verifier's authenticity in Stage 1 and encoded using one-time image  $C_r$  as  $EnS_1 = (2 \times C_r + Sample_1 + RG_4)$ , and  $EnS_2 = (2 \times C_r + Sample_2 + RG_4)$ . Instead of original samples, the encoded versions ( $EnS_1$ ,  $EnS_2$ ) are sent to verifier. The verifier decodes received samples using local share  $RG_4$  as  $Sample_1 = (EnS_1 - RG_4) \bmod 2$  and  $Sample_2 = (EnS_2 - RG_4) \bmod 2$ .

**Stage 3:** At this stage, matching is computed for biometric traits. The two encrypted shares  $RG_1$  and  $RG_2$  are available at the verifier and the third share  $RG_3$  is fetched from the other database server. The reference face  $Rec\_SI_2$  and fingerprint  $Rec\_SI_3$  templates are recovered as  $Rec\_SI_2(\text{face}) = RG_1 \oplus shifted(RG_2)$  and  $Rec\_SI_3(\text{fingerprint}) = RG_3 \oplus shifted(RG_2)$ .

For each modality, the recovered reference template is matched with the decoded query template to generate matching scores. The face templates ( $Rec\_SI_2$  and  $Sample_1$ ) are matched using Linear Discriminant Analysis (LDA) [15] and fingerprint samples ( $Rec\_SI_3$  and  $Sample_2$ ) are input to a fingerprint recognition algorithm [16] to compute similarity scores  $S_{\text{face}}$  and  $S_{\text{finger}}$ , respectively. Similarity scores are normalized and added to generate a combined score which is used to grant the access.

## 4 Experimental Results and Discussion

For experimental verification on multiple modalities, a virtual database is developed using CASIA-Face V5 [9] database and PolyU HRF fingerprint database [10]. CASIA-Face V5 contains color facial images for 500 subjects with 5 samples per

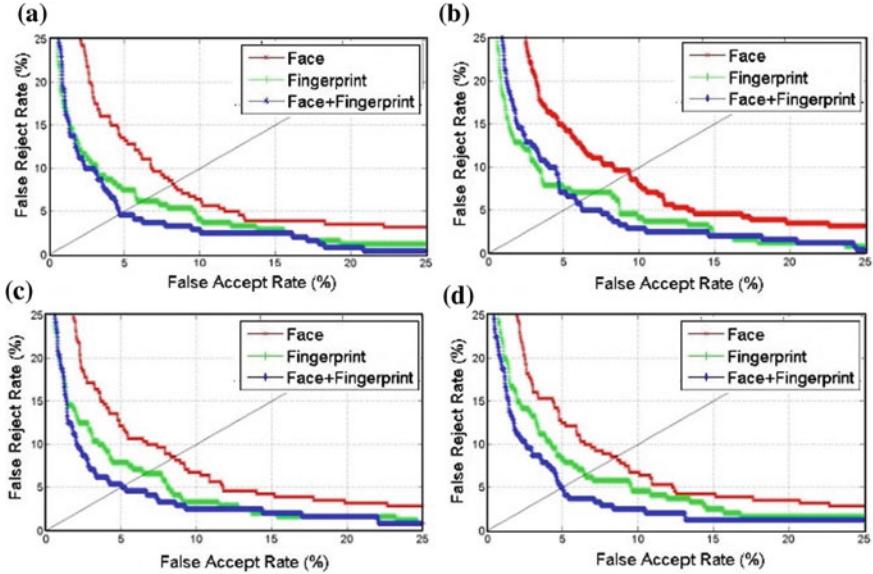
**Table 1** Equal Error Rates (EER) for original and reconstructed templates

Modality	(Original vs. original) (%)	(Original vs. reconstructed) (%)		
		Shift = 1/8	Shift = 1/16	Shift = 1/32
Face	8.22	9.33	8.56	8.55
Fingerprint	6.24	7.09	6.67	6.58
Face+Fingerprint	4.75	5.37	4.97	4.95

subject. The DBII set of PolyU HRF database contains grayscale images for 148 fingers with 5 images per finger. To develop the database, a one-to-one mapping is established for the first 148 subjects. Also, a public database is constructed by randomly selecting 300 images from Google images [11].

**Experiment 1—Performance with reconstructed images:** Performance depends on the quality of reconstructed images  $Rec\_SI_2$  and  $Rec\_SI_3$ . Matching performance is evaluated by determining Equal Error Rates (EER), where lower EER value indicates higher accuracy. Performance of the system should not degrade while operating on reconstructed images. Therefore, a baseline is established by calculating performance of the original templates (without encryption). To evaluate the proposed scheme, samples are reconstructed using the proposed algorithm for different shift factors,  $shift = 1/8, 1/16$ , and  $1/32$ . Table 1 compares performance of the system obtained for original templates match, and for original versus reconstructed templates match for different shift factors. Average results are reported after tuning the system for fivefold cross-validation, which shows that performance of the proposed system does not degrade much as compared to baseline results while providing increased security. EER of the system for original images match are 8.22%, 6.24%, and 4.75% for face, fingerprint, and face+fingerprint (using score-level fusion), respectively. It degrades a little when reconstructed templates are matched with original ones and observed as 8.56%, 6.67%, and 4.97% for face, fingerprint, and face+fingerprint, respectively, at shift factor  $shift = 1/16$ . The performance is slightly better at  $shift = 1/32$  with decreased loss of visual area, i.e., only  $1/32$  of width, but requiring extra computation. However, it starts degrading at  $shift = 1/8$  and beyond as more area is lost on recovery. The supporting ROC curves are shown in Fig. 5.

**Experiment 2—To check the security of shares:** It must be computationally hard to obtain the secret image by any individual share or shares less than the required number. Any individual share should not reveal any information about the secret images. Here, the secret face and fingerprint images are revealed only by stacking  $RG_1$  and  $RG_2$ , and  $RG_2$  and  $RG_3$ , respectively. The possibility of obtaining secret images is minimized by storing  $RG_2$  and  $RG_3$  on distributed servers and providing  $RG_1$  and  $RG_4$  as token to the user. Furthermore, the possibility of relating secret images to the encrypted shares is experimentally observed. Reconstructed biometric images  $Rec\_SI_2$  and  $Rec\_SI_3$  are independently matched with encrypted shares  $RG_1$ ,  $RG_2$ ,  $RG_3$ , and combinations of  $RG_1 \oplus RG_3$ ,  $RG_2 \oplus RG_3$ , and  $RG_1 \oplus RG_2 \oplus RG_3$ .



**Fig. 5** ROC curves, **a** original versus original, **b** original versus recovered,  $p = 8$ , **c** original versus recovered,  $p = 16$ , and **d** original versus recovered,  $p = 32$

Results provided in Table 2 show very high values of EER, which confirms that secret image cannot be estimated from the encrypted shares.

**Experiment 3—Revocability and diversity:** For different applications, a template is decomposed into different shares by using different cover images. Let  $RG_{1i}^k$ ,  $RG_{2i}^k$  and  $RG_{3i}^k$  be the shares created for  $i$ th template using  $k = 3$  different cover images, where  $1 \leq i \leq 740$ . Matching is performed between pairwise combinations of (1)  $RG_{1i}^1$ ,  $RG_{1i}^2$ , and  $RG_{1i}^3$ , (2)  $RG_{2i}^1$ ,  $RG_{2i}^2$ , and  $RG_{2i}^3$ , and (3)  $RG_{3i}^1$ ,  $RG_{3i}^2$ , and  $RG_{3i}^3$ . The results provided in Table 3 show high EER indicating no-match. It validates that new shares can be created by changing the cover image, thus providing good revocability and diversity.

**Experiment 4—Stolen Token Scenario:** This scenario assumes an attacker is always in possession of genuine users' tokens. It is simulated by assigning the same token  $RG_1$  and  $RG_4$  to each user in the database. Shares  $RG_2$  and  $RG_3$  are generated by assigning different cover images to each user. Let  $RG_{2i}$  and  $RG_{3i}$  be the two encrypted shares for an instance  $i$ . Matching is performed between (1)  $RG_{1i}$  and  $RG_{1j}$ ; (2)  $RG_{2i}$  and  $RG_{2j}$ ; and (3)  $RG_{1i}$  and  $RG_{2j}$  for  $1 \leq i, j \leq 740$ . Table 4 shows that there exists no correlation between shares created in this scenario.

**Table 2** EER when reconstructed images are matched with individual shares

Pairwise combinations	EER (%)	Pairwise combinations	EER (%)
$Rec\_SI_2$ versus $RG_1$	49.5	$Rec\_SI_2$ versus $RG_1 \oplus RG_3$	50.0
$Rec\_SI_2$ versus $RG_2$	48.0	$Rec\_SI_2$ versus $RG_2 \oplus RG_3$	41.2
$Rec\_SI_2$ versus $RG_3$	46.7	$Rec\_SI_3$ versus $RG_1 \oplus RG_3$	48.6
$Rec\_SI_3$ versus $RG_1$	49.5	$Rec\_SI_3$ versus $RG_2 \oplus RG_3$	47.2
$Rec\_SI_3$ versus $RG_2$	50.0	$Rec\_SI_2$ versus $RG_1 \oplus RG_2 \oplus RG_3$	47.1
$Rec\_SI_3$ versus $RG_3$	48.7	$Rec\_SI_2$ versus $RG_1 \oplus RG_2 \oplus RG_3$	46.7

**Table 3** EER on cross-matching diverse shares generated using different cover images

Pairwise combinations	EER (%)	Pairwise combinations	EER (%)
$RG_{1i}^1$ versus $RG_{1i}^2$	46.0	$RG_{2i}^3$ versus $RG_{2i}^1$	47.0
$RG_{1i}^2$ versus $RG_{1i}^3$	45.0	$RG_{3i}^1$ versus $RG_{3i}^2$	44.7
$RG_{1i}^3$ versus $RG_{1i}^1$	46.5	$RG_{3i}^2$ versus $RG_{3i}^3$	46.5
$RG_{2i}^1$ versus $RG_{2i}^2$	47.1	$RG_{3i}^3$ versus $RG_{3i}^1$	47.8
$RG_{2i}^2$ versus $RG_{2i}^3$	46.4		

**Table 4** EER on cross-matching shares generated in the stolen token scenario

Pairwise combinations	EER (%)
$RG_{1i}$ versus $RG_{1j}$	49.6
$RG_{2i}$ versus $RG_{2j}$	47.3
$RG_{1i}$ versus $RG_{2j}$	47.5

## 5 Security and Privacy Issues

Security of the system is its strength to resist various attacks while privacy requires that no additional information other than identity should be revealed during storage and transmission. The proposed system addresses security and privacy concerns for various attack scenarios as discussed below:

**Spoofing Attacks:** Authentication using multiple modalities and user token makes it difficult for an attacker to spoof using covert means. Tamper-resistant smart cards may be used as tokens for higher security against covert attacks. If a user loses his token, then the system may temporarily block his account.

**Phishing Attacks:** In stage 1 of the proposed architecture, the claimant receives a resized version of the cover image which allows him to check the authenticity of verifier site. If the verifying site is fake, the generated image will be different from the cover image generated during enrollment.

**Man-In-Middle Attacks:** For the user having identity  $ID$ ,  $(RG_4, ID)$  pair is stored locally at verifier and also  $RG_4$  is provided to the user. Each biometric transmis-

sion is encoded using one-time random image  $C_r$  sent by the verifier using  $RG_4$ . The knowledge of  $RG_4$  remains limited to claimant and verifier, and is never transmitted over the network. Even if an adversary obtains the encoded information, it is not possible for him to recover the original information.

**Replay Attacks and Attacks via Record Multiplicity:** One-time authentication and encoding of all images using captcha image  $C_p$  enhances the security against replay attacks. As the value of encoded images keeps on changing for each transmission, it prevents the attacker who is listening to the link to spoof the verifier or replay the cover image back to fool the user. Also, it prevents cryptanalysis of any recorded information.

**Database Attacks:** The original biometric templates collected during enrollment are encrypted as shares which are stored on distributed databases, and the original templates are destroyed. Even if an attacker gains access to one or more database, no information can be revealed. In case of any tampering with the stored shares, the value generated at the verifier after stacking the shares would be noisy and lead to the detection of such attacks. Experiment 3 verifies that diverse shares cannot be cross-matched for tracing and tracking.

## 6 Conclusion

The experimental data validates the usage of the proposed architecture in remote authentication environments. The proposed architecture is based on visual cryptography yet recovered images do not suffer from the pixel expansion and/or lost in contrast, which are the major limitations of visual cryptography especially while sharing multiple secrets. The recovered images are slightly distorted but the quality is not degraded. The reconstruction of images from encrypted shares requires simple XOR computation. The shares have noisy appearance and do not correlate to any of the secret images as verified through experimentation for various conditions. The cover images are used to provide revocability and diversity. The proposed approach addresses various security and privacy concerns and meets template protection criteria.

**Acknowledgements** This work is supported by BRNS, Dept. of Atomic Energy, Government of India, Grant. No: 36(3)/14/58/2016-BRNS.

## References

1. Naor, M., Shamir, A.: Visual cryptography. In: Advances in Cryptology-EUROCRYPT'94, pp. 1–12. Springer (1995)
2. Monoth, T., Anto, P.B.: Tamperproof transmission of fingerprints using visual cryptography schemes. Procedia Comput. Sci. **2**, 143–148 (2010)

3. Muhammed, R.P., et al.: A secured approach to visual cryptographic biometric template. ACEEE Int. J. Netw. Secur. **2** (2011)
4. Ross, A., Othman, A.: Visual cryptography for biometric privacy. IEEE Trans. Inf. Forensics Secur. **6**, 70–81 (2011)
5. Jain, A.K., Nandakumar, K., Nagar, A.: Biometric template security. EURASIP J. Adv. Signal Process. **2008**, 113 (2008)
6. Takur, V., Jaiswal, R., Sonawane, S., Nalavade, R., et al.: Biometric data security using recursive visual cryptography. Inf. Knowl. Manage. **2**, 32–36 (2012)
7. Patil, S., Tajane, K., Sirdeshpande, J.: Enhancing security and privacy in biometrics based authentication system using multiple secret sharing. In: 2015 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 190–194. IEEE (2015)
8. Nandakumar, K., Ratha, N., Pankanti, S., Darnell, S.: Secure one-time biometric tokens for non-repudiable multi-party transactions. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE (2017)
9. Biometrics Ideal Test: CASIA-FaceV5. <http://biometrics.idealtest.org> (2010)
10. The Hong Kong Polytechnic University: PolyU HRF. <http://www.comp.polyu.edu.hk/~biometrics/HRF/HRF.htm> (2008)
11. Google. <https://www.google.com>
12. Floyd, R.W.: An adaptive algorithm for spatial gray-scale. Proc. Soc. Inf. Disp. **17**, 75–77 (1976)
13. Wang, Z.H., Pizzolatti, M., Chang, C.C.: Reversible visual secret sharing based on random-grids for two-image encryption. Int. J. Innov. Comput. Inf. Control **9**, 1691–1701 (2013)
14. Chen, T.H., Tsao, K.H.: User-friendly random-grid-based visual secret sharing. IEEE Trans. Circuits Syst. Video Technol. **21**, 1693–1703 (2011)
15. Lu, J., Plataniotis, K.N., Venetsanopoulos, A.N.: Face recognition using kernel direct discriminant analysis algorithms. IEEE Trans. Neural Netw. **14**, 117–126 (2003)
16. He, Y., Tian, J., Luo, X., Zhang, T.: Image enhancement and minutiae matching in fingerprint verification. Pattern Recognit. Lett. **24**, 1349–1360 (2003)

# Caption-Based Region Extraction in Images



**Palash Agrawal, Rahul Yadav, Vikas Yadav, Kanjar De and Partha Pratim Roy**

**Abstract** Image captioning and object detection are some of the most growing and popular research areas in the field of computer vision. Almost every upcoming technology uses vision in some way, and with various people researching in the field of object detection, many vision problems which seemed intractable seem close to solved now. But there has been less research in identifying regions associating actions with objects. Dense Image Captioning [8] is one such application, which localizes all the important regions in an image along with their description. Something very similar to normal image captioning, but repeated for every salient region in the image. In this paper, we address the aforementioned problem of detecting regions explaining the query caption. We use edge boxes for efficient object proposals, which we further filter down using a score measure. The object proposals are then captioned using a pretrained Inception [19] model. The captions of each of these regions are checked for similarity with the query caption using the skip-thought vectors [9]. This proposed framework produces interesting and efficient results. We provide a quantitative measure of our experiment by taking the intersection over union (IoU) with the ground truth on the visual genome [10] dataset. By combining the above techniques in an orderly manner, we have been able to achieve encouraging results.

**Keywords** Image captioning · Region proposal network · Skip thought vectors · Long short-term memory · Inception networks

## 1 Introduction

Object detection is one of the most important, basic as well as challenging tasks in computer vision. After having achieved significantly higher accuracy (95%) on image classification tasks, more and more sophisticated and efficient algorithms are proposed every year to solve the object detection task which also involves predicting

---

P. Agrawal · R. Yadav · V. Yadav · K. De (✉) · P. Pratim Roy  
Indian Institute of Technology Roorkee, Roorkee, India  
e-mail: [kanjar.cspdf2017@iitr.ac.in](mailto:kanjar.cspdf2017@iitr.ac.in)

the region in the image where the object is located. Traditionally, object detection was formulated as a classification problem over a sliding window. This approach required evaluating the trained classifier over an exhaustive list of scales, aspect ratios, and positions. This had three basic limitations as follows:

- it only worked out specific regions with specific aspect ratios,
- it is computationally very expensive,
- it was able to detect specific objects, but not activities going on in the regions.

In this paper, Region Proposal Networks (RPN), which have recently emerged in the field of object detection, thereby replacing the traditional computationally expensive sliding window approach has been used. RPNs are an unsupervised approach to segment the images and find bounding boxes for useful regions in an image. Several new approaches to object detection like Fast-RCNN [4] and FasterRCNN [15] also use RPNs.

Image captioning is one of the open problems of computer vision integrated with Natural Language Processing (NLP) and has many interesting applications. With the advent of CNNs and GPUs (more computational power), harder problems like Dense Image Captioning [8] appear more feasible than ever. The traditional captioning algorithms describe the whole image as a single region, but there are certain other interesting regions which might contain some other activity and hence they need to be described as a separate entity. This is taken by the dense image captioning module, which generalizes object detection when single word descriptions are provided and when one predicted region covers the full image then Image captioning is done. However, there is still one important limitation with this method, given this Dense Image Captioning model, we can localize, extract, and describe the salient regions in the input image, however, given a caption or query, we are not able to locate a region which best matches the query description.

All the dense captioning tasks use either localization layers in their CNN or region proposal networks to extract the multiple regions of interest in an image. The proposed method uses the second technique aforementioned for dense captioning and combines it with the sentence similarity task to get the desired result of localizing the input query to the best possible region in an image.

We perform the following tasks : (i) Solve the prime task of Dense Image Captioning using RPNs, CNNs, and LSTMs. (ii) Provide a single training/testing pipeline to solve the region extraction problem as mentioned above. To achieve that, we use the skip-thought encoding for the captions to find similarity index between two given captions, and hence find the most suitable region for the provided caption. To measure the performance of the model, we have used the Visual Genome Dataset. The dataset consists of densely captioned images, we pick up random set of images and captions, provide the caption as the query caption, and measure the Intersection over Union ( $IOU$ ) area of the predicted region and the ground truth, if  $IOU$  is greater than some threshold, we consider it as a correct prediction. The rest of the paper is organized as follows: the background and related work is explained in Sect. 2,

followed by Sect. 3 where we describe the proposed technique and then we present the results of the experiments conducted in Sect. 4 and finally provide conclusion in Sect. 5.

## 2 Background and Related Work

The proposed work is based on some recent work in region proposal networks, image captioning, and sentence similarity matching. Here, we discuss the current approaches available for all three modules. The training and testing pipeline is discussed later on.

### 2.1 *Region Proposal Networks*

There has been some work on using sliding windows as RPNs for objection detection by overfeat [17]. Selective Search [16] used a different approach by applying bottom-up segmentation along with merging at multiple scales. Edge Boxes [21] is another very famous approach that uses window scoring approach instead of segmentation. Recently, Google introduced MultiBox [3], which uses CNNs to generate boxes and Facebook Research launched DeepMask [14], which generated segmentation instead of bounding boxes. After viewing results on various images from Visual Genome, we found out that Edge Boxes gave comparatively better results in case of dense images and was also significantly faster than selective search on high-resolution images. Hence, we use Edge Boxes approach in our proposed technique.

### 2.2 *Image Captioning*

Image Captioning has been attempted by various approaches some using NLP, some using more advanced neural networks. Autoencoders [5] are used vastly for tasks like Image-to-Text translation or text- to-text translation. We use a similar approach by first encoding the input image and then decoding it back to produce the text. Some more recent approaches like image-to-text (im2txt) [20] rely on advanced versions of RNNs like LSTMs along with some image information from CNNs or other image encoding networks, along with some soft attention policies. The main difference arrives from the image encoder model we use. VGG Net was a good choice for a long time, until Microsoft ResNet [6] and Google Inception [19] showed that increasing the layers along with interlayer connections among distant layers can also significantly improve accuracy. We use the same image-to-text approach, along with Google Inception [19] model at the front to encode the image, and LSTMs to produce the captions (decoder) in the proposed approach.

## 2.3 Sentence Similarity

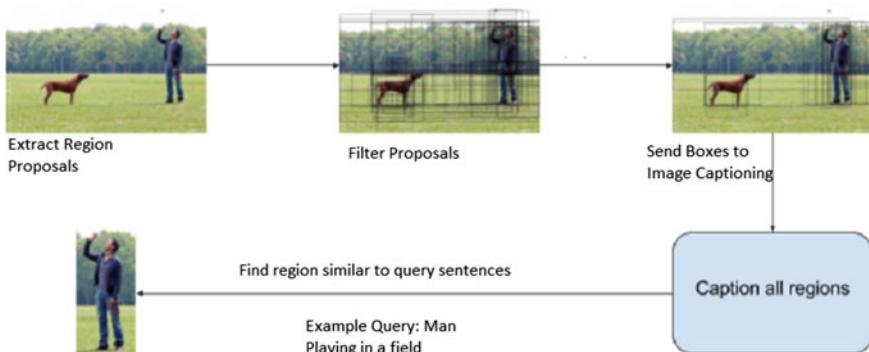
Some earlier approaches used similarity measures like Term Frequency–Inverse Document Frequency (TF-IDF), and other probabilistic approaches. After the introduction of word similarity by word2vec [12], many approaches used this along with some mathematical transformations were applied to provide sentence similarity. Recently introduced Skip-thought vectors [9] used unsupervised approaches on LSTMs to encode sentences. We use skip-thought vectors [9] to first encode the sentences and then find the distance between the encodings of two sentences to measure sentence similarity.

## 3 Proposed Approach

### 3.1 Overview

The outline of our proposed method is shown in Fig. 1. A single image is sent into the Edge Boxes [21] module, which produces outputs as bounding boxes where there is a highest probability to find an object. Edge boxes [21] method works by extracting the sparse representation which is informative provided by edges of an image. The main idea for doing this is because the number of contours which are completely contained inside a bounding box indicate the likelihood of that bounding box to contain an object. To choose the best bounding box, a simple box objectness score is used and this score is measured using the number of edges groups that exist inside the box subtracted by those that are members of contours which overlap the boxes boundary. Currently, we consider the top 30 bounding boxes and further run them through the image captioning module.

The top 30 bounding boxes that were found in Egde Boxes are fed into the Image Captioning module. This module produces one caption for each of the bounding box,



**Fig. 1** Overview and sample results of the proposed method

which has a decent probability of containing an object. All of these captions are then matched with the given query sentence. To compute the caption, the image is first fed into the Inception Net and then passed to the LSTM [7], which predicts the word embedding. Using the dictionary, we convert the word embedding to sentence. A similarity score between each of the captions and the query sentence is found via the Sentence Similarity module.

The highest similarity score won result in the best region, since there is a probability score for each region too. We combine the above two such that both matter equally in deciding the final region. The formula we use for final score is given by Eq. 1 keeping in mind the scaling of scores in each part. Since similarity of caption matters much more than the probability of the region containing a box, we scale the former with a higher power and this results in better bounding boxes.

$$score = \text{probability of object in box} \quad (1)$$

### 3.2 Modules

**Edge Boxes: Region Proposal Network** For tasks relating to object detection, there has been a recent development in the field of region proposals for improving computational efficiency of the work. EDGEBOXES [21] was proposed in a paper by authors from Microsoft. As the name suggests, the algorithm gives a method for generating object bounding boxes proposals using edges. Just like segments, edges provide a simple depiction of an image which is very informative. Additionally, line drawing of an image can show the high-level data contained in the image using only a fraction of the total information present in the image, with considerable accuracy. Since we only have to work with edges, which are efficiently computed, the resulting edge maps are sparse, which leads to many computational advantages.

*Edge Groups and Affinities* Edges usually correlate with object boundaries, and therefore bounding boxes which tightly hug those contours are likely to contain that object. Since the number of possible bounding boxes can be huge, to efficiently score the candidates, Structured Edge detector [1, 2], is used to obtain the starting edge map. The edge groups are formed using a simple greedy approach. Eight-connected edges are combined until the curvature of those edges remains below a certain threshold ( $\frac{\pi}{2}$ ).

Next, we need to compute the affinities between the different edge groups. Given a set of edge groups  $s_i \in S$ , affinity between each pair of edge group is computed. The calculation is based on the mean positions  $x_i$  and  $x_j$ , and the mean orientations  $\theta_i$  and  $\theta_j$ . Intuitively, we see that if the angle between the groups means is comparable to the orientation angle, then these two groups should have high affinity. The formula used to compute affinity is

$$a(s_i, s_j) = |\cos(\theta_j - \theta_{ij}) * \cos(\theta_j - \theta_{ij})|^\gamma \quad (2)$$

where  $\theta_{ij}$  is the angle between  $x_i$  and  $x_j$ .  $\gamma$  is a term used to regulate the responsiveness to the changes in orientation ( $\gamma = 2$  is used in practice). Groups disconnected by more than two pixels have an affinity of zero.

*Bounding Box Scoring* After computing the affinities of a set of edge groups  $S$ , we now compute the object proposal score for any candidate bounding box  $b$  using the following terms.

- $m_i$ —sum of magnitudes  $m_p$  for all edges  $p$  in the group  $s_i$ .
- $x_i$ —a random pixel position of some pixel  $p$  in each group  $s_i$ .
- $S_b$ —set of edge groups that overlap the bounding box boundary.

For each group, a continuous value  $w_b(s_i) \in [0, 1]$  is calculated, which indicates whether the group  $s_i$  is enclosed by the box  $b$  or not. For all groups  $s_i$ , such that  $s_i \in S_b$  or  $x_i \ni b$ , the value of  $w_b(s_i) = 0$ . The score for remaining boxes is found as follows:

$$w_b(s_i) = 1 - \max_T \prod_j^{|T-1|} a(t_j, t_j + 1) \quad (3)$$

where  $T$  is an ordered path of edge groups with length  $T$  that begins at some  $t_i \in S_b$  and ends at  $t_{|T|} = s_i$ . If there is no such path, then  $w_b(s_i) = 1$ . By making use of the computed values for  $w_b$ , the box score is defined as follows:

$$h_b = \frac{\sum_i w_b(s_i)m_i}{2(b_h + b_w)^{\kappa}} \quad (4)$$

where  $b_w$  and  $b_h$  are the boxes width and height, respectively. The variable  $\kappa = 1.5$  is used for offsetting the bias of larger boxes having more edges on average.

*Finding Intersecting Edge Groups* To find out the set of edge groups  $S_b$  that intersect the boundary of a box  $b$ , which we used in the above section, we find the groups which intersect the horizontal and the vertical boundaries of the box separately. To find the intersections along any horizontal boundary from  $(c_0, r)$  to  $(c_1, r)$ , we use two data structures:

- $L_r$ —used to store an ordered list of edge group indices for the row  $r$ .
- $K_r$ —which has the same size as the width of the image and it is used to store the corresponding index into  $L_r$  for each column  $c$  and row  $r$ .

If a pixel  $p$  with location  $(c, r)$  is member of the group  $s_i$ , then  $L_r(K_r(c)) = i$ . We can quickly find all of the coinciding edge groups by searching  $L_r$  from index  $K_r(c_0)$  to  $K_r(c_1)$ .

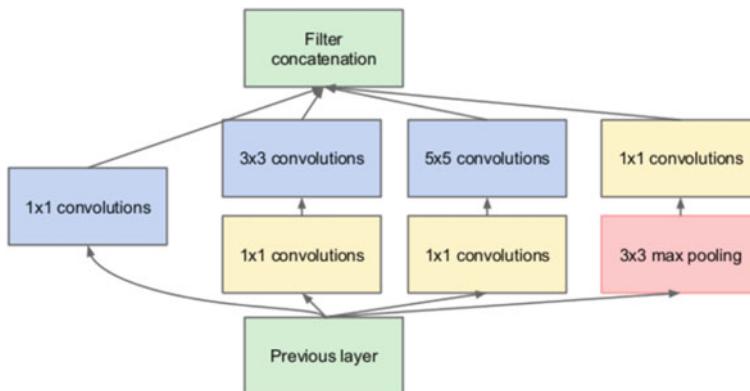
*Search Strategy* Sliding window search on position, scale, and aspect ratio is performed to search for bounding boxes. After the sliding window search is completed, all boxed found with a score  $h_{in}^b$  above a small margin are refined. The boxes are refined using a greedy iterative approach to maximize  $h_{in}^b$  across all the three variables (position, scale and aspect ratio). At the end of each repetition, the stride size

if reduced to half of the value in the previous iteration until it becomes less than 2 pixels. After the refinement, the candidate bounding boxes are recorded and sorted according to their maximum scores. As a final stage, a Non-Maximal Suppression (NMS) [13] is performed on the sorted boxes. As a result, a box which overlaps another box with more score, with an IoU of more than  $\beta$ , the box is removed. In practice, setting  $\beta = \delta + 0.05$  gives very good precision over all values of  $\delta$ .

**Image Captioning** The image captioning takes an image as input, and outputs a set of captions along with their respective probabilities. We use a similar approach as described in the image-to-text [20] paper. The Show and Tell Model is an example of an encoder–decoder neural network. Encoder–Decoder models work by first encoding an image into a vector representation of fixed size, and then decoding the representation into a natural language description.

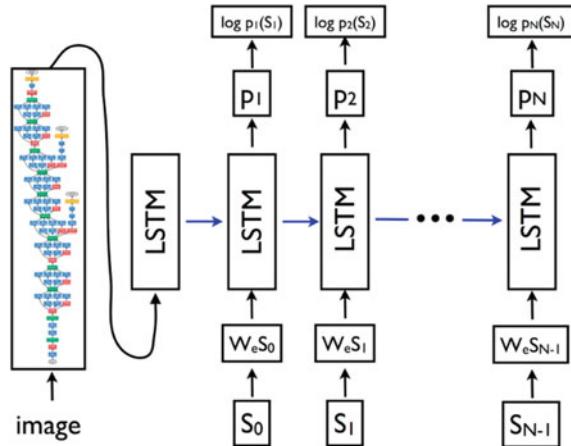
**Image Encoder** The job of the encoder is to translate the image into a vector representation. There is a wide range of CNN networks to choose from like VGGNet [18], AlexNet [11], ResNet [6], and many others. The network for the proposed model is the Inception v3 image recognition model. For this work, pretrained model is on the ILSVRC-2012-CLS image classification dataset is used. The inception model is based on inception modules. In normal CNNs, we need to decide at each layer what size of convolutional filters we want, however, inception module uses all of them. Figure 2 describes how inception layers work. A single inception module consists of  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  convolutions along with a  $3 \times 3$  max pooling. These inception modules combine together to create the Inception v3 model. The final structure of the model looks as shown below. Given the large number of convolutional layers, the number of computations in the model gets very high, hence, the model makes use of  $1 \times 1$  convolutions for dimensionality reduction. As we can see, after several convolutional layers, the model consists of all inception layers, which consist of 3–4 convolutions.

**Decoder** The decoder is a Long Short-Term Memory (LSTM) [7] network. Since, simple LSTM might not be able to predict how different words are related to each



**Fig. 2** Working of inception layers

**Fig. 3** The decoder model using LSTMs



other, instead of passing simple words to the LSTM, we pass the word2vec encodings of the words. This provides LSTM with information such as which words are similar and which are not. We use the word2vec embedding because this way, instead of being as discrete entities such as indexes in the dictionary or just strings, the words become related in an n-dimensional world, similar words have similar n-dimensional vector embeddings.

Now we train the LSTM on the Visual Genome image-caption dataset. Here instead of training the whole dataset from scratch, we used a pretrained LSTM Captioning model from Google, trained on ILSVRC-2012-CLS dataset. We finetune the above model, by running 3 iterations of the Visual Genome dataset. The final finetuned model is then used to caption every image encoding provided by the Inception Image Encoder model. As we can observe in Fig. 3, the image gets encoded by the inception model and then gets passed into the decoder which is the LSTM here.  $S_i$  is the  $i$ th word of the caption here, and  $W_eS_i$  is the word2vec embedding of the corresponding  $i$ th word. The output  $P_i$  is the probability distribution of the next word at the  $i$ th step. Negated sum of log-likelihoods is the minimization objective of the model. This way the LSTM is able to generate word-by-word caption, which has the maximum probability of being related to the image.

**Sentence Similarity** The above two parts together complete the Dense Image Captioning task. However, to predict the region which best matches the provided caption, we need to find out the similarity between a caption and a region. To measure the similarity between the query sentence and a region, we use the precomputed captions for each region. Let  $Sim(A, B)$  be the similarity score between two objects  $A$  and  $B$ . Now assuming  $X$  to be the caption of region  $R$ , and let  $Y$  be the query caption, then:-  $Sim(Y, R) \propto Sim(Y, X) * Sim(X, R)$ . Since, we have already received  $Sim(X, R)$  from the encoder-decoder image captioning model, we need to find similarity score between two sentences. To find the similarity between two sentences, we use skip-thought vectors. This approach is also similar to an encoder approach.

We first encode the sentences using skip-thought encoder, and then find the distance between the encodings.

*Skip-Thought Vectors* Skip-thought vectors are an unsupervised approach to learn generic discriminative features for text. Skip-thought model uses an encoder–decoder approach, where both encoder and decoder tasks are performed by a RNN. Given a sentence, an encoder–decoder RNN model tries to reconstruct the context of a sentence (related to the surrounding sentences). Sentences with similar syntactic and semantic properties are hence mapped to similar vector representations.

Assuming  $SK(X)$  to be the skip-thought vector representation of the sentence  $X$ .

$$Sim(X, Y) \propto \frac{1}{Distance(SK(X), SK(Y))^4} \quad (5)$$

After trying out different proportionality criteria, it was observed that the probability part ( $Sim(X, R)$ ) varies too much, while the  $Distance(SK(X), SK(Y))$  do not vary much, we decided to reduce the proportionality criteria for both of them, making  $Sim(X, Y)$  less proportional to both.

$$Sim(Y, R) = \frac{\sqrt{Sim(X, R)}}{Distance(SK(X), SK(Y))^4} \quad (6)$$

To compute the  $Distance(X, Y)$  between vectors  $X$  and  $Y$ , we used two different approaches:

- Euclidean Distance

$$Euclidean\ Similarity(X, Y) = \frac{1}{|X - Y|} \quad (7)$$

- Cosine Similarity Distance

$$Cosine\ Similarity(X, Y) = \frac{(X \cdot Y)}{(|X| * |Y|)} \quad (8)$$

The above formula pays the right weightage to both the party, making them equally responsible for finding the  $Sim(Y, R)$ . In this case cosine similarity tends to give better results, we have used this idea as an inspiration from TF-IDF similarity measure, which also uses angle difference as the measure, which in turn is equivalent to cosine similarity.

## 4 Experimental Results

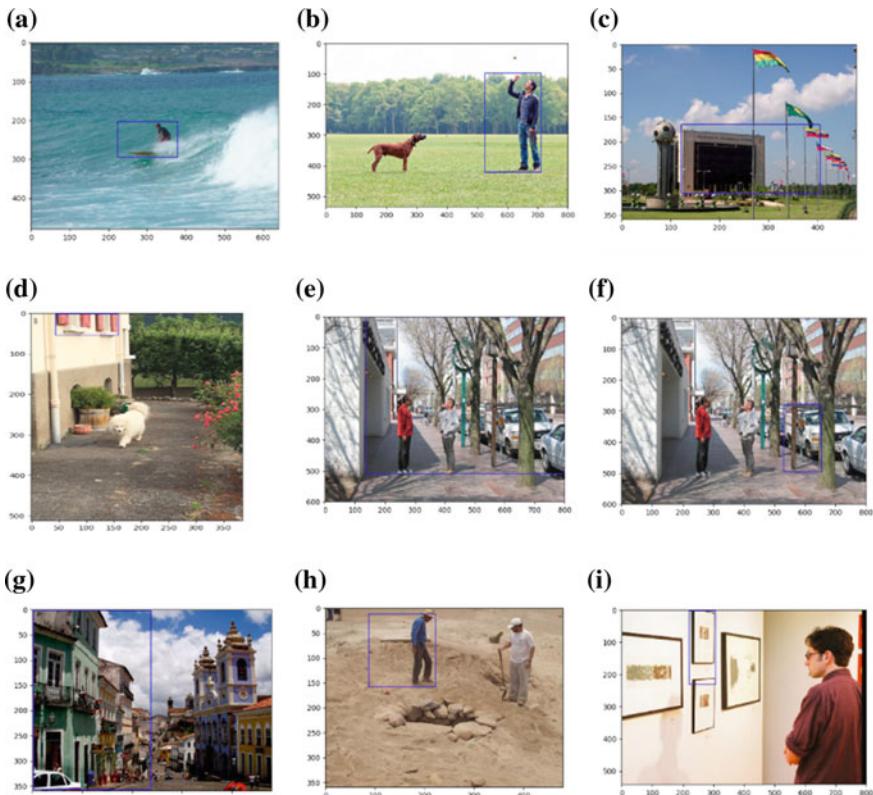
We have tested the proposed on the Visual Genome Dataset. To find accuracy of the prediction, we use Intersection over Union (IOU) methodology as used in edge boxes. For each caption instance, we measure the IOU of the predicted box and

**Table 1** Different accuracies on different IoUs

IoU threshold	Accuracy (%)
0.2	68.4
0.3	61.9
0.4	53.2

the truth value, and if IOU is above a certain threshold, we mark the prediction as correct, else incorrect. The test data contains a set of 5000 images randomly sampled from the Visual Genome test dataset. The images have an average size of  $500 \times 800$ . Processing each images has the following phases:

- Extracting region proposals takes around 1 s for each image to generate 300 region proposals, which are then filtered to 50 boxes.
- Dense Image Captioning, captioning each region takes around 0.6 s, hence in total it takes around 0.5 min for a single image



**Fig. 4** Results for queries—**a** Man surfing in the sea. **b** Man playing in the field. **c** Flagpoles in front of the building. **d** Window on the wall. **e** Two people talking to each other. **f** Car parked on the road. **g** A group of people walking down the street. **h** Man holding shovel in his hand. **i** Paintings on the wall

- Query-based region extraction takes around 0.1 s, since it only needs to perform similarity on the extraction regions.

However, initializing the skip-thought and the captioning model takes around 3–4 minutes. The results obtained for the caption-based region extraction task are as given in the Table 1. Examples of results of queries of certain images from the Visual Genome Dataset is shown in Fig. 4. To the best of our knowledge, we have not yet seen any other model performing caption-based region extraction task to compare this result.

## 5 Conclusion and Future Work

We introduced the caption-based region extraction task, and built a model to dense caption an image. For the above tasks, we used RNNs, CNNs, and LSTMs [7]. These three modules were combined to solve the dense image captioning task, whose solution was then extended to solve caption-based region extraction. Our experiments on the Visual Genome [10] Dataset show promising and visually pleasing results. This task can also be further extended to solve image-based search and database storing problems efficiently. We have also seen how to extract all the information in the provided image using Dense Image Captioning [8]. This can be used to make databases of images, which store the images with a more content- based approach to make many operations faster. Apart from storing and searching the images, this can be used by general web applications also, which aim to extract information from images. This can thus be used in many fields related to image processing, computer vision, and AI.

## References

1. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1841–1848 (2013)
2. Dollár, P., Zitnick, C.L.: Fast edge detection using structured forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(8), 1558–1570 (2015)
3. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2147–2154 (2014)
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)

8. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4565–4574 (2016)
9. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in Neural Information Processing Systems, pp. 3294–3302 (2015)
10. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **123**(1), 32–73 (2017)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. Pham, T.Q.: Non-maximum suppression using fewer than two comparisons per pixel. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 438–451. Springer (2010)
14. Pinheiro, P.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: Advances in Neural Information Processing Systems, pp. 1990–1998 (2015)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017)
16. Van de Sande, K.E., Uijlings, J.R., Gevers, T., Smeulders, A.W.: Segmentation as selective search for object recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1879–1886. IEEE (2011)
17. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. arXiv preprint [arXiv:1312.6229](https://arxiv.org/abs/1312.6229) (2013)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
19. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
20. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 652–663 (2017)
21. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: European Conference on Computer Vision, pp. 391–405. Springer (2014)

# Facial Expression Recognition Using Improved Adaptive Local Ternary Pattern



Sumeet Saurav, Sanjay Singh, Ravi Saini and Madhulika Yadav

**Abstract** Recently, there has been a huge demand for assistive technology for industrial, commercial, automobile, and societal applications. In some of these applications, there is a requirement of an efficient and accurate system for automatic facial expression recognition (FER). Therefore, FER has gained enormous interest among computer vision researchers. Although there has been a plethora of work available in the literature, an automatic FER system has not yet reached the desired level of robustness and performance. In most of these works, there has been the dominance of appearance-based methods primarily consisting of local binary pattern (LBP), local directional pattern (LDP), local ternary pattern (LTP), gradient local ternary pattern (GLTP), and improved local ternary pattern (IGLTP). Keeping in view the popularity of appearance-based methods, in this paper, we have proposed an appearance-based descriptor called Improved Adaptive Local Ternary Pattern (IALTP) for automatic FER. This new descriptor is an improved version of ALTP, which has been proved to be effective in face recognition. We have investigated ALTP in more details and have proposed some improvements like the use of uniform patterns and dimensionality reduction via principal component analysis (PCA). The reduced features are then classified using kernel extreme learning machine (K-ELM) classifier. In order to validate the performance of the proposed method, experiments have been conducted on three different FER datasets using well-known evaluation measures such as accuracy, precision, recall, and F1-Score. The proposed approach has also been compared with some of the state-of-the-art works in literature and found to be more accurate and efficient.

**Keywords** Facial expression recognition (FER) · Adaptive local ternary pattern (ALTP) · Principal component analysis (PCA) · Kernel extreme learning machine (K-ELM)

---

S. Saurav (✉) · S. Singh · R. Saini  
Academy of Scientific and Innovative Research (AcSIR), Chennai, India  
e-mail: [sumeetssaurav@gmail.com](mailto:sumeetssaurav@gmail.com)

CSIR-Central Electronics Engineering Research Institute, Pilani, India

M. Yadav  
Department of Electronics, Banasthali Vidyapith, Vidyapith, Rajasthan, India

## 1 Introduction

Rapid advancement in automation recently has resulted in huge demand for assistive technology for industrial, commercial, automobile, and societal applications. One such technology is the facial expression recognition (FER) system. This technology has many practical implications as, facial expression provides an important cue which reveals the actual intention and state of mind of a person.

As per the existing works related to FER in literature, the techniques available for automatic FER can be broadly classified into two main categories. These are the geometric-based methods and the appearance-based methods and are discussed in more details in [1, 2]. Common techniques under the category of appearance-based feature extraction methods consist of local binary patterns (LBP), local ternary pattern (LTP), local derivative pattern (LDP), local directional number pattern (LNDP), local directional texture pattern, local directional ternary pattern. Deployment of LBP for FER for the first time was done in [3]. Although LBP is very effective and computationally efficient feature descriptor, it has been found to perform poorly under the presence of non-monotonic illumination variation and random noise. To overcome this limitation, Sobel-LBP [4] was proposed. The performance of this operator was found to outperform the traditional LBP operator in terms of recognition accuracy. However, this operator also fails in uniform and near-uniform regions. To overcome this, LDP [5] was developed wherein a different texture coding scheme is used which comprises of directional edge response values instead of gray-level intensity values as in LBP. Although LDP has been proved to be superior to LBP but it also face issues similar to Sobel-LBP. In order to overcome the limitations of LDP and Sobel-LBP, LTP was developed. LTP uses the ternary code as opposed to binary codes in LBP. More recently, a technique called gradient local ternary pattern (GLTP) [6] has been developed for the purpose of FER which combines Sobel operator with LTP operator. GLTP uses a three-level discrimination ternary coding scheme like LTP of gradient magnitudes obtained after Sobel operation to encode the texture of an image. As expected, GLTP has proved to be more effective for FER task compared to the earlier discussed operators. Another feature descriptor which was developed to overcome the limitations of the LBP is the Weber local descriptor (WLD) [7] which has been adopted for the purpose of FER in [8]. A more recent face descriptor called local directional ternary pattern (LDTP) has been developed for FER [2]. LDTP efficiently encodes information of emotion-related features by using the directional information and ternary pattern. Another recent method for FER which has been motivated by GLTP is improved gradient local ternary patterns (IGLTP) proposed by the authors in [9]. The improvements over GLTP includes the use of an input preprocessing step, a more accurate edge detection scheme, use of PCA to reduce the feature dimension, and discriminate feature extraction from facial components.

The remainder of the paper is organized as follows: In Sect. 2, we have provided a brief description of the proposed methodology used in our work which is followed by experimental results and discussion in Sect. 3. Finally, Sect. 4 concludes the paper.

## 2 Proposed Methodology

The algorithmic pipeline used for the implementation of the proposed automatic FER consists of a sequence of steps which involves: Face detection and registration, feature extraction, feature dimensionality reduction, and classification. All of these steps have been discussed briefly below.

### 2.1 Face Detection and Registration

Face detection and registration step comprise face detection and landmark detection unit. The face detector takes an input image and provides the location of human faces and for this, Viola and Jones frontal face detector [10] have been used with cascade classifier trained using Multi-Block Local Binary Pattern (MB-LBP) features [11]. The detected face is passed to the facial landmark detection unit [12], which marks the location of different landmarks on the face. Finally, using coordinates of different landmarks from the left and right eyes, the positions of the eyes center is calculated. Based on the location of the eye's center, the image is rotated and in the subsequent step, the area of interest is cropped and scaled to the specified size in order to obtain the registered facial image.

### 2.2 Facial Feature Extraction and Dimensionality Reduction

The sequence of steps involved in the proposed IALTP facial feature extraction has been shown in Fig. 1. The first step of the proposed IALTP involves the use of a uniform version of Adaptive Local Ternary Pattern (ALTP) proposed in [13] for face recognition. Once the Uniform ALTP coded lower and upper images are obtained, they are then divided into different cells from which histograms are calculated and concatenated to get the final facial features. Finally, a dimensionality reduction via principal component analysis (PCA) is applied to get the reduced uniform ALTP features called IALTP.

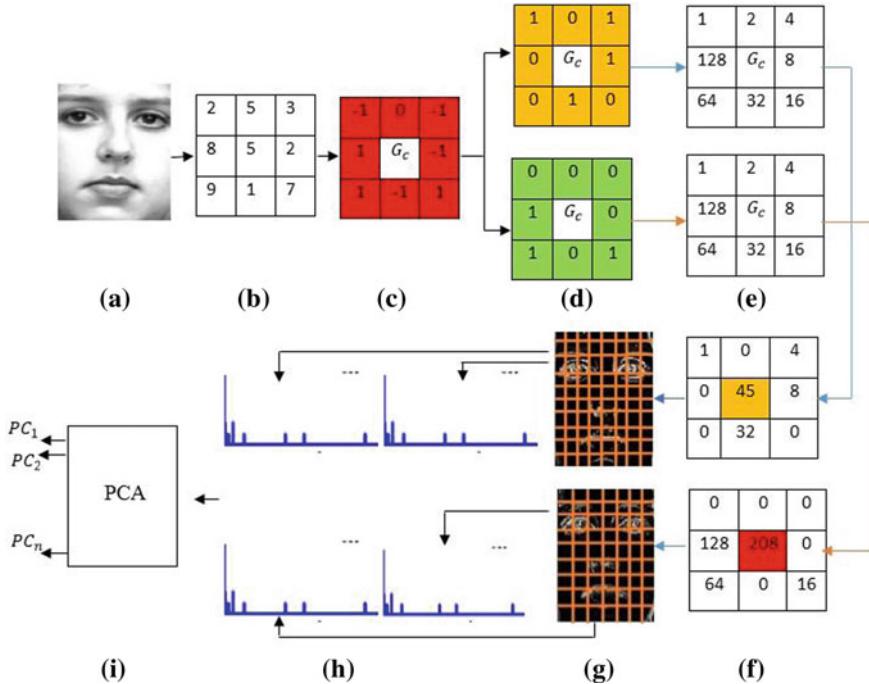
Major motivation behind using the uniform version of ALTP (called Uniform Adaptive Local Ternary Patterns) has been taken from the work of [14]. Using uniform patterns helps in reducing the length of the feature vector from 256-bins in original ALTP to 59-bins in uniform ALTP. This, in turn, facilitates improving the performance of classifiers both in terms of accuracy and computational burden without any loss of recognition accuracy. Furthermore, ALTP has been used in comparison to other similar descriptors because this descriptor makes use of an adaptive way of threshold determination without any manual intervention using Weber's law [15] given in (1). Weber's law states that the size of a just noticeable difference (i.e.,  $\Delta I$ ) is a constant proportion of the original stimulus value.

$$\frac{\Delta I}{I} = k \quad (1)$$

Thus, in ALTP the threshold is automatically set using Weber's law and therefore the method is called adaptive local feature descriptor. In (1), the parameter  $k$  is known as Weber's parameter and is determined experimentally. Once  $k$  is determined, the threshold is set according to (2).

$$t = G_c \times k \quad (2)$$

The threshold is then applied around a center pixel value  $G_c$  of  $3 \times 3$  pixel neighborhoods throughout the input facial image as shown in Fig. 1b. Neighbor pixels falling in between  $G_c + t$  and  $G_c - t$  are quantized to 0, while those below  $G_c - t$  to -1 and finally those above  $G_c + t$  to 1 using (3). In (3),  $S_{ALTP}$  are the quantized values of the surrounding neighbors as shown in Fig. 1c. The resulting eight  $S_{ALTP}$  values for each result in a much higher number of possible patterns when compared to that of LBP, therefore to reduce the dimensionality, each ALTP code is split into its positive and negative parts and treated as individual codes as shown in Fig. 1d. The formula for converting each binary ALTP code to positive  $P_{ALTP}$  and negative  $N_{ALTP}$  decimal codes are given in (4), (5) and (6), (7), respectively, where in each of these



**Fig. 1** Representation of sequence of steps involved in the proposed IALTP with  $t = 2$

patterns are multiplied to some fixed weights shown in Fig. 1e and then summed to give the positive and negative decimal coded value as shown in Fig. 1f.

$$S_{ALTP}(G_c, G_i) = \begin{cases} -1, & G_i < G_c - t, \\ 0, & G_c - t \leq G_i \leq G_c + t, \\ +1, & G_i > G_c + t. \end{cases} \quad (3)$$

$$P_{ALTP} = \sum_{i=0}^7 S_P(S_{ALTP}(i)) \times 2^i, \quad (4)$$

$$S_P(v) = \begin{cases} 1, & \text{if } v > 0, \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$$N_{ALTP} = \sum_{i=0}^7 S_N(S_{ALTP}(i)) \times 2^i \quad (6)$$

$$S_N(v) = \begin{cases} 1, & \text{if } v < 0, \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

Once the positive  $P_{ALTP}$  and negative  $N_{ALTP}$  decimal coded images are available, these are then converted into their respective uniform version using a lookup table. Since in our case we are using 8 sampling points, the number of different output labels for mapping the patterns of 8 bits is 59 wherein out of 256 patterns, 58 uniform patterns are given output labels from 0 to 57 and the rest of the nonuniform patterns are grouped with a single output label 58. Finally, the uniform positive and negative ALTP coded image is divided into  $m \times n$  regions of some specified size as shown in Fig. 1g. A positive ( $H_{P_{UALTP}}$ ) and negative ( $H_{N_{UALTP}}$ ) uniform ALTP histogram is computed for each region using (8)–(10).

$$H_{P_{UALTP}}(\tau) = \sum_{r=1}^M \sum_{c=1}^N f(P_{ALTP}(r, c), \tau) \quad (8)$$

$$H_{N_{UALTP}}(\tau) = \sum_{r=1}^M \sum_{c=1}^N f(N_{ALTP}(r, c), \tau) \quad (9)$$

$$f(a, \tau) = \begin{cases} 1 & a = \tau \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

In (8), (9), M and N are the width and height of the uniform ALTP coded image whereas r and c denote the dimension of the encoded image. The value of  $\tau$  ranges from 0 to 58 (as compared to 0–255 in the case of ALTP) and for which the frequency of occurrence is calculated using the above-listed equations. Finally, the positive and negative histograms for each region are concatenated together to form the feature vector as in Fig. 1h. Since the size of the feature vector obtained is quite high in

dimension, therefore we have proposed another improvement over traditional ALTP wherein we have made use of dimensionality reduction technique, namely, PCA as shown in Fig. 1i to reduce the dimension of the uniform ALTP feature vector and called the reduced vector as improved ALTP (IALTP) feature vector.

### 2.3 Kernel Extreme Learning Machine (K-ELM) Classifier

In order to classify the facial attributes into their corresponding emotion label, kernel extreme learning machine (K-ELM) multi-class classifier has been used in our proposed work. K-ELM is a popular classifier and is usually treated as the kernelized variant of extreme learning machine (ELM) classifier used for fast training a Single-Layer Feed-Forward Neural Network (SLFN) [16]. The major benefit of using K-ELM as compared to the traditional backpropagation algorithm based neural network architecture is that the training using K-ELM does not involve any iteration and the output weights are calculated using a direct solution. K-ELM classifier is mostly used in the cases where the mapping function is not known, and it uses kernel technique based on Mercer's condition [17]. The output vector  $f(x)$  of K-ELM can be represented as shown in (11).

$$f(x) = h(x)\beta = h(x)H^T \left( \frac{I}{C} + HH^T \right)^{-1} T = \begin{bmatrix} \phi(x, x_1) \\ \vdots \\ \phi(x, x_{N_k}) \end{bmatrix} \left( \frac{I}{C} + \Phi \right)^{-1} T \quad (11)$$

where

$$\Phi = HH^T = \begin{bmatrix} \phi(x_1, x_1) & \cdots & \phi(x, x_{N_k}) \\ \vdots & \ddots & \vdots \\ \phi(x_{N_k}, x_1) & \cdots & \phi(x_{N_k}, x_{N_k}) \end{bmatrix}$$

In this paper, Gaussian function is used as the kernel  $\phi$  which is represented as in (12), where  $\sigma$  denotes the spread (i.e., standard deviation) of the Gaussian function.

$$\phi(x, x_i) = \exp \left( -\frac{\|x_i - x_j\|^2}{\sigma^2} \right) \quad (12)$$

### 3 Experimental Results and Discussion

In this section, we discuss our experimental setup and various experiments which were performed on different FER datasets. All the experiments have been performed using Matlab 2015a running on a windows platform with 64 GB RAM.

#### 3.1 Datasets

In order to validate the performance of our proposed approach, we have used three FER datasets in our experiments. The first one is the extended Cohn-Kanade (CK+) dataset [18]. This dataset is an extended version of the CK dataset. In our experimental setup, we have used both 6 class and 7 class expression images which were obtained from 309 labeled sequences selected from 106 subjects. For 6-class expression recognition, from each labeled sequence, we selected the three most expressive images resulting in 927 images and for 7-class expression, we simply added the first image of neutral expression from each of the 309 sequences to the 6-class dataset, resulting in a total of 1236 images as is done in [9]. The second dataset used in the experiments is the recently introduced Radbound Faces database (RFD) [19]. The dataset contains images of 67 subjects performing 8 facial expression (anger, disgust, fear, happiness, contemptuous, sadness, surprise, and neutral) with 3 gaze directions. However, in our experiments, we have only used frontal gaze direction images comprising 7 expressions (anger, disgust, fear, happy, neutral, sad, and surprise) for a total of 469 images. Finally, the third dataset used is the Japanese female facial Expression (JAFFE) dataset [20]. The dataset contains 7 different prototypic facial expression images of 10 female subject consisting of a total of 213 images.

#### 3.2 Parameter Selection

Designing an efficient FER system usually, involve a number of parameters. Therefore, the optimal value of these parameters is often desired to achieve good recognition accuracy. First, we tried to determine the optimal facial image size and cell size and for this, we experimented with two different facial image resolution of size  $65 \times 59$  as in [9] and  $147 \times 108$  pixels as in [21]. We experimented with 8 different cell sizes in which the facial image is divided. The experiments were performed on CK+ 7 expression dataset with tenfold cross-validation strategy which was repeated 10-times using K-ELM classifier with regularization parameter C and kernel parameter  $\gamma$  value of 100 and 200, respectively. The value of Weber's parameter k fixed in the experiment is 0.12. Tables 1 and 2 depicts the results of the experiment for  $65 \times 59$  and  $147 \times 108$  resolution facial images, respectively. Based on the experimental results, we found that the facial image with a resolution of  $147 \times 108$  and cell size of

**Table 1** Performance of different cell sizes on  $65 \times 59$  pixels facial image

	[65, 59]	[5, 4]	[6, 5]	[7, 6]	[8, 7]	[9, 8]	[10, 9]	[11, 10]	[12, 11]
Avg. Acc. 10 runs	99.1 $\pm$ 0.2	99.2 $\pm$ 0.2	98.9 $\pm$ 0.2	99.0 $\pm$ 0.2	98.5 $\pm$ 0.3	97.6 $\pm$ 0.4	96.5 $\pm$ 0.5	96.7 $\pm$ 0.1	
Feature Dim.	19,824	<b>12,980</b>	9558	6608	5782	4248	2950	2950	
Avg. Acc.	99.35	<b>99.35</b>	99.35	99.19	99.03	98.30	97.17	97.01	
Avg. Prec.	99.29	<b>99.27</b>	99.33	99.12	98.91	98.66	97.88	97.57	
Avg. Rec.	99.34	<b>99.34</b>	99.33	99.23	99.08	97.86	96.10	96.09	
Avg. F1-S	99.32	<b>99.30</b>	99.33	99.17	98.99	98.24	96.94	96.76	

**Table 2** Performance of different cell sizes on  $147 \times 108$  pixels facial image

	[147, 108]	[7, 6]	[8, 7]	[9, 8]	[10, 9]	[11, 10]	[12, 11]	[13, 12]	[14, 13]
Avg. Acc. 10 runs	99.1 $\pm$ 0.2	99.1 $\pm$ 0.2	<b>99.3 <math>\pm</math> 0.2</b>	98.9 $\pm$ 0.2	99.2 $\pm$ 0.1	99.0 $\pm$ 0.2	98.6 $\pm$ 0.2	98.7 $\pm$ 0.2	
Feature Dim.	40,120	31,860	<b>24,544</b>	18,172	15,340	12,744	10,384	9440	
Avg. Acc.	99.35	99.35	<b>99.51</b>	99.11	99.27	99.27	98.95	98.95	
Avg. Prec.	99.30	99.33	<b>99.46</b>	99.04	99.26	99.15	99.09	98.75	
Avg. Rec.	99.46	99.45	<b>99.70</b>	99.31	99.28	99.38	98.98	99.05	
Avg. F1-S	99.38	99.39	<b>99.58</b>	99.17	99.27	99.26	99.03	98.89	

$9 \times 8$  performed well compared to other possible combinations and therefore, this value of facial image size and cell size was used in all our further experiments.

The second experiment involved the determination of the optimal value of Weber's parameter k for every dataset, whose results have been tabulated in Tables 3, 4, and 5 and the third experiment dealt with the determination of the K-ELM

**Table 3** Performance of different k on CK+ 7 expression database

K	0.01	0.03	0.06	0.09	0.12	0.15	0.18	0.21
Avg. Acc.	<b>99.35</b>	99.09	99.20	99.21	99.31	98.97	98.77	98.59

**Table 4** Performance of different k on JAFFE 7 expression database

K	0.01	0.03	0.06	0.09	0.12	0.15	0.18	0.21
Avg. Acc.	<b>95.53</b>	94.97	95.44	94.52	92.86	93.32	92.71	91.45

**Table 5** Performance of different k on RFD 7 expression database

K	0.01	0.03	0.06	0.09	0.12	0.15	0.18	0.21
Avg. Acc.	96.43	<b>97.42</b>	96.47	95.43	95.15	94.36	93.46	93.39

parameters, i.e., the optimal value of regularization coefficient C and kernel parameter  $\gamma$ . For this experiment, we used the CK+ 7 expression dataset with fixed image size, cell size, and k obtained in our earlier experiment. The range of both C and  $\gamma$  taken here was [1:10] in a logarithmic scale of base 2. The result of the experiment has been tabulated in Table 6.

Further experiment involved the determination of the optimal number of principal components for all the three FER datasets. To determine this, we fixed the value of all other parameters involved in the optimal value determined earlier. The experimental result for CK+ 7 expression dataset has been tabulated in Table 7. Similar experiments were also carried out for JAFFE and RFD dataset but we have not shown the results due to limited paper length. From the experiments, we found that the optimal number

**Table 6** Determination of K-ELM parameter

Performance measure/datasets		CK+ 7 expressions						
Avg. Accuracy 10-runs		$99.47 \pm 0.08$						
Kernel parameter ( $\gamma$ )		1024						
Regularization parameter (C)		256						

**Table 7** Determination of no. of principal component of CK+ 7 expression dataset

No. of PC	32	64	96	128	160	192	224	256
Avg. Acc. 10 runs	$98.1 \pm 0.2$	$98.6 \pm 0.2$	$99.2 \pm 0.2$	$99.4 \pm 0.1$	$99.4 \pm 0.2$	$99.5 \pm 0.1$	<b><math>99.5 \pm 0.1</math></b>	$99.4 \pm 0.1$
Avg. Acc.	98.3	98.8	99.4	99.6	99.6	99.6	<b>99.7</b>	99.5
Avg. Prec.	98.5	99.0	99.4	99.5	99.6	99.6	<b>99.6</b>	99.4
Avg. Rec.	98.3	99.0	99.6	99.7	99.7	99.7	<b>99.8</b>	99.7
Avg. F1-S	98.4	99.0	99.5	99.6	99.6	99.6	<b>99.7</b>	99.6

of principal components is 224 for CK+, 96 for JAFFE and 256 for RFD dataset which clearly demonstrate the usefulness of dimensionality reduction for which PCA has been used in the proposed work.

### 3.3 Results on CK+ Dataset

In order to determine the performance of the proposed FER pipeline on CK+ dataset, we performed tenfold cross-validation which was repeated 10 times. On CK+ 6 expression the accuracy achieved using uniform ALTP is  $99.9 \pm 0.1$  and that using IALTP is  $99.9 \pm 0.1$ . On CK+ 7 expression dataset the FER pipeline achieved accuracy of  $99.4 \pm 0.2$  and  $99.5 \pm 0.1$  using uniform ALTP and IALTP, respectively. Tables 8 and 9 depict the result of our experiment using different performance mea-

**Table 8** Confusion matrix of IALTP on CK+ 6 expressions

Actual/predicted	An	Di	Fe	Ha	Sa	Su	Recall
An	135	0	0	0	0	0	100
Di	0	177	0	0	0	0	100
Fe	0	0	75	0	0	0	100
Ha	0	0	0	207	0	0	100
Sa	0	0	0	0	84	0	100
Su	0	0	0	0	0	249	100
Precision	100	100	100	100	100	100	
F1-Score	100	100	100	100	100	100	

Avg. performance: recall = 100, precision = 100, accuracy = 100, F1-Score = 100

**Table 9** Confusion matrix of IALTP on CK+ 7 expressions

Actual/predicted	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	135	0	0	0	0	0	0	100
Di	0	177	0	0	0	0	0	100
Fe	0	0	75	0	0	0	0	100
Ha	0	0	0	207	0	0	0	100
Ne	1	0	0	0	307	1	0	99.3
Sa	0	0	0	0	0	84	0	100
Su	0	0	0	0	2	0	247	99.2
Precision	99.3	100	100	100	99.3	98.8	100	
F1-Score	99.6	100	100	100	99.3	99.4	99.6	

Avg. performance: recall = 99.8, precision = 99.6, accuracy = 99.7, F1-Score = 99.7

sures corresponding to the best tenfold cross-validation for both the categories of the dataset.

### 3.4 Results on Jaffe Dataset

In order to determine the performance of the proposed FER pipeline on JAFFE, similar experiments were performed as discussed above. On JAFFE 6 expression, the accuracy achieved using uniform ALTP is  $95.8 \pm 0.9$  and that using IALTP is  $95.9 \pm 0.8$ . On JAFFE 7 expression dataset, the FER pipeline achieved accuracy of  $95.5 \pm 0.7$  and  $95.9 \pm 0.9$  using uniform ALTP and IALTP, respectively. The performance of the best tenfold cross-validation in terms of different measures has been shown in Tables 10 and 11.

**Table 10** Confusion matrix of IALTP on JAFFE 6 expressions

Actual/predicted	An	Di	Fe	Ha	Sa	Su	Recall
An	30	0	0	0	0	0	100
Di	0	28	0	0	1	0	96.5
Fe	0	0	31	0	0	1	96.9
Ha	0	0	0	31	0	0	100
Sa	0	0	1	1	29	0	93.5
Su	0	0	0	1	0	29	96.7
Precision	100	100	96.9	93.9	96.7	96.7	
F1-Score	100	98.2	96.9	96.9	95.1	96.7	

Avg. performance: recall = 97.3, precision = 97.4, accuracy = 97.3, F1-Score = 97.3

**Table 11** Confusion matrix of IALTP on JAFFE 7 expressions

Actual/predicted	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	30	0	0	0	0	0	0	100
Di	0	28	0	0	0	1	0	96.5
Fe	0	0	31	0	0	0	1	96.9
Ha	0	0	0	31	0	0	0	100
Ne	0	0	0	0	30	0	0	100
Sa	0	0	1	1	0	29	0	93.5
Su	0	0	0	1	0	0	29	96.7
Precision	100	100	96.9	93.9	100	96.7	96.7	
F1-Score	100	98.2	96.9	96.9	100	95.1	96.7	

Avg. performance: recall = 97.7, precision = 97.7, accuracy = 97.6, F1-Score = 97.7

### 3.5 Results on RFD Dataset

Performance of the proposed FER pipeline on RFD dataset again in terms of different performance measures corresponding to the best tenfold cross-validation run has been tabulated in Table 12. The overall average accuracy of the 10 runs of the tenfold cross-validation using uniform ALTP and IALTP is  $97.4 \pm 0.4$  and  $97.8 \pm 0.4$ , respectively. From the confusion matrix, we find the classifier performed well in classifying the expressions of disgust (Di), fear (Fe), happy (Ha), neutral (Ne), and surprise (Su).

Finally, the proposed FER pipeline using the IALTP features has also been compared with some of the state-of-the-art works available in the literature. Comparison result has been shown in Table 13. From the table, one can find that the proposed approach has achieved superior performance and is more accurate and effective. This

**Table 12** Confusion matrix of IALTP on RFD 7 expressions

Actual/predicted	An	Di	Fe	Ha	Ne	Sa	Su	Recall
An	66	0	0	0	0	1	0	98.5
Di	0	67	0	0	0	0	0	100
Fe	0	0	67	0	0	0	0	100
Ha	0	0	0	67	0	0	0	100
Ne	0	0	0	0	67	0	0	100
Sa	1	0	1	0	4	61	0	91.0
Su	0	0	0	0	0	0	67	100
Precision	98.5	100	98.5	100	94.4	98.4	100	
F1-Score	98.5	100	99.3	100	97.1	94.6	100	

Avg. performance: recall = 98.5, precision = 98.5, accuracy = 98.5, F1-Score = 98.5

**Table 13** Comparison of recognition accuracy (%) on different datasets

Method	CK+ 6 expression	CK+ 7 expression	RFD 7 expression	JAFFE 6 expression	JAFFE 7 expression
LBP [22]	90.1	83.3	–	–	NA
LDP [22]	93.7	88.4	–	–	NA
LTP [22]	93.6	88.9	–	–	NA
GLTP [6]	97.2	91.7	–	77.0	74.4
Improved GLTP [9]	99.3	97.6	–	83.3	81.7
HOG [21]	95.8	94.3	94.9	–	–
<b>Uniform ALTP</b>	<b>99.9</b>	<b>99.4</b>	<b>97.4</b>	<b>95.8</b>	<b>95.5</b>
<b>IALTP proposed</b>	<b>100</b>	<b>99.5</b>	<b>97.8</b>	<b>95.9</b>	<b>95.9</b>

clearly indicates the powerfulness of the ALTP descriptor for solving FER problem as compared to the other texture and shape descriptors. Moreover, it can also be found from the table that the performance of the IALTP in terms of recognition accuracy is comparable to that of the uniform version of the ALTP with much smaller feature dimension.

## 4 Conclusion

In the presented paper, a new facial feature descriptor named improved adaptive local ternary pattern has been proposed, which is a modified version of ALTP developed for the purpose of face recognition application. We investigated ALTP from FER problem perspective and proposed some improvements like the use of uniform ALTP patterns and dimensionality reduction via PCA. Both of these improvements were done to enhance the recognition accuracy and processing speed of the proposed FER pipeline. K-ELM classifier has been used for classifying the facial expressions. For a fair comparison with the existing works, the performance of the proposed approach has been validated using 10-fold cross-validation which was repeated 10 times. The experiments were performed on three FER datasets, viz., CK+, JAFFE, and RFD and performance was observed using precision, recall, accuracy, and, F1-score measures. Performance of the proposed approach was also compared with some of the state-of-the-art works available in literature and found to be effective both in terms of accuracy and efficiency which clearly indicates the usefulness of the ALTP descriptor.

## References

1. Rivera, A.R., Castillo, J.R., Chae, O.O.: Local directional number pattern for face analysis: face and expression recognition. *IEEE Trans. Image Process.* **22**(5), 1740–1752 (2013)
2. Ryu, B., Rivera, A.R., Kim, J., Chae, O.: Local directional ternary pattern for facial expression recognition. *IEEE Trans. Image Process.* **26**(12), 6006–6018 (2017)
3. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* **27**(6), 803–816 (2009)
4. Zhao, S., Gao, Y., Zhang, B.: Sobel-lbp. In: 15th IEEE International Conference on Image Processing, pp. 2144–2147 (2008)
5. Jabid, T., Kabir, M. H., Chae, O.: Facial expression recognition using local directional pattern (LDP). In: 17th IEEE International Conference on Image Processing, pp. 1605–1608 (2010)
6. Ahmed, F., Hossain, E.: Automated facial expression recognition using gradient-based ternary texture patterns. *Chin. J. Eng.* (2013)
7. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: a robust local image descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(9), 1705–1720 (2010)
8. Alhussein, M.: Automatic facial emotion recognition using weber local descriptor for e-Healthcare system. *Clust. Comput.* **19**(1), 99–108 (2016)
9. Holder, R.P., Tapamo, J.R.: Improved gradient local ternary patterns for facial expression recognition. *EURASIP J. Image Video Process.* **2017**(1), 42 (2017)
10. Viola, P., Jones, M.J.: Robust real-time face detection. *Int. J. Comput. Vision* **57**(2), 137–154 (2004)

11. Martin, K.: Efficient Metric Learning for Real-World Face Recognition. [http://lrs.icg.tugraz.at/pubs/koestinger\\_phd\\_13.pdf](http://lrs.icg.tugraz.at/pubs/koestinger_phd_13.pdf)
12. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 532–539 (2013)
13. Yang, W., Wang, Z., Zhang, B.: Face recognition using adaptive local ternary patterns method. Neurocomputing **213**, 183–190 (2016)
14. Lahdenoja, O., Poikonen, J., Laiho, M.: Towards understanding the formation of uniform local binary patterns. ISRN Mach. Vis. (2013)
15. Jain, A.K.: Fundamentals of Digital Signal Processing (1989)
16. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **42**(2), 513–529 (2012)
17. Huang, Z., Yu, Y., Gu, J., Liu, H.: An efficient method for traffic sign recognition based on extreme learning machine. IEEE Trans. Cybern. **47**(4), 920–933 (2017)
18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In IEEE Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
19. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.D.: Presentation and validation of the Radboud Faces Database. Cogn. Emot. **24**(8), 1377–1388 (2010)
20. Lyons, M.J., Akamatsu, S., Kamachi, M., Gyoba, J., Budynek, J.: The Japanese female facial expression (JAFFE) database. In Proceedings of Third International Conference on Automatic Face and Gesture Recognition, pp. 14–16 (1998)
21. Carcagnì, P., Coco, M., Leo, M., Distante, C.: Facial expression recognition and histograms of oriented gradients: a comprehensive study. SpringerPlus **4**(1), 645 (2015)
22. Ahmed, F., Kabir, M.H.: Directional ternary pattern (dtp) for facial expression recognition. In IEEE International Conference on Consumer Electronics (ICCE), pp. 265–266 (2012)

# Cell Extraction and Horizontal-Scale Correction in Structured Documents



Divya Srivastava and Gaurav Harit

**Abstract** Preprocessing techniques form an important task in document image analysis. In structured documents like forms, cheques, etc., there is a predefined space called frame field/cell for the user to fill the entry. When the user is writing, the nonuniformity of inter-character spacing becomes an issue. Many times, the starting characters of the word are written with sparse spacing between the characters and then gradually with a more compact spacing so as to accommodate the word within the frame field. To deal with this variation in intra-word spacing, horizontal-scale correction is applied to the extracted form fields. The effectiveness of the system is proved by applying it as a preprocessing step in a recognition system proposed in (Almazán et al. in Pattern Anal Mach Intell 36(12):21552–2566, 2014 [2]). The recognition framework results in reduced error rates with this normalization.

**Keywords** Region growing · Cell extraction · Horizontal-scale correction · Bilingual forms

## 1 Introduction

Offline handwritten text recognition is still a challenging task even after recent advances in the technology. There are various challenges involved in this task, major of all is writer variability. Other challenges include degradation in the text, slant, skew, etc. [3, 4, 21, 22]. While considering structured documents like forms, cheques, etc., there arise other challenges like localizing handwritten text among printed text, text crossing their respective field boundary, etc. [6, 20, 23]. While doing the handwritten text recognition, preprocessing steps such as slant and skew correction have significant advantages. They reduce inter- and intra-writer variability leading to increased

---

D. Srivastava (✉) · G. Harit  
Indian Institute of Technology Jodhpur, Jodhpur, India  
e-mail: [srivastava.5@iitj.ac.in](mailto:srivastava.5@iitj.ac.in)

G. Harit  
e-mail: [gharit@iitj.ac.in](mailto:gharit@iitj.ac.in)

recognition accuracy. A form can be considered as a structured document composed of three main elements, viz., horizontal and vertical layout lines, preprinted data, and user filled-in data. Field frame is a predefined space for the user to fill the entry [19]. When a user writes within a form field, it is often seen that the starting characters of the word are written in a sparse manner but toward the end, it gets squeezed to a compact form as the user attempts to accommodate the word/sentence within the frame field. For applications like word spotting and word recognition, this can affect the performance. Hence the variation in horizontal scale needs to be corrected. This paper addresses the problem of cell extraction and horizontal-scale correction of words written in form fields. The proposed methodology is divided into two tasks, namely, cell extraction and horizontal-scale correction. Dataset for this task is not available, so we have created our own dataset comprising 200 words obtained from multi-writer bilingual (English and Devanagari) forms of 10 different layouts with 10 forms of each layout capturing a variety of within word inter-character spacings.

## 2 Related Work

Related work is divided into two subsections. The first part briefs about the form of frame lines and cell extraction while the second part discusses various text normalization techniques.

### 2.1 Cell Extraction

A form consists of various frame lines which form boundaries of cells in which the user fills the desired text. These lines exhibit the structural layout of the form document. The data filled in a cell have well-defined semantics. In [12], a method based on Box-Driven Reasoning(BDR) is developed for structural analysis in table-form documents. The method recognizes cells in tabular documents with broken lines and touching characters. Extracting these frame lines is also useful in classification and recognition of form. In an organization, similar forms have a similar layout. Extracting the form layout and exploring them can lead to their application in form classification and recognition. Hierarchical representation of various cell forming the structure of form documents is proposed in [7] for identification and retrieval of forms. A heuristic algorithm based on the XY-tree method [5, 16] is used to transform the geometric structure of the form document to a hierarchical structure by using the horizontal and vertical layout lines. The hierarchical structure of the tree corresponds to the hierarchy of the blocks of the form document. This representation can handle the minor variation of the form in the retrieval process.

## 2.2 Text Normalization

Text normalization is an important preprocessing step in the recognition system. Text normalization is of two types, vertical and horizontal text normalization, corresponding to the word being scaled in the vertical and horizontal direction, respectively. In the literature, various approaches for vertical normalization are proposed. A local extrema based normalization approach for the vertical direction is described in [10] where a supervised learning method is used to classify local extrema into a set of five classes. From these classes, four reference lines, namely, upper baseline, lower baseline, ascender, and descender are extracted [13]. These four reference lines comprise the three zones which are normalized using linear scaling applied to each column of the image to a fixed height. Their method fails to maintain the aspect ratio of the word image. Vertical normalization for short sentences or isolated words is proposed in [18] where HMM/ANN-based method is used and ANN is trained using April-ANN toolkit [24]. HMM hybridized with ANN is applied column-wise to segment a word into three zones depending upon the four reference lines, namely, upper baseline, lower baseline, ascender, and descender. According to these zones, the word image is further normalized. A CNN-based approach for vertical normalization is proposed in [17]. Using CNN, pixels of scanned text line are classified for their belongingness to the main body area. The reference lines demarcating the pixels belonging to the main body area are obtained from local estimates by means of Dynamic Programming. According to the area delimited by these reference lines, scaling is done to a fixed height using linear interpolation. ANN-based approach for various preprocessing steps is proposed in [8]. For text normalization, local extrema of text contours are classified into five classes of reference lines using Multilayer Perceptrons. These lines comprise the zones which are normalized by linearly scaling them to a fixed height.

A technique for horizontal normalization is proposed in [14], where the number of black–white transitions in the horizontal direction is determined for each image line. Using the maximum number of transitions and the image width, the mean number of strokes per pixel is estimated. This number is set in relation to a reference value for the mean number, which is computed offline by statistics over the set of all the text lines in the image database. Scaling factor for horizontal normalization is set according to the relation between the actual and the reference value calculated by the abovementioned approach. In this approach, uniform horizontal normalization is carried out over the whole text line. This same method is adopted in [11, 25] for horizontal normalization but for the word and not on complete text lines. These methods provide uniform horizontal normalization such that the width of the text lines as in [14] and width of the word [11, 25] is of the same size depending on the scaling factor. This method does not resolve the issue of the nonuniformity of inter-character space between characters within a word. To address this issue, a method that applies nonuniform scaling is required.

The proposed method addresses the issue of inconsistent spacing between characters within a word. This scenario is common in handwritten words, however, more

specifically when the handwritten text is written in frame fields. In such cases, characters are sparsely written at the beginning of the word and compactly written toward the end of the word to accommodate text in frame fields. In our approach, interest points are marked to locate the sparsely and densely written part. Nonuniform scaling is used such that only horizontal strokes are scaled with least distortions to vertical strokes. The next section describes the proposed methodology. Section 4 illustrates the experimental setup established for our approach. Finally, the paper concludes with some concluding remarks and future work.

### 3 Methodology

The proposed methodology finds application in structured documents such as forms. Before proceeding to scale correction of handwritten word images, we need to locate and extract the handwritten words in a form. Our workflow comprises two main steps, namely, cell extraction, and horizontal-scale correction. Using the method proposed in [8] and modifying the region growing algorithm, form layout is identified for cell extraction. We make the assumption that out of the printed and handwritten words located in various frames of the form, we know the location of the frames having handwritten text written in it. So the scale correction is performed directly on the handwritten words.

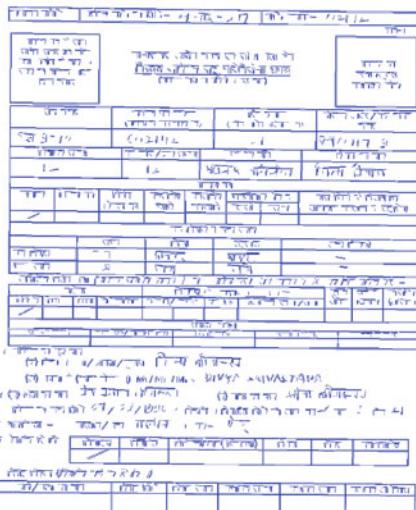
#### 3.1 Cell Extraction

The method given by [1] is used to detect and extract all the line segments present in an image. In the case of a structured document, it returns the line segments which form the cell boundaries. However, since the algorithm is very sensitive, all the line segments including straight strokes of character and shirorekha in case of Devanagari, are also extracted as a noise. During digitization, document image gets degraded and the cell boundary is distorted. It leads to extraction of the line segment in forms but with noise and distorted cell boundaries. Figure 1a shows a sample input form along with the line segments extracted in Fig. 1b using the method proposed in [1]. The enlarged view of a part of the form as shown in Fig. 2 shows the distorted cell boundaries termed as leaky gaps and extra line segments as a noise.

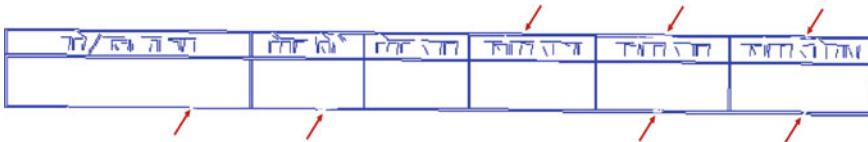
Region growing method described in [9] is used with some modification to extract various cells present in a form. In the region growing algorithm, there are two crucial aspects, seed point selection and criteria for selecting a pixel in a region. A seed point is needed to initiate the procedure. In this work, the conventional method of seed point selection is considered in which the first pixel is considered as a seed point. Modification is made in the second aspect of selection of pixel which is responsible for region expansion. Using the method given in [1] noisy gaps present in the frame due to broken lines allows the region to spill out from those gaps. As a result, cells

(a)

(b)



**Fig. 1** **a** Sample input form. **b** Lines segment extracted after applying [1] in form image

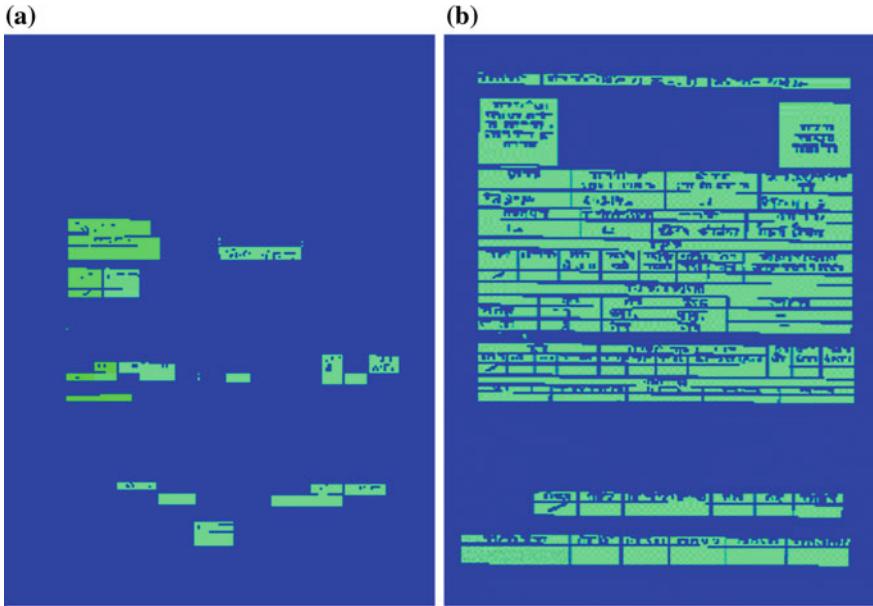


**Fig. 2** Enlarged image of a part of Fig. 1b showing noisy gaps (arrows) in frame lines

get merged and more than one cell forms a single region. This happens because for a pixel of some predefined type (background pixels in our case). If any of its 8 connected neighbors belongs to its type the region expands through that pixel. In our modified approach we impose the condition that, if the number of pixels determined by constant  $\alpha$  multiplied with 8 connected neighbors belongs to background pixels then adjacent pixel is set to new seed point and region expands. Value of  $\alpha$  is found to be 0.9 experimentally. Let the current seed point be denoted as  $g(x, y)$ , the next seed point as  $h(x, y)$ , the  $P(BG)$  be the number of pixels belonging to Background pixels in the 8 connected neighborhood. For modified region growing is given in Eq. 1:

$$h(x, y) = \begin{cases} BG, & \text{if } P(BG) \geq \alpha \times 8 \\ FG, & \text{otherwise} \end{cases} \quad (1)$$

This criterion prevents the region growing process to leak through noisy gaps in the lines such that all the cells of the frames are extracted as a separate region. Figure 3 shows cell extracted from original and modified region growing method.



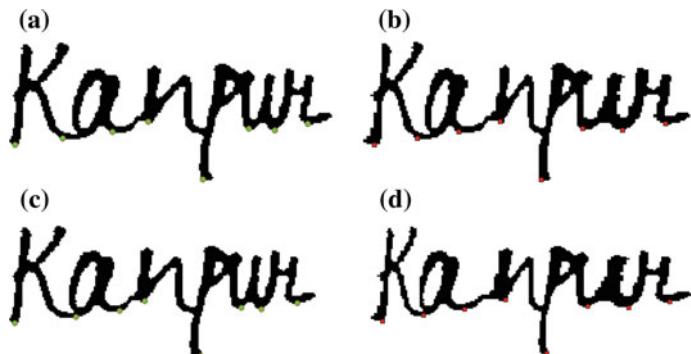
**Fig. 3** **a** Cell extraction using region growing approach in [1]. **b** Cell extraction using proposed modified region growing approach

### 3.2 Horizontal-Scale Correction

After extracting the cell, its content, i.e., the handwritten word is normalized so that the subsequent tasks such as word spotting, word recognition can be carried out efficiently. Inconsistent spacing within word leading to variations in horizontal scale needs to be corrected. Essentially there will be a need to shrink the sparsely written text and stretch the compactly written text within a word, ensuring that only the white space within the word is scaled, i.e., only horizontal-scale correction is done so that vertical strokes are minimally blurred.

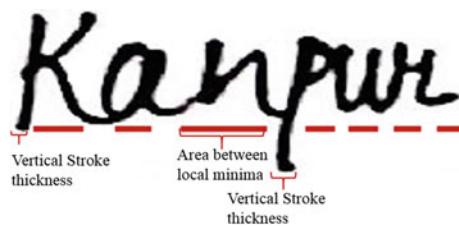
We need to scale the word in such a manner that scaling is done at locations, where the inter-character spacings are small. These locations where scaling is required are called key locations. We need to find these locations and for this purpose, interest points are computed. These interest points in a word are identified by extracting local minima of the lower contour of a word.

Scaling is applied to the area between the local minima using a window-based approach. This window has a height same as that of the image and a width equal to the distance between consecutive local minima. This results in windows that can have a variable width. Scaling factor is calculated from the relationship between actual distance and reference distance between the interest points. Reference distance is computed by taking the mean of all the distances between adjacent local minima.



**Fig. 4** **a** Input image with starting index of local minima marked with asterisk. **b** Horizontal scaled corrected obtained when starting index of local minima is considered. **c** Input image with last index of local minima marked with asterisk. **d** Horizontal scaled corrected obtained when last index of local minima is considered

**Fig. 5** Width of vertical strokes and selective scaling windows between location minima



Horizontal-scale correction inside a window for a given scale factor is done by the standard bicubic interpolation method.

While performing horizontal-scale correction, the vertical stroke should not be distorted. The local minima in a word image can correspond to the lowest coordinate of a vertical stroke or a horizontal stroke joining two characters or can be within a character. In a word image, each stroke has some thickness. Therefore for vertical strokes, the location of the minimum will extend to its thickness. Considering the starting index of local minimum of vertical stroke will lead to expansion or shrinkage of vertical stroke when scaled to the next local minimum. Considering the last index of local minimum of vertical stroke will also lead to this distortion when scaled to the previous local minimum. Both the cases where the starting index and where the last index of local minimum is considered are shown in Fig. 4. For this purpose, starting index of local minimum is taken as a key location when scaling is done with previous key location and end index is taken when scaling is done with next location. Hence, selective scaling is applied between these local minimum and not for the width of local minimum. This ensures that only horizontal strokes are scaled and vertical strokes are least distorted during the process. Figure 5 shows the width of the vertical stroke. Horizontal lines show the region where scale correction will be applied.

## 4 Experiment

We start by describing the dataset used for experiments. This is followed by a discussion on the implementation details. Finally, the efficiency of our system is described in the result section.

### 4.1 Dataset

The proposed approach is tested on two scripts, namely, English and Devanagari. The dataset for this specific problem is not available, so we have created our dataset comprising 200 words obtained from bilingual (English and Devanagari Script) forms of 10 different layouts with 10 forms of each layout written by 10 writers to incorporate inter writer variability to our dataset. Various inter-character space variations are made at different locations in our dataset. The method for word recognition in [2] for English script is adopted to validate our work. For training the recognition system, IAM dataset [15] is used. The system is trained on IAM dataset and then tested in our images of English scripts. For the purpose of training, 1000 images of IAM are used and 100 images of IAM and 100 images of our dataset are used for testing. Figure 6 shows sample images of our dataset.

### 4.2 Implementation

The handwritten text after horizontal normalization is fed into the word recognition system proposed in [2]. It verifies the utility of our framework in the recognition task. A quantitative and qualitative analysis is performed on the results. Recognition model [2] works only for English script, therefore, quantitative analysis is made for English script only whereas qualitative results are shown for both English and

**Fig. 6** Sample images from our dataset

Devanagari script. Word recognition method is trained with IAM images. Testing is done using 100 IAM and 100 English words images from our dataset.

Three types of experiments are performed for quantitative analysis. First, without applying horizontal normalization, word recognition is carried out using the method given in [2]. Next, the horizontal normalization proposed in [14] is performed on testing images and then the recognition of these normalized images is done. Then the same procedure is repeated but with normalization performed by the proposed approach. Experimental conditions are set similar to those mentioned in [2]. To evaluate the performance, Character Error Rate (CER) is calculated for each system. The CER between two words is defined as the edit distance or Levenshtein distance between them. It signifies the minimum number of character insertions, deletions, and substitutions needed to transform one word into the other, normalized by the length of the words.

The qualitative analysis of result is done for word images of both English and Devanagari script. The method proposed in [14] for horizontal normalization is taken as a baseline method.

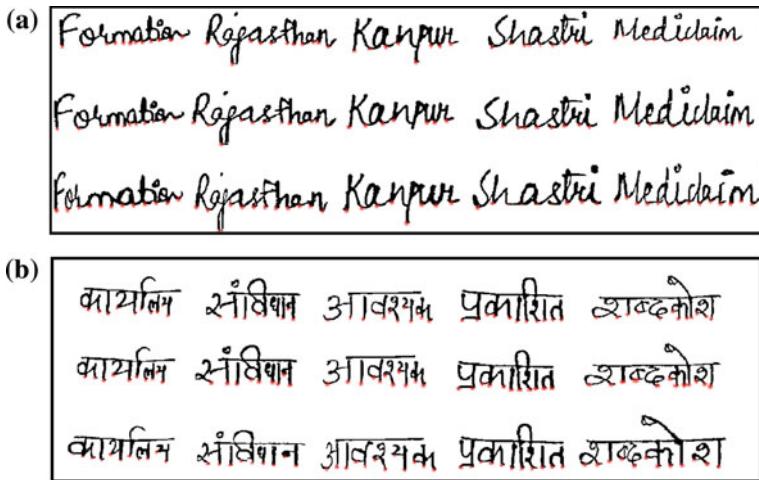
### 4.3 Results and Discussions

Table 1 illustrates the results obtained from word recognition task. CER metric is used for comparison where the lower score is better. CER metric is compared for the baseline method and for the method that does not use any horizontal normalization. Results prove that our horizontal-scale normalization step reduces the error rate in comparison to other approaches.

Qualitative results of horizontal-scale correction are shown in Fig. 7. The first row corresponds to the input image, middle row for results obtained by applying method proposed in [14], and the last row shows the resultant images obtained by applying our proposed method. Figure 7a shows results for English script and Fig. 7b for Devanagari script. Red asterisk depicts the interest point locations. It can be seen that in the first and middle row, the consecutive distance between adjacent interest points is unequal which is normalized in the last row. In the method proposed in [14], the maximum number of black to white transitions along the various text lines is considered which signifies the vertical strokes present in a word. This method has two shortcomings, first, the maximum number of transitions lead to an unnecessary number of strokes. Due to writer variability, the number of strokes for a word can vary among writers. The second shortcoming is that they are using uniform scaling, which

**Table 1** Recognition error rates

Method	CER
Without normalization Almazan et al. [2]	33.67
Using normalization of Bunke et al. [14]	29.58
Using proposed normalization	<b>26.31</b>



**Fig. 7** **a** Horizontal-scale correction in English script. **b** Horizontal-scale correction in Devanagari Script. Top row signifies the input word images, middle row the result obtained using the approach of [14], and bottom row shows the result from our approach. Red asterisk denotes the interest points

does not have the effect on scale correction within a word. The proposed method eliminates these by using nonuniform scale correction using the local minima of the lower contour of a word. As can be seen in Fig. 7 that in an upper and middle row, these interest points are at unequal distance to each other and in the bottom row all the interest points are at a constant distance to their adjacent interest points.

## 5 Conclusion

Preprocessing steps in document analysis are of great importance, which leads to increased accuracy for many important tasks like handwritten text recognition, word spotting, etc. The paper has proposed a method for horizontal-scale correction. This step is useful in every document analysis task. We have focused on the structured document, where to accommodate the handwritten text within frame, words are written with variable inter-character spacing. To address this issue, horizontal-scale correction approach is proposed so that the inter-character spacing is approximately equal. The proposed method works in two steps. In the first step the form fields called cells are extracted using the modified region growing algorithm, after which horizontal-scale correction is applied to the handwritten content of each cell. This horizontal-scale correction step is applied in the recognition system proposed by [2]. The reduced error rate shows its effectiveness in the recognition task. Also, our method is compared with the horizontal normalization technique proposed in [14] and by qualitative approach analysis, our method has proved to give a better

result. The proposed system has made an assumption that the location of printed and handwritten words in form is known in advance. For future work, handwritten and printed cell will be segmented and located automatically.

**Acknowledgements** The authors would like to thank the research scholars at IIT Jodhpur for their contribution toward dataset creation. All procedures performed in dataset creation involving research scholars at IIT Jodhpur were in accordance with the ethical standards of the institute.

## References

1. Akinlar, C., Topal, C.: Edlines: real-time line segment detection by edge drawing (ed). In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 2837–2840. IEEE (2011)
2. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(12), 2552–2566 (2014)
3. Bal, A., Saha, R.: An improved method for text segmentation and skew normalization of handwriting image. In: Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, pp. 181–196. Springer (2018)
4. Brink, A., Niels, R., Van Batenburg, R., Van den Heuvel, C., Schomaker, L.: Towards robust writer verification by correcting unnatural slant. *Pattern Recognit. Lett.* **32**(3), 449–457 (2011)
5. Cesarini, F., Gori, M., Marinai, S., Soda, G.: Structured document segmentation and representation by the modified XY tree. In: 1999 Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99, pp. 563–566. IEEE (1999)
6. Chen, J.L., Lee, H.J.: Field data extraction for form document processing using a gravitation-based algorithm. *Pattern Recognit.* **34**(9), 1741–1750 (2001)
7. Duygulu, P., Atalay, V.: A hierarchical representation of form documents for identification and retrieval. *Int. J. Doc. Anal. Recognit.* **5**(1), 17–27 (2002)
8. Espana-Boquera, S., Castro-Bleda, M.J., Gorbe-Moya, J., Zamora-Martinez, F.: Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(4), 767–779 (2011)
9. Gonzalez, R.C.: Woods: Digital Image Processing Using MATLAB, vol. 2. Pearson-Prentice-Hall Upper Saddle River, New Jersey (2009)
10. Gorbe-Moya, J., Boquera, S.E., Zamora-Martínez, F., Bleda, M.J.C.: Handwritten text normalization by using local extrema classification. *PRIS* **8**, 164–172 (2008)
11. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5), 855–868 (2009)
12. Hori, O., Doermann, D.S.: Robust table-form structure analysis based on box-driven reasoning. In: ICDAR, p. 218. IEEE (1995)
13. Likforman-Sulem, L., Zahour, A., Taconet, B.: Text line segmentation of historical documents: a survey. *Int. J. Doc. Anal. Recognit. (IJDAR)* **9**(2–4), 123–138 (2007)
14. Martí, U.V., Bunke, H.: Using a statistical language model to improve the performance of an hmm-based cursive handwriting recognition system. *Int. J. Pattern Recognit. Artif. Intell.* **15**(01), 65–90 (2001)
15. Martí, U.V., Bunke, H.: The iam-database: an english sentence database for offline handwriting recognition. *Int. J. Doc. Anal. Recognit.* **5**(1), 39–46 (2002)
16. Nagy, G., Seth, S.: Hierarchical Representation of Optically Scanned Documents (1984)
17. Pastor-Pellicer, J., Espana-Boquera, S., Castro-Bleda, M., Zamora-Martinez, F.: A combined convolutional neural network and dynamic programming approach for text line normalization. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 341–345. IEEE (2015)

18. Pastor-Pellicer, J., Espana-Boquera, S., Zamora-Martínez, F., Castro-Bleda, M.J.: Handwriting normalization by zone estimation using hmm/anns. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 633–638. IEEE (2014)
19. Perez-Cortes, J.C., Andreu, L., Arlandis, J.: A model-based field frame detection for handwritten filled-in forms. In: 2008 The Eighth IAPR International Workshop on Document Analysis Systems. DAS'08, pp. 362–368. IEEE (2008)
20. Sharma, D.V., Lehal, G.S.: Form field frame boundary removal for form processing system in Gurmukhi Script. In: 2009 10th International Conference on Document Analysis and Recognition. ICDAR'09, pp. 256–260. IEEE (2009)
21. Singh, C., Bhatia, N., Kaur, A.: Hough transform based fast skew detection and accurate skew correction methods. *Pattern Recognit.* **41**(12), 3528–3546 (2008)
22. Uchida, S., Taira, E., Sakoe, H.: Nonuniform slant correction using dynamic programming. In: 2001 Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 434–438. IEEE (2001)
23. Xi, D., Lee, S.W.: Extraction of reference lines and items from form document images with complicated background. *Pattern Recognit.* **38**(2), 289–305 (2005)
24. Zamora-Martínez, F.: April-ANN Toolkit a Pattern Recognizer in Lua Artificial Neural Networks Module (2013)
25. Zamora-Martinez, F., Frinken, V., España-Boquera, S., Castro-Bleda, M.J., Fischer, A., Bunke, H.: Neural network language models for off-line handwriting recognition. *Pattern Recognit.* **47**(4), 1642–1652 (2014)

# DeepAttent: Saliency Prediction with Deep Multi-scale Residual Network



Kshitij Dwivedi, Nitin Singh, Sabari R. Shanmugham and Manoj Kumar

**Abstract** Predicting where humans look in a given scene is a well-known problem with multiple applications in consumer cameras, human–computer interaction, robotics, and gaming. With large-scale image datasets available for human fixation, it is now possible to train deep neural networks for generating a fixation map. Human fixations are a function of both local visual features and global context. We incorporate this in a deep neural network by using global and local features of an image to predict human fixations. We sample multi-scale features of the deep residual network and introduce a new method for incorporating these multi-scale features for the end-to-end training of our network. Our model DeepAttent obtains competitive results on SALICON and iSUN datasets and outperforms state-of-the-art methods on various metrics.

**Keywords** Saliency • Neural networks • Multi-scale

## 1 Introduction

Visual attention mechanism in humans helps to selectively process and prioritize important parts of a massive amount of visual information in a given scene. This mechanism has evolutionary origins and aided early primates and humans in

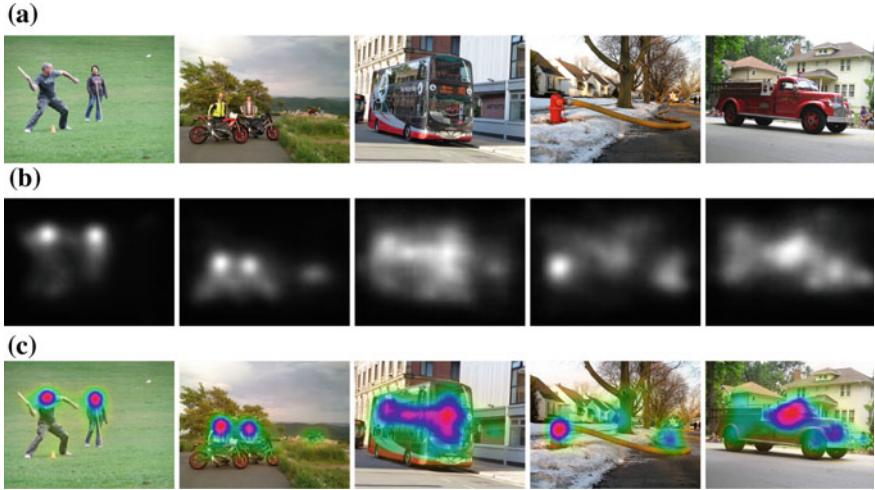
---

K. Dwivedi  
Singapore University of Technology and Design, Singapore, Singapore  
e-mail: [kshitijdwivedi93@gmail.com](mailto:kshitijdwivedi93@gmail.com)

N. Singh (✉) · M. Kumar  
Samsung Research Institute, Bengaluru, India  
e-mail: [nitin.ks@samsung.com](mailto:nitin.ks@samsung.com); [nitinksingh99@gmail.com](mailto:nitinksingh99@gmail.com)

M. Kumar  
e-mail: [manoj.kumar5@samsung.com](mailto:manoj.kumar5@samsung.com)

S. R. Shanmugham  
DataRobot, Singapore, Singapore  
e-mail: [sabari.shanmugam@datarobot.com](mailto:sabari.shanmugam@datarobot.com)



**Fig. 1** Image from COCO test set are shown in **a**, their saliency maps generated by our method are shown in **b** and the saliency map overlaid on the image is shown in **c**

detecting prey, predators, food, and other important objects required for survival. Emulating selective visual attention using computational models can be useful in several computer vision tasks such as autofocus, image compression, localization, thumbnail generation, multiple object detection, and active vision in robotics.

The computational models of attention mechanism aim at generating saliency maps which encode the relative saliency or conspicuity of objects in a visual environment [9]. In previous works, saliency has been defined as a salient object segmentation task [1] as well as a fixation prediction task [9]. In this work, we have focused on fixation prediction task by training our models on fixation ground truth from large-scale image datasets (Fig. 1).

To predict the saliency map, Itti et al. [9] used low-level features of an image such as color, intensity, and orientations. Zhang et al. [6] used Fourier statistic of an image. Later, Cerf et al. [4], Judd et al. [13] incorporated semantics in their model with prior information of faces and other objects of saliency prediction. The early success of deep neural networks for image classification and object detection attracted researchers to apply these features for fixation prediction. eDN [25] was the first method to use DNNs for predicting fixations. Their method uses features from three different convolutional layers. Deepgaze [15] used transfer learning for the fixation prediction task due to lack of labeled fixation datasets. Their model utilized features from each convolution layer of Alexnet trained on IMAGENET dataset. Using these feature maps, a linear classifier is trained to predict fixation maps. Jiang et al. [12] introduced a new fixation dataset which was then used in different works to improve saliency benchmark. Pan et al. [20] used a completely data-driven approach by training a convolutional neural network for saliency prediction. Salicon [8] emphasized on using information from multiple scales. Kruthiventi et al. [14] introduce Deepfix model which uses VGG [23] and inception [24] architecture.

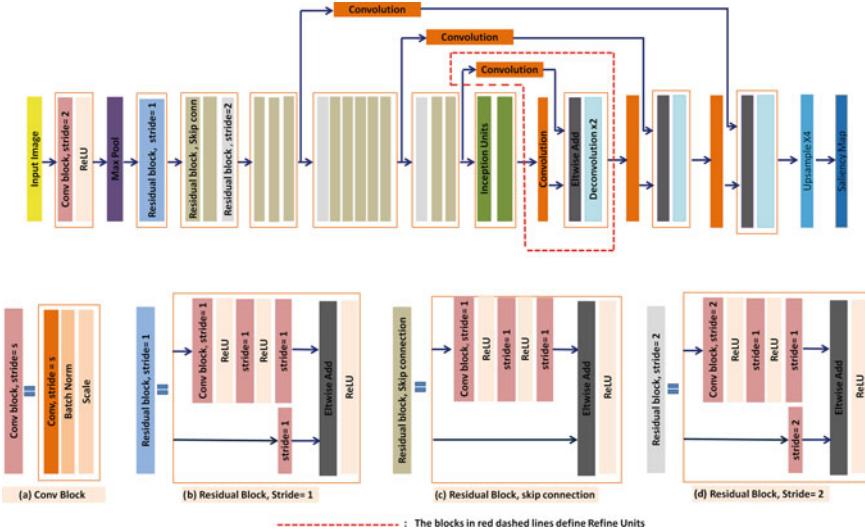
To utilize both local image features and global context, classical models of Achanta et al. [1] and Itti et al. [9] use information from multiple scales. The Deepfix model uses inception layer, which has multiple kernels of varying sizes and thus has multiple receptive fields. The inception layer merges information from multiple scales at the output layer. However, the inception layer receives a feature map of low resolution as input and thus lack the high-frequency information of high-resolution feature maps.

In this work, we explore refine units [22] to use features from multiple scales in a deep neural network and generate high-resolution saliency maps. Our algorithm does not require any extra information and it integrates both low-level and high-level features in an end-to-end framework to generate saliency maps. We evaluate the performance of our model on SALICON [12] and iSUN [27] datasets against the DNN based state of the art methods for fixation prediction.

The paper is arranged as follows. In Sect. 2, we explain our model for saliency detection. In Sect. 3, we provide the experimentation details and analyze the results of the network with state-of-the-art methods. Finally in Sect. 4, we conclude our paper with the summary of the approach and the results.

## 2 Model

The model we develop, DeepAttent (shown in Fig. 2) is a fully convolutional neural network and consists of three building blocks. We use a 50-layer Residual network as a backbone network, Inception units for context aggregation, and Refine units to



**Fig. 2** DeepAttent: CNN network for saliency prediction

combine low-level and high-level feature maps. Below we describe these building blocks in detail:

## 2.1 Residual Network

One way to improve the performance of a deep neural network is to increase its size which can be increased both by increasing the number of layers or the number of neurons at each layer. However, training very deep networks suffer from optimization difficulties due to vanishing gradients. He et al. [7] addressed this problem by using an identity connection that bypasses a few convolution layers at a time. A residual unit is formed by each such bypass in which convolution layer predicts a residual, which is added to next residual units input tensor. He et al. [7] trained 152 layers deep residual architecture to achieve state-of-the-art accuracy in image classification, object detection, and segmentation. In purview of the memory constraints, we use a smaller 50-layer version of the residual model trained on Imagenet [5] dataset, for fixation prediction.

## 2.2 Inception Units

We use inception modules [24] similar to Deepfix model [14] to capture the multi-scale information from the last convolution layer of the residual network. The inception module takes a convolutional feature map as input and uses kernels with receptive field of  $1 \times 1$ ,  $3 \times 3$ ,  $5 \times 5$  in parallel to capture a variety of structures. Thus, we obtain multi-scale information from the hierarchical deep features using the inception units. In Deepfix model, though inception architecture captures multi-scale information, it only receives a feature map of low resolution as input and thus lacks the high frequency information from the high resolution features. To address this issue, we use refine units which merge features from multiple scales to generate a high resolution saliency map.

## 2.3 Refine Units

We use refine units proposed by Pedro et al. [22] to simultaneously upsample the fixation maps while utilizing the rich information from hierarchical feature maps. The refine units stated in Pedro et al. [22] merge feature maps and segmentation maps to generate higher resolution segmentation maps. We use a similar strategy for fixation map refinement. The refine units (as shown in red-dashed line in Fig. 2) generate accurate upsampled fixation maps by merging information from feature maps of different scales. The input to refine units is a high-level fixation feature map

and low-level convolutional feature map of same spatial dimensions. The fixation map and the feature map then undergo a convolution operation to match the number of channels for element-wise addition. The merged feature map is then upsampled using a deconvolution layer to generate a higher resolution fixation map.

### 3 Experiments and Results

In this section, we analyze the results obtained using the proposed architecture. We first describe the dataset used for the experiment. We then provide information of the frameworks used in the experiment and the experimentation details. We finally evaluate the results of our model in comparison to other state-of-the-art algorithms on multiple datasets.

#### 3.1 Dataset

We use SALICON [12] and iSUN [27] datasets to train and evaluate the performance of our model. The two datasets have different methods for creating the ground truth annotations of saliency. The SALICON dataset (subset of MS COCO [18] dataset) contains 10000 training, 5000 validation, and 5000 test images. The ground truth saliency annotations of these images have been generated using mouse trajectories instead of eye trajectories, thus enabling large-scale data collection of saliency ground truths. iSUN dataset (subset of SUN [26] database) has 6000 training, 926 validation, and 2000 test images. This dataset has been annotated using a webcam-based gaze tracking system that supports large-scale, crowdsourced eye tracking deployed on Amazon Mechanical Turk (AMTurk) [3].

#### 3.2 Training

We train our model using the Caffe [11] framework on a NVIDIA GeForce GTX TITAN X system, which supports 6 GPU Cards each with 12GB of memory. We use a Euclidean loss which computes the difference between the annotated ground truth and the output to train the network. Thus, our network tries to minimize the difference between the ground truth and output at every pixel. We use a batch size of 5 and the following hyperparameters,  $lr: 10^{-5}$ ;  $lr\ policy: step$ ;  $stepsize = 1000$ ;  $momentum = 0.9$  and  $gamma = 0.1$ , to train our network. Our model takes 3 hours for training for SALICON train dataset and 50 ms for generating fixation maps for a image size of  $480 \times 640$  on the TitanX GPU.

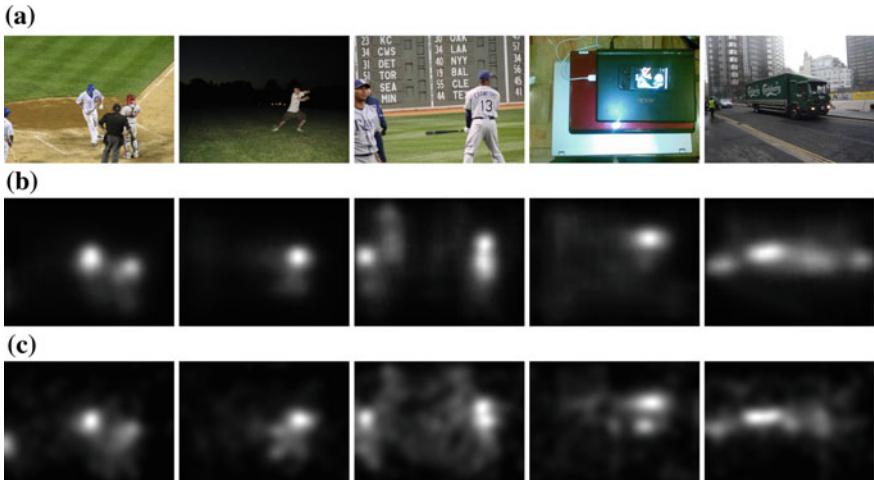
### 3.3 Quantitative Comparison

We do a comparative study of our network with other networks on the SALICON validation dataset. The quantitative comparison on various metrics on this dataset is shown in Table 1. The qualitative results on the COCO test dataset and COCO validation dataset are shown in Figs. 1 and 3, respectively. The quantitative comparison of the results on the test datasets of SALICON and iSUN are shown in Tables 2 and 3, respectively. The results on SALICON and iSUN test datasets were submitted to Large-Scale Scene Understanding Challenge held in CVPR 2016. These results are publicly available on the LSUN leaderboard [19].

As shown in Table 1, our model improves over DeepFix [14] (VGG + Inception) model by 5.8 % in AUC, 2.6% in sAUC, 0.237 in CC and 0.641 in NSS metric. These metrics were evaluated on the 5000 images of SALICON validation dataset. Tables 2 and 3 show that our model outperformed various models in the LSUN challenge. The

**Table 1** Results on SALICON validation dataset

Model	AUC [13] (%)	CC [17]	NSS [21]	sAUC [2] (%)
Deepfix (VGG+Inception)	83.2	0.531	1.843	75.3
ResNet+Inception	87.3	0.678	2.026	77.1
ResNet+Inception+Skip connections	87.5	0.690	2.076	77.6
<b>DeepAttent</b> (ResNet+Inception+Refine Units)	<b>89.0</b>	<b>0.768</b>	<b>2.484</b>	<b>77.9</b>



**Fig. 3** Image from COCO validation set are shown in **a**, their saliency maps generated by our network are shown in **b** and the saliency ground truth is shown in **c**

**Table 2** Results on SALICON test dataset

Model	AUC (%)	CC	IG [16]	sAUC (%)
<b>DeepAttent</b>	<b>76.7</b>	<b>0.890</b>	<b>0.326</b>	63.1
VAL [14]	76.1	0.804	0.315	63.0
NPU [19]	75.6	0.775	0.318	<b>63.7</b>
XRCE [10]	75.6	0.822	0.304	63.2
UPC [20]	75.5	0.797	0.292	63.6
ML-Net [19]	74.8	0.724	0.274	63.3
Donders [19]	74.8	0.767	0.247	62.7
HUCVL [19]	74.7	0.825	0.197	59.8
SDUVSIS [19]	74.4	0.735	0.179	59.9
VLL [19]	73.2	0.766	0.161	59.9

**Table 3** Results on iSUN test dataset

Model	AUC (%)	CC	IG	sAUC (%)
<b>DeepAttent</b>	<b>86.2</b>	0.814	0.174	55.0
VAL	86.2	0.815	<b>0.179</b>	<b>55.0</b>
NPU	86.1	<b>0.815</b>	0.156	55.0
UPC	86.0	0.798	0.136	54.1
XRCE	85.5	0.787	0.102	53.8
SDUVSIS	85.0	0.788	0.051	52.1
Donders	83.5	0.668	0.002	54.4

models were tested on the SALICON test dataset and iSUN test dataset. Our model shows an average increase by 1.73% in AUC, 0.110 in CC and 0.072 in IG metric on the SALICON test dataset. On the iSUN dataset we show an average increase of 0.8% in AUC and 0.94% in sAUC metric.

## 4 Conclusion

In this paper, we propose a model to emulate human visual attention mechanism using a multi-scale deep residual network which predicts eye fixations for a given scene. We show that merging information from representations at different scales improves the fixation prediction significantly. We benchmark our model with the current state of the art models of fixation prediction and outperform them in several metrics. Our model, DeepAttent performed very well in the Saliency challenge of Large-Scale Scene Understanding Challenge (LSUN) workshop hosted in Computer Vision and Pattern Recognition (CVPR) conference 2016.

## References

1. Achanta, R., Hemami, S., Estrada, F., Sussstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 1597–1604. IEEE (2009)
2. Borji, A., Tavakoli, H.R., Sihite, D.N., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 921–928. IEEE (2013)
3. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon’s mechanical turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* **6**(1), 3–5 (2011)
4. Cerf, M., Frady, E.P., Koch, C.: Using semantic content as cues for better scanpath prediction. In: Proceedings of the 2008 symposium on Eye tracking research and applications, pp. 143–146. ACM (2008)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 248–255. IEEE (2009)
6. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008, pp. 1–8. IEEE (2008)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 262–270 (2015)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
10. Jetley, S., Murray, N., Vig, E.: End-to-end saliency mapping via probability distribution prediction. In: Proceedings of Computer Vision and Pattern Recognition 2016, pp. 5753–5761 (2016)
11. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
12. Jiang, M., Huang, S., Duan, J., Zhao, Q.: Salicon: saliency in context. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1072–1080. IEEE (2015)
13. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113. IEEE (2009)
14. Kruthiventi, S.S., Ayush, K., Babu, R.V.: Deepfix: a fully convolutional neural network for predicting human eye fixations. *IEEE Trans. Image Process.* **26**(9), 4446–4456 (2017)
15. Kümmeler, M., Theis, L., Bethge, M.: Deep gaze i: boosting saliency prediction with feature maps trained on imagenet. [arXiv:1411.1045](https://arxiv.org/abs/1411.1045) (2014)
16. Kümmeler, M., Wallis, T.S., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. *Proc. Nat. Acad. Sci.* **112**(52), 16054–16059 (2015)
17. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level visual features. *Vision Res.* **47**(19), 2483–2498 (2007)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
19. LSUN’16: Large-scale scene understanding challenge: Leaderboard (2016). [http://lsun.cs.princeton.edu/leaderboard/index\\_2016.html#saliencySalicon](http://lsun.cs.princeton.edu/leaderboard/index_2016.html#saliencySalicon). Last accessed 20 Apr 2018
20. Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., O’Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 598–606 (2016)

21. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. *Vision Res.* **45**(18), 2397–2416 (2005)
22. Pinheiro, P.O., Lin, T.Y., Collobert, R., Dollár, P.: Learning to refine object segments. In: European Conference on Computer Vision, pp. 75–91. Springer (2016)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
24. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going Deeper with Convolutions. In: CVPR (2015)
25. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2798–2805 (2014)
26. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3485–3492. IEEE (2010)
27. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: TurkerGaze: Crowd-sourcing saliency with webcam based eye tracking. arXiv preprint [arXiv:1504.06755](https://arxiv.org/abs/1504.06755) (2015)

# Copy–Move Image Forgery Detection Using Gray-Tones with Texture Description



Anuja Dixit and Soumen Bag

**Abstract** Copy–move forgery is a well-known image forgery technique. In this image manipulation method, a certain area of the image is replicated and affixed over the same image on different locations. Most of the times replicated segments suffer from multiple post-processing and geometrical attacks to hide sign of tampering. We have used block-based method for forgery detection. In block-based proficiencies, image is parted into partially overlapping blocks. Features are extracted corresponding to blocks. In the proposed scheme, we have computed Gray-Level Co-occurrence Matrix (GLCM) for blocks. Singular Value Decomposition (SVD) is applied over GLCM to find singular values. We have calculated Local Binary Pattern (LBP) for all blocks. The singular values and LBP features combinedly construct feature vector corresponding to blocks. These feature vectors are sorted lexicographically. Further, similar blocks discovered to identify replicated section of image. To ensure endurance of the proposed methods, Detection Accuracy (DA), False Positive Rate (FPR), and F-Measure are calculated and compared with existing methods. Experimental results establish the validity of proposed scheme for precise detection, even when meddled region of image sustain distortion due to brightness change, blurring, color reduction, and contrast adjustment.

**Keywords** Copy–move image forgery · Feature extraction · Image forensics · Local binary pattern · Singular value decomposition

---

A. Dixit · S. Bag

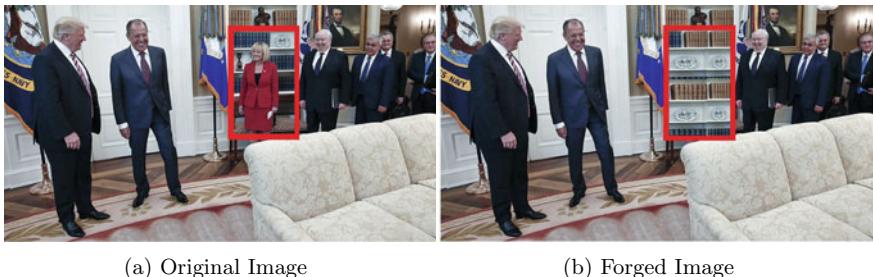
Department of Computer Science and Engineering, Indian Institute of Technology  
(Indian School of Mines), Dhanbad, India  
e-mail: [anu2010cse1@gmail.com](mailto:anu2010cse1@gmail.com)

S. Bag  
e-mail: [bagsoumen@gmail.com](mailto:bagsoumen@gmail.com)

## 1 Introduction

Copy-move forged [1] image is obtained by imitating a part of image and gluing it at different locations upon the same image as shown in Fig. 1. The principal motive behind such kind of forgery technique is to replicate the objects of image or to conceal the selective information rendered through image. Copy-move forgery follows above perception, because of which forged portion has a constitutional resemblance to rest of the image. As forged section has implicit resemblance to rest portion of the image which made detection of such kind of forgery a complex process. Copy-move forged images can be fabricated using image processing software without leaving any trace of forgery. Tampered region of the image may go through several geometrical and post-processing attacks to make detection of forgery a complicated task. Detection of copy-move forged images is integrative part of image forensics [2]. Significant amount of research work is practiced in this field. Fridrich et al. [3] designed pioneer technique for sleuthing copy-move forgery. Their method extracts discrete cosine transform (DCT) features of overlapping blocks. The features are sorted and resemblance among feature vectors is computed to detect fiddled areas. Alkawaz et al. [4] suggested a colligated algorithm that also uses DCT coefficients as features for distinct blocks of varying sizes. Both of the techniques possess high computational complexity and inappropriate identification of altered areas when manipulated areas of image suffer from post-processing operations. Zandi et al. [5] proposed a method utilizing adjustive similarity threshold. Their method utilized standard deviation of the image blocks for computation of threshold proportions. In [6], Lee et al. suggested a strategy employing histogram of orientated gradients (HOG) to spot tampered areas. Their algorithm performed well when altered image suffer from small degree rotations, blurring, brightness adjustment, and color reduction. Silva et al. [7] applied point of interest as well as blocks of pixels for forged region detection. They utilized voting process in multiscale space.

From the literature survey, we observed that copy-move forgery detection results may sustain false matches (which are incorrectly detected as forged even if primitives they are not forged region of image). To increase the detection accuracy and



**Fig. 1** An instance of region duplication image forgery (tampered region is outlined with red color rectangle)

reduce false matches while forged image enduring several post-processing attacks (blurring, contrast modification, luminance change, color diminution, etc.) is an open-research subject to cultivate. To minimize the presence of fictitious matches, we have used lexicographical sorting with Euclidean distance calculation accompanied with computation of shift vector. For precise localization of forged region, we have used hybridized feature extraction technique in which we have combined singular values (obtained from decomposition over GLCM) with LBP features. Proposed approach is invariant to post-processing attacks and obtained improved detection accuracy as well as reduced false matches than former techniques.

GLCM [8] is a texture feature extraction technique, which represents the frequency of occurrence of one gray tone, in a delimited spatial linear relationship with other gray tones in image. SVD [9] possess algebraic and geometric invariant properties. For a given matrix, unique singular values [10] are obtained which is useful in firm representation for image blocks. Most important data of image block is possessed by largest singular values (LSV) while small singular values are sensitive towards noise. LSV possess effective stability even when an image endure minor distortions. LBP [11] operator facilitates integrated description for a texture patch with structural and statistical features. LBP is highly discriminative texture operator [12] which analyze numerous patterns corresponding to each pixel with its neighborhood. The signed difference between center and neighboring pixel value is invariant to variation in mean luminance as well as change in grayscale.

This paper is structured as follows. Section 2 illustrates the proposed algorithm. Section 3 shows the experimental outcome validating the legitimacy of the proposed approach by utilizing evaluation metrics. Finally, the proposed study is concluded in Sect. 4.

## 2 Proposed Methodology

Figure 2 presents the fundamental framework followed for copy–move forgery detection. The proposed technique utilize following steps:

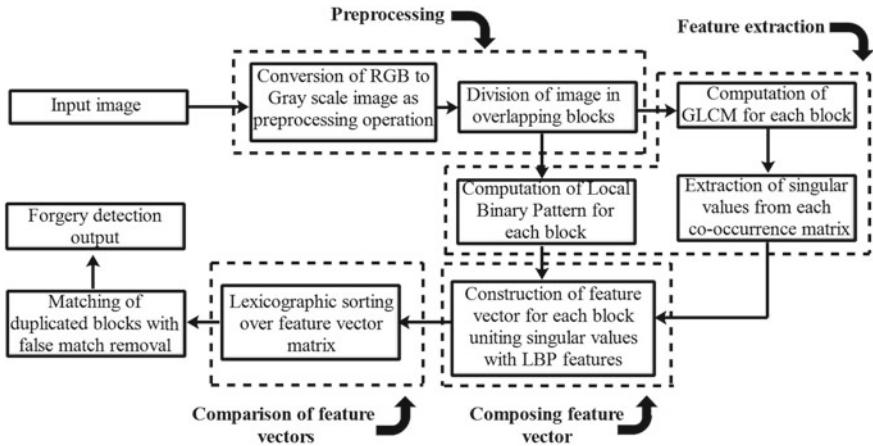
### 2.1 Preprocessing

#### 1. Conversion of color channel:

Initially, if the input image is RGB then it is converted to grayscale image.

#### 2. Division of image in overlapping blocks:

After preprocessing step, image is parted into partially overlapping blocks. If the gray scale image is of dimension  $M \times N$  and block size used for division of image is  $B \times B$  then total number of blocks are  $(M - B + 1) \times (N - B + 1)$ .



**Fig. 2** Framework of the proposed methodology

## 2.2 Feature Extraction

### 1. Computing gray-level co-occurrence matrix:

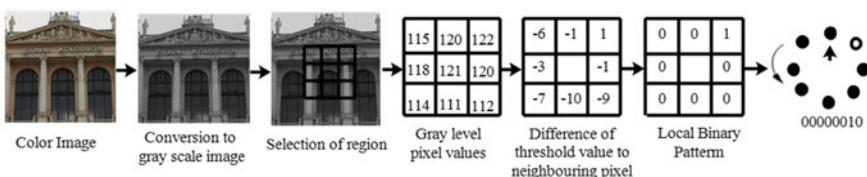
For each block, GLCM [8] is obtained to represent spatial relationship between the gray tones within each image block. The size of GLCM depends on maximum variation in gray level values of pixel within a block. For a block of dimension  $B \times B$ , co-occurrence matrix of size  $B \times B$  is incurred.

## 2. Calculation of singular values:

SVD [9] is applied over co-occurrence matrix. As result of decomposition, GLCM decomposed in three components: left unitary, right unitary, and diagonal matrix. Diagonal matrix contains singular values [10] on diagonal positions. Singular values represent feature vectors for all blocks of image. For a  $B \times B$  dimensional matrix feature vector obtained is of length  $B$ .

### 3. Extraction of LBP features and fusion with singular values:

Blocks are processed using LBP [11] as shown in Fig. 3. LBP features [12] are stored for each block. Singular values prevailed using SVD, and LBP features combinedly formulate feature vector to represent each block of image.



**Fig. 3** Working rationale of local binary pattern operator

### 2.3 Comparison of Feature Vectors

Feature vectors for overlapping blocks are placed in matrix  $FM$ . The dimension of feature vector matrix  $FM$  is  $(M - B + 1)(N - B + 1) \times len$ , where ‘len’ represents length of feature vector. For robust localization of similar feature vectors of duplicated blocks, Lexicographical sorting is applied over feature matrix. As a resultant, we achieve similar feature vector settled at neighboring locations.

$$FM = \left\{ \begin{array}{c} FV_1 \\ FV_2 \\ FV_3 \\ \vdots \\ FV_{(M-B+1)(N-B+1)} \end{array} \right\}$$

### 2.4 Similarity Detection and False Match Removal

#### 1. Duplicate region detection:

Euclidean distance is measured to find similarity between feature vectors as indicated in Eq. 2.

$$D(F, F') = \left( \sum_{i=1}^{n_f} (F_i - F'_i)^2 \right)^{\frac{1}{2}} \quad (2)$$

where  $F$  and  $F'$  shows feature vectors obtained from blocks and  $n_f$  represent the length of feature vector. To detect similar feature vectors from sorted feature matrix threshold value is used, such that  $D(F, F') \leq T_{dis}$ . To detect similar regions of image, shift vectors between similar blocks are measured. The top left corner location of a block is deliberated as it’s emplacement within image. Let,  $(i_1, j_1)$  and  $(i_2, j_2)$  be the coordinates of blocks  $FM$ . Shift vector between two co-ordinates can be obtained as in Eq. 3.

$$S = (s_1, s_2) = (i_1 - i_2, j_1 - j_2) \quad (3)$$

Both Shift vectors  $-S$  and  $S$  symbolize same order of magnitude for shifting. So, displacement vectors are normalized by multiplying with  $-1$  to maintain  $S \geq 0$ . A counter value  $C$  is initialized with zero. For similar shift vector between neighboring blocks, counter value increased by 1 as in Eq. 4.

$$C(s_1, s_2) = C(s_1, s_2) + 1 \quad (4)$$

## 2. Removal of false matches:

Counter value corresponding to different shift vector represents the frequency of similar shifting between block pairs. The proposed method detects all normalized shift vectors  $S_1, S_2, S_3, \dots, S_r$ . Shift vectors whose occurrence is higher than user-defined threshold  $T_{sh}$  shows forged blocks.  $C(S_k) > T_{sh}$  for  $k = 1, 2, \dots, R$ . High value of  $T_{sh}$  may leave altered regions undetected, while low value of  $T_{sh}$  can give rise to false matches. We employ a color map for localizing forged blocks of image.

## 3 Experimental Results and Discussion

### 3.1 Dataset

The source images for our experiments accumulated from CoMoFoD dataset [13]. Experiments are carried out on images with dimension  $512 \times 512$ . We selected 60 images suffering from color reduction, 60 images bearing contrast adjustment, 60 images with blurring attack, and 60 images with brightness change attack. In CoMoFoD dataset, color reduction images are divided into three categories with color channels contracted from 256 to 128, 64, and 32. Images with contrast adjustment also have three ranges [0.01, 0.95], [0.01, 0.9], and [0.01, 0.8]. Image with blurring attacks also have three categories based on average filter size as  $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ , and images with illumination change attack classified in three categories [0.01, 0.95], [0.01, 0.9], and [0.01, 0.8].

### 3.2 Experimental Setup

All experiments are executed over MATLAB 2016a, installed upon a platform equipped with 64-bit Windows (8GB RAM) and Intel core i7 processor. We set every parameter as  $B = 8$ ,  $t_{dis} = 50$ , and  $t_{sh} = 50$ . Various size of forged images can be used as input to proposed method, so values of  $M$  and  $N$  may vary. As here we consider all input images with size  $512 \times 512$  so,  $M = 512$  and  $N = 512$ . From experimental results, we found that when  $B = 4$ , too many false matches appear. when  $B$  increased to 16 then detection accuracy is compromised.  $T_{sh}$  is used for locating groups of blocks with similar shifting and meaningful forged region detection.  $t_{dis}$  is used for spotting similar feature vectors. Higher value of  $t_{dis}$  results in detection of different feature vectors as similar whereas smaller values of  $t_{dis}$  perform rigorous detection of similar feature vectors which may stipulate slightly dissident feature vectors as different. Here, length of extracted feature vector ‘len’ is 67.

### 3.3 Performance Evaluation

To demonstrate the outcomes of proposed technique, we have computed DA, FPR, and F-Measure by comparing the forgery detection outcomes with ground truth. DA, FPR, and F-Measure can be calculated as in Eqs. 5, 6, and 7 respectively.

$$DA = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$FPR = \frac{FP}{TN + FP} \quad (6)$$

$$F\text{-Measure} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (7)$$

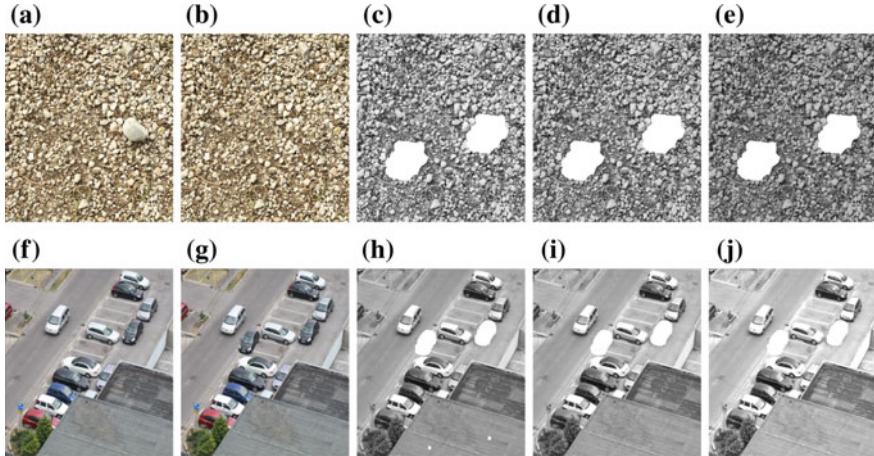
True Positive (TP) symbolizes the number of pixels correctly perceived as spoofed. False Positive (FP) represents the number of pixels which are falsely identified as faked but primitively they are not forged. False Negative (FN) shows the number of pixels which are manipulated but not detected. True Negative (TN) indicates number of pixels correctly detected as not forged. DA symbolizes the measure to which forgery detection algorithm can detect the altered region of image correctly, whereas FPR represents the percentage of pixels falsely detected as altered. F-Measure expresses the alliance of number of accurately detected forged pixels with respect to number of forged pixels perceived, and relevant forged region detection (i.e., proportion of correct detection of tampered pixels with respect to number of altered pixels acquainted by ground truth of forged image.).

### Contrast Adjustment and Brightness Change Attacks

Table 1 illustrates average quantitative results obtained through the proposed method for forgery detection when images are enduring contrast adjustment and brightness change alterations. Qualitative outcomes incurred applying the proposed method are pictured in Fig. 4.

**Table 1** Average DA and FPR results for contrast adjustment and brightness change attack

Contrast adjustment attack (Range)	DA (%)	FPR (%)	F-Measure (%)	Brightness change attack (Range)	DA(%)	FPR (%)	F-Measure (%)
[0.01, 0.95]	99.6647	1.0154	77.074	[0.01, 0.95]	97.8172	1.2816	68.269
[0.01, 0.90]	99.0486	2.1989	75.163	[0.01, 0.90]	95.5021	2.1989	67.725
[0.01, 0.80]	98.2311	2.6115	74.006	[0.01, 0.80]	94.9635	2.646	65.318



**Fig. 4** **a, f** shows untampered image. **b, g** shows corresponding forged images. **c, d**, and **e** shows results for forgery detection in case of contrast modification category range as [0.01–0.8], [0.01–0.9], and [0.01–0.95] respectively. **h, i**, and **j** shows results for forgery detection in case of tampered images with luminance change range as [0.01–0.8], [0.01–0.9], and [0.01–0.95] respectively

**Table 2** Average DA and FPR results for image blurring and color reduction attack

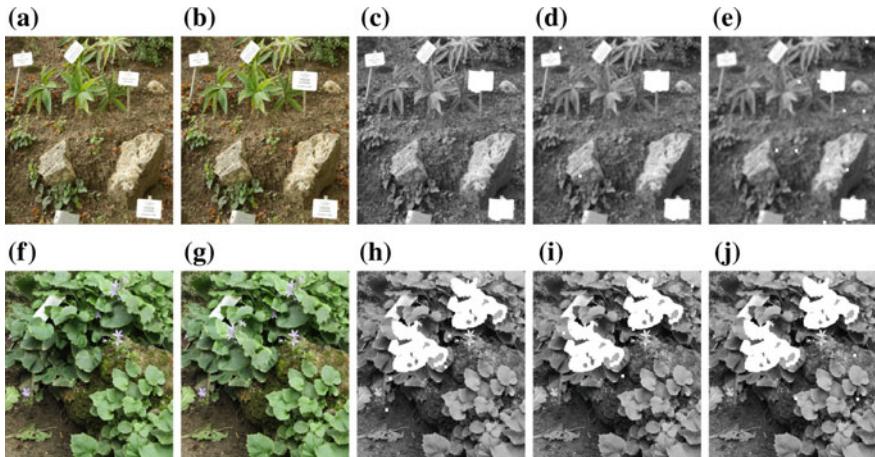
Image blurring attack (Filter size)	DA (%)	FPR (%)	F-Measure (%)	Color reduction attack (Level)	DA (%)	FPR (%)	F-Measure (%)
3 × 3	98.6736	0.9127	65.077	32	97.044	2.5684	64.810
5 × 5	97.0562	1.2478	64.008	64	97.8699	1.3162	67.295
7 × 7	96.7445	2.6115	61.016	128	98.8578	0.7159	71.627

### Image Blurring and Color Reduction Attacks

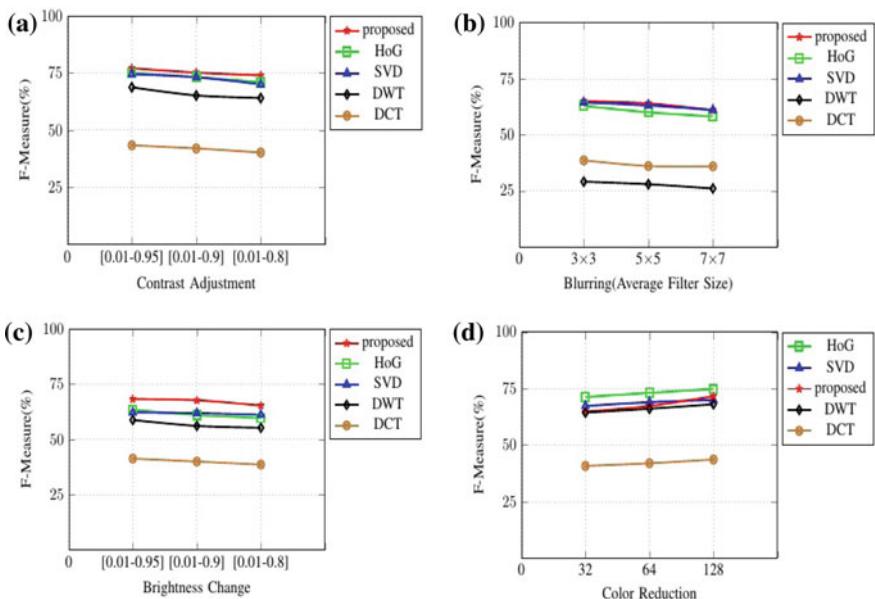
Table 2 shows average quantitative results obtained using the proposed method for forgery detection when images are suffering from image blurring and color reduction attack. Qualitative results incurred using proposed method are shown in Fig. 5.

### 3.4 Comparative Analysis

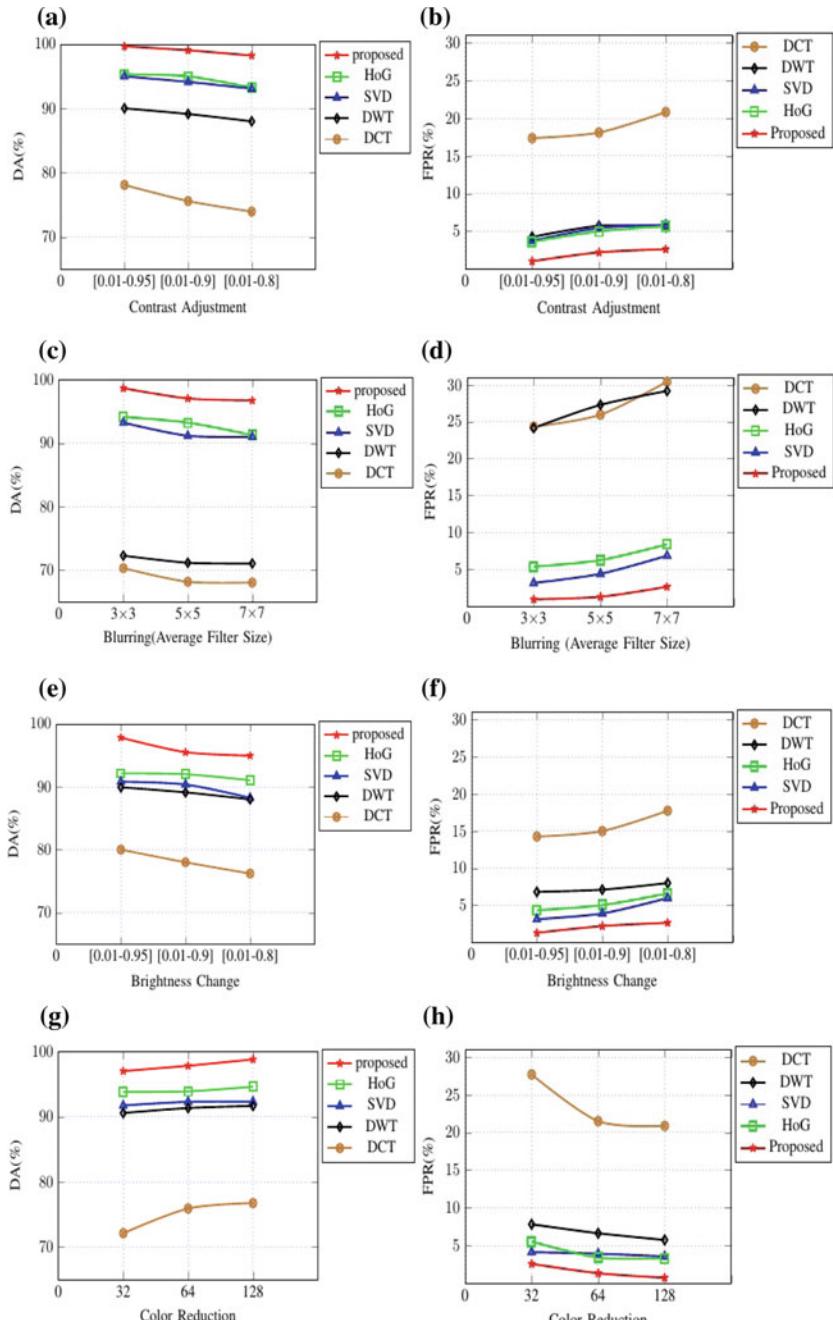
Experimental outcomes establish the fact that our method is an efficacious technique for copy-move forgery detection. In comparison to other features based on DCT [3], SVD [11], HOG [6], and discrete wavelet transform (DWT) [14], our method



**Fig. 5** **a, f** displays original image. **b, g** represents forged images. **c, d**, and **e** show results for forgery detection in case of image blurring with filter size  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$ , respectively. **h, i**, and **j** shows results for forgery detection in case of color reduction with level 32, 64, and 128 respectively



**Fig. 6** **a, b, c**, and **d** portray comparative results for F-measure when image endures contrast adjustment, image blurring, brightness change, and color reduction attacks, respectively



**Fig. 7** **a, c, e, and g** show comparative results for DA, whereas **b, d, f, and h** represent comparative outcomes for FPR when forged image bears contrast adjustment, blurring, luminance change, and color diminution attacks, respectively

achieved highest DA and FPR rates for forgery detection, when altered image is suffering from contrast adjustment, image blurring, brightness change, and color reduction attacks. For color reduction attack, HoG-based features obtained highest F-measure for all color reduction levels. SVD features achieved better results for F-Measure than proposed method toward color reduction levels 32 and 64. Figure 6 displays comparative results obtained for F-measure when image endures post-processing attacks, whereas Fig. 7 shows comparative results for DA and FPR.

The complexity of copy-move forgery detection algorithms primarily depends on number of blocks accessed and length of feature vector generated, which are represented by  $n_b$  and  $n_f$  respectively. For input image size  $512 \times 512$ , the proposed algorithm obtains  $n_f = 67$  and  $n_b = 255,055$ . DCT-based feature extraction method possess  $n_f = 64$  and  $n_b = 255,055$ . For SVD-based method  $n_f = 8$  and  $n_b = 255,055$ . Forgery detection scheme using HoG features obtains  $n_f = 4$  and  $n_b = 247,009$ . DWT-based method acquires  $n_b = 62,201$  and  $n_f = 64$ . For recording forgery detection time, we performed forgery detection operation over image size  $256 \times 256$ . The detection time for proposed algorithm is 98.65 s. For HoG, DWT, DCT, and SVD based methods detection times are 17.55 s, 1.69 s, 46.91 s, and 10.99 s respectively. Due to fusion of features obtained through singular value decomposition over GLCM and LBP, the proposed algorithm takes high computational time as compared to other cited methods, but facilitates high detection accuracy with low rate of occurrence of false matches.

## 4 Conclusion

Images have their applications in various fields like criminal investigation, courts of law, medical imaging, document analysis, etc. On account of speedy growth of technology, lot of potent computer applications are available which have made forging process easier. Forgery over images are practiced for denigration, blackmailing, harassment, political disputation, fun-making, etc. Digital images are not adequate in courts of law for witness, without forensic investigation over evidences. Such concerns related to multimedia security and forensic probes resulted in evolution of various advancements in image tampering detection techniques. As a little contribution in field of image forgery detection, our method obtained substantially high DA and low FPR when forged image suffer from blurring, color reduction, contrast adjustment, and brightness change attack. As future work, we will search for methods invariant to geometrical attacks with greater robustness than state-of-the-art methods of copy-move image forgery detection.

## References

1. Zandi, M., Mahmoudi-Aznaveh, A., Talebpour, A.: Iterative copy-move forgery detection based on a new interest point detector. *IEEE Trans. Inf. Forensics Secur.* **11**(11), 2499–2512 (2016). <https://doi.org/10.1109/TIFS.2016.2585118>
2. Chen, C., Ni, J., Shen, Z., Shi, Y.Q.: Blind forensics of successive geometric transformations in digital images using spectral method: theory and applications. *IEEE Trans. Image Process.* **26**(6), 2811–2824 (2017). <https://doi.org/10.1109/TIP.2017.2682963>
3. Fridrich, J., Soukal, D., Lukas, J.: Detection of copy-move forgery in digital images. In: Digital Forensic Research Workshop, pp. 55–61. IEEE Computer Society (2003)
4. Alkawaz, M.H., Sulong, G., Saba, T., Rehman, A.: Detection of copy-move image forgery based on discrete cosine transform. *Neural Comput. Appl.* 1–10 (2016). <https://doi.org/10.1007/s00521-016-2663-3>
5. Zandi, M., Mahmoudi-Aznaveh, A., Mansouri, A.: Adaptive matching for copy-move forgery detection. IEEE International Workshop on Information Forensics and Security (WIFS), pp. 119–124 (2014). <https://doi.org/10.13140/RG.2.1.2189.5200>
6. Lee, J.C., Chang, C.P., Chen, W.K.: Detection of copy-move image forgery using histogram of orientated gradients. *Inf. Sci.* **321**(C), 250–262 (2015). <https://doi.org/10.1016/j.ins.2015.03.009>
7. Silva, E., Carvalho, T., Ferreira, A., Rocha, A.: Going deeper into copy-move forgery detection: exploring image telltales via multi-scale analysis and voting processes. *J. Vis. Commun. Image Represent.* **29**, 16–32 (2015). <https://doi.org/10.1016/j.jvcir.2015.01.016>
8. Shen, X., Shi, Z., Chen, H.: Splicing image forgery detection using textural features based on grey level co-occurrence matrices. *IET Image Process.* **11**(1), 44–53 (2017). <https://doi.org/10.1049/iet-ipr.2016.0238>
9. Tai, Y., Yang, J., Luo, L., Zhang, F., Qian, J.: Learning discriminative singular value decomposition representation for face recognition. *Pattern Recognit.* **50**, 1–16 (2016). <https://doi.org/10.1016/j.patcog.2015.08.010>
10. Zhang, T., Wang, R.: Copy-move forgery detection based on SVD in digital images. In: International Congress on Image and Signal Processing, pp. 1–5 (2009). <https://doi.org/10.1109/CISP.2009.5301325>
11. Li, L., Li, S., Zhu, H., Chu, S.C., Roddick, J.F., Pan, J.S.: An efficient scheme for detecting copy-move forged images by local binary patterns. *J. Inf. Hiding Multimed. Signal Process.* **4**(1), 46–56 (2013)
12. Li, Z., Liu, G.Z., Yang, Y., You, Z.Y.: Scale and rotation-invariant local binary pattern using scale-adaptive texton subuniform-based circular shift and sub uniform-based circular shift. *IEEE Trans. Image Process.* **21**(4), 2130–2140 (2012). <https://doi.org/10.1109/TIP.2011.2173697>
13. Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: CoMoFoD—new database for copy-move forgery detection. In: International Symposium Electronics in Marine, pp. 49–54 (2013)
14. Khan, S., Kulkarni, A.: An efficient method for detection of copy-move forgery using discrete wavelet transform. *Int. J. Comput. Sci. Eng.* **2**(5), 1801–1806 (2010)

# Object Labeling in 3D from Multi-view Scenes Using Gaussian–Hermite Moment-Based Depth Map



Sadman Sakib Enan, S. M. Mahbubur Rahman, Samiul Haque,  
Tamanna Howlader and Dimitrios Hatzinakos

**Abstract** Depth as well as intensity of a pixel plays a significant role in labeling objects in 3D environments. This paper presents a novel approach of labeling objects from multi-view video sequences by incorporating rich depth information. The depth map of a scene is estimated from focus-cues using the Gaussian–Hermite moments (GHMs) of local neighboring pixels. It is expected that the depth map obtained from GHMs provides robust features as compared to that provided by other popular depth maps such as those obtained from Kinect and defocus cue. We use the rich depth and intensity values of a pixel to score every point of a video frame for generating labeled probability maps in a 3D environment. These maps are then used to create a 3D scene wherein available objects are labeled distinctively. Experimental results reveal that our proposed approach yields excellent performance of object labeling for different multi-view scenes taken from RGB-D object dataset, in particular showing significant improvements in precision–recall characteristics and *F1*-score.

**Keywords** Depth map · Object labeling · 3D scene

---

S. S. Enan (✉) · S. M. Mahbubur Rahman (✉)  
Department of EEE, BUET, Dhaka 1205, Bangladesh  
e-mail: [sadmansakib.enan@gmail.com](mailto:sadmansakib.enan@gmail.com)

S. M. Mahbubur Rahman  
e-mail: [mahbubur@eee.buet.ac.bd](mailto:mahbubur@eee.buet.ac.bd)

S. Haque  
Department of ECE, North Carolina State University, Raleigh, NC 27606, USA  
e-mail: [shaque2@ncsu.edu](mailto:shaque2@ncsu.edu)

T. Howlader  
ISRT, University of Dhaka, Dhaka 1000, Bangladesh  
e-mail: [tamanna@isrt.ac.bd](mailto:tamanna@isrt.ac.bd)

D. Hatzinakos  
Department of ECE, University of Toronto, Toronto, ON M5S 2E4, Canada  
e-mail: [dimitris@comm.utoronto.ca](mailto:dimitris@comm.utoronto.ca)

## 1 Introduction

Object recognition is one of the fundamental problems in computer vision and machine learning. Over the years, an ample amount of research have been performed to detect objects from 2D images. Various algorithms have been proposed for this problem of 2D object recognition considering only the image intensities [4]. In recent times, ubiquitous use of 3D display devices and access to 3D data pave the way for object labeling in 3D environments. For example, 3D scenes captured with laser scanners enable high-impact applications such as geographical mapping [7]. Different machine learning approaches have been proposed for labeling 3D objects from laser data (see, for example, [18, 24]). But, the laser-based 3D object detection techniques suffer from poor resolution.

On the other hand, object labeling in 3D point clouds has shown to be performed well for complex scenes, wherein histogram of oriented gradient (HOG) features are calculated for segments of point clouds [15, 25]. At the expense of huge size training dataset, the deep convolutional neural networks have been employed on 3D point cloud data for object classification [8]. In general, object labeling in 3D scenes shows more challenge than its 2D counterpart due to a broad range of variations with regard to object shapes, background, lighting, and viewing condition. For an increasing accuracy of detection of objects of various shapes and sizes in a 3D scene, reliable depth information is crucial along with the intensity information of individual objects in the scene [6, 13]. Depth descriptors estimated from local patches of an image show promising results while detecting an object in 3D scene [1]. Combining intensity and depth information works well when features in terms of two sets of HOGs are learnt to detect objects in a 3D scene [5, 11, 20]. Thus, recent advancements in RGB-D mapping, make it feasible to reconstruct a labeled 3D scene with a high degree of accuracy [10].

Since depth maps give significant information to a scene, there remains an opportunity to improve this 3D object detection problem using a better estimate of the depth of a scene instead of using depth from traditional methods. In tradition, the depth estimation from multi-view settings is a complex process. Existing methods of depth estimation use traditional visual cues such as the motion, geometry, and defocus of an object in a scene [23, 26, 27]. But, these methods are mostly limited by the shape of the objects and their relative distances. In many cases, the objects remain in the blurred background, which makes it difficult to estimate the depth map. Moreover, depth map from Kinect needs a recursive median filtering to fill the missing values of depth data [12].

In a recent work, it has been shown that the focus-cue expressed in terms of the Gaussian–Hermite moments (GHMs) of local neighboring pixels of a scene can provide a robust estimate of depth map as compared to the traditional depth estimation techniques [9]. Motivated by this work, we concentrate on developing an object labeling technique that combines the HOG features of RGB image and depth map calculated from GHMs. These features are used to train a support vector machine (SVM)-based classifier for detecting objects in a scene. In particular, the classified

scores of pixel can be used to generate a probability map for a video frame. Finally, these probability maps of all the frames can be integrated using the RGB-D mapping algorithm [10] to create an overall multi-view labeled 3D scene for visualization. Further refinements of this 3D scene can be achieved by using a graph cut technique [2], which minimizes the energy of a multi-class pairwise Markov random field. Therefore, the introduction of a better estimate of depth map using the GHMs is expected to ameliorate the overall quality of object labeling in a 3D scene. In this context, the main objective of this paper is to present an object labeling technique in 3D from multi-view scenes using GHM-based depth map. The overall contributions of the paper are

- Introducing focus-cue based depth map of a scene estimated from GHMs for object labeling in 3D environment
- Adopting two sets of HOG features—one for GHM-based depth map and another for RGB frame to represent an object
- Applying SVM-based classifiers to create labeled probability maps as per the categories of different types of objects in a scene.

The rest of the paper is organized as follows. Section 2 provides a description of the proposed method. The experimental setup and the results obtained are described in Sect. 3. Finally, Sect. 4 provides the conclusion.

## 2 Proposed Method

Let  $I(i, j, k)$  ( $(i, j, k) \in \mathbb{Z}^3$ ) be a pixel intensity at the location  $(i, j)$  in  $k$ -th frame of an RGB video. We analyze the local neighboring region of this pixel, and then estimate a corresponding depth value  $D(i, j, k)$  using GHMs. The HOG features are extracted from the neighboring regions of  $I(i, j, k)$  and  $D(i, j, k)$  to classify the objects in the  $k$ -th frame. Finally, the probability maps are generated from all these classified intensities to create labeled objects in 3D scene.

### 2.1 Estimation of Depth Map

To get a complete depth map of an RGB image, focus-cues of all the pixels are estimated first by using the GHMs of local neighboring pixels. In this stage, a sparse focus map is generated for each of the frame [9]. Then, a Bayesian matting is applied to refine this sparse focus map [3]. Let  $I^W(i, j, k)$  be selected by an  $N \times N$  square-shaped window  $W$ . Focus-cue of a pixel is estimated from lower- and higher order moments that capture significant variations and minute details around the pixel of interest. These moments can be considered as low- and high-pass filtered components, respectively. The energy of reconstructed region obtained from these moments gives the focus-cue of a particular pixel and is defined as

$$F_{\gamma}^W(i, j, k) = \frac{\|\mathbb{H}(I^W(i, j, k), \gamma)\|}{\|\mathbb{L}(I^W(i, j, k), \gamma)\|} \quad (1)$$

where  $\mathbb{L}(I^W(i, j, k), \gamma)$  and  $\mathbb{H}(I^W(i, j, k), \gamma)$  are the low- and high-pass components of the local neighboring region  $W$  reconstructed using lower- and higher order GHMs,  $\|\cdot\|$  represents the  $\mathcal{L}_2$  norm and  $\gamma(\gamma \ll N^2)$  determines the boundary between the lower and higher order moments. The energy of the low-pass components of the neighboring region of a particular pixel is calculated using GHMs as [9]

$$\|\mathbb{L}(I^W(i, j, k), \gamma)\| = \left\| \sum_{p=0}^{\gamma} \sum_{q=0}^{\gamma} \eta_{p,q}^W \hat{H}_p(i; \sigma) \hat{H}_q(j; \sigma) \right\| \quad (2)$$

where  $\eta_{p,q}^W$  are the GHMs of order  $(p, q)$  ( $p, q \in 0, 1, 2, \dots, N - 1$ ) and  $\hat{H}_p(i; \sigma)$ ,  $\hat{H}_q(j; \sigma)$  are generalized  $p$ th and  $q$ th order Gaussian–Hermite (GH) polynomials with spread parameter  $\sigma$  ( $\sigma > 0$ ) on the real line  $x$  and  $y$  ( $x, y \in \mathbb{R}^1$ ), respectively, that maintain the orthogonality condition [19]. The generalized GH polynomial of  $p$ th order can be expressed as

$$\hat{H}_p(x; \sigma) = (2^p p! \sqrt{\pi} \sigma)^{-1/2} \exp(-x^2/2\sigma^2) H_p(x/\sigma) \quad (3)$$

where

$$H_p(x) = (-1)^p \exp(x^2) \frac{d^p}{dx^p} \exp(-x^2) \quad (4)$$

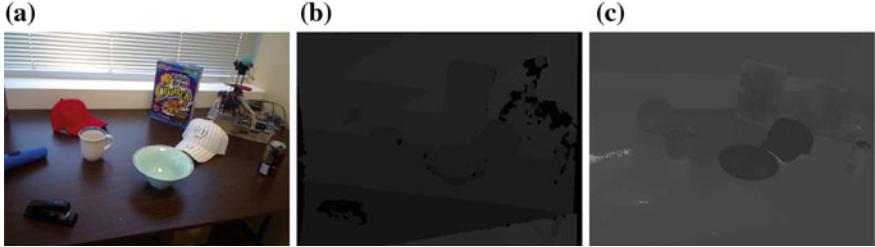
The generalized GH polynomials form basis functions and the GHMs of an image signal can be considered as the projections of the signal onto the functions. Therefore, these moments and GH polynomials are used to characterize a local neighboring region of an image. The GHMs can be calculated from pixels of local neighboring region  $W$  and expressed as [22]

$$\eta_{p,q}^W(i, j, k) = \sum_{m,n \in W} I^W(m, n, k) \hat{H}_p(m; \sigma) \hat{H}_q(n; \sigma) \quad (5)$$

Finally, the energy of the high-pass components of the local neighboring region of the concerned pixel is calculated using GHMs as

$$\|\mathbb{H}(I^W(i, j, k), \gamma)\| = \|I^W(i, j, k)\| - \|\mathbb{L}(I^W(i, j, k), \gamma)\| \quad (6)$$

where  $\|I^W(i, j, k)\|$  is the total energy of the local neighboring region  $W$  which is characterized by a finite set of GHMs. Focus-cues estimated using GHMs are found to be sparse because of the edge and textures of the objects in the local neighboring region  $W$ . Thus, a matting process is adopted to get a smoother depth map for a frame. Therefore, the complete depth map of the  $k$ th frame is expressed as [27]



**Fig. 1** Visual outputs of depth map of a typical video frame. **a** Original intensity image. Depth maps are obtained from **b** Kinect camera [12] and **c** GHMs of local neighboring pixels [9]

$$\mathbf{D} = \lambda \mathbf{dF}(\mathbf{L} + \lambda \mathbf{d})^{-1} \quad (7)$$

where  $\mathbf{F}$  is the estimated sparse focus vector,  $\mathbf{L}$  is the Laplacian matting matrix,  $\mathbf{d}$  is the diagonal matrix and  $\lambda (\lambda < 1)$  balances between the sparsity of focus map and smoothness of the interpolation [16]. Figure 1 shows an example of visual outputs of depth map of a frame that has different objects such as bowl, cap, cereal box, coffee mug, and soda can. It is seen from this figure that depth obtained from Kinect shows poorly defined edges of the objects, whereas depth using GHMs demonstrates well-defined boundaries of objects.

## 2.2 Feature Extraction

Let  $c (c \in 1, 2, 3, \dots, C)$  be an object class among  $C$  number of classes. We consider the object class is background, when  $c = 0$ . Let  $x$  be a pixel that belongs to a certain class  $c$  ( $0 \leq c \leq C$ ) in the  $k$ -th frame. Let  $f_c^h(x_I)$  be the intensity feature vector at pixel location  $(i, j, k)$  for an object class  $c$  ( $1 \leq c \leq C$ ) or background class  $c_B$  ( $c = 0$ ) at a scale  $h$ . Sliding window detectors have been used in multiple scaled versions of the intensity image because an object usually appears in different sizes based on its distance from a camera. A feature vector  $f_c^h(x_I)$  is estimated for each pixel  $x$  by extracting HOG features from the intensity of local neighboring region. In particular, the gradients of orientations in each square-shaped neighborhood are encoded into a histogram-based feature vector [14]. In the same way, the depth feature vector  $f_c^h(x_D)$  for the same pixel location  $(i, j, k)$  is estimated from the GHM-based depth map. The complete feature vector  $f_c^h(x)$  is found by concatenating the intensity and depth feature vectors, namely,  $f_c^h(x_I)$  and  $f_c^h(x_D)$ . Finally, this feature vector  $f_c^h(x)$  is fed to an object detector to classify the pixel into one of the different types of objects or background in the scene.

### 2.3 Object Labeling and Probability Maps

Since each scene can have multiple objects, the proposed method employs  $C$  number of linear SVM object detectors to get the probability maps corresponding to different types of objects. Each detector is a binary classifier between an object class  $c$  ( $1 \leq c \leq C$ ) and background class  $c_B$  ( $c = 0$ ). The advantage of such training of  $C$  number of binary classifiers instead of adopting a multi-class classifier is that each detector can be trained independently by considering different window sizes and aspect ratios as per the size of objects. A particular detector gives a specific score to the point  $x$  according to the class of object. Thus, the score map of an object to be in class  $c$  ( $1 \leq c \leq C$ ) in the  $k$ th frame is given by [14]

$$s_c(x) = \max_h \{\mathbf{w}_c^\top \mathbf{f}_c^h(x) + b_c\} \quad (8)$$

where  $\mathbf{w}_c, b_c$  are the weight vector and the bias of the detector  $c$  ( $1 \leq c \leq C$ ) that are learned from the training data. The learned weight vectors and bias terms are then employed to all the frames of multi-view video sequence to obtain the score maps.

In order to visualize different types of objects in 3D environment, the probabilities of all the points to be in a class  $c$  ( $1 \leq c \leq C$ ) are calculated. Let  $p(c|x)$  be the probability of a point  $x$  belonging to the object class  $c$ . This probability can be calculated from the entire set of scores  $s_c(x)$  estimated by  $C$  number of object detectors, a binary classifier that recognizes object from the background [21]. A complete probability map of the  $k$ th frame can be found by using Platt scaling as [17]

$$p(c|x) = \frac{1}{1 + \exp\{us_c(x) + v\}} \quad (9)$$

where  $u$  and  $v$  are the parameters of the sigmoid function. In order to obtain the probability map of the background, the points that do not belong to any of the  $C$  classes of objects are found. Thus, the probability map of the background can be defined as

$$p(c_B|x) = \alpha \min_{1 \leq c \leq C} \{1 - p(c|x)\} \quad (10)$$

where  $\alpha$  ( $0 < \alpha \leq 1$ ) controls the inconstancy of probability maps estimated for the objects. Finally, the probability maps of the objects and background are integrated for all the frames of a given multi-view video sequence. In particular, the alignments of probability maps through RGB-D mapping algorithm [10] and the refinements of maps through graph-based energy minimization [2] result in labeled 3D scene for stereotype visualization.

### 3 Experiments and Results

Experiments are carried out to evaluate the performance of the proposed GHM-based object labeling technique as compared to that of the existing methods. In this section, first we give an overview of the dataset used in the experiments, then we describe our training and testing partitions, experimental setup, and parameter settings of the proposed system. The comparing methods are introduced next for performance evaluation. Finally, the results are presented and evaluated in terms of common performance metrics.

#### 3.1 Dataset

To carry out the experiments, RGB-D object database [12] is chosen. The database subsumes a huge number of commonly seen 3D objects with different shapes and sizes such as bowl, cap, flashlight, battery, rubber, coffee mug, etc. The objects are captured in a multi-view environment by Kinect camera from PrimeSense. The dataset contains  $2.5 \times 10^5$  segmented RGB-D images of 300 objects in 51 categories. It also includes RGB-D video dataset, which consists of 8 multi-view image sequences with different number of objects, and 11 background videos without any objects. The scenes of images and videos are captured from office and kitchen environments. For the purpose of labeling multi-view scenes in 3D, 25 objects in 5 categories such as bowl (4 types), cap (4 types), cereal box (5 types), coffee mug (6 types), soda can (6 types) and 11 background videos are trained to develop the linear SVM object detectors. For testing, four multi-view scenes, namely, the meeting room, the kitchen room and two types of office room are chosen. These testing videos include a wide variety in terms of number, size, shape, and occlusion of objects in the scene.

#### 3.2 Setup

For estimating focus-cue of a pixel from the local neighboring GHMs, threshold value  $\gamma$  is chosen to be 2 to rely mostly on the lower order moments. The regularization parameter  $\lambda$  is chosen to be 0.002 for balancing between the sparsity of focus map and the smoothness of the interpolation as prescribed in [9]. The HOG features of the intensity images and depth maps are estimated from gradient orientations of sliding window of size  $8 \times 8$ , and the histograms are encoded into a 108-dimensional vector [14]. The parameter  $\alpha$  which controls the upper bound of the inconstancy is chosen as 0.1. For visualization, five categories of objects, namely, the bowl, cap, cereal box, coffee mug, and soda can are colored as red, green, blue, yellow and cyan,

respectively. The background is colored as gray. The MeshLab v1.3.3 is used to visualize the labeled objects in the scenes.

### 3.3 Comparing Methods

The proposed 3D object labeling technique uses the features obtained from GHM-based depth map, and hence, we denote the proposed method as D(GHM). The labeling performance of the proposed method is compared with two methods that employ commonly referred depth maps. The methods compared are briefly described as follows:

- Kinect-based labeling D(Kinect) [12]: In this method, the depth map is generated from Kinect camera and the depth feature vector is estimated by extracting HOG features from the depth map.
- Defocus-based labeling D(Defocus) [27]: This depth estimation technique assumes that blurred objects of a scene usually remain the farthest from the camera. The feature vector of this method is extracted in the same way as the other comparing methods.

### 3.4 Evaluation Metrics

The performance of the three comparing methods is evaluated in terms of precision-recall characteristics curve, average precision, and  $F1$ -score. Precision is the fraction of correctly predicted labels of pixels with respect to the ground truths, and recall implies the fraction of correctly predicted labels with respect to the total number of correct labels. A precision-recall curve shows the trade-off between precision and recall for different thresholds. A system with low precision and high recall returns many results but most of them are incorrectly predicted. On the other hand, a system with high precision and low recall returns fewer results but most of them are correctly predicted. Average precision reveals the accuracy of labeling by considering the availability of ground truth.

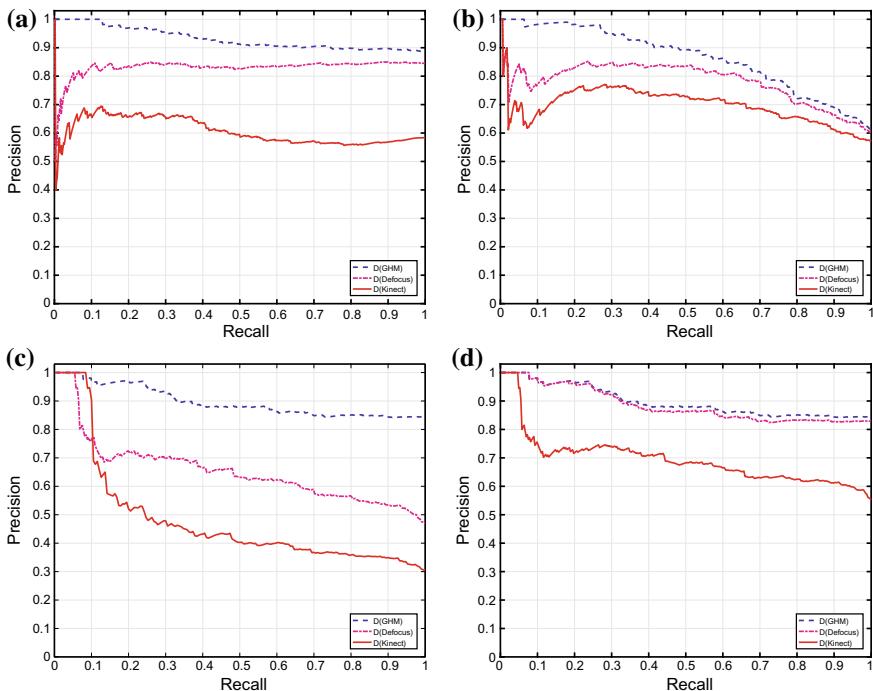
$F1$ -score gives an idea about the close matching between the labeled scene in 3D prepared by the GHM-based depth map and that by the ground truth. In order to measure the  $F1$ -score, a common threshold 0.45 is selected for each of the comparing methods. The metric  $F1$ -score is given as

$$F1\text{-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

### 3.5 Results

Figure 2 shows precision–recall characteristics curves for labeling of four multi-view testing video sequences, namely, meeting room, kitchen room, and office rooms 1 and 2. It is seen from this figure that the proposed GHM-based method maintains the high level of precision as compared to the methods using depth from Kinect and defocus cue independent of the recall rate. In particular, the meeting room and office room 2 maintains higher level of precision while comparing with the ground truth labeling mainly because of superior lighting conditions of these two scenes.

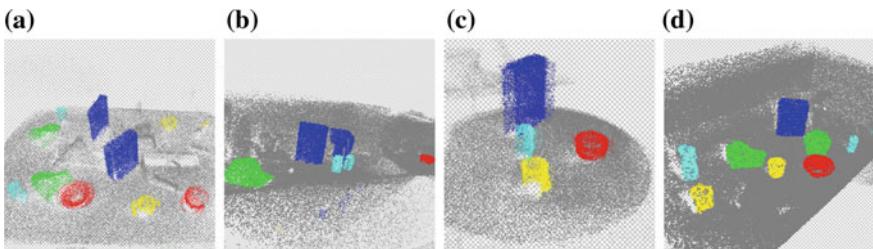
Table 1 shows the overall performance in terms of average precision and  $F1$ -score for predicting the objects appeared in the testing videos. It is seen from this table that the proposed GHM-based method performs the best by providing the highest average precision and  $F1$ -score for all the scenes. The proposed GHM-based method shows an improvement of average precision as high as 29.5% and 26.9% as compared to the D(Defocus)- and D(Kinect)-based methods, respectively. It can also be found



**Fig. 2** Precision–recall characteristics curves for labeling of four multi-view video sequences taken from RGB-D object dataset [12]. The video sequences are **a** meeting room, **b** kitchen room, **c** office room 1, and **d** office room 2. The blue lines represent the curves corresponding to the D(GHM) technique, the performance of which is superior in terms of precision as compared to that of the D(Defocus) and D(Kinect) techniques represented in magenta and red lines, respectively

**Table 1** Performance comparisons of predicting labels of 3D objects in multi-view scenes

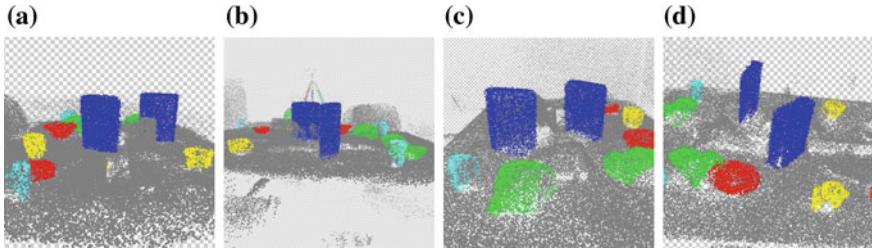
Scene	Technique					
	D(Kinect)		D(Defocus)		D(GHM)	
	Avg. Prec. (%)	F1-score (%)	Avg. Prec. (%)	F1-score (%)	Avg. Prec. (%)	F1-score (%)
Meeting room	62.40	58.76	85.03	67.90	<b>93.29</b>	<b>72.04</b>
Kitchen room	71.54	66.48	79.42	71.01	<b>80.54</b>	<b>74.46</b>
Office room 1	55.49	52.12	65.66	59.67	<b>85.04</b>	<b>71.53</b>
Office room 2	69.88	66.19	88.70	70.42	<b>89.93</b>	<b>70.92</b>



**Fig. 3** Visual outputs of labeled objects in 3D using the proposed GHM-based depth maps for multi-view scenes. The test video sequences are **a** meeting room, **b** kitchen room, **c** office room 1, and **d** office room 2. The 3D objects are colored by their categories, namely, bowl is red, cap is green, cereal box is blue, coffee mug is yellow and soda can is cyan

from Table 1 that the proposed method provides the highest *F1*-score as compared to others by reaching as high as 74.46% for the scene of kitchen room. These results reveal that the proposed method provides the lowest error while predicting the labels of objects.

Figure 3 shows the visual outputs of labeled objects in 3D using the proposed method for four testing scenes, namely, meeting room, kitchen room, office rooms 1 and 2. It is seen from this figure that the scenes widely vary according to the number of objects, their sizes, shapes, and viewing positions. In all these scenarios, the proposed GHM-based method is capable of labeling most of the pixels of the 3D objects correctly so that the objects can be recognized distinctively. Figure 4 shows visual outputs of the labeled 3D objects of meeting room at four different viewing angles, namely, 0°, 90°, 180° and 270°. This figure reveals that the occlusion of viewing angle has little impact on the labeling performance of the proposed GHM-based method. It can be seen from Fig. 4 that the objects such as the two yellow-colored coffee mugs appear in 0° and 270° viewing angles, although one of the mugs is occluded in the other viewing positions.



**Fig. 4** Visual outputs of labeled objects in multiple angles of 3D using the proposed GHM-based depth maps for the scene of meeting room. The viewing angles are **a** 0°, **b** 90°, **c** 180°, and **d** 270°. The proposed method labels 11 objects in 5 categories, namely, bowl, cap, cereal box, coffee mug, and soda can

## 4 Conclusion

With the current trend toward automation, robotic industry is growing profusely and encountering complex robotic tasks every day. To have a reliable idea about the surroundings, robust performance of object labeling is in demand for machine vision. Because of the limitations of 2D labeling, a technique of labeling objects in 3D from multi-view video sequence has been proposed in this paper. The proposed method has employed HOG-based features from the GHM-based depth maps and the intensity values of the scene. The score map of each type of objects has been predicted by a linear SVM-based classifier using the estimated features. The labeling of 3D objects has been performed by estimating the probability maps from class-dependent predicted scores. Experimental results have revealed that the proposed GHM-based method of 3D object labeling performs better than the Kinect and defocus cue-based labeling both in terms of precision–recall characteristics and *F1*-score. It has been reported that the GHM-based depth map can contribute significantly to achieve average precision and *F1*-score as high as 93.29% and 74.46%, respectively, for labeling objects in 3D from multi-view scenes. In the future, this work can be extended to label nonstationary objects in dynamic scenes.

## References

1. Bo, L., Ren, X., Fox, D.: Depth kernel descriptors for object recognition. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems, pp. 821–826. San Francisco, CA, USA (2011)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(11), 1222–1239 (2001)
3. Chuang, Y.Y., Curless, B., Salesin, D.H., Szeliski, R.: A Bayesian Approach to Digital Matting, vol. 2, pp. 264–271. Kauai, Hawaii (2001)
4. Collet, A., Berenson, D., Srinivasa, S.S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: Proceedings of IEEE International

- Conference on Robotics and Automation, pp. 48–55. Kobe, Japan (2009)
- 5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 886–893. Washington, DC, USA (2005)
  - 6. Das, S., Koperski, M., Bremond, F., Francesca, G.: Action recognition based on a mixture of RGB and depth based skeleton. In: Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 1–6. Lecce, Italy (2017)
  - 7. Douillard, B., Fox, D., Ramos, F., Durrant-Whyte, H.: Classification and semantic mapping of urban environments. *Int. J. Robot. Res.* **30**(1), 5–32 (2011)
  - 8. Engelcke, M., Rao, D., Zeng Wang, D., Hay Tong, C., Posner, I.: Vote3Deep: fast object detection in 3D point clouds using efficient convolutional neural networks. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1355–1361. Singapore (2017)
  - 9. Haque, S., Rahman, S.M.M., Hatzinakos, D.: Gaussian-Hermite moment-based depth estimation from single still image for stereo vision. *J. Vis. Commun. Image Represent.* **41**, 281–295 (2016)
  - 10. Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D.: RGB-D mapping: using depth cameras for dense 3D modeling of indoor environments. In: Khatib, O., Kumar, V., Sukhatme, G. (eds.) *Experimental Robotics: Springer Tracts in Advanced Robotics*, vol. 79, pp. 477–491. Springer (2014)
  - 11. Lai, K., Bo, L., Fox, D.: Unsupervised feature learning for 3D scene labeling. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 3050–3057. Hong Kong, China (2014)
  - 12. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1817–1824 (2011)
  - 13. Lai, K., Bo, L., Ren, X., Fox, D.: Sparse distance learning for object recognition combining RGB and depth information. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 4007–4013. Shanghai, China (2011)
  - 14. Lai, K., Bo, L., Ren, X., Fox, D.: Detection-based object labeling in 3D scenes. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 1330–1337. Saint Paul, MN, USA (2012)
  - 15. Lai, K., Fox, D.: Object recognition in 3D point clouds using web data and domain adaptation. *Int. J. Robot. Res.* **29**(8), 1019–1037 (2010)
  - 16. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 61–68. Washington, DC, USA (2006)
  - 17. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P.J. (eds.) *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press (1999)
  - 18. Quigley, M., Batra, S., Gould, S., Klingbeil, E., Le, Q., Wellman, A., Ng, A.Y.: High-accuracy 3D sensing for mobile manipulation: improving object detection and door opening. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 2816–2822. Kobe, Japan (2009)
  - 19. Rahman, S.M.M., Lata, S.P., Howlader, T.: Bayesian face recognition using 2D Gaussian-Hermite moments. *EURASIP J. Image Video Process.* **2015**(35), 1–20 (2015)
  - 20. Ren, X., Ramanan, D.: Histograms of sparse codes for object detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3246–3253. Portland, OR, USA (2013)
  - 21. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multi-class object detection. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1481–1488. Los Alamitos, CA, USA (2011)
  - 22. Shen, J., Shen, W., Shen, D.: On geometric and orthogonal moments. *Int. J. Pattern Recognit. Artif. Intell.* **14**(07), 875–894 (2000)

23. Su, H., Huang, Q., Mitra, N.J., Li, Y., Guibas, L.: Estimating image depth using shape collections. *ACM Trans. Graph.* **33**(4), 37:1–37:11 (2014)
24. Triebel, R., Schmidt, R., Mozos, O.M., Burgard, W.: Instance-based AMN classification for improved object recognition in 2D and 3D laser range data. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2225–2230. Hyderabad, India (2007)
25. Xiong, X., Munoz, D., Bagnell, J.A., Hebert, M.: 3-D scene analysis via sequenced predictions over points and regions. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 2609–2616. Shanghai, China (2011)
26. Xu, Y., Hu, X., Peng, S.: Sharp image estimation from a depth-involved motion-blurred image. *Neurocomputing* **171**(C), 1185–1192 (2016)
27. Zhuo, S., Sim, T.: Defocus map estimation from a single image. *Pattern Recognit.* **44**(9), 1852–1858 (2011)

# Zernike Moment and Mutual Information Based Methods for Multimodal Image Registration



**Suraj Kumar Kashyap, Dinesh Jat, M. K. Bhuyan, Amit Vishwakarma and Prathik Gadde**

**Abstract** Image registration enables joint operations between images obtained from diverse sources. However, there have been limited advances in the registration of multichannel images. The accuracy of registration is a significant concern for medical applications, among others. Two methods, PCA–ZM and CED–ZM, have been proposed for registration based on Zernike moment and enhanced mutual information. Edge detection by Zernike moment and identification of common features in multichannel images are used as a foundation to improve accuracy over single-channel registrations. Single-channel registration accuracy for MRI and SPECT brain images is found to surpass the methods compared against. PCA–ZM demonstrates good accuracy for MR-MR registration, while CED–ZM has good accuracy for MR-SPECT registration. These measures improve upon accurate registration for images, especially where many modalities are available, such as in medical diagnosis.

**Keywords** Multimodal · Multichannel · MRI · SPECT · Medical image registration · Zernike moment · Enhanced mutual information · Principal component analysis · Canny edge detector · Gradient information · Edge overlap

---

S. K. Kashyap · D. Jat · M. K. Bhuyan · A. Vishwakarma (✉)  
Department of Electronics and Electrical Engineering, Indian Institute  
of Technology Guwahati, Guwahati 781039, Assam, India  
e-mail: [amitvishwakarma2625@gmail.com](mailto:amitvishwakarma2625@gmail.com)

S. K. Kashyap  
e-mail: [suraj.pathak@iitg.ac.in](mailto:suraj.pathak@iitg.ac.in)

D. Jat  
e-mail: [j.dinesh@iitg.ac.in](mailto:j.dinesh@iitg.ac.in)

M. K. Bhuyan  
e-mail: [mkb@iitg.ac.in](mailto:mkb@iitg.ac.in)

P. Gadde  
School of Informatics and Computing, Indiana University, Purdue University,  
535 West Michigan St., Indianapolis, IN 46202, USA  
e-mail: [pgadde@iupui.edu](mailto:pgadde@iupui.edu)

## 1 Introduction

Image registration is the process of transforming the floating or moving image coordinate system to the reference image coordinate system. This enables finding correspondences between images. This process is applied to images of a subject that have been captured by different cameras, conditions, standard specifications, positions, and times. These are advantageous for analysis and information fusion in medical imaging, remote sensing, astrophotography, natural image processing for computer vision tasks, etc.

Medical images of a patient's anatomy are obtained under various standard specifications to obtain specific information. Magnetic Resonance Imaging (MRI) characterizes hydrogen-rich molecules, such as water and fat in the human brain. Under T1-weighted parameters for MRI, fat gives a high signal and water gives a low signal. The opposite is true under T2-weighted MRI. In PD weighted MRI, fat has a somewhat higher signal intensity than water. Infusing gadolinium shortens the response of water and gives a high signal in T1-weighted MRI. Computed Tomography (CT) images use X-ray radiation and help in identifying regions of dense matter, such as bone or fiber. Single Photon Emission Computed Tomography (SPECT) incorporates tracer material, which flows with blood while emitting gamma rays, while Positron Emission Tomography (PET) tracer is absorbed by tissues instead.

While registration between same and similar modalities has been tackled with great success, less work has been done toward registration between images of very different modalities, such as those mentioned above. Furthermore, with the availability of 3-D imaging data and 2-D image slices thereof, multichannel registration techniques are suitable for such inputs. After describing examples of previous methods for registration as below, we propose two methods which perform registration by extracting common features of multichannel image sets. This is followed by experiments, comparative results, and our conclusion drawn from the same. The detailed information regarding image registration and applications can be found in [1–3, 5–7, 9, 10, 12].

## 2 Previous Methods

Among the many image registration procedures in existence, the broad classes of intensity-based, gradient- or edge-based, texture-based, and more complex patch-based methods can be identified. Maes et al. [7] proposed an intensity-based Mutual Information metric (MI) for registration of single-channel medical images. Mutual information is the difference between the sum of individual entropies of images and their joint entropy.

An extension of this is given by Pluim et al. [9], where the mutual information metric is replaced by Normalized Mutual Information (NMI), said to be less affected by the area of overlap between images. A gradient information metric has been given

as a weighted average of correlation in a direction between gradients of the images. The product of gradient information and NMI is used for registration. Accurate registration becomes possible where the images had excellent gradient information despite a poor MI. This evidently reflects the importance held by edges and gradients with respect to image registration.

A simultaneous registration-fusion process proposed by Li and Verma [5] utilizes a measure of information obtained from multichannel images after fusing them. Images were fused by first decomposing each channel image with Gabor wavelets transforms, with each wavelet yielding a component image to be fused. Independent component analysis is applied for separating the underlying features, such as water, blood, white matter, and muscle, for each multichannel set. A dissimilarity-based symmetric measure of registration is given, which tends to be lower when the same component in each set has maximum and equal values at a point.

Registration on the basis of region descriptors has been explored. One procedure proposed by Zhu et al. [12] uses Zernike polynomials as moment functions over a circular patch. In this manner, the Zernike Moment-Based Local Descriptor (ZMLD) is defined as a 3-vector of complex moments obtained using the Zernike polynomials  $V_{00}$ ,  $V_{11}$ , and  $V_{22}$ . The Euclidean distance between the descriptors of pixels is used as the dissimilarity metric for registration of single-channel images.

Pradhan and Patra [10] proposed a registration measure named enhanced mutual information ( $E_{MI}$ ) to extend a model of quantitative–qualitative measure of mutual information ( $Q_{MI}$ ). The process described the saliency of each pixel of an image using entropy of a neighborhood, which contribute to the weight or utility for calculation of  $E_{MI}$ . These weights are gradually changed to one over multiple iterations. Geometric splines were applied to obtain reliable intensity values, saliency, and  $E_{MI}$ .

Chen et al. [2] proposed a purely gradient-based metric named as Normalized Total Gradient (NTG) for registration of multispectral images, also shown to have potential for multimodal registration of medical images. The metric is the ratio of the pixel-wise sum of gradient magnitude of the difference image to the pixel-wise sum of gradient magnitudes of the individual images. An image pyramid scheme for registration was also given, which is said to improve the discovery and computation time.

Arce-Santana et al. [1] proposed a purely intensity-based registration metric and procedure based on the expectation–maximization methodology. Here, intensities of the floating image are assumed to have a distribution which is conditional on the intensities of corresponding reference pixels. A parametric model of the distribution is assumed, such as Gaussian distribution, whose unknown means and variances are estimated by an iterative process. For the affine transformation model and low misalignment, Newton–Raphson method is also detailed for fast registration.

### 3 Proposed Methods

We propose two measures of registration which are based on the methods given above. Both methods are aimed for registration of multichannel images, and also have good single-channel registration capability. We have utilized some common features in both of our methods. These are the Zernike moment  $A_{11}$  and enhanced mutual information ( $E_{MI}$ ).

#### 3.1 Common Features

The Zernike polynomial  $V_{11}$  is a complex-valued function defined over a circular patch, whose convolution with an image yields the Zernike moment  $A_{11}$  (ZM). For a  $(2S + 1) \times (2S + 1)$  sized patch, ZM for an image  $R_c$  of the multichannel reference image set  $R$  having intensity  $r_c(\mathbf{x}_r)$  at  $\mathbf{x}_r$  is

$$A_{11}[R_c](\mathbf{x}_r) = \sum_{\|\mathbf{x}\| \leq S} \frac{V_{11}^*(\mathbf{x})}{S} \cdot r_c(\mathbf{x}_r - \mathbf{x}) , \quad (1)$$

$$\text{where } V_{11}(\rho, \theta) = \rho \exp(i\theta), \text{ or equivalently, } V_{11}(x, y) = x + iy . \quad (2)$$

$V_{11}$  over a  $5 \times 5$  patch defines a filter similar to the  $3 \times 3$  Sobel mask filter used widely for edge detection. The  $A_{11}$  moment exhibits rotational invariance and incorporates patch size as a parameter. Larger patches capture more of the local variations, resolve gradients better, and have reduced noise sensitivity.

Another complex-valued filter for edge and feature extraction is the Gabor wavelets transform, whose filter mask is comprised of irrational coefficients, contrasted with the  $V_{11}$  mask having simple complex numbers. The band-pass nature of Gabor wavelet masks both enable and require multiscale decomposition techniques for feature extraction, as edges of a given orientation and width respond strongly to a certain mask. This issue is less of a concern with Zernike polynomials which are of a high-pass nature. While any one wavelet can be stored to generate all wavelet masks, the transform must be computed for many such masks. With these motivations, we consider Zernike moments as a compromise between the speed of simpler masks and accuracy of large, irrational masks.

MI-based measures give good registration results when images have similar intensity variations, but this is often not true for multimodal images which represent a different subset of information about the subject, such as different properties of the anatomy. For example, MR images describe the resonance characteristics of tissue, whereas CT images mainly represent tissue density.

We, therefore, use the enhanced mutual information measure to include a characterization of pixels. In  $E_{MI}$ , the utility of an intensity pair is indirectly obtained from saliency of image pixels, where saliency may be determined using local entropy and

local intensity probabilities, as given by [6]. However, unlike [10], we have kept the saliencies constant, instead of gradually varying them over the course of registration.

We use the definition of  $E_{\text{MI}}$  for two images  $R_c$  and  $F_c$ , having intensities represented by  $r_c$  and  $f_c$ , and utility  $w[R_c, F_c](r_c, f_c)$  as

$$E_{\text{MI}}(R_c, F_c) = \sum_{R_c, F_c} \left[ p(r_c, f_c) \log \frac{p(r_c, f_c)}{p(r_c)p(f_c)} + w[R_c, F_c](r_c, f_c)p(r_c, f_c) \right] , \quad (3)$$

$$w[R_c, F_c](r_c, f_c) = \sum \delta[R_c](\mathbf{x}_r) \cdot \sum \delta[F_c](\mathbf{x}_f) , \quad (4)$$

where the first sum is over pixels of intensity  $r_c$  of  $R_c$ , which lie in the region of overlap with  $F_c$ . Using local entropy  $H_D[R_c]$  and self-dissimilarity measure  $U_D[R_c]$  in a radius of  $s_p[R_c]$ , we have the saliency  $\delta[R_c]$  at  $\mathbf{x}_r$  as

$$\delta[R_c](\mathbf{x}_r) = H_D[R_c](s_p, \mathbf{x}_r) \cdot U_D[R_c](s_p, \mathbf{x}_r) , \quad (5)$$

$$s_p[R_c](\mathbf{x}_r) = \arg \max \{H_D[R_c](s, \mathbf{x}_r)\} , \quad (6)$$

$$H_D[R_c](s, \mathbf{x}_r) = - \sum_{R_c} p[R_c](r_c, s, \mathbf{x}_r) \log p[R_c](r_c, s, \mathbf{x}_r) , \quad (7)$$

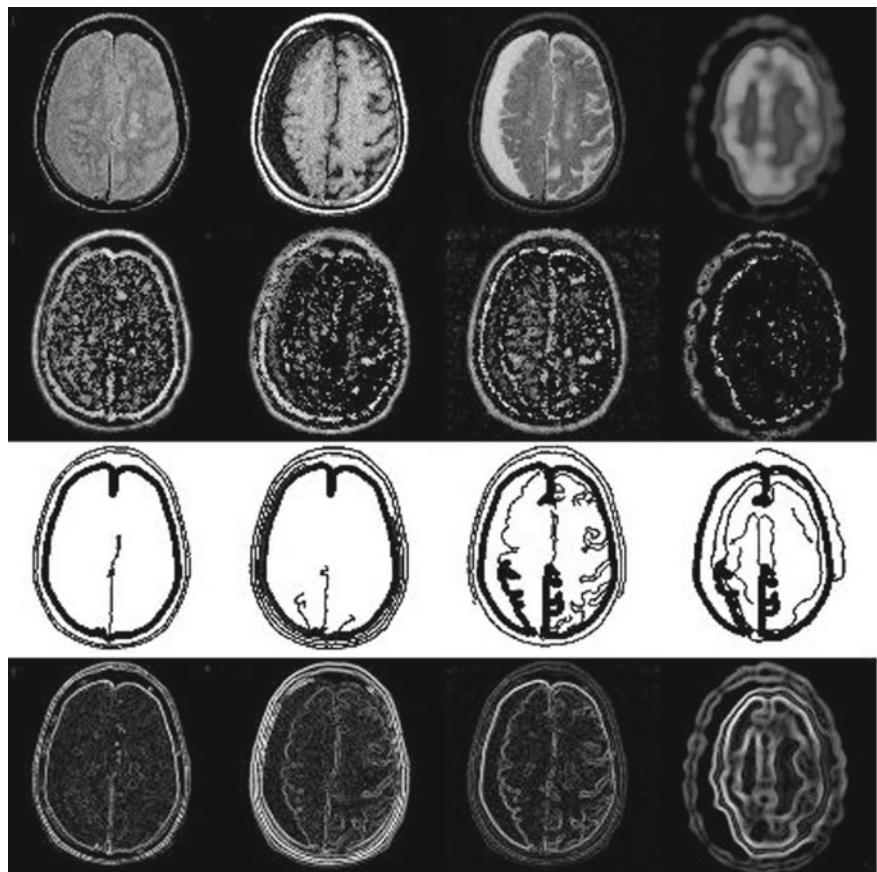
$$U_D[R_c](s, \mathbf{x}_r) = \sum_{R_c} \left| \frac{\partial}{\partial s} p[R_c](r_c, s, \mathbf{x}_r) \right| s , \quad (8)$$

where  $p[R_c](r_c, s, \mathbf{x}_r)$  is probability of intensity  $r_c$  in the disk of radius  $s$  around  $\mathbf{x}_r$  in  $R_c$  (Fig. 1).

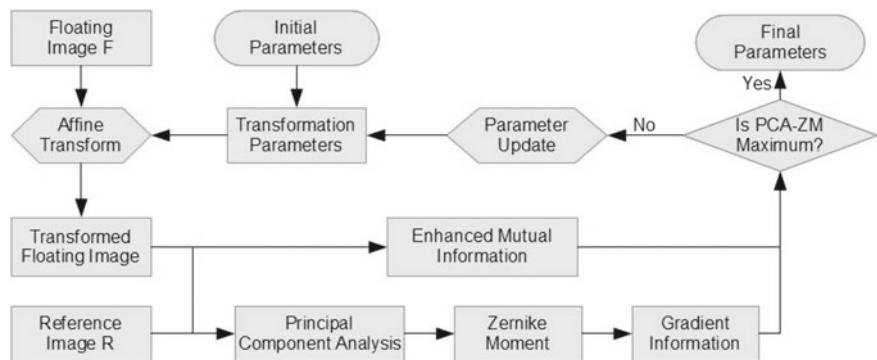
### 3.2 Proposed Principal Component Analysis and Zernike Moment-Based Method (PCA-ZM)

This method operates on a pair of multichannel images by performing Principal Component Analysis (PCA) on each multichannel image. The resulting set of images tend to have common texture and edges (common intensity variance) concentrated in the first few components. The features of each channel which contribute less to the variance will appear in the last few components. Furthermore, the resulting components are orthogonal to each other. If the method is applied to single-channel images, we consider each image to be its own principal component (Fig. 2).

Given a matrix  $\mathbf{X}$  whose rows are mean-subtracted intensities of images of a multichannel set, eigenvectors of  $\mathbf{X}\mathbf{X}^T$  after pre-multiplying by  $\mathbf{X}^T$  are proportional to eigenvectors for nonzero eigenvalues of the covariance matrix  $\mathbf{X}^T\mathbf{X}$ .



**Fig. 1** Single-channel registration features, from top to bottom: Original MR and SPECT images [4], saliency map (scaled) for PCA–ZM, saliency map for CED–ZM, Zernike moment (magnitude, scaled)



**Fig. 2** Flowchart for PCA–ZM [8]

We proceed to find the Zernike moment  $A_{11}$  for each of the principal components. We observe that the  $A_{11}$  moment is a good feature for edge detection, is rotationally invariant, and exhibits resilience to noise. Computing the moment is efficient, as it is linear with respect to the intensities of each pixel within a region. We obtain the magnitude of the moments of each component and use them as an approximation for a map of edges for each channel. These maps are then used for computing the gradient information as defined by [9]:

$$\cos \theta(R_c, F_c, \mathbf{x}_r) = \frac{\operatorname{Re}\{A_{11}[R_c]A_{11}^*[F_c](\mathbf{x}_r)\}}{|A_{11}[R_c]A_{11}[F_c](\mathbf{x}_r)|} \quad (9)$$

$$G_{\text{ZM}}(R_c, F_c) = \sum_{\mathbf{x}_r \in R_c} \cos^2 \theta(R_c, F_c, \mathbf{x}_r) \cdot \min(|A_{11}[R_c](\mathbf{x}_r)|, |A_{11}[F_c](\mathbf{x}_r)|). \quad (10)$$

We have opted for computing the Zernike moments of principal components, instead of obtaining principal components of Zernike moments. Our reasoning for this was, if the channels of the reference set have a region of common intensity variations, there must be a common underlying anatomy to have caused this. In [5], independent component analysis was done after Gabor wavelets transform, which we see as obtaining the regions of similar transitions in tissue, rather than aiming to find regions of similar tissues. Furthermore, PCA is a deterministic decomposition for multimodal images and is fundamentally oriented toward the output intensities as observed, instead of input tissue characteristics as predicted by ICA.

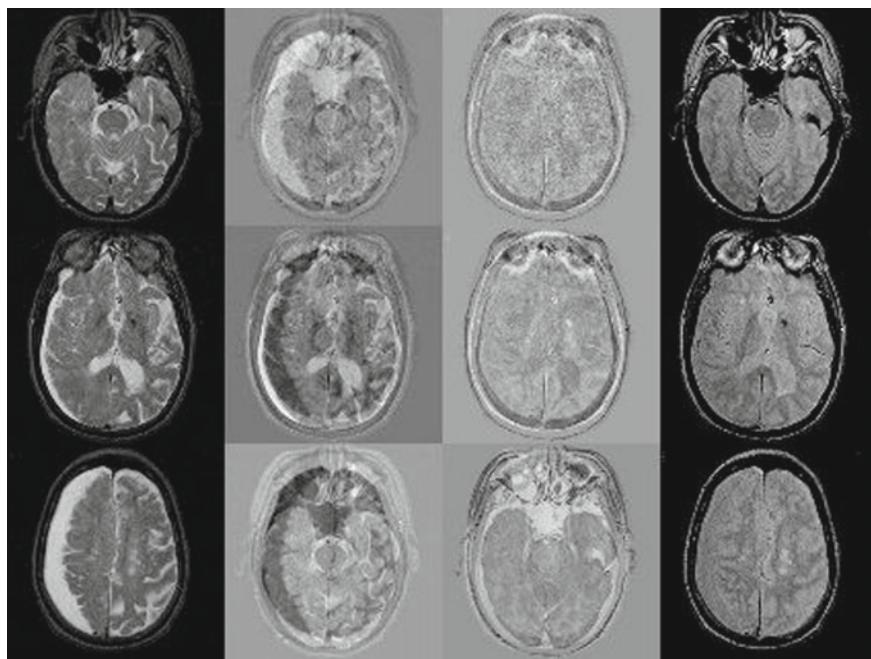
Separately, for each image, we calculate the saliency of each pixel as per [6], to be used for the channel-wise calculation of  $E_{\text{MI}}$  between the multichannel sets. This computation is done as a preprocessing step for each image.

$$\text{PCA-ZM} = \left( \sum_{c=1}^N E_{\text{MI}}(R_c, F_c) \right) \cdot \left( \sum_{i=1}^N G_{\text{ZM}}(\text{PCA}(R)_i, \text{PCA}(F)_i) \right). \quad (11)$$

The final registration measure is a product of the channel-wise sum of  $E_{\text{MI}}$  and component-wise sum of gradient information. Addition was not used to avoid issues with modeling of typical values of the metrics. We will refer to this method as PCA-ZM for future reference (Fig. 3).

### 3.3 Proposed Canny Edge Detector and Zernike Moment-Based Method (CED-ZM)

This method relies on the segmentation of multichannel sets on the basis of gradients. Canny edge detection process is used for identifying the most distinct edge pixels of each image in a multichannel set. A common threshold is used for all images, which we have tuned by a PI control scheme. The intersection of neighborhoods of these

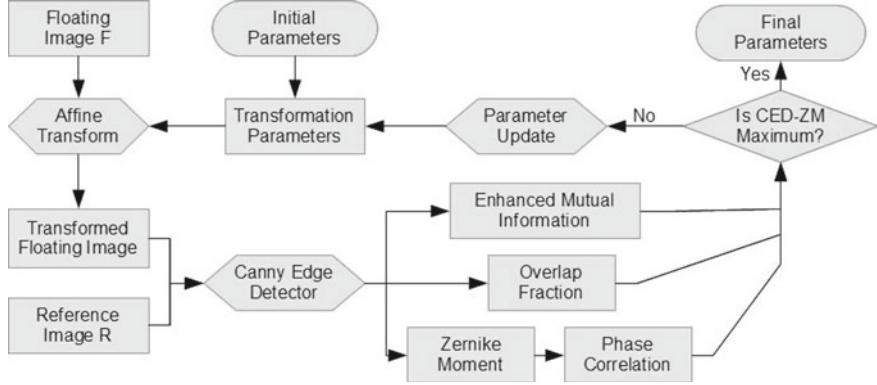


**Fig. 3** Multichannel features for PCA-ZM, from left to right: MR-T2 channel images, principal components for MR-T2 (scaled), principal components for MR-PD (scaled), MR-PD channel images [4]

distinct edges of each image is termed as the Strong Edge Neighborhood (SEN) for the entire multichannel set. Then, using a lower threshold, weaker edges in each image are identified. Voxels are not in the SEN but lie in the union of neighborhoods of weak edges form the Weak Edge Neighborhood (WEN) for the set. Voxels not lying in either neighborhood are referred to as bulk voxels for the set.

We use the above segmentation to define saliency  $\delta$  for all voxels for computation of  $E_{MI}$ . We use a definition of saliency which is different from the one given in [10]. We have set the saliency voxels in SEN as 0, that of bulk voxels to be 1, and the saliency of voxels in WEN to an intermediate value of 0.1. This guides the registration of bulk voxels of the floating set to those of the reference, with a penalty on bulk voxels being registered to weak or strong edges. Furthermore, regions of high-intensity variance and high-intensity gradient are less likely to be registered to regions with high variance but a low gradient, a common characteristic of texture. The first metric based on  $E_{MI}$  is the channel-wise sum of  $E_{MI}$  measures between  $R_c$  and  $F_c$  (Fig. 4).

Overlap fraction of SEN between multichannel sets is used as the second metric to coarsely guide registration. Our reasoning for this was, if the channels of  $R$  have a common region of strong gradients, there must be a common underlying anatomy to have caused this. If the set  $F$  also has a strong edge neighborhood of similar geometry, then both neighborhoods of both sets have the same cause. Furthermore,



**Fig. 4** Flowchart for CED–ZM [11]

if this particular causal anatomy is absent in  $R$ , then the SEN is likely to be very different, and perhaps even absent, in  $F$ . In cases of absence, registration on the basis of overlap of SEN would be inappropriate, and the overlap fraction should be set to 0 to avoid interfering with the registration process. Given the number of voxels in SEN of reference set as  $n(\text{SEN}(R))$ , overlap fraction is

$$O_{\text{SEN}}(R, F) = n(\text{SEN}(R) \cap \text{SEN}(F)) / n(\text{SEN}(R)). \quad (12)$$

With these metrics as a foundation, we use a third metric based on Phase Correlation ( $PC_{\text{ZM}}$ ) for gradients between the weak edge neighborhoods of the image sets to further improve accuracy. First, channel-wise Zernike moments are found. The direction of the gradient at any point is taken as the phase of the Zernike moment at that point. As part of the weak edge neighborhood, it is known that gradients at these voxels are fairly strong, and no further coefficient is used to scale the contribution of phase correlation at pixels with high gradients vis-a-vis that at pixels with low gradients. The phase correlation metric we have used is a channel-wise sum of phase correlation between voxels in  $\text{WEN}(R)$  and corresponding voxels in  $F$ .

We will refer to the final objective function as CED–ZM for future reference. Multiplication is used over addition to avoiding issues with modeling of typical values of the individual metrics. As both edge overlap and, more rarely, phase correlation can be zero despite both multichannel image sets being registered to each other, one is added to both of them.

$$PC_{\text{ZM}}(R_c, F_c) = \frac{1}{n(\text{WEN}(R))} \sum_{x_r \in \text{WEN}(R)} \left( \frac{\text{Re}\{A_{11}[R_c]A_{11}^*[F_c](x_r)\}}{|A_{11}[R_c]A_{11}[F_c](x_r)|} \right)^2 \quad (13)$$

$$\text{CED-ZM} = (1 + O_{\text{SEN}}(R, F)) \cdot (1 + \sum_{c=1}^N PC_{\text{ZM}}(R_c, F_c)) \cdot \sum_{c=1}^N E_{\text{MI}}(R_c, F_c). \quad (14)$$

## 4 Result

We implemented PCA–ZM and CED–ZM in MATLAB 2017a, along with Normalized MI with Gradient information (NMI-G) by [9], NTG measure by [2] and the Expectation–Maximization Global Parametric Registration (EMGPR) measure by [1], for the purpose of performance comparison. NMI-G, NTG, and EMGPR are single-channel registration measures, and their testing was limited accordingly.

Registration was performed on a dataset of medical images from five different patients. These images were sourced from the Whole Brain Atlas [4], a site hosting preregistered brain images. In summary, 79 single-channel and 26 multichannel cases were prepared from 53 axial images, with a mix of MR-MR and MR-SPECT registration pairs. For each patient, 3–4 nonadjacent axial images were used per modality to form one multichannel image set. Images from submodalities under MR have been presented as multimodal MR test cases. Six CT images from one patient were also used, which have been grouped under MR images for uniformity.

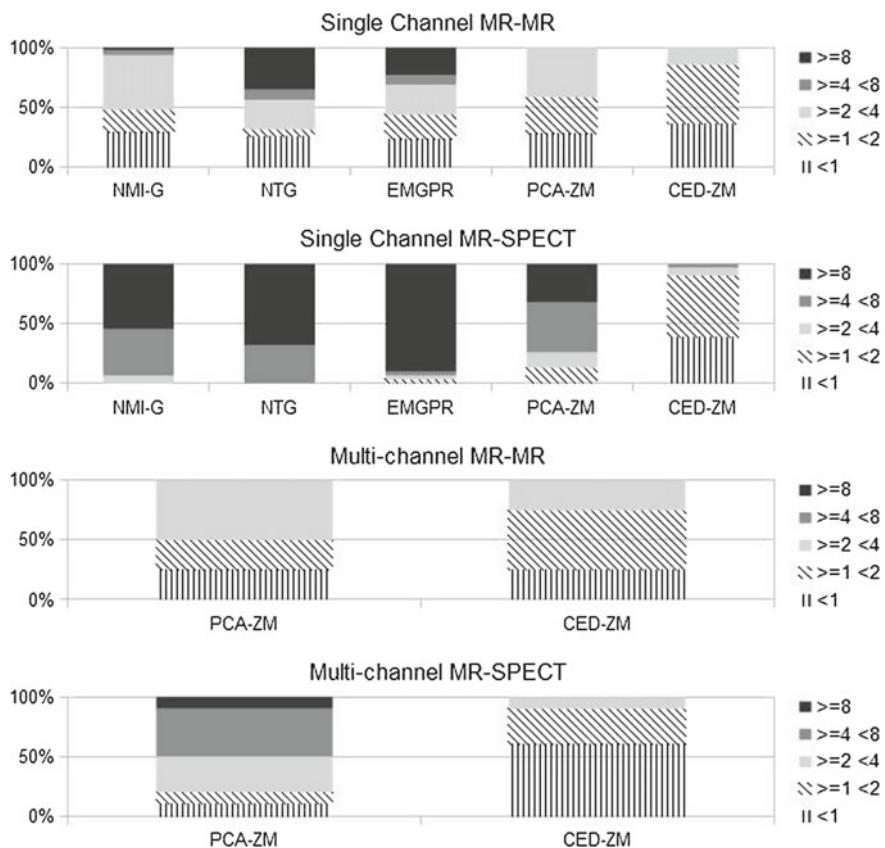
For each test case, the Mean Registration Error (MRE) metric was used to obtain the error in pixels, between the optimized transformation  $\mathbf{h}$  and the gold standard transformation  $\mathbf{g}$ . For our dataset,  $\mathbf{g}(\mathbf{x}) = \mathbf{x}$ .

$$\text{MRE}[\mathbf{h}] = \frac{1}{n(F)} \sum_{\mathbf{x} \in F} \|\mathbf{h}(\mathbf{x}) - \mathbf{g}(\mathbf{x})\|. \quad (15)$$

We considered registration to be unsuccessful when MRE is above eight pixels, where the end user (a doctor or a computer vision algorithm) could identify adverse registration by inspection. Hence, average MRE for successful registration has been given in Table 1 along with the number of “successful” test cases. The overall distribution of MRE is given in Fig. 5, where it can be seen that our methods have given stable transformations, i.e., within eight pixels in most of the test cases, even when other methods had failed with MRE of 100 pixels (point-sized  $F$ ) or 200 pixels (inverted  $F$ ).

**Table 1** Average MRE for successful registration

Method	NMI-G	NTG	EMGPR	PCA-ZM	CED-ZM
MR-MR (Single)	<b>1.947(47)</b>	<b>2.304(31)</b>	<b>2.004(37)</b>	<b>1.723(48)</b>	<b>1.309(48)</b>
MR-SPECT (Single)	5.702(14)	6.179(10)	3.416(3)	4.357(21)	<b>1.305(31)</b>
MR-MR (Multi)				<b>1.734(16)</b>	<b>1.471(16)</b>
MR-SPECT (Multi)				3.585(9)	<b>1.122(10)</b>



**Fig. 5** Percentage distribution of MRE for all test cases

Registration was started from the gold standard transformation. The affine transformation model was used for floating images, with four parameters for rotation, translations, and scaling. Optimization of the measures was performed by Powell's method of multidimensional optimization. Inbuilt MATLAB functions were used for the one-dimensional optimization steps. The testing was done on a workstation with 64-bit 3.2 GHz processor and 8 GB RAM.

PCA-ZM was implemented with pre-calculated saliences for each image. Saliency was determined by finding a local maximum of  $H_D(s, x)$  for  $s \leq 10$  pixels. For CED-ZM, the inbuilt Canny edge detector was used with the lower threshold set at 0.1. We empirically set a target of 2% of image pixels to be detected as edges, which was achieved by tuning the threshold for Strong edges by a PI controller. Upper threshold for Weak edges was set as 0.2 lower than this value. SEN was taken as four pixels radially from Strong edges and one pixel radially from Weak edges for WEN. The saliency of voxels in WEN was set to 0.1. In the case of NTG, the associated image pyramid scheme for registration was implemented with four layers and

a sampling factor of 2, where each layer used Powell’s method. For EMGPR, both expectation and maximization steps were repeated five times to limit computation time.

EMGPR in many cases proceeded toward a point-sized floating image with the scale parameter close to zero. For a point-sized transformation  $\mathbf{h}(\mathbf{x}) = \mathbf{x}_0$ , all pixels of  $F$  correspond to some pixel in  $R$ , and all other pixels of  $R$  do not correspond to any image pixel. Hence, almost intensities  $r$  of  $R$  correspond to the background intensity of  $F$ . Here, conditional variance of  $f$  for almost all  $r$  becomes zero, and the EMGPR cost function is minimized to zero. This contradicts registration, as the gold standard transformation should be a local minimum, if not global.

Unsuccessful registration for NTG may be due to the nature of total gradient. For multispectral images, [2] empirically observed that sharp intensity gradients  $\nabla r(\mathbf{x})$  and  $\nabla f(\mathbf{g}(\mathbf{x}))$  often occur together and their directions are parallel. If two modalities have gradients which are antiparallel, such as due to inverted colors or intensity, the gold standard transformation will lead to maximization of the NTG measure. This was the case with some MR-MR pairs, where significant deviations from  $\mathbf{g}$  occurred due to the inverted contrast between T1 and T2, and the relative absence of gradients in MRPD images. For MR-SPECT pairs, gradients generally do not occur at the same  $\mathbf{x}$ , which contributed to inversion of  $F$  in multiple cases.

## 5 Conclusion and Future Work

We have developed and demonstrated two methods, PCA-ZM and CED-ZM, which incorporate intensity- and gradient-based features. Both methods were developed for registration on the basis of common features of multichannel sets while retaining single-channel performance. These methods were compared with previous methods on a dataset of brain MR and SPECT images. PCA-ZM demonstrated equal accuracy for both single-channel and multichannel MR-MR tests and gave stable results for most multichannel MR-SPECT tests. Performance improvements may be possible by considering the eigenvalues of various components in the objective function, such as when a component corresponds to an infinitesimal eigenvalue. CED-ZM performed well in both single- and multichannel MR-SPECT registration, but its MR-MR performance deteriorated from single-channel to multichannel test cases. For performance in specific applications, tuning of parameters such as saliency, the target for PI controller, and sizes of SEN and WEN may be feasible. We hope these methods will be useful for medical, remote sensing, and natural image registration, where high multimodal registration accuracy is beneficial for diagnosis.

## References

1. Arce-Santana, E.R., Campos-Delgado, D.U., Reducindo, I., Mejia-Rodriguez, A.R.: Multi-modal image registration based on the expectation-maximisation methodology. *IET Image Process.* **11**(12), 1246–1253 (2017)
2. Chen, S.J., Shen, H.L., Li, C., Xin, J.H.: Normalized total gradient: a new measure for multi-spectral image registration. *IEEE Trans. Image Process.* **27**(3), 1297–1310 (2018)
3. Iscen, A., Tölias, G., Gosselin, P.H., Jgou, H.: A comparison of dense region detectors for image search and fine-grained classification. *IEEE Trans. Image Process.* **24**(8), 2369–2381 (2015)
4. Johnson, K.A., Becker, J.A.: The whole brain atlas. [www.med.harvard.edu/aalib/home.html](http://www.med.harvard.edu/aalib/home.html) (1999)
5. Li, Y., Verma, R.: Multichannel image registration by feature-based information fusion. *IEEE Trans. Med. Imaging* **30**(3), 707–720 (2011)
6. Luan, H., Qi, F., Xue, Z., Chen, L., Shen, D.: Multimodality image registration by maximization of quantitative-qualitative measure of mutual information. *Pattern Recognit.* **41**(1), 285–298 (2008)
7. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **16**(2), 187–198 (1997)
8. Nor'aini, A.J., Raveendran, P., Selvanathan, N.: A comparative analysis of zernike moments and principal component analysis as feature extractors for face recognition. In: Ibrahim, F., Osman, N.A.A., Usman, J., Kadri, N.A. (eds.) 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, pp. 37–41. Springer, Berlin, Heidelberg (2007)
9. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Med. Imaging* **19**(8), 809–814 (2000)
10. Pradhan, S., Patra, D.: Enhanced mutual information based medical image registration. *IET Image Process.* **10**(5), 418–427 (2016)
11. Qu, Y.-D., Cui, C.-S., Chen, S.-B., Li, J.-Q.: A fast subpixel edge detection method using sobelzernike moments operator. *Image Vis. Comput.* **23**(1), 11–17 (2005). <http://www.sciencedirect.com/science/article/pii/S0262885604001660>
12. Zhu, F., Ding, M., Zhang, X.: Self-similarity inspired local descriptor for non-rigid multi-modal image registration. *Inf. Sci.* **372**, 16–31 (2016). <http://www.sciencedirect.com/science/article/pii/S0020025516305965>

# A Novel Deep Learning Approach for the Removal of Speckle Noise from Optical Coherence Tomography Images Using Gated Convolution–Deconvolution Structure



Sandeep N. Menon, V. B. Vineeth Reddy, A. Yeshwanth, B. N. Anoop and Jeny Rajan

**Abstract** Optical coherence tomography (OCT) is an imaging technique widely used to image retina. Speckle noise in OCT images generally degrades the quality of the OCT images and makes the clinical diagnosis tedious. This paper proposes a new deep neural network despeckling scheme called gated convolution–deconvolution structure (GCDS). The robustness of the proposed method is evaluated on the publicly available OPTIMA challenge dataset and Duke dataset. The quantitative analysis based on PSNR shows that the results of the proposed method are superior to other state-of-the-art methods. The application of the proposed method for segmenting retinal cyst from OPTIMA challenge dataset was also studied.

**Keywords** OCT · GCDS · Speckle noise · Denoising

## 1 Introduction

Optical coherence tomography (OCT) is a noninvasive medical imaging technique. OCT comes under optical tomography where it is an optical equivalent of ultrasound (US) imaging. In OCT, light echoes are used as opposed to sound echoes in US imaging [1]. OCT images are mainly used to detect ophthalmic diseases and skin disorders.

In ophthalmology, OCT images are used in clinical diagnosis of different retinal diseases like glaucoma, age-related macular edema, cataract, etc. The visual quality of OCT is often degraded by speckle noise [2]. Speckle noise gives OCT images a grainy appearance and obscures small-intensity features. The presence of speckle hinders the ability to distinguish and measure the layers present in the retina efficiently.

---

S. N. Menon · V. B. Vineeth Reddy · A. Yeshwanth · B. N. Anoop (✉) · J. Rajan  
Department of Computer Science and Engineering, National Institute of Technology  
Karnataka, Surathkal, India  
e-mail: [anoopcem@gmail.com](mailto:anoopcem@gmail.com)

Noise also influences other image processing operations like image segmentation, registration, etc. Hence, despeckling is a crucial preprocessing step for the analysis of OCT images.

Despeckling of OCT images comes under the umbrella of image denoising. The following model is considered to express speckle noise [3]

$$f(u, v) = g(u, v) \cdot \gamma_m(u, v), \quad (1)$$

where  $f(u, v)$  is the noisy image,  $g(u, v)$  is the noise-free image, and  $\gamma_m(u, v)$  is the multiplicative component of the speckle noise. From the literature, it is evident that the noise in OCT follows the Gamma probability distribution [4, 5].

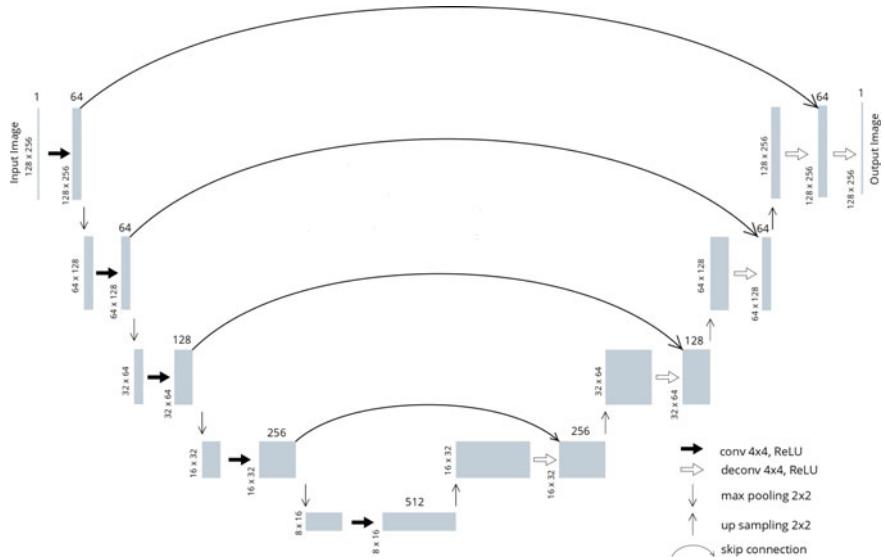
Several methods have been introduced to remove speckle from OCT images. They vary from modeling the speckle noise using probabilistic and statistical methods [6, 7], reconstruction in Fourier domain [8], wavelet domain [9–11], variational image decomposition [12–14], patch-based methods [15–18], bilateral filters [4, 19], shock filters [20], low-rank image decomposition and completion [21–24].

With the advent of deep neural network architectures such as deep convolutional neural networks and stacked denoising autoencoders, promising results in noise removal from natural images [25–32] as well as medical images [28, 33] were observed. The aforementioned papers discuss removal of noise from natural images which is modeled using the Gaussian distribution. Similar to OCT, the synthetic-aperture-radar (SAR) images have an inherent speckle noise. There have been both conventional [34–38] and deep learning [39] methods for removal of speckle noise from SAR images. But deep learning methods for removal of speckle noise from OCT images are not available in the literature. In this paper, we present empirical evidence and qualitative analysis that a conv–deconv stacked structure performs exceptional despeckling (removal of a multiplicative noise) of OCT images.

## 2 Methodology

### 2.1 Model Architecture

The GCDS architecture uses a chain of five convolutional layers followed by five deconvolution (upsampling with convolution) layers. The convolution and deconvolution effectively act as encoding and decoding image features as done in denoising autoencoders [26, 30, 32, 33]. Encoding involves extracting image features that preserve fundamental components while eliminating the noise. The deconvolution layers decode these features to restore the image. The encoding–decoding is evaluated with an assumed noise-free image. This trains the encoder–decoder functions to generate noise-free image from a noisy image. Since encoding does not preserve noise components, the output of the deconvolution layers presents the noise-free image. Inspired by previous models that tried to use symmetric skip connections



**Fig. 1** 10-layer model architecture

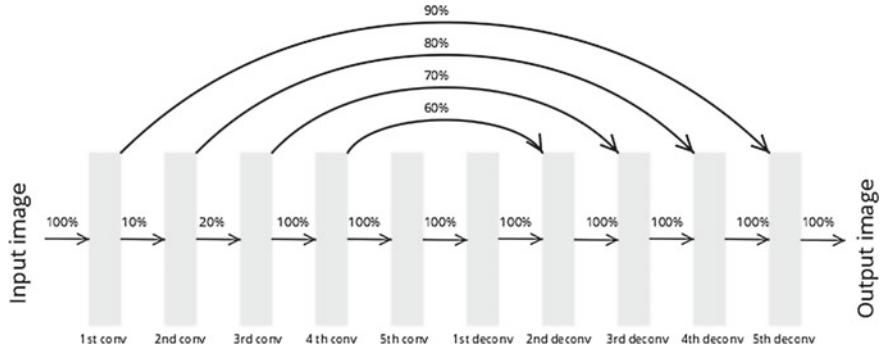
[31] for denoising, we have developed an architecture that despeckles OCT images. The main purpose of the skip connections is to reduce the number of weights that are required when we use a 10-layer model. Thus, the skip connections in turn also provide speed up for the learning. In the architecture, every convolution layer has a skip connection to corresponding deconvolution layer.

As shown in Fig. 1, the Conv\_1 (Conv\_x is the xth convolution layer from left in Fig. 1) receives the initial input grayscale image of size  $128 \times 256 \times 1$ . The convolution layers (Conv\_2 and Conv\_3) receive partial information. Conv\_4 and Conv\_5 layers receive full information from the previous layers. This is so that the inner layers do not get affected by the compounded gating factors and receive meager levels of features [25].

$$X_i = \text{Conv}(\delta * X_{i-1}, W_i) + b_i \quad i \in \{2, 3\} \quad (2)$$

$$X_i = \text{Conv}(X_{i-1}, W_i) + b_i \quad i \in \{4, 5\}. \quad (3)$$

In Eqs. 2 and 3,  $X_i$  denotes the output after the  $i$ th convolution layer,  $W_i$  represents the weight matrix, while  $b_i$  denotes the bias of the  $i$ th layer,  $\delta$  represents the gating factor which determines the degree of split of the information between the next convolution layer and the corresponding deconvolution layer via the skip connection.



**Fig. 2** Dataflow diagram

$$X'_{i+1} = \text{Deconv}(X'_i + (1 - \delta) * X_i, W'_i) + b'_i \quad i \in \{2, \dots, 5\}. \quad (4)$$

In Eq. 4,  $X'_i$  denotes the output after the  $i$ th deconvolution layer,  $W'_i$  represents the weight matrix, while  $b'_i$  denotes the bias of the  $i$ th layer. Figure 2 shows the gating factor and skip connection information. Equations 2 and 3 explain the information that is received by the second four layers from the preceding layer.

Max pooling layers with filter size  $2 \times 2$  are added after each convolution and deconvolution. The deconvolution layers receive the concatenated information from the symmetrically opposite Conv layers and the preceding Deconv layer which is mathematically represented in Eq. 4.

The down-sampled feature set after the encoding layers is decoded to original input dimension using upsampling layers before every deconvolution layers. A  $4 \times 4$  kernel with ReLU activation was used in all convolution and deconvolution layers. The model uses the *adam* optimizer for convergence.

## 2.2 Loss Function

In case of denoising OCT images, we need the denoised images to look as close to the ground truth in terms of the structure. Thus, structural similarity (SSIM) value of the denoised and the ground truth should be high. Hence, we use DSSIM loss to make sure the structure similarity is preserved as shown [40] below:

$$DSSIM = \frac{(1 - SSIM)}{2}. \quad (5)$$

SSIM metric is defined as [40] follows:

$$SSIM(f, g) = l(f, g) \cdot c(f, g) \cdot s(f, g), \quad (6)$$

where

$$l(f, g) = \frac{2\mu_f\mu_g + C_1}{\mu_f^2 + \mu_g^2 + C_1} \quad c(f, g) = \frac{2\sigma_f\sigma_g + C_2}{\sigma_f^2 + \sigma_g^2 + C_2} \quad s(f, g) = \frac{\sigma_{fg} + C_3}{\sigma_f\sigma_g + C_3}. \quad (7)$$

The first term in Eq. 7 is used to compare the luminance between the two images' mean luminance ( $\mu_f$  and  $\mu_g$ ). The second term is used for contrast comparison which is obtained by the standard deviation ( $\sigma_f$  and  $\sigma_g$ ), whereas the last term is used for structure comparison. Here  $\sigma_{fg}$  denotes the covariance matrix between  $f$  and  $g$ . To avoid a null denominator, the positive coefficients  $C_1$ ,  $C_2$ , and  $C_3$ , where  $C_i = k_i L$  ( $k_i$  is of the order of  $10^{-2}$  and  $L$  is the dynamic range of the pixel values) are used.

### 3 Results and Discussion

#### 3.1 Dataset

To evaluate the proposed model, images from different datasets were acquired. The different datasets used along with the number of frames per dataset are shown in Table 1.

Duke dataset [23] contains 18 patients' information, and each patient's information is associated with synthetically generated ground truth images by averaging multiple frames. The OPTIMA challenge dataset [41] has OCT images from four different vendors, namely, Zeiss Cirrus, Nidek, Spectralis Heidelberg, and Topcon.

#### 3.2 Noise Model

The noise in the OCT images is approximated to follow gamma distribution. The gamma distribution with respect to the shape ( $\rho$ ) and scale ( $\beta$ ) is described as [4] follows:

$$f(x; \rho, \beta) = \frac{x^{\rho-1} e^{-\frac{x}{\beta}}}{\beta^\rho \Gamma(\rho)} \quad \text{for } x > 0 \text{ and } \rho, \beta > 0, \quad (8)$$

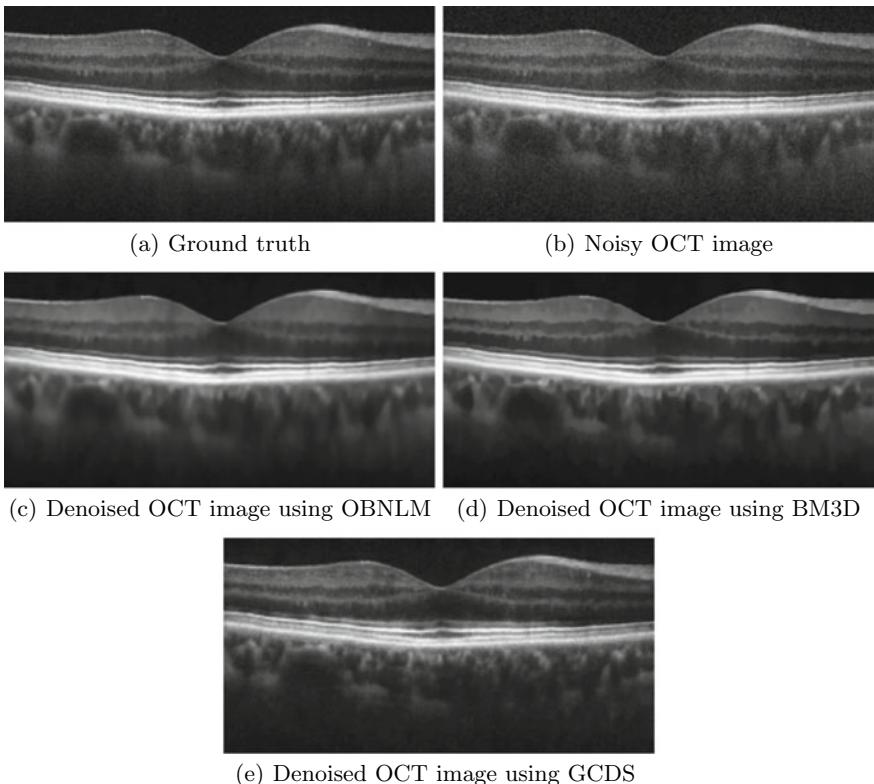
**Table 1** Dataset description

Vendor	No. of frames
Cirrus	128–200
Nidek	7–128
Spectralis	7–49
Topcon	128
Duke	18

where  $\Gamma(\rho)$  is a gamma function defined as [42]:  $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ . The 18 ground truth images obtained from the Duke dataset were extended to 882 by adding different levels of gamma noise, in which both shape ( $\rho$ ) and scale ( $\beta$ ) were varied from 2 to 8.

### 3.3 Experiments on Duke Dataset

This section discusses quantitative and qualitative results of despeckling synthetic OCT images using GCDS and its comparison with optimized Bayesian nonlocal means filter (OBNL) [43] and SAR-BM3D [34]. Figure 3a shows the synthetically generated ground truth (by capturing multiple frames from a unique position followed by averaging [23, 44]) of the image, Fig. 3b depicts the same image corrupted with gamma noise ( $\rho = 4, \beta = 4$ ). It can be observed that the image denoised



**Fig. 3** Results obtained with different denoising algorithms on synthetic OCT image of Duke data set with added Gamma noise ( $\rho = 4, \beta = 4$ )

**Table 2** PSNR (db) comparison

Noise levels	$\rho = 3, \beta = 2$	$\rho = 4, \beta = 3$	$\rho = 4, \beta = 4$	$\rho = 6, \beta = 6$	$\rho = 8, \beta = 7$
OBNLM	<b>32.484</b>	<b>30.462</b>	<b>28.950</b>	25.286	23.514
SAR-BM3D	32.331	30.345	28.633	24.365	21.966
GCDS	28.261	27.346	28.014	<b>26.060</b>	<b>25.616</b>

**Table 3** SSIM comparison

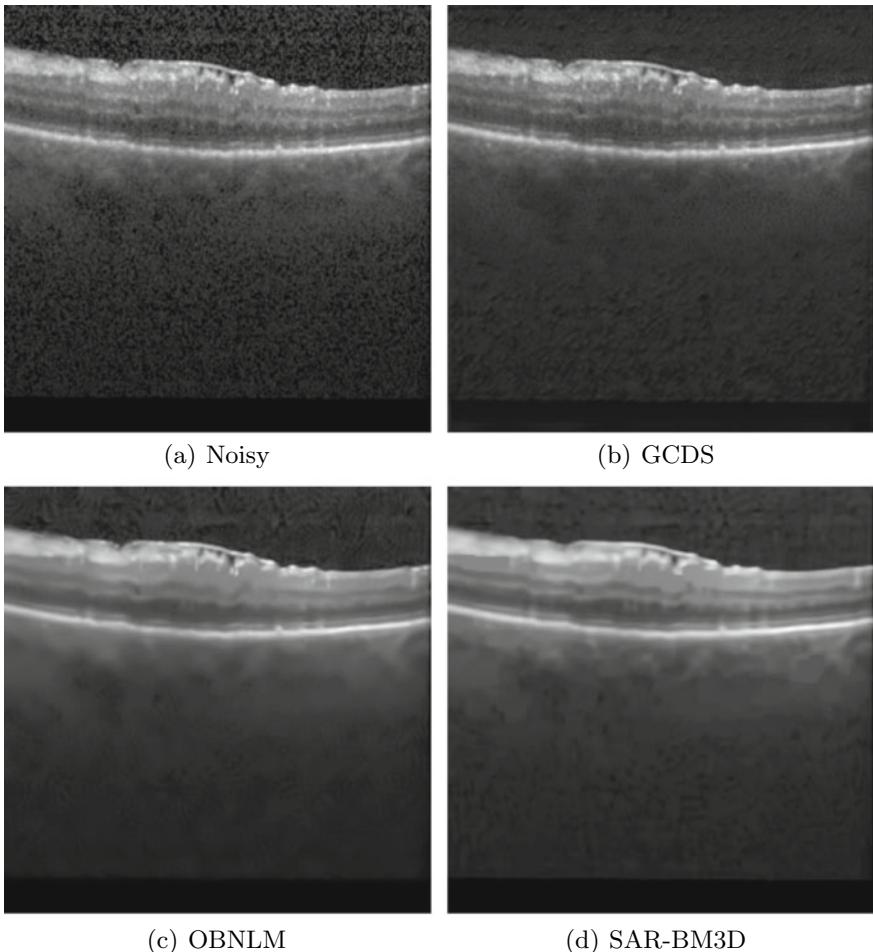
Noise levels	$\rho = 3, \beta = 2$	$\rho = 4, \beta = 3$	$\rho = 4, \beta = 4$	$\rho = 6, \beta = 6$	$\rho = 8, \beta = 7$
OBNLM	0.906	0.903	0.899	0.853	0.782
SAR-BM3D	0.908	0.902	0.894	0.847	0.787
GCDS	<b>0.967</b>	<b>0.942</b>	<b>0.953</b>	<b>0.921</b>	<b>0.912</b>

with OBNLM (Fig. 3c) and SAR-BM3D (Fig. 3d) is more blurred than the image denoised with GCDS. The proposed GCDS model performs well in denoising the image completely while retaining the edge information as depicted in Fig. 3e. Qualitatively the GCDS model provides better gamma noise removal, better preservation of high-frequency regions, and good contrast between regions, compared to other denoising algorithms. The quantitative analysis is done by comparing the peak signal-to-noise ratio (PSNR) [45] and structural similarity index matrix (SSIM) [46]. It can be inferred that the quantitative results shown in Tables 2 and 3 match well with the qualitative results discussed above.

Five different noise levels were chosen that are found to be close with the noise characteristics present in OCT images [4]. We use these noise levels for testing.

### 3.4 Experiments on OPTIMA Dataset

The despeckling algorithms were also applied on the OPTIMA dataset (vendor-spectralis). Figure 4 shows the noisy input frame and its denoised version using GCDS, OBNLM, and SAR-BM3D, respectively. Figure 5 depicts the zoomed portions of noisy and denoised spectralis image. From the figures (Figs. 4 and 5), it is clear that the quality of the image denoised with GCDS is improved and it preserves the details in the edge regions. Also, from the denoised results of OBNLM and SAR-BM3D, it is clear that the methods over smoothen the image and it leads to the blurring of edges in the OCT layer. Even though the quality of the denoised images using GCDS is superior, the formation of artifacts (undesirable intensity variations) introduced by the proposed architecture in the insignificant regions is the matter of concern. Further, we proceed to analyze the cyst segmentation [47] accuracy of the denoised dataset in terms of dice-coefficient (mean and standard deviation). Segmentation results of spectralis data denoised using OBNLM, SAR-BM3D, proposed

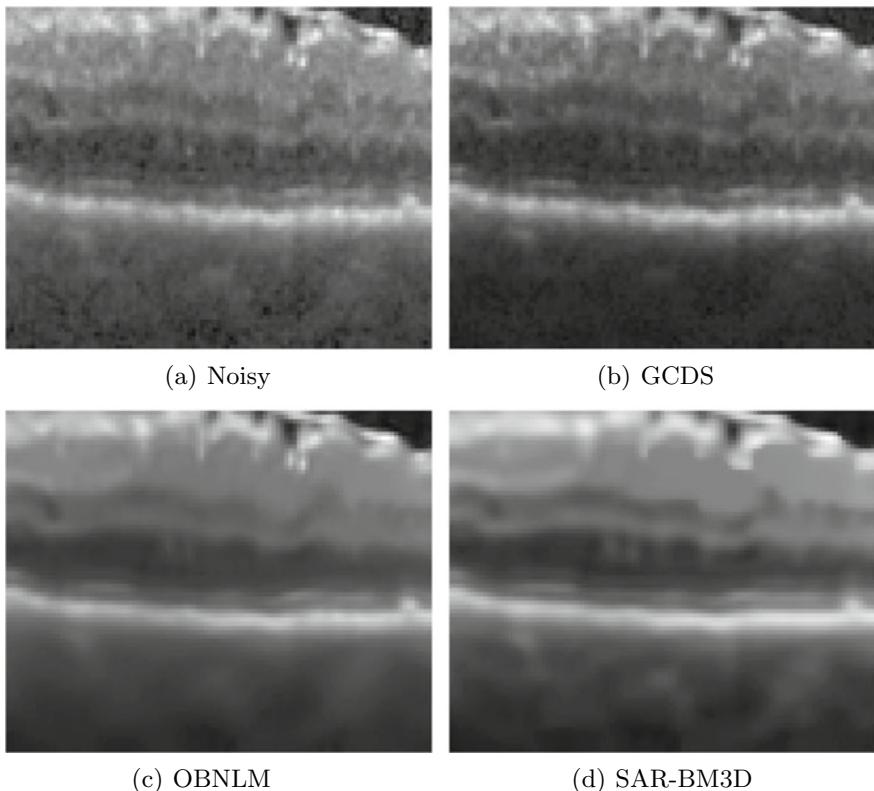


**Fig. 4** Noisy as well as denoised spectralis data

GCDS are shown in Table 4. This proclaims the fact that the proposed method outperforms all the current methods.

### 3.5 *Experimental Setup*

The proposed architecture was implemented in TensorFlow on a workstation with a 64-bit Ubuntu operating system, Intel Xeon Processor E5-2600 (Intel, Mountain View, CA), solid-state hard drive, 128 GB of RAM, and the NVIDIA Quadro K2000 GPU with 2 GB dedicated memory.



**Fig. 5** Zoomed portion of noisy as well as denoised spectralis data

**Table 4** Comparison of segmentation accuracy in terms of dice-coefficient on OPTIMA (Spectralis) dataset

Method	Mean	Standard deviation
SARBM3D	0.66	0.11
OBNLM	0.77	0.15
GCDS	0.82	0.09

## 4 Conclusion

A deep convolutional neural network (CNN) model with additional skip connections called gated convolution–deconvolution structure(GCDS) to denoise the OCT images was presented in this paper. Efficiency of the model was tested on the optima challenge dataset and Duke data set with the underlying assumption that the speckle follows the gamma distribution. Obtained results show that the proposed model

despeckles the OCT images flawlessly. The values of the PSNR and SSIM evidence the above statement.

**Acknowledgements** This work was supported by the Science and Engineering Research Board (Department of Science and Technology, India) through project funding EMR/2016/002677.

## References

1. Fujimoto, J.G., Pitriss, C., Boppart, S.A., Brezinski, M.E.: Optical coherence tomography: an emerging technology for biomedical imaging and optical biopsy. *Neoplasia* **2**(1–2), 9–25 (2000)
2. Li, M., Idoughi, R., Choudhury, B., Heidrich, W.: Statistical model for OCT image denoising. *Biomed. Opt. Express* **8**(9), 3903–3917 (2017)
3. Wong, A., Mishra, A., Bizheva, K., Clausi, D.A.: General Bayesian estimation for speckle noise reduction in optical coherence tomography retinal imagery. *Opt. Express* **18**(8), 8338–8352 (2010)
4. Sudeep, P., Niwas, S.I., Palanisamy, P., Rajan, J., Xiaojun, Y., Wang, X., Luo, Y., Liu, L.: Enhancement and bias removal of optical coherence tomography images: an iterative approach with adaptive bilateral filtering. *Comput. Biol. Med.* **71**, 97–107 (2016)
5. Tao, Z., Tagare, H.D., Beaty, J.D.: Evaluation of four probability distribution models for speckle in clinical cardiac ultrasound images. *IEEE Trans. Med. Imaging* **25**(11), 1483–1491 (2006)
6. Amini, Z., Rabban, H.: Optical coherence tomography image denoising using Gaussianization transform. *J. Biomed. Opt.* **22**(8), 086011 (2017)
7. Rajabi, H., Zirak, A.: Speckle noise reduction and motion artifact correction based on modified statistical parameters estimation in OCT images. *Biomed. Phys. Eng. Express* **2**(3), 035012 (2016)
8. Meiniel, W., Gan, Y., Olivo-Marin, J.-C., Angelini, E.: A sparsity-based simplification method for segmentation of spectral-domain optical coherence tomography images. In: Wavelets and Sparsity XVII, vol. 10394, p. 1039406. International Society for Optics and Photonics (2017)
9. Isar, C.S.-C.A.: Optical coherence tomography speckle reduction in the wavelets domain. Editorial Board 3
10. Du, Y., Liu, G., Feng, G., Chen, Z.: Speckle reduction in optical coherence tomography images based on wave atoms. *J. Biomed. Opt.* **19**(5), 056009 (2014)
11. Mayer, M.A., Borsdorf, A., Wagner, M., Hornegger, J., Mardin, C.Y., Tornow, R.P.: Wavelet denoising of multiframe optical coherence tomography data. *Biomed. Opt. Express* **3**(3), 572–589 (2012)
12. Duan, J., Tench, C., Gottlob, I., Proudlock, F., Bai, L.: New variational image decomposition model for simultaneously denoising and segmenting optical coherence tomography images. *Phys. Med. Biol.* **60**(22), 8901 (2015)
13. Duan, J., Lu, W., Tench, C., Gottlob, I., Proudlock, F., Samani, N.N., Bai, L.: Denoising optical coherence tomography using second order total generalized variation decomposition. *Biomed. Signal Process. Control* **24**, 120–127 (2016)
14. Ren, H., Qin, L., Zhu, X.: Speckle reduction and cartoon-texture decomposition of ophthalmic optical coherence tomography images by variational image decomposition. *Optik-Int. J. Light Electron Opt.* **127**(19), 7809–7821 (2016)
15. Varnousfaderani, E.S., Vogl, W.-D., Wu, J., Gerendas, B.S., Simader, C., Langs, G., Waldstein, S.M., Schmidt-Erfurth, U.: Geodesic denoising for optical coherence tomography images. In: Medical Imaging 2016: Image Processing, vol. 9784, p. 97840K. International Society for Optics and Photonics (2016)
16. Aum, J., Kim, J.-H., Jeong, J.: Effective speckle noise suppression in optical coherence tomography images using nonlocal means denoising filter with double Gaussian anisotropic kernels. *Appl. Opt.* **54**(13), D43–D50 (2015)

17. Chen, Q., de Sisternes, L., Leng, T., Rubin, D.L.: Application of improved homogeneity similarity-based denoising in optical coherence tomography retinal images. *J. Digit. Imaging* **28**(3), 346–361 (2015)
18. Liu, X., Yang, Z., Wang, J.: A novel noise reduction method for optical coherence tomography images. In: International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 167–171. IEEE (2016)
19. Anantrasirichai, N., Nicholson, L., Morgan, J.E., Erchova, I., Mortlock, K., North, R.V., Albon, J., Achim, A.: Adaptive-weighted bilateral filtering and other pre-processing techniques for optical coherence tomography. *Comput. Med. Imaging Graph.* **38**(6), 526–539 (2014)
20. Liu, G., Wang, Z., Mu, G., Li, P.: Efficient OCT image enhancement based on collaborative shock filtering. *J. Healthc. Eng.* (2018)
21. Baghaie, A., D'souza, R.M., Yu, Z.: Sparse and low rank decomposition based batch image alignment for speckle reduction of retinal OCT images. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 226–230. IEEE (2015)
22. Luan, F., Wu, Y.: Application of RPCA in optical coherence tomography for speckle noise reduction. *Laser Phys. Lett.* **10**(3), 035603 (2013)
23. Fang, L., Li, S., McNabb, R.P., Nie, Q., Kuo, A.N., Toth, C.A., Izatt, J.A., Farsiu, S.: Fast acquisition and reconstruction of optical coherence tomography images via sparse representation. *IEEE Trans. Med. Imaging* **32**(11), 2034–2049 (2013)
24. Cheng, J., Duan, L., Wong, D.W.K., Akiba, M., Liu, J.: Speckle reduction in optical coherence tomography by matrix completion using bilateral random projection. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 186–189. IEEE (2014)
25. Zhao, A.: Image denoising with deep convolutional neural networks
26. Cho, K.: Boltzmann machines and denoising autoencoders for image denoising. [arXiv:1301.3468](https://arxiv.org/abs/1301.3468)
27. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Advances in Neural Information Processing Systems, pp. 341–349 (2012)
28. Agostinelli, F., Anderson, M.R., Lee, H.: Adaptive multi-column deep neural networks with application to robust image denoising. In: Advances in Neural Information Processing Systems, pp. 1493–1501 (2013)
29. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* **26**(7), 3142–3155 (2017)
30. Mao, X.-J., Shen, C., Yang, Y.-B.: Image restoration using convolutional auto-encoders with symmetric skip connections. [arXiv:1606.08921](https://arxiv.org/abs/1606.08921)
31. Mao, X., Shen, C., Yang, Y.-B.: Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: Advances in Neural Information Processing Systems, pp. 2802–2810 (2016)
32. Kim, M., Smaragdis, P.: Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures. In: International Conference on Latent Variable Analysis and Signal Separation, pp. 100–107. Springer (2015)
33. Gondara, L.: Medical image denoising using convolutional denoising autoencoders. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 241–246. IEEE (2016)
34. Murali, Y., Babu, M., Subramanyam, M., Giriprasad, M.: A modified BM3D algorithm for SAR image despeckling. *Procedia Comput. Sci.* (Elsevier) **70**(1), 69–75 (2015)
35. Xu, L., Li, J., Shu, Y., Peng, J.: SAR image denoising via clustering-based principal component analysis. *IEEE Trans. Geosci. Remote Sens.* **52**(11), 6858–6869 (2014)
36. Parrilli, S., Poderico, M., Angelino, C.V., Verdoliva, L.: A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage. *IEEE Trans. Geosci. Remote Sens.* **50**(2), 606–616 (2012)
37. Achim, A., Tsakalides, P., Bezerianos, A.: SAR image denoising via Bayesian wavelet shrinkage based on heavy-tailed modeling. *IEEE Trans. Geosci. Remote Sens.* **41**(8), 1773–1784 (2003)

38. Kovaci, M., Isar, D., Isar, A.: Denoising SAR images. In: International Symposium on Signals, Circuits and Systems, 2003. SCS 2003, vol. 1, pp. 281–284. IEEE (2003)
39. Chierchia, G., Cozzolino, D., Poggi, G., Verdoliva, L.: SAR image despeckling through convolutional neural networks. [arXiv:1704.00275](https://arxiv.org/abs/1704.00275)
40. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 2366–2369. IEEE (2010)
41. OPTIMA cyst segmentation challenge (2015). <https://optima.meduniwien.ac.at/research/challenges/>
42. Dubey, S.D.: Compound gamma, beta and f distributions. *Metrika* **16**(1), 27–31 (1970)
43. Coupé, P., Hellier, P., Kervrann, C., Barillot, C.: Nonlocal means-based speckle filtering for ultrasound images. *IEEE Trans. Image Process.* **18**(10), 2221–2229 (2009)
44. Thapa, D., Raahemifar, K., Lakshminarayanan, V.: Reduction of speckle noise from optical coherence tomography images using multi-frame weighted nuclear norm minimization method. *J. Modern Opt.* **62**(21), 1856–1864 (2015)
45. Fisher, Y.: Fractal Image Compression: Theory and Application. Springer Science & Business Media (2012)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)
47. Girish, G., Kothari, A.R., Rajan, J.: Marker controlled watershed transform for intra-retinal cysts segmentation from optical coherence tomography B-scans. *Pattern Recognit. Lett.* <https://doi.org/10.1016/j.patrec.2017.12.019>. <http://www.sciencedirect.com/science/article/pii/S0167865517304658>

# ***D<sup>2</sup>ehazing: Real-Time Dehazing in Traffic Video Analytics by Fast Dynamic Bilateral Filtering***



**Apurba Das, Shashidhar Pai, Vinayak S. Shenoy, Tanush Vinay and S. S. Shylaja**

**Abstract** Traffic video analytics is highly dependent on accurate determination of traffic density in terms of state, position, and velocity of each vehicle in each lane. The accuracy of vehicle detection gets impacted by adverse weather condition like fog/haze and rain. There are many high-accuracy algorithms available for haze detection and removal for still images using bilateral filtering like guided filtering. But unfortunately they are slow and far from real-time dehazing and detection performance in video. This current work has proposed a detailed algorithm and system architecture based on R-CNN and fast bilateral filtering. The proposal of dynamic triggering of detector and dehaze engine has promised to achieve real-time performance in intelligent traffic video analytics in haze environment. The results have demonstrated that the proposed *Dynamic Dehazing (D<sup>2</sup>ehazing)* algorithm is competitive with existing algorithms in terms of both accuracy and performance.

**Keywords** Bilateral filter · Guided filters · R-CNN · Deep learning

---

A. Das · S. S. Shylaja  
PES University, Bengaluru, India  
e-mail: [apurba\\_das1@hotmail.com](mailto:apurba_das1@hotmail.com)

S. S. Shylaja  
e-mail: [shylaja.sharath@pes.edu](mailto:shylaja.sharath@pes.edu)

S. Pai (✉) · V. S. Shenoy · T. Vinay  
PESIT, Bengaluru, India  
e-mail: [shashidharpa195@gmail.com](mailto:shashidharpa195@gmail.com)

V. S. Shenoy  
e-mail: [vinayak96shenoy@gmail.com](mailto:vinayak96shenoy@gmail.com)

T. Vinay  
e-mail: [tanush.vinay@gmail.com](mailto:tanush.vinay@gmail.com)

## 1 Introduction

Intelligent traffic video analytics is an important component of smart city to employ Intelligent Transportation Systems (ITS). Real-time video analytics of intelligent traffic is derivative of vehicle density measurement, localization, and state [5] in each lane. Both dust and water droplets degrade the quality of sensing outdoor environment. The scattering of light through atmosphere leads hazing effect in sensed scene [13]. In traffic surveillance, the video sequence is often captured from a long distance and suffers from poor visibility in deteriorated weather conditions. Hence, a natural tendency is to dehaze the scene which is essentially a computationally complex method. Need of ITS is always real time when the traffic analytics is regarded. Video dehazing is applied post-capturing the same to extract a clear video free of fog, mist from the deteriorated video feed. Dehazing an image is described as follows [7, 13]:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (1)$$

where  $I$  is the observed intensity,  $J$  is the scene radiance,  $A$  is the global atmospheric light, and  $t$  is the medium transmission describing the light that reaches the camera without suffering from scatter due to the dust or water particle in the medium. The goal of any dehazing algorithm is to extract the scene radiance  $J$  from a hazy input image  $I$ .

Sun et al. [13] observed that, in outdoor environment, most of the local patches have lowest intensity in at least one color channel. Based on this observation they proposed the Dark Channel Prior (DCP) model which is considered to be the traditional model for dehazing an image since then. In their later work [7], the use of a guided filter [1] is proposed to improve the quality of the results. However, the guided filter adds to the computational cost further thus limiting it from achieving real-time dehazing. To improve the performance, Zhu et al. [16] have shown improvement in guided filter based performance by treating edges as guide image whereas in the same year another work [15] introduced the color attenuation prior, where the haze is identified from the change in the brightness and contrast of pixels in an image. Very recently Ren et al. [12] used deep neural net based learning of scene transition match to improve performance of dehazing. In traffic video surveillance, dehazing is integral part of the vehicle detection whereas most of the previous works, a subsample of which is cited, focusses on dehazing still image. Cheng et al. [2] attempted to use improved DCP for video dehazing just by treating every frame as uncorrelated image ignoring the time correlation between adjacent frames. We propose a dynamic dehazing ( $D^2ehazing$ ) algorithm based on fast bilateral filtering. The dynamic/adaptiveness has been achieved in two tiers of WHAT and WHERE:

1. Identifying/deciding the key frame (i.e., “WHAT” frame) to trigger transmission/atmosphere parameter estimation.
2. Automated identification of Region of Interest (RoI) for dehazing (i.e., “WHERE” of the video frame).

This aforementioned adaptiveness ensured our proposed algorithm to operate in real time.

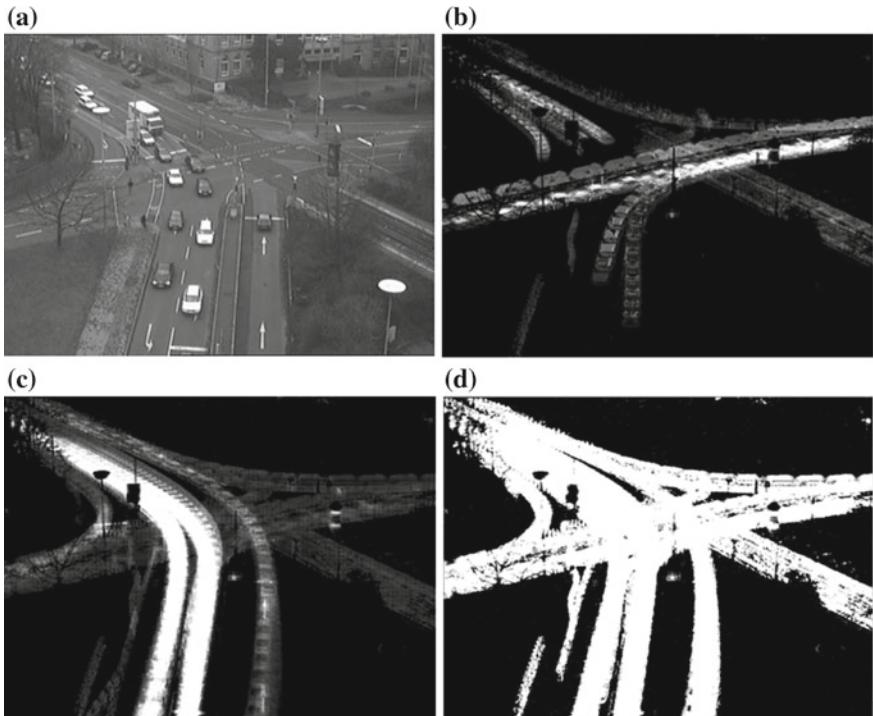
In Sect. 2, two basic components of ITS have been discussed where the intelligent adaptation of Region of Interest (RoI) and R-CNN-based vehicle detection has been discussed to be used as basic detector component in traffic video analytics. The same framework has also been used to validate our dehazing accuracy. Next, in Sect. 3, the proposal of real-time dehazing algorithm of traffic video has been depicted in detail. The experimental results have been shown to illustrate the effectivity of our proposed dehazing algorithm in Sect. 4. Finally, in Sect. 5, we have concluded our findings with a direction to the future research.

## 2 Detection of Vehicles in Adaptive RoI

Recent development of deep convolutional neural network has shown impressive results in the area of vehicle detection of different types through R-CNN networks like You Only Look Once (YOLO) [10] and Single Shot Detector (SSD) [9]. For all our experiments, we have used YOLO and similar results should be obtained from SSD, too. Our proposal of dehazing can be considered as a plug-in to any vehicle detection system. The YOLO darknet model [11] has been already trained on 80 predefined classes. As the model is trained keeping automotive domain also in mind, it can detect and classify cars, buses, trucks, motorbikes, and so on.

In this work, our objective is not only to improve the accuracy of the vehicle detection in hazy traffic video but also ensuring the same in real time. Hence, choosing proper region of interest where the algorithm of vehicle detection to be applied as very important. This would not only help the system to achieve desired performance but also will reduce the dependencies of sensor (i.e., camera) specifications. The production agility would definitely be improved if the algorithm can adaptively understand the region of interest and learn the perspective matrix automatically through the calculation of vanishing point.

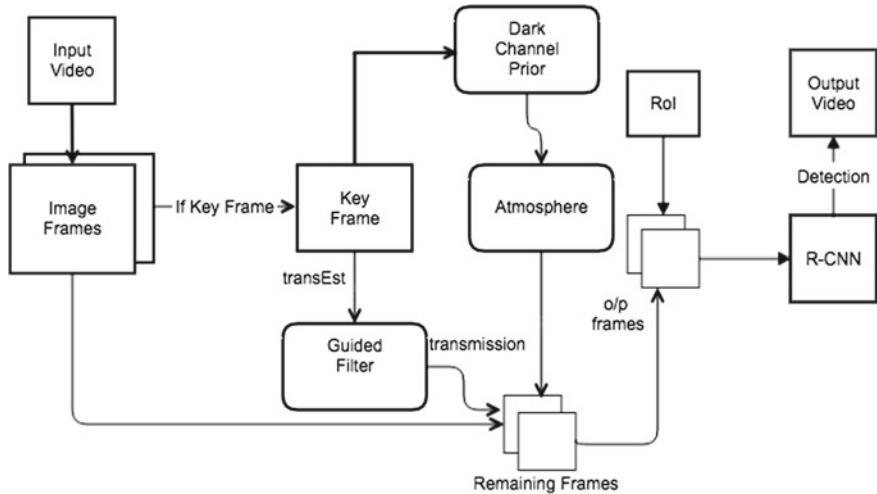
To achieve the same, we have used Gaussian Mixture Model (GMM) Learning [8] based background modeling followed by hole filling in the extracted region of route. The idea is to understand the moving objects first through GMM and apply simple frame averaging across the region of moving blob detection through GMM. After approximately 1 min of run, most of the route is averaged and we get (1) the background model, (2) the RoI with some holes as the averaging is done only at the position of the moving blobs. Next the intermediate holes are filled up by simple morphological closing operation [3]. The complete process has been depicted in Fig. 1. This operation is like unsupervised learning of the background model and needs to be triggered once every 4–6 h based on change in environmental illumination. The same RoI boundaries have been used to calculate perspective map from vanishing point derived.



**Fig. 1** Adaptive ROI (AROI) determination for triggering vehicle detector and dehazing algorithm online: **a** original frame, **b** moving object segregation, **c** AROI after 2 s, **d** AROI after 15 s of video run

### 3 Detailed Algorithm for Dehazing Traffic Video in Real Time

The proposed solution of dehazing a video feed has been depicted in Fig. 2. There are three principal modules of the dehazing algorithm. First, the *darkchannel* [13] is extracted as haze map. Next the atmospheric color of the haze, called the *atmosphere*, is extracted. The amount of light that passes through the haze from the object, called as *transmission* parameter, is calculated finally. As we know the environmental condition does not drastically change momentarily, we have exploited this behavior of the environment. The said environmental behavior inspired us to determine the *transmission* and *atmosphere* parameters only for key frames. The calculation for determining the aforementioned (*transmission* and *atmosphere*) parameters is maximally computationally complex. Once parameters are determined, the



**Fig. 2** Architecture of the proposed system of real-time dehazing of traffic video

dehazing filters are applied on all frames based on the calculated parameters from the key frame. Here the key frames are decided by squared difference between gray converted histograms ( $H_{time}(intensity = i)$ ) of periodic image frames as depicted in Eqs. 2 and 3. The default period is defined as 5 s and user has provision to change the interval ( $T$ ) from dynamic configuration file used during execution. The frame at  $time = (t + T)$ th is key frame if Eq. 2 satisfies, i.e., the squared difference between periodic histogram is higher than the predefined threshold,  $Th$ .

$$\sum_{i=0}^{256} (H_t(i) - H_{t+T}(i))^2 > Th \quad (2)$$

$$t = t + T \quad (3)$$

Primary module in dehazing an image is extracting the haze map called the dark channel. The dark channel identifies and maps how hazy each pixel is. Sun et al. [13] have shown that in most local regions which do not cover the sky, it is very often that some pixels (called “dark pixels”) have very low intensity in at least one color (RGB) channel. Therefore, these dark pixels can directly provide accurate estimation of the haze’s transmission.

In Algorithm 3.1, the arguments or parameters used to compute dark channel are input frame (*image*), spatial shift of image filter (*stride*), and size of kernel (*winSize*).

---

**Algorithm 3.1:** DARKCHANNEL(*image*, *stride*, *winSize* = 15)
 

---

**comment:** Output = darkChannel

```
(m, n) := size(image)
padSize := winSize/2
darkChannel := zeros(m × n)
paddedImage := pad(image, ∞)

for i ∈ (1 … m)
  do { for j ∈ (1 … n)
    do { patch = getPatch(paddedImage, windowSize)
      darkChannel(i … i + stride, j … j + stride) = min(patch)
      j := j + stride
    }
    i := i + stride
  }
return (darkChannel)
```

---

We have modulated the spatial shift of image filter (i.e., the kernel), *stride* to improve the performance of the filter as follows. The input image is first padded with zeros based on the window size. For every (*i*, *j*)th pixel minimum from the three channels from a local region around the pixel is extracted and output is stored in a 2D array of same position. This computation is traditionally performed over every pixel of an image which makes getting the dark channel an expensive component of the dehazing algorithm. The function *getPatch()* picks a patch of the image of size *winSize* for further operations. A stride is used to skip over computation of a single or multiple rows and columns of pixels to improve computational efficiency leveraging already computed filter response and interpolate. This gives us a quicker solution to derive the dark channel map. Pseudocode 3.1 has depicted the proposed idea.

The dark channel gives us a map of the hazy pixels, and the next step is to get the RGB color of the light from these hazy pixels. This is called the atmospheric light and is the second main component of the dehaze algorithm. The RGB color of the *atmospheric* light can be derived from the RGB color of the haziest pixels in the dark channel, and this is where the haze is dense, so all the color is from the reflected atmosphere rather than from the objects behind the haze. The top 0.1% of the brightest pixels (with high intensity) from the dark channel are taken, which gives us the haziest pixels in the image as these are maximally haze opaque [13]. Finally, the input image pixels (RGB) within this aforementioned group of haze opaque pixels are treated as the “atmospheric light” [13].

The third main component involves extracting the *transmission* parameter, which is the amount of light that passes through the haze from the object. It will mostly look opposite to the dark channel picture. Direct use of the transmission leads to certain halo effect or artifacts in the dehazed image, and hence the transmission is passed through a guided filter to smoothen fantom color of the image. The guided filter is an edge-preserving smoothing filter which uses a guide image to enhance the input image [4]. The result of the guided filter is a smoothed transmission, which is then used to dehaze the input image. With an estimate of the transmission and the atmospheric/ambient light the original image has been recovered. Essentially the atmospheric light is subtracted from each pixel in proportion to the transmission at that pixel.

The identification/configuration of key frame, determining *transmission* and *atmosphere* parameters from the key frame, and applying the same to all frames in ROI to achieve dehazed video have been depicted in Pseudocode 3.2. In Fig. 3, one sample hazy frame has been shown along with the dehazed frame by our proposed algorithm.

---

**Algorithm 3.2:** DEHAZVIDEO(*VideoFrames*, *stride*)

---

**comment:** Output = Dehazed Video

(*Interval*, *radius*, *regularization*) := InitializeParams(*interAdap*, 15, 0.001)  
*keyFrame* = ExtractKeyFrame(*VideoFrames*)

**while** *frame* ∈ *VideoFrames*

**if** *frame* == *keyFrame*

darkChannel := GetDarkChannel(*frame*, *radius*, *stride*)

(Atmosphere, *transEst*) := getAtmo(*frame*, *darkChannel*,

...  $\omega$  = 0.95, *radius*, *stride*)

Transmission = GuidedFilter(grey(*frame*), *transEst*,

rad, *regularization*)

**do** {

*frame* := ROI(*frame*)

*DehazedFrame* = ((*frame* - Atmosphere)./*Max(Transmission)*)

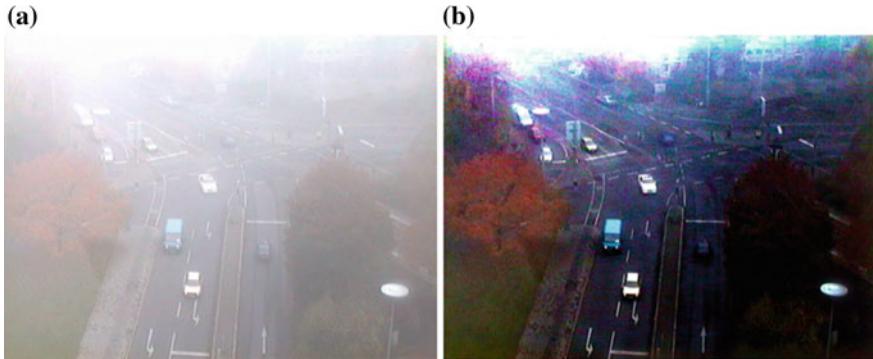
... + AtmosphereMatrix

*OutputVideo* = Video(*DehazedFrames*)

---

## 4 Experimental Results

Earliest visibility is always the key in intelligent transportation systems, especially the traffic video analytics. Fog or haze has always been a hinderance to that. Haze generally impacts on the contrast sensitivity of a scene nonlinearly with respect to distance from camera. Hence, visibility of vehicles coming from really far is always poor and hence the detection accuracy also is very low. Table 1 has depicted comparative duration of visibility before and after applying our proposed dehazing algorithm.



**Fig. 3** Dehazing by our proposed algorithm: **a** Hazy image from a traffic video [6], **b** dehazed frame

**Table 1** Comparative visibility duration ( $VT$ ) before and after applying our proposed dehazing algorithm. The result has been depicted on sample video [6] of 15 s with total 11 cars in the scene

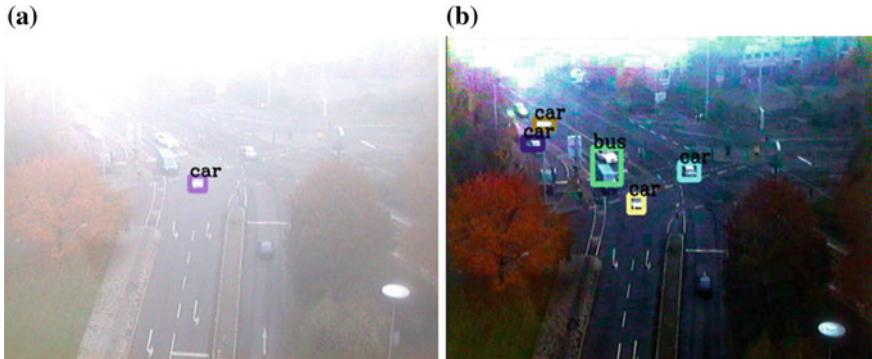
Car ID→	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
$VT_{Raw}$ (s)	1	0.5	0	0	0	3	4	1.5	0	0	0
$VT_{ARoIkFrm}$ (s)	4	3	4	1.5	3	3	6	6	1	2	2
$VT_{GT}$ (s)	4.2	3	4	1.5	3	3	6	6	1.3	2	2

The result has been depicted on sample video [6] of 15 s having total 11 cars available in the scene. Frame size is  $768 \times 576$  square pixels. The terms  $VT_{ARoIkFrm}$  and  $VT_{GT}$  are presenting the count of our proposed dehazing algorithm based on adaptive ROI and ground truth, respectively. As mentioned in the previous section, the standard vehicle detection algorithm considered is darknet implementation [11] of YOLO [10] for both the stages of raw and dehazed video. Table shows almost 100% accuracy with respect to the ground truth ( $VT_{GT}$ ).

Figure 4 has illustrated detection improvement by our proposed dehazing algorithm.

As we described, the dynamic fast bilateral filtering based dehazing and YOLO-based vehicle detection have been integrated employing two-tier adaptation: (a) Adaptive ROI to reduce the area of triggering the filter, (b) Calculation of transmission/atmospheric parameters once and apply always. The aforementioned idea has helped us to achieve improved accuracy (Table 1) and performance simultaneously (Table 2). On the other hand, Fig. 5 has depicted the comparative result of our proposed algorithm for stride 3 and 6 for single image dehazing. The test images of different sizes (legends of the graph) have been derived from the FRIDA (Foggy Road Image DAtabase) dataset [14].

The last row of Table 2 has been calculated based on experiment in an Intel Core i7-7700 HQ @ 2.70 GHz having NVIDIA GeForce GTX 1070 GPU with 8GB RAM, whereas first two rows of the aforementioned table are presenting same CPU time



**Fig. 4** Detection improvement by YOLO after dehazing through our proposed algorithm: **a** YOLO-based vehicle detection on hazy video [6], **b** YOLO-based vehicle detection over dehazed video by our proposed algorithm

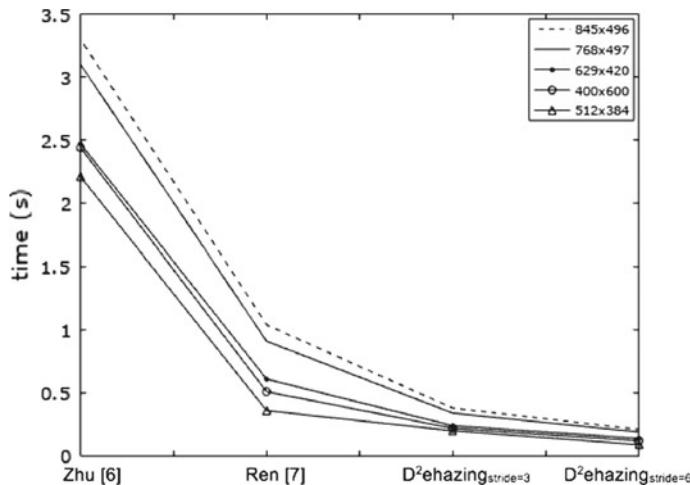
**Table 2** Performance comparison for image-, video-based dehazing, and dehazing+vehicle detection with state-of-the-art algorithms

	Sun et al. [13]	Zhu et al. [15]	Ren et al. [12]	<i>Proposed</i> <sub>Stride=3</sub>	<i>Proposed</i> <sub>Stride=6</sub>
<i>Dehaze</i> <sub>image</sub> (s)	39.87	2.21	0.36	0.19	0.098
<i>Dehaze</i> <sub>video</sub> (s)	17941.5	994.5	162	3.78	3.76
<i>Dehaze</i> + <i>VD</i> <sub>video</sub> (s)	6120	97.5	25.74	22.28	22.27

without GPU. The second and third rows present number for processing 450 frames. It is to be noted that for single image dehazing the performance is close to the state of the art whereas with increasing number of frames, the enhancement in performance has been prominent. Processing 450 frames (i.e., 15 s video of 30 FPS) in 22 s infers around 20.5 FPS processing time which is near real time.

## 5 Conclusion

In today's world of Internet of Things (IoT), smart city concept is coming up really fast to the realization. To realize a proper smart city, the core module is traffic video analytics with very high accuracy in any environmental condition. The same should run in real time as well to generate alert for violations, anomalies from general statistics of the traffic flow, and so on. The current paper has proposed a dynamic fast bilateral filtering concept for traffic video analytics in haze environment. The dynamic/adaptiveness has been brought into the system in two tiers. Primarily the WHERE of the algorithm to be triggered is analyzed through an unsupervised machine learning approach through motion modeling. This has drastically reduced the RoI for vehicle detection. On the other hand, detection of transmission



**Fig. 5** Performance comparison for single image dehazing

and atmospheric parameters, being of highest computational complexity, has been triggered only once per predefined duration (e.g., 15 min) and calculated aforementioned parameters have been applied on the following frames real time. The proposed algorithm is not tightly coupled to any vehicle detector engine or dehazing engine. The proposed algorithm is agile enough to be treated as a plug-in to any generic available infrastructure. The experimental results have shown promises in terms of improvement of vehicle detection accuracy and performance both for dehazing and vehicle detection with dehazing. In future, both accuracy and performance can be further improved employing deep neural network for offline parallel processing for determining aforementioned parameters and offline adaptation of pretrained vehicle detection module.

## References

1. Caraffa, L., Tarel, J.P., Charbonnier, P.: The guided bilateral filter: when the joint/cross bilateral filter becomes robust. *IEEE Trans. Image Process.* **24**(4), 1199–1208 (Apr 2015). 10.1109/TIP.2015.2389617, <http://perso.lcpc.fr/tarel.jean-philippe/publis/ip15.html>
2. Cheng, Y., Niu, W., Zhai, Z.: Video dehazing for surveillance unmanned aerial vehicle. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pp. 1–5 (2016)
3. Das, A.: Guide to Signals and Patterns in Image Processing. Springer (2015)
4. Das, A., Nair, P., Shylaja, S.S., Chaudhury, K.N.: A concise review of fast bilateral filtering. In: 2017 Fourth International Conference on Image Information Processing (ICIIP), pp. 1–6 (2017)
5. Das, A., Ruppin, K., Dave, P., Pv, S.: Vehicle boundary improvement and passing vehicle detection in driver assistance by flow distribution. In: 2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6 (2017, November)
6. Hazy Traffic Videos Datasets. [http://i21www.ira.uka.de/image\\_sequences/](http://i21www.ira.uka.de/image_sequences/). Last accessed Nov 2017

7. He, K., Sun, J., Tang, X.: Guided image filtering. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1397–1409 (2013)
8. Lee, D.S.: Effective gaussian mixture learning for video background subtraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(5), 827–832 (2005)
9. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision—ECCV 2016*, pp. 21–37. Springer, Cham (2016)
10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016, June)
11. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv (2018)
12. Ren, W., Liu, S., Zhang, H., Pan, J., Cao, X., Yang, M.H.: Single image dehazing via multi-scale convolutional neural networks. In: European Conference on Computer Vision, pp. 154–169. Springer (2016)
13. Sun, J., He, K., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 2341–2353 (2010)
14. Tarel, J.P., Hautire, N., Cord, A., Gruyer, D., Halimaoui, H.: Improved visibility of road scene images under heterogeneous fog. In: IEEE Intelligent Vehicles Symposium (IV’10) (2010)
15. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* **24**(11), 3522–3533 (2015)
16. Zhu, X., Li, Y., Qiao, Y.: Fast single image dehazing through edge-guided interpolated filter. In: 14th IAPR International Conference on Machine Vision Applications, MVA 2015, pp. 443–446. IEEE (2015)

# Joint Bit Allocation for 3D Video with Nonlinear Depth Distortion—An SSIM-Based Approach



Y. Harshalatha and Prabir Kumar Biswas

**Abstract** Perceptual quality improvement approach for 3D video through bit allocation is presented in this paper. Bit allocation between texture video and depth map plays an important role in deciding quality of synthesized views at the decoder end. To have better visual quality, structural similarity (SSIM) index is used as a distortion metric in rate distortion optimization (RDO) of the 3D video. In this paper, we used the nonlinear relationship of depth distortion with synthesis distortion in computing rate distortion cost resulting in better mode decision. Using the same depth map RDO in bit allocation algorithm, more accurate results are obtained when compared to the linear relation of depth distortion with synthesis distortion.

**Keywords** 3D video · Perceptual quality · Bit allocation

## 1 Introduction

The 3D video is a motion picture format that gives the real depth perception and thus gained huge popularity in many application fields. The depth perception is possible with two or more videos. With two views, the stereoscopic display produces a 3D vision and the viewer needs to wear special glasses. Multiview acquisition, coding, and transmission are necessary for an autostereoscopic display to provide 3D perception to the viewers without special glasses. Instead of multiple views, representation formats are used and only a few views along with the depth maps are coded and transmitted. Virtual view synthesis [2] is used to render intermediate views. These views are distorted and thus affect the end display.

Normally, rate distortion optimization (RDO) computes the rate distortion (RD) cost for a macroblock (MB). If distortion is measured using metrics like structural

---

Y. Harshalatha (✉) · P. K. Biswas  
Indian Institute of Technology Kharagpur, Kharagpur, India  
e-mail: [harshalatha.y@ece.iitkgp.ernet.in](mailto:harshalatha.y@ece.iitkgp.ernet.in)

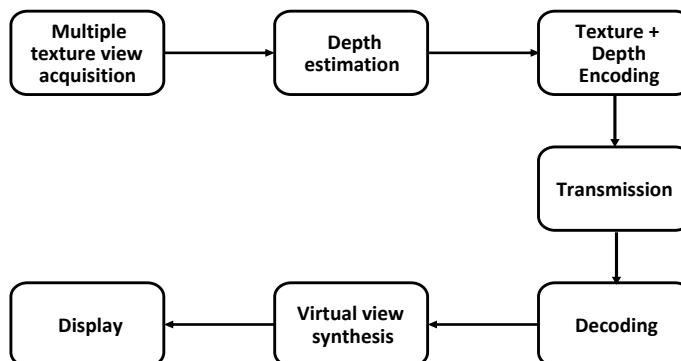
P. K. Biswas  
e-mail: [pkb@ece.iitkgp.ernet.in](mailto:pkb@ece.iitkgp.ernet.in)

similarity (SSIM) index, the quality of the reconstructed block matches the human vision. In [5], the authors extended the perceptual RDO concept to 3D video by considering the linear relationship between depth distortion and synthesis distortion. However, depth map distortion is nonlinearly related to view synthesis distortion, i.e., depth distortion remains the same whereas synthesis distortion varies according to the details in the texture video. In this paper, a suitable Lagrange multiplier is determined for RDO using nonlinear depth distortion. We used this framework of RDO in SSIM-based bit allocation algorithm [6].

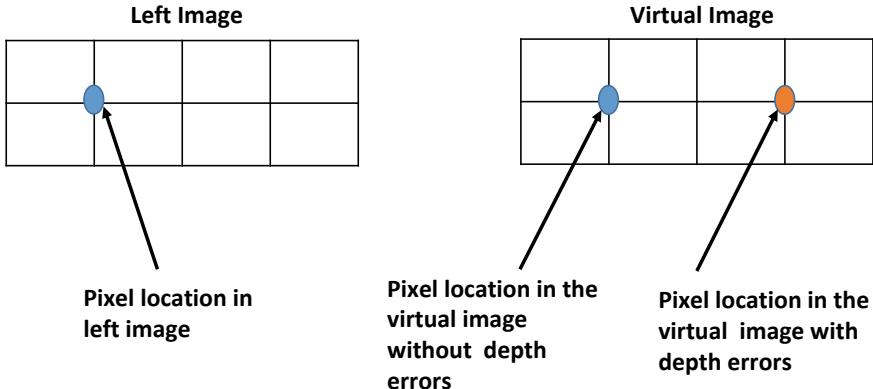
Section 2 gives a brief explanation of view synthesis distortion and its relation with texture and depth distortion. In Sect. 3, SSIM index is discussed briefly. In Sect. 4, we discuss the RDO process with linear and nonlinear depth distortion. Bit allocation criteria for 3D video is explained in Sect. 5. Performance evaluation of the proposed method is discussed in Sect. 6 and concluding remarks are given in Sect. 7.

## 2 View Synthesis Distortion

3D video system shown in Fig. 1 has multiple videos as inputs and uses multiview plus depth (MVD) representation format that is more economical compared to other representation formats. View synthesis process generates a new intermediate view (target view) from the available texture views (reference views) and depth maps. It consists of two steps: warping and blending [10]. Warping is a process to convert the reference viewpoint to 3D point and then to target viewpoint. This pixel mapping is not one-to-one mapping, and thus holes are created. These holes are filled by the blending process. Warping uses depth data in converting a reference viewpoint to target viewpoint and accuracy of conversion depends on depth data. As lossy compression method is used in encoding depth map, it affects the warping process and causes distortion in the synthesized view.



**Fig. 1** Block diagram of 3D video system



**Fig. 2** Geometric error due to inaccurate depth data

Distortion in synthesized views is mainly due to texture distortion and depth inaccuracy. Distortion model derived in [14] as well as in [8] assumed that texture and depth distortion ( $D_t$  and  $D_d$ ) are linearly related to synthesis distortion ( $D_v$ ) as in Eq. 1.

$$D_v = AD_t + BD_d + C \quad (1)$$

During the synthesis process, geometric errors will be minimum if the depth data is accurate. Inaccurate depth data leads to change in pixel position as shown in Fig. 2. This position error is linearly proportional to depth error as given in Eq. 2 [7].

$$\Delta P = \frac{f \cdot L}{255} \left( \frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right) \quad (2)$$

where  $f$  represents camera focal length,  $L$  is the baseline distance, and  $Z_{near}$  and  $Z_{far}$  are nearest and farthest depth values, respectively. However, synthesis distortion depends on details in the texture video and is not the same in all the regions with same position error. This implies, even with the same amount of geometric error, degradation of synthesized view will be different for textured and textureless areas. Texture area with edge information will have more error compared to smooth regions. Considering these factors, synthesis distortion caused by depth distortion is formulated as in Eq. 3 [13].

$$D_{d \rightarrow v} = \Delta P \cdot D_d \cdot [D_{t_{(x-1)}} + D_{t_{(x+1)}}] \quad (3)$$

where  $D_{t_{(x-1)}}$  and  $D_{t_{(x+1)}}$  are the horizontal gradients computed between collocated texture blocks.

### 3 SSIM Index

Traditional methods for image quality measurement use objective evaluations and most of the metrics do not match with human visual characteristics. Human vision is sensitive to structural information in the scene and the quality metric must measure the structural degradation. Wang et al. [12] proposed structural similarity (SSIM) that measures structural degradation and thus evaluates according to human vision. For measuring SSIM, two images are required and measurement is done at the block level. For each block  $x$  and  $y$ , three different components, namely, luminance ( $l(x, y)$ ), contrast ( $c(x, y)$ ), and structure ( $s(x, y)$ ) are measured and therefore SSIM is expressed as in Eq. 4.

$$SSIM(x, y) = f(l(x, y), c(x, y), s(x, y)) \quad (4)$$

Instead of similarity measure, we need distortion based on SSIM in RDO. Therefore, dSSIM is used and is defined as  $dSSIM = \frac{1}{SSIM}$ .

### 4 SSIM-Based RDO with Nonlinear Depth Distortion

Rate distortion optimization helps to reduce the distortion of reconstructed video with minimum rate while increasing the computation complexity. SSIM is used instead of sum of squared error (SSE) in mode decision and motion estimation to improve the visual quality. In our previous work [5], the linearity of texture and depth distortions with view synthesis and suitable Lagrange multiplier is determined as in Eq. 5.

$$\lambda_{new} = \frac{2\sigma_{x_i}^2 + C_2}{S_f \left( \exp\left(\frac{1}{M} \sum_{j=1}^M \log(2\sigma_{x_j}^2 + C_2)\right) \right)} \lambda_{SSE} \quad (5)$$

where  $S_f$  is the scaling factor,  $\sigma_{x_i}^2$  is the variance of  $i$ th macroblock,  $M$  is the total number of macroblocks, and  $C_2$  is constant to limit the range of SSIM.

Depth map RDO is performed by computing RD cost as in Eq. 6.

$$J_d = \Delta P \cdot D_{t_G} \cdot D_d + \lambda_{SSE} R_d \quad (6)$$

where  $R_d$  is the depth map rate and  $D_{t_G} = D_{t_{(x-1)}} + D_{t_{(x+1)}}$  is horizontal gradient computed from texture video. For depth map RDO, Eq. 6 is minimized and Lagrange multiplier is derived as in Eq. 7.

$$\lambda_{i(d)} = \frac{\lambda_{new}}{S_f \cdot \kappa} \lambda_{SSE} \quad (7)$$

where  $\kappa = \Delta P \cdot D_{t_G}$ .

## 5 Bit Allocation Algorithm

In 3D video, bit rate must be set properly between the views to improve the virtual view quality. In the literature [3, 9, 13–16], many joint bit allocation methods are proposed, and all these methods improve the PSNR of synthesized views. Visual quality enhancement can be achieved by using dSSIM as distortion metric ( $dSSIM_v$ ) and bit allocation to improve SSIM is given by Eq. 8.

$$\begin{aligned} & \min_{(R_t, R_d)} dSSIM_v \\ & \text{s.t. } R_t + R_d \leq R_c \end{aligned} \quad (8)$$

In terms of  $dSSIM$ , a planar model for synthesis distortion is determined as in Eq. 9.

$$dSSIM_v = a \cdot dSSIM_t + b \cdot dSSIM_d + c \quad (9)$$

Using SSIM-MSE relation, the distortion model in Eq. 9 is converted into Eq. 10.

$$dSSIM_v = \frac{a}{2\sigma^2_{x_t} + C_2} D_t + \frac{b}{2\sigma^2_{x_d} + C_2} D_d + z \quad (10)$$

$$dSSIM_v = p_1 D_t + p_2 D_d + c \quad (11)$$

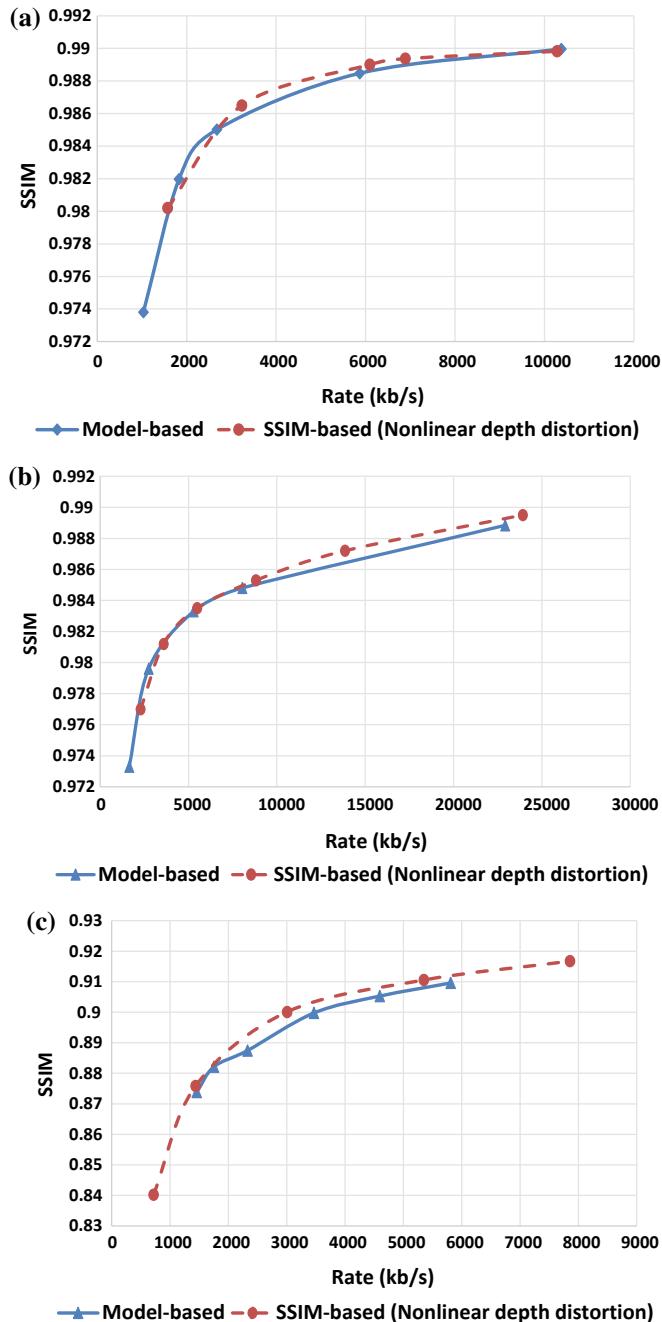
where  $p_1 = \frac{a}{2\sigma^2_{x_t} + C_2}$  and  $p_2 = \frac{b}{2\sigma^2_{x_d} + C_2}$ .  $Q_t$  (Eq. 13a) and  $Q_d$  (Eq. 13b), quantization steps of texture video and depth map determined by minimizing distortion-quantization model (Eq. 12).

$$\begin{aligned} & \min \quad (p_1 D_t + p_2 D_d) \\ & \text{s.t. } (a_t Q_t^{-1} + b_t + a_d Q_d^{-1} + b_d) \leq R_c \end{aligned} \quad (12)$$

$$Q_t = \frac{a_t + \sqrt{\frac{K_1 a_t a_d}{K_2}}}{R_c - b_t - b_d} \quad (13a)$$

$$Q_d = \sqrt{\frac{K_2 a_d}{K_1 a_t}} Q_t \quad (13b)$$

where  $K_1 = p_1 \alpha_t$  and  $K_2 = p_2 \alpha_d$ .



**Fig. 3** SSIM-based bit allocation with nonlinear depth distortion **a** Kendo sequence, **b** Balloons sequence, and **c** Breakdancer sequence

**Table 1** BD-rate comparison

Sequence	Proposed VS model-based bit allocation [14]		Proposed VS SSIM-based bit allocation [6]	
	$\Delta$ SSIM	$\Delta$ Rate	$\Delta$ SSIM	$\Delta$ Rate
Kendo	0.0003	-6.9196	0.0002	-3.8245
Balloons	0.0002	-2.8413	0.00004	-1.0327
Breakdancer	0.0034	-12.6183	0.0019	-7.1566

## 6 Results

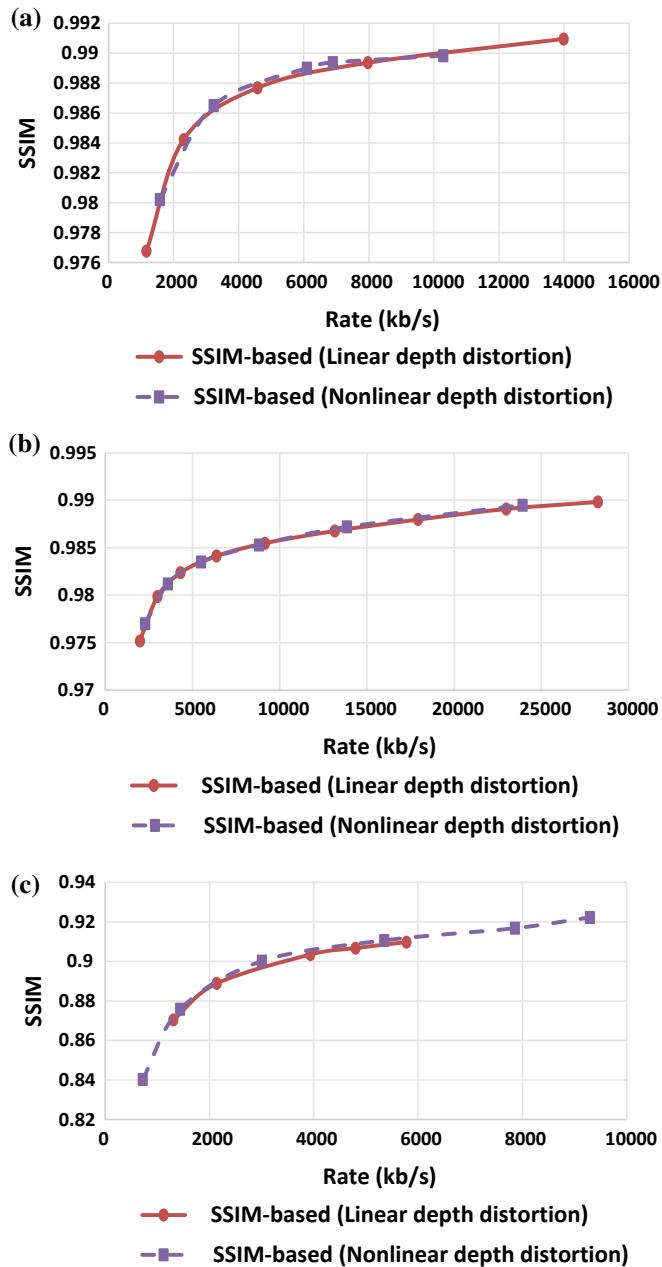
We conducted experiments to check the performance of the joint bit allocation algorithm with nonlinear depth distortion. Encoding is done using 3DV-ATM reference software [1] and VSRS 3.0 [11] reference software is used for virtual synthesis. The test sequences used are Kendo, Balloons [4], and Breakdancer [17] with a frame rate of 30 frames/s. Kendo and Breakdancer sequences have 100 frames whereas Balloons sequence has 300 frames.

Experiments were conducted to evaluate SSIM-based bit allocation where nonlinear depth distortion is implemented in RDO. SSIM is computed between synthesized views and original views. For comparison, we utilized model-based algorithm of Yuan et al. [14] and Harshalatha and Biswas's algorithm [6]. RD curves in Fig. 3 give a comparison between our proposed algorithm and bit allocation with model parameters. Bjontegaard distortion-rate (BD-rate) calculations are done and tabulated in Table 1.

Further, bit allocation algorithm with linear and nonlinear effect of depth distortion RDO is compared as in Fig. 4 and also using BD-rate (Table 1). Nonlinear effect of depth distortion considered in RDO gives more accurate bit allocation results.

## 7 Conclusions

Rate distortion optimization improves the efficiency of an encoder and we proposed depth map RDO for 3D video by considering nonlinear relation of depth map distortion with view synthesis distortion. To improve the visual quality of synthesized views, dSSIM is used as distortion metric. Bit allocation algorithm is verified by using nonlinear depth distortion RDO and gives better performance over linear depth distortion RDO.



**Fig. 4** SSIM-based bit allocation with nonlinear depth distortion compared with linear distortion  
**a** Kendo sequence, **b** Balloons sequence, and **c** Breakdancer sequence

## References

1. 3DV-ATM Reference Software 3DV-ATMv5.lrv2. <http://mpeg3dv.nokiaresearch.com/svn/mpeg3dv/tags/3DV-ATMv5.lrv2/>. Accessed 29 July 2018
2. Fehn, C.: A 3D-TV approach using depth-image-based rendering (DIBR). In: Proceedings of VIIP, vol. 3 (2003)
3. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: Electronic Imaging 2004, pp. 93–104. International Society for Optics and Photonics (2004)
4. Fujii Laboratory, Nagoya University. <http://www.fujii.nuee.nagoya-u.ac.jp/multiview-data/>. Accessed 29 July 2018
5. Harshalatha, Y., Biswas, P.K.: Rate distortion optimization using SSIM for 3D video coding. In: 23rd International Conference on Pattern Recognition (ICPR), pp. 1261–1266. IEEE (2016)
6. Harshalatha, Y., Biswas, P.K.: SSIM-based joint-bit allocation for 3D video coding. Int. J. Multimedia Tools Appl. **77**(15), 19051–19069 (2018). <https://doi.org/10.1007/s11042-017-5327-0>
7. Kim, W.S., Ortega, A., Lai, P., Tian, D.: Depth Map Coding Optimization Using Rendered View Distortion for 3-D Video Coding (2015)
8. Shao, F., Jiang, G.Y., Yu, M., Li, F.C.: View synthesis distortion model optimization for bit allocation in three-dimensional video coding. Opt. Eng. **50**(12), 120502–120502 (2011)
9. Shao, F., Jiang, G., Lin, W., Yu, M., Dai, Q.: Joint bit allocation and rate control for coding multi-view video plus depth based 3D video. IEEE Trans. Multimed. **15**(8), 1843–1854 (2013)
10. Tian, D., Lai, P.L., Lopez, P., Gomila, C.: View synthesis techniques for 3D video. In: Applications of Digital Image Processing XXXII, Proceedings of the SPIE 7443, 74430T–74430T (2009)
11. View Synthesis Reference Software VSRS3.5. <ftp://ftp.merl.com/pub/avetro/3dv-cfp/software/>. Accessed 20 May 2018
12. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
13. Yang, C., An, P., Shen, L.: Adaptive bit allocation for 3D video coding. Circuits, Systems, and Signal Processing, pp. 1–23 (2016)
14. Yuan, H., Chang, Y., Huo, J., Yang, F., Lu, Z.: Model-based joint bit allocation between texture videos and depth maps for 3-D video coding. IEEE Trans. Circuits Syst. Video Technol. **21**(4), 485–497 (2011)
15. Yuan, H., Chang, Y., Li, M., Yang, F.: Model based bit allocation between texture images and depth maps. In: International Conference On Computer and Communication Technologies in Agriculture Engineering (CCTAE), vol. 3, pp. 380–383. IEEE (2010)
16. Zhu, G., Jiang, G., Yu, M., Li, F., Shao, F., Peng, Z.: Joint video/depth bit allocation for 3D video coding based on distortion of synthesized view. In: IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–6. IEEE (2012)
17. Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High-quality video view interpolation using a layered representation. In: ACM Transactions on Graphics (TOG), vol. 23, pp. 600–608. ACM (2004)

# A Modified FCM-Based Brain Lesion Segmentation Scheme for Medical Images



Anjali Gautam, Debanjan Sadhya and Balasubramanian Raman

**Abstract** Segmentation of brain lesion from medical images is a critical problem in the present day. In this work, we have proposed a new distance metric for fuzzy clustering based classification of different brain regions via acquiring accurate lesion structures. The modified distance metric segments the images into different regions by calculating the distances between the cluster centers and object elements, and subsequently classify them via fuzzy clustering. The proposed method can effectively remove noise from the images, which results in a better homogeneous classification of the image. Our method can also accurately segment stroke lesion where the results are near to the ground truth of the stroke lesion. The performance of our method is evaluated on both magnetic resonance images (MRI) and computed tomography (CT) images of brain. The obtained results indicate that our method performs better than the standard fuzzy  $c$ -means (FCM), spatial FCM (SFCM), kernelized FCM methods (KFCM), and adaptively regularized kernel-based FCM (ARKFCM) schemes.

**Keywords** Brain stroke · Hemorrhage · Segmentation · Fuzzy  $c$ -means (FCM) · Metric

## 1 Introduction

Hemorrhagic stroke is a condition when blood vessels inside the brain burst due to different medical disorders including hypertension and amyloidosis [1]. It damages the brain within a very short interval and can even cause the death of a patient in the worst case. This medical condition accounts for about 13% of all stroke cases

---

A. Gautam (✉) · D. Sadhya · B. Raman  
Indian Institute of Technology Roorkee, Roorkee, Uttarakhand 247667, India  
e-mail: [agautam@cs.iitr.ac.in](mailto:agautam@cs.iitr.ac.in); [anjali31gautam@gmail.com](mailto:anjali31gautam@gmail.com)

D. Sadhya  
e-mail: [debanjan.sadhy@gmail.com](mailto:debanjan.sadhy@gmail.com)

B. Raman  
e-mail: [balarfcs@iitr.ac.in](mailto:balarfcs@iitr.ac.in)

and 30% of all cerebrovascular diseases. In 2013, approximately 3.4 million people suffered from hemorrhagic stroke. Hence, imaging methods like CT and MRI have been developed for proper medical treatment of hemorrhagic strokes [2]. Kidwell et al. [3] discovered that in order to detect chronic intracerebral hemorrhage, especially microbleeds, MRI may be more precise than CT. However, for initial diagnosis, physicians prefer CT over MRI because its imaging takes less time, is more cost-effective, and is able to detect the presence of hemorrhage. Hence, we have used CT scan images for the prognosis of hemorrhagic strokes in this work. The CT images have been used for the early identification of abnormalities by using appropriate computer-assisted diagnosis tools which have been previously used in many works [4–11].

Many researchers have focused in areas of stroke lesion detection using semiautomatic and automatic methods. Phillips et al. [12] performed radiological examination of medical images followed by a fuzzy clustering based segmentation. Some morphological operations and region growing methods were also used in many schemes where expert system-based segmentation has been used [5]. Loncaric et al. [4] used spatially weighted  $k$ -means histogram based clustering method by using morphological operations for segmentation. Active contour models were used in [6] and [11] for segmenting acute head trauma and intracranial hemorrhage. Bardera et al. [10] quantified hematoma and edema region from the brain using level set methods. However, the segmentation results obtained by the previous methods were not close to the manual results as required in medical practice. Due to these problems, automatic segmentation is considered as a difficult approach in medical terminology (for which physicians prefer manual segmentation). Hence, delineated lesion should be close to the manual results for the development of an automatic segmentation method for hemorrhage detection. In this paper, a new fuzzy clustering based method has been developed where the Euclidean distance between cluster center and the data points is replaced by a new distance metric. The proposed metric is inspired from the notion of *Canberra distance* [13], which is basically a weighted version of the  $L_1$  (Manhattan) distance.

The rest of the article is organized as follows. Section 2 discusses the background of fuzzy  $c$ -means. In Sect. 3, our proposed methodology is presented to delineate the stroke lesion. In Sect. 4, experimental analysis of results is given. Finally, the paper is concluded in Sect. 5.

## 2 Background

### 2.1 Fuzzy C-Means Clustering

Bezdek [14] proposed the FCM clustering in order to enhance the segmentation results. In this technique, a degree of fuzziness is imparted to each data point depend-

ing on the similarity to every class. This scheme partitions the image into  $C$  clusters by using the objective function given by

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

such that  $\sum_{j=1}^C u_{ij} = 1, \forall j$  and  $0 < \sum_{i=1}^N u_{ij} < N, \forall i$  where  $u_{ij}$  represents the degree of fuzziness ranging in  $[0-1]$ ,  $m$  is the weighting exponent,  $x$  represents the data point in region  $R^n$ , and the total number of data points in  $R^n$  are represented by  $N$ .  $x_i$  is the data point and  $c_j$  is the cluster center. Finally,  $\gamma^2(x_i, c_j)$  represents the Euclidean distance between  $x_i$  and  $c_j$ .

## 2.2 Extensions of FCM

The FCM clustering technique has been widely used in several research works. Consequently, it has many extensions which reduce the noise level from the input images. Many researchers have modified the objective function given in Eq. 1 for increasing the robustness of their proposed method. Pham et al. [15] present the robust FCM algorithm (RFCM) by incorporating a penalty term in the objective function for restricting the associated membership functions. Ahmed et al. [16] brought a new variant of FCM by utilizing the spatial constraints (FCM\_S) where they added a regularization term to the objective function as the weighted average of Euclidean distance from cluster center to gray level of neighbor pixels. However, a major disadvantage of FCM\_S is the requirement to compute neighborhood labels in each step of iteration. Chen and Zhang [17] modified FCM\_S by introducing mean and median neighborhood action. The authors named these schemes as FCM\_S1 and FCM\_S2. They also present the kernelized FCM algorithm (KFCM) [18] by introducing a kernel distance metric and spatial penalty on the membership functions.

Chuang et al. [19] developed the spatial FCM (SFCM) technique which also incorporates the spatial constraint in the membership function  $u_{ij}$ . In this work, the objective function is represented as the sum of all  $u_{ij}$  from the neighborhood of every pixel. Another variant of the FCM algorithm was given by Wang et al. [20] which utilizes local and nonlocal spatial constraints. The fast generalized FCM (FGFCM) algorithm was proposed by Cai et al. [21] for measuring the local similarity. The authors fused spectral and spatial information to form a new mean-filtering image which enables better classification of image elements. Recently, Elazab et al. [22] proposed a new clustering method named adaptively regularized kernel-based FCM (ARKFCM). The authors have used a Gaussian radial basis kernel functions instead of the standard Euclidean distance in the objective function. Some other variants of FCM can be found in [23–25].

### 3 Proposed Methodology

This section presents a new variant of fuzzy  $c$ -means clustering by modifying the objective function given in Eq. 1. We replace the original Euclidean distance by a normalized form of the Canberra distance. The new objective function is represented by

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

where  $\vartheta^2(x_i, c_j)$  is given as

$$\vartheta^2(x_i, c_j) = \sum_{i=1}^N \frac{\|x_i - c_j\|^2}{\|x_i\|^2 + \|c_j\|^2} \quad (3)$$

The proposed metric  $\vartheta^2(x, c)$  is basically a squared Euclidean distance which is normalized by the sum of the two data points  $x$  and  $c$ . The formulation of this metric is inspired by the Canberra distance, which is a weighted version of Manhattan distance. Due to its intrinsic properties,  $\vartheta^2(x, c)$  is biased for measures around the origin and very sensitive for values close to zero. This property consequently facilitates in increasing the segmentation accuracy of CT and MRI medical images. It can be observed from Eq. 3 that  $\vartheta^2(x, c)$  follows the properties of nonnegativity ( $\vartheta^2(x, c) \geq 0$ ), identity of indiscernibles ( $\vartheta^2(x, c) = 0 \iff x = c$ ), and symmetry ( $\vartheta^2(x, c) = \vartheta^2(c, x)$ ).

The partition of a region  $R$  is brought out through the iterative improvement of the objective function defined through Eq. 2. The term  $u_{ij}$  will update in every iteration according to

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left[ \frac{\vartheta(x_i, c_j)}{\vartheta(x_i, c_k)} \right]^{m-1}} \quad (4)$$

The cluster center  $c_j$  will be subsequently updated using Eq. 5:

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (5)$$

The objective function will stop when  $\max_{ij} |u_{ij}^{(t+1)} - u_{ij}^{(t)}| < \epsilon$ , where  $\epsilon$  lies between 0 and 1. The iteration steps are represented by  $t$ ,  $m$  is chosen to be 2, and  $j$  depends on total number of clusters.

In this work, we have delineated hemorrhagic lesions from CT images for proper medical treatment of the patient suffering from a hemorrhagic stroke. The detailed description of the brain lesion segmentation process using our proposed method is given in Algorithm 1.

---

**Algorithm 1:** Hemorrhagic lesion segmentation using modified distance metric of fuzzy clustering algorithm

---

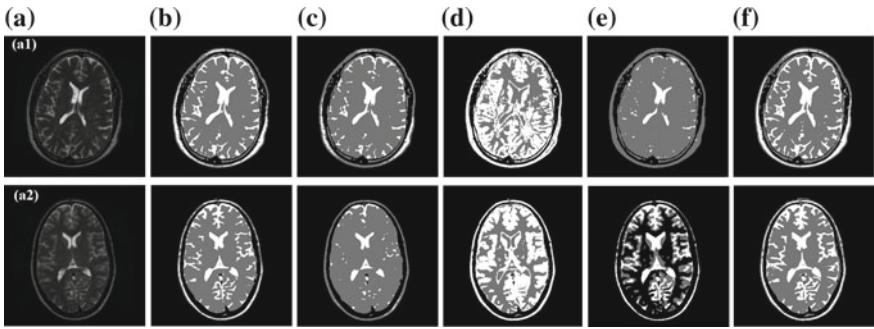
**Input :** Input image  $I$   
**Output:** Hemorrhagic lesion segmented image  
**1 Function**  $I = \text{Seg\_FCM\_D}(I, C)$

- 2 Randomly initialize the cluster centers  $c_j$  and fuzzy membership function  $u_{ij}$  with termination criterion  $\epsilon$  to  $\exp(-8)$
- 3 Update the  $u_{ij}$  using Eq. 4
- 4 Update new clustering centers using Eq. 5
- 5 Goto step 3 until  $\max_{ij} |u_{ij}^{(t+1)} - u_{ij}^{(t)}| < \epsilon$
- 6 Consider the cluster which likely to have a brain region
- 7 Reconstruct the image  $I$  classified into different clusters using  $u_{ij}$  (shown in Fig. 1f and Fig. 2f)
- 8 Construct  $j$  images of homogeneous regions from  $I$
- 9 Input the cluster no. containing the ROI as  $K_1$  and cluster no. containing background of the image as  $K_2$ . Subsequently, remove largest connected component from both the images
- 10  $K_2 = \text{complement}(K_2)$
- 11  $result = \text{subtract}(K_1, K_2)$
- 12 Apply some morphological operations on  $result$  in order to get the finally segmented image of hemorrhagic stroke.

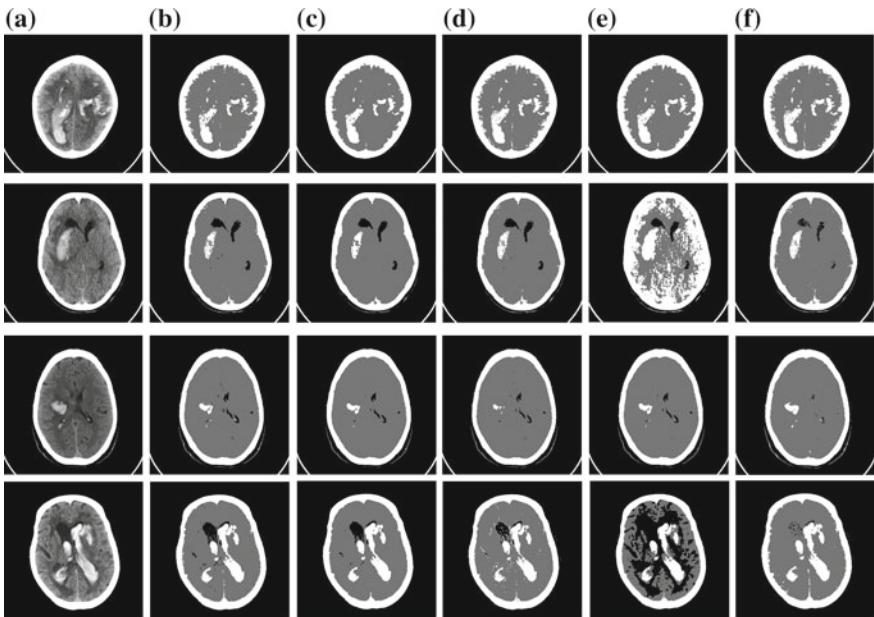
---

## 4 Experimental Results and Analysis

The performance of our method has been tested on MRI and CT scan images, wherein we have classified every input image into the black, gray, and white matter. The MRI dataset was downloaded from IXI dataset [26], whereas 35 CT scan images of hemorrhagic strokes were collected from the Himalayan Institute of Medical Sciences, Jollygrant, Dehradun, India. All the MRI images were in NIfTI format, whereas the CT images were in DICOM format. Noticeably, we have considered the T2-weighted MRI images for all our experiments. The image shown in Fig. 1a1 is of a male patient which has a label id 101\_Guys, whereas the image in Fig. 1a2 is of a female patient with label id 102\_HH.

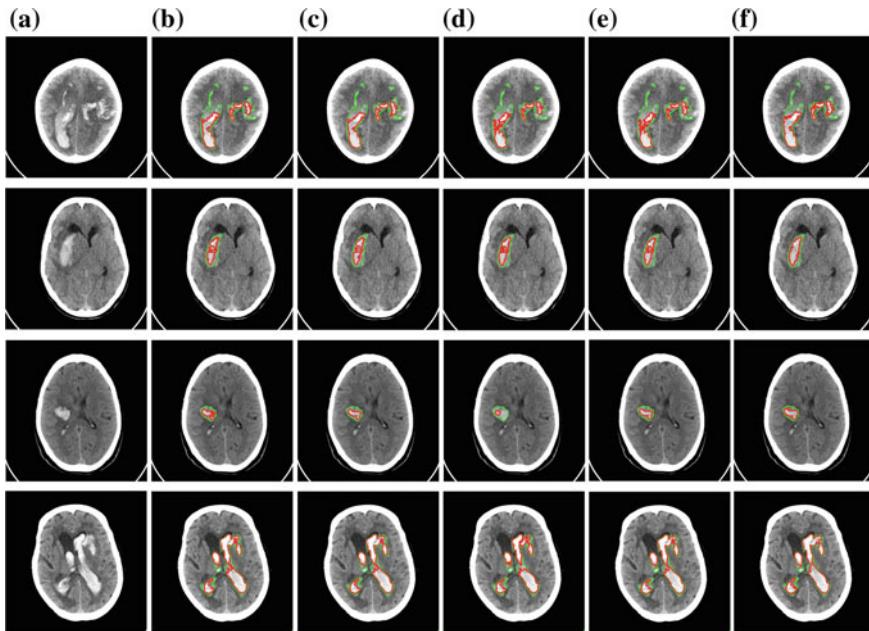


**Fig. 1** Clustering results of MRI images in **a** using **b** FCM, **c** SFCM, **d** KFCM, **e** ARKFCM, and **f** proposed method



**Fig. 2** Clustering results of CT images in **a** using **b** FCM, **c** SFCM, **d** KFCM, **e** ARKFCM, and **f** proposed method

The proposed technique can suppress most of the noise from input images, which subsequently results in better segmentation. The clustering and segmentation results using various algorithms are shown in Figs. 1, 2, and 3. The segmentation accuracy using proposed method is found to be much better than the existing ones [14, 18, 19]. Figures 2 and 3 portray the CT images of hemorrhagic stroke. In these images, we have segmented only the stroke lesion using Algorithm 1. The red and green color contours in Fig. 3 represent the segmentation results and ground truth, respectively.



**Fig. 3** Segmentation results of stroke lesions in **a** original image, using **b** FCM, **c** SFCM, **d** KFCM, **e** ARKFCM, and **f** proposed method

**Table 1** Hemorrhagic stroke’s segmentation results using standard evaluation metrics

Method	DC [0, 1]	JSI [0, 1]	HD (mm)	Precision [0, 1]	Recall [0, 1]
FCM	$0.68 \pm 0.19$	$0.54 \pm 0.19$	$7.11 \pm 14.30$	$0.97 \pm 0.11$	$0.55 \pm 0.19$
SFCM	$0.68 \pm 0.19$	$0.54 \pm 0.19$	$7.16 \pm 14.17$	$0.97 \pm 0.11$	$0.55 \pm 0.19$
KFCM	$0.65 \pm 0.21$	$0.51 \pm 0.21$	$7.96 \pm 14.26$	$0.97 \pm 0.11$	$0.52 \pm 0.21$
ARKFCM	$0.70 \pm 0.16$	$0.56 \pm 0.17$	$5.93 \pm 8.07$	$0.93 \pm 0.12$	$0.61 \pm 0.23$
Proposed	<b><math>0.75 \pm 0.15</math></b>	<b><math>0.62 \pm 0.17</math></b>	<b><math>5.91 \pm 11.91</math></b>	<b><math>0.97 \pm 0.08</math></b>	<b><math>0.63 \pm 0.17</math></b>

The segmentation results obtained by FCM, SFCM, KFCM, ARKFCM, and our proposed method are quantitatively compared with their ground truth by using the following performance measures—Dice coefficient (DC), Jaccard similarity index (JSI), precision, recall, and Hausdorff distance (HD). All the performance measures are indicated in Table 1.

For two images  $I_1$  (ground truth) and  $I_2$  (segmentation), these evaluation metrics can be calculated by using Eqs. 6–10 [27–29]. In these equations,  $d(\cdot)$  represents the Euclidean distance,  $FN$  is the false negative,  $FP$  is the false positive, and  $TP$  is the true positive. The Dice, Jaccard, precision, and recall measures indicate perfect segmentation if their value is near to 1. The Hausdorff distance is measured in mm, and it signifies good segmentation when its value is low.

$$DC = \frac{2 |I_1 \cap I_2|}{|I_1| + |I_2|} \quad (6)$$

$$JSI = \frac{|I_1 \cap I_2|}{|I_1| + |I_2| - |I_1 \cap I_2|} \quad (7)$$

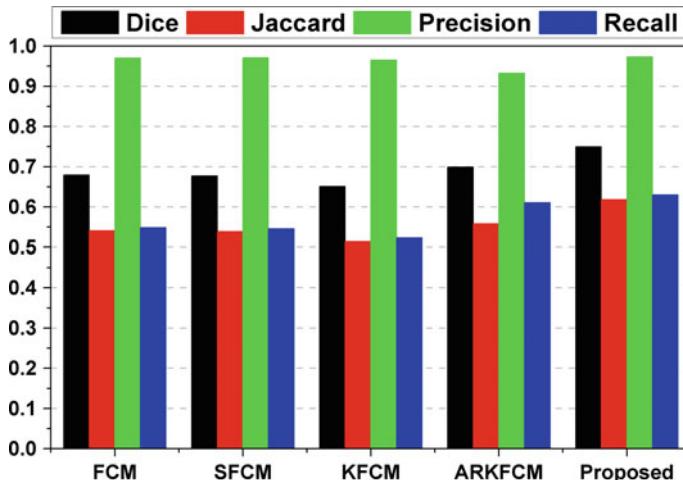
$$HD = \max \left\{ \frac{1}{N_a} \sum_{a \in I_1} \min_{b \in I_2} d(a, b), \frac{1}{N_b} \sum_{b \in I_2} \min_{a \in I_1} d(b, a) \right\} \quad (8)$$

where  $N_a$  and  $N_b$  represent number of pixels in  $I_1$  and  $I_2$ , respectively [30].

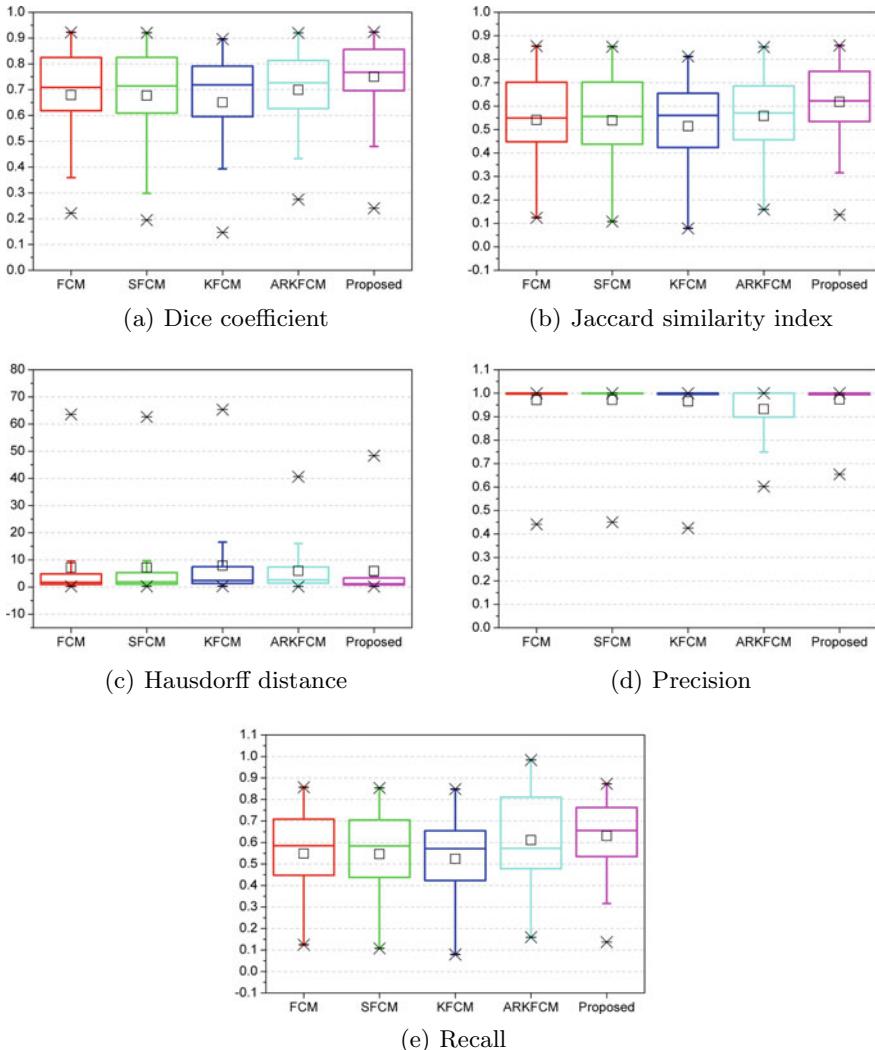
$$precision = \frac{TP}{TP + FP} \quad (9)$$

$$recall = \frac{TP}{TP + FN} \quad (10)$$

Table 1 shows that the proposed method performs better than FCM, SFCM, KFCM, and ARKFCM in terms of the evaluation metrics (average  $\pm$  standard deviation). The average and standard deviation of Dice score are 0.75 and 0.15, respectively; for Jaccard similarity, their values are 0.62 and 0.17. For Hausdorff distance, these values are 5.91 and 11.91. The precision and recall in terms of average values



**Fig. 4** Average segmentation accuracy comparison using different evaluation metrics



**Fig. 5** Boxplot comparison of different metrics for the proposed approach

calculated after segmentation of all images are 0.97 and 0.63 with standard deviations 0.08 and 0.17, respectively. These values are also visually depicted via a bar graph in Fig. 4. Finally, the boxplots for the individual metrics are depicted in Fig. 5.

## 5 Conclusion

Fuzzy  $c$ -means clustering is an exceptionally famous method which has been used in a variety of applications including segmentation of medical images. However, it suffers from noise-induced errors which substantially diminish the segmentation accuracy. In this paper, we have modified the standard FCM algorithm by replacing the Euclidean distance with a variant of the Canberra distance. The clustering approach based on our proposed metric allows pixels of the same category to form homogeneous regions while suppressing the noise level. We have implemented our proposed approach on CT scan images containing hemorrhagic stroke by segmenting the stroke lesions from the images and also on normal brain MR images. The comparison of segmentation results was done with previous methods. Subsequently, we found that our proposed approach performs better than the standard FCM and its variants like SFCM, KFCM, and ARKFCM in terms of the segmentation accuracy.

**Acknowledgements** We thank Institute Human Ethics Committee (IHEC) of Indian Institute of Technology Roorkee, India for allowing us to collect the CT image dataset of hemorrhagic stroke with its ground truth information from Himalayan Institute of Medical Sciences (HIMS), Dehradun, Uttarakhand, India. The consent to obtain CT scan images of patients has already been taken by the radiologists of HIMS. We also thank Dr. Shailendra Raghuwanshi, Head of Radiology Department, HIMS for providing us his useful suggestions.

## References

1. Kase, C.S.: Intracerebral hemorrhage: non-hypertensive causes. *Stroke* **17**(4), 590–595 (1986)
2. Stroke. <https://emedicine.medscape.com/article/338385-overview>. Last accessed 18 Jan 2017
3. Kidwell, C.S., Chalela, J.A., Saver, J.L., Starkman, S., Hill, M.D., Demchuk, A.M., Butman, J.A., Patronas, N., Alger, J.R., Latour, L.L., Luby, M.L.: Comparison of MRI and CT for detection of acute intracerebral hemorrhage. *JAMA* **292**(15), 1823–1830 (2004)
4. Loncaric, S., Dhawan, A.P., Broderick, J., Brott, T.: 3-D image analysis of intra-cerebral brain hemorrhage from digitized CT films. *Comput. Methods Programs Biomed.* **46**(3), 207–216 (1995)
5. Cosic, D., Loucaric, S.: Computer system for quantitative: analysis of ICH from CT head images. In: Proceedings of the 19th Annual International Conference Engineering in Medicine and Biology Society, vol. 2, pp. 553–556. IEEE (1997)
6. Maksimovic, R., Stankovic, S., Milovanovic, D.: Computed tomography image analyzer: 3D reconstruction and segmentation applying active contour models ‘snakes’. *Int. J. Med. Inform.* **58**, 29–37 (2000)
7. Hu, Q., Qian, G., Aziz, A., Nowinski, W.L.: Segmentation of brain from computed tomography head images. In: Proceedings of the 27th Annual International Conference Engineering in Medicine and Biology Society, pp. 3375–3378. IEEE (2005)
8. Yoon, D.Y., Choi, C.S., Kim, K.H., Cho, B.M.: Multidetector-row CT angiography of cerebral vasospasm after aneurysmal subarachnoid hemorrhage: comparison of volume-rendered images and digital subtraction angiography. *Am. J. Neuroradiol.* **27**(2), 370–377 (2006)
9. Chan, T.: Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain. *Comput. Med. Imaging Graph.* **31**(4–5), 285–298 (2007)

10. Bardera, A., Boada, I., Feixas, M., Remollo, S., Blasco, G., Silva, Y., Pedraza, S.: Semi-automated method for brain hematoma and edema quantification using computed tomography. *Comput. Med. Imaging Graph.* **33**(4), 304–311 (2009)
11. Bhaduria, H.S., Singh, A., Dewal, M.L.: An integrated method for hemorrhage segmentation from brain CT imaging. *Comput. Electr. Eng.* **39**(5), 1527–1536 (2013)
12. Phillips II, W.E., Velthuizen, R.P., Phuphanich, S., Hall, L.O., Clarke, L.P., Silbiger, M.L.: Application of fuzzy c-means segmentation technique for tissue differentiation in MR images of a hemorrhagic glioblastoma multiforme. *Magn. Reson. Imaging* **13**(2), 277–290 (1995)
13. Lance, G.N., Williams, W.T.: Mixed-data classificatory programs I—agglomerative systems. *Aust. Comput. J.* **1**(1), 15–20 (1967)
14. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10**(2–3), 191–203 (1984)
15. Pham, D.L.: Spatial models for fuzzy clustering. *Comput. Vis. Image Underst.* **84**(2), 285–297 (2001)
16. Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A., Moriarty, T.: A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. *IEEE Trans. Med. Imaging* **21**(3), 193–199 (2002)
17. Chen, S., Zhang, D.: Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure. *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)* **34**(4), 1907–1916 (2004)
18. Zhang, D.Q., Chen, S.C.: A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artif. Intell. Med.* **32**(1), 37–50 (2004)
19. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c-means clustering with spatial information for image segmentation. *Comput. Med. Imaging Graph.* **30**(1), 9–15 (2006)
20. Wang, J., Kong, J., Lu, Y., Qi, M., Zhang, B.: A modified FCM algorithm for MRI brain image segmentation using both local and non-local spatial constraints. *Comput. Med. Imaging Graph.* **32**(8), 685–698 (2008)
21. Cai, W., Chen, S., Zhang, D.: Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recognit.* **40**(3), 825–838 (2007)
22. Elazab, A., Wang, C., Jia, F., Wu, J., Li, G., Hu, Q.: Segmentation of brain tissues from magnetic resonance images using adaptively regularized kernel-based fuzzy-means clustering. *Comput. Math. Methods Med.* (2015)
23. Kannan, S.R., Devi, R., Ramathilagam, S., Hong, T.P., Ravikumar, A.: Effective kernel FCM: finding appropriate structure in cancer database. *Int. J. Biomath.* **9**(02), 1650018 (2016)
24. Kannan, S.R., Devi, R., Ramathilagam, S., Hong, T.P.: Effective fuzzy possibilistic c-means: an analyzing cancer medical database. *Soft Comput* **21**(11), 2835–2845 (2017)
25. Farahani, F.V., Ahmadi, A., Zarandi, M.H.F.: Hybrid intelligent approach for diagnosis of the lung nodule from CT images using spatial kernelized fuzzy c-means and ensemble learning. *Math. Comput. Simul.* **149**, 48–68 (2018)
26. IXI Dataset. <https://www.nitrc.org/ir/data/projects/ixi>, last accessed 2017/7/8
27. Kang, S.J.: Multi-user identification-based eye-tracking algorithm using position estimation. *Sensors* **17**(1), 41 (2016)
28. Zhuang, A.H., Valentino, D.J., Toga, A.W.: Skull-stripping magnetic resonance brain images using a model-based level set. *NeuroImage* **32**(1), 79–92 (2006)
29. Maier, O., Schröder, C., Forkert, N.D., Martinetz, T., Handels, H.: Classifiers for ischemic stroke lesion segmentation: a comparison study. *PloS ONE* **10**(12), e0145118 (2015)
30. Dubuisson, M.P., Jain, A.K.: A modified Hausdorff distance for object matching. In: *Proceedings of the 12th International Conference on Pattern Recognition*, pp. 566–568. IEEE (1994)

# Word Spotting in Cluttered Environment



Divya Srivastava and Gaurav Harit

**Abstract** In this paper, we present a novel problem of handwritten word spotting in cluttered environment where a word is cluttered by a strike-through with a line stroke. These line strokes can be straight, slant, broken, continuous, or wavy in nature. Vertical Projection Profile (VPP) feature and its modified version, which is the combinatorics Vertical Projection Profile (cVPP) feature is extracted and aligned by modified Dynamic Time Warping (DTW) algorithm. The dataset for the proposed problem is not available so we prepared our dataset. We compare our method with Rath and Manmath [6], and PHOCNET [17] for handwritten word spotting in the presence of strike-through, and achieve better results.

**Keywords** Word spotting · Dynamic time warping · Vertical projection profile · Combinatorics vertical projection profile

## 1 Introduction

Digitization of a large number of documents raises the issue of indexing and retrieval in an efficient manner. For this reason, research has been emphasized on word spotting which refers to localization of word images of interest in the dataset without actually interpreting its content [3]. Some of the issues with handwritten word spotting in documents' images are handwriting variability, ink bleed, language variability, degradation caused by aging, strains, repetitive use, etc. We are here considering a specific case where a handwritten word has been distorted accidentally or intentionally by a strike-through over it. The line used for striking can be straight, slant, broken, continuous, or wavy. Especially in some scripts, like Devanagari, Sanskrit, Punjabi, Bengali, etc., the shirorekha can mistakenly be passed through the word

---

D. Srivastava (✉) · G. Harit  
Indian Institute of Technology Jodhpur, Jodhpur, India  
e-mail: [srivastava.5@iitj.ac.in](mailto:srivastava.5@iitj.ac.in)

G. Harit  
e-mail: [gharit@iitj.ac.in](mailto:gharit@iitj.ac.in)

leading to a clutter. In structured document images like form, cheques, etc., handwritten words are often overlapping with cell boundaries. To resolve this issue, we propose a word spotting technique that works in a cluttered medium where a word is cluttered by a strike-through over it. Vertical Projection Profile (VPP) feature is considered as a feature vector for uncluttered query word and a new feature called combinatorics Vertical Projection Profile (cVPP) feature is used for the cluttered candidate words. CVPP selects or discards the various combinatorics of possible strokes that contribute to VPP along each of its column. The similarity measure between the words is computed using a Dynamic Time Warping (DTW) approach which has been extensively used in word spotting. To the best of our knowledge word spotting for such handwritten words cluttered by strike is a novel problem which has not been addressed so far.

In the following section, related work for word spotting and DTW is discussed. Section 3 explains the VPP and its combinatorial version cVPP used as a feature descriptor. Section 4 explores the working and various constraints associated with DTW in our method. Section 5 discusses the matching procedure for the proposed word descriptor. Results and discussion are presented in Sect. 6 followed by conclusion.

## 2 Related Work

### 2.1 Word Spotting

Analysis of offline handwritten documents faces various challenges, such as inter-and intrawriter variability, ink bleed, language variability, document degradation, etc. Word spotting in offline handwritten documents is a difficult task. The problem of touching and broken characters due to degradation is addressed in [12], where word spotting is done by decomposing text line into character primitives. A word image is encoded into a string of coarse and fine primitives chosen according to the codebook and then a string matching approach based on dynamic programming is used to retrieve a similar feature sequence in a text line from the collection.

HMM-based methods have been used extensively for this purpose. Lexicon-free keyword spotting using the filler or garbage Hidden Markov Models (HMM) has the issue of high computational cost of the keyword-specific HMM Viterbi decoding process required to obtain confidence score of each word. To deal with this issue, [9] has proposed a novel way to compute this confidence score, directly from character lattice which is produced during a single Viterbi decoding process using only the “filler” model, thus saving the time-consuming keyword-specific decoding step. Later they extended their work in [4] by using context-aware character lattices obtained by Viterbi decoding with high-order character N-gram models. In another approach by [13], lexicon-free keyword spotting system is presented which employs trained character HMMs. The statistical framework for word spotting is proposed in

[8] where HMMs model keywords and Gaussian Mixture Model (GMM) do score normalization.

Various approaches represent the handwritten word in graphical representation by their different notions and employ the various graph matching algorithms for spotting purpose. In [2], a handwritten word is represented as a graph such that graphemes are extracted from shape convexity that are used as stable units of handwriting. These graphemes are associated with graph nodes and spatial relationships between them are considered as graph edges. Spotting is done using bipartite graph matching as error-tolerant graph matching. In another approach using graph-based method [5], invariants which are the collection of writing pieces automatically extracted from the old document collection are used as a descriptor to characterize the word and the graph edit distance is used as dissimilarity between the word images.

Various fusion frameworks where more than one methods are used to solve word spotting with higher accuracy have been proposed. To deal with the issue of large variation due to unconstrained writing style, a method based on Heat Kernel Signature (HKS) and triangular mesh structure is proposed in [20] for keyword spotting. HKS captures local features while triangular mesh structure is used to represent global characteristics. Notably, this method does not require preprocessing. In [10], both topological and morphological features are employed for word spotting. Skeleton-based graphs with shape context labeled vertices are constructed for connected components and each word is represented as a sequence of graphs. For fast retrieval, region of interest is found using graph embedding. For refined result graph, edit distance based on DTW approach is used.

In [5], word spotting problem is addressed using a shape-based matching scheme where segmented word images are represented by local contour features. A segmentation-free approach is given in [18] and SIFT features are extracted and encoded into visual words which are accumulated in a set of local patches. These features are projected to a topic space using latent semantic analysis. Further compressing the feature with product quantization method leads to efficient indexing of document information both in terms of memory and time. Another segmentation-free approach is proposed in [19] where again the spotting is done in two steps of selecting candidate zone followed by refining the results. Both the steps are based on the process of accumulation of votes obtained by application of generalized Haar-like features. Deep learning based approach for word spotting PHOCNET is proposed in [17]. CNN is used to extract PHOC features for word images which are then compared using Bray–Curtis dissimilarity [16].

## 2.2 Dynamic Time Warping

To overcome the shortcomings of Euclidean metric in measuring similarity measure between sequences with variation in stretches and compression, DTW was introduced as similarity metric by Berndt [7]. DTW works on the principle of dynamic programming to find the minimal distance between two sequences which are warped

by stretching or shrinking in their subsection. DTW similarity measure finds its application whenever the feature vector is in a sequential format. It has been therefore widely used in speech processing, bioinformatics, online handwriting recognition, etc. DTW is used in offline handwriting recognition where the sequential features like projection profile are considered as a feature vector. In [14], DTW was used for handwritten word spotting for robustness against intrawriter variability. DTW was used for word spotting in noisy historical documents in [11] using projection profile, word profiles, and background/ink transitions as features. By comparing DTW with six matching techniques, it was found that DTW performs better. DTW has quadratic complexity and therefore is computationally expensive for large-scale images. Its faster version was proposed in [21] which provides 40 times speedup. DTW-based statistical framework is proposed in [8] for a challenging multi-writer corpus.

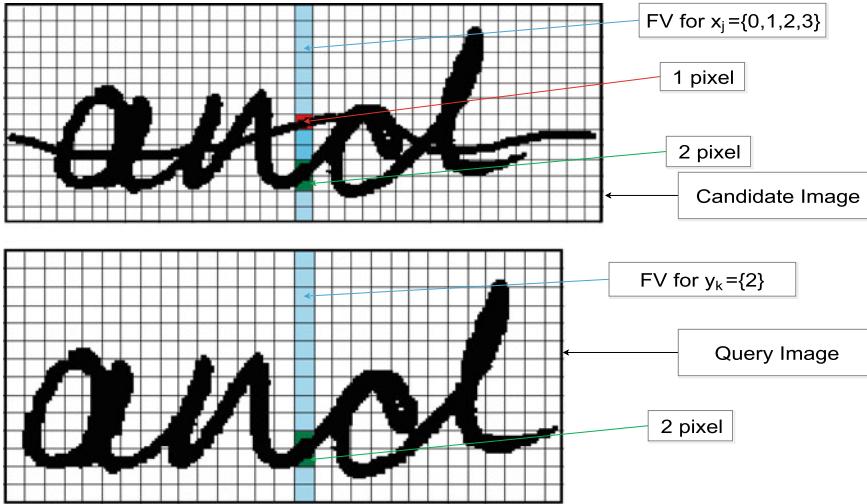
### 3 Features

Word spotting using DTW approach is carried out using sequential features and one among them is projection profile feature. Projection profile corresponds to the histogram of the number of black pixels accumulated along the horizontal and vertical axis of the image. Depending upon the axis considered they are termed as Horizontal Projection Profile (HPP) or Vertical Projection Profile (VPP) for the accumulation of black pixels along horizontal and vertical axes, respectively.

The proposed feature is a slight variation of VPP feature and uses combinatorics in VPP. Candidate word in our method is a set of cluttered and uncluttered word present in our dataset. As shown in Fig. 1, the cluttered candidate word has a strike-through over it and the query word is an uncluttered word without any strike-through. VPP in the query word is computed by counting all the black pixels across each column in a black and white image with foreground black pixels and background white pixels. Several strokes can intersect with a column of pixels (i.e., the vertical axis). Here, a stroke refers to a run of black pixels along a column in an image. For a candidate word cluttered by a single strike-through line, we have an extra stroke which corresponds to the pixels along the column contributed by the strike-through. We consider summation of black pixels for various combinations of strokes which accounts for the profile feature of the candidate image.

Consider an image  $I$  with  $M$  rows and  $N$  columns, and let  $I(x, y)$  denote a pixel in the image such that  $1 \leq x \leq M$  and  $1 \leq y \leq N$ . VPP value for the uncluttered query word at  $y$ th column can be computed as shown in Eq. 1, where  $y$  indexes the column and  $i$  indexes the row.

$$V P P(y) = \sum_{i=1}^M I(i, y) \quad (1)$$



**Fig. 1** Computation of feature vector for candidate and query image

A candidate word image can be cluttered with strike-through and therefore we compute the combinatorics VPP (cVPP) feature. Consider  $n$  as the total number of strokes in a column  $y$  and let  $i$  be the row index. cVPP is a vector (indexed by  $k$ ) computed for every column as follows:

$$cVPP(k, y) = \sum_{s_m \in S_k} \sum_{i=s_m^{x_1}}^{s_m^{x_2}} I(i, y) \quad (2)$$

Here,  $S_k$  is one of the subsets of the set of strokes that cuts across column  $y$ .  $s_m$  is a member of the subset  $S_k$ . The  $x$ -limits (coordinates of the vertical black run) of the stroke  $s_m$  are  $s_m^{x_1}$  and  $s_m^{x_2}$ . Index  $k$  varies over all possible subsets of strokes, including the null set. cVPP leads to a feature vector where for each column of strokes we have multiple entries depicting black pixel counts for various possible combinations of strokes. In further section by using modified DTW, selection will be done for each column of cVPP to find the best combination of strokes which matches with the query image.

## 4 Feature Matching Using DTW

Let  $X$  denote the feature sequence (VPP) for the query image and  $Y$  denote the feature sequence (cVPP) for the candidate image. The feature sequences  $X$  and  $Y$  are defined as  $X = x_1, x_2, \dots, x_{M_1}$  of length  $M_1$  and  $Y = y_1, y_2, \dots, y_{M_2}$  of length  $M_2$ . As described

in previous section, each entry in the cVPP feature vector denotes a vector whose length depends on the number of strokes in the corresponding column of the image. The alignment from X to Y is described in terms of a warp path  $W = w_1, w_2, \dots, w_K$ , where K is the warp path length whose range is  $\max(M_1, M_2) \leq k \leq M_1 + M_2$ , where  $w_k = (i_k, j_k)$  indicates the indices of feature sequences X and Y at  $k$ th instance of the matching. To find this alignment in terms of warp path consider an  $(M + 1) \times (N + 1)$  cost matrix for the sequences X and Y. While doing the optimal alignment, it is assumed that the starting and ending indices of the two sequences align with each other, and hence we need to compute the path from lower left to top right corner of the cost matrix. Computing all possible paths between the two extremes of the cost matrix and finding out the minimum cost path among them is computationally very expensive. To overcome this, DTW distance measure which is based on dynamic programming makes use of recursive nature of distance function. Hence, the entries of the cost matrix need to be filled by distance function which is essentially based on dynamic programming approach.

The mathematical expression for DTW distance between the feature vector X and Y is given in Eq. 3:

$$D^{l,m}(X_i, Y_j) = d^{l,m}(X_i, Y_j) + \min_{\forall a,b} \begin{cases} D^{a,b}(X_{i-1}, Y_j) \\ D^{a,b}(X_i, Y_{j-1}) \\ D^{a,b}(X_{i-1}, Y_{j-1}) \end{cases} \quad (3)$$

where  $d^{l,m}(X_i, Y_j)$  is the local cost of a cell at index  $(i, j)$  by selecting the  $l$ th and  $m$ th combination for candidate and query image, respectively. Local cost is responsible for the selection of stroke combination in candidate feature vector such that it matches closely to the corresponding candidate feature vector. For this purpose, Manhattan distance is computed between VPP for a column in query image and various cVPP for a column in candidate image. Best matched stroke is selected and cVPP leading to minimum distance is selected and considered as the local distance for corresponding entry of the cost matrix for DTW algorithm.

Each entry in a cell of the cost matrix is computed by addition of local cost and minimum cost among the previously traversed adjacent cell.  $D^{a,b}(X_i, Y_j)$  signifies the total distance for a cell at index  $(i, j)$  of the cost matrix by selecting the  $a$ th and  $b$ th combination for candidate and query image, respectively.  $D^{l,m}(X_i, Y_j)$  has the same interpretation as  $D^{a,b}(X_i, Y_j)$ , difference lies in that  $l$  and  $m$  stands for the combination referring to the local cost of current cell while  $a$  and  $b$  stand for the combination referring to the total cost of the pre-traversed adjacent cell.

The total cost computed by the DTW approach is the summation of the cost incurred during the whole warp path which needs to be normalized to make it invariant to the variable length sequences. Therefore, the DTW cost corresponding to the optimal warp path is given in Eq. 4, where  $K$  is a normalizing factor equal to the optimal warp path length and  $\sum_{k=1}^K D(X_k, Y_k)$  corresponds to the summation of all the entries in the DTW cost matrix along the optimal warp path.

$$D(X, Y) = \frac{1}{K} \sum_{k=1}^K D(X_k, Y_k) \quad (4)$$

The optimal warp path is identified by backtracking the cost matrix  $D$  along the minimum cost index pairs  $(i_k, j_k)$  starting from  $(M + 1, N + 1)$  to  $(0, 0)$  in  $O(MN)$  time.

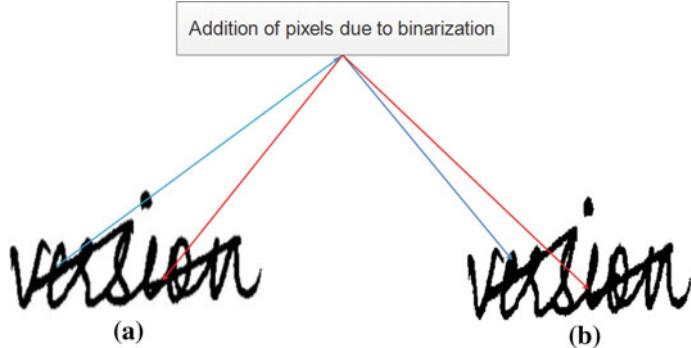
Some constraints followed by the warp path [15] are mentioned below:

- Boundary Constraint: The starting point  $(0, 0)$  and the ending point  $(M + 1, N + 1)$  of the warping path are fixed, i.e.,  $w_1 = (0, 0)$  and  $w_k = (M + 1, N + 1)$ .
- Continuity Constraint: It ensures that the path is continuous and no break is permitted, i.e.,  $i_k - i_{k-1} = 1$ .
- Monotonicity Constraint: The sequence of characters in a word is constant. It requires the characters in a word to be clustered in monotonically increasing order, which means  $j_k - j_{k-1} = 1$ . Summarizing the continuity and monotonicity constraint, we have  $w_k - w_{k-1} \in (1, 0), (0, 1), (1, 1)$ .
- Global Path Constraint: It is used to restrict the extent of expansion or compression of the features extracted from the word image. The spatial variability is considered to be limited which means that search space can be pruned. To ensure global path constraint, we have considered Sakoe–Chiba Band of window width 20.

## 5 Methodology

We are considering the segmented word as a grayscale image. The word image is binarized using the Otsu's thresholding method [1] which yields two classes of intensity values to have minimal intraclass variance and maximal interclass variance. VPP feature is computed for the query image and cVPP is computed for the candidate images as described in the previous section.

Our aim is to devise a matching procedure which is more likely to account for the correct set of strokes and is able to ignore the strokes arising due to strike-through. For this purpose, as illustrated in Fig. 1, the cVPP is computed for the candidate image, where for each column VPP is computed for all the possible combinations of strokes present in that column. For each column, the objective is to find the combination of strokes in the candidate image which gives a projection value closest to that for the corresponding column in the query image. Applying Brute Force method for this combination and selection problem is computationally costly; therefore, we apply modified DTW algorithm as explained in Sect. 4. DTW cost is normalized two times. First it is normalized by the path length to account for the word length and then it is normalized by the cluttered word length after adding penalty to the DTW cost to make it invariant to the occurrence of blob induced due to clutter.



**Fig. 2** **a** Grayscale cluttered image without blob. **b** Presence of blob in binarized image

### 5.1 Computing the Penalty Term

Whenever a word is struck-through, the extra stroke of the line renders some extra pixels termed as blob at the intersecting region in a word due to binarization as shown in Fig. 2. The strokes are horizontally oriented; therefore, a word containing more number of characters will have more intersection bleeds leading to more number of blobs getting formed in comparison to a word containing less number of characters. This results in misalignment and some non-diagonal steps in the DTW path.

To resolve this issue, we use penalty term in DTW cost. The blobs produce large projection values in their respective columns. A penalty term is computed on the basis of possible number of occurrences of intersection points present in the candidate word. This helps to minimize the effect of blobs. We find the mean of VPP values and the count of VPP values exceeding their mean are computed as penalty term. The penalty term is subtracted from the DTW cost which is then again normalized by the optimal warp path length  $K$ .

Given  $K$  as the DTW path length and  $p$  as the penalty term, the aggregate match cost is given as follows:

$$D(X, Y) = \frac{1}{K} \left[ \sum_{k=1}^K D(X_k, Y_k) - p \right]. \quad (5)$$

## 6 Dataset

The performance of the proposed approach is evaluated in the retrieval step, where for each of the uncluttered query word, its best possible match among candidate words is found. To test our approach, we need a dataset, where some words are

**Fig. 3** Samples for query images from our dataset

Honey industrial version tomorrow demand  
Delaney churchyard every back recognition

**Fig. 4** Samples for candidate images from our dataset

elusively something Honey industrial recognition  
Delaney shortest churchyard back promiscuous

(a) Samples of Candidate images from our Dataset 1

Honey industrial version tomorrow gather  
Delaney every but demand become

(b) Samples of Candidate images from our Dataset 2

rewritten and cluttered by strike-through or line segments. We have developed our own dataset in which word images are written and then the same image is cluttered with the strike-through line. These are type 1 images. Words are also rewritten and cluttered with the line, i.e., strike-through on a different instance of the word. We refer to them as type 2 images.

We collected 50 query word images and 100 candidate word images of type 1 (referred to as dataset 1), and 100 candidate word images of type 2 (referred to as dataset 2). Candidate word images contain 50 cluttered and 50 uncluttered word images. The line drawn to induce the clutter is of free form, it can be straight, slant, waveform, broken, etc. In the dataset, we have considered clutter by a single line so that along with a column there is only one extra stroke. Sample images from our dataset are shown in Figs. 3 and 4.

## 7 Experiments

Experiments are conducted on our dataset, where for each word in the query image set, its best match in the dataset is retrieved based on the dissimilarity measure computed as the matching cost obtained from DTW algorithm. To verify our approach, results are compared to the word spotting results obtained from the DTW approach given by Rath et al. in [6]. Also, we compare our method with PHOCNET [17]. It is a pretrained model on George Washington dataset [2]. We use this model and test it on our images.

There are three sets of experiments conducted. First, DTW cost is computed using [6] and our approach on word images from dataset 1. Second, DTW cost is

computed using [6] and our approach on word images from dataset 2. In the third set of experiments, precision at rank 1 is computed for [6], PHOCNET [17], our approach for both the types of our dataset.

## 8 Results

Tables 1 and 2 show the matching cost computed using our approach and that developed by [6] for five example images from Dataset 1 and Dataset 2, respectively. It can be seen that the matching cost is smaller in our approach. In Table 3, the average precision at rank 1 is computed for our approach and two other approaches [6] and

**Table 1** Matching cost for five examples from Dataset 1

Query image	Candidate image	Proposed method	DTW approach [6]
clumsily	clumsily	<b>1.26</b>	6.07
surroundings	surroundings	<b>0.74</b>	6.46
something	something	<b>1.05</b>	8.91
advantages	advantages	<b>0.85</b>	4.88
out	out	<b>0.37</b>	1.48

**Table 2** Matching cost for five examples from Dataset 2

Query image	Candidate image	Proposed method	DTW approach [6]
honey	honey	<b>2.5</b>	12.04
industrial	industrial	<b>1.57</b>	9.82
tomorrow	tomorrow	<b>1.38</b>	11.13
gather	gather	<b>0.62</b>	13.46
solitude	solitude	<b>0.68</b>	12.87

**Table 3** Average precision at rank 1

Dataset	Proposed method	DTW approach [6]	PHOCNET [17]
Dataset 1	<b>0.92</b>	0.86	0.36
Dataset 2	<b>0.80</b>	0.76	0.24

[17]. The average precision at rank 1 computed for our approach is more than that computed for the approaches [6] and [17]. PHOCNET is an attribute-based approach where for each word, PHOC (Pyramidal Histogram of Characters) is computed. This network has been trained for clean images with minimal distortion to the structure of the word. In our approach, cluttering the word distorts its structure affecting the performance of PHOCNET.

## 9 Conclusion

In this paper, we have proposed a novel problem of handwritten word spotting in the cluttered environment where a word is cluttered by line/lines. These lines can be straight, slant, broken, continuous, or wavy in nature. VPP feature and its modified version cVPP feature are considered as feature vectors. cVPP considers the possible combinations of VPP in a cluttered word which lead to the selection of appropriate projection value for a column, and hence the rejection of cluttered pixels, thereby retrieving the matched word. DTW is used for aligning the two features and obtaining the cost of matching. A penalty term helps to overcome the problem of addition of extra pixels due to the intersection of a word with clutter. The dataset for the proposed problem is not present so we prepared our dataset. We have compared our method with the method proposed in [6] and in [17] and have achieved improved performance over both the methods. Also, the comparison of matching cost has been made between our method and [6]. Our method performs better and we have obtained smaller matching cost for struck-through words.

**Acknowledgements** The authors would like to thank the research scholars at IIT Jodhpur for their contribution toward dataset creation. All procedures performed in dataset creation involving research scholars at IIT Jodhpur were in accordance with the ethical standards of the institute.

## References

1. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop, vol. 10, pp. 359–370. Seattle, WA (1994)
2. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character hmms. Pattern Recognit. Lett. **33**(7), 934–942 (2012)

3. Ghorbel, A., Ogier, J.M., Vincent, N.: A segmentation free word spotting for handwritten documents. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 346–350. IEEE (2015)
4. Giotis, A.P., Gerogiannis, D.P., Nikou, C.: Word spotting in handwritten text using contour-based models. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 399–404. IEEE (2014)
5. Lavrenko, V., Rath, T.M., Manmatha, R.: Holistic word recognition for handwritten historical documents. In: Proceedings of First International Workshop on Document Image Analysis for Libraries, 2004. pp. 278–287. IEEE (2004)
6. Mammatha, R., Croft, W.: Word Spotting: Indexing Handwritten Manuscripts, Intelligent Multimedia Information Retrieval (1997)
7. Mammatha, R., Han, C., Riseman, E.M.: Word spotting: a new approach to indexing handwriting. In: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 631–637 (1996). <https://doi.org/10.1109/CVPR.1996.517139>
8. Nagendar, G., Jawahar, C.: Efficient word image retrieval using fast DTW distance. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 876–880. IEEE (2015)
9. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
10. Rath, T.M., Manmatha, R.: Word image matching using dynamic time warping. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003, vol. 2, pp. II–II. IEEE (2003)
11. Rath, T.M., Manmatha, R.: Word spotting for historical documents. *Int. J. Doc. Anal. Recognit. (IJDAR)* **9**(2–4), 139–152 (2007)
12. Rodríguez-Serrano, J.A., Perronnin, F.: Handwritten word-spotting using hidden markov models and universal vocabularies. *Pattern Recognit.* **42**(9), 2106–2116 (2009)
13. Roy, P.P., Rayar, F., Ramel, J.Y.: Word spotting in historical documents using primitive code-book and dynamic programming. *Image Vis. Comput.* **44**, 15–28 (2015). <https://doi.org/10.1016/j.imavis.2015.09.006>
14. Rusiñol, M., Aldavert, D., Toledo, R., Lladós, J.: Efficient segmentation-free keyword spotting in historical document collections. *Pattern Recognit.* **48**(2), 545–555 (2015)
15. Shanker, A.P., Rajagopalan, A.: Off-line signature verification using DTW. *Pattern Recognit. Lett.* **28**(12), 1407–1414 (2007)
16. Sudholt, S., Fink, G.A.: A modified isomap approach to manifold learning in word spotting. In: German Conference on Pattern Recognition, pp. 529–539. Springer (2015)
17. Sudholt, S., Fink, G.A.: Phocnet: a deep convolutional neural network for word spotting in handwritten documents. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 277–282. IEEE (2016)
18. Toselli, A.H., Puigcerver, J., Vidal, E.: Context-aware lattice based filler approach for key word spotting in handwritten documents. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 736–740. IEEE (2015)
19. Toselli, A.H., Vidal, E.: Fast hmm-filler approach for key word spotting in handwritten documents. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 501–505. IEEE (2013)
20. Wang, P., Eglin, V., Garcia, C., Largeron, C., Lladós, J., Fornés, A.: A coarse-to-fine word spotting approach for historical handwritten documents based on graph embedding and graph edit distance. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 3074–3079. IEEE (2014)
21. Zhang, X., Tan, C.L.: Handwritten word image matching based on heat kernel signature. *Pattern Recognit.* **48**(11), 3346–3356 (2015)

# Physical Intrusion Detection System Using Stereo Video Analytics



G. Aravamuthan, P. Rajasekhar, R. K. Verma, S. V. Shrikhande, S. Kar  
and Suresh Babu

**Abstract** Physical Intrusion Detection System (PIDS) detects unwanted elements (person or object) entering into a restricted zone. There are various single-camera solutions for surveillance available in the market. The key problem with video analytics based solution is false alert. With the addition of dimensional information of intruding object, it is possible to reduce false alerts without compromising an increase in true negative. We propose deployment of a pair of cameras in stereo configuration for facilitating dimensional measurements. This method also will be useful to detect and characterize flying objects entering in restricted zone. Experiments were conducted on various scenarios emphasizing the usability of stereo in surveillance.

**Keywords** Intrusion detection · Wide baseline · Far range stereo · Stereo video · Sparse depthmap

---

P. Rajasekhar · R. K. Verma · S. Babu  
EISD, Bhabha Atomic Research Centre, Mumbai, India  
e-mail: [rajs@barc.gov.in](mailto:rajs@barc.gov.in)

R. K. Verma  
e-mail: [vermark@barc.gov.in](mailto:vermark@barc.gov.in)

S. Babu  
e-mail: [subabu@barc.gov.in](mailto:subabu@barc.gov.in)

G. Aravamuthan (✉) · S. V. Shrikhande · S. Kar  
Homi Bhabha National Institute (HBNI), Mumbai, India  
e-mail: [amuthan@barc.gov.in](mailto:amuthan@barc.gov.in)

S. V. Shrikhande  
e-mail: [svs@barc.gov.in](mailto:svs@barc.gov.in)

S. Kar  
e-mail: [skar@barc.gov.in](mailto:skar@barc.gov.in)

## 1 Introduction

High security installations may contain large boundary and they need to be protected from unwanted elements entering the boundary. One solution would be to provide CCTV cameras at the boundaries and monitor from a centralized control room by security personnel. Continuous manual monitoring of the mostly eventless video feed is tiring, and hence there is a chance of missing some event. Hence automated solutions are recommended.

Proposed PIDS is one of the automated solutions used to detect the movements of an intruder attempting to breach a security wall or region and alert security. In PIDS video feed is analyzed (Video Analytics) to detect intrusions. The key problem with video analytics based solution is false alerts [1–3] which is due to inherent complications of understanding of the object detected in the video especially if the object is far from the camera. The object may appear very small in the image that makes recognition of the object more difficult.

To reduce the false alert, if sensitivity of the system is reduced, then true negative (Specificity) will be increased (Sensitivity Specificity trade-off) which means system may lose some important intrusion alerts. To reduce false alert without increase of true negative, additional information needs to be given to the system. Dimension of the object that is a potential intrusion (here after object/intrusion) could be one such information.

Based on the dimension of intrusion, false alerts can be filtered. In single-camera solution with the help of landscape layout, in principle it is possible to estimate the dimension of the ground object, which is not applicable for flying object. There are also focus-based depth estimation methods [4] available; but they are not accurate for long range for PIDS applications.

By applying stereo camera, one can estimate the depth and hence dimension which can be used to reduce the false alert. Some literature [5, 6] attempt to use stereo for surveillance, most of the solutions try to generate a depth map of the scene in front of the camera which is very complicated inherently due to correspondence problem.

We propose to use stereo video with sparse correspondence, in which left blob is matched with corresponding right blob to obtain sub-pixel level disparity and hence higher depth resolution even at greater depths. With depth information, it is possible to estimate size which can help in reduction of false alerts.

We do not process inside the intrusion to find what it contains, for example, animal or human, which requires large number of pixels. We use intrusion object only for the purpose of correspondence between left and right images which does not require large number of pixels. We argue that even less than a pixel intrusion can be detected and characterized.

The following section describes application of stereo camera in the field of security. Section three explains our methodology and Section four describes about various experiments we did to prove the basic concepts.

## 2 Related Work

There are few stereo video based installations [7] which cover only indoor or short-range applications. Few US patents [8, 9] describe the concept of stereo video analytics for security applications. Surprisingly, there is no reference of these patents in stereo video literature. There are many vehicle-based stereo video analytics [10, 11] which focus on dense depth map covering a range of 50 m. With dense depth map it would be difficult to get high depth resolution which could be one of the reasons for unpopularity of stereo-based intrusion detection systems.

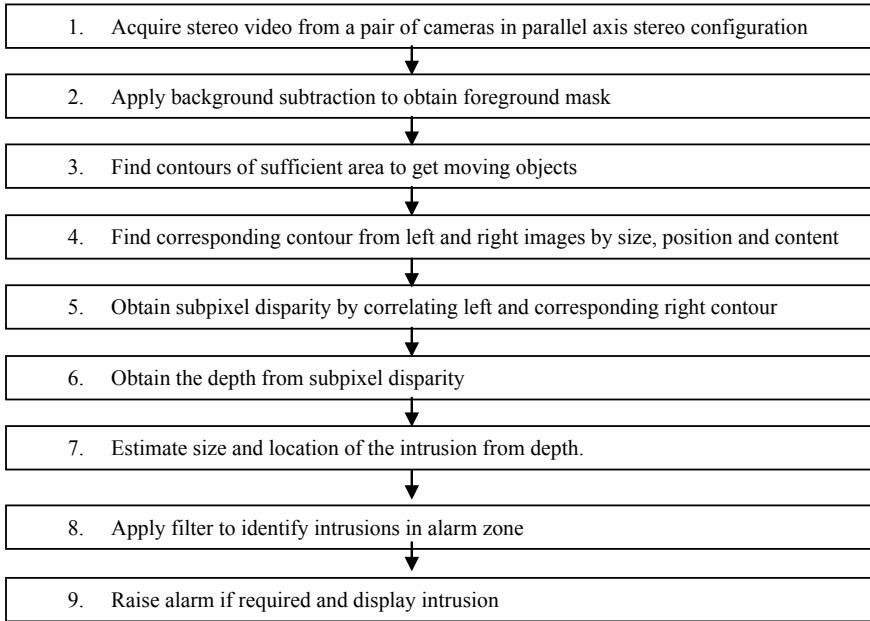
Generally stereo video applications look for dense depth map which faces problem of stereo correspondence and researches have not focused on sparse depth map [6]. We use wide baseline, far range sparse stereo video analytics for intrusion detection which has not gained attention among the research community. We use one depth per object. Since the entire object is checked for the correspondence, we will be able to get accurate sub-pixel correspondence and hence better depth resolution at far distance.

## 3 Our Method

As mentioned above, we use stereo video with sparse correspondence to get depth and dimensional information of intrusion which can help in the reduction of false alerts. First, a moving object is detected using background subtraction and then with the left-right correspondence, the depth is estimated. With the depth obtained, position and size of intrusion are estimated. The position coordinates and size are used for alarm filtration.

Please note that here we use object correspondence which helps to get sub-pixel correspondence and hence better depth resolution. We do not use correspondence to get position and dimensional information, instead use depth and pixel position information in the image to obtain intrusion position in the 3D space and the size. These are our two main differences with respect to standard stereo processing algorithms.

Flowchart of the proposed PIDS is given below:



Depth estimation, size and position estimation, and alarm filtering steps are given in more detail below.

### **3.1 Depth Estimation**

Once moving objects are detected from both left and right camera, by comparing size and position we can find corresponding object. This reduces the typical complicated dense stereo correspondence problem. Now we can get the disparity by calculating the horizontal distance between left and right corresponding object center.

When we did the above, due to variation in the left and right view we did not get the disparity accurately. We took a reduced version of an object in left image and cross-correlated it with its corresponding object in right image. This gave better-matched inner square with object in left frame. Then we calculated disparity by comparing sample points of sub-objects in left and right frame which improved correspondence.

With correspondence data, we can estimate depth by expression (1)

$$\text{depth} = \left( \frac{\text{focal\_pixel} * \text{baseline}}{\text{disparity}} \right) \quad (1)$$

where *focal\_pixel* is focal length in pixel units and *baseline* is horizontal distance between the two cameras, *disparity* is the distance between corresponding points in pixels.

### 3.2 Estimation of Position and Size of Intrusion

With depth and horizontal number of pixels and vertical number of pixels we can estimate horizontal size and vertical size of the object by using expression (3). Note that the width and height of the intrusion estimated are as perceived by camera and it may not be same as the object's actual width and height.

$$\text{angle} = \tan^{-1} \left( \frac{(\text{pix})}{\text{res}/2} * \tan \left( \frac{\text{fov}}{2} \right) \right) \quad (2)$$

$$\text{size} = \text{depth} * ((\tan(\text{angle1})) - (\tan(\text{angle2}))) \quad (3)$$

where *angle* is the angle made by a point in the scene represented by pixel in the image with respect to optical axis of the camera; for example *angle1* is left most point of an object and *angle2* is right most point of the same object for the calculation of width of the object. *res* and *fov* are resolution and field of view of camera.

With depth and position of object on the screen obtained in pixels, the horizontal and vertical position of the object can be estimated using expression (4). By associating GPS position and orientation of camera, with estimated position coordinates of the object one can obtain exact global position of the object.

$$\text{position} = \text{depth} * (\tan(\text{angle})) \quad (4)$$

where *angle* is, angle made by object (center point) with respect to optical axis of the camera.

### 3.3 Alarm Filtering and Visualization

We can compare position coordinates and dimensions of the object with pre-defined values to find if the object is sufficiently large and present in the alarm zone. Based on the result, visual alarm may be raised on the display and optionally an audio alarm also may be produced.

## 4 Experiments and Results

Applications of PIDS are further elaborated in this section. We have performed various experiments for each application to demonstrate the proof of concept. All the experiments estimate depth, dimension and position as mentioned in Sect. 3. Scenarios considered for experiments were similar to the real-life application of surveillance viz., museums, hazardous industrial zone, perimeter protection, etc.

### 4.1 Application of PIDS as Virtual Wall

PIDS can be deployed for detecting objects crossing a virtual wall. In the following experiment, we have demonstrated this application.

In this experiment we employed face-based intrusion detection to locate face in both left and right cameras. We could skip first four steps of flow chart given in Sect. 3 and by applying steps 4–7, the position and location of the face are obtained. They are displayed on the monitor shown in Fig. 1. We then created a virtual wall at two meters from the camera by applying the filter. If the person comes closer than two meters we give visual indication and audio sound. See the second part of the above image with red visual alarm.

### 4.2 Virtual Volume

PIDS can be deployed for detecting objects entering or present in a virtual volume. To demonstrate this we carried out the following experiment.

This experiment makes use of the position and size of the intrusion. For that we constructed a virtual volume in space. To visualize the virtual volume a wire frame



**Fig. 1** Virtual wall in action: Left image is person standing outside the virtual wall, right image is person standing inside the virtual wall

**Fig. 2** Virtual volume:  
When the blue cube is inside  
the cube alarm is generated



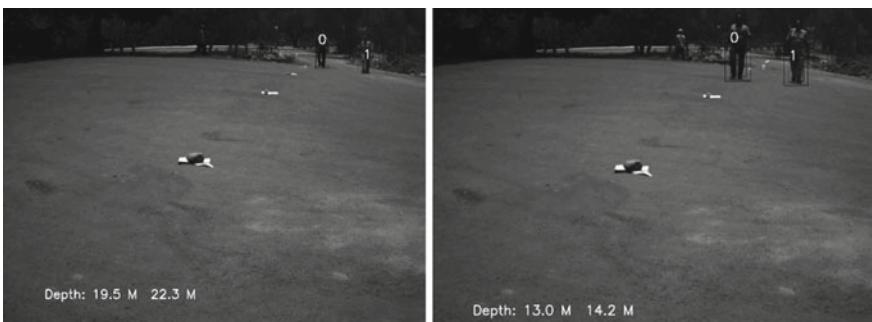
was placed. It was not required for the algorithm to detect intrusion. This time we made color-based object detection and matching.

When intrusion object with size more than 1 cm by 1 cm enters the virtual volume, an alarm is generated and displayed on the monitor as shown in Fig. 2.

#### 4.3 Narrow Baseline PIDS for Near Range (20 M)

PIDS is deployed to detect persons approaching a protected zone. For this a stereo camera setup with baseline of 12 cm is used.

This experiment involves all the steps given in flowchart. Two personnel are shown walking toward the camera. Markers are kept at every 5 m. As they approach, the depth of the personnel are displayed as shown in Fig. 3.



**Fig. 3** Picture shows intrusions and its depth. Left image shows two people walking at distance of around 20 m and right shows at depth of around 15 m



**Fig. 4** Image shows the checkerboard at 10, 100 and 200 m

#### 4.4 Wide Baseline Far Range Checker Board Experiments

In the previous experiment, we estimated the depth of the intrusion in the near range of around 20 m. To estimate the depth and size information from the image at far distance, we performed the following experiment.

In this experiment, we used one meter baseline for stereo camera. We moved Checkerboard in depth and captured image at every meter for 250 m. Checkerboard corners are used to calculate correspondence and checkerboard square dimension. Figure 4 shows three sample images at varying depths.

To reduce the error in depth estimation, the pan angle between two cameras were estimated using the above images and then used for estimation of depth and size of the Checkerboard. Depth and size table is given below.

In Table 1, please note at 10 m depth, dimension accuracy is higher than depth accuracy. It could be because of inaccuracy of depth measurement by tape (experi-

**Table 1** Actual and estimated depth and size of checker board

Depth measured using tape (meters)	Depth estimated using PIDS (meters)	Accuracy in depth estimated using PIDS (%)	Checkerboard size (0.15 m) estimation (meters)	Accuracy in checkerboard size estimation (%)
10	11.06	89.41	0.149	99.40
20	20.13	99.34	0.150	99.97
30	29.94	99.80	0.148	98.89
43	42.36	98.50	0.146	97.56
53	51.59	97.33	0.145	96.76
63	60.76	96.44	0.144	95.73
73	71.35	97.74	0.146	97.27
84	82.61	98.35	0.146	97.56
94	92.55	98.46	0.147	98.25
112	109.87	98.10	0.146	97.37

(continued)

**Table 1** (continued)

Depth measured using tape (meters)	Depth estimated using PIDS (meters)	Accuracy in depth estimated using PIDS (%)	Checkerboard size (0.15 m) estimation (meters)	Accuracy in checkerboard size estimation (%)
170	167.08	98.28	0.148	98.48
187	183.25	98.00	0.147	98.05
198	194.49	98.23	0.148	98.36
209	205.94	98.54	0.149	99.22
220	218.13	99.15	0.149	99.12
230	229.30	99.70	0.151	99.66

mental error). It can be also seen that the accuracy is increasing with depth this can be explained with the fact that the used pan angle is estimated from far image pairs.

## 5 Conclusion

In this paper, a novel way to reduce false alerts using stereo video analytics is described. Also it is shown how a typical problem of correspondence associated with stereo will naturally get resolved by use of intrusions. The experiments show usability of the idea of stereo video for obtaining position and dimension information even at large depths. This method also enables the detection of flying intrusions such as drone.

**Acknowledgements** The person present in Fig. 1 is Shri R. K. Verma, one of the authors of this paper. Consent has been taken from him.

## References

1. Thornton, J., Baran-Gale, J., Yahr, A.: An assessment of the video analytics technology gap for transportation facilities. In: IEEE Conference on Technologies for Homeland Security (2009). <https://doi.org/10.1109/ths.2009.5168025>
2. Norman, B.C.: Assessment of video analytics for exterior intrusion detection applications. In: IEEE International Carnahan Conference on Security Technology (2012). <https://doi.org/10.1109/ccst.2012.6393585>
3. Honovich, J.: Top 3 problems limiting the use and growth of video analytics (2008). <https://ipvm.com/reports/top-3-problems-limiting-the-use-and-growth-of-video-analytics>
4. Kulkarni, J.B., Sheelarani, C.M.: Generation of depth map based on depth from focus—A survey. In: International Conference on Computing Communication Control and Automation (2015). <https://doi.org/10.1109/iccubea.2015.146>
5. Hamzah, R.A., Ibrahim, H.: Literature survey on stereo vision disparity map algorithms. J. Sens. **2016**, Article ID 8742920, 23 (2016). <https://doi.org/10.1155/2016/8742920>

6. Lazaros, N., Sirakoulis, G.C., Gasteratos, A.: Review of stereo vision algorithms software to hardware. *Int. J. Optomechatronics* (2008). <https://doi.org/10.1080/15599610802438680>
7. Stepanov, D., Tishchenko, I.: The concept of video surveillance system based on the principles of stereo vision. In: 18th Conference of Open Innovations Association and Seminar on Information Security and Protection of Information Technology (2016). <https://doi.org/10.1109/fruct-ispit.2016.7561546>
8. Jin, M.W., Jin, L.Y., Murynin, AB., Kuznetsov, V.D., Ivanov, P.A., Jeong, I.-J.: United States Patent 7,088,243 B2 (2006)
9. Hassapis, C., Nishihara, H.K.: United States Patent 8,432,448 B2 (2013)
10. Marita, T., Oniga, F., Nedevschi, S., Graf, T., Schmidt, R.: Camera calibration method for far range stereovision sensors used in vehicles. In: IEEE Intelligent Vehicles Symposium (2006). <https://doi.org/10.1109/ivs.2006.1689654>
11. Ernst, S., Stiller, C., Goldbeck, J., Roessig, C.: Camera calibration for lane and obstacle detection. In: IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems (1999). <https://doi.org/10.1109/itsc.1999.821081>

# Identification of Fraudulent Alteration by Similar Pen Ink in Handwritten Bank Cheque



Priyanka Roy and Soumen Bag

**Abstract** In the research field of document image analysis, especially in handwritten documents, fraudulent alteration identification is a crucial task due to several forgery activities that are happening for few decades which affect a nation economically. In this paper, we are differentiating visually identical ink of different pens used for alteration in such documents by considering this problem as a binary classification problem. Here, we have used  $YC_bC_r$  color model. To formulate the problem a number of statistical and texture features are extracted to create feature vectors. These feature vectors are accommodated to generate data set which are classified by multilayer perceptron technique. The method has been executed on both blue and black ink samples. On average, the proposed method produces an efficient result of accuracy of more than 93% for known and 82.30% for unknown pen data, respectively. This performance measurement shows the efficacy of the proposed method comparing with other existing methods.

**Keywords** Fraudulent alteration · Handwritten document · Multilayer perceptron · Texture features ·  $YC_bC_r$  color model

## 1 Introduction

Forgery activities in bona fide documents especially the legitimate handwritten documents are intended to deceive existence of another body or entity. A fraudulent addition or alteration in financial documents, such as bank cheques, business contracts, medical bills alteration, etc. not only causes unrecoverable redress of individual concerned but also demolishes the financial condition of the commonwealth.

---

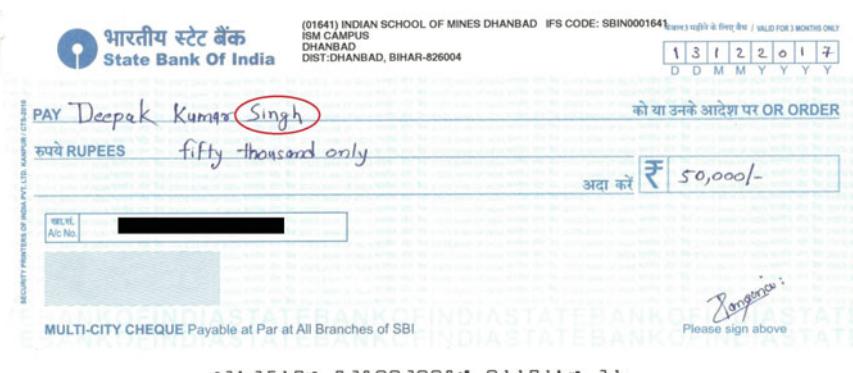
P. Roy · S. Bag

Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad 826004, Jharkhand, India

e-mail: [priyankaroy@gmail.com](mailto:priyankaroy@gmail.com)

S. Bag

e-mail: [bagsoumen@gmail.com](mailto:bagsoumen@gmail.com)



**Fig. 1** Image of a tampered handwritten bank cheque. The deception is indicated by red circle

A most common example is raising of payment amount in a bank cheque. Suppose a person who wants to donate an amount of five thousand by drafting cheque in favor of an organization and he has issued the cheque. But while the cheque has been processed by the bank authority, the payable amount is changed to twenty-five thousand. Figure 1 demonstrates an exemplar of another fraud deed which is marked by red circle in the image. In the original cheque, the person for whom this cheque was meant his name was changed from Deepak Kumar to Deepak Kumar Singh. Since, all the added words are of identical color ink, identification of authenticated and unauthenticated words happens to be quite intractable. Therefore, this directs us to mark the differentiation of similar pen ink in bank cheque.

Typically destructive and nondestructive are two separate units of ink analysis techniques. Merril and Bratick [1] have proposed a destructive method based on Fourier transform-Infrared spectrum analysis of Thin-Layer Chromatography plate of dye components of ballpoint pen ink. Taylor [2] has illustrated three different methods, viz., stereomicroscope, distilled water, and wax lift method to detect which of the given intersected lines is lastly drawn. The validity of these three methods has been examined by Chi-square test and their relative potency has been estimated based on line density and color. Few other destructive techniques are also found in the context of other compounds of the ink [3]. One major limitation of these methods is that original document has been blotted out which may be required in future. Contrarily, chromatographic analysis of hyperspectral imaging, analysis using microspectrometer, and combination of other image processing techniques with the aid of machine learning techniques are some example of nondestructive techniques. Khan et al. [4] have proposed a method by constructing spectral responses in 33 bands of hyperspectral images and then used K-mean clustering to find out ink dissimilarity. Khan et al. [5] have also presented more superior technique using principal component analysis, which has been applied to select bands that have most statistical information not all to boost the performance of the previous work. Dasari and Bhagavati [6] have created feature set consisting of total 11 statistical and texture features that have

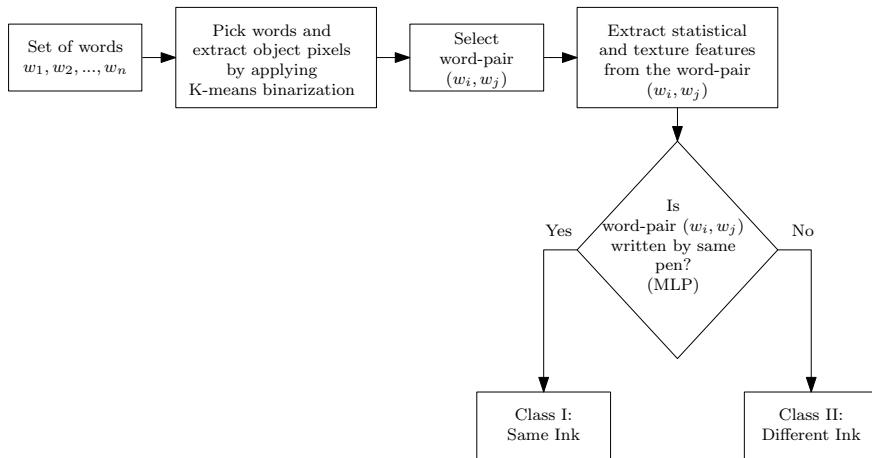
been taken from three color space: hue, saturation, and value space. To distinguish ink the features have been classified on the basis of distance measurements between features. Two types of distance, within same pen and between different pens, have been involved here. Kumar et al. [7, 8] have suggested some other techniques of identification of pen ink. From  $YC_bC_r$  and its opponent color model, they have extracted some texture features from gray-level cooccurrence matrix (GLCM), Legendre, and geometric moment features from two intersecting ink strokes and selected a discriminating set of feature using Multilayer Perceptron (MLP) classifier. Three different classifiers, namely, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and MLP have been used to test the accuracy. Gorai et al. [9] have introduced a histogram matching score based identification procedure to identify ink dissimilarity. From each RGB plane and corresponding gray plane, all total 12 feature images including LBP and Gabor filter with fixed orientation based feature images have been built. Normalized histogram of each feature image is compared together using a pre-settled threshold value to correctly identify the separation of the ink. Dansena et al. [10] have designed a binary classification technique for ink differentiation. Five statistics, namely, mean, variance, kurtosis, skewness, and mean absolute deviation have been used to extract 15 statistical features from RGB planes. Generated features are classified using MLP classifier.

Hyperspectral image acquisition techniques and spectrometric analysis are very costly due to application of expensive hardware. Moreover, it is too difficult to avail such instruments in markets. It is noted that the use of normal scanning device for image acquisition and application of classifier on feature set to distinguish the similar types of ink, the image processing methodology notably reduces the overhead of expensive appliances used in other operational modes to examine handwritten forensic documents. The main motto is to design an automation system to identify fraudulent alteration by analyzing ink pixels of handwritten words in bank cheques. In this paper, we have projected a nondestructive method that is effectively less expensive. The method has adequate potentiality for distinguishing perceptually similar ink of different pens. The problem has been defined as a two-class classification problem where one class is considered for those words that are written by same ink of a particular pen and another class containing words of different pens. The taken handwritten samples are simply scanned by the scanner, usually available in the market. The ink samples, basically the handwritten words are cropped manually. K-mean binarization technique has been applied to the word images for isolating ink pixels. Statistical and Gabor filter based features are computed and MLP classifier has been used to analyze the performance of the proposed method.

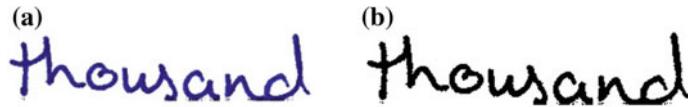
The remaining portions of this research study are documented accordingly: Sect. 2 narrates the proposed methodology for the detection of forgery in a handwritten bank cheque by ink analysis. The corresponding outcomes of the experiment and the conclusion of this study are illustrated in Sect. 3 and Sect. 4, respectively.

## 2 Proposed Methodology

The procedure is started with generating data set. The words of scanned color cheque images are cropped manually. Separated word images are converted into  $YC_bC_r$  model and K-mean binarization is used to separate foreground pixel from background. Some statistical and texture features are extracted exclusively from foreground pixel of each and every segmented word. At the end, differentiation of ink incompatibility testing is achieved by fetching generated feature sets into MLP classifier. All the manually cut words of a cheque are labeled into two classes. One class holds the words that are written by one type of pen ink whereas other one contains the remaining words of another. The concept of the proposed design is sketched in Fig. 2. The problem is designed as a binary classification problem. The objective is to identify whether two selected words are written by same pen or not. If the word-pair is written by same pen ink it is to be classified as class-I to indicate no forgery is there between the word-pair; otherwise, it is to be labeled to class-II. In this way, if all word-pairs of a particular cheque are classified as class-I then assuredly no deception has occurred. Separation of object pixel values from background is a foremost task. Using K-mean clustering based thresholding algorithm foreground pixels are excluded from grayscale images. A sample binarized output of this algorithm is shown in Fig. 3. Binarization reveals that there are two groups of assignment of gray levels of pixel; one group assigns object gray-level pixel intensity and the second group adapts the pixel values of the background area. A clear idea about K-mean binarization technique is also stated in consecutive subsection.



**Fig. 2** Proposed conceptual system diagram



**Fig. 3** A sample output of K-mean binarization technique. **a** A word before binarization and **b** after binarization

## 2.1 K-mean Binarization

The steps of K-mean thresholding algorithm are briefly described as follows:

1. Initialize the value of K by 2 and randomly select the centers of both partitions. The need of K is set to 2 for separating image pixels into two labels, one label for object pixels and another label for background pixels.
2. Generate new cluster by setting down gray-level intensity of pixel into nearest center. Assignment of pixel values to the nearest cluster is based on Euclidean distances of the pixel values to the centers of the clusters.
3. Again calculate the new centers of the clusters from the mean of the pixel intensity assigned to it.
4. Repeat step 2 and step 3 until the center values remain unchanged.
5. The threshold value is calculated as the average of two centers.

## 2.2 Feature Extraction

Features are the most essential keys used to build a system of classification problem. The intended problem is also not exempted from this. Though various manufacturers use different chemical substances to make ink, it is hard to differentiate color ink of two different pens of same color by enlarging or enhancing processes. Feature set helps to come out from this kind of hurdles. Moreover, use of color model also affects the path of identification of ink differentiation. The  $YC_bC_r$  color model is being appointed in this study. Images of every word are divided into cells of fixed size. As, for example, if size of an image is  $120 \times 250$  pixels and block size or cell size is  $10 \times 10$  pixels then the image will be partitioned into 300 blocks. Partitions that have equal to or more than 50% object pixels than the background pixels are taken under consideration. Features have been taken out exclusively from these partitions. In this juncture, some statistical features and textural features have to be introduced and those are briefly described in consecutive points.

1. **Mean:** The mean ( $\mu$ ) is a standard statistical measurement of the average brightness of the image. In this problem, the mean is calculated as

$$\mu = \frac{\sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_o(x, y)}{N_o} \quad (1)$$

$$\text{where, } I_o(x, y) = I(x, y)|I(x, y) \neq 0 \quad (2)$$

$N_o \subseteq M \times N$ , the total number of foreground pixels.

2. **Standard Deviation:** Standard deviation ( $\sigma$ ) is the measurement of variability of pixel intensity. Mathematically it can be written as follows:

$$\sigma = \sqrt{\frac{1}{N_o} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} (I_o(x_i, y_j) - \mu)^2} \quad (3)$$

3. **Moment:** Moments are the type of projection of image intensity function on the basis of different polynomials.

(a) **Legendre Moments:** Teague [11] gave a set of orthogonal moments based on Legendre polynomial for image analysis. Since a Legendre polynomial  $P_m(x)$  is defined on the interval  $[-1, 1]$ , pixel coordinates of an image of size  $M \times N$  are also mapped in the same range, i.e.,  $-1 < x, y < 1$ . The two-dimensional discrete Legendre moments of order  $(p, q)$  are defined in Eq. (4) [12]:

$$L_{pq} = \lambda_{pq} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} P_m(x_i) P_n(y_j) I(i, j) \quad (4)$$

where the normalizing constant  $\lambda_{pq} = \frac{(2m+1)(2n+1)}{MN}$ ,  $(x_i, y_j)$  represents the normalized pixel coordinates  $x_i = \frac{2i}{M-1} - 1$  and  $y_j = \frac{2j}{N-1} - 1$ . The Legendre polynomial  $P_m(x)$  of order  $m$  is given by

$$P_m(x) = \sum_{k=0}^m (-1)^{\frac{m-k}{2}} \frac{1}{2^m} \frac{(m+k)!x^k}{\frac{m-k}{2}! \frac{m+k}{2}! k!}; m - k = \text{even} \quad (5)$$

(b) **Geometric Moments:** Algebraic form of two-dimensional geometric moments of order  $(p, q)$  is

$$G_{pq} = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x^p y^q I(x, y) \quad (6)$$

By calculating Legendre and Geometric moments up to third order for each partition of the images of separated color channels  $Y$ ,  $C_b$ , and  $C_r$ , eight color features are obtained from each color space.

- 4. Gabor Filters:** Gabor filters are a set of band-pass filters. Gabor filter bank is widely involved in image processing and texture analysis and also for feature extraction problems [13]. A Gabor filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope. It can be written as follows:

$$G(x, y, \sigma, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} e^{j\pi(x\cos\theta+y\sin\theta)} \quad (7)$$

where  $\sigma$  and  $\theta$  represent the standard deviation and orientation of Gabor filter, respectively, and  $j = \sqrt{-1}$ . Though it indicates that it is a complex production; in this work, we evaluate magnitude of the filter along four different orientations,  $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ , keeping other parameters fixed. Applying the Gabor bank filters only on the  $Y$ -color space images, four magnitudes have been taken from each partition as four different features.

According to above knowledge, average for every single categorical feature has been obtained from each partition. Thus, these values are grouped to form a feature vector of 34 length. All these features are then normalized in  $[0, 1]$  range for further processing.

### 2.3 Feature Classification

In this investigation, MLP classifier has been used to identify diversity of individual but similar color ink of multiple pens. MLP classifier has been trained with training data set. During validation testing all the parameters of the classifier are tuned. The final assignment of the parameters is one hidden layer of 35 nodes, 1000 iterations, 0.3 learning rate, and all nodes are Sigmoid types. Feature vectors of different word-pairs are taken as input to train the computational neural network to achieve a desired level of accuracy for known data set. After completion of training stage, the same MLP model is applied to test the unknown data set with validation set testing.

## 3 Experimental Result and Discussion

How an adequate amount of data sets is created for training and testing purpose to simulate the behavior of the pen ink and the corresponding results is discussed in consecutive subsections.

### 3.1 Experimental Data Set Accumulation

In this experiment, IDRBT bank cheque database [14] is used for the creation of data set. There are 112 number of cheque bills written by 9 volunteers with the help of 14 pens. In these 14 pens, 7 pens are of blue ink and rest of 7 pens are of black ink. Moreover, each and every single cheque is drafted by two volunteers with two similar color ink. It has been found that the cheques are from four different Indian banks. For the sake of clear understanding, for each cheque  $C_i$  there are  $|w_{ik}|$  and  $|w_{il}|$  number of words written by pen  $P_k$  and  $P_l$ , respectively, where  $P_k$  and  $P_l$  either belongs to blue pen set or black pen set. Thus, all the handwritten entries are excluded manually from each other and are separated into two groups with respect to their pen type. The problem is framed as a two-class classification problem. The word-pairs that are written by same pens are labeled as class-I and the words-pair that are of different pens are denominated as class-II. The data set has been formulated for two cases; Case-I which considers the word-pair is written by same pen ink by defining a Cartesian product within the set of words of the corresponding pen itself. Contrarily, Case-II treats the word-pairs of different pens. These two conditions have been constructed by Cartesian product between two set of words of any of the two different pens of similar color ink. Hence, for the cheque set  $C = \{C_1, C_2, C_3, \dots, C_{112}\}$  and if say,  $w_k$  and  $w_l$  be the set of words written by pen  $P_k$  and  $P_l$ , respectively, we can mathematically express the number of combined instances of words under these two cases as follows:

**Case-I:** Here, we assume the word-pairs of the same pen of the same cheque with the exclusion of same word-pair instances. For 112 number of cheques, the number of combinations can be structured in terms of Cartesian product of two entities; therefore, for class-I the overall count of word-pairs is

$$\{|w_{ik}| \times (|w_{ik}| - 1)\} + \{|w_{il}| \times (|w_{il}| - 1)\} \quad (8)$$

**Case-II:** In this case, on account to all the cheque bills, the corresponding instances are formed by associating word-pairs of different pens of similar ink. To remove the biasness of the classification operation, we reorder the occurrences of word-pair in the Cartesian product. Hence, the number of word-pairs in respect of the class-II can be expressed as follows:

$$(|w_{ik}| \times |w_{il}|) + (|w_{il}| \times |w_{ik}|) \quad (9)$$

Based on the above aspects feature vectors are arranged together with a class value in favor of the two cases. Generated data sets are simulated using MLP classifier. We have performed the present experiment by considering two-pen left out strategy. In this situation, every time the data set is sorted by keeping two-pen combination out for testing. The data set is separated into three sets called training set, testing set, and validation set. Whenever it is said that a pen-pair ( $P_k, P_l$ ) is kept out that means the training data set contains only instances of all other pens except the ( $P_k, P_l$ )

pen-pair. Whereas the corresponding instances of the pen-pair ( $P_k, P_l$ ) is either from blue pen set or from black pen set are kept as test data set. In this way, for 14 pens 42 number of training and testing data sets are constructed. To classify known data set, 10-fold cross-validation test is performed on all training data sets and the MLP classifier is learnt from it. After this, the unknown data set is tested on MLP neural network model with the help of validation set formation. Validation set is created by keeping out 20% data set from the training data set during training period of the MLP model. For each pen-pair combination, the training data set is divided into ten equal partitions. Out of the ten partitions, nine partitions are taken for a particular pen-pair to form ten different training data sets. According to this order, for all 42 combinations of pen-pair, training data sets have been prepared. Evaluation of all these training data sets has been carried out with the help of MLP classifier. The same trained MLP model has been tested on unknown data sets to achieve a promising level of accuracy of the system owing to unroll the differences of similar color ink.

### **3.2 Performance Accuracy and Comparison**

The performances of the proposed method are precised in Tables 1 and 2 for both blue ink pen set and black ink pen set separately. The second column is the representative of the two-pen kept out scheme. With respect to this approach, the third and fourth columns depict the efficiency of the proposed system for known data set and unknown data set successively. As an example, for known data, the pen-pair  $P_1P_2$  is kept out means the training data contains remaining pen combinations and corresponding test data is one of the ten parts of the training data. The kept out sets have been used as test data for unknown data. Confusion matrix is used to summarize the performance of the classifier. It allows us to describe the accuracy in terms of calculating the ratio of truly predicted cases to total number of predictions made. It has been found that the average accuracy of this technique for blue ink pens and black ink pens are 94.31% and 91.93% for known data sets, respectively. Similarly, for unknown data sets, we have reported here the rate of correct classification of ink differentiation for both blue and black ink pens that are 83.06% and 81.54%, respectively. Higher accuracy for known data is quite possible due to the model that learns from known data and is also tested on known data. One of the ten parts of the training set is kept as testing purpose for known data. Over and above, each and every time for unknown data the MLP builds a new model and evaluates the model based on completely unfamiliar test data. Due to nonsimilarity, evaluation of unknown data is free from biased results of classification.

In this article, a comparison between our proposed methodology and two other existing techniques has been made and recorded in Table 3 with respect to the database [14]. Gorai et al. [9] have built a technique using word-based histogram matching score. In addition to this, for creating data set they have not considered the writer biasness condition. Dasari and Bhagavati [6] introduced a hue, saturation, and value color space based ink matching identification technique. Their method is also exempted

**Table 1** Accuracy for identification of blue pen differentiation

Sl. no.	Pen-pair kept out	Known data accuracy (%)	Unknown data accuracy (%)
1	$P_1 P_2$	93.11	85.94
2	$P_1 P_3$	94.31	87.06
3	$P_1 P_4$	91.80	87.72
4	$P_1 P_5$	93.93	85.19
5	$P_1 P_6$	92.85	89.72
6	$P_1 P_7$	93.32	84.47
7	$P_2 P_3$	94.84	81.00
8	$P_2 P_4$	93.43	83.87
9	$P_2 P_5$	95.34	78.66
10	$P_2 P_6$	94.20	81.23
11	$P_2 P_7$	95.57	79.50
12	$P_3 P_4$	94.80	83.52
13	$P_3 P_5$	94.49	81.33
14	$P_3 P_6$	94.49	81.98
15	$P_3 P_7$	95.93	79.51
16	$P_4 P_5$	95.25	82.67
17	$P_4 P_6$	92.93	86.03
18	$P_4 P_7$	93.56	84.98
19	$P_5 P_6$	96.01	80.81
20	$P_5 P_7$	96.42	79.75
21	$P_6 P_7$	93.94	79.34
	Average	94.31	83.06

from the assumption of unbiased writer consideration. Our proposed method, attentively, has looked and processed this matter and offers better result compared to those two procedures stated above. Besides, in context to the proposal of Dansena et al. [10], though they have shown a qualitative result for the same database they have performed only cross validation. Another thing of their work is that for classification they have considered a large number of iteration of 5000 to get a higher level of accuracy. More number of iterations slow down the process of classification. Consequently, to concretize the performance of our proposed method, classification has been done with validation testing with only 1000 iterations. A validation set is 20% of the training data set, and hence the MLP model is learnt from less number of instances that reduces the accuracy. In spite of the classification rate being decreased due to the validation set, the model is more generalized toward unknown data set. Furthermore, validation testing helps to stop the training process earlier before the completion of the total number of iterations assigned. As soon as validation set error

**Table 2** Accuracy for identification of black pen differentiation

Sl. no.	Pen-pair kept out	Known data accuracy (%)	Unknown data accuracy (%)
1	$P_8 P_9$	92.50	86.70
2	$P_8 P_{10}$	91.87	82.49
3	$P_8 P_{11}$	92.00	88.60
4	$P_8 P_{12}$	92.23	81.55
5	$P_8 P_{13}$	91.65	88.19
6	$P_8 P_{14}$	90.91	85.53
7	$P_9 P_{10}$	93.00	76.31
8	$P_9 P_{11}$	92.41	84.56
9	$P_9 P_{12}$	93.53	72.96
10	$P_9 P_{13}$	91.93	83.80
11	$P_9 P_{14}$	91.30	80.37
12	$P_{10} P_{11}$	91.89	81.22
13	$P_{10} P_{12}$	92.00	75.86
14	$P_{10} P_{13}$	91.76	83.15
15	$P_{10} P_{14}$	92.01	78.28
16	$P_{11} P_{12}$	92.10	78.32
17	$P_{11} P_{13}$	91.66	87.03
18	$P_{11} P_{14}$	91.79	76.93
19	$P_{12} P_{13}$	92.29	80.84
20	$P_{12} P_{14}$	91.42	76.22
21	$P_{13} P_{14}$	90.21	83.45
	Average	91.93	81.54

**Table 3** Relative comparison of the proposed one with other existing methods

Sl. no.	Method	Blue ink accuracy (%)	Black ink accuracy (%)	Average accuracy (%)
1	Gorai et al. [9]	54.8	54.71	54.76
2	Dasari and Bhagavati [6]	80.58	80.93	80.76
3	Proposed method	83.06	81.54	82.30

increases compared to the previous step error, the training process then and there stopped itself. On a final note, the proposed method seems to be qualitatively sound and has considerable strength to engage further research.

## 4 Conclusion

In this study, we have explored a nondestructive image processing technique to find the differences between similar ink of pens of different brands. In a nutshell, it can be said that it is an effective technique to identify fraudulent alteration in handwritten forensic documents. This procedure has been applied on both blue and black color ink differentiation. To obtain and to analyze the feature set, we involve  $YC_bC_r$  color space conversation of object pixel. Thirty-four color and texture features including moments and Gabor filter based features have been computed from each extracted word. Comparison and classification of several word-pairs based on the feature vector have been undertaken with the aid of MLP classifier. Performance of the classifier has been checked out on both the known and unknown data sets, respectively. Besides that, the efficacy of this method over the two other methods has also been discussed.

**Acknowledgements** The authors would like to thank the sponsor of the project “Design and Implementation of Multiple Strategies to Identify Handwritten Forgery Activities in Legal Documents” (No. ECR/2016/001251, Dt.16.03.2017), SERB, Govt. of India. The authors are thankful to the research scholar Prabhat Dansena of Indian Institute of Technology (ISM) Dhanbad for providing the cheque which is used for problem demonstration at Fig. 1. He holds no conflicts of interest by any means and gives full consent for the publication of the article.

## References

1. Merrill, R.A., Bartick, E.G.: Analysis of ball pen inks by diffuse reflectance infrared spectrometry. *J. Forensic Sci.* **29**(1), 92–98 (1992). <https://doi.org/10.1520/JFS13260J>
2. Taylor, L.R.: Intersecting lines as a means of fraud detection. *J. Forensic Sci.* **37**(2), 528–541 (1984). <https://doi.org/10.1520/JFS11639J>
3. Chen, H.-S., Meng, H.-H., Cheng, K.-C.: A survey of methods used for the identification and characterization of inks. *J. Forensic Sci.* **1**(1), 1–14 (2002)
4. Khan, Z., Shafait, F., Mian, A.: Hyperspectral imaging for ink mismatch detection. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 877–881 (2013). <https://doi.org/10.1109/ICDAR.2013.179>
5. Khan, Z., Shafait, F., Mian, A.: Automatic ink mismatch detection for forensic document analysis. *Pattern Recognit.* **48**(11), 3615–3626 (2015). <https://doi.org/10.1016/j.patcog.2015.04.008>
6. Dasari, H., Bhagvati, C.: Identification of non-black inks using HSV color spaces. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 486–490 (2007). <https://doi.org/10.1109/ICDAR.2007.4378757>
7. Kumar, R., Pal, N.R., Sharma, J.D., Chanda, B.: A novel approach for detection of alteration in ball pen writings. In: Proceedings of the International Conference on Pattern Recognition and Machine Intelligence, pp. 400–405 (2009). [https://doi.org/10.1007/978-3-642-11164-8\\_65](https://doi.org/10.1007/978-3-642-11164-8_65)
8. Kumar, R., Pal, N.R., Sharma, J.D., Chanda, B.: Forensic detection of fraudulent alteration in ball-point pen strokes. *IEEE Trans. Inf. Forensics Secur.* **7**(2), 809–820 (2012). <https://doi.org/10.1109/TIFS.2011.2176119>
9. Gorai, A., Pal, R., Gupta, P.: Document fraud detection by ink analysis using texture features and histogram matching. In: Proceedings of the International Joint Conference on Neural Networks, pp. 4512–4517 (2016). <https://doi.org/10.1109/IJCNN.2016.7727790>

10. Dansena, P., Bag, S., Pal, R.: Differentiating pen inks in handwritten bank cheques using multi-layer perceptron. In: Proceeding of the International Conference on Pattern recognition and Machine Intelligence, vol. 10597, pp. 655–663 (2017). [https://doi.org/10.1007/978-3-319-69900-4\\_83](https://doi.org/10.1007/978-3-319-69900-4_83)
11. Teague, M.R.: Image analysis via the general theory of moments. *J. Opt. Soc. Am.* **70**, 920–930 (1980). <https://doi.org/10.1364/JOSA.70.000920>
12. Teh, C.H., Chin, R.T.: On image analysis by the methods of moments. *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, 496–513 (1988). <https://doi.org/10.1109/CVPR.1988.196290>
13. Kong, W.K., Zhang, D., Li, W.: Palmprint feature extraction using 2-D gabor filters. *Pattern Recognit.* **36**(10), 2339–2347 (2003). [https://doi.org/10.1016/S0031-3203\(03\)00121-3](https://doi.org/10.1016/S0031-3203(03)00121-3)
14. IDRBT Cheque Image Dataset: <http://www.idrbt.ac.in/icid.html>

# Faster RCNN-CNN-Based Joint Model for Bird Part Localization in Images



Arjun Pankajakshan and Arnav Bhavsar

**Abstract** Bird species classification is a challenging task in the field of computer vision because of its fine-grained nature, which in turn can lead to high interclass similarities. An important aspect for many fine-grained categorization problems involves processing of local-level semantics. This highlights the need for accurate part detection/localization. In this work, we propose a two-step approach to address the problem of bird part localization from an input image. In the first step, a Faster RCNN (FRCNN) is learnt to suggest possible bird part regions. However, the part region proposals given by Faster RCNN are not always precise. To refine these, a second step involving a CNN-based part classifier, trained only on bird part segments is used. Both FRCNN and CNN part classifiers are trained separately in a supervised manner. The part classifier effectively builds upon the FRCNN region proposals, as it is trained on more specific data as compared to FRCNN. We evaluate the proposed framework on the standard CUB-200-2011 bird dataset, as well as on a newly collected IIT Mandi bird dataset, where the latter is used only during testing.

**Keywords** Fine-grained classification · Bird part detection · Faster RCNN · CNN

## 1 Introduction

Fine-grained recognition involves identifying discriminative features to successfully distinguish between closely related entities (e.g., species of birds, flowers, plants, etc.). The task of visual fine-grained classification can be extremely challenging due to the similarity of the general appearance among different species of the same entity. Thus, it is important to consider differences in local regions, in order to identify some distinguishing regions/features. As a result, most of the methods [1, 2] defined for

---

A. Pankajakshan · A. Bhavsar (✉)

School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, India  
e-mail: [arnav@iitmandi.ac.in](mailto:arnav@iitmandi.ac.in)

A. Pankajakshan

e-mail: [arjunp@projects.iitmandi.ac.in](mailto:arjunp@projects.iitmandi.ac.in)

such a task generally follow the hierarchy of locating the parts in the image and then using these for classification. Thus, the classification performance depends on the accuracy and precision in localizing different parts in an image.

Some of the traditional local feature extractors in Computer Vision includes SURF [3], HOG [4], LBP [5], etc. A straightforward approach to part-based recognition is to extract local features using any of these methods and build a classifier. However, these standard feature extractors are unlikely to give an optimal solution, as the feature extraction is data independent, and do not use the highly domain-specific nature of the task.

Deep convolutional descriptors have proven to be more effective for part-based fine-grained recognition [6, 7] since these networks learn data-specific regional features [8]. Region proposal based models such as RCNN [9], Fast RCNN [10], and Faster RCNN [11] are examples of such robust part detection models.

In this work, we first demonstrate that even with a limited amount of data (few thousand images), the FRCNN model can be used for effective bird part localization. However, in this process, we also notice that it is also prone to make some detection errors, as the task involves relatively small regions. Thus, we also propose to augment the FRCNN-based part localization with a CNN-based part classifier, which is trained on part-specific regions. Unlike FRCNN, instead of considering the complete image, the part classifier is trained on only local part regions and fine-tuned on the region proposals from FRCNN. Thus, for the CNN part classifier the training is more specific, the fine-tuning requires much fewer examples. This helps it to correct some of the incorrect estimates by the FRCNN. We demonstrate the effectiveness of such a two-step approach on the standard CUB-200-2011 bird dataset [12], and an indigenous IIT Mandi bird dataset, where the latter is not used at all in the training or validation process.

The main contributions in this work are listed as follows: (1) We demonstrate the effectiveness of FRCNN for bird part localization. (2) We propose a joint model approach, utilizing the effectiveness of FRCNN in bird part localization and augmenting it with a CNN-based bird part classifier. (3) We demonstrate a simple and efficient fine-tuning strategy of the CNN part classifier based on the FRCNN region proposals. (4) We show that the proposed joint model can give good quality bird part localization for images from an unseen dataset, not used for training or validation.

## 2 Related Works

Resemblance of body parts and variations due to pose, background, and occlusions in the image, respectively, cause interclass similarity and intra-class variability between bird species. Thus, various approaches for fine-grained classification include the task of part localization. For instance, the work in [13] discusses about bird part localization using exemplar-based models. In [14, 15], the authors explore human-in-the-loop methods to identify the most discriminative regions for classifying fine-grained categories. A bilinear CNN model [16] which employed a two-stage pipeline

for part detection and part-based object classification has been introduced. The part-based fine-grained recognition methods [6, 17] used both bounding boxes of the birds and part annotations during training to learn an accurate part localization model. In [18], a learning framework that is able to connect text terms to its relevant bird parts has been proposed. Convolutional attention neural networks [19, 20] are state-of-the-art models used for fine-grained recognition of bird parts.

However, in the above methods, part localization is only reported as a part of the overall classification process, and the performance of part localization is not discussed. In this paper, we focus on the part localization problem and its performance with a joint model of well-known deep learning frameworks. Specifically, we propose a supervised approach, jointly using Faster RCNN and a CNN-based part classifier to localize different parts from an input bird image, where the CNN part classifier can refine the region proposals given by Faster RCNN, and needs a small amount of data to be trained.

### 3 Proposed Framework

The block diagram of the proposed joint model framework is shown in Fig. 1, which involves an FRCNN stage and a CNN-based part classifier. We discuss these in detail below in context of this work.

#### 3.1 *Bird Parts Region Extraction Using Faster RCNN*

Faster RCNN [11] is a popular contemporary deep learning model for object detection problems. FRCNN can be decomposed into two modules, the first being a Region Proposal Network (RPN) that gives region proposals, while the second module is the Fast RCNN detector [10]. Input to the Faster RCNN model is the full image dataset with part annotations and corresponding bounding box coordinates. The RPN has two channels, one for object classification and other for object bounds. The final classification yields refined object bounds through a regression process. Once the Faster RCNN model is trained, at the evaluation stage for each of the test images the model gives part annotations and the corresponding bounding box coordinates.

In the proposed method, the trained Faster RCNN is used as a part region extractor. Full bird images are used to train Faster RCNN model. We have used three different scales (64, 128, 256) and three different aspect ratios ([1, 1], [1, 2], [2, 1]) to train Faster RCNN model (Please see [11] for details on the scale and aspect ratio parameters). The model is optimized using Adam optimizer and trained for 350 epochs.

While FRCNN generally yields a reasonably good performance by itself, we observed that similar part regions such as end portion of tail and the beak, and the presence of background plus variations due to pose and occlusions in the training dataset can result in wrong predictions by the Faster RCNN model. Even though the

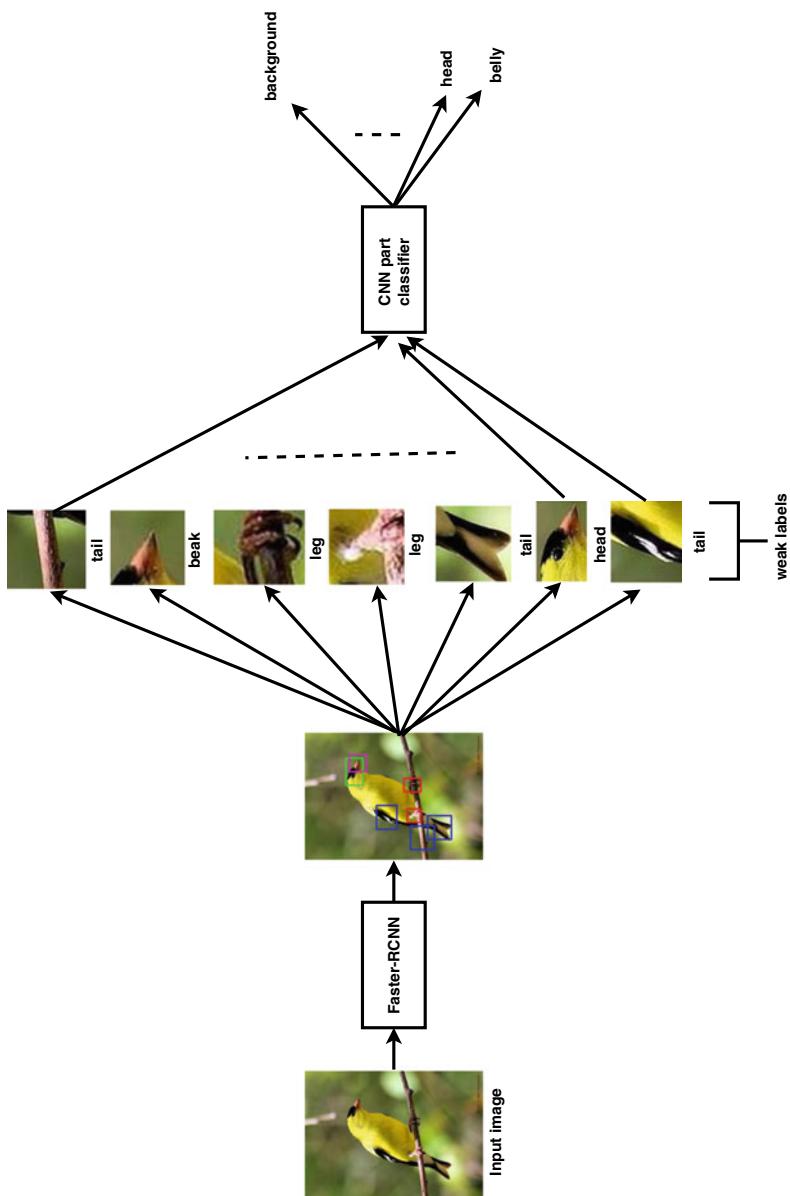


Fig. 1 Block diagram of proposed joint framework

bird class is a part of the Imagenet dataset, on which the Faster RCNN is originally trained, it does not yield precise prediction for bird part regions. This can be due to the fact that Faster RCNN is trained on full images, so the region proposals still contain a part of background which may be pose ambiguities with the actual parts, in the classification module.

### 3.2 Refining FRCNN Regions Using CNN Part Classifier

A deep convolutional part classifier is used to refine the region proposals given by the FRCNN model. Inspired from VGG19 [21] architecture, the proposed CNN part classifier has three pairs of stacked convolutions—max-pooling layers followed by three dense layers including the softmax layer. The input to the network is  $120 \times 120$  RGB images of bird parts. For regularization, batch normalization has been used after convolutional layers and dropout with the probability of 0.5 has been used after dense layers. The network is optimized using Adam optimizer with a learning rate of 0.00002 and categorical cross entropy as the loss function. The architecture of the model is shown in Fig. 2.

The CNN part classifier is first trained separately on annotated part regions from the same training images as used for the FRCNN model. Such part-based region of interest (ROI) used for training the part classifier are specifically cropped from the training images. However, during the decision-making process, it so happens that the region proposals from FRCNN often involve a larger/smaller field-of-view or are differently scaled, relative to the part ROIs on which the CNN part classifier is trained. Thus, as we demonstrate, while the baseline performance of the CNN-based

**Fig. 2** CNN part classifier architecture

<b>Input shape (120x120x3)</b>
<b>3x3 Conv2D 16 ReLU</b>
<b>3x3 Conv2D 16 ReLU</b>
<b>2x2 MaxPooling2D, stride (2x2)</b>
<b>3x3 Conv2D 32 ReLU</b>
<b>3x3 Conv2D 32 ReLU</b>
<b>2x2 MaxPooling2D, stride (2x2)</b>
<b>3x3 Conv2D 64 ReLU</b>
<b>3x3 Conv2D 64 ReLU</b>
<b>2x2 MaxPooling2D, stride (2x2)</b>
<b>Dense 256 ReLU</b>
<b>Dense 256 ReLU</b>
<b>Dense 5 softmax</b>

part classifier is very good, it is not robust to variations in the FRCNN output region proposals, relative to its own training data.

To mitigate this issue, CNN part classifier is fine-tuned with 20% data that is not used for training. Thus, this 20% data plays a similar role to validation data in pattern recognition applications. For fine-tuning, we select all the correct region proposals given by Faster RCNN model for each part within the 20% validation data. Finally, the part region proposals from Faster RCNN model on evaluation data is verified using fine-tuned CNN part classifier.

Importantly, we note that the fine-tuning process is scalable to new data, as it involves much lesser data (few FRCNN proposals) than that in the original training process for itself and FRCNN. Thus, in an active learning-based system-level application (e.g., a web-based bird identification tool), such a fine-tuning can be carried out easily, as required.

## 4 Experimentation and Results

We now describe various aspects of the experimentation followed by quantitative and qualitative results.

### 4.1 Datasets and Implementation Details

Performance analysis of the proposed method has been carried out using two datasets. For training both components of the joint model, the Caltech-UCSD Birds-200-2011 (CUB) [12] dataset is used. For evaluation, in addition to the test images from the CUB data, we also use the IIT Mandi bird dataset consisting of Indian birds. Thus, using the IIT Mandi bird dataset enables us to demonstrate the effectiveness of the proposed approach when testing on samples from a completely unseen dataset. The descriptions about datasets are given below.

**CUB-200 dataset** The dataset contains 11,788 images of 200 bird species with bounding boxes for bird region. Also, each image in the dataset has 15 parts annotations. As each individual part is very small, we consider larger semantic regions consisting of neighboring parts. These regions include head, tail, belly, and leg for training and evaluation.

**IIT Mandi bird dataset** We have created our own dataset for exclusive evaluation of the proposed joint model framework. While we are still in the process further building up the dataset, presently, it consists of around 30 species with a total of 600 images. While the current relatively small with respect to the classification task (with low per-class images), we believe it is still reasonable for a part localization problem.

**Training FRCNN and CNN part classifier** For training the FRCNN model, bounding boxes of fixed size were created with respect to each of the abovementioned part

regions. The size of the bounding boxes ranges from  $80 \times 80$  to  $110 \times 110$  depending on the part region size. For training the CNN part classifier model, image patches corresponding to each part were cropped out using the center pixel locations for parts given in the dataset. We have also considered background patches in the training process. Overall the model is trained with five classes including background. All the image patches are reshaped into a fixed size of  $120 \times 120$ . Some parts of certain images were not visible that results in variations between total number of examples for each parts.

For the CUB dataset, the FRCNN is trained with 60% of the data. It has been ensured that the number of examples for each part is balanced in training and testing phase. For the CNN part classifiers, 60% of the data is used for training (same as FRCNN), 20% for fine-tuning, and 20% for testing. The testing results for both models (only FRCNN and joint model) are reported on the remaining 20% data. As mentioned earlier, the IIT Mandi dataset is neither used in training nor during the validation process.

## 5 Experimental Results and Discussion

We now discuss our experimental results on the two datasets, highlighting some important observations.

### 5.1 On CUB Dataset

In Table 1, we first demonstrate the baseline performance of the CNN part classifier. The baseline performance indicates that performance when the CNN part classifier is tested on the part regions which are specifically cropped (similar to its training data) without any geometric variations of scales and aspect ratio (as induced in the FRCNN region proposals). The experimental results showed that CNN part classifier trained on bird part images can effectively distinguish different parts with a very high overall accuracy. This demonstrates that trained CNN part classifier can be effective in distinguishing parts across bird species and a variety of background.

**Table 1** Baseline performance of the CNN part classifier

Bird part	CNN part classifier (%)
Leg	96.8
Tail	97.7
Belly	99.2
Head	97.9

**Table 2** CNN part classifier accuracy for individual parts on FRCNN part region proposals (without fine-tuning)

Bird part	CNN part classifier (%)
Leg	80.7
Tail	80.7
Belly	67
Head	44.4

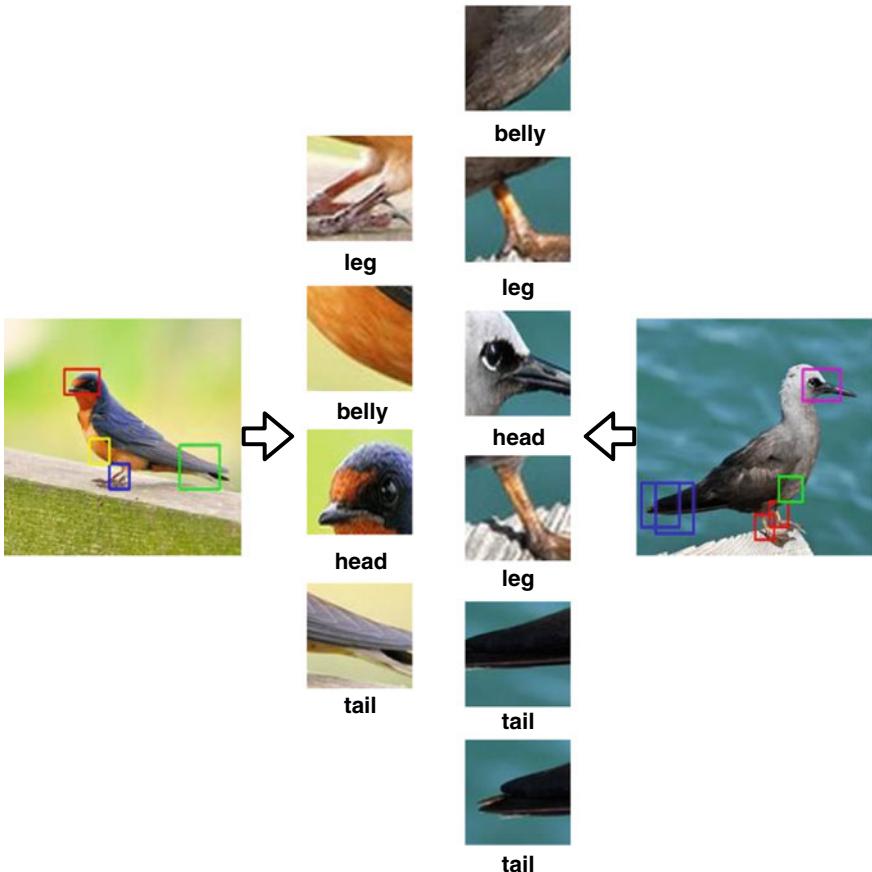
**Table 3** Individual part classification accuracy for Faster RCNN and the proposed joint model (with fine-tuning)

Bird part	Faster RCNN (%)	Proposed joint model (%)
Leg	93.2	94.5
Tail	93.5	94.3
Belly	88.8	90.2
Head	76.2	78.6

However, as discussed in Sect. 3.2, we observe that, while the baseline CNN part classifier performs very well, the accuracy significantly drops when the FRCNN region proposals are used as inputs (Table 2). By analyzing the predicted label for each part by CNN classifier and the actual label returned by Faster RCNN model, we observe that this is due to different scaling and field-of-views of the region proposals in Faster RCNN model. As a result, fine-tuning the CNN classifier was carried out as discussed earlier.

Following the fine-tuning, the CNN part classifier accuracy is improved and is shown in Table 3. First, we note that the FRCNN yields a high accuracy, for this task, even with less retraining data. Further, we also note that the CNN stage further improves the performance over the FRCNN classifier. Note that this improvement is achieved with fine-tuning, using data which is even lower than that used for originally retraining the FRCNN and CNN. We believe that a straightforward extension of the proposed approach with a better CNN architecture can yield further improvement over the FRCNN results.

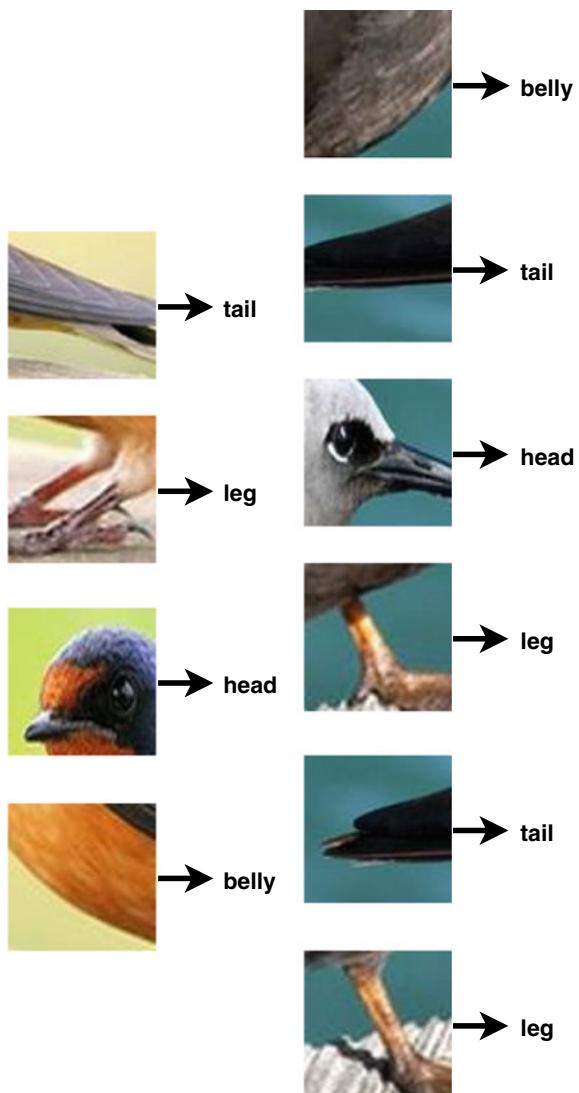
In addition to the quantitative results, our final interest is in the predicted labels returned by CNN part classifier, as an incorrect part label can directly affect the classification. Note that even one correction in a part label among the four part labels can have useful implications for a misclassified images in a bird species identification task. This is demonstrated in the qualitative examples discussed below.



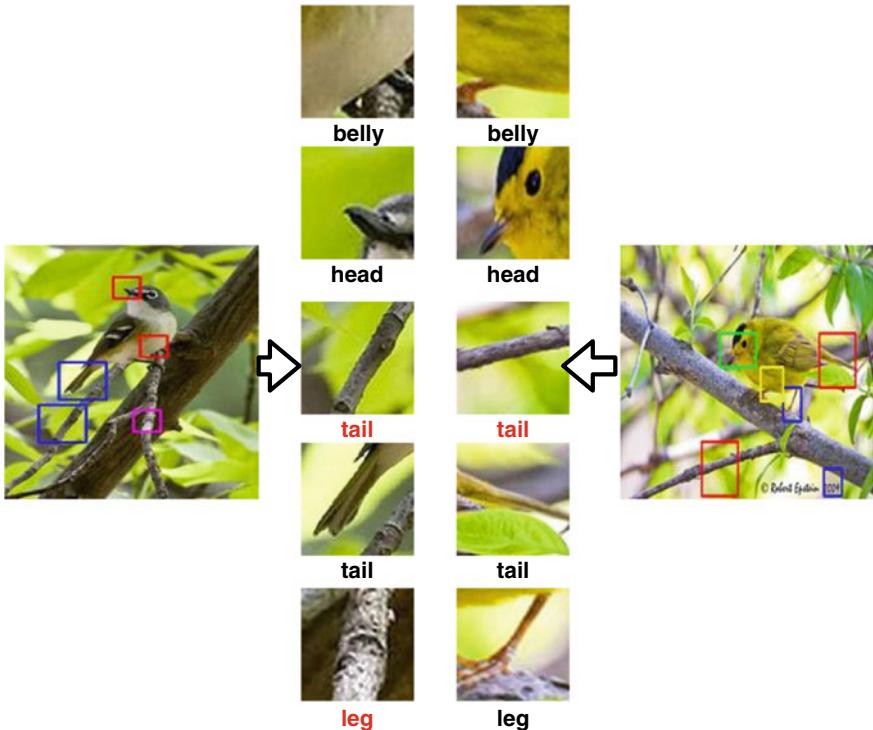
**Fig. 3** Examples for correct Faster RCNN region proposals on Caltech dataset

Figure 3 shows two examples for correct bird part region proposals from Faster RCNN on Caltech dataset and Fig. 4 represents the corresponding CNN classifier predictions. Since all region proposals from FRCNN is correct, there are no issues for classifier. Figure 5 shows two examples in which there are wrong bird part region proposals from Faster RCNN on Caltech dataset and Fig. 6 represents the corresponding

**Fig. 4** CNN part classifier predictions for region proposals in Fig. 3



CNN classifier predictions. Analyzing Fig. 6 it is clear that CNN classifier can predict the correct labels for wrong weak labels given by Faster RCNN model. Some region proposals are still predicted wrongly. Correct predictions are marked in green color and wrong predictions in red color.



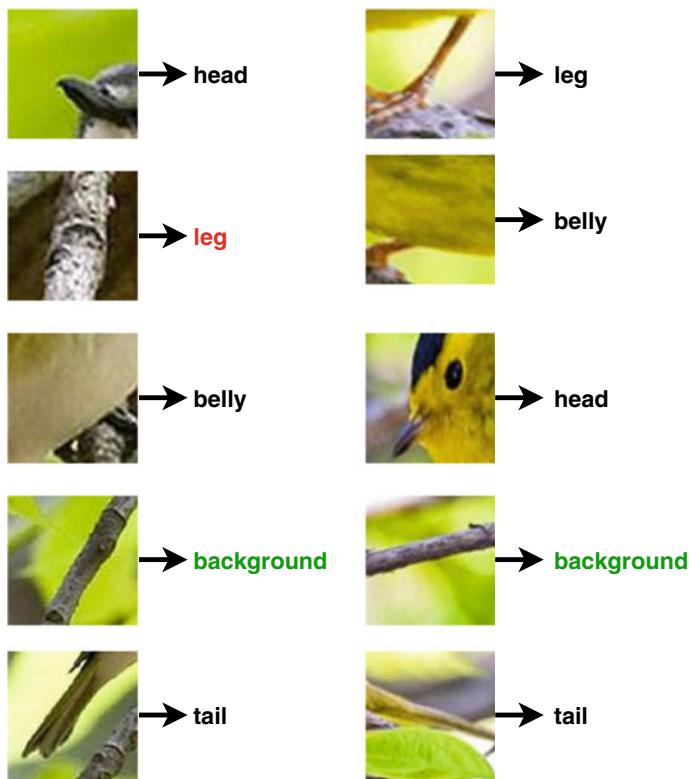
**Fig. 5** Examples for wrong Faster RCNN region proposals on Caltech dataset

## 5.2 On IIT Mandi Dataset

As mentioned earlier, this dataset is exclusively used for evaluation purpose. On this dataset, the overall accuracy after the CNN part classifier operates on the FRCNN region proposals which are as follows:

- Without fine-tuning: 51.3%,
- With 20% data for fine-tuning (using CUB data): 73.6%, and
- With 40% data for fine-tuning (using CUB data): 82.1%.

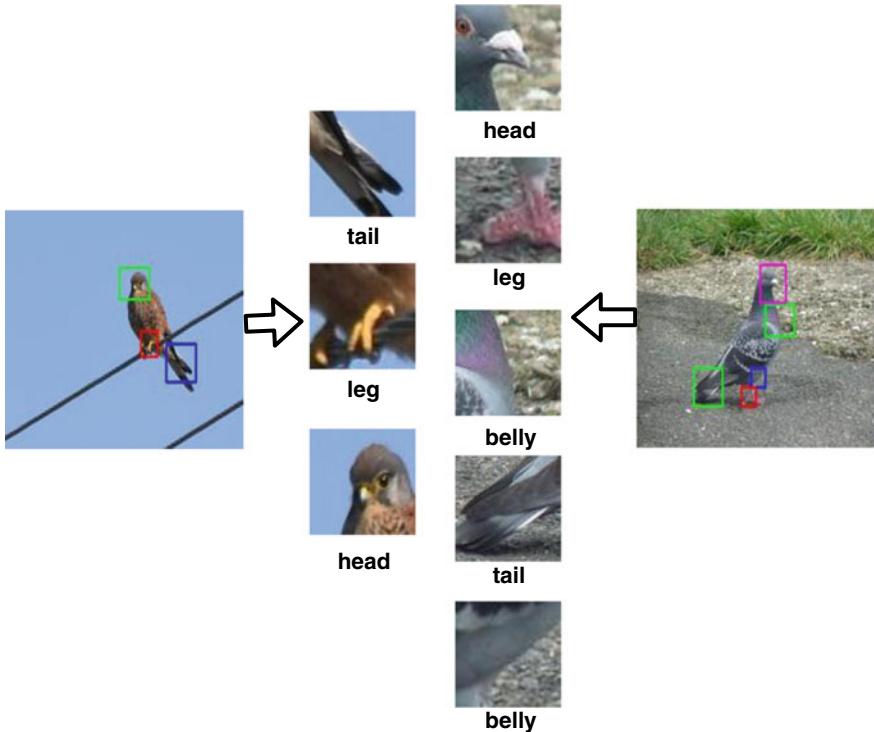
Note that in this case, we can afford to use all the remaining 40% data from the CUB dataset for fine-tuning, as its images are not used for evaluation. We believe that this is a useful result, which indicates an encouraging performance on a completely unseen dataset. Table 4 shows the classification score with Faster RCNN and proposed joint model for individual parts.



**Fig. 6** CNN part classifier predictions for region proposals in Fig. 5

**Table 4** Individual part classification accuracy for Faster RCNN and the proposed joint model (with 40% of CUB data for fine-tuning)

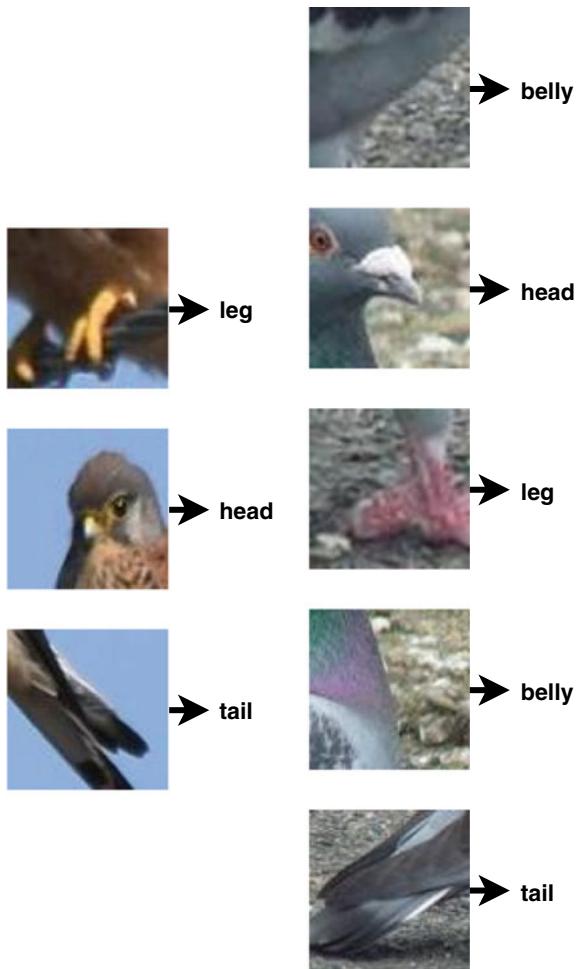
Bird part	Faster RCNN (%)	Proposed joint model (%)
Leg	84.5	86.7
Tail	83.3	85.8
Belly	79.2	82.6
Head	70.8	73.4



**Fig. 7** Examples for correct Faster RCNN region proposals on Mandi dataset

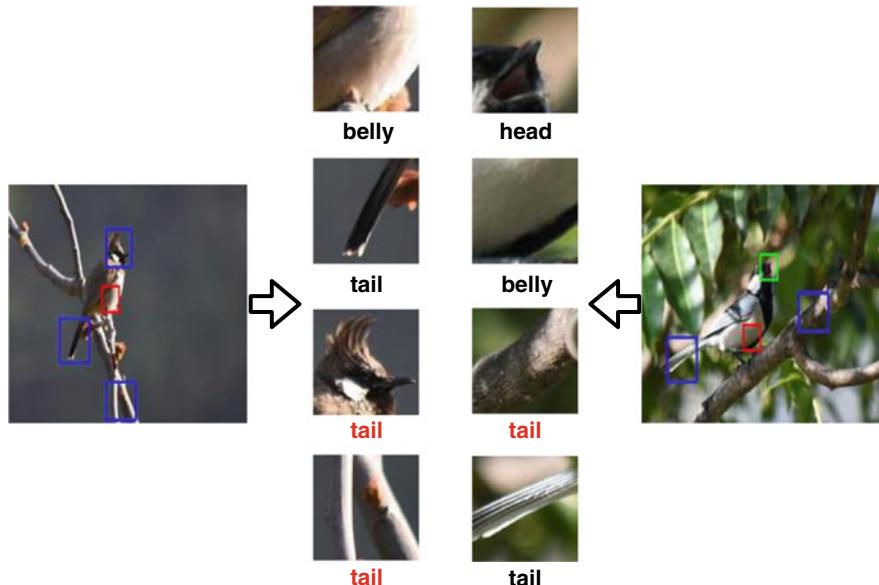
Figure 7 shows two examples for correct bird part region proposals from Faster RCNN on IIT Mandi dataset and Fig. 8 represents the corresponding CNN classifier predictions. Figure 9 shows two examples in which there are wrong bird part region proposals from Faster RCNN on IIT Mandi dataset and Fig. 10 represents the corresponding CNN classifier predictions. From Fig. 10, it is clear that the CNN classifier can correct some of the labels given by Faster RCNN model. This is due to the effect of including background in training the model. Correct predictions are marked in green color and wrong predictions in red color.

**Fig. 8** CNN part classifier predictions for region proposals in Fig. 7



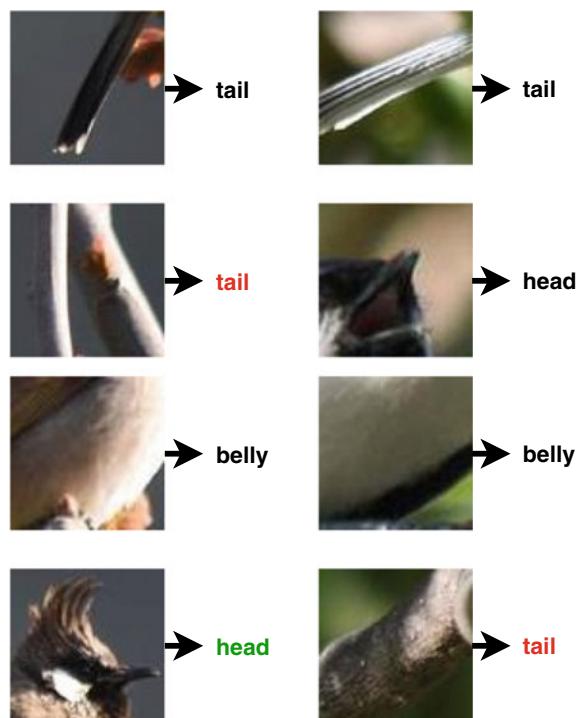
## 6 Conclusion

In this work, we proposed a joint approach for bird part localization, wherein the CNN-based part classifier augments the performance of the well-known FRCNN model. We devised a simple and efficient fine-tuning strategy of the CNN part classifier based on the FRCNN region proposals. Experimental results show that the proposed joint model approach can be effectively used for accurate bird part segmentation. We also demonstrate that the proposed approach generalizes well to an unseen dataset, which was not used during training or fine-tuning. Thus, we believe that the proposed joint framework can be used as an effective part extractor in fine-grained bird classification task.



**Fig. 9** Examples for wrong Faster RCNN region proposals on Mandi dataset

**Fig. 10** CNN part classifier predictions for region proposals in Fig. 9



## References

1. Zhang, X., et al.: Picking deep filter responses for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
2. Zhang, X., et al.: Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Trans. Image Process.* **25**(2), 878–892 (2016)
3. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: European Conference on Computer Vision. Springer, Berlin, Heidelberg (2006)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, CVPR 2005, vol. 1. IEEE (2005)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
6. Huang, S., et al.: Part-stacked CNN for fine-grained visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
7. Zhang, Y., et al.: Weakly supervised fine-grained categorization with part-based image representation. *IEEE Trans. Image Process.* **25**(4), 1713–1725 (2016)
8. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 IEEE International Conference on Computer Vision (ICCV). IEEE (2011)
9. Girshick, R., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
10. Girshick, R.: Fast R-CNN. arXiv preprint [arXiv:1504.08083](https://arxiv.org/abs/1504.08083) (2015)
11. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (2015)
12. Wah, C., et al.: The Caltech-UCSD birds-200-2011 dataset (2011)
13. Liu, J., Belhumeur, P.N.: Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In: 2013 IEEE International Conference on Computer Vision (ICCV). IEEE (2013)
14. Wah, C., et al.: Similarity comparisons for interactive fine-grained categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
15. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2013)
16. Lin, T.-Y., RoyChowdhury, A., Maji, S.: Bilinear CNN models for fine-grained visual recognition. arXiv preprint [arXiv:1504.07889](https://arxiv.org/abs/1504.07889) (2015)
17. Lin, Di, et al.: Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2015)
18. Elhoseiny, M., et al.: Link the head to the beak: zero shot learning from noisy text description at part precision. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
19. Liu, X., et al.: Fully convolutional attention networks for fine-grained recognition. arXiv preprint [arXiv:1603.06765](https://arxiv.org/abs/1603.06765) (2016)
20. Peng, Y., He, X., Zhao, J.: Object-part attention model for fine-grained image classification. *IEEE Trans. Image Process.* **27**(3), 1487–1500 (2018)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)

# Structural Analysis of Offline Handwritten Mathematical Expressions



Ridhi Aggarwal, Gaurav Harit and Anil Kumar Tiwari

**Abstract** Structural analysis helps in parsing the mathematical expressions. Various approaches for structural analysis have been reported in literature, but they mainly deal with online and printed expressions. In this work, two-dimensional, stochastic context-free grammar is used for the structural analysis of offline handwritten mathematical expressions in a document image. The spatial relation between characters in an expression has been incorporated so that the structural variability in handwritten expressions can be tackled.

**Keywords** Mathematical expression · Structural analysis

## 1 Introduction

Mathematics plays an essential role in scientific and technical documents. It provides an exact interpretation of ideas, theories, and conclusions in the form of formulas, equations, graphs, tables, and charts. Apart from this, numerals in the form of account number, date and amount on bank cheques, postal zip and phone numbers on address envelopes, etc. are present in day-to-day documents. Automatic recognition of these mathematical expressions has been an important topic of research in pattern recognition and document image analysis. The recognition of mathematical expressions in a document is substantially different from the recognition of normal text. The main challenges in handwritten mathematical expressions are as follows:

---

R. Aggarwal (✉) · G. Harit · A. K. Tiwari  
Indian Institute of Technology Jodhpur, Jodhpur, India  
e-mail: [pg201384012@iitj.ac.in](mailto:pg201384012@iitj.ac.in)

G. Harit  
e-mail: [gharit@iitj.ac.in](mailto:gharit@iitj.ac.in)

A. K. Tiwari  
e-mail: [akt@iitj.ac.in](mailto:akt@iitj.ac.in)

1. A mathematical expression has a complex two-dimensional structure. In normal text, the characters are systematically written from left to right while in mathematical expressions, the layout structure of characters is not fixed. Hence, the characters can be written in almost all the directions such as  $a + b$ ,  $a^b$ ,  $a_b$ ,  $\frac{a}{b}$ ,  $a^{b^2}$ ,  $a_{b^2}$ ,  $\sqrt{ab}$ ,  $\frac{a}{b^2}$ ,  $\sqrt{ab^2}$ ,  $\sqrt[3]{27}$ .
2. The mathematical expressions contain special symbols and Greek letters in addition to English letters and numbers.
3. In contrast to the printed expressions which are more regular and constrained due to uniform spacing between characters, the handwritten expressions depict nonuniformity in spacings between characters due to different writing styles of people and variations in the format of some expressions.
4. The mathematical expressions contain variant scales of math symbols, and hence the size of symbols in an expression is not fixed.
5. The interpretation of symbols in an expression depends on the context. For example, a horizontal line can be a fraction, a bar, or a minus operator depending on the location of the symbols.

The recognition process of a mathematical expression can be broadly divided into two major steps: symbol recognition and structural analysis. The purpose of symbol recognition is to separate all the characters in the numeral string so that they can be easily recognized and classified into the correct symbol category. Structural analysis deals with determining the relation among mathematical symbols in order to parse a complete mathematical expression. Our contribution in this work lies in making use of explicit modeling of the spatial relation between components using two-dimensional stochastic context-free grammar for structural parsing of offline handwritten mathematical expression. We have obtained improved results compared to those obtained when using a grammar that does not consider the variability in the spatial location of characters.

This paper consists of five sections. Section 2 presents a review of literature. Section 3 describes the context-free grammar. Section 4 describes our approach. The details of experiments along with analysis and discussion of results are reported in Sect. 5. Section 6 presents the concluding remarks.

## 2 Related Work

The purpose of structural analysis is to obtain the structure of an expression by determining the spatial relations between the symbols. For example, the “subscript” structure of expression “ $x_2 + 1$ ” is determined based on the symbol arrangements of “ $x$ ” and “ $2$ ”. The semantics of a mathematical expression are derived from the relative arrangement of the symbols. The common spatial relations found in an expression are subscript ( $x_2$ ), superscript ( $x^2$ ), vertical placement of symbols in fraction ( $\frac{x}{2}$ ), inside ( $\sqrt{x}$ ), horizontal ( $2x$ ), radicals ( $\sqrt[3]{64}$ ), left scripts ( $\mathcal{L}x$ ), and complex structure of matrices. In literature, various methods for structural analysis are used based on the type of parsing required.

In [5] and [15], recognition of mathematical expressions was attempted by using only the layout structures of symbols without actually parsing them. A symbol relation tree was built for each text line. The tree represents the relationships among the symbols in a mathematical expression which also reflects the structure of the expression. The main disadvantage of this approach is that it fails to detect the baseline due to slope variations in writing, and therefore some heuristics are used to correct the recognition error. In [7, 11, 20, 21], the white space between characters is used to segment the characters and symbols. The spatial relations between neighboring components are identified using a bottom-up process that facilitates merging and splitting of components in the expression tree. However, this bottom-up method based on the layout structure of an expression is sensitive to the skew of an input image. A small amount of skew changes the pixel density in horizontal and vertical directions, due to which the location of valley gets modified and affects the segmentation point. Also, the segmentation based on the histogram distribution of black pixels requires some other methods for characters containing multiple-connected components (e.g., j, i,  $\geq$ ,  $\leq$ ,  $\pm$ , etc.) and for characters which contain other characters within their zone (e.g.,  $\sqrt{}$ ). Hence, this projection profile approach fails to deal with subscript, superscript, and inclusion relationships as it becomes difficult to separate the characters. In [9] and [12], a number of penalty values for each spatial relation were calculated based on the possible different configurations of mathematical expressions in the target expression. Finally, the relation which attains the minimum penalty value is treated as the correct relation. This method is limited to simple expressions and often leads to misrecognition for deep-nested structure such as fractions in subscript terms, continued fractions, etc. In online handwritten expressions, the temporal alignment of symbol strokes along with the geometric properties of bounding box helps in the extraction of spatial alignment of symbols [14]. The arrangement of bounding box characterizes the spatial relation between components in an expression. However, various constraints are applied in this method such as user should write the expression in a certain order, i.e., from left to right and from top to bottom. Such constraint can easily be violated in real applications due to different writing styles. Also, this method is limited to simple expressions and higher order root operations, such as  $a^b$ , are not allowed due to size variability of symbols. All the above approaches require only layout structures of symbols.

Top-down parsing based approaches [4, 13, 17] involve first determining the high-level structures, i.e., subexpressions and then the low-level structures, i.e., symbols. The main advantage of this approach is that the global properties related to the structure of an expression are analyzed at an early stage. This helps in the interpretation of components in different spatial arrangements. For example, the symbol “ $\int$ ” at initial position in an expression “ $\int \cos x dx$ ” hypothesizes syntactic meaning of “ $dx$ ” which is different in “ $cx + dx$ ” expression. One of the earliest method for top-down parsing is syntax-directed parsing [4]. The syntax rules guide the partitioning of an input element set into subsets, and assign a syntactic goal to each of these subsets. This procedure is repeated until either all sub-goals have been satisfied or all possibilities have failed. However, this top-down approach is not very efficient because the partitioning strategy involves two non-terminal symbols on the right-hand side

of the production rules. Hence, each partition generates further more partitions and leads to high computational complexity.

Recognition of symbols and structure of an expression can also be done simultaneously by using grammar constraints. In stochastic context-free grammar (SCFG) [6] and [3], every production rule contains an associated probability. Hence, the grammar generates each output derivation with an associated probability. However, the output derivation with the maximum probability is selected as the final output. The probability of spatial relation (spr) between symbols in online handwritten expressions is defined by the grammar as  $Pr(A \xrightarrow{\text{spr}} BC)$  in [10, 19, 22], where  $A$  is a non-terminal symbol,  $\text{spr}$  denotes the spatial relation, and  $B, C$  are the regions in an image containing symbol. Hence, it represents the probability of spatial relations between regions  $B$  and  $C$ . For example, for an ideal horizontal relation, the vertical centers of bounding box for regions  $B$  and  $C$  align at the same horizontal line. Hence, the value of  $Pr(A \xrightarrow{\text{hor}} BC)$  is 1. Hence, the computation of spatial relation takes into account the writing order and the two-dimensional arrangement of symbols in an expression. For the structural analysis of printed expressions, the grammar-based approach defined in [22] is used along with the penalty factor associated with the occurrence of symbol in a rectangular area of the image. The image is divided into small regions and the occurrences of symbols are detected by using OCR tool in each region. The OCR tool detects the symbols using a variable size window which moves through the image and assigns a penalty if the window stores a particular symbol. The penalties associated with all the symbol occurrences help in determining the quality of the recognized symbol. For structural analysis of an expression image, the production rule which derives the component is recorded. Then based on the penalty values, these components are combined to produce larger components until the desired structure in an image is recognized successfully. The limitation of this method is the accuracy of the OCR tool. Not all symbols in an expression are separated efficiently by the rectangles and hence this affects the recognition results.

Another approach similar to SCFG is Fuzzy Relational Context-Free Grammar (Fuzzy r-CFG). In this method [8, 17], instead of probability  $Pr(A \xrightarrow{\text{spr}} BC)$ , fuzzy function  $r_T$  is associated with each production rule. In ambiguous situations, it constructs all the possible interpretations of the user's writing so that the intended or more reasonable interpretation can be selected. It provides the confidence score for each interpretation and the most reasonable interpretation is selected as the final output. Finally, all these interpretations are parsed by forest construction and the individual parse trees are extracted from the forest tree in decreasing ranked order. The score of the tree is computed by two types of relations: geometric relation [16, 18] between recognized symbols and spatial relation between terminal symbols. However, the computational complexity of this method is quite large at  $O(n^4)$ .

### 3 Context-Free Grammar

A context-free grammar is a 4-tuple

$$G = (V, \Sigma, R, S) \quad (1)$$

where

1.  $V$  is a finite set, whose elements are called *variables*.
2.  $\Sigma$  is a finite set, whose elements are called *terminals*. It makes the actual content of an expression.
3.  $S$  is an element of  $V$ ; it is called as *start variable*.
4.  $R$  is a finite set, whose elements are called rules. Each rule has the form  $A \rightarrow w$ , where  $A \in V$  and  $w \in (V \cup \Sigma)^*$ .

The grammar is specified by writing down its rules. The left-hand side of the rule is identified as the symbol variable. The terminal symbols are written only on the right-hand side. By convention, the start variable is the variable on the left-hand side of the first rule. These grammar rules implicitly use spatial relations.

Example:

```
< EXP > → < EXP > < SYM >
< EXP > → < SYM > < EXP >
< EXP > → < SYM > < SYM >
< EXP > → < OP > < SYM >
< SYM > → 0|1|2|3|4|5|6|7|8|9
< OP > → +|-|*|/
```

The above grammar rules are defined for a simple expression. It has 3 variables (the capitalized grammatical terms written inside brackets: EXP, SYM, OP), 10 terminals (0–9 digits), and 6 rules.

Let the string  $L(G)$  be:  $2^3 + 1$

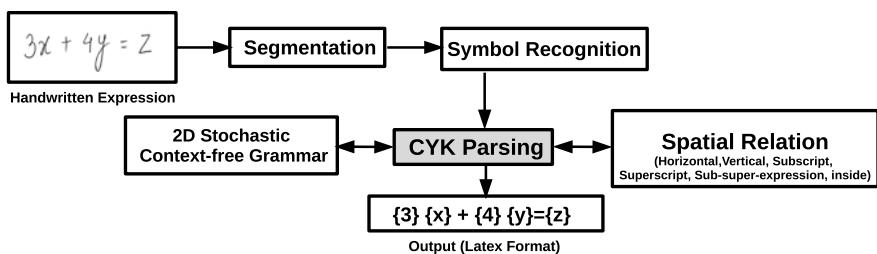
By using the above grammar rules, the derivation of this expression is as follows:

```
< EXP > → < EXP > < SYM >
< EXP > → < SYM > < SYM >
< SYM > → 2
< SYM > → 3
< EXP > → < OP > < SYM >
< OP > → +
< SYM > → 1
```

## 4 Methodology

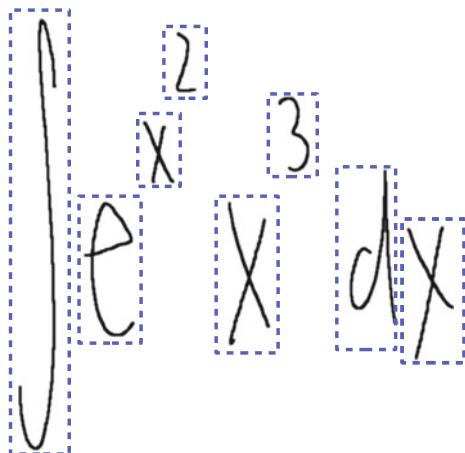
In this work, we have proposed a system based on 2D SCFG for recognition of offline handwritten mathematical expressions. The formulation of this work is motivated from existing work on recognizing printed mathematical expressions by using grammar rules [1]. The processing steps involved in our structural analysis framework are shown in Fig. 1.

The input to the system is the handwritten mathematical expression in image form. The first step is to segment the image into individual components. To segment the image, eight neighborhood-based connected components are computed for an expression as shown in Fig. 2. After doing the segmentation, we have a list of regions that can contain a mathematical symbol. For all these segmented regions, a mathematical symbol classifier is used to determine the class of each character. In our case, a recurrent neural network with offline feature is used to classify the handwritten math symbols [2]. Due to different writing styles and ambiguities, a character can belong to multiple symbol classes. Hence, the symbol recognition process classifies

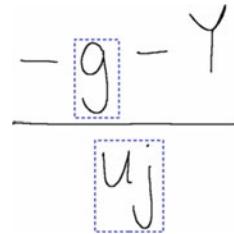


**Fig. 1** Block diagram of proposed work

**Fig. 2** Connected component of an expression



**Fig. 3** Handwritten expression



each terminal into several non-terminals because of the variations in handwriting. For example, in Fig. 3, the symbol in the numerator can be classified as digit “9” or letter “g”.

Finally, the most probable interpretation is decided by the parsing process.

For parsing offline handwritten mathematical expressions, a 2D SCFG grammar is defined to model all the expressions that appeared in our dataset. In [1], the production rules were defined for printed expressions. Motivated from [1] we have defined the production rules of the grammar for handwritten mathematical expressions in Sect. 3. Unlike printed expressions, the handwritten expressions do not contain uniform spacing between characters. Hence, the spatial relation between the components can be interpreted in multiple ways. Also, the expression written by the same writer can have symbols of different sizes. To overcome this challenge, we have proposed to update the spatial probabilities associated with the grammar rules to accommodate the handwriting variations. As shown in Fig. 3, the denominator can be interpreted as horizontal expression ( $uj$ ) or subscript expression ( $u_j$ ). Hence, the most likely expression is selected based on the probability associated with the spatial relationship between characters “ $u$ ” and “ $j$ ”. The spatial relation is modeled by using the geometric properties of the neighboring regions. The updated probability ( $\hat{Pr}$ ) assigned for the spatial relation ( $spr$ ) between two neighboring regions,  $R_1$  and  $R_2$ , is defined as follows:

$$\hat{Pr}(R_1, R_2, spr) \leftarrow Pr(R_1, R_2, spr) - \left( \frac{\Delta_v}{100} \right) \quad (2)$$

where  $R_1$  and  $R_2$  are the regions in which the symbol is present,  $Pr(R_1, R_2, spr)$  denotes the predefined spatial probability as formulated in [1], and  $\Delta_v$  is the deviation from the perfect spatial relation. Every relation such as horizontal, vertical, subscript, superscript, etc. has a predefined spatial probability [1]. The more the difference in the geometric properties of the two regions from predefined values, lesser will be the probability value associated with the spatial relation between them. For example, in a perfect horizontal relation, the vertical distance between the centers of two neighboring regions perfectly align, and hence  $\Delta_v = 0$ . As the two neighboring regions deviate from the predefined position, their vertical centers also shift. Hence,  $\Delta_v$  increases and the associated probability for horizontal relation decreases (Eq. 2).

There are total 21 non-terminals : Exp, Sym, ExpOp, OpUn, ROpUn, OPEExp, OpBin, OverExp, Over, OverSym, LeftPar, RightPar, RPExp, SSEExp, BigOpExp, BigOp, Sqrt, Func, 2Let, Let, and SupSym. The initial state must be either *Exp*

**Table 1** Terminal productions

Non-terminal	Terminal productions
Sym	Symbols like $\alpha, \beta, \chi, \gamma, \infty, \iota, \lambda, \omega, \kappa, \epsilon, \phi, \pi, \sigma, \nu$ etc.
Let	<b>52</b> : alphabets in capital letters and small letters
Over	Fraction line: –
BigOp	$\sum, \bigcup, \cap, \cup, f, \prod$
OpUn	–, ¬, +
ROpUn	Exclamation (!)
OpBin	\, , $\bigcup$ , $\cdot\cap$ , $\circ$ , : , $\cup$ , . , =, $\geq$ , $>$ , $\in$ , $\int$ , $\leq$ , $<$ , $\mapsto$ , –, $\neq$ , $\notin$ , $\oplus$ , $\otimes$ , $\perp$ , +, $\prod$ , $\rightarrow$ , ~, /, $\subset$ , $\subseteq$ , $\supset$ , $\times$ ,  , $\wedge$
OverSym	<i>check</i> ( $\check{x}$ ), <i>hat</i> ( $\hat{x}$ ), <i>overline</i> ( $\overline{x}$ ), <i>tilde</i> ( $\tilde{x}$ )
SupSym	*, !, /, +
LeftPar	[, (, {, <,
RightPar	], }, ),
Sqrt	square root ( $\sqrt{x}$ )

(e.g.,  $x^2 + y^2 + z^2$ ) or *Sym* (e.g.,  $\pi = 3.141$ ). The non-terminals: Sym, Let, Over, BigOp, OpUn, ROpUn, OpBin, OverSym, SupSym, LeftPar, RightPar, and Sqrt have terminal production rules (Table 1).

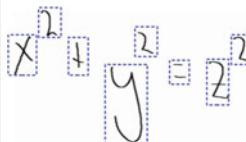
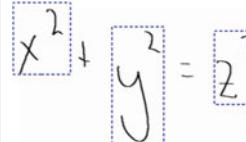
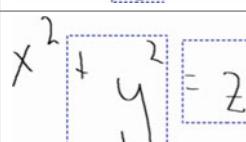
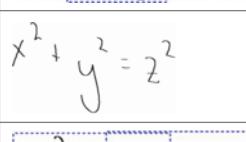
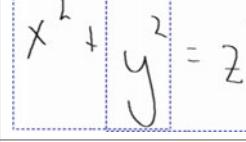
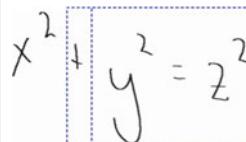
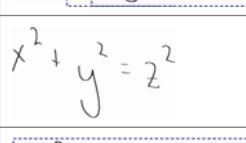
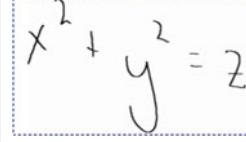
Finally, in order to build a complete structure of an expression, the spatial relation between two symbols is determined by using the Cocke–Younger–Kasami (CYK) algorithm. CYK is a parsing algorithm for context-free grammar. It is based on bottom-up dynamic programming. The CYK table is initialized by the number of connected components and their associated classifier probability (Table 2). Once the CYK parsing table is initialized, the algorithm builds new hypotheses of increasing size. The probability for merging two neighboring regions is computed as follows:

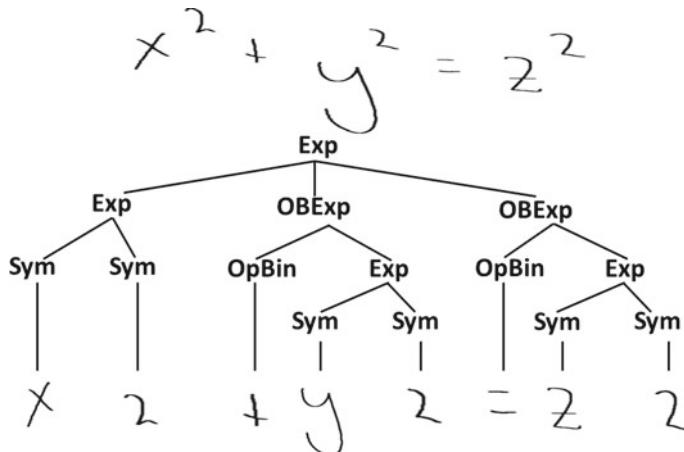
$$Pr(R_1) \cdot Pr(R_2) \cdot Pr(A \xrightarrow{\text{spr}} R_1 R_2) \quad (3)$$

where  $Pr(R_1)$  and  $Pr(R_2)$  denote the probability of symbols present in the region  $R_1$  and  $R_2$  belonging to the symbol class under consideration and  $Pr(A \xrightarrow{\text{spr}} R_1 R_2)$  denotes the probability associated with spatial relation between regions  $R_1$  and  $R_2$ . It is computed using Eq. 2.

The proposed system is illustrated in Table 2. The expression is segmented into individual components and each component has an associated class probability (refer t[1] in Table 2). To parse an expression, each region is combined to form a larger region by using the predefined grammar and spatial relation probability. The grouping of  $k$  terminal symbols is denoted by t[k]. In Table 2, regions containing four terminal symbols and seven terminal symbols cannot be formed, and hence t[4] and t[7] have  $\phi$  value. Finally, the recognized expression based on the maximum probable interpretation of symbols and the associated structure is obtained. The parse tree for the illustration presented in Table 2 is shown in Fig. 4.

**Table 2** An example of CYK parsing

t[1]: (Sym(2), $R_1$ , 0.92), (Sym(2), $R_2$ , 0.89), (Sym(2), $R_3$ , 0.89), (Sym(x), $R_4$ , 0.95), (Op(+), $R_5$ , 0.91), (Sym(y), $R_6$ , 0.97), (Op(=), $R_7$ , 0.98), (Sym(z), $R_8$ , 0.88)	
t[2]: $Exp \Rightarrow Sym \quad Sym (Exp(x^2), R_{14}, 0.79), (Exp(y^2), R_{26}, 0.76), (Exp(z^2), R_{38}, 0.75)$	
t[3]: $OBExp \Rightarrow OpBin \quad Exp (Exp(+y^2), R_{526} = R_5 + R_{26}, 0.55), (Exp(z^2), R_{738} = R_7 + R_{38}, 0.53)$	
t[4]: $=\phi$	
t[5]: $Exp \Rightarrow Exp \quad Exp (Exp(= x^2 + y^2), R_{14526} = R_{14} + R_{526}, 0.44), (Exp(y^2 = z^2), R_{26738} = R_{26} + R_{738}, 0.41)$	
t[6]: $OBExp \Rightarrow OpBin \quad Exp (Exp(+y^2), R_{526738} = R_{526} + R_{738}, 0.29)$	
t[7]: $=\phi$	
t[8]: $Exp \Rightarrow Exp \quad Exp (Exp(+y^2), R_{14526738} = R_{14} + R_{526738}, 0.02)$	

**Fig. 4** Parse tree

## 5 Results

Since any standard public dataset of offline handwritten mathematical expressions is not available yet, we use the CROHME 2016 dataset of Task-1 Formula Recognition from handwritten strokes. This dataset contains online handwritten mathematical expressions. We convert them into offline images by using the xy coordinates of the traces made by ink. To evaluate our technique, 8,836 expressions are used which are collected by CROHME organizers from five different databases: HAMEX, MfrDB, ExpressMatch, KAIST, and MathBrush. We have converted all these expressions from InkML files to offline images and labeled them.

To evaluate the performance of our structural analysis, we compute the accuracy of the obtained spatial relations. We compared our work with the grammar-based structural analysis method [1]. By updating the grammar production rules and the probability associated with the structural relation of characters in an expression, the proposed system is able to incorporate the handwriting variations. Hence, it performs well for the structural analysis of handwritten expression. A comparison of spatial relation parsing performance on offline handwritten expressions and printed expressions is presented in Table 3. The results of the proposed method are illustrated in Fig. 5. Since the spatial probability associated with the grammar rules are updated, the structure of a large number of writing variations can be easily parsed. However,

**Table 3** Spatial relation accuracy

Type of expressions	Spatial variability accounted (SVA) grammar-based method (%)	Grammar-based method (%)
Handwritten (offline)	<b>96.12</b>	89.12
Printed	<b>97.12</b>	95.32

$\left(\frac{a}{e}\right)$ <code>\frac{a}{e}</code>	$a^x + b^x + \frac{c}{2}$ <code>a^{x}+b^{x}+\frac{c}{2}</code>	$a + \frac{\sqrt{b+c}}{2}$ <code>a+\frac{\sqrt{b+c}}{2}</code>
$5.3P_X$ <code>5.3\{P\}_{x}</code>	$\int e^{x^2} x^3 dx$ <code>\int e^{x^2} x^3 dx</code>	$\frac{i+i^\beta}{[a]}$ <code>\frac{i+i^{\beta}}{[a]}</code>
$\frac{4}{3}\pi r^3$ <code>\frac{4}{3}\pi r^3</code>	$\int x^2 dx$ <code>\int x^2 dx</code>	$\frac{g+g}{x-y}$ <code>\frac{g+g}{x-y}</code>

**Fig. 5** Example results of proposed method: the handwritten expression(input) and the output in LATEX format

it must be mentioned that the trigonometric functions are not covered in our study. Hence, these functions can be interpreted as sin or individual characters *s,i* and *n*. Also, the ambiguity associated with handwriting is difficult to eliminate for similar looking symbols such as {X, x}, {O, 0, o}, {l, 1}, {g, 9}, etc.

## 6 Conclusions

In this paper, we have presented a stochastic context-free grammar-based structural analysis of offline handwritten mathematical expressions. The proposed system is robust to the spatial variability in the relative placement of symbols and operators. As part of future work, we will extend this system to other 2D structures such as tables, figures, charts, etc. in handwritten documents.

## References

1. Alvaro, F., Benedi, J.M., et al.: Recognition of printed mathematical expressions using two-dimensional stochastic context-free grammars. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), pp. 1225–1229. IEEE (2011)
2. Álvaro, F., Sánchez, J.A., Benedí, J.M.: Offline features for classifying handwritten math symbols with recurrent neural networks. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 2944–2949. IEEE (2014)
3. Álvaro, F., Sánchez, J.A., Benedí, J.M.: Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden Markov models. *Pattern Recognit. Lett.* **35**, 58–67 (2014)
4. Anderson, R.H.: Two-dimensional mathematical notation. In: Syntactic Pattern Recognition, Applications, pp. 147–177. Springer (1977)
5. Chen, L.: A system for on-line recognition of handwritten mathematical expressions. *Comput. Process. Chin. Orient. Lang.* **6**(1), 19–39 (1992)
6. Chou, P.A.: Recognition of equations using a two-dimensional stochastic context-free grammar. In: Visual Communications and Image Processing IV, vol. 1199, pp. 852–866. International Society for Optics and Photonics (1989)
7. Faure, C., Wang, Z.X.: Automatic perception of the structure of handwritten mathematical expressions. In: Computer Processing of Handwriting, pp. 337–361. World Scientific (1990)
8. Fitzgerald, J.A., Geiselbrechtinger, F., Kechadi, M.T.: Structural analysis of handwritten mathematical expressions through fuzzy parsing. *ACST* **6**, 151–156 (2006)
9. Fukuda, R., Sou, I., Tamari, F.: A technique of mathematical expression structure analysis for the handwriting input system. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, ICDAR'99, pp. 131–134. IEEE (1999)
10. Garain, U., Chaudhuri, B.B.: Recognition of online handwritten mathematical expressions. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **34**(6), 2366–2376 (2004)
11. Ha, J., Haralick, R.M., Phillips, I.T.: Understanding mathematical expressions from document images. In: Proceedings of the Third International Conference on Document Analysis and Recognition, 1995, vol. 2, pp. 956–959. IEEE (1995)
12. Inoue, K., Miyazaki, R., Suzuki, M.: Optical recognition of printed mathematical documents. In: Proceedings of the Third Asian Technology Conference on Mathematics, pp. 280–289 (1998)
13. Julca-Aguilar, F., Mouchère, H., Viard-Gaudin, C., Hirata, N.S.: Top-down online handwritten mathematical expression parsing with graph grammar. In: IberoAmerican Congress on Pattern Recognition, pp. 444–451. Springer (2015)
14. Kosmala, A., Rigoll, G.: On-line handwritten formula recognition using statistical methods. In: Proceedings of the Fourteenth International Conference on Pattern Recognition, 1998, vol. 2, pp. 1306–1308. IEEE (1998)
15. Lee, H.J., Wang, J.S.: Design of a mathematical expression understanding system. *Pattern Recognit. Lett.* **18**(3), 289–298 (1997)
16. Liang, P., Narasimhan, M., Shilman, M., Viola, P.: Efficient geometric algorithms for parsing in two dimensions. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition, 2005, pp. 1172–1177. IEEE (2005)
17. MacLean, S., Labahn, G.: A new approach for recognizing handwritten mathematics using relational grammars and fuzzy sets. *Int. J. Doc. Anal. Recognit. (IJDAR)* **16**(2), 139–163 (2013)
18. Miller, E.G., Viola, P.A.: Ambiguity and constraint in mathematical expression recognition. In: AAAI/IAAI, pp. 784–791 (1998)
19. Mitra, J., Garain, U., Chaudhuri, B., HV, K.S., Pal, T.: Automatic understanding of structures in printed mathematical expressions. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 540–544. IEEE (2003)

20. Okamoto, M.: Recognition of mathematical expressions by using the layout structure of symbols. In: Proceedings of the 1st International Conference on Document Analysis and Recognition, 1991, pp. 242–250 (1991)
21. Okamoto, M., Miyazawa, A.: An experimental implementation of a document recognition system for papers containing mathematical expressions. In: Structured Document Image Analysis, pp. 36–53. Springer (1992)
22. Yamamoto, R., Sako, S., Nishimoto, T., Sagayama, S.: On-line recognition of handwritten mathematical expressions based on stroke-based stochastic context-free grammar. In: Tenth International Workshop on Frontiers in Handwriting Recognition. Suvisoft (2006)

# L1-Regulated Feature Selection and Classification of Microarray Cancer Data Using Deep Learning



B. H. Shekar and Guesh Dagnew

**Abstract** The microarray cancer data obtained through microarray technology poses a lot of challenges during classification since the sample size is very small and the dimensionality of the data is very high. It is noticed that usually, the number of classes in multiclass datasets are highly imbalanced. In order to reduce the dimensionality thereby enabling accurate classification, in this work, we propose an L1-regulated feature selection and deep learning is applied for classification. The L1-regulated feature selection is based on Linear Support Vector Machine (LSVM) which is characterized by adding a penalty term to the prediction error in order to reduce the weight of the irrelevant features and to make the relevant features having nonzero weights. For classification purpose, deep learning neural network is initialized with sigmoid activation function in the input and hidden layers and to accommodate multiclass classification, the softmax activation function is used in the output layer. In order to demonstrate the suitability of the proposed approach, experiments are conducted on the six numbers of standard multiclass cancer datasets and to argue the predictive capability of the proposed approach, experiments are conducted on imbalanced class datasets such as 5-class lung cancer dataset, and 4-class Leukemia cancer dataset. Comparative study is also provided with state-of-the-art approaches and the results are presented considering classification accuracy, precision, recall, f-measure, confusion matrix, average precision, and ROC metrics to exhibit the performance of the proposed approach.

**Keywords** L1-regularization · Feature selection · Deep learning · Classification

---

B. H. Shekar · G. Dagnew (✉)

Department of Computer Science, Mangalore University, Mangalore, India  
e-mail: [guesh.nanit@gmail.com](mailto:guesh.nanit@gmail.com)

B. H. Shekar  
e-mail: [bhshekar@gmail.com](mailto:bhshekar@gmail.com)

## 1 Introduction

Due to the innovation of the microarray technology, a massive amount of microarray cancer data is produced. Microarray cancer data is characterized by its big curse of dimensionality, small sample size, and imbalanced classes in the multiclass problem. Consequently, a wide multidisciplinary research area is opened in the disciplines such as Genomic studies, Computational Biology, Statistics and Machine learning. Conducting a research in the domain of microarray cancer data has significances such as diagnosis of cancer patients, identification, and differentiation among cancer types [14, 16, 23].

Very recently, several feature selection and classification approaches are proposed by different researchers. Garro et al. [7] proposed artificial bee colony based feature selection. Chen et al. [4] introduced particle swarm and decision tree based feature selection and variants of ridge regression methods for classification. Aziz et al. [2] proposed a combination of fuzzy backward feature elimination and independent component analysis for feature selection. Nguyent et al. [18] introduced an aggregate feature selection method based on statistical ranking methods. Feature selection based on game theory is introduced by Sasikala et al. [20]. Moayedikia et al. [15] introduced symmetric uncertainty and harmony search based feature selection. Sharabaf et al. [21] proposed filter-based feature selection. According to Ravi et al. [19], the deep learning is playing major role in optimizing many of the algorithms in Biomedical and Health Informatics. In our work, to handle the curse of dimensionality with respect to the small sample size, we introduce the L1-regulated feature selection followed by classification of multiclass microarray cancer data using deep learning techniques.

Six standard multiclass microarray cancer datasets are used in our experimentation and the results indicate that the proposed approach is comparable to many of the state-of-the-art works. To evaluate the performance of the model, classification accuracy, confusion matrix, average precision, and ROC curve are used.

The rest of the paper is outlined as follows. Section 2 discusses recent related works in the domain. Section 3 deals with the proposed methodology and Sect. 4 describes the datasets, Sect. 5 is about the evaluation methods. Section 6 presents the experimental results and discussion. Finally, the conclusion is drawn in Sect. 7.

## 2 Related Works

In this section, state-of-the-art works in the domain of microarray cancer data are described with respect to feature selection and classification. Numerous feature selection and classification methods for microarray cancer data are investigated and still there is a challenge with respect to the small sample size, the high curse of dimensionality, and imbalanced class problem.

Kar et al. [9] proposed Particle Swarm Optimization (PSO) with adaptive KNN based gene selection, whereas Chen et al. [4] proposed an integration of PSO and

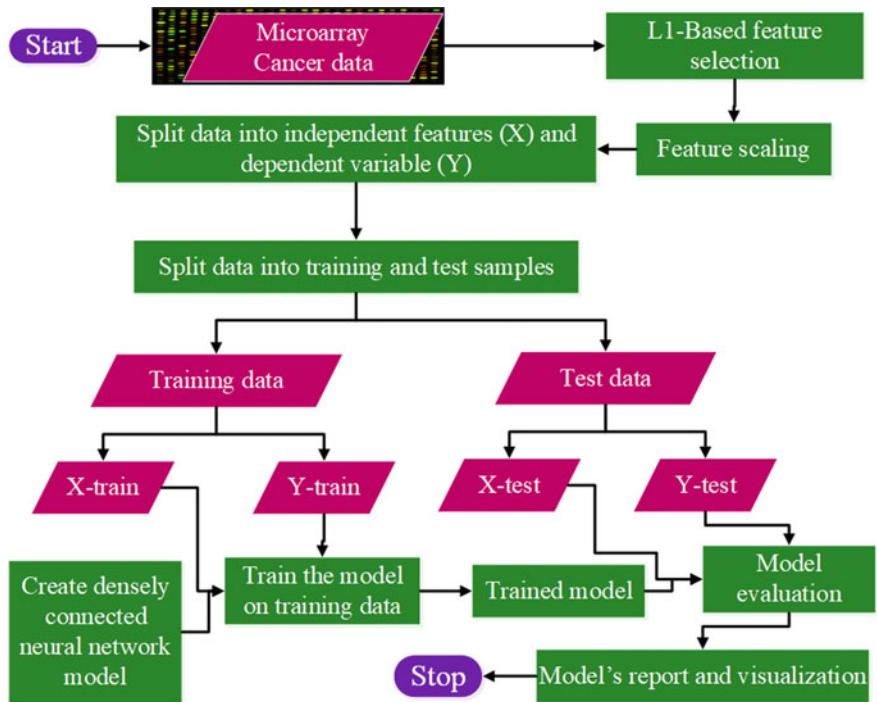
C4.5 based feature selection. Tabakhi et al. [22] proposed an unsupervised gene selection method called MGSACO, which incorporates the Ant Colony Optimization (ACO) algorithm into the filter approach. Guo et al. [8] proposed two-step L1-regularization method for classification of microarray data. A statistical test ANOVA based on MapReduce is proposed by Kumar et al. [10] to select the relevant features. An ensemble approach based on Maximum Relevancy and Minimum Redundancy (MRMR) based feature selection method using Hesitant Fuzzy Sets (HFSs) is proposed by Ebrahimpour and Eftekhari [5]. Al-Rajab et al. [1] introduced a three-phase approach which includes feature detection, classification, and performance evaluation. Medjahed et al. [14] proposed a complete cancer diagnostic process through kernel-based learning and feature selection. Mollae and Moattar [17] proposed a hybrid ensemble approach for cancer diagnosis and classification.

### 3 Proposed Methodology

Motivated from [8] which uses L1-based regularization with logistic regression, we propose an SVM-based L1-regulated feature selection in microarray cancer data for classification purpose. The method uses linear SVC with parameters  $\alpha$  regularizer which controls the sparsity of the data. The smaller the value of  $\alpha$  in linear SVC, the fewer features selected. A fully connected neural network architecture is created along with its parameters such as the number of epochs, batch size, and the activation function which are initialized to sigmoid in the input and hidden layers of the network. Since all the datasets considered in this work are multiclass, the softmax activation function is initialized in the output layer to yield the ratio of the probability of a given class to the summation of all classes in a given dataset. Moreover, the random state is initialized to a constant seed value 7 so as to keep results of the model reproducible. The model is trained on the basis of fivefold cross-validation. The workflow of the proposed model is shown in Fig. 1, which shows the steps carried out mainly for feature selection, splitting of the data into training and test data, model creation, evaluation, and visualization are described in the following subsection.

#### 3.1 L1-Regularized Feature Selection

Regularization is an important technique to control model complexity and feature selection in machine learning and artificial neural networks. In this work, L1-regularizer feature selection method is employed for feature selection. It applies shrinking strategy by adding penalizing term to the least square errors in a linear regression and hence to assign zero coefficient to the irrelevant features to discard them from the model. Only nonzero coefficient variables are considered so as to minimize prediction error by tackling overfitting and simplified the model complexity and computationally stability. L1-regularization is a method which helps in reducing the complexity of the model by adding a penalty term (See Eq. 1) to the least square



**Fig. 1** Workflow of the model

errors (See Eq. 2) and shrinking some of the weights of data points so as to overcome overfitting. It allows removing certain features from the model if they are not helpful in training the model.

$$p = \alpha \sum_{i=1}^d |w| \quad (1)$$

where  $p$  is the magnitude of the penalty,  $d$  is the dimension of the features,  $\alpha$  is the control parameter, and  $w$  is the weight of each feature.

$$E(w) = \sum_j^n \left( y_j - \sum_{i=0}^d w \cdot x_i \right)^2 \quad (2)$$

L1-regularization produces sparse coefficients. For example, the 3-class Leukemia dataset in this paper has 7130 features but only 27 features which are nonzero coefficients are selected and the rest of the features are discarded as a result of shrinking of their coefficients to zero.

L1-regularized feature selection is the newly introduced feature selection method to the microarray data analysis. It works with a classifier model such as SVMs and logistic regression to select an optimal number of features. L1-based feature selection

uses LSVM to fit the data and returns the best fit hyperplane that divides the data into categories. It uses local optimal solutions to remove features with zero coefficients. It uses the control parameter  $\alpha$  to control the sparsity of the data. Sparsity allows few features in a matrix to have large nonzero coefficient values. As the control parameter  $\alpha$  gets maximum value such as  $\alpha = 0.1, \alpha = 0.01$ , many features are selected and as the value of  $\alpha$  become small such as  $\alpha = 0.001, \alpha = 0.0001$  up to a certain limit which gives the optimal features [3, 6, 8, 25].

$$E(w) = \sum_j^n \left( y_j - \sum_{i=0}^d w \cdot x_i \right)^2 + \alpha \sum_{i=1}^d |w| \quad (3)$$

where  $E$  is the error,  $w$  is a weight coefficient,  $y$  is the label,  $x$  is the input features,  $\alpha$  is the controller parameter,  $d$  is the dimension of features, and  $n$  is number of samples.

### **3.2 Fully Connected Neural Network Method for Multiclass Microarray Cancer Data Classification**

In this work, a fully connected neural network model is proposed. The model contains the conventional structure of the neural network, namely, input, hidden, and output layers, except it is deep in the sense number of hidden layers goes up to five layers to make it deep learning.

The parameters such as the number of nodes in the input layer which is set equal to the number of features in the input data are used. The Kernel initializer parameters are set to the normal distribution to show if data are normally distributed. The kernel at each layer is initialized as a normal distribution and activation functions are sigmoid in the input and hidden layers and softmax in the output layer. Since multiclass classifier is working in terms of numerous binary classifiers in the hidden layers it is sufficient to assign sigmoid activation function in the hidden layer and due to the reason that all intermediate binary classifiers are integrated at the output layer, softmax is initialized to accommodate the multiclass behavior of the data. It shall be noted here that the number of neurons at the output layer is equal to the number of classes in the dataset.

The argmax function is applied to get the predicted classes. Argmax checks the input values of each class and picks the maximum probability as an index of the particular class.

To train and test the model, the fivefold cross-validation (CV) technique is applied which divides the full dataset into five folds. In a CV method, the data is nearly evenly distributed among the folds and each fold has a chance of being test data which helps the model to overcome overfitting. Fold-1 is a test case in the first phase, fold-2 in the second phase, and fold-k is a test case in the  $k$ th phase. The ultimate goal of this technique is to create out-of-sample prediction in fivefolds CV and five neural network models.

The proposed model takes an input vector  $x = [x_1, x_2, x_3 \dots x_n]$  and each input vector is multiplied by its corresponding weight. Hence, the weight vector for the input data is represented as  $w = [w_1, w_2, w_3 \dots w_n]$  and the bias  $b$  is added to the weighted input vectors. In each layer of the model, the weighted input vectors are multiplied by the sigmoid activation function to yield the intermediate probabilistic results in the hidden layers based on Eq. 4.

$$y_i = f \left( \sum_{i=1}^n w_i \cdot x_i + b \right) \quad (4)$$

where  $y_i$  is the dependent variable to be predicted,  $w_i$  are the weight matrix and  $x_i$  are feature vectors, and  $f$  is the sigmoid activation function based on Eq. 5.

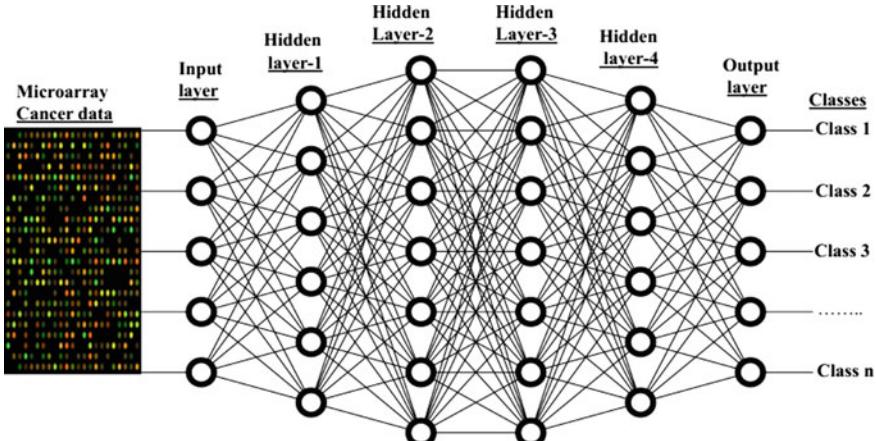
$$f(x) = \frac{1}{1 + e^{-(x_i w_i)}} \quad (5)$$

The softmax activation function is applied in the output layer to predict the class of each sample to yield the final prediction result as shown in Eq. 6.

$$P(y = j|x) = \frac{e^{x^T w_j}}{\sum_{k=1}^K e^{x^T w_k}} \quad (6)$$

where  $k \in \{1, \dots, K\}$  ranges over all the classes and  $x^T w_i$  stands for the inner product of each feature and corresponding weight.

As shown in Fig. 2, selected features of microarray cancer data are fed to the input layer. Once the data passes to the hidden layer, the model understands the relationship between each feature and their respective class labels. The model uses



**Fig. 2** Model architecture

**Table 1** Dataset description

Dataset	Sample size	Number of features	Number of classes
Leukemia_3	72	7129	3
Leukemia_4	72	7129	4
SRBCT_4	83	2308	4
MLL_3	72	7129	3
Tumor_11	174	12534	11
Lung_5	203	12600	5

the one-versus-rest approach to classify the data in the hidden layer using sigmoid activation function. The result of hidden layer is integrated into the output layer using the softmax activation function to give the prediction class levels.

## 4 Dataset Description

In this section, the detailed description of datasets used in this work is presented. Five publicly available multiclass microarray cancer datasets downloaded from Shenzhen University data repository [26] are used in this work and their description is shown in Table 1. Three of the datasets are from Leukemia cancer family. Leukemia-4 is 4-class dataset and the remaining two Leukemia are three class datasets, namely, Leukemia-3 and Mixed-Lineage Leukemia (MLL) but with different class names. The Small Round Blue Cell Tumors (SRBCTs) dataset is a childhood tumor having four classes and Tumor 11-class dataset belongs to various human tumor types and the lung cancer dataset with five classes is also used.

## 5 Evaluation Metrics

Every performance metric has its own pros and cons. To alleviate this constraint, performance measures such as accuracy, recall, precision, f-score, average precision, confusion matrix, micro-average, and macro-average ROC are used.

$$\text{Macro-Avg-ROC} = \frac{AUC_{c1} + AUC_{c2} + \dots + AUC_{Cn}}{n} \quad (7)$$

$$AP = \sum_n (R_n - R_{n-1}) \times P_n \quad (8)$$

where  $R_n$  and  $P_n$  are the recall and precision at the threshold.

The accuracy of the model is computed by considering the ratio of the sum of  $TP$  and  $TN$  to the total sample size in the test data. Equation 7 shows macro-average of the ROCs of each class which assigns the same weight to each label where AUC

is the area under the curve for each class and  $c$  is a class and  $n$  is the number of classes. Average precision, which is the area under the curve of precision–recall curve, compares the recall on the x-axis and precision on the y-axis at each threshold (See Eq. 8), where the ideal average precision of a model is at the top right corner of a plot.

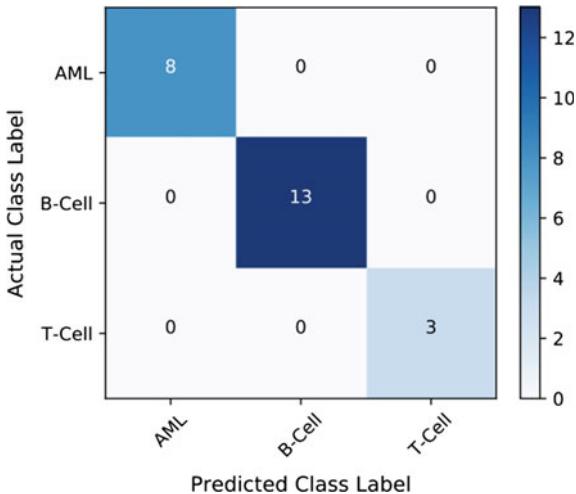
## 6 Experimental Results and Discussion

In this section, detailed discussion of experimental results is covered. Extensive experiments are conducted on the six standard multiclass microarray cancer datasets and results are presented in Table 2. The fivefold cross-validation method is employed to evaluate the performance of the proposed model. The average of the fivefold cross-validation gives the classification accuracy along the standard deviation which measures the deviation of the predicted class from the mean of the classification accuracy on the fivefold evaluation results. 100% classification accuracy is achieved on three of the datasets, namely, Leukemia 3-class, SRRCT, and MLL and the standard deviation for these particular datasets is zero as the model performs perfection in all the fivefold iterations. An accuracy of 93.09% and 93.57% is achieved on tumor and lung cancers data with a standard deviation of 2.94% and 2.57%, respectively, as shown in Table 2.

**Table 2** Classification report of the model: fivefold CV, average accuracy of the fivefold and std, average precision, micro- and macro-average AUC

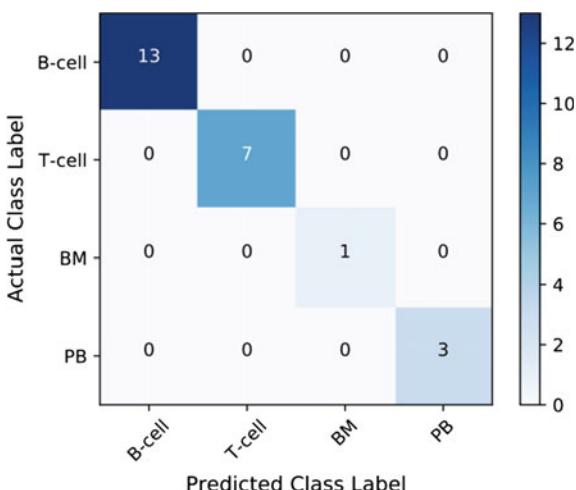
Datasets	Selected features	Accuracy of fivefold CV	Accuracy and std	AP	Mic-avg AUC	Mac-avg AUC
Leukemia_3	27	[1, 1, 1, 1, 1]	100.00% (0.00%)	1.00	1.00	1.00
Leukemia_4	37	[1, 1, 1, 0.93, 1]	98.57% (2.86%)	1.00	1.00	1.00
SRRCT	53	[1, 1, 1, 1, 1]	100.00% (0.00%)	1.00	1.00	1.00
MLL_3	31	[1, 1, 1, 1, 1]	100.00% (0.00%)	1.00	1.00	1.00
Tumor	132	[0.97, 0.94, 0.89, 0.94, 0.91]	93.09% (2.94%)	0.93	0.96	0.95
Lung	38	[0.93, 0.98, 0.95, 0.90, 0.93]	93.57% (2.57%)	0.92	0.97	0.95

**Fig. 3** Confusion matrix for 3-class Leukemia

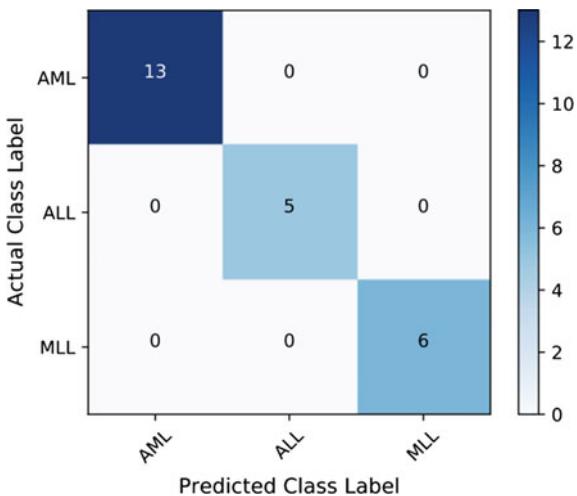


Results of the proposed model are also evaluated using confusion matrix. In a multiclass classification, the confusion matrix contains  $n^2$  rows and columns where  $n$  is the number of classes. In a confusion matrix, correctly classified samples are presented along the diagonal line and wrongly classified samples are located off-diagonal positions of the confusion matrix. As shown in Figs. 3, 4, 5, and 6, all samples in each class are correctly classified as they are presented in the diagonal line and the rest of the diagonal elements of the confusion matrix are showing zeros to indicate no element was wrongly classified. In the case of Tumor 11-class dataset, which has 58 elements in the test data, 54 of the elements are correctly classified and are presented along with the diagonal and the wrongly classified four elements are

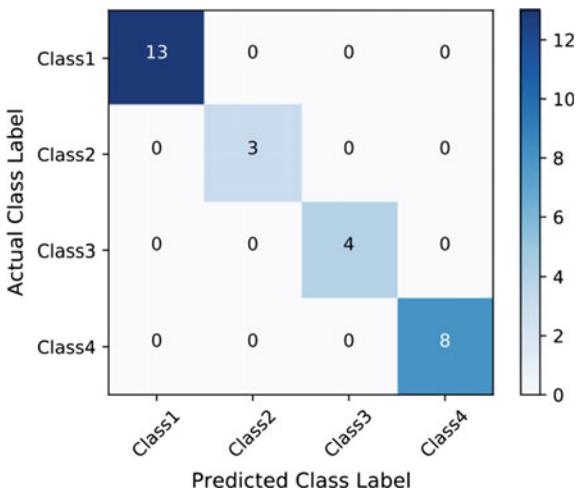
**Fig. 4** Confusion matrix for 4-class Leukemia



**Fig. 5** Confusion matrix for MLL



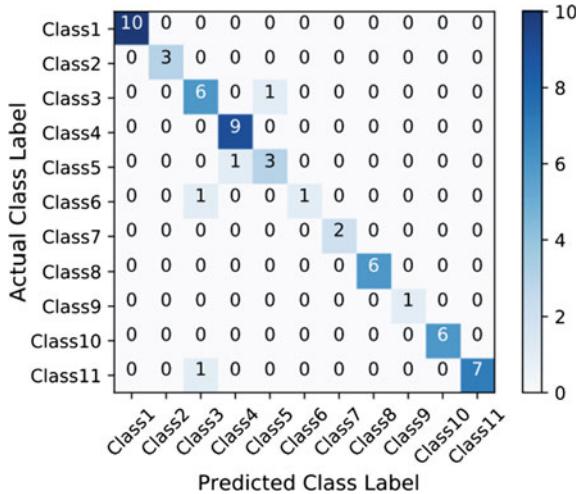
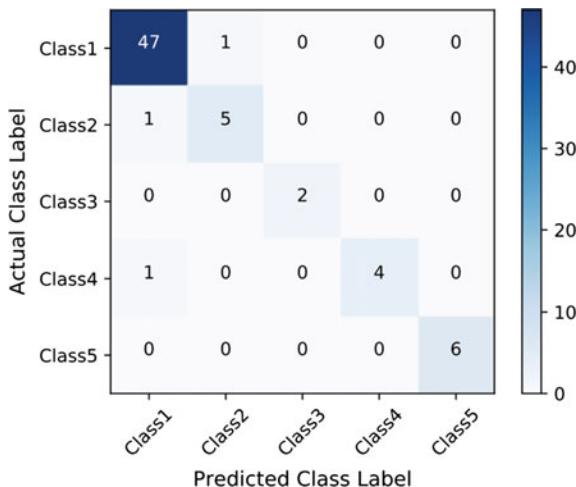
**Fig. 6** Confusion matrix for SRBCT cancer



distributed in the off-diagonal positions and this is shown in Fig. 7. For lung cancer, the confusion matrix shows 67 samples which are correctly classified out of 70 test samples as shown in Fig. 8. The vertical lines associated to each confusion matrix indicate the elements in the class with a maximum number of elements.

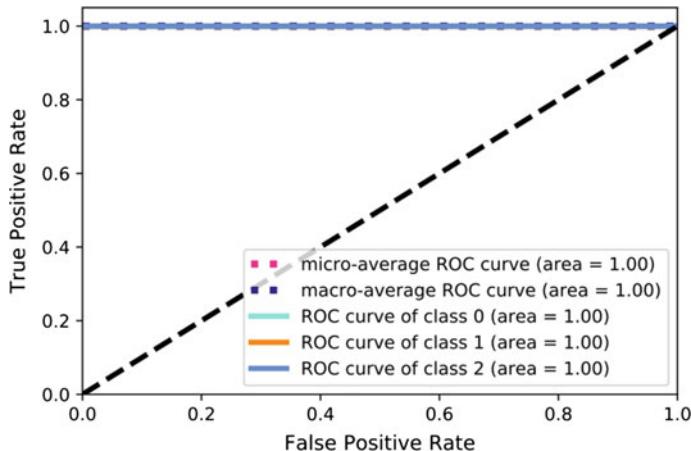
The ROC curve represents the True Positive Rate (TPR) on the y-axis and False Positive Rate (FPR) on the x-axis. The top left corner of the plot is the ideal point to be achieved which indicates an FPR of zero on the y-axis and TPR of 1 on the x-axis. The closer the graph to the top left corner, the larger Area Under the Curve (AUC) and the ROC curve of our work is presented in Figs. 9, 10, 11, 12, 13, and 14.

Average precision of the proposed model is considered and the results indicate 100% performance on four of the datasets, namely, Leukemia 3-class, Leukemia

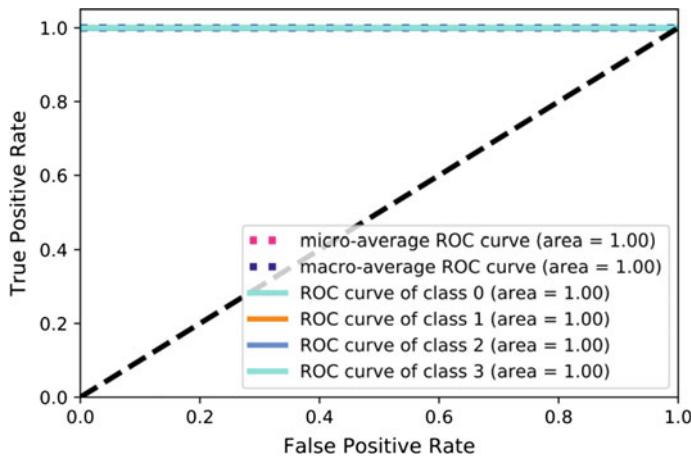
**Fig. 7** Confusion matrix for tumor cancer**Fig. 8** Confusion matrix for lung cancer

4-class, SBRCT, and MLL. The model scores 0.93 average precision and recall on Tumor 11-class dataset. Macro-average of the ROC is computed as shown in Eq. 7.

Moreover, the proposed model is evaluated in using precision, recall and f-measure, and results indicated that 100% precision, recall, and f-measure is achieved on three datasets, namely, *Leukemia*<sub>3</sub>, SRBCT, and *MLL*<sub>3</sub>. Leukemia 4-class dataset, the precision, recall, and f-measure achieved is 0.98. In the case of Leukemia 4-class, Tumor and lung cancer datasets, 0.98, 0.95, and 0.94 precision, recall, and f-measure is achieved and this is presented in Table 3.

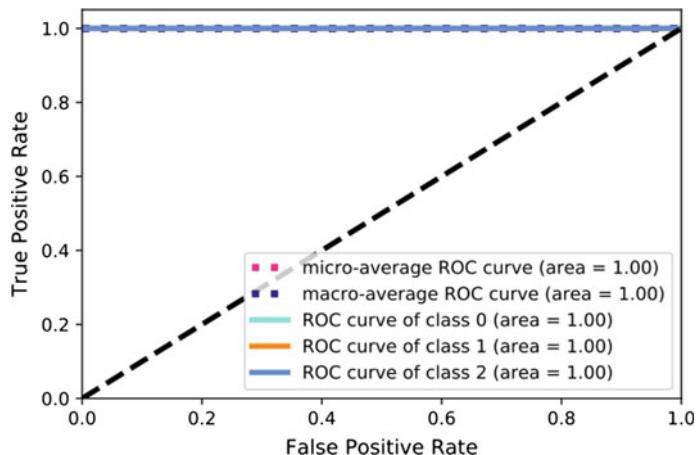


**Fig. 9** ROC and AUC for 3-class Leukemia

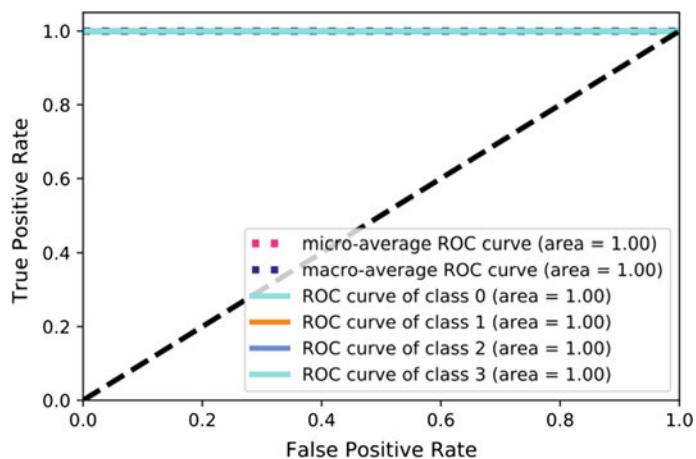


**Fig. 10** ROC and AUC for 4-class Leukemia

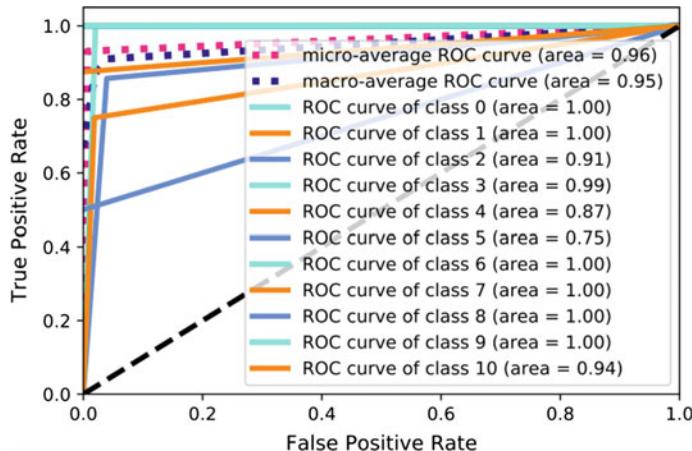
Comparative study, in terms of classification accuracy, is carried out and better results are achieved in four of the datasets, namely, *Leukemia\_3*, *Leukemia\_4*, *SRBCT*, and *MLL*. In the case of *Lung\_5* and *Tumor<sub>11</sub>*, the proposed model scores 93.57% and 93.09% and we observe that the works of Liu et al. [12] and Wang et al. [24] scored an accuracy of 95% and 99.34%, respectively, as presented in Table 4.



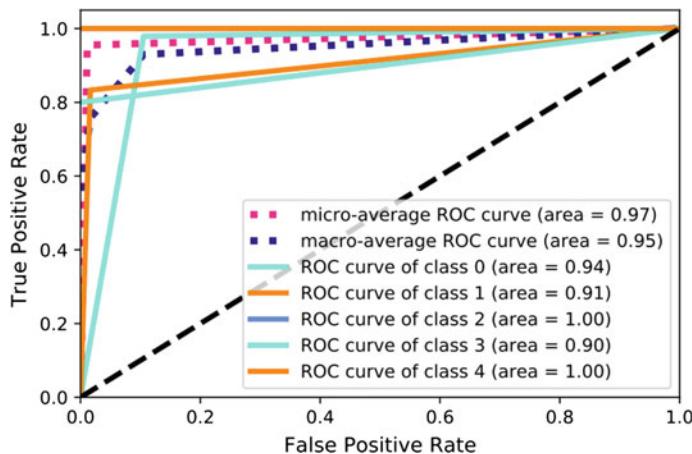
**Fig. 11** ROC and AUC for MLL class



**Fig. 12** ROC SRBCT



**Fig. 13** ROC and AUC for tumor cancer



**Fig. 14** ROC and AUC lung cancer

**Table 3** Precision, recall, and F-measure on all the datasets

Dataset	Precision	Recall	F-measure
Leukemia_3	1.00	1.00	1.00
Leukemia_4	0.98	0.98	0.98
SRBCT	1.00	1.00	1.00
MLL_3	1.00	1.00	1.00
Tumor_11	0.95	0.95	0.95
Lung_5	0.94	0.94	0.94

**Table 4** Comparative study of the proposed method with other related works

Authors	Method	Dataset					
		Leukemia_3	Leukemia_4	SRBCT	MLL	Tumor_11	Lung
Liu et al. [12]	WELM	–	–	99		<b>95</b>	97
Lin et al. [11]	GASS	–	–	100		–	–
Lv et al. [13]	MOEDA	100	–	91	–	–	88
Wang et al. [24]	BCO	–	–	100	–	89.62	<b>99.34</b>
<b>Our work</b>	<b>LFSDL</b>	<b>100</b>	<b>98.57</b>	<b>100</b>	<b>100</b>	93.09	93.57

## 7 Conclusion

To address the challenges, such as the curse of dimensionality, small sample size, and imbalanced class, L1-regulated feature selection based on the linear SVC model with parameters,  $\alpha$  regularizer which controls the sparsity of the data and deep learning classification method is introduced. The fully connected neural network architecture is created along with its parameters such as a number of epochs, batch size, and the activation function which is initialized to sigmoid in the input and hidden layers of the network. Experiments on six standard multiclass microarray cancer datasets show that the proposed method is effective enough in predictive capability even in the case of highly imbalanced class datasets such as Lung and Leukemia 4-class datasets. Comparative study of our method, using the standard metrics, with some selected works shows that our method achieves a better result in four of the datasets.

## References

1. Al-Rajab, M., Joan, L., Qiang, X.: Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Comput. Methods Programs Biomed.* **146**, 11–24 (2017)
2. Aziz, R., Verma, C.K., Srivastava, N.: A fuzzy based feature selection from independent component subspace for machine learning classification of microarray data. *Genomics Data* **8**, 4–15 (2016)
3. Bühlmann, P., Van De Geer, S.: *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media (2011)
4. Chen, K.-H., Wang, K.-J., Wang, K.-M., Angelia, M.-A.: Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. *Appl. Soft Comput.* **24**, 773–780 (2014)
5. Ebrahimpour, M.K., Eftekhari, M.: Ensemble of feature selection methods: a hesitant fuzzy sets approach. *Appl. Soft Comput.* **50**, 300–312 (2017)
6. Fonti, V., Belitsier, E.: Feature selection using LASSO, VU Amsterdam Research Paper in Business Analytics (2017)
7. Garro, B.A., Rodríguez, K., Vázquez, R.A.: Classification of DNA microarrays using artificial neural networks and ABC algorithm. *Appl. Soft Comput.* **38**, 548–560 (2016)

8. Guo, S., Guo, D., Chen, L., Jiang, Q.: A l1-regularized feature selection method for local dimension reduction on microarray data. *Comput. Biol. Chem.* **67**, 92–101 (2017)
9. Kar, S., Sharma, K.D., Maitra, M.: Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive k-nearest neighborhood technique. *Expert Syst. Appl.* **42**(1), 612–627 (2015)
10. Kumar, M., Rath, N.K., Swain, A., Rath, S.K.: Feature selection and classification of microarray data using mapreduce based ANOVA and k-nearest neighbor. *Procedia Comput. Sci.* **54**, 301–310 (2015)
11. Lin, T.-C., Liu, R.-S., Chen, C.-Y., Chao, Y.-T., Chen, S.-Y.: Pattern classification in DNA microarray data of multiple tumor types. *Pattern Recognit.* **39**(12), 2426–2438 (2006)
12. Liu, Z., Tang, D., Cai, Y., Wang, R., Chen, F.: A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. *Neurocomputing* **266**, 641–650 (2017)
13. Lv, J., Peng, Q., Chen, X., Sun, Z.: A multi-objective heuristic algorithm for gene expression microarray data classification. *Expert Syst. Appl.* **59**, 13–19 (2016)
14. Medjahed, S.A., Saadi, T.A., Benyettou, A., Ouali, M.: Kernel-based learning and feature selection analysis for cancer diagnosis. *Appl. Soft Comput.* **51**, 39–48 (2017)
15. Moayedikia, A., Ong, K.-L., Boo, Y.L., Yeoh, W.G.S., Jensen, R.: Feature selection for high dimensional imbalanced class data using harmony search. *Eng. Appl. Artif. Intell.* **57**, 38–49 (2017)
16. Mohapatra, P., Chakravarty, S., Dash, P.K.: Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm Evol. Comput.* **28**, 144–160 (2016)
17. Mollaee, M., Moattar, M.H.: A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. *Biocybern. Biomed. Eng.* **36**(3), 521–529 (2016)
18. Nguyen, T., Khosravi, A., Creighton, D., Nahavandi, S.: A novel aggregate gene selection method for microarray data classification. *Pattern Recognit. Lett.* **60**, 16–23 (2015)
19. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.-Z.: Deep learning for health informatics. *IEEE J. Biomed. Health Inform.* **21**(1), 4–21 (2017)
20. Sasikala, S., Appavu alias Balamurugan, S., Geetha, S.: A novel adaptive feature selector for supervised classification. *Inf. Process. Lett.* **117**, 25–34 (2017)
21. Sharbaf, F.V., Mosafer, S., Moattar, M.H.: A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. *Genomics* **107**(6), 231–238 (2016)
22. Tabakhi, S., Najafi, A., Ranjbar, R., Moradi, P.: Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing* **168**, 1024–1036 (2015)
23. Tarek, S., Elwahab, R.A., Shoman, M.: Gene expression based cancer classification. *Egypt. Inform. J.* **18**(3), 151–159 (2017)
24. Wang, H., Jing, X., Niu, B.: A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. *Knowl. Based Syst.* **126**, 8–19 (2017)
25. You, W., Yang, Z., Ji, G.: Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination. *Expert Syst. Appl.* **41**(4), 1463–1475 (2014)
26. Zhu, Z., Ong, Y.-S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognit.* **40**(11), 3236–3248 (2007)

# Image Embedding for Detecting Irregularity



M. K. Sharma D. Sheet and Prabir Kumar Biswas

**Abstract** Detecting irregularity in an image or video is an important task in quality control or automatic visual inspection. This paper presents an image embedding technique for detecting an irregularity or abnormality in images. This can further be utilized in image screening application. In the proposed architecture, deep adversarial autoencoder is trained to extract the features from images. Using these features and skip-gram model, we develop the image2vec architecture to capture contextual probability in an image. Various score aggregation techniques are explored and its performance is reported. As a case study, we present a scenario of foreign body object detection in clinical-grade X-ray images. The proposed approach is found to correctly detect and localize abnormality in images.

**Keywords** Visual inspection · Image embedding · Contextual abnormality detection · Adversarial autoencoder · Skip-gram · Negative sampling

## 1 Introduction

Detecting abnormality is one of the important aspects of images and video. Abnormality detection in image plays an important role in many real-life applications [5] such as in the visual inspection, medical image screening, and video surveillance. An accurate and reliable method for detecting abnormality will help in healthcare image

---

M. K. Sharma ()

Advanced Technology Development Centre, IIT Kharagpur, Kharagpur, India  
e-mail: [manojsharma.net@gmail.com](mailto:manojsharma.net@gmail.com)

D. Sheet

Electrical Engineering, IIT Kharagpur, Kharagpur, India  
e-mail: [debdoott@ee.iitkgp.ernet.in](mailto:debdoott@ee.iitkgp.ernet.in)

P. K. Biswas

Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur, India  
e-mail: [pkb@ece.iitkgp.ernet.in](mailto:pkb@ece.iitkgp.ernet.in)

screening. However, learning and detecting abnormality is very challenging due to high-dimensional data, presence of noise, illumination, and complex background.

Features extraction process in deep architecture is found to perform better compared to hand-designed feature extraction such as HOG and SIFT. In this paper, we proposed a framework to detect the contextual abnormality in an image. The framework consists of three components (a) feature extraction, (b) context prediction, and (c) abnormality localization. To extract the relevant feature, we train adversarial autoencoder on the image. It learns the features directly from raw data and handles the distribution of data accordingly [13]. The contextual learning is performed based on skip-gram with negative sampling from the features learnt by adversarial autoencoder for images. Experiments with real-life foreign body object detection in clinical-grade X-ray images show that it can successfully identify the abnormality in an image. To the best of our knowledge, this is the first work to incorporate skip-gram with negative sampling for detecting and localizing an abnormality in the image. The contributions of the paper are as follows:

- Developed an *image2vec* model based on skip-gram with negative sampling to find the probability of image patch,  $c_i$ , to occur in the context of given patch,  $k$ , i.e.,  $P(c_i/k)$ .
- A framework is developed to detect and localize the contextual abnormality in an image.
- The model is used for screening the images for visual inspection. The performance of the model to detect the foreign body object is explored and reported.

The rest of the paper is organized as follows. Section 2 describes the related work in abnormality detection. In Sect. 3, we describe some preliminaries of adversarial autoencoder and skip-gram. Section 4 describes image embedding and abnormality map generation. Section 5 presents a discussion and conclusion.

## 2 Related Work

Nowadays, abnormality or irregularity detection systems are becoming popular [3, 6]. It has a wide range of application such as visual inspection, video surveillance, etc. Most of the existing work identifies the features, learns a normal model, and uses anomaly detection technique such as one-class SVM, Gaussian of mixture model (GMM), isolation forest, etc. [3, 9]. In video, several different features are used in literature to detect abnormality like using optical flow, the histogram of an oriented gradient, histogram of optical flow, mixture of dynamic texture, etc. [9, 12]. Depending upon the algorithm, different features are used. The choice of applying contextual abnormality depends upon the application. In some of the cases, defining a context is straightforward like walking in restricted areas, where region belonging to the restricted area can be identified and then task reduces to identifying pedestrian in this area. However, there are many situations where defining region is not very easy.

In comparison with rich literature in abnormality detection, work on the contextual abnormality is very limited [5]. Broadly speaking the contextual abnormality detection approach is classified into two categories. In the first category, contextual abnormality problem is reduced to point anomaly detection, whereas in the second category, the structure of data is used to detect an abnormality [5]. On the other hand, identifying saliency in images is useful for quality control and automatic inspection, whereas in the video it is used for drawing the attention of users. In order to detect image saliency, Itti et al. [8] computed the dissimilarity between image and its neighborhood. For example, an area having large changes in the contrast is detected as salient image regions. Further, Irani et al. [4] use the graph-based Bayesian model to detect an abnormality. It detects large ensembles of patches by using the relative geometric arrangement of these patches.

### 3 Some Preliminaries

This section briefly discusses adversarial autoencoder and skip-gram with negative sampling.

#### 3.1 Adversarial Autoencoder

Adversarial autoencoder [13] consists of two parts—autoencoder and adversary. The task of autoencoder is to update encoder and decoder to minimize the reconstruction error, whereas tasks of the adversarial part are to first update its discriminative network to tell apart true sample from generated one. Here, true samples are those which are taken from prior, whereas generated data samples are taken from the encoder. Further, the model updates the encoder to confuse discriminator. After training, the decoder will act as a generative model that maps the imposed prior of  $p(z)$  to the data distribution shown in Eq. 1.

$$q(z) = \int_x q(z/x) p_d(x) dx \quad (1)$$

Here,  $z$  is latent code vector,  $x$  is input,  $q(z/x)$  be encoding distribution,  $P_d(x)$  be data distribution, and  $q(z)$  is aggregate posterior distribution. Hence, adversarial autoencoder guides  $q(z)$  to match  $p(z)$ .

#### 3.2 Skip-Gram in NLP

The skip-gram model was introduced in Mikolov [14] to predict the context from given word. The objective of the skip-gram model was to learn high-quality vector

representation of words from a large amount of unstructured text data which will be good at predicting nearby words in the associated contexts. Training of skip-gram model does not involve dense metric multiplication. This makes training simple and efficient.

Skip-gram model was successfully used in *word2vec*<sup>1</sup> architecture in Natural Language Processing (NLP) [7], where the task was to learn word embedding. It is shallow, two-layer neural network to learn the context of the word. It takes a corpus of text and produces vector space of hundreds of dimension. Each unique word is associated with a vector space. Those vectors which have common context are located close to one other. For example, the vectors for the word king and queen are similar and close to each other.

*Negative Sampling:* A large number of weights are used in the neural network and need to be updated every time a training data is selected, which increases the training time. Negative sampling in the skip-gram addresses this by modifying the only small amount of weight vector, instead of taking all of them for each training sample. Here, data are chosen based on its uni-gram frequency, as selecting the word from vocabulary is related to its frequency of use. Those words which are more frequent are more probable to be selected as a negative sample.

## 4 Proposed Approach

This section describes the framework to develop image embedding based abnormality detection system. A feature extractor module is developed which takes an input image and convert them into the feature vector. These feature vectors are used to build the vocabulary of image patches. Further, we build an image embedding module, represented as *image2vec*, which is used to compute the probability of context, given a patch (as shown in Fig. 1). Finally, we aggregate the score for inferencing the abnormality in the image.

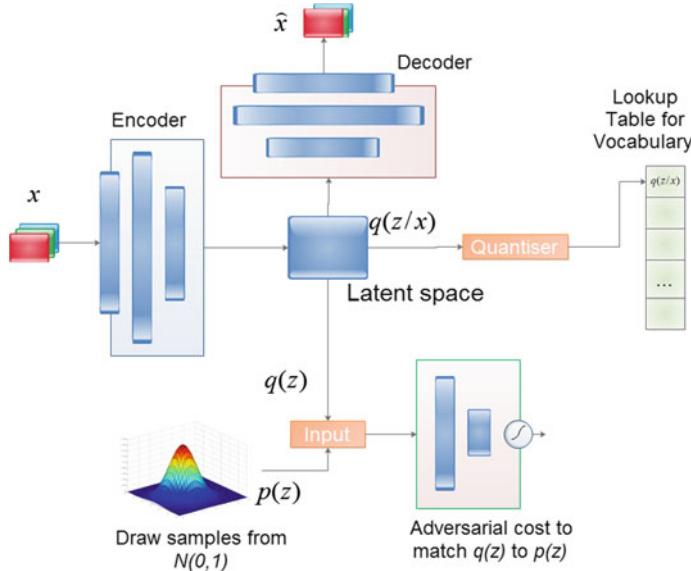
### 4.1 Image Embedding

Generative adversarial autoencoder is trained for the feature extraction process. These features are further used for building the vocabulary. The block diagram of feature extraction and vocabulary building is shown in Fig. 2. Let us assume the input patch size is very small, in our case it is  $3 \times 3$  pixels. Hence, most of the image patch exists in vocabulary.

**Network Structure:** Generative adversarial autoencoder consists of *encoder*, *decoder*, and *adversarial* module. Further, encoder contains three blocks, namely,

---

<sup>1</sup>C implementation available at: <https://code.google.com/archive/p/word2vec/>, and Lua at [https://github.com/yoonkim/word2vec\\_torch](https://github.com/yoonkim/word2vec_torch), respectively.

**Fig. 1** Basic building block**Fig. 2** Feature extraction and vocabulary building block

*EB-1*, *EB-2* and *EB-3*. In decoder side, also there are three blocks, namely, *DB-1*, *DB-2*, and *DB-3*. For adversary, two blocks are used, namely, *AB-1* and *AB-2*. Further description of each of such blocks is presented in Table 1.

**Feature Extraction:** Due to computation limitation, we cannot feed the entire image as an input to the adversarial network. Considering the size of an input image as  $240 \times 210$ , the number of pixels is 50,400. Here, a large number of nodes are required to represent them in the hidden layer. In order to process them, we divide the images

**Table 1** Network details

Type	Block name	Weight	Description
Encoder	<i>EB-1</i>	9–512	Linear +BatchNormalization+ReLU
Encoder	<i>EB-2</i>	512–32	Linear +BatchNormalization+ReLU
Encoder	<i>EB-3</i>	32–1	Linear
Decoder	<i>DB-1</i>	1–32	Linear +BatchNormalization+ReLU
Decoder	<i>DB-2</i>	32–512	Linear +BatchNormalization+ReLU
Decoder	<i>DB-3</i>	512–9	Linear +BatchNormalization+Sigmoid
Adversary	<i>AB-1</i>	1–16	Linear +BatchNormalization+ReLU
Adversary	<i>AB-2</i>	16–1	Linear +BatchNormalization+Sigmoid

into a set of patches of size  $h \times w$ . However, the given image size cannot be fully divisible by the number of patch size. Hence, some prepossessing is required to divide the image into patches. The simple option will be to pad the image; however, padding the image will introduce the noise due to the boundary condition, where sudden dark sub-patch will appear within the patch. In order to overcome these issue, we divide the image such that small amount of pixel from all the side can be adjusted. This is computed using Algorithm 1. This algorithm takes input frame and the patch size represented as  $src$  and  $s$ . Left, right, up, and down represent the amount of pixel shift required in given direction. Processed frame represented as  $data$  in line 22, Algorithm 1 is further used to divide the large image into the collection of images having required patch size.

In this case, patch size is  $3 \times 3$ . More specifically, we take patches and flatten them into a vector of size  $1 \times 9$ , and these vectors are considered as input to the deep learning module. The role of deep learning is to learn good feature representation directly from data and to reduce its dimension from 9 to 1. However, it uses the over-complete network as it maps features from 9 to 512 in the first hidden layer followed by reducing it to 32 and then to 1.

The adversarial network learns the Gaussian distribution while training. Given patch  $x_i$  its output  $y_i = F(x_i)$  is 99% times lies in range of  $3\sigma$  from mean ( $\mu$ ). As an infinite number of points are present in the given range, we quantize them into fixed bins. The number of bins depends upon the level of precession we want to keep.

---

**Algorithm 1:** Preprocessing frame

---

```

1 begin
2   < H, W >=< src : size(1), src : size(2) >;
3   < h, w >=< H% $s$ , W% $s$  >;
4   if w%2 == 0 then
5     left = w/2;
6     right = w/2;
7   else
8     left = (w - 1)/2;
9     right = (w + 1)/2;
10  end
11 if h%2 == 0 then
12   up = h/2;
13   down = h/2;
14 else
15   up = (h - 1)/2;
16   down = (h + 1)/2;
17 end
18 ih1 = 1 + up;
19 ih2 = H - down;
20 iw1 = 1 + left;
21 iw2 = W - right;
22 data = src[{ih1, ih2}, {iw1, iw2}]];
23 end

```

---



---

**Algorithm 2:** Preprocessing for negative sampling

---

```

1 begin
2   N = indexer size;
3   n = vocabulary size;
4   T = 0;
5   foreach  $f_i$  in  $V$  do
6     |  $T = T + f_i^\alpha$ ;
7   end
8   pos = 1;
9    $c_{f_i} = V[i2img(pos)]^\alpha / T$ ;
10  for idx = 1,  $N$  do
11    |  $\Theta[idx] = pos$ ;
12    | if  $\delta > c_{f_i}$  then
13      |   |  $c_{f_i} =$ 
14      |   |  $c_{f_i} + V[i2img(pos)]^\alpha / T$ ;
15      |   |  $pos = pos + 1$ ;
16    | end
17    | if pos > n then
18      |   |  $pos = pos - 1$ ;
19  end
20 end

```

---

**Vocabulary Building:** Let each bin be represented as one entry into the dictionary or lookup table. When the training phase is over, the dictionary, represented as  $V$ , contains the feature vector and its frequency of occurrence. For building the image vocabulary, we remove that entry which has little frequency of occurrence. Say less than  $k$  times, where  $k$  is the frequency threshold. A new lookup table, represented

**Table 2** Frequency of vocabulary words

Vocab	$f$	$f^\alpha$	$\frac{f^\alpha}{\sum f^\alpha}$	$cf_i$
$v_1$	2	1.682	0.151	0.151
$v_2$	3	2.280	0.205	0.356
$v_3$	1	1.00	0.090	0.446
$v_4$	5	3.344	0.300	0.746
$v_5$	4	2.828	0.254	1.000

**Table 3** Selection of samples

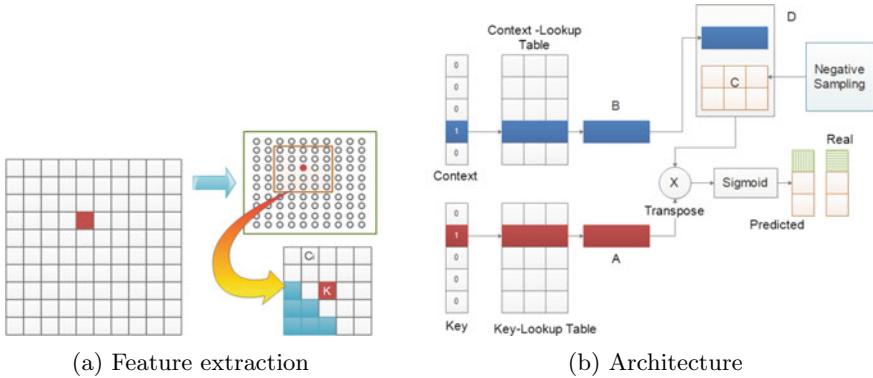
idx	$\delta$	$\Theta$	idx	$\delta$	$\Theta$
1	0.05	1	11	0.55	4
2	0.10	1	12	0.60	4
3	0.15	1	13	0.65	4
4	0.20	1	14	0.70	4
5	0.25	2	15	0.75	4
6	0.30	2	16	0.80	5
7	0.35	2	17	0.85	5
8	0.40	2	18	0.90	5
9	0.45	3	19	0.95	5
10	0.50	3	20	1.00	5

as  $image2index$ , is created and it takes the image feature and returns its position in it. Another lookup table, represented as  $index2image$ , is created to get the feature vector from the given index.

Preprocessing for negative sampling data is discussed with example. Suppose we have five entries  $v_1$  to  $v_5$  in vocabulary having frequency  $f_1$  to  $f_5$  as 2, 3, 1, 5, and 4, respectively, as shown in Table 2. Further assume  $\alpha = 0.75$ . For example, frequency of  $v_i = 2$ , and hence  $f_1^\alpha = 1.682$ . The sum of total  $f^\alpha$  is  $\sum f^\alpha = 11.13343$ . Next we compute the  $cf_i$  by adding the previous entry with the current element. Hence, total will be 1 for last entry.

Assuming there are  $N$  (say, 20) entry reserved to compute the  $indexer$  represented as  $\Theta$ . In Table 3,  $\delta$ , which is  $idx/N$ , is computed for each entry and compared to the  $cf_i$ . In case of 1<sup>st</sup> vocabulary, we insert 1 in Table 3 and compare if  $\delta_1 > cf_1$  if not we repeat the step by checking  $\delta_2 > cf_1$ . However, when it becomes equal, it first updates table and then starts comparing with  $cf_2$  and so on. Finally, we have collection of  $\Theta$  which can be used to select the vocabulary. Here, its occurrence is proportional to the frequency of occurrence of word in vocabulary.

This section illustrates the negative sampling. Let the index of vocabulary word present in the context be represented as context-id and denoted as  $C$  and list of context to store for this  $C$  be  $L$ . Suppose we are interested in the  $n$  number of a negative sample whose weight should be simultaneously updated with the single



**Fig. 3** Generating data and training image2vec

positive sample. We retrieve  $\Theta_i$  where  $i = \text{random}(1, N)$  from Table 3. If retrieved value is not same as context-id, i.e.,  $C \neq \Theta_i$ , then  $\Theta_i$  is stored into list of contexts  $L_2 = \Theta_i$ . We repeat the steps until we get the  $n$  number of context. The first value ( $L_1 = C$ ) is reserved for the context-id and other  $n$  for negative samples.

**Generating Training Data for image2vec:** Suppose image,  $I$ , is divided into grid of patches,  $p_i$ , as shown in Fig. 3a, for each  $p_i$ , we compute the feature,  $F$ , and then quantize them represented as  $\tilde{F}$ . A new image,  $I_1$ , is created by filling the code or feature vector  $\tilde{F}$ , generated from the previous step. Assuming the size of context window be  $W \times W$ , we extract a contextual patch and represent them as  $\rho$  from image  $I_1$ . We store *key* as center value of patch, i.e.,  $\text{key} = \rho_{(W+1)/2}$ . Further, we do not want to compute the probability of occurring of key given key, i.e.,  $P(\text{key}/\text{key})$ , and hence we zeros out the content of center patch. Also, those patches which do not exist in vocabulary are also zeros out. Let the number of nonzero entry in  $\rho$  be denoted as  $n_z$ . We took only those nonzero entries,  $n_z$ , in context patch  $\rho$  and randomize their order. Let  $rw$  be *reduced\_window* size computed as  $rw = \min(\text{random}(1, n_z), W)$ . For each element in  $rw$ , we store *key* and its corresponding context ( $L$ , described in previous section) into lists *train-key* and *train-context*.

**Training image2vec:** The training phase is illustrated with the help of an example shown in Fig. 3b. Let there be five words present in vocabulary. Also, assume that index of key in vocabulary be two and that of context be four, hence hot vector of *key* will be  $[0, 0, 0, 0, 1]$  and that for context will be  $[0, 0, 0, 0, 1]$ . This will activate the second and fourth element in the key lookup table and context lookup table represented as  $A$  and  $B$ , respectively. We also sample the element corresponding to the context (say two negative samples as in the figure) and represent them as  $C$ . Let  $D$  represent one vector with the positive sample and two negative samples. We multiply the transpose of  $A$  with that of  $D$  and represent them as  $E$ . This will create  $3 \times 1$  metric. Then we take the sigmoid of  $E$ , i.e.,  $\text{Sigmoid}([B_k, C]^t \times A_j^t)$ . We compare the output with the input label, where the label for the first entry is one for positive sample and rest are zeros.

**Computing Contextual Probability:** We use binary cross-entropy criteria to train the model.

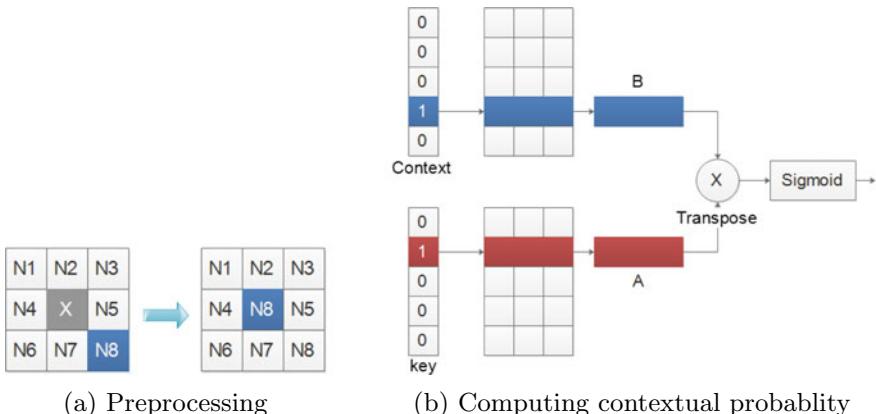
$$\text{loss}(o, t) = -\frac{1}{n} \sum_i (t[i] \times \log(o[i]) + (1 - t[i]) \times \log(1 - o[i])) \quad (2)$$

Here,  $o[i]$  is output and  $t[i]$  is target label. The output of a sigmoid layer is interpreted as the probability of predicting  $t[i] = 1$ .

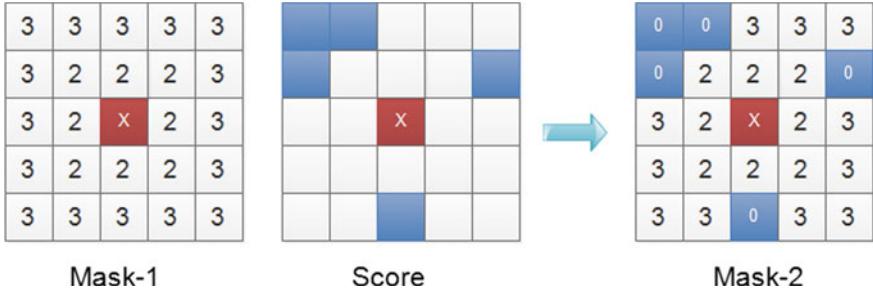
## 4.2 Abnormality Map

This section presents the description of how to generate the abnormality map. We start with the preprocessing part followed by contextual probability computation and score aggregation.

**Preprocessing:** In order to reduce the noise in the abnormality map, one level of preprocessing is done on the frame data (shown in Figs. 4 and 5). Take a patch convert them into feature vector  $F$ , if  $F$  is found in vocabulary  $V$  store the  $F$  at its corresponding position, if not found put 0 in place of  $F$ . Considering center patch  $p_k$  and its eight neighbor patches  $p_1$  to  $p_8$ , as shown in Fig. 4, we compute the mean square error between  $p_k$  and  $p_i$ , i.e.,  $MSE(p_k, p_i)$ , where  $k$  is center patch. Let  $p_j$  be the most similar patch compared to  $p_k$ , if  $p_k$  does not exist in vocabulary, we replace the value of  $p_k$  with that of  $p_j$  and proceed. As shown in Fig. 4a, here  $X$  is replaced by  $N_8$ . After preprocessing, all the data are stored in temporal storage represented as  $\xi$ .



**Fig. 4** Abnormality map generation



**Fig. 5** Masking image patch

**Computing Probability:** From  $\xi$ , we take the *key* as a center of patch and context element say  $c_i$  and compute the probability of occurrence of  $p(c_i/\text{key})$ , which is computed as follows: weight vector corresponding to *key* and *context* is extracted, let it be represented as  $A_j$  and  $B_k$ . We further take the transpose of  $A_j$  and multiply it by  $B_k$  then take sigmoid. The output of sigmoid layer is considered as the probability of  $p(c_i/\text{key})$  as shown in Fig. 4b.

$$P(c_i/\text{key}) = \text{Sigmoid}(B_k \times A_j^T) \quad (3)$$

For those points whose *key* is not defined even after preprocessing, we considered its probability to be zero and handle them separately.

**Detecting Abnormality:** We conducted the experiments on the image taken from the *Radiopedia* [10] and Refs. [2, 11, 15, 16], as depicted in Table 4. The samples of X-ray images containing the foreign body object are presented to adversarial autoencoder. Thereafter, the vocabulary is built using feature quantization and the probability of occurring image patch in a context is computed. In our case, context window size is  $25 \times 25$ , and each of the entities consists of image patch of size  $3 \times 3$ . Hence, the proposed approach captures a total of  $75 \times 75$  pixels in terms of context.

Aggregated abnormality scores for seven different approaches are evaluated and reported in Table 4. Here,  $R1$  represents the image under investigation. In  $R2$ , we take a simple average of the contextual probability within the context window. On the other hand,  $R3$  indicates the performance of taking average with the preprocessing filter (as mentioned in Fig. 4a). Performance of the weighted average taking *mask-1* (as shown in Fig. 5),  $\sum \log(w_j) \cdot x_i / M_1$  is shown in  $R4$ . We note the vignetting effects within  $R4$ , and the difference in corner is compared to  $R5$ , in which the performance of weighted average,  $\sum \log(w_j) \cdot x_i / M_2$ , with *mask-2* is considered. We found that  $R5$  performs better than the other aforementioned approaches. Further, the output from adversarial autoencoder from the adversary is represented in  $R6$ . The performance of one-class SVM and isolation forest is represented as  $R7$  and  $R8$ , respectively. As discussed in Fig. 4a, we applied preprocessing to all approaches except  $R2$ . In order to obtain the binary predicted image from this abnormality map, we corrected the

**Table 4** Abnormality map for different approaches

ID	Image1	Image2	Image3	Image4	Image5	Image6	Image7
R1							
R2							
R3							
R4							
R5							
R6							
R7							
R8							

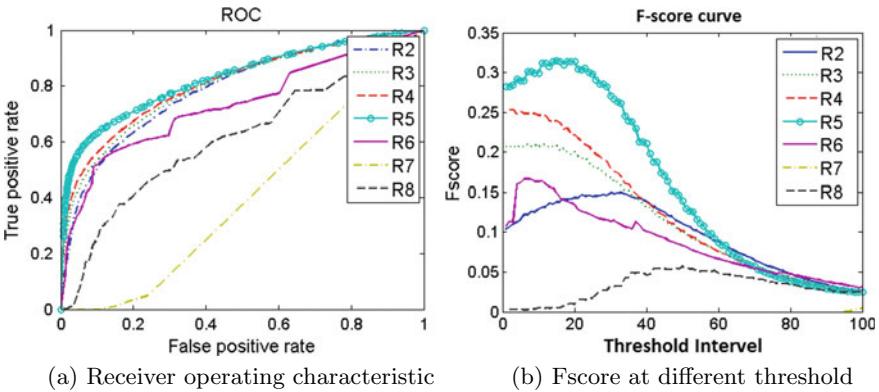
Image according to ID, *R1* image under investigation [2, 10, 11, 15, 16], *R2* average score, *R3* average with preprocessing, *R4* weighted average taking *mask-1*, *R5* weighted average taking *mask-2*, *R6* adversarial autoencoder, *R7* one-class SVM, *R8* isolation forest

nonuniform background while executing the steps mentioned in Matlab Tool Box [1].

For illustration, ground truth and predicted data observed from different approaches for a specific image (*Image2*) at a fixed threshold are presented in Table 5. Here, blue, yellow, and brown indicate ground truth, pixel belonging to correctly predicted, and false alert regions, respectively. We yield that both *R4* and *R5* perform

**Table 5** Predicted irregularity and its ground truth for image2 at fixed threshold

R2	R3	R4	R5	R6	R7	R8

**Fig. 6** Performance measure

better than other approaches. However, we observed less false alert in *R5* in terms of pixel level.

Performance in terms of *ROC* and *F-score* is shown in Fig. 6. From *ROC*, we observed that *R5* outperforms other approaches. Additionally, *R5* has high *F-score* value compared to other approaches and found to be perceptually better. To summarize, we observe that the performance of *R5* is best in localizing contextual abnormality or irregularity.

## 5 Discussion and Conclusion

This paper presents how adversarial autoencoder and skip-gram with negative sampling can be utilized to build the framework for abnormality detection system. The proposed system performs the medical screening and localizes if some abnormality exists in X-ray images. Here, the presence of abnormality means the presence of a foreign body object in clinical-grade X-ray images. We extract the feature from X-ray images with the help of adversarial autoencoder and then build the image vocabulary. These vocabularies are utilized to learn the contextual probability of image patch to occur within the context. However, we learn from the experiment that when patch size is very small in an image, it is very hard to distinguish between the

constituting element of normal versus abnormal region. We overcome this with the help of context prediction and demonstrated that the proposed approach can distinguish between these image regions. The performances of various score aggregation techniques are compared. As a case study, we utilize the scenario of foreign body object detection. The proposed approach is found to correctly detect and localize abnormality in images.

## References

1. Correct nonuniform illumination and analyze foreground objects. <https://in.mathworks.com/help/images/correcting-nonuniform-illumination.html;jsessionid=3cc5aed8bc168e712e7558b041c>. Accessed 2018
2. Swallowed coin. <https://www.topsimages.com/images/swallowed-coin-02.html>. Accessed Sept 2018
3. Aggarwal, C.C.: Outlier Analysis. Springer Science & Business Media (2013)
4. Boiman, O., Irani, M.: Detecting irregularities in images and in video. *Int. J. Comput. Vis.* **74**(1), 17–31 (2007)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. *ACM Comput. Surv.* **41**(3), 15 (2009)
6. Einarsdóttir, H., Emerson, M.J., Clemmensen, L.H., Scherer, K., Willer, K., Bech, M., Larsen, R., Ersbøll, B.K., Pfeiffer, F.: Novelty detection of foreign objects in food using multi-modal X-ray imaging. *Food Control* **67**, 39–47 (2016)
7. Goldberg, Y., Levy, O.: word2vec explained: deriving Mikolov et al.’s negative-sampling word-embedding method (2014). [arXiv:1402.3722](https://arxiv.org/abs/1402.3722)
8. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
9. Javan-Roshtkhari, M.: Visual event description in videos. Ph.D. thesis, McGill University (2014)
10. Jones, D.J., Bickle, D.I.: Ingested foreign bodies in children. <https://radiopaedia.org/>. Accessed Sept 2018
11. Kemp, C.: Be on the lookout for subtle signs of foreignbody ingestion. <https://www.aappublications.org>. Accessed 2018
12. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(1), 18–32 (2014)
13. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders (2015). [arXiv:1511.05644](https://arxiv.org/abs/1511.05644)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
15. Sakhavar, N., Teimoori, B., Ghasemi, M.: Foreign body in the vagina of a four-year-old-girl: a childish prank or sexual abuse. *Int. J. High Risk Behav. Addict.* **3**(2) (2014)
16. Sterling, J.: Straight, no chaser: when foreign bodies are ingested. <https://www.jeffreysterlingmd.com/tag/surgery/>. Accessed 2018

# Optimal Number of Seed Point Selection Algorithm of Unknown Dataset



Kuntal Chowdhury, Debasis Chaudhuri and Arup Kumar Pal

**Abstract** In the present world, clustering is considered to be the most important data mining tool which is applied to huge data to help the futuristic decision-making processes. It is an unsupervised classification technique by which the data points are grouped to form the homogeneous entity. Cluster analysis is used to find out the clusters from a unlabeled data. The position of the seed points primarily affects the performances of most partitional clustering techniques. The correct number of clusters in a dataset plays an important role to judge the quality of the partitional clustering technique. Selection of initial seed of K-means clustering is a critical problem for the formation of the optimal number of the cluster with the benefit of fast stability. In this paper, we have described the optimal number of seed points selection algorithm of an unknown data based on two important internal cluster validity indices, namely, Dunn Index and Silhouette Index. Here, Shannon's entropy with the threshold value of distance has been used to calculate the position of the seed point. The algorithm is applied to different datasets and the results are comparatively better than other methods. Moreover, the comparisons have been done with other algorithms in terms of different parameters to distinguish the novelty of our proposed method.

**Keywords** Clustering · Cluster validity indices · Data mining · Seed point · K-means · Shannons entropy

---

K. Chowdhury (✉) · A. K. Pal

Department of Computer Science and Engineering, Indian Institute of Technology (Indian School of Mines), IIT(ISM), Dhanbad 826004, Jharkhand, India  
e-mail: [ikuntal09@gmail.com](mailto:ikuntal09@gmail.com)

A. K. Pal  
e-mail: [arupkpal@gmail.com](mailto:arupkpal@gmail.com)

D. Chaudhuri  
DRDO Integration Centre, Panagarh, West Bengal, India  
e-mail: [deba\\_chaudhuri@yahoo.co.in](mailto:deba_chaudhuri@yahoo.co.in)

## 1 Introduction

Clustering is an unsupervised technique to group homogeneous data objects to identify a particular class considering certain features. Clustering analysis is considered to be a significant platform for data mining domain in the current era. The objective of the clustering is to maximize the similarity in intracluster and dissimilarity among inter-cluster. Pattern recognition, genetics, etc. are the popular real-life application areas of clustering. Feature selection, clustering algorithm design, cluster validation, and result interpretation are the basic steps of any clustering process. Clustering algorithms are broadly classified as hierarchical approaches, partitional approaches, density-based approaches, grid-based approaches, and model-based approaches. In hierarchical clustering, partition takes place in a bottom-up manner and it is appropriate for information processing. On the other hand, partitional clustering algorithms can be applied to huge datasets and the whole datasets are subdivided into disjoint clusters. The selection of the appropriate number of clusters and the position of its initial seed points play an important role to produce quality clusters. K-means is one of the most famous partitional clustering algorithms in terms of popularity and simplicity. This algorithm needs the value of the number of clusters in advance which is not possible to know for real-life datasets, which is one of the demerits of it [1]. Literature surveys reveal the techniques to find out the correct value of K [2]. The value of K has been decided through the different cluster validity indices [3]. Random initial seed selection in this algorithm affects the global optimum results. It is also sensitive to the shape of the cluster. Mahalanobis distance has been introduced to solve the problems for the hyper-ellipsoidal clusters [4]. Moreover, it has been noticed that this algorithm is more sensitive to noise and outlier points. Outlier pruning is a technique to solve this problem proposed by Zhang [5]. The existence of more than one seed points in a cluster is also possible and defined in the literature [6]. This concept is used to tackle the cluster shape problems. This type of clustering is known as multi-seed clustering. In the literature, different approaches regarding initial seed selection known as initialization of K-means have been proposed by mentioning in a comparative manner. The Voronoi diagram has also been used in initial seed point detection of K-means [7]. Weighted clustering technique has been established to get better initial seed of the cluster [8]. Better initial seed selection algorithm has also been proposed by Celebi for better performance [9]. The rough set model has also been used to get better results in seed selection [10]. Pal et al. [11] have used fuzzy logic for the detection of seed points. In image segmentation, automatic seed selection algorithm has also been used [12]. In the literature, other methods of seed selection approaches have been proposed using density concept [13, 14]. Highest density point can be taken as the best initial seed point [13]. Cumulative density based analysis has also been taken for the selection of seed point [14]. Similarly, another density-based approach has been used by Bai et al. to select the initial seed [15]. This method can give the effective optimal solution provided the number of clusters is less. Sometimes statistical parameter mean is used to select the seed points [16]. This method is sensitive to data ordering. Forgy [17] allocates each point

of the dataset into K clusters uniformly in a random manner. This method has no theoretical basis and suffers from the absence of internal homogeneity [18]. Janceys [19] selects each center to randomly generated data in the data space and leads to the generation of empty clusters. Macqueens [20] has described two different methods for the selection of the initial center. The first method is based on quick clustering technique of IBM SPSS Statistics [21] but also sensitive to data ordering. The second method randomly chooses the initial point so that the points belong to dense regions but no such methodology exists for the outlier detection in that method. Spath [22] has suggested a selection of points to particular clusters in a cyclical fashion identical toForgys method [17] but sensitiveness to data ordering. Max-min method [23] selects the initial center randomly and the next seed is chosen based on the maximum among the minimum distance from the previously selected centers.

## 2 Main Contribution and Organization of Paper

The paper concerns the seed point selection algorithm using entropy with the correct number of clusters. The paper is presented as follows. In Sect. 3, the mathematical definition of two cluster validity indices has been defined upon which our optimal number of the seed points selection (*ONSP*) algorithm stands. The optimal number of the seed points selection (*ONSP*) algorithm 1 is discussed in Sect. 4. The proposed seed point selection algorithm named as SPD (Algorithm 2) is represented in Sect. 4. Experimental results of different datasets are presented in Sect. 5. Finally, we have represented conclusion with future scope in Sect. 6.

## 3 Optimal Number of Seed Points

Determination of the exact number of clusters in an unknown dataset is a very critical job and the cluster quality depends on the number of clusters in the dataset. Clustering quality measures can also be applied to the suitability of the clustering model by comparing different clustering algorithms over the same dataset. Different internal criteria have been formulated to measure the cluster quality and to find out the appropriate number of clusters in data sets [24].

### 3.1 Preliminaries

In real-life unknown dataset, it is very challenging to find out the suitable number of clusters ( $Opt_S$ ) with an objective to improve the cluster quality. In different literature, entropy approach has already been used to find out the exact number of clusters of an unknown real-life dataset [25]. We have observed in different kinds of literature

regarding the importance of different internal and external cluster validity indices for the determination of the number of clusters ( $Opt_s$ ). Clustering quality measure is a good choice for the development of a clustering model to compare different clustering algorithms over the same dataset. Different internal and external criteria have been formulated to measure the cluster quality and also to predict the actual number of clusters in datasets [24]. In the literature, different approaches have been presented to find out the actual number of clusters in a dataset. The major used approaches are listed below. Following approaches are the basic to find out the number of clusters in a dataset like the rule of thumb, Elbow method, Information criteria, Information theory, and cross-validation. The first method is the simple and there is no restriction regarding the application of the dataset. Second method [26] is the oldest method where the optimal number of clusters is determined by visual identification. Information criteria have also been used as a parameter using the model selection approach to determine the suitable number of clusters [27, 28]. Rate-distortion theory has also been applied as an alternative approach for finding out the optimal number of clusters in a dataset. Sugar et al. [29] have used Jump Statistics to propose the optimal number of cluster detection for the Gaussian distribution model. Cross-validation has also been used to calculate the number of clusters [30]. Wang [31] has suggested modified cross-validation approach to find out the number of clusters. Evanno et al. [32] have already proposed a software simulation based scheme to find out the number of clusters in a dataset. Trupti et al. [33] have also proposed the optimal number of cluster detection algorithm. Xu et al. [34] have proposed different models for the determination of the number of clusters. The performance of our algorithm depends on the optimal number of the clusters in the dataset by using these two scalar cluster validity indices. The appropriate number of clusters for a particular dataset has been denoted throughout the paper as  $Opt_s$ . The mathematical definitions of these indices are represented below.

### 3.2 Dunn Index

Dunn index [35] is one of the important cluster validity indexes. The maximum value of Dunn index gives qualitative clustering solution. The mathematical definition of this index is given as

$$D = \min_{m=1 \dots K} \left[ \min_{n=m+1 \dots K} \left( \frac{d(c_m, c_n)}{\max_{i=1 \dots K} \text{diam}(c_i)} \right) \right] \quad (1)$$

where  $d(c_m, c_n) = \min_{x \in c_m, y \in c_n} (d(x, y))$  and  $\text{diam}(c_m) = \min_{x, y \in c_m} (d(x, y))$   $d(c_m, c_n)$  denotes inter-cluster distance between cluster  $m$ th and  $n$ th clusters  $c_m$  and  $c_n$ , respectively,  $d(x, y)$  is the distance between two data element, and  $K$  be the number of clusters.

### 3.3 Silhouette Index

This index [36] is a very popular method which is a combined form of both cohesion and separation. The dataset is clustered through a clustering technique and for each datum  $i$  calculate its average distance, say  $a_i$  to all other points in its cluster. Assume  $b_i$  be the minimum average distance of  $i$  to any other clusters points, of which  $i$  is not a member of those clusters. The silhouette coefficient for the  $i$ th point is

$$S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)} \quad (2)$$

It is to be highlighted that the average value of  $S_i$  for all points of a cluster is a measure of compactness of all the points in the cluster. So the average value of  $S_i$  for all points of the entire dataset is a measure of how properly the points have been clustered. The higher value indicates a good clustering solution. These two internal validity indices give the optimal number of cluster of a particular dataset.

## 4 Proposed Seed Point Selection Algorithm

Clustering is primarily an unsupervised technique and it is very difficult to predict the actual number of the cluster in an unknown dataset. In the clustering technique, there are unlabeled classes, and therefore it is very difficult to find a suitable metric for measuring if the found cluster construction is satisfactory or not. There are many measures of indices available in the literature [3]. Among them, Dunn and Silhouette indices are used for measuring the goodness of a clustering result. The quality of a clustering result depends on two factors (i) the number of seed point and (ii) the position of the first seed point. The clustering results are qualitative provided the exact number of clusters (seed points) are decided. But the difficulty is to get an idea about the appropriate number of seed points. Here, we have taken the maximum value of Dunn and Silhouette indices as the decision-maker for finding the exact number of seed points of an unknown dataset. Already we have discussed that the quality of clustering result depends on the appropriate number of seed points and the initial position of those points. Now the algorithm for finding the optimal number of seed points (*ONSP*) of an unknown dataset is represented as Algorithm 1.

---

**Algorithm 1 Optimal Number of Seed Points (ONSP) Algorithm**


---

**Input:** Image Data or Discrete Data

**Output:** Optimal Number of Seed Points( $Opt_S$ )

- 1: First, compute the Dunn and Silhouette indices assuming the total number of pixels as  $N$  for cluster number  $k = 2$ .
  - 2: Next increase  $k = k + 1$  and find the Dunn and Silhouette indices for current cluster number.
  - 3: Continue 2 until there is a tendency of decreasing mode for both the values of indices for a certain amount of  $k$  from the beginning.
  - 4: Find the number of clusters  $k$  for which both the values of Dunn and Silhouette indices are maximum. Let the maximum attain for Dunn index is at  $k = k_1$  and the same for Silhouette index is at  $k = k_2$ . Let the maximum values of Dunn and Silhouette indices are denoted as  $M_D$  and  $M_S$ , respectively.
  - 5: If  $k_1 == k_2$  then the number of clusters present in the dataset is  $k = k_1$ . Otherwise, find  $Max_k = \max [M_D, M_S]$  and corresponding number of clusters  $k = Opt_S = Max_k$ .
  - 6: Stop.
- 

## 4.1 Seed Point Detection (SPD)

In the previous subsection, we have already discussed the detection of the number of seed points for an unknown dataset. This subsection concerns regarding the position of the initial seed. The cluster quality and faster convergence of the clustering technique depends on the positions of initial seed points. Two cases are considered for the seed point detection: (i) dot pattern (spatial domain) and (ii) image pattern (spectral domain). A unique novel entropy-based seed point selection algorithm is proposed where the seed points are selected on the basis of maximizing the entropy on both these cases. For two dimensional, the number of variables is two and for the three dimensional, the number of variables is three.

## 4.2 Multidimensional Entropy Estimation of Image Data

In this subsection, we are representing the mathematical definition of a multidimensional entropy estimation method for image data. Here we have used mutually independence properties of wavelengths of bands. We have assumed images are of  $q$  bands, namely,  $X_1, X_2, X_3 \dots X_q$ . We have assumed the corresponding intensity values of these  $q$  bands as  $(x_{1j}, x_{2j}, x_{3j} \dots x_{qj})$  respectively, for a particular pixel where  $x_{ij} \in [0, 255]$ ,  $i = 1, 2 \dots q$ ,  $j = 0, 1 \dots 255$  for 8-bit data. So, we can write  $P((X_1 = x_{1j}), (X_2 = x_{2j}) \dots (X_q = x_{qj})) = P(X_1 = x_{1j})P(X_2 = x_{2j}) = \dots = P(X_q = x_{qj})$ ,  $j = 0, 1, \dots, 255$ . Now  $P(X_1 = x_{1j}) = \frac{n_{1j}}{L}$ ,  $P(X_2 = x_{2j}) = \frac{n_{2j}}{L}, \dots P(X_q = x_{qj}) = \frac{n_{qj}}{L}$ ;  $j = 0, 1, 2, \dots, 255$ . Here  $n_{1j}, n_{2j}, \dots n_{qj}$  are the number of points corresponding to the intensity levels  $x_{1j}, x_{2j}, \dots, x_{qj}$ , respectively, for  $j = 0, 1, \dots, 255$ . Here,  $L$  is the total number of points in the input data. So, we can define the entropy ( $E_i$ ) of the pixel as given below:

$$\begin{aligned}
E_i &= -P((X_1 = x_{1j}), (X_2 = x_{2j}), \dots, (X_q = x_{qj})) \\
&\quad \log P((X_1 = x_{1j}), (X_2 = x_{2j}), \dots, (X_q = x_{qj})) \\
&= -P(X_1 = x_{1j})P(X_2 = x_{2j}) \dots P(X_q = x_{qj}) \\
&\left[ \log(P(X_1 = x_{1j})) + \log(P(X_2 = x_{2j})) + \dots + \log(P(X_q = x_{qj})) \right], j = 0, 1, \dots, 255
\end{aligned} \tag{3}$$

The representation of the seed point detection algorithm (*SPD*) is given in Algorithm 2.

---

**Algorithm 2 Proposed Seed Point Selection Algorithm (*SPD*)**


---

**Input:** Input Data and Optimal Number of Seed Points( $Opt_S$ )

**Output:** Detected Seed Points

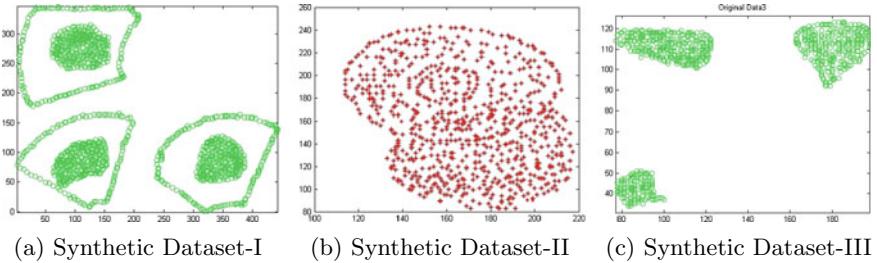
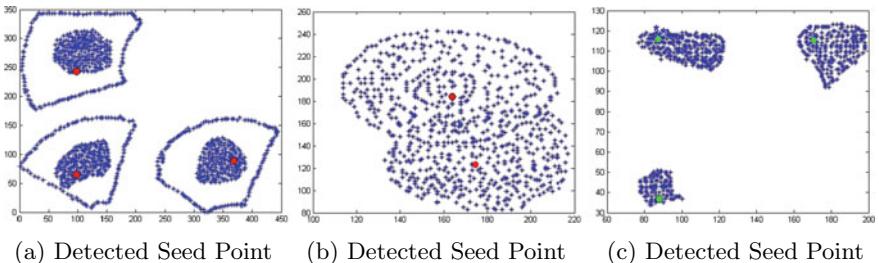
- 1: Find the exact number of seed points in the unknown dataset by applying *ONSP* algorithm.  
Calculate entropy ( $E_i$ ) of the input data using equation (3).
  - 2: We have denoted number of seed point as  $NO_{SP}$  and it is assumed to be zero in the initial stage( $NO_{SP}=0$ ).
  - 3: Sort the datasets in descending order of entropy values calculated using equation (3).
  - 4: Consider maximum entropy value is as the first seed point. Include the first seed point in set S.  
Perform  $NO_{SP} = (NO_{SP} + 1)$ .
  - 5: Find the next point in the sorted dataset after the first point according to the descending order of entropy and find the distance denoted as  $d$  between this point and the first point in the set S. If ( $d < T$ ) goto 6. If( $d \geq T$ ) then include the point in S. Update the set S using  $NO_{SP} = (NO_{SP} + 1)$ . Here,  $T$  is known as the threshold. if ( $NO_{SP} == Opt_S$ ) then goto 8.
  - 6: Consider the next point in the order of highest entropy among the all other points. Find the minimum distance ( $min_d$ ) among the different distances calculated from the current point and the points in the set S. If ( $min_d < T$ ) then goto 7.  $T$  is known as threshold distance. it is heuristic and depend on the distribution of the data. If( $min_d \geq T$ ) then the point will be added in the seed point set S. Perform  $NO_{SP} = NO_{SP} + 1$ .
  - 7: Repeat Step 6 until ( $NO_{SP} == Opt_S$ ).
  - 8: Stop.
- 

## 5 Experimental Results

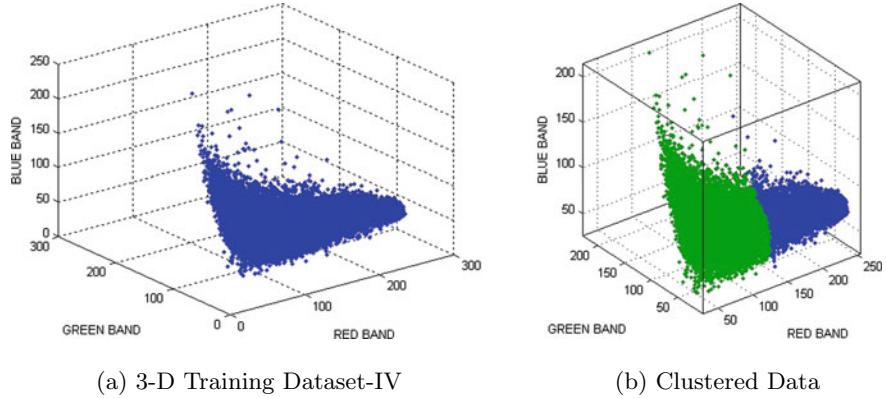
The set of synthetic 2-D data, as well as real-life 3-D data, is used to test the performance of the algorithm. These experiments are performed in MATLAB 7.5 having processor configuration Intel Pentium (R) 2.16 GH with 4GB RAM. The datasets with the optimal number of clusters and the threshold values ( $T$ ) are being represented in Table 1. We have also shown the comparison of these two indices with the other two indices using these datasets represented Figs. 4a–5b. Dunn index and Silhouette index values with respect to the number of clusters and on close observation we have seen that maximum value occurred for both the indices is three for data in Fig. 1a. So, the number of clusters for the Dataset-I is three. The position of the seed points

**Table 1** Description of datasets with optimal number of clusters

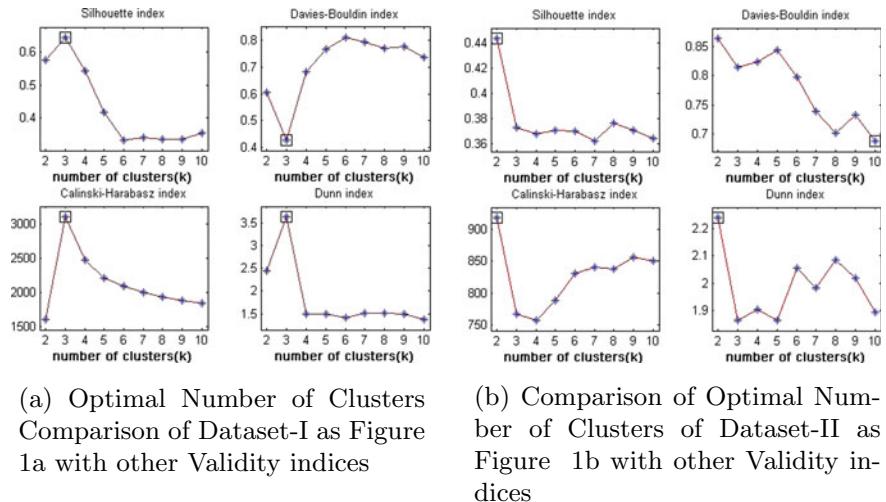
Data	Figure No.	Dimension	Threshold[T]	<i>Opts</i>
Dataset-I	Fig. 1a	(1019 × 2)	70	3
Dataset-II	Fig. 1b	(783 × 2)	65	2
Dataset-III	Fig. 1c	(675 × 2)	30	3
Dataset-IV	Fig. 3a	(64516 × 3)	230	2

**Fig. 1** Three different synthetic datasets**Fig. 2** Detected seed points of three synthetic datasets

detected by our proposed algorithm is highlighted with different colors (green and red) using dots in Fig. 2a. Similarly, it is observed that maximum value that occurred for both the indices is 2 for data in Fig. 1b. So, the exact number of clusters is 2. The dataset in Fig. 1b is an overlapped data and it is very critical to handle such kind of data for clustering. The position of the seed points is highlighted with different colors in Fig. 2b. In Fig. 1c, it is observed that maximum value occurred for both the indices is three. Similarly, the exact number of clusters is three. The position of the seed points is highlighted with different colors in Fig. 2c. Another 3D dataset and the clustered datasets with the optimal number of clusters using the proposed seed selection method are shown in Fig. 3a and b, respectively. We have already discussed that the selection of the optimal number of seed points and their positions are equally important for final clustering result and the computational cost. In the whole process of clustering technique, the time for selection of seed point is also contributed to the computational cost. The computational cost of a clustering technique for a large data

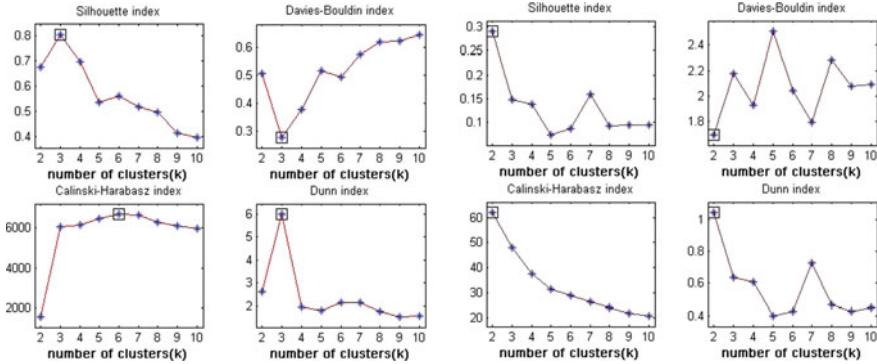


**Fig. 3** 3-D training dataset and its corresponding clustered data



**Fig. 4** Comparison of *Opts* of Dataset-I and Dataset-II

depends on various factors. Among them, the optimal number of seed points, appropriate positions of seed points that reduces the iteration of a clustering technique, and the computational cost for selection of seed points are three important contributors for the total cost of the clustering procedure. Now we have compared the proposed seed selection method as PS with other seed selection methods, which are reported as Astrahan seed selection method as AS [13], Forgy seed selection method as FS [17], Ball and Hall seed selection method as BS [16], Macqueen's seed selection method as MS [20], Max-min seed selection method as MaS [23], Chaudhuri seed selection method as CS [14], min-max seed selection method as MMS [37], and modified min-max seed selection method as MMMS [38]. The estimated time for selection



(a) Comparison of Optimal Number of Clusters of Dataset-III as Figure 1c with other Validity indices

(b) Comparison of Optimal Number of Clusters of Dataset-IV as Figure 3a with Other Validity Indices

**Fig. 5** Comparison of  $Opt_S$  of Dataset-III and Dataset-IV

**Table 2** Comparison of seed selection time between the proposed and other seed selection methods (in Sec)

Figures	AS	BS	CS	FS	MS	MaS	PS	MMS	MMMS
	[13]	[16]	[14]	[17]	[20]	[23]		[37]	[38]
Dataset-I 1a	50.35	16.83	35.13	37.09	0.03	27.02	0.21	0.23	0.20
Dataset-II 1b	22.65	10.01	18.24	29.02	0.03	21.03	0.14	0.18	0.16
Dataset-III 1c	15.66	08.01	11.54	21.02	0.03	16.03	0.08	0.12	0.1
Dataset-IV 3a	600.3	390.5	420.2	500.3	385.7	385.1	374.29	380.2	376.4

of seed points of four datasets Dataset-I (Fig. 1a), Dataset-II (Fig. 1b), Dataset-III (Fig. 1c), and Dataset-IV (Fig. 3a) are shown in Table 2. It is noticed from Table 2 that the computational cost of the proposed method is very less than other methods. Finally, the comparison of the proposed method with the other old and classical methods has been presented in tabular format. These seed point selection methods are denoted like Astrahan method using K-means as AK [13], Ball and Hall method as BK [16], Chaudhuri's method as CK [14], Forgy method as FK [17], Macqueens method as MKM [20], Max-min method as MaKM [23], Min-max method as MMK [37], and modified min-max method as MMMK [38]. The comparisons of K-means clustering algorithm by using various seed selection method are in terms of number of iterations using K-means clustering algorithm and CPU time. The number of iterations by using various seed selection methods and CPU time of all those methods are represented in Tables 3 and 4, respectively. It is pointed out that the number iterations and CPU time are less for convergence of K-means clustering algorithm by the proposed method than the other methods.

**Table 3** Number of iterations using K-means applied by the proposed and other seed selection methods

Figures	AS	BS	CS	FS	MS	MaS	PS	MMS	MMMS
	[13]	[16]	[14]	[17]	[20]	[23]		[37]	[38]
Dataset-I 1a	6	4	5	4	4	4	2	4	3
Dataset-II 1b	9	11	10	11	11	9	4	6	5
Dataset-III 1c	2	3	3	2	2	2	2	2	2
Dataset-IV 3a	21	15	18	19	17	13	11	14	12

**Table 4** CPU time by the proposed and other seed selection methods (in Sec)

Figure	AK	BK	CK	FK	MK	MaK	PK	MMK	MMMK
	[13]	[16]	[14]	[17]	[20]	[23]		[37]	[38]
Dataset-I 1a	0.03	0.33	0.23	0.39	0.031	0.03	0.01	0.04	0.02
Dataset-II 1b	0.03	0.03	0.03	0.05	0.05	0.06	0.01	0.03	0.02
Dataset-III 1c	0.02	0.03	0.03	0.03	0.05	0.03	0.02	0.03	0.02
Dataset-IV 3a	1.35	1.23	0.98	1.15	1.16	1.12	0.83	0.92	0.88

## 6 Conclusion and Future Scope

The optimal seed point selection method on unknown data based on two important internal cluster validity indices, namely, Dunn Index and Silhouette index using Shannons entropy concept with threshold distance has been discussed here. The proposed algorithm produces better performance compared to other algorithms. The proposed algorithm has been tested on different real-life data and the results appear better and qualitative. The proposed algorithm may be tested on some standard datasets like UCI repository and used in developing of a supervised classification scheme for class data, the reduction of training samples in future.

## References

1. Jain, A.K.: Data clustering: 50 years beyond k-means. *Pattern Recognit. Lett.* **31**(8), 651–666 (2010)
2. Chen, K., Liu, L.: The “best k” for entropy-based categorical data clustering (2005)
3. Vendramin, L., Campello, R.J., Hruschka, E.R.: Relative clustering validity criteria: a comparative overview. *Stat. Anal. Data Min. ASA Data Sci. J.* **3**(4), 209–235 (2010)
4. Mao, J., Jain, A.K.: A self-organizing network for hyperellipsoidal clustering (hec). *IEEE Trans. Neural Netw.* **7**(1), 16–29 (1996)
5. Zhang, J.S., Leung, Y.W.: Robust clustering by pruning outliers. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **33**(6), 983–998 (2003)
6. Chaudhuri, D., Chaudhuri, B.: A novel multiseed nonhierarchical data clustering technique. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **27**(5), 871–876 (1997)
7. Reddy, D., Jana, P.K., et al.: Initialization for k-means clustering using voronoi diagram. *Proc. Technol.* **4**, 395–400 (2012)

8. Lu, J.F., Tang, J., Tang, Z.M., Yang, J.Y.: Hierarchical initialization approach for k-means clustering. *Pattern Recognit. Lett.* **29**(6), 787–795 (2008)
9. Celebi, M.E., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst. Appl.* **40**(1), 200–210 (2013)
10. Cao, F., Liang, J., Jiang, G.: An initialization method for the k-means algorithm using neighborhood model. *Comput. Math. Appl.* **58**(3), 474–483 (2009)
11. Pal, S.K., Pramanik, P.: Fuzzy measures in determining seed points in clustering (1986)
12. Kansal, S., Jain, P.: Automatic seed selection algorithm for image segmentation using region growing. *Int. J. Adv. Eng. Technol.* **8**(3), 362 (2015)
13. Astrahan, M.: Speech analysis by clustering, or the hyperphoneme method. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, Tech. rep. (1970)
14. Chaudhuri, D., Murthy, C., Chaudhuri, B.: Finding a subset of representative points in a data set. *IEEE Trans. Syst. Man Cybern.* **24**(9), 1416–1424 (1994)
15. Bai, L., Liang, J., Dang, C., Cao, F.: A cluster centers initialization method for clustering categorical data. *Exp. Syst. Appl.* **39**(9), 8022–8029 (2012)
16. Ball, G.H., Hall, D.J.: Isodata, a novel method of data analysis and pattern classification. Tech. rep., Stanford research inst Menlo Park CA (1965)
17. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* **21**, 768–769 (1965)
18. Anderberg, M.R.: Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks, vol. 19. Academic Press, New York (2014)
19. Jancey, R.: Multidimensional group analysis. *Aust. J. Bot.* **14**(1), 127–130 (1966)
20. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA (1967)
21. Norušis, M.J.: IBM SPSS Statistics 19 Statistical Procedures Companion. Prentice Hall, Upper Saddle River (2012)
22. Späth, H.: Computational experiences with the exchange method: applied to four commonly used partitioning cluster analysis criteria. *Eur. J. Oper. Res.* **1**(1), 23–31 (1977)
23. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.* **38**, 293–306 (1985)
24. Milligan, G.W.: A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* **46**(2), 187–199 (1981)
25. Yan, M., Ye, K.: Determining the number of clusters using the weighted gap statistic. *Biometrics* **63**(4), 1031–1037 (2007)
26. Ng, A.: Clustering with the k-means algorithm. *Machine Learning* (2012)
27. Bozdogan, H.: Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In: Proceedings of the first US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, pp. 69–113. Springer, Berlin (1994)
28. Xu, L.: Byy harmony learning, structural rpcl, and topological self-organizing on mixture models. *Neural Netw.* **15**(8–9), 1125–1151 (2002)
29. Sugar, C.A., James, G.M.: Finding the number of clusters in a dataset: an information-theoretic approach. *J. Am. Stat. Assoc.* **98**(463), 750–763 (2003)
30. Smyth, P.: Clustering using monte carlo cross-validation. *Kdd* **1**, 26–133 (1996)
31. Wang, J.: Consistent selection of the number of clusters via crossvalidation. *Biometrika* **97**(4), 893–904 (2010)
32. Evanno, G., Regnaut, S., Goudet, J.: Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**(8), 2611–2620 (2005)
33. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. *Int. J.* **1**(6), 90–95 (2013)
34. Hu, X., Xu, L.: Investigation on several model selection criteria for determining the number of cluster. *Neural Inf. Process.-Lett. Rev.* **4**(1), 1–10 (2004)
35. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **4**(1), 95–104 (1974)

36. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
37. Tzortzis, G., Likas, A.: The minmax k-means clustering algorithm. *Pattern Recognit.* **47**(7), 2505–2516 (2014)
38. Wang, X., Bai, Y.: A modified minmax-means algorithm based on PSO. *Comput. Intell. Neurosci.* **2016**, (2016)

# TexFusionNet: An Ensemble of Deep CNN Feature for Texture Classification



**Swalpa Kumar Roy, Shiv Ram Dubey, Bhabatosh Chanda,  
Bidyut B. Chaudhuri and Dipak Kumar Ghosh**

**Abstract** The texture classification from images is one of the important problems in pattern recognition. Several hand-designed methods have been proposed in last few decades for this problem. Nowadays, it is observed that the convolutional neural networks (CNN) perform extremely well for the classification task mainly over object and scene images. This improved performance of CNN is caused by the availability of huge amount of images in object and scene databases such as ImageNet. Still, the focus of CNN in texture classification is limited due to non-availability of large texture image data sets. Thus, the trained CNN over Imagenet database is used for texture classification by fine-tuning the last few layers of the network. In this paper, a fused CNN (TexFusionNet) is proposed for texture classification by fusing the last representation layer of widely adapted AlexNet and VGG16. On the top of the fused layer, a fully connected layer is used to generate the class score. The categorical cross-entropy loss is used to generate the error during training, which is used to train the added layer after the fusion layer. The results are computed over several well-known Brodatz, CURET, and KTH-TIPS texture data sets and compared with the state-of-the-art texture classification methods. The experimental results confirm outstanding performance of the proposed TexFusionNet architecture for texture classification.

---

S. K. Roy (✉)

Jalpaiguri Government Engineering College, Jalpaiguri 735102, India

e-mail: [swalpa@cse.jgec.ac.in](mailto:swalpa@cse.jgec.ac.in)

S. R. Dubey

Indian Institute of Information Technology, Sri City 517646, Andhra Pradesh, India

e-mail: [srdubey@iiits.in](mailto:srdubey@iiits.in)

B. Chanda · B. B. Chaudhuri

Indian Statistical Institute, Kolkata 700108, India

e-mail: [bbc@isical.ac.in](mailto:bbc@isical.ac.in)

B. Chanda

e-mail: [chanda@isical.ac.in](mailto:chanda@isical.ac.in)

D. K. Ghosh (✉)

Adamas University, Kolkata 700126, India

e-mail: [dipak@ieee.org](mailto:dipak@ieee.org)

**Keywords** Convolutional neural network (CNN) · Deep learning · Texture classification · Fusion

## 1 Introduction

One of the underlying problems in pattern recognition is the automatic texture image classification. The texture classification has huge applications in different areas such as content-based image retrieval, ground analysis through satellite imagery, biomedical and industrial inspection, etc. Though the human is good recognizer to identify the texture in a real scenario, the definition of texture is still ambiguous [1]. Despite attempts in the past few decades, the texture classification problem is still challenging. The common practice in texture classification is to first compute the suitable features and then use those features with some classifier for training and classification. Numerous texture feature descriptors have been discovered by various researchers [2]. The main three desired properties of any texture feature descriptor are robustness to deal with the intraclass variability, discriminativeness to distinguish between different categories, and low dimensionality to reduce the processing time.

In earlier days, the popular texture classification methods are based on the statistical features such as co-occurrence matrix [3] and Markov random fields [4]. After that, there was an era of filtering-based approaches where the images were converted into feature vectors by applying the bank of filters such as wavelet filters [5], Gabor filters [6], etc. Later on, the macro-pattern-based approaches as local binary pattern (LBP) [7] are proved to be discriminative and robust for texture analysis such. The LBP became one of the state-of-the-art approach for feature extraction due to its simplicity and local relationship representation ability. Several variants of LBP are investigated for different applications such as biomedical image analysis [8–10], face recognition [11, 12], image retrieval [13, 14], pedestrian detection [15], local patch matching [16], etc. Other LBP-based feature descriptors are also proposed for texture classification such as completed local binary pattern (CLBP) [17], LBP variance (LBPV) [18], binary rotation invariant and noise-tolerant descriptor (BRINT) [19], fractal weighted local binary pattern local (FWLBP) [20], complete dual cross pattern (CDCP) [21], jet pattern (LJP) [22], local directional zig-zag pattern (LDZP) [23], local morphological pattern (LMP) [24], etc. These hand-crafted feature descriptors generally perform well in controlled conditions in practice. This is due to the lack of robustness against different types of geometric and photometric variations in a single descriptor.

Recently, the deep learning methodology has changed the research direction in machine learning, computer vision, natural language processing, and pattern recognition area. It learns from massive amount of data and improves the performance with great margin [25]. The first and revolutionary work using deep learning in computer vision was witnessed in 2012 by Krizhevsky et al. [26] for image classification task over a largest and challenging ImageNet database [27] by winning the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Later on, this system

is named as *AlexNet* which is basically a convolutional neural network (CNN) with seven learning layers. After AlexNet, several CNN architectures were proposed for image classification such as VGG16 (16 learning layers) in 2014 by Simonyan and Zisserman of Oxford University [28], GoogleNet (22 learning layers) in 2014 by Szegedy et al. of Google [29], ResNet (152 learning layers) in 2015 by He et al. of Microsoft Research [30], etc. A trend to use more number of learning layers in CNN is observed over the years after 2012. The CNN-based approaches have also shown very exciting performance in other problems such as object detection (Faster R-CNN [31]), image segmentation (Mask R-CNN [32]), biomedical image analysis (Colon cancer detection [33]), etc.

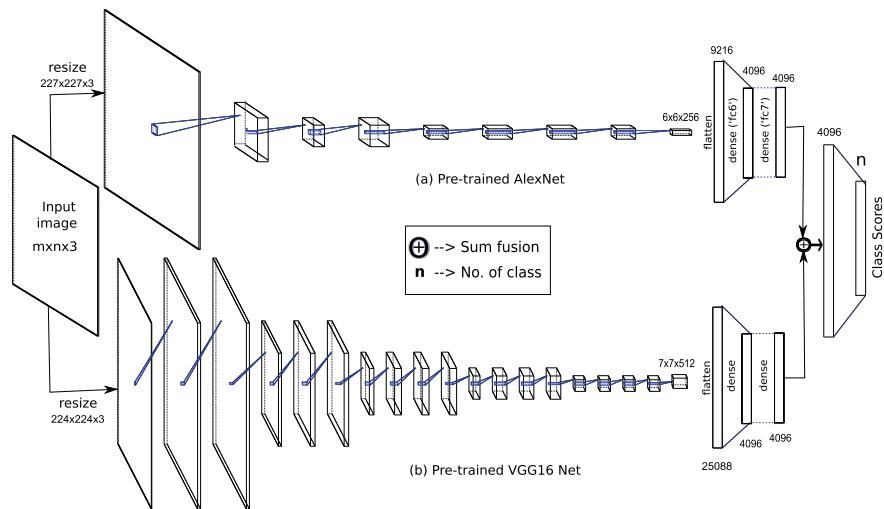
The hierarchical representation of in deep convolution neural networks (CNN) is a key and universal characteristic which is directly formed from the data set and utilized to classify the images. As a universal representation, deep CNNs have shown their recognition power. However, the lack of geometric invariance of globally used CNN activations has limited robustness for recognition. Due to lack of large-scale texture databases, the deep learning based work is not witnessed much for texture classification. Recently, some researchers used the CNN-based methods for texture classification using transfer learning [34, 35]. The transfer learning is a way to use the already trained CNN model for small-scale databases by considering the same weights and fine-tuning the last few layers. Filter banks are also used in the architecture of convolutional neural networks named as texture CNN and applied for classification in texture databases [36]. Cimpoi et al. proposed the Fisher Vector pooling of a CNN filter bank (FV-CNN) for texture recognition and segmentation [37]. Liu et al. [38] have done a performance analysis over LBP and deep texture descriptors in terms of the robustness against different geometric and photometric changes such as rotation, illumination, scale, etc. There is little penetration into the behavior and internal operation of the network although deep CNN models significantly progressed. In 2013, ScatNet was introduced by Mallat et al. [39, 40] where no learning process is needed and convolution filters are predefined as wavelet. Inspired by ScatNet Chan et al. [41] have proposed PCANET, a simple network of deep learning, where network is formed by cascading of multistage principle component analysis (PCA), histogram pooling, and binary hashing. Chan et al. also proposed RandNet [41], a simple variation of PCANET where the cascaded filters are randomly selected and no learning is required. Nowadays, use of more and more complex networks becomes a major trend in deep CNN research community. However, a powerful computer with large memory and GPUs are needed to train the networks.

Most of the existing CNN-based models for texture classification are based on the single model architecture. In this paper, we leverage the power of multiple CNN architectures and design a fused convolutional neural network model for texture classification named as the *TexFusionNet*. The proposed TexFusionNet architecture first fuses the last representation layer of AlexNet and VGG16 and then uses the fully connected layer and softmax layer on top of that for fine-tuning and classification. To evaluate the classification performance, the experiments are conducted on benchmark Brodatz, CUReT, and KTH-TIPS texture databases in support of proposed TexFusionNet architecture.

The rest of the paper is structured as follows: Sect. 2 presents the proposed TexFusionNet architecture; Sect. 3 depicts the experimental results and analysis; and Sect. 4 provides the conclusion.

## 2 Proposed TexFusionNet

The proposed TexFusionNet architecture for texture classification is illustrated in Fig. 1 by fusing two CNN models. The TexFusionNet is composed of the AlexNet [26] and VGG16 [28] architectures of image classification. The AlexNet and VGG16 architectures are the state-of-the-art methods and widely adapted for classification task. The fusion between these models is performed at the last representation layer. The last representation layer is referred to as the last layer of these models after removing the class score layer. In the original AlexNet and VGG16, the last layer is class score layer for 1000 classes of ImageNet database which is removed in our architecture. We could merge well at last representation layer because both AlexNet and VGG16 models produce the same dimensional features (i.e., 4096 dimensional) at the last layer. Moreover, the weights of filters from first layer to last representation layer in both AlexNet and VGG16 are transferred from the pretrained AlexNet and VGG16 models, respectively. Let us consider  $I$  as the input color texture image of resolution  $m \times n$ ,  $\text{alex}$  is a function representing the combination of convolutional layers, max-pooling layers, and fully connected layers from first layer to last representation layer of AlexNet model, and similarly  $\text{vgg16}$  is a function for VGG16 model. Note that the input resolutions for AlexNet and VGG16 are  $227 \times 227$  and



**Fig. 1** Proposed TexFusionNet architecture for texture classification

$224 \times 224$ , respectively. So, the input image  $I$  is converted into  $I_{alex}$  and  $I_{vgg}$  having the resolution of  $227 \times 227$  and  $224 \times 224$  for AlexNet and VGG16, respectively, as follows:

$$I_{alex} = \tau(I, [227, 227]) \quad (1)$$

$$I_{vgg} = \tau(I, [224, 224]) \quad (2)$$

where  $\tau(I, [\beta, \beta])$  is a transformation function to resize any image  $I$  into the resolution of  $(\beta \times \beta)$ .

Let us consider  $\Phi_{alex}$  and  $\Phi_{vgg}$  represent the features or values of last representation layer of AlexNet and VGG16 models, respectively. The  $\Phi_{alex}$  and  $\Phi_{vgg}$  are defined as

$$\Phi_{alex} = alex(I_{alex}) \quad (3)$$

$$\Phi_{vgg} = vgg(I_{vgg}) \quad (4)$$

where the dimensions of  $\Phi_{alex}$  and  $\Phi_{vgg}$  representation features are  $D_{alex}$  and  $D_{vgg}$ , respectively, with  $D_{alex} = D_{vgg}$ .

In the proposed work, the  $\Phi_{alex}$  and  $\Phi_{vgg}$  texture features of input image  $I$  using AlexNet and VGG16 models are fused by addition to produce a combined texture feature representation denoted as  $\Phi_{fused}$  and defined as follows:

$$\Phi_{fused}(i) = \Phi_{alex}(i) + \Phi_{vgg}(i) \quad \forall i \in [1, D] \quad (5)$$

where  $D$  is the dimension of the fused feature  $\Phi_{fused}$  with  $D = D_{alex} = D_{vgg}$ ,  $\Phi_{fused}(i)$  is the  $i$ th element of fused feature  $\Phi_{fused}$  while  $\Phi_{alex}(i)$  and  $\Phi_{vgg}(i)$  are the  $i$ th elements of input features  $\Phi_{alex}$  and  $\Phi_{vgg}$ , respectively.

The computed fused features  $\Phi_{fused}$  are considered as the input to a fully connected layer which produces  $n$  number of outputs as the class scores for  $n$  classes of any texture database (see Fig. 1). Suppose  $S_j|_{j \in [1, n]}$  represents the class score for  $j$ th class of the texture database, where  $n$  is the number of classes. Mathematically, the class score  $S_j$  for  $j$ th class is defined as

$$S_j = \sum_{k=1}^D w_{k,j} \times \Phi_{fused}(k) \quad (6)$$

where  $j \in [1, n]$  and  $w_{k,j}$  is a weight connecting  $k$ th element of feature map  $\Phi_{fused}$  to  $j$ th class score.

During training the pretrained weights of first to last representational layer of both AlexNet and VGG16 are freezed (i.e., not trained over texture database) to utilize the already trained layers. The last fully connected layer (mapping from  $D$  to  $n$ ) is trained by computing the categorical cross-entropy loss over the class scores ( $S$ ) and

backpropagating to last added layer only. Suppose  $c$  is the ground truth class and  $S$  is the computed class scores for image  $I$ . Then the categorical cross-entropy loss ( $L_I$ ) for input image  $I$  is defined as follows:

$$L_I = -\log \left( \frac{e^{S_c}}{\sum_{j=1}^n e^{S_j}} \right) = -S_c + \log \sum_{j=1}^n e^{S_j}. \quad (7)$$

The training loss is computed batchwise and the weights of last added fully connected layer are updated in the opposite direction of its gradients.

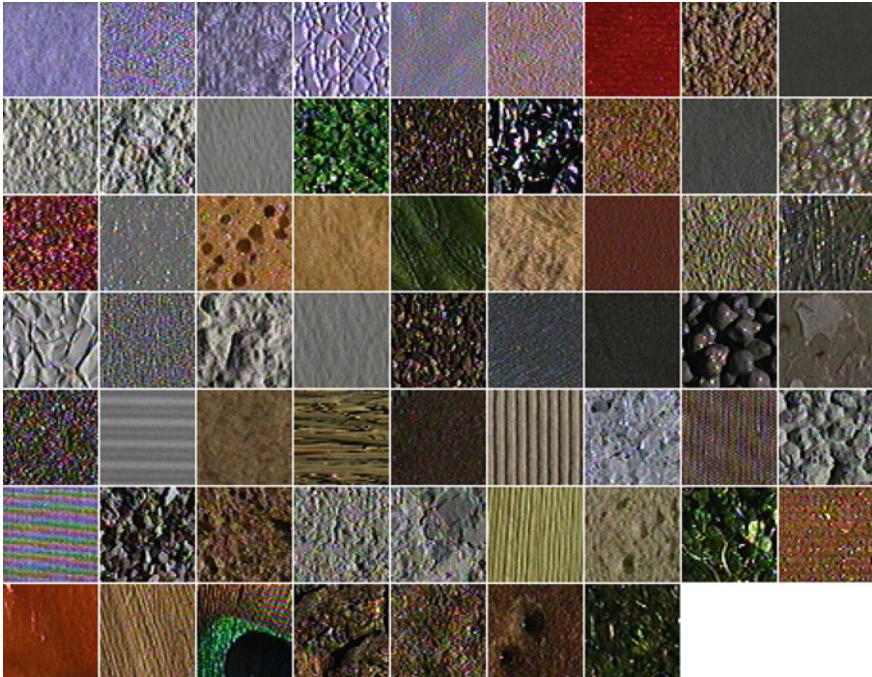
Once the added layer in TexFusionNet is trained over training texture image database, it is used as a trained classifier over test cases. At test time, the input to TexFusionNet is an image itself and the output is class scores. First, the features are computed automatically in intermediate layers by AlexNet and VGG16 separately then these features are fused to produce a combined feature map which is finally used as the input to the final fully connected layer to produce the class scores.

### 3 Experimental Results and Discussion

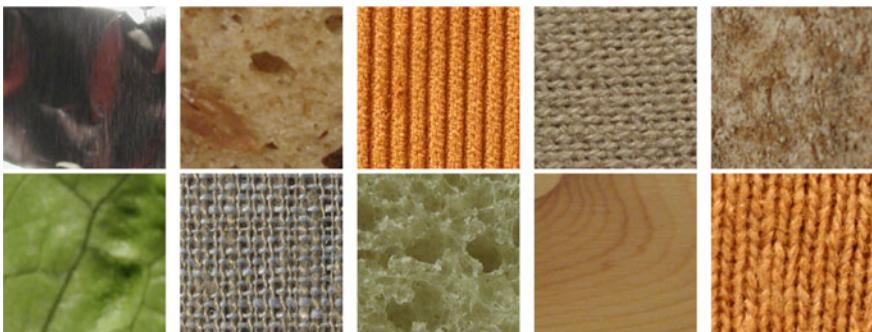
In this section, we examine and evaluate the power of features representation using proposed TexFusionNet which combines two pretrained CNN models for texture classification. We evaluate the effectiveness of proposed TexFusionNet CNN features on the following three publicly available texture databases: first the details of used Brodatz, CUReT, and KTH-TIPS texture databases are discussed in detail. Then the experimental setup and evaluation criteria are illustrated and finally the classification results are analyzed over texture databases using proposed TexFusionNet and compared with the state-of-the-art results.

#### 3.1 Databases Used

Three benchmark texture databases including Brodatz album [42], CUReT [43], and KTH-TIPS [44] are used in this paper to justify the improved performance of introduced TexFusionNet architecture. The presence of various categories, variable number of images in each category, and high degree of intraclass variations are some of the difficulties which make these databases very challenging. **Brodatz** [42] texture database is opted to facilitate a fair comparison with the state-of-the-art results [45]. There are 32 homogeneous texture categories in this database. Each image is partitioned into 25 nonoverlapping sub-regions of size  $128 \times 128$ , and each sub-image is downsampled to  $64 \times 64$  pixels. The same subset of images of **CUReT** database [43] as used in [46] is also chosen in this paper. It contains 61 texture categories having the large intraclass variations with 92 images per category



**Fig. 2** Sample images of the ColumbiaUtrecht (CUReT) texture database



**Fig. 3** Sample images from KTH-TIPS texture database

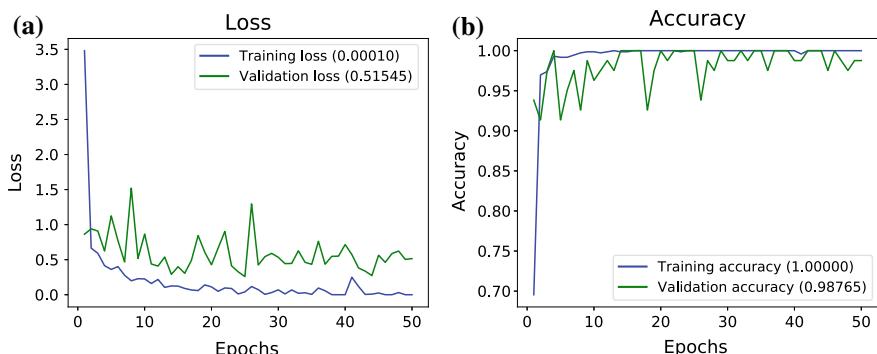
cropped into  $200 \times 200$  pixels region. The sample images are shown in Fig. 2. The images are taken with varying illumination and viewing points with constant scale. The **KTH-TIPS** database [44] is extended by imaging new samples of ten CUReT textures as shown in Fig. 3. It contains texture images with three different poses, four illuminations, and nine different scales of size  $200 \times 200$  and hence each class contains 81 samples.

### 3.2 Experimental Setup and Evaluation Criteria

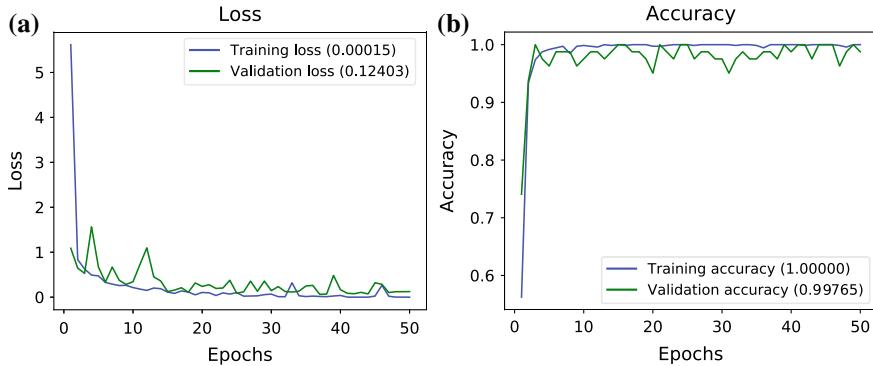
We use Keras to implement the proposed TexFusionNet which is derived from two widely adapted AlexNet and VGG16 models by fusing at the last representation layer. As both of adapted AlexNet and VGG16 models require the predefined size of the input texture image, all images are resized to  $227 \times 227 \times 3$  for AlexNet and  $224 \times 224 \times 3$  for VGG16 where number of feature maps, kernel sizes, etc. are kept same. The mean subtraction preprocessing, a prior stage to computing CNN activation is used in the implementation where the pixel mean value is subtracted from RGB channels through the whole training set corresponding to each pixel. In TexFusionNet, the pretrained convolutional neural network is further trained using *Adadelta* optimizer where the base learning rate is 0.001 and the weight decay is 0.0006 during the performance evaluation using the proposed TexFusionNet. The classification performance of CNN feature is evaluated in terms of the classification accuracy using  $K$ -fold cross-validation test. The ROC is also measured to investigate the performance in terms of the True Positive Rate (TPR) and False Positive Rate (FPR).

### 3.3 Texture Classification Results

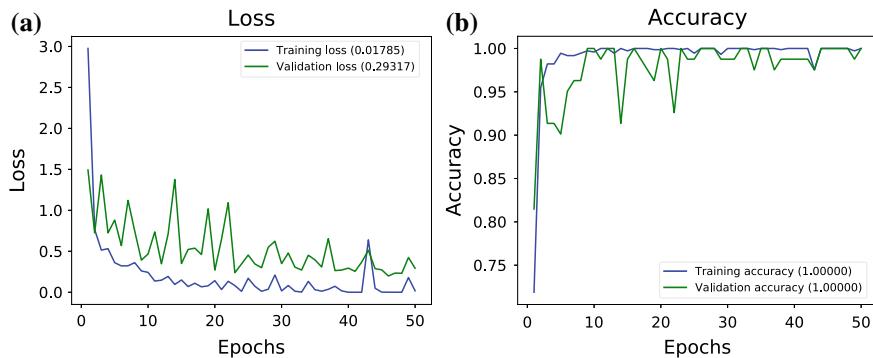
The training and testing texture classification results in terms of the loss and accuracy against the number of epochs using proposed TexFusionNet are depicted in Fig. 4 over Brodatz texture database. The first plot shows the loss versus epochs, whereas the second plot demonstrates the accuracy versus epochs. The similar results over CURET and KTH-TIPS texture databases are illustrated in Figs. 5 and 6, respectively. It can be observed in the plots of loss and accuracy versus epochs that the loss is decreasing and accuracy is increasing over the epochs. Roughly within 10 epochs of



**Fig. 4** The performance of texture classification using the TexFusionNet method over Brodatz texture database, **a** Loss versus epochs and **b** Accuracy versus epochs



**Fig. 5** The performance of texture classification using the TexFusionNet method over CUReT texture database, **a** Loss versus epochs, and **b** Accuracy versus epochs



**Fig. 6** The performance of texture classification using the TexFusionNet method over KTH-TIPS texture database, **a** Loss versus epochs, and **b** Accuracy versus epochs

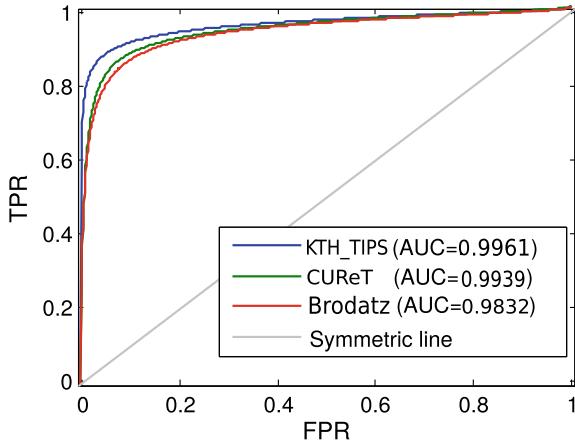
training, the proposed TexFusionNet attains a very high classification result which depicts the benefit of using the pretrained weights of AlexNet and VGG16 models.

The performance comparison of proposed TexFusionNet with state-of-the-art hand-crafted as well as deep learning based methods is also carried out. Table 1 summarized the classification results over Brodatz, CUReT, and KTH-TIPS databases for different methods. The results for hand-crafted descriptors such as LBPV [18], BRINT [19], CLBP [17], PRICOLBP [47], COALBP[48], CDCP [21], and RI-LBD [49] are compared in Table 1. The learning-based methods such as ScatNet [39, 40], PCANet [41], AlexNet [26], FV-VGGM [37], FV-VGGVD [37], and RandNet [41] are also compared in Table 1 with our method. It is observed from Table 1 that the proposed TexFusionNet outperforms the hand-crafted methods such as LBPV [18], BRINT [19], CDCP [21], RI-LBD [49], and also provides better classification performance compared to AlexNet [26], FV-VGGM [37], and other state-of-the-art deep learning based methods. Though the number of training samples is very less,

**Table 1** Comparison of proposed TexFusionNet CNN feature with other variants of LBPs and state-of-the-art deep CNN methods in terms of the texture classification accuracy

Hand-craft methods	Classification rates (%)			Deep CNN Methods	Classification rates (%)		
	Brodatz	KTH-TIPS	CUReT		Brodatz	KTH-TIPS	CUReT
LBPV [18]	93.80	95.50	94.00	ScatNet [39, 40]	84.46	99.40	99.66
BRINT [19]	98.22	97.75	97.06	PCANet [41]	90.89	59.43	92.03
CLBP [17]	94.80	97.19	97.40	AlexNet [26]	98.20	99.60	98.50
PRICOLBP [47]	96.90	98.40	98.40	FV-VGGM [37]	98.60	99.80	98.70
COALBP[48]	94.20	97.00	98.00	FV-VGGVD [37]	98.70	99.80	99.0
CDCP [21]	97.20	97.90	–	RandNet [41]	91.14	60.67	90.87
RI-LBD [49]	97.80	99.30	98.60	TexFusionNet	98.76	100	99.76

**Fig. 7** The area under the curve (AUC) indicates the probability that a model provides classification score between 0 and 1 ranks a randomly chosen true sample larger than a randomly chosen false sample



the proposed TexFusionNet-based CNN features achieve the outstanding average classification accuracy of 98.76%, 100 %, and 99.76% for Brodatz, KTH-TIPS, and CUReT test suits, respectively.

To further visualize the performance of the proposed TexFusionNet in terms of the receiver operating characteristics (ROC), the area under curve (AUC) is measured and depicted in Fig. 7 over Brodatz, CUReT, and KTH-TIPS databases. The true positive rate (TPR) and false positive rate (FPR) values are plotted along y- and x-axes, respectively. It is observed from Fig. 7 that the the performance of proposed model is reasonable over each texture database and proposed CNN feature achieves AUC values of 98.32%, 99.61% and 99.39%, respectively for Brodatz, KTH-TIPS, and CUReT test suits.

## 4 Conclusion

In this paper, a TexFusionNet model is proposed for the texture classification. The TexFusionNet used the existing pretrained CNN models of AlexNet and VGG16 and fused at last representation layer after removing the original class layer of these networks. The fusion is performed by adding the last representation layer of both networks. A fully connected layer is placed over the fusion layer to generate the class scores which is used to generate the loss during training with categorical cross-entropy loss function. After training, the fused model is used for the testing over texture images in classification framework. Three benchmark Brodatz, CUReT, and KTH-TIPS texture databases are used to judge the performance of proposed model. The results are compared with the state-of-the-art hand-crafted and learning-based methods. The experimental results suggest that the introduced TexFusionNet outperforms the hand-crafted methods and also shows very promising performance as compared to the deep learning based methods.

## References

1. Tuceryan, M., Jain, A.K., et al.: Texture analysis. In: *Handbook of Pattern Recognition and Computer Vision*, vol. 2, pp. 207–248 (1993)
2. Xie, X., Mirmehdi, M.: A galaxy of texture features. In: *Handbook of Texture Analysis*, pp. 375–406. World Scientific, Singapore (2008)
3. Haralick, R.M., Shanmugam, K., et al.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**(6), 610–621 (1973)
4. Cross, G.R., Jain, A.K.: Markov random field texture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **5**(1), 25–39 (1983)
5. Chang, T., Kuo, C.-C.J.: Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans Image Process.* **2**(4), 429–441 (1993)
6. Idrissa, M., Acheroj, M.: Texture classification using gabor filters. *Pattern Recognit. Lett.* **23**(9), 1095–1102 (2002)
7. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
8. Dubey, S.R., Singh, S.K., Singh, R.K.: Local wavelet pattern: a new feature descriptor for image retrieval in medical ct databases. *IEEE Trans. Image Process.* **24**(12), 5892–5903 (2015)
9. Dubey, S.R., Singh, S.K., Singh, R.K.: Local bit-plane decoded pattern: a novel feature descriptor for biomedical image retrieval. *IEEE J. Biomed. Health Inform.* **20**(4), 1139–1147 (2016)
10. Dubey, S.R., Singh, S.K., Singh, R.K.: Local diagonal extrema pattern: a new and efficient feature descriptor for ct image retrieval. *IEEE Signal Proces. Lett.* **22**(9), 1215–1219 (2015)
11. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **19**(6), 1635–1650 (2010)
12. Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE Trans. Image Process.* **19**(2), 533–544 (2010)
13. Murala, S., Maheshwari, R.P., Balasubramanian, R.: Local tetra patterns: a new feature descriptor for content-based image retrieval. *IEEE Trans. Image Process.* **21**(5), 2874–2886 (2012)
14. Dubey, S.R., Singh, S.K., Singh, R.K.: Multichannel decoded local binary patterns for content-based image retrieval. *IEEE Trans. Image Process.* **25**(9), 4018–4032 (2016)

15. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV, pp. 32–39. IEEE (2009)
16. Dubey, S.R., Singh, S.K., Singh, R.K.: Rotation and illumination invariant interleaved intensity order-based local descriptor. *IEEE Trans. Image Process.* **23**(12), 5323–5333 (2014)
17. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **19**(6), 1657–1663 (2010)
18. Guo, Z., Zhang, L., Zhang, D.: Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recognit.* **43**(3), 706–719 (2010)
19. Liu, L., Long, Y., Fieguth, P.W., Lao, S., Zhao, G.: Brint: binary rotation invariant and noise tolerant texture classification. *IEEE Trans. Image Process.* **23**(7), 3071–3084 (2014)
20. Roy, S.K., Bhattacharya, N., Chanda, B., Chaudhuri, B.B., Ghosh, D.K.: FWLBP: a scale invariant descriptor for texture classification. [arXiv:1801.03228](https://arxiv.org/abs/1801.03228) (2018)
21. Roy, S.K., Chanda, B., Chaudhuri, B., Ghosh, D.K., Dubey, S.R.; A complete dual-cross pattern for unconstrained texture classification. In: 4th Asian Conference on Pattern Recognition (ACPR 2017), Nanjing, China, pp. 741–746 (2017)
22. Roy, S.K., Chanda, B., Chaudhuri, B.B., Banerjee, S., Ghosh, D.K., Dubey, S.R.: Local jet pattern: a robust descriptor for texture classification. [arXiv:1711.10921](https://arxiv.org/abs/1711.10921) (2017)
23. Roy, S.K., Chanda, B., Chaudhuri, B.B., Banerjee, S., Ghosh, D.K., Dubey, S.R.: Local directional zigzag pattern: a rotation invariant descriptor for texture classification. *Pattern Recognit. Lett.* **108**, 23–30 (2018)
24. Roy, S.K., Chanda, B., Chaudhuri, B.B., Ghosh, D.K., Dubey, S.R.: Local morphological pattern: a scale space shape descriptor for texture classification. *Digit. Signal Process.* (In Press) (2018). Elsevier
25. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT press, Cambridge (2016)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 (2012)
27. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR, pp. 248–255. IEEE (2009)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. In: CVPR (2015)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
32. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV, pp. 2980–2988. IEEE (2017)
33. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R.J., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **35**(5), 1196–1206 (2016)
34. Hafemann, L.G., Oliveira, L.S., Cavalin, P.R., Sabourin, R.: Transfer learning between texture classification tasks using convolutional neural networks. In: IJCNN, pp. 1–7. IEEE (2015)
35. Basu, S., Mukhopadhyay, S., Karki, M., DiBiano, R., Ganguly, S., Nemani, R., Gayaka, S.: Deep neural networks for texture classification a theoretical analysis. *Neural Netw.* **97**, 173–182 (2018)
36. Andrearczyk, V., Whelan, P.F.: Using filter banks in convolutional neural networks for texture classification. *Pattern Recognit. Lett.* **84**, 63–69 (2016)
37. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: CVPR, pp. 3828–3836 (2015)
38. Liu, L., Fieguth, P., Wang, X., Pietikäinen, M., Hu, D.: Evaluation of lbp and deep texture descriptors with a new robustness benchmark. In: ECCV, pp. 69–86. Springer, Berlin (2016)

39. Bruna, J., Mallat, S.: Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1872–1886 (2013)
40. Sifre, L., Mallat, S.: Rotation, scaling and deformation invariant scattering for texture discrimination. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1233–1240 (2013)
41. Chan, T.-H., Jia, K., Gao, S., Jiwen, L., Zeng, Z., Ma, Y.: Pcanet: a simple deep learning baseline for image classification? *IEEE Trans. Image Process.* **24**(12), 5017–5032 (2015)
42. Brodatz, P.: *Textures: A Photographic Album for Artists and Designers*. Dover publications, New York (1966)
43. Dana, K.J., Van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real-world surfaces. *ACM TOG* **18**(1), 1–34 (1999)
44. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.-O.: On the significance of real-world conditions for material classification. In: *European Conference on Computer Vision*, pp. 253–266. Springer, Berlin (2004)
45. Liao, S., Law, M.W.K., Chung, A.C.S.: Dominant local binary patterns for texture classification. *IEEE Trans. Image Process.* **18**(5), 1107–1118 (2009)
46. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. *Int. J. Comput. Vis.* **62**(1–2), 61–81 (2005)
47. Qi, X., Xiao, R., Chun-Guang Li, Y., Qiao, J.G., Tang, X.: Pairwise rotation invariant co-occurrence local binary pattern. *IEEE TPAMI* **36**(11), 2199–2213 (2014)
48. Nosaka, R., Suryanto, C.H., Fukui, K.: Rotation invariant co-occurrence among adjacent lbps. In: *ACCV*, pp. 15–25. Springer, Berlin (2012)
49. Duan, Y., Jiwen, L., Feng, J., Zhou, J.: Learning rotation-invariant local binary descriptor. *IEEE Trans. Image Process.* **26**(8), 3636–3651 (2017)

# Person Identification Using Footprint Minutiae



Riti Kushwaha and Neeta Nain

**Abstract** Increasing demand of biometric security system is leading us to evaluate more novel approach which could recognize and identify humans using their unique traits. Like fingerprint, footprint of human also share the unique traits which could be used as additional biometric trait in absence of major biometric trait (fingerprint, face). This could be installed at the places where it is mandatory to remove shoe for security or religious reasons, for example, airports and places of worship like mosques, pilgrimage, etc. We propose a minutiae-based footprint recognition system for person identification. Furthermore, credibility of fingerprint matching matrices is used for the evaluation of our footprint approach. Minutiae features of foot fingers are used to evaluate the performance of person identification using footprint. We have collected dataset of 120 persons including 78 males and 42 females at different resolution (500, 600, and 1200 dpi). Collected dataset involves two samples of each person for every foot finger. A series of dataset preprocessing operations are applied for getting good quality images which also ensures the correct region of interest detection for getting the large number of minutiae in a small region. Concept of crossing number is used for ridge termination and bifurcation extraction. The method achieves *False Non Match Rate (FNMR)* of 0.4% with the *False Match Rate (FMR)* of 0.2%.

**Keywords** Footprint · Biometrics · Local orientations · Local frequency · Level 2 features · Minutiae

---

R. Kushwaha · N. Nain

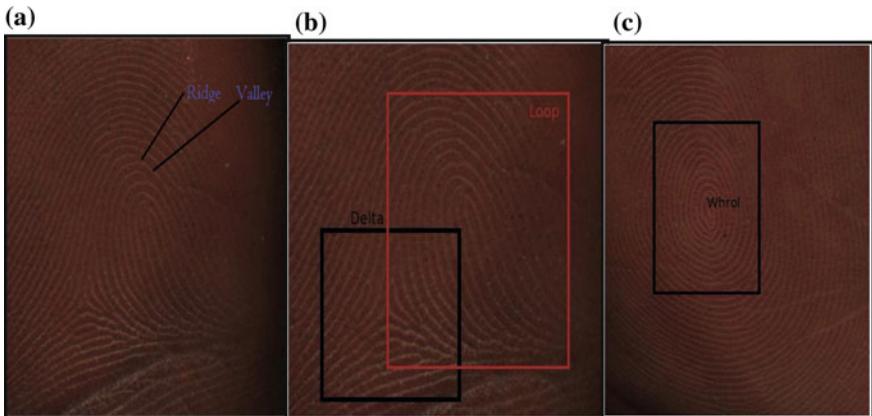
Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur 302017, India  
e-mail: [riti.kushwaha07@gmail.com](mailto:riti.kushwaha07@gmail.com)

N. Nain  
e-mail: [nnain.cse@mnit.ac.in](mailto:nnain.cse@mnit.ac.in)

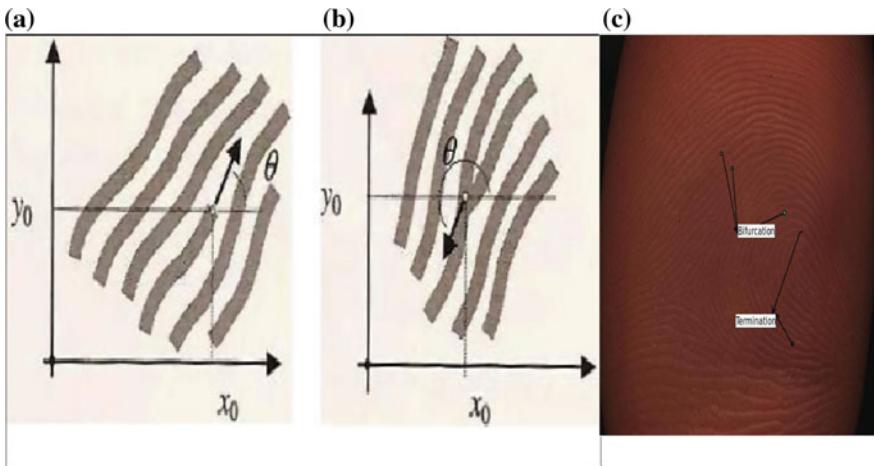
## 1 Introduction

Similar to fingerprint, a typical scan of footprint at 500 dpi also consists of two levels of features. At level 1, it has ridges and valley structure, we can get local orientations and local frequencies. At level 2, it has minutiae features which consist of ridge bifurcation, ridge ending, and the angle between them. The level 3 features are visible at 1200 dpi, which are pores and ridge shape details as given below:

- Level 1: Analysis of the footprint at global level gives us two main features which are local orientations and local frequency. Position  $(x, y)$  gives the local orientation, and for getting the angle of the ridge a reference horizontal axis is taken. The angle made between horizontal axis and the ridge gives the orientation of ridge. The local frequency is defined at any given point  $(x, y)$  as the number of ridges per unit length along a hypothetical segment centered at  $(x, y)$  and orthogonal to the local orientation. We follow the ridge pattern in this level. Ridge is the outermost structural part of epidermis. Pattern may have atleast one region also known as ridgelines structure, which are assumed of distinctive shape. These regions are called singularities. It could be classified into three major topologies, Loop ( $O$ ), Delta ( $\Delta$ ), and Whorls ( $U$ ). We first find the core points in the structure and then align those core points to match the shape. Figure 1 shows the level-1 features of a sample foot. All the five foot fingers and ball region of foot have these singularities. Because of using a flatbed paper scanner, singularity at foot thumb and ball region is completely clear. We achieve less singularity in other three fingers of foot because of its swirling behavior.
- Level 2: At the local level of a footprint scan, minutiae features can be found in the pattern of foot, and these patterns are visible because of ridges (emerged



**Fig. 1** Level-1 features: **a** ridges and valleys in a foot thumb, **b** delta is represented by black color rectangle and loop is represented by red color rectangle, **c** whorl in the ball region as black color rectangle



**Fig. 2** Level-2 features: **a** termination, **b** bifurcation, **c** a typical scan of foot thumb shows both minutiae features (termination and bifurcation)

portion) structure. We tap the significant change in ridge pattern, classified into two categories as ridge termination and ridge bifurcation. A ridge comes to an end is called termination, a ridge is divided in to two parts, and is called ridge bifurcation. We also find the angle between minutiae. Figure 2 shows the level-2 features of a sample foot.

- Level 3: It could be found at every local level of footprint scan. These details include width, shape, edge contour, and all the dimensional attributes of the ridges, sweat pores, breaks and scars, and incipient ridges. These features could be helpful in latent footprint examination for forensic applications.

The paper structure is given as follows: Sect. 2 explains the literature survey of fingerprint. Section 3 explains dataset collection process along with the problems while capturing the footprint. It also describes the initial effort to check whether footprint has minutiae features or not. Section 4 explains the proposed methodology, Sect. 5 illustrates the experimental analysis, and Sect. 6 provides the conclusion.

## 2 Related Work/Background

Kennedy [5] checks the uniqueness of the barefoot impressions. He collected barefoot dataset from 19 volunteers using ink. He tracked the impressions and also checks the uniqueness of the barefoot impressions. Now they have over 6000 impressions. The study of Barker [2] shows that there are three main factors which contribute in the discrimination of foot. (1) shape and size, (2) minutiae features, and (3) environment condition. The approach developed by Nakajima et al. [14] is unique. He captured

the pressure distribution of a foot using pressure-sensing mat. These pair of footprint got normalized in direction and position along with the geometric information of foot. The distance measure is Euclidean distance. Daniel et al. [18] collected the dataset of footprint from 106 infants using high-resolution optical sensor. It captures enough minutiae information for the identification. He captures both palm print and footprint of newborn and concluded that palm prints are better than the footprints in terms of recognition accuracy.

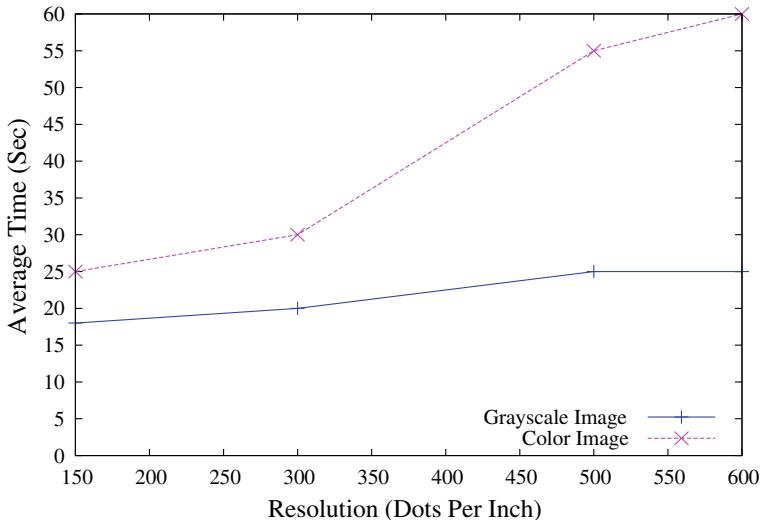
Uhl and Wild [17] did extensive analysis on footprint biometric and found out the shape features, eigen features, and minutiae features. The proposed techniques provide satisfactory results. Kapil et al. [13] used pressure-sensing mat to capture footprint. Their algorithm is based on the normalization of foot using the Haar wavelet energy features. Kumar and Ramakrishnan [7], proposed an algorithm based on wavelet transform and Gabor filter for foot feature extraction. Detailed description of Gabor filtering is given in the article [9]. Article [8] proposes the possible application of footprint. He explains that this could be employed at holy places for providing security. The another application could be baby swapping prevention in hospitals. The footprint's minutiae features of the infant could be associated with the biometric identity of mother. Rohit et al. [6] used independent component analysis and principal component analysis to extract the texture- and shape-based features. Both the techniques are very efficient. PCA is mainly used for data reduction and ICA is a blind source separation technique for revealing hidden factors. Different distance computation techniques such as city block, Euclidean distance, correlation, and cosine are computed as distance metric. Paper [10] details the computation of geometrical features of foot, distance from centroid to contour foot polygon which is matched with the template foot. Authors [11] further explore the footprint geometry by slicing foot image horizontally and foot descriptors are derived for feature extraction and matching.

Extensive studies have been conducted on the human fingerprint for identification purpose. It is very popular because the minutiae features are permanent even after doing hand-labor-intensive work and also data collection is very simple [1, 3, 4, 12, 16]. The time complexity of fingerprint recognition process is very less in real time which makes it the first choice of user identification. By taking the credit of fingerprint algorithm, we are going to check the accuracy of footprint biometric trait for person identification.

### 3 Capturing Footprint Images

For the modern application involving identification of human using their footprints, the initial challenge is to collect footprint images with good quality. This is essential because of the following reasons:

- *Less-cooperative device:* There is no traditional footprint-capture sensor available. We captured the foot images using a normal paper scanner at different resolutions.



**Fig. 3** Resolution versus time graph: *X*-axis shows the resolution and *Y*-axis shows the time

Time is varying according to the resolution of the image. For example, if we capture the color image at 150 dpi then it will take 20s, for 300 dpi it takes approximately 30–40 s, for 600 dpi it takes approx 1 min, and for 1200 dpi it takes approx 3 min. Hence, it is difficult to place foot at a particular position till this duration of time. As a result, often there is insufficient time for the footprint sensors to capture good quality images. Typically, one has to be very patient for capturing the footprint at 1200 dpi. Figure 3 shows the graph of time taken by the scanner at different resolutions. Dirt, cracked heels, soiled foot, with an added disadvantage of removal of footwear, and small-sized paper-print scanner are some of the challenges in capturing good footprints.

Next we acquired footprint images with the traditional paper scanner on 500, 600, and 1200 dpi resolution. Based on our experience, sensible sensor for capturing footprints should have fast capture speed, because it is difficult to hold the foot at one position for a very long time. Compact and comfortable sensor platen is required to place the person's foot properly. Distribution of weight is required so that whole weight of body should not fall on the foot and not cause deform pattern and crack in scanner. We used the manual way of weight distribution by using a chair to sit upon, such distribution allows most of the body weight to fall in the chair, and hence we can place the foot comfortably on the scanner. There is a need to regularly clean the sensor platen to prevent the residue buildup from previous footprint capture which may appear as background noise in the image. The foot is to be placed very lightly on the scanner; otherwise, any external pressure on the foot causes distortion in the minutiae pattern by flattening the ridges. Need to wipe off the extra dirt from foot surface after removing shoes or socks. This could capture sufficiently good quality

**Fig. 4** Collected foot samples using paper scanner at 500 dpi



images using this arrangement and we were able to build out footprint dataset using this setup. Usage of photographic cameras, to obtain a high-resolution image will also be considered in the future.

### 3.1 Data Collection

480 Footprint images from 78 males and 42 females were captured using the “HP Scanjet G3110” and “Canon IMAGEclass MF3010” paper scanner at department of computer science, MNIT Jaipur, India. Resolution of the image is 500 dpi (Canon IMAGEclass MF3010), 600 dpi (HP Scanjet G3110), and 1200 dpi (HP Scanjet G3110). Two samples of each person for both feet are captured, and the age variation of subjects is [18–35] years.

A total of 480 images from 120 persons are captured and labeled. Impressions of five fingers and one ball region are cropped from the image. A total of 2880 footprint fingers and ball region impressions were taken. Data was collected over two samples per foot in a session. Method is cost-effective, because it is captured on paper scanner which is cheap and could be cleaned using a simple wipe of cloth. Figure 4 shows footprint samples of the Foot dataset.

## 4 Methodology to Compute Minutiae Features

Footprint patterns consist of ridges and valleys. We use ridges for the extraction of features. Ridge flow is of two types: (1) pseudo-parallel ridge flow and (2) high curvature ridge flows which are located around the core point or delta points.

This technique is based on minutiae extraction which consists of two main points, ridge ending and ridge bifurcation. When a ridge ends it is considered as ridge ending and when a ridge is bifurcating into two (converge or diverge) it is known as ridge bifurcation. Minutiae extraction has three main stages: preprocessing, minutiae point extraction, and postprocessing.

After acquiring the footprint dataset, we perform dataset preprocessing, which is further divided in to five steps: histogram equalization, enhancement using FFT, binarization using local adaptive threshold. Ridge detection is done by block direction estimation and extraction of ROI (region of interest) using morphological operations.

Minutiae extraction are mainly divided into two steps. The first step involves thinning and minutiae marking and second step involves removal of false minutiae. Second step is also known as post-processing. After performing all the steps, we perform minutiae matching. Detailed description of every step is given below:

- *Database preprocessing*: Extraction of correct minutiae relies remarkably on the quality of input footprint image. In the ideal case of footprint image, ridge and valley flows alternate in a constant direction, so that ridges can be easily detected for correctly locating minutiae points. However, many times we are not able to get the desired image due to skin conditions, sensor limitations, and incorrect footprint pressure. Skin condition such as wet or dry footprint, cuts, and bruises, sensor quality by not using the dedicated sensor for footprint acquisition, and incorrect foot pressure these points leads to image degradation and flattens the ridge and valley structure. Imaged degradation could be shown by the ridges that are not strictly continuous, and parallel ridges which are not well separated. These conditions make ridge extraction extremely difficult, and create problems like missing a significant number of genuine minutiae and addition of spurious minutiae. To ensure good performance of the algorithm, an enhancement technique is necessary. Histogram equalization followed by Fourier transformation is used to connect the broken ridge structure and increase the contrast between ridges and valley.

1. *Histogram equalization*: It expands the pixel value distribution in the range of 0–255, so that visualization is enhanced.
2. *Fast Fourier transform to enhance footprint image*: Footprint image is divided into small blocks of  $8 \times 8$ ,  $16 \times 16$ , and  $32 \times 32$  pixels to remove noise. Fourier transform is given by

$$f(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-2j\pi(\frac{ux}{M} + \frac{vy}{N})} \quad (1)$$

for  $u = 0, 1, \dots, 31$  and  $v = 0, 1, \dots, 31$ . To enhance the specific block by its frequency, the FFT of the block will be multiplied by its magnitude. The magnitude of the original FFT is given by  $\text{abs}(F(u, v)) = |F(u, v)|$ . The enhanced block is obtained by

$$g(x, y) = F^{-1} \{ F(u, v) \times |F(u, v)|^k \} \quad (2)$$

and  $F^{-1}(F(u, v))$  is

$$f(x, y) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(u, v) e^{\left(2j\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)\right)} \quad (3)$$

for  $x = 0, 1, \dots, 31$  and  $y = 0, 1, \dots, 31$ . The value of  $k$  is constant and determined by experiment, we have chosen  $k = 0.45$  empirically. A very small value of  $k$  creates small gaps and a very large value of  $k$  contributes in joining false ridges.

3. *Image binarization:* Input image is transformed from 8-bit gray image to the value in the range of [0, 1]. Ridges are represented by black color and valleys are represented by white color. Local adaptive binarization method is used to decide threshold.
4. *Image segmentation:* We cannot use the whole foot finger image to extract minutiae because it has background information as well, and hence extraction of ROI is very important to reduce the complexity. To extract the ROI, a two-step method is used: (1) block direction estimation and (2) direction variety check.
5. *Block direction estimation:* Direction of each block of footprint image is estimated. The size of the block should be  $W \times W$  and the size of  $W$  as default is taken as 16. Steps are given below:
  - To compute the gradient of each pixel block along x-axis ( $g_x$ ) and y-axis ( $g_y$ ) Sobel filter is used.
  - A nonlinear least square approximation is done for all the pixels of each block as

$$\theta = \frac{1}{2} \tan^{-1} \left[ \frac{\sum_{i=1}^W \sum_{j=1}^W 2g_x(i, j) * g_x(i, j)}{\sum_{i=1}^W \sum_{j=1}^W 2g_x^2(i, j) - g_y^2(i, j)} \right] \quad (4)$$

which is formulated by

$$\theta = \frac{1}{2} \tan^{-1} \sin\theta * \cos\theta / (\cos^2\theta - \sin^2\theta) \quad (5)$$

Later, the blocks which do not have significant information on ridge and valleys are discarded and this process is called as calculation of certainty given by

$$E = \frac{2 \sum \sum (g_x g_y) + \sum \sum (g_x^2 - g_y^2)}{W^2 \sum \sum (g_x^2 + g_y^2)} \quad (6)$$

- If the value of certainty  $E$  is less than the decided threshold then the block will be known as background block. The value of  $E$  should be greater than 0.05.
- 6. *Extraction of region of interest:* OPEN and CLOSE are the morphological operations which are used for the extraction of region of interest. The peak generated by background noise is removed by OPEN operation. Close operation fills the

generated small cavities, which gives only the relevant area and the bound of the fingerprint image. Subtraction of closed area to open area is known as bound. To get the region only containing the bound and inner area, we discard the leftmost, rightmost, uppermost, and bottommost blocks.

- *Minutiae extraction:* It involve two steps: (1) ridge thinning and (2) minutiae marking.

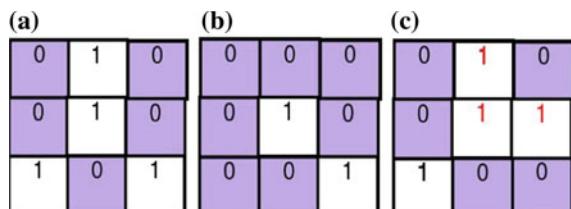
1. *Ridge thinning:* Iterative removal of redundant pixels until it is not left with single pixel is the process of ridge thinning. A window of  $3 \times 3$  is chosen as the structuring element for thinning followed by pruning to remove spikes, island points, and H break.

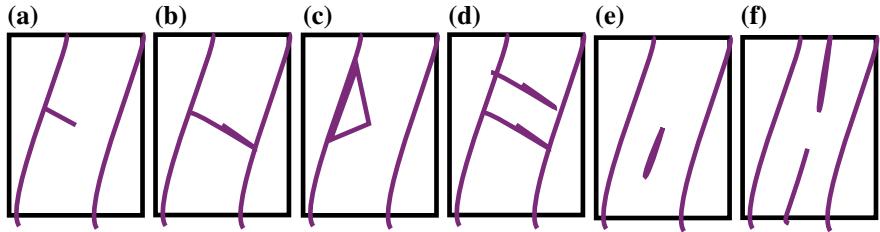
2. *Minutiae marking:* Minutiae marking is performed by the concept of crossing number ( $CN$ ). Check a  $3 \times 3$  neighborhood. If the central pixel is  $p = 1$ , and has exactly three one-value neighbors, then the central pixel is a ridge bifurcation, i.e.,  $CN(p) = 3$ . If the central pixel is 1 and has only one one-value neighbor, then the central pixel is a ridge ending, i.e.,  $CN(p) = 1$  for pixel  $p$ .

Figure 5a depicts the bifurcation, Fig. 5b depicts the termination, and Fig. 5c depicts the triple counting edge. When two vertical pixels have the value 1 and the rightmost pixel from second vertical pixel has neighbor outside the  $3 \times 3$  window then these two pixels will be marked as branches too. It adds a branch in the small region, which makes it necessary to check that none of the neighbors of a branch is added which is a spurious minutiae. Average distance between two neighboring ridges known as average inter-ridge width  $D$  is also computed by summing up the thinned ridge image having value 1. This sum is divided by row length to get inter-ridge width. Now all the inter-ridge widths are averaged to get the  $D$ . This marks the minutiae and all the thinned ridges in the fingerprint image are labeled with a unique ID for further operation.

- *Minutiae postprocessing:* Preprocessing step does not remove the noise completely and also intermediate steps add some noise which contribute to the construction of false minutiae. False minutiae are also added due to broken ridgelines and when two ridgelines get falsely connected forming bifurcation. Figure 6 shows the types of false minutiae. Figure 6a shows a spike piercing into valley Fig. 6b shows a spike connecting two ridges. Figure 6c shows two bifurcation points located at the same ridge. Figure 6d shows two spikes piercing into valley. Figure 6e represents one short ridge and Fig. 6f represents two ridge broken point. False minutiae affect

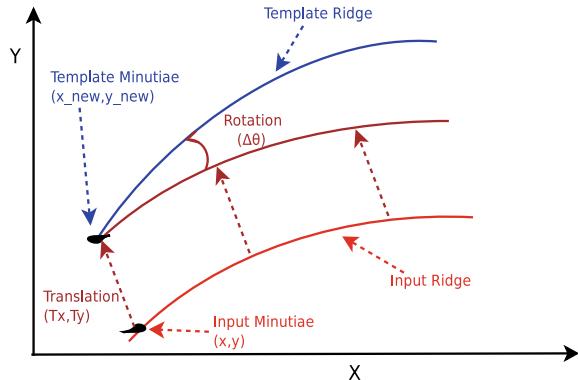
**Fig. 5** Minutiae marking: **a** bifurcation, **b** termination, **c** triple counting branch





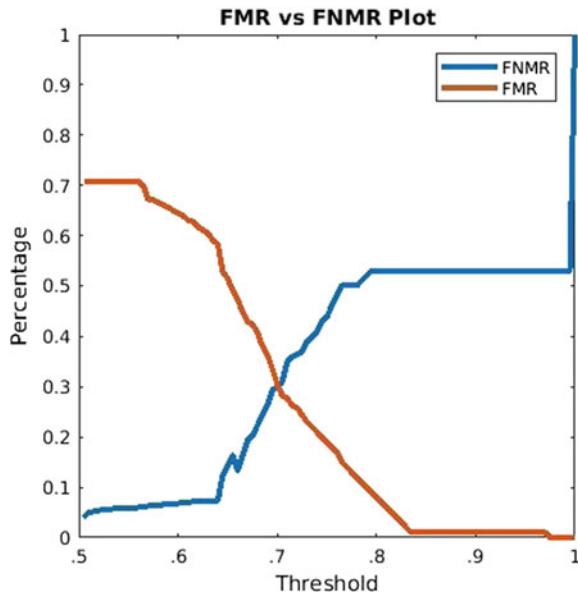
**Fig. 6** Types of false minutiae

**Fig. 7** Two minutiae points are translated and rotated for the matching

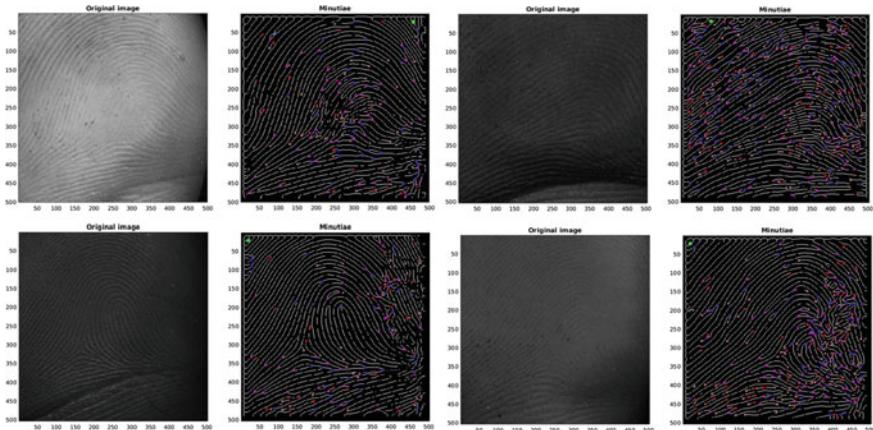


the accuracy of algorithm. Hence, detection and removal of false minutiae are important, and we have removed false minutiae using value of  $D$  as proposed by [15].

- *Minutiae matching:* Minutiae matching is done by aligning two footprint images. One minutiae point will be chosen as reference point and now the similarity score will be computed by taking each and every point. Minutiae will be transformed to a new xy coordinate. This process will be repeated until we find the maximum similarity score. This similarity score will be computed by using Euclidean distance as similarity metric. Translation and rotation operations are used to transform minutiae (Fig. 7).



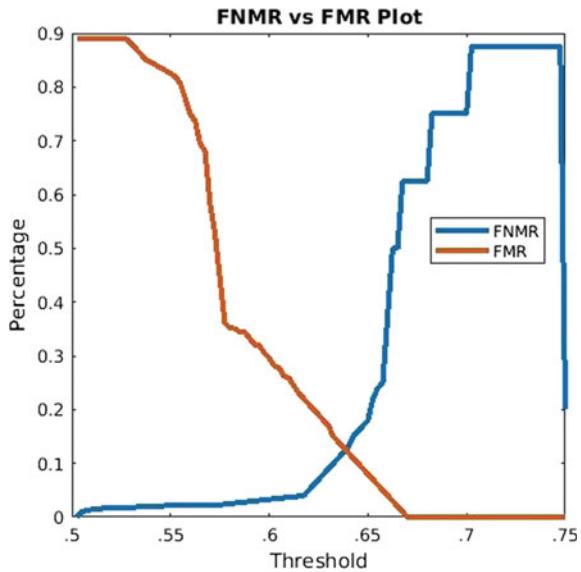
**Fig. 8** Experiment 1: plot between FMR and FNMR versus threshold



**Fig. 9** Test results show the cases where the algorithm is not able to recognize foot thumb images due to imperfection in dataset collection

The alignment-based matching algorithm is sufficient to find the similarity score between two minutiae points. The algorithm works efficiently. Figure 9 shows the cases where our algorithm does not identify well. In the first image, minutiae feature got flattened due to pressure applied on the foot. Second image shows bad skin condition. Third image shows the uneven lightening conditions and the fourth case has both uneven lighting condition and flat ridges.

**Fig. 10** Experiment 2: plot between FMR versus threshold and FNMR versus threshold



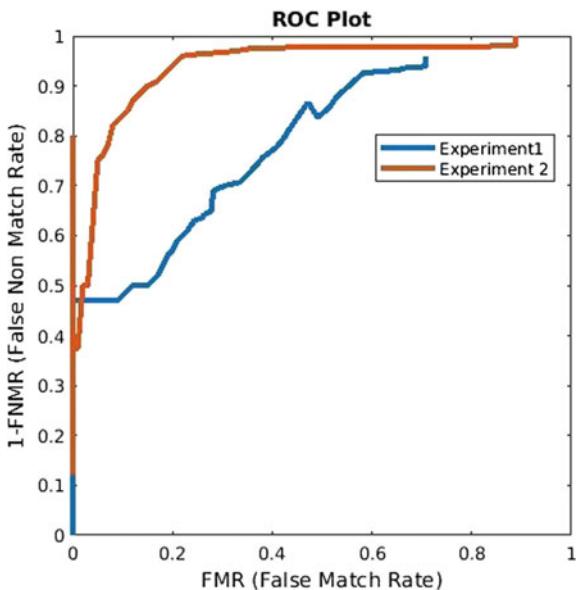
## 5 Performance Evaluation

Captured dataset is of 480 foot samples of 120 people. It involves two samples of each foot. The database was collected over a period of 2 weeks without any interval difference. The dataset size is increased by applying data augmentation technique (rotation by 15° and 30°) to get six samples of each foot.

**Experiment 1:** We used (83%) of the data for training, i.e., five samples of 100 people for training (500 samples). One sample of each (100) was used for testing. One person was randomly selected from the set of 100 people and enrolled in the database as a client. The remaining 99 people were considered impostors. Total number of genuine attempts are  $100 \times C_5^2 = \frac{(100 \times 99)}{(3! \times 2!)} = 1000$  and total number of imposter attempts are  $5 \times 5 \times C_{100}^2 = \frac{(100 \times 99 \times 25)}{(98! \times 2!) = 123750}$ . Figure 8 shows that the system achieves an EER (Equal Error Rate) of  $FMR = FNMR = 0.3\%$  at threshold  $T = 0.7$ . A minimum TER (Total Error Rate,  $TER = FMR + FNMR$ ) of 0.6% is achieved at  $T = 0.692$ . Our system can achieve an  $FMR = 0.0\%$  with an  $FNMR = 0.53\%$  with  $T = 0.83$ . We could not achieve an  $FNMR = 0.0\%$ . We get  $FNMR = 0.2\%$  with the  $FMR = 0.4\%$ .

**Experiment 2:** For this experiment, we have chosen best two samples from 50 people (100 samples). One person was randomly selected from the set of 50 people to act as a client and was enrolled in the database. The remaining 49 people were considered impostors. The results were recorded after performing the identification. Total number of recorded genuine attempts are 50 and total number of imposter attempts are 4900. Figure 10 shows that the system achieves an EER (equal error rate) of  $FMR = FNMR = 0.12\%$  at threshold  $T = 0.645$ .

**Fig. 11** ROC: plot between FMR and FNMR for experiment 1 and experiment 2



A minimum *TER* (*Total Error Rate*,  $TER = FMR + FNMR$ ) of 0.24% is achieved with  $T = 0.645$ . Our system can achieve an  $FMR = 0.0\%$  with an  $FNMR = 0.48\%$  with  $T = 0.66$ . We can achieve  $FNMR = 0.0\%$  at  $FMR = 0.9\%$  at  $T = 0.51$ , and  $FNMR = 0.2\%$  with the  $FMR = 0.1\%$ .

Figure 11 shows the ROC of two experiments. Experiment 1 with all images, it shows Genuine Match Rate (1-False Non-match Rate) 0.5% only with  $FMR = 0.2\%$ . As we observe, with the increase in Genuine Match Rate the False Match Rate also increases. For experiment 2, we have considered only the best samples from the database. Observed value of Genuine Match Rate (1-False Non-match Rate) 0.9% with  $FMR = 0.2\%$ . We also report the  $FAR@FRR = 2.0\%$  for both the experiments as shown in Table 1.

**Table 1** Value of False Accept Rate (FAR) at 2% False Reject Rate (FRR)

	$FAR@FRR = 2.0\% (0.2\%)$
Experiment 1	0.42%
Experiment 2	0.13%

## 6 Conclusions and Future Work

We conclude that the reliability of minutiae-based footprint identification primarily relies on the size of the dataset, sample quality which depends on sensor capabilities and extraction of correct minutiae points. Dataset collection of foot samples is presently difficult: (a) no traditional footprint-capture sensor, (b) uneven pressure on sensor platen, (c) cracked heels, (c) dirt in footprint, and (d) removal of footwear contribute to inconvenience to subjects. Given these limitations and complications we have captured 480 images for the analysis. To extract minutiae, image segmentation is done using morphological operation, triple counting branch is considered for minutiae localization to increase the precision. Two more things can be added to improve the accuracy. (1) We can increase the hardware efficiency which might give us more accuracy because of better image quality. (2) Image enhancement technique could be improved for the better extraction of ROI. These things can surely improve the accuracy of the system.

**Acknowledgements** The necessary equipments/tools for dataset collection is provided by computer science department of Malaviya National Institute of Technology, Jaipur, India. This dataset has been collected with the full consent of all the participants.

## References

1. Alonso-Fernandez, F., Bigun, J., Fierrez, J., Fronthaler, H., Kollreider, K., Ortega-Garcia, J.: Fingerprint recognition. In: Guide to Biometric Reference Systems and Performance Evaluation, pp. 51–88. Springer, London (2009)
2. Barker, S., Scheuer, J.: Predictive value of human footprints in a forensic context. *Med. Sci. Law* **38**, 341–346
3. Cappelli, R., Ferrara, M.: A fingerprint retrieval system based on level-1 and level-2 features. *Exp. Syst. Appl.* **39**(12), 10465–10478 (2012)
4. Hong, L., Wan, Y., Jain, A.: Fingerprint image enhancement: algorithm and performance evaluation. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(8), 777–789 (1998). <https://doi.org/10.1109/34.709565>
5. Kennedy, R.B.: Uniqueness of bare feet and its use as a possible means of identification. *Forensic Sci. Int.* **82**(1), 81–87 (1996)
6. Khokher, R., Singh, R.C., Kumar, R.: Footprint recognition with principal component analysis and independent component analysis. In: Macromolecular Symposia. vol. 347 (2015)
7. Kumar, V.A., Ramakrishnan, M.: Manifold feature extraction for foot print image. *Indian J. Bioinform. Biotechnol.* **1**, 28–31
8. Kumar, V., Ramakrishnan, M.: Employment of footprint recognition system. *Indian J. Comput. Sci. Eng. IJCSE* **3**(6) (2013)
9. Kushwaha, R., Nain, N.: Facial expression recognition. *Int. J. Curr. Eng. Technol.* **2**(2), 270–278 (2012)
10. Kushwaha, R., Nain, N., Gupta, S.K.: Person identification on the basis of footprint geometry. In: 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 164–171. IEEE (2016)
11. Kushwaha, R., Nain, N., Singal, G.: Detailed analysis of footprint geometry for person identification. In: 2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 229–236. IEEE (2017)

12. Liu, E., Cao, K.: Minutiae extraction from level 1 features of fingerprint. *IEEE Trans. Inf. Forensics Secur.* **11**(9), 1893–1902 (2016)
13. Nagwanshi, K.K., Dubey, S.: Biometric authentication using human footprint. *Int. J. Appl. Inf. Syst. IJAIS* **3**, 1–6
14. Nakajima, K., Mizukami, Y., Tanaka, K., Tamura, T.: Footprint-based personal recognition. *IEEE Trans. Biomed. Eng.* **47**, 1534–1537
15. Sharma, K., Rajput, R.: Spurious minutia removal technique using euclidean distance approach. *Int. J. Eng. Comput. Sci.* **3**(11) (2014)
16. Tatar, F., Machhout, M.: Improvement of the fingerprint recognition process
17. Uhl, A., Wild, P.: Footprint-based biometric verification. *J. Electron. Imaging* **17**, 011016
18. Weingaertner, D., Bellon, O.R.P., Silva, L., Cat, M.N.: Newborn's biometric identification: can it be done? In: VISAPP, pp. 200–205 (2008)

# Target Tracking Based Upon Dominant Orientation Template and Kalman Filter



Nikhil Kumar, Puran Dhakrey and Neeta Kandpal

**Abstract** Due to the proliferation of surveillance systems in defense as well as civilian sectors, target tracking field has always remained adorable to researchers of the computer vision community. Template matching based approaches are a conventional way to solve target tracking problems. Ironically such approaches are computationally complex due to a requirement of repetitive arithmetic operations. A number of attempts have been made to reduce this computational overhead; Dominant Orientation Template (DOT) based template matching has given a new dimension to solve problems of such a kind using logical operations only. It has been observed that the performance of template matching based approaches degrades drastically in the presence of occlusion. DOT-based template matching is also not an exception to this. Proposed approach introduces a novel target tracking framework having severe occlusion handling capacity by integrating DOT-based template matching and Kalman Filtering. In the present approach, a small target search window is formulated around the Kalman estimated location of the considered frame; rather than exploring the whole frame. DOT-based template matching works inside this window only. A parameter known as Kalman error is defined here which is considered as the measure of occlusion. Initially, DOT calculated location is assumed as measured value but in case of occlusion when this Kalman error becomes sufficiently large the Kalman estimated location is regarded as the measurement. A customized dataset having occluded targets in FLIR as well as visible video sequences is generated to provide a robust testbed for the proposed approach.

**Keywords** Dominant Orientation Template (DOT) · Kalman Filter · Target tracking · Template matching · Occlusion · Gradient orientation

---

N. Kumar (✉) · P. Dhakrey · N. Kandpal  
Instruments Research and Development Establishment, Defence Research  
and Development Organization, Dehradun 248008, India  
e-mail: [nikhilkumar@irde.drdo.in](mailto:nikhilkumar@irde.drdo.in)

## 1 Introduction

Present work is an attempt to solve the challenging problem of target tracking by combining DOT [5] based template matching and Kalman Filter [10] on one platform. Template matching based approaches rely on brute force practices of target detection, in such approaches the best matching location of template is explored in any image by calculating any similarity measure like cross-correlation [15] or SSD (sum of the square of distances) [13]. The problem of template matching got a promising direction by the introduction of DOT [6] based template matching approach in which orientation of gradients of the considered image is quantized to introduce a compact binary representation. The beauty of this compact representation is that one can get rid of lengthy mathematical expressions for computing the correlation between image and template. One can understand this as the only logical operation based correlation measurement. Although in terms of computational overhead it's alone a very promising scheme; yet its poor occlusion handling capacity makes it a measurable choice for researchers. In a real-world scenario, occlusion is very common. To handle such scenarios a Kalman Filter based approach is integrated [16] in the present framework which runs simultaneously with DOT-based template matching. For speeding up template matching operation, a search window is formulated inside the considered image for measuring the similarity. As the size of the chosen search window is sufficiently smaller than image and a couple of times bigger than the size of the template, hence a lot of computation is reduced. The center of this search window is estimated with the help of Kalman Filter. DOT-based template matching is utilized for searching the exact location of the target inside the window; this data is used as measurement during correction step [4] of Kalman Filter. If the discrepancy between Kalman estimated location and DOT calculated location is more than a threshold value it is assumed that either there is occlusion or DOT-based detection is being failed due to lack of strong gradients or presence of clutter; in such scenarios Kalman estimated location is treated as measurement till next availability of exact target location from DOT-based template matching. In this way even if DOT-based detection fails for few frames, the performance of the algorithm does not get hampered.

In the remainder of paper first related work is discussed then in subsequent sections methodology of the current approach is explained.

## 2 Related Work

For target tracking, there has always remained search of an algorithm with robust performance and less computational overhead. Template matching based techniques have always remained an adjacent solution to such problems. A number of attempts have been made in the past to speed up template matching techniques.

Techniques using contour-based template representation and chamfer distance [3] as a dissimilarity measure remained very popular in previous years, but due to fragile and illumination-dependent contour extraction performance of these methods remained debatable.

Comaniciu and Meer [1] represented object with a weighted histogram computed from a circular region. In place of brute force, search similarity is maximized by calculating Bhattacharya distance between both histograms. At each iteration, a mean shift vector is computed which is the measure of similarity and iterations are repeated till convergence.

In HOG-based technique [2], local distribution of gradient information of the image is quantized on a regular grid. In spite of promising results, computational complexity always remained a hurdle in real-time implementation of this approach.

Recently, DOT has added a new dimension [8] in template matching domain by speeding it up significantly, but its performance in the cluttered and partially occluded scenario has always remained questionable [8]. Video sequences with lack of strong gradients have also remained a bottleneck in the performance of DOT-based techniques [5].

Target tracking in FLIR [7, 9], video sequences has remained a tough problem to solve due to considerable differences in IR [12] and visible signals [14]. There are a lot of shortcomings in IR signals like lack of color information, poor SNR [11]; resultant of all these artifacts generates the scarcity of strong gradient information.

### 3 Methodology

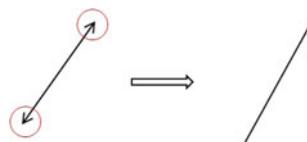
The methodology adopted is elaborated in the following sections.

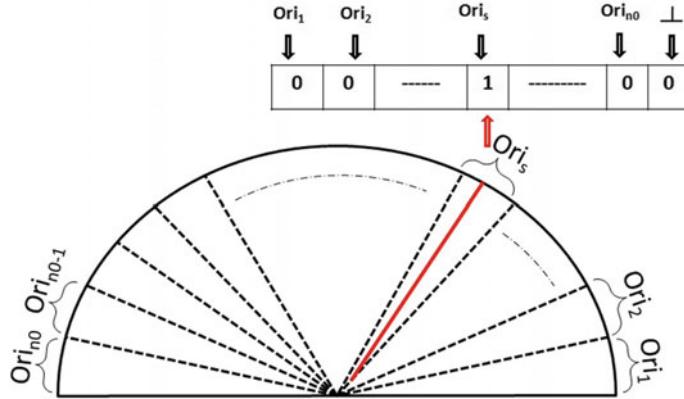
#### 3.1 DOT-Based Template Matching

The target location in any  $k$ th frame  $\{x_{dot}(k), y_{dot}(k)\}$  can be evaluated by following steps

1. Considering the fact that gradient information of an image is more immune and reliable than intensity values, gradient information of image and template is considered for further processing.
2. From gradient orientation map of the considered image, the direction is omitted as in Fig. 1.

**Fig. 1** Ignoring direction vector from orientation





**Fig. 2** Quantization of orientation map of considered image  $I$  and compact binary representation of one location

3. As in Fig. 2, orientation space of image is quantized in  $n_0$  equal divisions, where  $ori_s$  is  $s$ th quantization level of gradient orientation. In this way all locations of image are assigned  $n_0 + 1$  bit binary number.
4. An image  $I$  having size  $M \times N$  is divided in similar square regions of size  $n_0 \times n_0$ . Consider any square region  $\zeta_{i,j}$ ,  $\forall i \in \{0, 1, 2, 3, \dots, m-1\}$  and  $\forall j \in \{0, 1, 2, 3, \dots, n-1\}$  within image  $I$

where

$$m = \lfloor \frac{M}{n_0} \rfloor \quad (1)$$

and

$$n = \lfloor \frac{N}{n_0} \rfloor \quad (2)$$

There will be  $m \times n$  such similar square regions of size  $n_0 \times n_0$ . If  $\mu_{i,j}$  and  $\psi_{i,j}$  are sets of magnitudes and quantized binary orientations for square region  $\zeta_{i,j}$  defined as following:

$$\mu_{i,j} = \{\mu_{ij,1}, \mu_{ij,2}, \dots, \mu_{ij,n_0^2}\} \quad (3)$$

and

$$\psi_{i,j} = \{\psi_{ij,1}, \psi_{ij,2}, \dots, \psi_{ij,n_0^2}\} \quad (4)$$

where any element  $\psi_{ij,r}$ ,  $\forall r \in \{1, 2, 3, \dots, n_0^2\}$  is quantized orientation data represented as  $n_0 + 1$  bits binary number with all zeros except  $(n_0 - r)_{th}$  bit as 1. Then

$$DOIMG_{i,j} = \begin{cases} \perp & \text{if } \max[\mu_{ij}] < \tau \\ \psi_{ij,t} & \text{otherwise, } \forall t: \mu_{ij,t} = \max[\mu_{ij}] \end{cases}$$

where  $DOIM G_{i,j}$  is introduced dominant orientation representation of  $ij_{th}$  region of the considered image  $I$ ,  $\tau$  is a threshold magnitude and  $\perp$  is a  $n_0 + 1$  bits binary number with all zeros except LSB as 1.

5. Similarly any template  $T$  having size  $P \times Q$  is divided in similar square regions of size  $n_0 \times n_0$ . Let  $\xi_{i',j'}, \forall i' \in \{0, 1, 2, 3, \dots, p-1\}$  and  $\forall j' \in \{0, 1, 2, 3, \dots, q-1\}$  be any such square region. Where,

$$p = \lfloor \frac{P}{n_0} \rfloor \quad (5)$$

and

$$q = \lfloor \frac{Q}{n_0} \rfloor \quad (6)$$

There will be  $m \times n$  such similar square regions of size  $n_0 \times n_0$ . If  $\varphi_{i',j'}$  is a set of quantized binary orientations for square region  $\xi_{i',j'}$  defined as follows:

$$\varphi_{i',j'} = \{\varphi_{i',j',1}, \varphi_{i',j',2}, \dots, \varphi_{i',j',n_0^2}\} \quad (7)$$

where for any  $\varphi_{i',j',s}, \forall s \in \{1, 2, 3, \dots, n_0^2\}$  is quantized orientation data represented as  $n_0 + 1$  bits binary number with all zeros except  $(n_0 - r)$ th bit as 1. Then

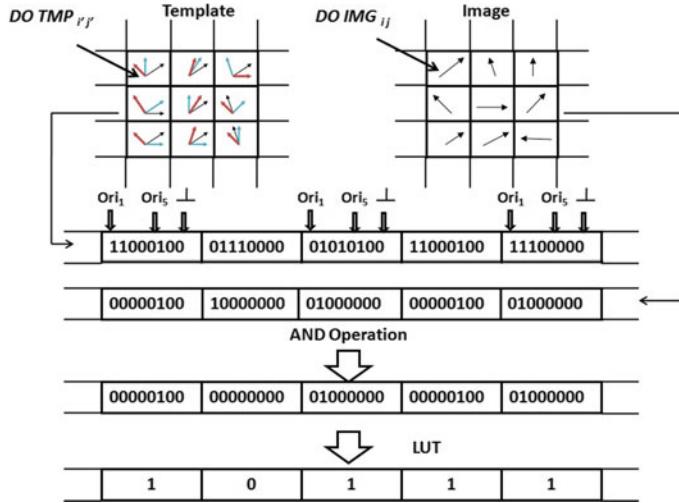
$$DO TMP_{i',j'} = \{\varphi_{i',j',1} OR \varphi_{i',j',2} OR \dots OR \varphi_{i',j',n_0^2}\} \quad (8)$$

Where  $DO TMP_{i',j'}$  is introduced dominant orientation representation of  $i'j'$ th region of the considered template  $T$ , and  $OR$  is bitwise logical  $OR$  operation.

6. In this way, all square regions of image  $I$  and template  $T$  are presented in binary dominant orientation representation.
7. As in Fig. 3 logical  $AND$  operation is performed by sliding template over all locations of the window and a lookup table is generated by masking nonzero elements.
8. In this way the target location  $\{x_{dot}(k), y_{dot}(k)\}$  in any  $k$ th frame of the considered video sequence can be measured by thresholding generated lookup table.

### 3.2 Kalman Filtering

The Kalman Filter is a predictor-corrector type optimal estimator [4], which can predict the location of the considered target in next frame based upon motion model and its locations in past. In the current approach, Kalman filter is utilized for estimating



**Fig. 3** Matching score calculation

target co-ordinates in the next frame. These estimated co-ordinates are used as the center of a search window for DOT-based template matching. For implementing the Kalman Filter for target tracking, there is a need to model the motion of the considered target. Its trivial to mention that accurate modeling of an object's motion in a realistic scenario is next to impossible, but the beauty of the Kalman Filter is that an acceptable approximation of the motion also shows remarkable performance. Its necessary to mention the presumption that motion along  $X$  and  $Y$  axes are totally uncorrelated and rotational variation in pose is neglected. The motion of target can be described as state-space equation 9.

$$\begin{bmatrix} x(k+1) \\ y(k+1) \\ v_x(k+1) \\ v_y(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x(k) \\ y(k) \\ v_x(k) \\ v_y(k) \end{bmatrix} + \begin{bmatrix} \frac{T_s^2}{2} & 0 \\ 0 & \frac{T_s^2}{2} \\ T_s & 0 \\ 0 & T_s \end{bmatrix} \times \begin{bmatrix} a_x(k) \\ a_y(k) \end{bmatrix} \quad (9)$$

Assume that  $\{x(k), y(k)\}$  represent location of target,  $\{v_x(k), v_y(k)\}$  represent  $X$  and  $Y$  components of velocity and  $\{a_x(k), a_y(k)\}$  represent  $X$  and  $Y$  components of acceleration in  $k$ th frame  $\forall k \in \{1, 2, 3, \dots, F\}$  where considered video sequence has  $F$  frames and  $\frac{1}{T_s}$  frame rate. Above state space model can be represented as follows:

$$X(k+1) = AX(k) + Bu(k) \quad (10)$$

To handle uncertainty generated by inaccuracy of the model, white noise can be introduced in this model in the following manner:

$$X(k+1) = AX(k) + Bu(k) + w \quad (11)$$

where  $w$  is white noise

$$P(w) \sim N(0; Q) \quad (12)$$

where  $Q = BQ_bB^T$  and

$$Q_b = \begin{bmatrix} \sigma_{a_x}^2 & 0 \\ 0 & \sigma_{a_y}^2 \end{bmatrix} \quad (13)$$

Since there is a measurement data generated from DOT-based template matching, this data can be expressed as follows:

$$Y(k+1) = CX(k+1) + \nu \quad (14)$$

where

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (15)$$

and  $\nu$  is measurement noise introduced during measurement process.

$$P(\nu) \sim N(0; R) \quad (16)$$

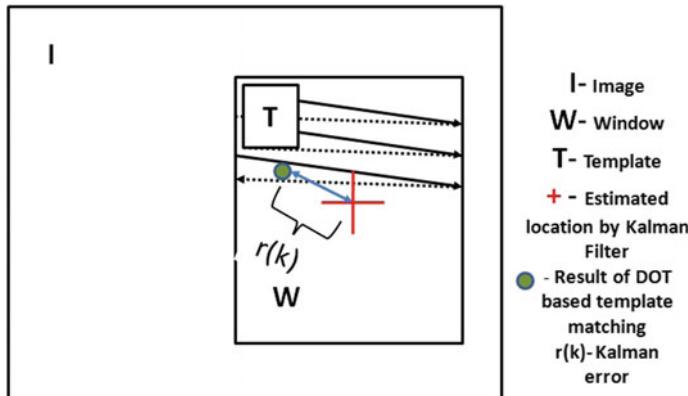
where  $R = CVC^T$  and

$$V = \begin{bmatrix} \sigma_{z_x}^2 & 0 \\ 0 & \sigma_{z_y}^2 \end{bmatrix} \quad (17)$$

Using above equations and measured location of the target from DOT-based template matching  $\{(x_{dot}(k), y_{dot}(k))\}$  in any  $k$ th frame Kalman Filter can be formulated and location of target  $\{x_{kalm}(k+1), y_{kalm}(k+1)\}$  in  $k+1$ th frame of the considered sequence can be estimated.

### 3.3 Proposed Approach

An overview of the adopted methodology is depicted in Fig. 4. First Kalman filter estimates the location of a probable target in the current frame from its location in previous frames, the search window is formulated around this location and DOT-based template matching is performed on this window. As DOT-based template matching is dependent only upon logical AND operations and chosen search window is relatively smaller than image hence computational overheads are reduced significantly



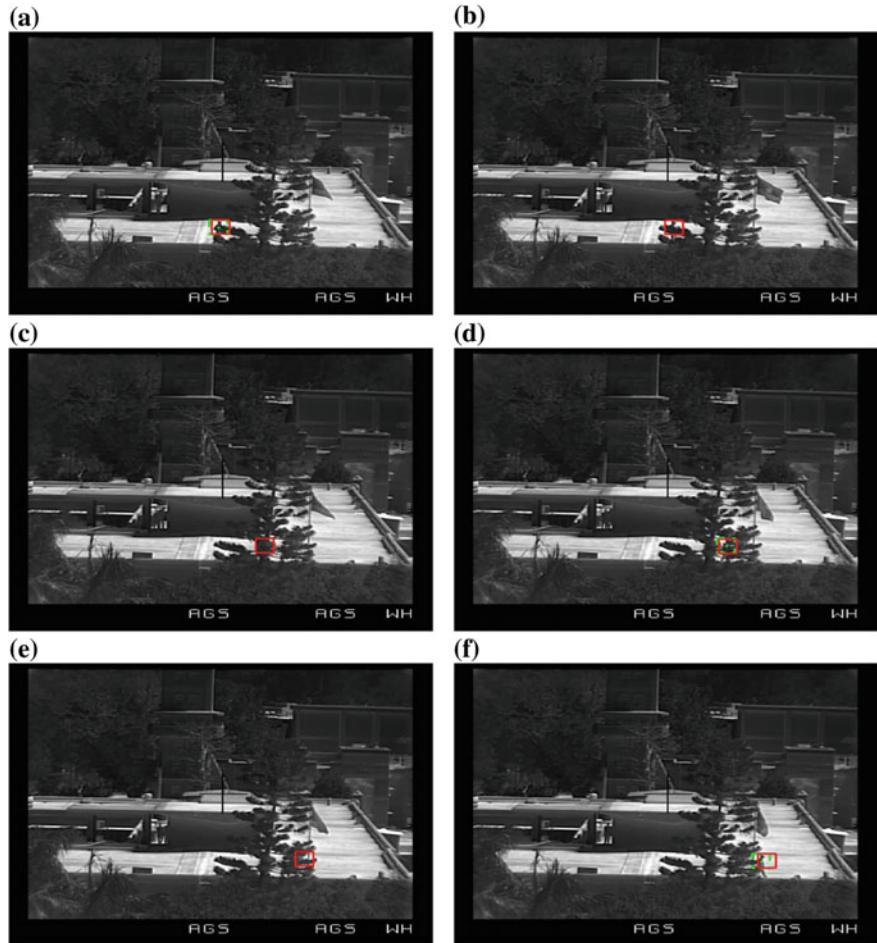
**Fig. 4** Methodology of the proposed approach

as compared to conventional template matching approaches. Let  $\{x_{kalm}(k), y_{kalm}(k)\}$  is estimated target location in the current frame and  $\{x_{dot}(k), y_{dot}(k)\}$  represent measured target location by DOT-based template matching then Kalman error [16]  $r(k)$  may be defined as following Eq. 18

$$r(k) = \sqrt{(x_{kalm}(k) - x_{dot}(k))^2 + (y_{kalm}(k) - y_{dot}(k))^2} \quad (18)$$

This parameter plays an important role in occlusion handling. Proposed approach is elaborated as follows

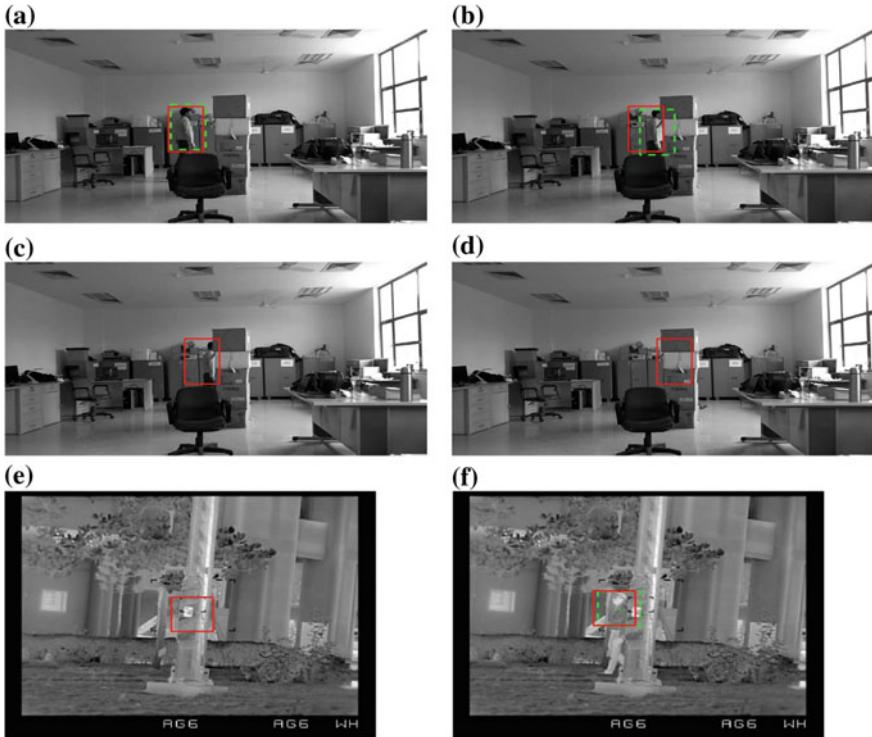
1. Kalman Filter estimates the location of a probable target in the considered frame.
2. A search window is formulated around this estimated location. Size of this search window may be chosen relatively smaller than image and larger than template.
3. DOT-based binary template matching is performed inside this window by sliding template at all locations.
4. As in Fig. 3  $\{x_{dot}(k), y_{dot}(k)\}$  are calculated inside window.
5.  $r(k)$  is calculated as per Eq. 18.
6. If  $r(k) < \lambda$  then consider  $\{x_{dot}(k), y_{dot}(k)\}$  as measured data while estimating target location in next frame through Kalman Filter. Here  $\lambda$  is a threshold distance which can be chosen experimentally.
7. If  $r(k) > \lambda$  then consider  $\{x_{kalm}(k), y_{kalm}(k)\}$  as measured data while estimating target location in next frame through Kalman Filter.
8. For cases when  $\{x_{dot}(k), y_{dot}(k)\}$  is not defined for example failure of DOT- based template matching due to lack of sufficient gradient information or clutter then also consider  $\{x_{kalm}(k), y_{kalm}(k)\}$  as measured data while estimating target location in next frame through Kalman Filter.



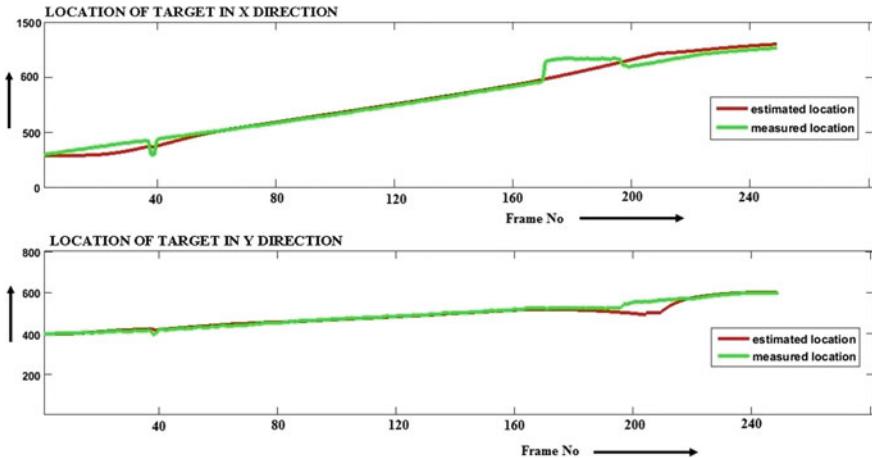
**Fig. 5** Results (Rectangle with red solid lines is representing target detected using the proposed approach, rectangle with broken green lines is representing target detected using DOT-based template matching alone) of the proposed approach on ROOF-JR sequence captured using an indigenous FLIR camera developed at IRDE, Dehradun having  $640 \times 512$  InSb detecting elements working in MWIR **a** target is detected by the proposed approach and DOT both, **b** due to lack of strong gradients and presence of background clutter DOT alone fails but the proposed approach is working, **c** although the target is under full occlusion, yet the proposed approach is working, **d** target is partially occluded and both approaches are working, **e** target is partially occluded and only the proposed approach is working and **f** target is coming out of occlusion and results of both approaches are converging

## 4 Results

Although DOT-based detection and use of Kalman Filter for target tracking are very popular techniques independently this is merely the first time that both approaches are integrated into one common framework. A dataset of IR video sequences is generated for establishing a testbed to present approach. Out of 67 frames of *PARK-IR* sequence as in Fig. 6. DOT-based template matching alone fails in 18 frames due to occlusion or clutter but the proposed approach has detected target almost in each frame. In case



**Fig. 6** Results (*Rectangle with red solid lines is representing target detected using the proposed approach, rectangle with broken green lines is representing target detected using DOT-based template matching alone*) of proposed approach on *ROOM-VIS* sequence **a** target is detected by the proposed approach and DOT-based template matching both, **b** target is partially occluded and both approaches are working, **c** target is partially occluded and only the proposed approach is working, **d** although the target is under full occlusion, yet the proposed approach is working. Results (*Rectangle with red solid lines is representing target detected using the proposed approach, rectangle with broken green lines is representing target detected using DOT-based template matching alone*) of proposed approach on *PARK-IR* sequence captured using an indigenous FLIR camera developed at IRDE, Dehradun having  $640 \times 512$  InSb detecting elements working in MWIR **e** target is 100% occluded and only the proposed approach is working, **f** target is detected by the proposed approach and DOT-based template matching both



**Fig. 7**  $X$  and  $Y$  co-ordinates of estimated and measured target locations from *IR-CAR* sequence having 250 frames of size  $1280 \times 720$

of full occlusion, if the estimated location is considered as true positive, it converges towards the actual location. Out of 146 frames of *ROOF-IR* sequence as in Fig. 5. DOT-based template matching alone fails in 50 frames due to occlusion or clutter but proposed approach has detected target almost in each frame again. Figure 7 is depicting  $X$  and  $Y$  coordinates of measured and estimated locations of target in *IR-CAR* sequence. In this sequence, DOT-based template matching alone fails in 48 frames out of the total 250 frames. It can be concluded from these results that the performance of the proposed approach under clutter and occlusion is remarkable.

## 5 Conclusion

In this paper, DOT-based template matching and Kalman Filtering is used in a common framework. As template matching depends only upon logical operations and the brute force search is done in a small window of the considered frame, computational overhead is reduced significantly as compared to conventional template matching approaches. If the performance of the present approach is compared with DOT-based template matching alone, then it can be concluded that there is a remarkable gain in target detection especially in those scenarios where the target is severely occluded. Even in the case of IR sequences where it is assumed that scarcity of strong gradients is common, the proposed approach is detecting the targets almost in each frame. The results are promising and since the model is extremely simple, it can be applied to real-time tracking applications.

**Acknowledgements** The authors take this opportunity to express their sincere gratitude to Mr. Benjamin Lionel, Director IRDE for his constant motivation and support as well as permission to publish this work. He has always inspired the authors towards innovation and adopting a creative and simple approach for solving difficult problems.

## References

1. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: IEEE Conference on Computer Vision and Pattern Recognition, 2000. Proceedings, vol. 2, pp. 142–149. IEEE (2000)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
3. Gavrila, D.M., Philomin, V.: Real-time object detection for “smart” vehicles. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, vol. 1, pp. 87–93. IEEE (1999)
4. Grewal, M.S. Kalman filtering. In: International Encyclopedia of Statistical Science, pp. 705–708. Springer, Berlin (2011)
5. Hinterstoisser, S., Lepetit, V., Ilic, S., Fua, P., Navab, N.: Dominant orientation templates for real-time detection of texture-less objects, 2257–2264 (2010)
6. Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, Vincent: Gradient response maps for real-time detection of textureless objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(5), 876–888 (2012)
7. Holst, Gerald C.: Common sense approach to thermal imaging. SPIE Optical Engineering Press, Washington (2000)
8. Hong, C., Zhu, J., Song, M., Wang, Y.: Realtime object matching with robust dominant orientation templates. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1152–1155. IEEE (2012)
9. Hudson, R.D.: Infrared System Engineering, vol. 1. Wiley-Interscience, New York (1969)
10. Kalman, R.E.: A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**(1), 35–45 (1960)
11. Kumar, N., Venkatesh, K.S., Namboodiri, V.P.: Regularity flow inspired target tracking in FLIR imagery. In: 2016 27th Irish Signals and Systems Conference (ISSC), pp. 1–6. IEEE (2016)
12. Kumar, N., Kumar, A., Kandpal, N.: Video synopsis for IR imagery considering video as a 3D data cuboid. In: Proceedings of International Conference on Computer Vision and Image Processing, pp. 227–237. Springer, Singapore (2017)
13. Martin, J., Crowley, J.L.: Comparison of correlation techniques. In: Intelligent Autonomous Systems, pp. 86–93 (1995)
14. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 1 (2013)
15. Yilmaz, A., Javed, O., Shah, M.: Object tracking: a survey. *ACM Comput. Surv. (CSUR)* **38**(4), 13 (2006)
16. Zhao, J., Qiao, W., Men, G.-Z.: An approach based on mean shift and Kalman filter for target tracking under occlusion. In: 2009 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2058–2062. IEEE (2009)

# Robust Single Image Super Resolution Employing ADMM with Plug-and-Play Prior



V. Abdu Rahiman and Sudhish N. George

**Abstract** Image super resolution is a signal processing technique to post-process a captured image to retrieve its high-resolution version. Majority of the conventional super resolution methods fail to perform in presence of noise. In this paper, a noise robust reconstruction based single image super resolution (SISR) algorithm is proposed, using alternating direction method of multipliers (ADMM) and plug-and-play modeling. The plug-and-play prior concept is incorporated to the two variable update steps in ADMM. Therefore, a fast SISR model and a denoiser are used in ADMM to implement the proposed robust SISR scheme. The experimental results show that the noise performance of the proposed approach is better than the conventional methods. The impact of parameter selection on the performance of the algorithm is experimentally analyzed and the results are presented.

**Keywords** Single image super resolution • Plug-and-play prior • Noise robust

## 1 Introduction

High-resolution (HR) images are always desired in any image applications since they contain more details. However, HR images need not be available in all situations due to the limitations in capturing devices, etc. The process of reconstructing HR images from low-resolution (LR) images is called super resolution (SR). LR images have less details than corresponding HR images, thus super resolution algorithm will regenerate the HR images by incorporating these missing details [1]. Therefore, SR

---

Abdu Rahiman V thank TEQIP Phase II for the funding provided to attend the conference.

---

V. Abdu Rahiman ()  
Government College of Engineering Kannur, Kannur, India  
e-mail: [vkarahim@gmail.com](mailto:vkarahim@gmail.com)

S. George  
National Institute of Technology Calicut, Kozhikode, India  
e-mail: [sudhish@nitc.ac.in](mailto:sudhish@nitc.ac.in)

is an inverse problem to estimate the HR image from LR observations. Equation (1) gives a simplified image acquisition model.

$$\mathbf{y}_k = \mathbf{S}_k \mathbf{H}_k \mathbf{x} + \mathbf{n}_k \quad (1)$$

where  $\mathbf{x} \in \mathbb{R}^{M_h N_h \times 1}$  is the vectorized HR ground truth image of size  $M_h \times N_h$ ,  $\mathbf{y}_k \in \mathbb{R}^{M_l N_l \times 1}$  is the vectorized LR observations of size  $M_l \times N_l$ ,  $\mathbf{H}_k$  denotes a blurring operator,  $\mathbf{S}_k$  is the downsampling operator, and  $\mathbf{n}_k$  represents the noise effects on  $k$ th observation. Finding  $\mathbf{x}$  from  $\mathbf{y}_k$  is the problem of super resolution which has infinitely many solutions. There are different categories of super resolution algorithms, viz., single image super resolution (SISR), multi-frame super resolution (MFSR), face hallucination, text super resolution, etc.

Number of super resolution algorithms are reported in literature to address various aspects of the problem. But most of them fail to perform, if the input LR observation is affected with noise [2]. Objective of our research work is to develop noise robust image super resolution algorithm. In SISR, the number of low-resolution observation is limited to one, i.e.,  $k = 1$  and hence the problem of super resolution becomes more difficult. An SISR problem model is given in Eq. (2).

$$\mathbf{x} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{x}) \quad (2)$$

where  $R(\mathbf{x})$  is a regularization term used to limit the number of solutions and  $\lambda$  is the regularization parameter controlling the weightage of each term. Various methods were attempted for getting the solution of this problem.

Broad category of SISR methods include learning-based algorithms and reconstruction-based algorithms. In learning-based algorithm, the model parameters are computed from a set of LR-HR training image pairs [3]. Training-based methods generally show implicit dependence on the training data. Whereas in reconstruction-based methods tries to minimize the artifacts caused by undersampling, which does not require any training examples. Therefore, such algorithms will not have any dependence on training data. In this paper, a reconstruction-based approach is presented to recover  $\mathbf{x}$  from noisy observations.

## 1.1 Related Works

Super resolution is a well-studied research area and various algorithms are proposed in the literature. The presence of noise is one of the major challenges in SISR. Even though a vast number of techniques are available in the literature, the presence of noise is not well investigated. A few approaches were proposed to tackle the presence of noise in LR observation. The simplest one was to apply a denoising algorithm before performing the super resolution. But in this approach, the denoising step may destroy the useful details in the image which will limit the performance of SR algorithm.

Sparse representation-based super resolution algorithm proposed by Yang et al. [4] claims that it is robust to noise if the noise level is relatively less. But performance of this method degrades considerably as the noise level increases.

Rahiman et al. [5] proposed a learning-based super resolution and a noise robust version of the algorithm was presented in [6]. It was a denoised patch-based algorithm to perform simultaneous super resolution and noise removal. The hybrid wavelet fusion based super resolution approach (WFSR) proposed in [6] combines the result obtained from multiple super resolution algorithms to produce a better result. Super resolution using convolutional neural network (SRCNN) by Dong et al. [7] is a deep learning based method for SISR, which consist of three convolutional layers. These approaches are learning-based methods which require a training phase in which the model parameters are determined. ADMM [8] is a recently popular convex optimization algorithm. Many reconstruction-based super resolution techniques are proposed using ADMM. FSR-ADMM is a reconstruction-based fast robust SISR scheme proposed by Zhao et al. [9] which uses Tikhonov regularization and ADMM optimization to reconstruct the HR image. The concept of plug-and-play prior was introduced by Venkatakrishnan et al. [10] and Brifman et al. [11] proposed an ADMM-based robust super resolution algorithm which utilizes an image denoising step as a prior in the super resolution algorithm. Stability and convergence of the method was the major issue of this approach. Chan et al. [12] proposed the methods to improve the convergence of this algorithm. In [12], they proposed the conditions on denoising prior to improve convergence.

## 1.2 Our Contribution

In this paper, a robust SISR method is proposed by extending the concept of plug-and-play prior presented in Chan et al. [12].

- In the proposed method, two variable update steps in the ADMM (Eqs. (5) and (6)) are updated using off-the-shelf algorithms.
- The inversion step in the ADMM, i.e., Eq. (9) is implemented using an SISR algorithm with similar objective function and it is made robust to noise, by including a denoising algorithm.
- The convergence of the proposed algorithm is studied by using sample variance of the performance metric, which helps to improve the parameter selection.

Rest of this paper is organized as follows. A detailed description of the proposed method is given in Sect. 2. The experimental procedure and results are presented in Sect. 3 and the paper is concluded in Sect. 4.

## 2 Proposed Method

The proposed reconstruction-based SISR algorithm is formulated as an ADMM-based optimization problem. Using variable splitting, the SISR problem model given in Eq. (2) can be written as

$$\mathbf{x} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{v}), \text{s.t. } \mathbf{x} = \mathbf{v} \quad (3)$$

where  $R(\mathbf{v})$  is the regularization function and  $\lambda$  is the regularization parameters. Considering its augmented Lagrangian function, the above objective function can be written as an unconstrained optimization problem as given in Eq. (4).

$$\mathcal{L}_{\rho, \lambda}(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{v}) + \frac{\rho}{2} \|\mathbf{u} + \mathbf{x} - \mathbf{v}\|_2^2 \quad (4)$$

where  $\mathbf{u}$  is the vector of scaled Langrangian multipliers and  $\rho$  is the parameter of optimization.

The minimum of the function  $\mathcal{L}_{\rho, \lambda}(\mathbf{x}, \mathbf{v}, \mathbf{u})$  can be found using ADMM iteration. In ADMM, the problem is split to three sub-problems. Each sub-problem minimizes one variable by keeping the other two as constants. In every iteration, these variables are updated. The update equation for each variable is given by

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \frac{\rho^k}{2} r_x^2 \quad (5)$$

$$\mathbf{v}^{k+1} = \arg \min_{\mathbf{v}} \frac{\rho^k}{2} r_v^2 + \lambda R(\mathbf{v}) \quad (6)$$

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{x}^{k+1} - \mathbf{v}^{k+1} \quad (7)$$

$$\rho^{k+1} = \alpha \rho^k \quad (8)$$

where  $r_x^2 = \|\mathbf{x} - \mathbf{v}^k + \mathbf{u}^k\|_2^2$ ,  $r_v^2 = \|\mathbf{x}^{k+1} - \mathbf{v} + \mathbf{u}^k\|_2^2$ , and  $\mathbf{u}^k$  scaled Langrangian parameters in  $k$ th iteration. By making the substitution  $\tilde{\mathbf{x}} = \mathbf{v}^k - \mathbf{u}^k$  and  $\tilde{\mathbf{v}} = \mathbf{x}^{k+1} + \mathbf{u}^k$  in Eqs. (5) and (6), respectively, we get

$$\mathbf{x}^{k+1} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \frac{\rho^k}{2} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 \quad (9)$$

$$\mathbf{v}^{k+1} = \arg \min_{\mathbf{v}} \frac{\rho^k}{2} \|\mathbf{v} - \tilde{\mathbf{v}}\|_2^2 + \lambda R(\mathbf{v}) \quad (10)$$

Equation (10) has the form of a denoising problem with  $R(\mathbf{v})$  regularization, where  $\tilde{\mathbf{v}}$  is the noisy image with noise level  $\sqrt{\frac{\lambda}{\rho^k}}$ . Therefore, the SR method proposed by Chan et al. [12] uses a denoising algorithm for updating the  $\mathbf{v}$  in the ADMM algorithm.

A similar analysis of Eq. (9) reveals that it is an SISR algorithm with  $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$  term as the regularization function. This can be implemented using a closed-form solution FSR-ADMM as suggested by Zhao et al. [9]. The objective function in FSR-ADMM [9] has the form  $\mathbf{x} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{SHx} - \mathbf{y}\|_2^2 + \frac{\rho}{2} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$  which is very similar to the variable update step in Eq. (9).

Use of a robust super resolution algorithm at this step may improve the performance of the overall algorithm. A basic method used for super resolving noisy image is to perform denoising on the LR observation followed by the super resolution [2]. In this paper, this update process is implemented in two steps. Noise level in  $\tilde{\mathbf{x}}$  is determined using algorithm proposed by Liu et al. [13] for noise-level estimation. To avoid oversmoothing due to denoising algorithm, a manually selected constant  $0 < c < 1$  is employed to scale down the estimated noise level.  $\tilde{\mathbf{x}}$  is passed through a denoising scheme. The FSR-ADMM [9] is then applied on the result. This sequence ensures that the noise is removed in each iteration.

The convergence of plug-and-play algorithm is major challenge in the implementation. It is achieved by slowly increasing the value of  $\rho$  in each iteration, which forces the noise level to approach zero as  $k$  increases [12]. Rate of increase of  $\rho$  is controlled by the parameter  $\alpha$ . Analysis of the convergence of off-the-shelf algorithms poses a big challenge. Therefore, an experimental study is conducted for choosing proper value of  $\alpha$  which is one of the parameters that controls the convergence in this paper. If the proposed SISR algorithm converges to a fixed point, there exist  $\mathbf{x}^k$  such that  $\|\mathbf{x}^k - \mathbf{x}\|_2^2 \rightarrow 0$  as  $k \rightarrow \infty$ . Thus, executing the algorithm multiple times will give same results and thus the variance of the quality metric will be close to zero. Relaxation in convergence results in slightly different results in each execution of the algorithm. Hence, the variance of performance metric is a measure of convergence and it is used to quantify the convergence of the proposed algorithm. Algorithm 1 summarizes the procedural flow of the proposed method for a magnification factor  $s$ .  $\mathcal{D}_\sigma(\mathbf{x})$  in the algorithm denotes the denoising operation on  $\mathbf{x}$  with noise level  $\sigma$ .

---

**Algorithm 1** Proposed plug-and-play ADMM-based robust super resolution algorithm

---

```

1: procedure ROBUSTSR( $\mathbf{y}, s$ )
2:   Initialize  $\mathbf{v}^0 = \mathbf{y} \uparrow s$ ,  $\mathbf{x}^0 = \mathbf{v}^0$ ,  $\mathbf{u}^0 = \mathbf{0}$ ,  $\rho^0, \alpha, \lambda$ 
3:   for  $k = 1 \dots K$  do
4:      $\tilde{\mathbf{x}} = \mathbf{v}^{k-1} - \mathbf{u}^{k-1}$ 
5:      $\sigma$  = noise level in  $\tilde{\mathbf{x}}$ 
6:     Denoising;  $\tilde{\mathbf{x}} = \mathcal{D}_{c\sigma}(\tilde{\mathbf{x}})$ 
7:      $\mathbf{x}^k = \text{FSR-ADMM}(\mathbf{y}, \tilde{\mathbf{x}}, \frac{\rho^{k-1}}{2})$  [9]
8:      $\sigma = \sqrt{\frac{\lambda}{\rho^{k-1}}}$ 
9:      $\tilde{\mathbf{v}} = \mathbf{x}^k + \mathbf{u}^{k-1}$ 
10:    Denoising;  $\mathbf{v}^k = \mathcal{D}_\sigma(\tilde{\mathbf{v}})$ 
11:     $\mathbf{u}^k = \mathbf{u}^{k-1} + \mathbf{x}^k - \mathbf{v}^k$ 
12:     $\rho^k = \alpha \rho^{k-1}$ 
13:   Return  $\mathbf{x} = \mathbf{x}^k$                                  $\triangleright$  Super resolved image

```

---

### 3 Experimental Results

Experiments are conducted to evaluate the performance of proposed algorithm with the closely related techniques. The experiments are conducted using MATLAB R2014a on a Windows 7 professional operating system in a laptop with 4GB RAM and 1TB HDD. The proposed method is evaluated in terms of peak signal-to-noise ratio (PSNR) and sharpness index (SI) [14]. A high value of these metrics indicate a better result. Proposed method is compared with plug-and-play denoiser prior (PPD) based SISR algorithm proposed by Chan et al. [12] and FSR-ADMM by Zhao et al. [9] for which the MATLAB implementation is available in public domain. These algorithms are recently reported reconstruction-based SISR methods. The result of proposed method is also compared with the deep learning based SR algorithm, SRCNN [7]. Except in the case of zero noise level, SRCNN [7] is implemented after applying BM3D denoising on the noisy LR image and it is denoted as Dn + SRCNN in Table 1. Popular test images used in the SR experiments are employed for reporting the performance of proposed algorithm. All the images are converted to gray scale with pixel values ranging from 0 to 255. LR observations are prepared by smoothing and downsampling the HR examples. Noisy images are prepared by adding white Gaussian noise with the specified standard deviation to LR image. Few images used for the experiments are shown in Fig. 1. We used the following parameter values in our implementation. Magnification factor  $s = 2$ , HR image size is  $256 \times 256$ , and  $\lambda = 0.15$ . BM3D [15] denoiser was used in all the experiments. The proposed algorithm is applied on the test images 10 times and the average of the PSNR and SI values are reported in Tables 1 and 2. Table 1 gives a comparison in terms of PSNR of the existing algorithms with the proposed one and it shows improvements in the result especially when noise level in the LR observation is above  $\sigma = 5$ . Figure 2 gives a visual comparison of the results of proposed method with FSR-ADMM by Zhao et al. [9] and the PPD-based SR algorithm by Chan et al. [12]. At low noise levels, Dn + SRCNN is giving better metrics while the other SR methods also have comparable performance, especially in highly textured images like doll and flower. The quantitative comparison in Table 2 shows that proposed method has high sharpness index (SI) [14] at all noise levels.

To study the effect the choice of parameter  $\alpha$  on the performance of the algorithm, the experiments are repeated ten times on each images and the average value and sample variance of PSNR are calculated. PSNR is selected as a metric for this analysis because it shows systematic variation with the changes in image. Sharpness index, on the other hand, shows significant fluctuation in the values for a change in the image. A low value of sample variance indicates the convergence of the result of proposed method to a target image. By choosing  $\alpha = 1$  in the algorithm, value of  $\rho$  remains same in all iterations which affects the convergence of the algorithm. This characteristic is indicated by a high value in the variance of PSNR. When  $\alpha > 1$ , the value of  $\rho$  increases in every iteration which forces the algorithm to converge as  $\sigma = \sqrt{\frac{\lambda}{\rho^k}} \rightarrow 0$ . Figure 4 shows the change in the PSNR of the proposed algorithm for different values of  $\alpha$ .

**Table 1** The performance comparison of the SISR methods for different noise levels in terms of PSNR values in dB. Noise level  $\sigma$  is the standard deviation of noise present in LR image

Method	$\sigma = 0$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	$\sigma = 25$
<i>Butterfly</i>						
Proposed	<b>24.044</b>	<b>23.976</b>	<b>23.714</b>	<b>23.095</b>	<b>22.405</b>	<b>21.126</b>
PPD	23.563	23.517	23.390	22.972	21.621	20.066
FSR-ADMM	23.675	23.356	22.556	21.471	20.276	19.134
Dn + SRCNN	23.661	23.066	21.886	21.131	20.248	19.322
Bicubic	22.567	22.360	21.821	21.055	20.174	19.255
<i>Flower</i>						
Proposed	24.357	24.337	<b>24.267</b>	<b>24.040</b>	<b>23.377</b>	<b>22.060</b>
PPD	24.663	24.345	23.530	22.447	21.279	20.131
FSR-ADMM	24.580	24.198	23.518	22.558	21.066	19.748
Dn + SRCNN	<b>25.037</b>	<b>24.536</b>	23.536	22.443	21.266	20.109
Bicubic	23.753	23.365	22.922	21.827	20.520	19.814
<i>Lena</i>						
Proposed	26.182	<b>26.139</b>	<b>26.053</b>	<b>25.784</b>	<b>24.884</b>	<b>23.003</b>
PPD	26.112	25.897	25.292	24.563	23.765	22.935
FSR-ADMM	26.082	25.845	24.915	23.213	21.550	20.128
Dn + SRCNN	<b>26.485</b>	25.371	24.577	23.231	21.854	20.552
Bicubic	26.064	25.647	24.593	23.261	21.895	20.601
<i>Doll</i>						
Proposed	23.210	23.216	<b>23.197</b>	<b>23.051</b>	<b>22.526</b>	<b>21.405</b>
PPD	23.539	23.318	22.686	21.794	20.786	19.759
FSR-ADMM	23.705	23.230	23.139	21.926	20.643	19.375
Dn + SRCNN	<b>24.344</b>	<b>23.596</b>	22.629	21.716	20.696	19.661
Bicubic	22.808	22.519	22.269	21.120	20.321	19.045
<i>Zebra</i>						
Proposed	22.232	21.720	<b>21.694</b>	<b>21.599</b>	<b>21.245</b>	<b>20.431</b>
PPD	22.293	<b>21.746</b>	20.739	20.388	20.004	19.595
FSR-ADMM	<b>22.349</b>	21.720	21.511	20.592	19.635	18.656
Dn + SRCNN	21.616	21.203	20.443	19.879	19.194	18.445
Bicubic	20.942	20.812	20.440	19.881	19.202	18.460

(continued)

**Table 1** (continued)

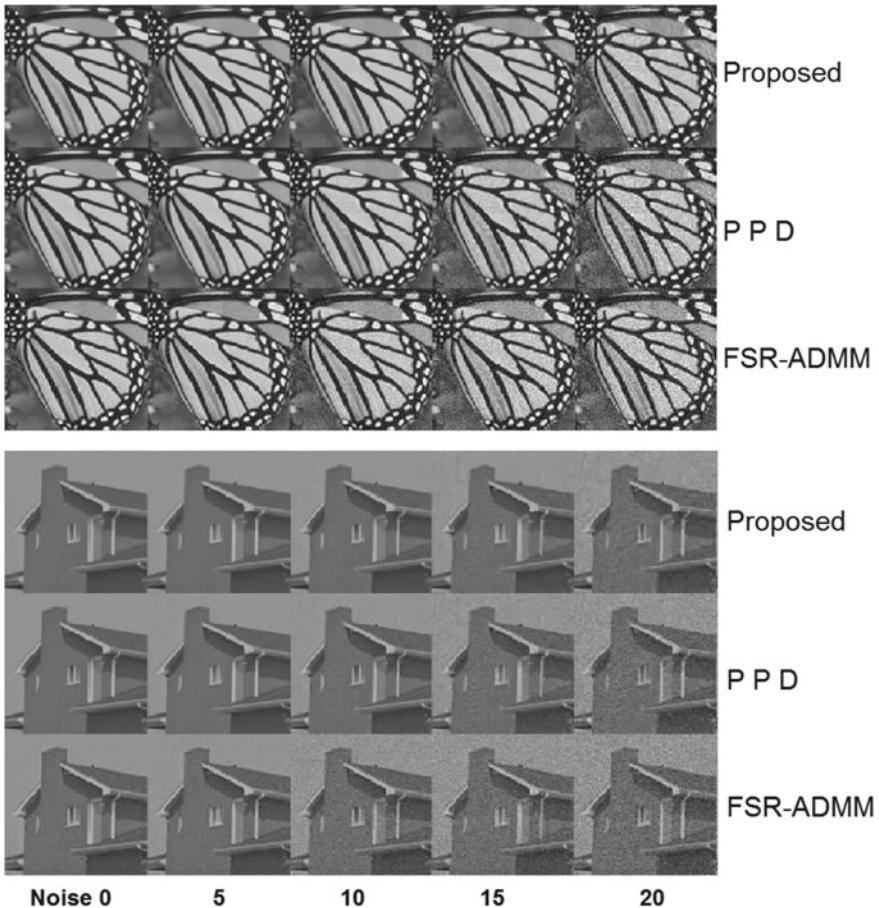
<i>House</i>						
Proposed	<b>30.288</b>	<b>30.263</b>	<b>29.661</b>	<b>28.898</b>	<b>27.223</b>	<b>24.393</b>
PPD	29.842	29.775	29.547	27.665	24.593	21.878
FSR-ADMM	30.047	28.818	26.509	24.190	22.235	20.596
Dn + SRCNN	29.502	27.492	26.334	24.433	22.685	21.146
Bicubic	28.888	28.090	26.361	24.475	22.737	21.204
<i>Average</i>						
Proposed	25.052	<b>24.942</b>	<b>24.765</b>	<b>24.411</b>	<b>23.610</b>	<b>22.070</b>
PPD	25.002	24.766	24.197	23.305	22.008	20.727
FSR-ADMM	25.073	24.528	23.691	22.325	20.901	19.606
Dn + SRCNN	<b>25.174</b>	24.210	23.234	22.139	20.990	19.872
Bicubic	24.170	23.799	23.068	21.937	20.808	19.730

**Fig. 1** Some of the images used in the experiments

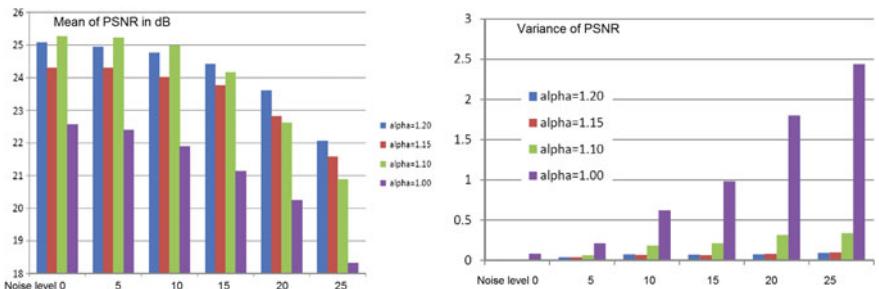
The experiment is conducted with a magnification factor  $s = 3$  and the average values of PSNR are given in Table 3. At this magnification level also, proposed method is outperforming other techniques, if the noise level in the input is high. Figure 3 gives the visual comparison of the results of different methods for a magnification factor  $s = 3$ .

## 4 Conclusion

In this paper, the problem of SISR is addressed by employing the concept of plug-and-play prior in ADMM optimization. The two variable update steps in the ADMM are modified with the help of two separate algorithms. By introducing a robust SISR step in the variable update step, it was possible to achieve improvement in the noise performance of the final result. Parameter selection and the convergence are the major



**Fig. 2** Comparison of results obtained with proposed method, PPD, FSR-ADMM methods. The noise values shown are corresponding to the standard deviation of the noise present in the input LR observation of respective columns



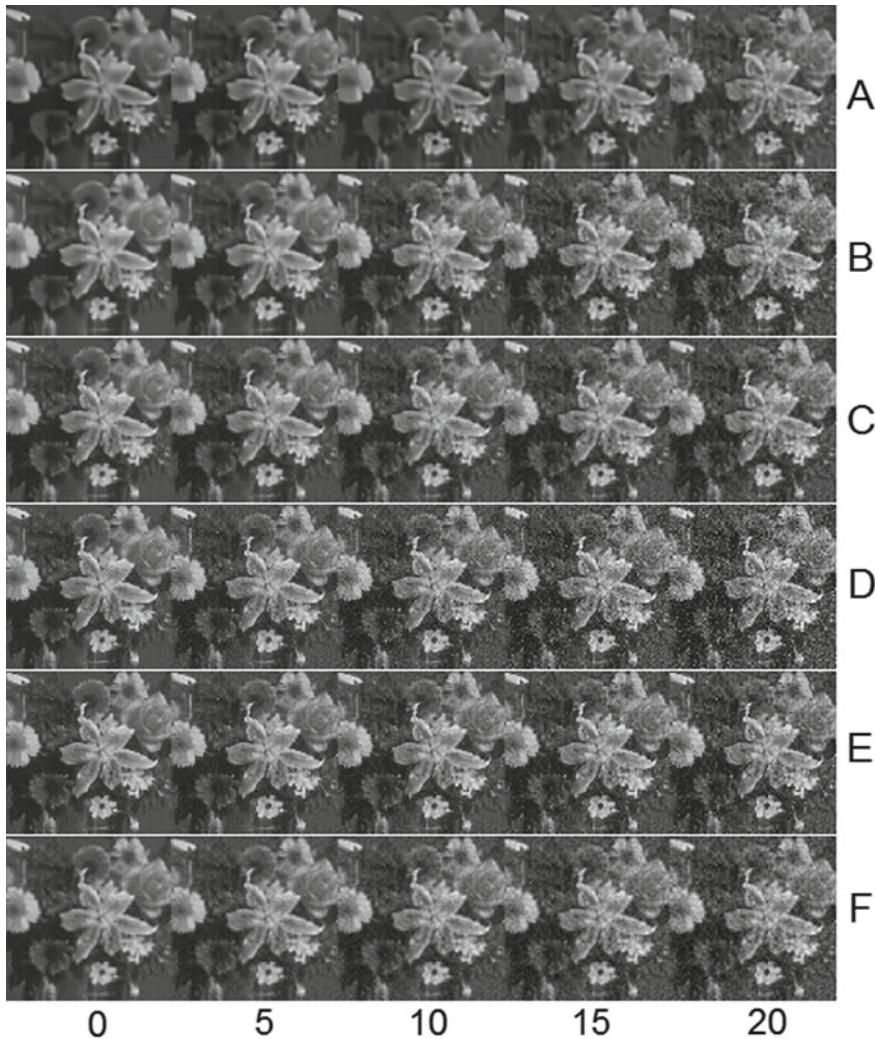
**Fig. 3** Results obtained with the flower image for a scaling factor of 3. Rows A-Proposed method, B-PPD, C-FSRADMM, D-SRCNN without denoising step, E-Dn + SRCNN and (5)-Bicubic interpolated. Columns correspond to different noise levels ( $\sigma$ ) in the input image

**Table 2** The performance comparison of the SISR methods for different noise levels in terms of the sharpness index (SI) [14]. Noise level  $\sigma$  is the standard deviation of noise present in LR image

Method	$\sigma = 0$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
<i>Butterfly</i>					
Proposed	<b>76.9</b>	<b>74.8</b>	<b>69.0</b>	<b>60.7</b>	<b>45.3</b>
PPD	72.7	72.1	64.9	44.6	24.6
FSR-ADMM	71.4	68.0	48.2	33.7	21.7
<i>Flower</i>					
Proposed	<b>97.2</b>	<b>94.1</b>	<b>87.7</b>	<b>74.8</b>	<b>54.0</b>
PPD	96.1	90.8	75.3	48.8	25.9
FSR-ADMM	81.0	67.1	46.4	31.8	17.7
<i>Lena</i>					
Proposed	<b>115.1</b>	<b>117.9</b>	<b>118.0</b>	<b>101.7</b>	<b>67.8</b>
PPD	110.4	113.5	99.8	56.8	24.5
FSR-ADMM	88.3	106.0	83.9	60.8	31.9
<i>Doll</i>					
Proposed	<b>66.1</b>	<b>65.7</b>	<b>63.0</b>	<b>55.4</b>	<b>42.4</b>
PPD	58.9	57.6	50.8	36.8	22.5
FSR-ADMM	53.8	41.1	25.3	18.5	16.6
<i>Zebra</i>					
Proposed	<b>114.9</b>	<b>113.7</b>	<b>109.3</b>	<b>97.2</b>	<b>74.1</b>
PPD	106.1	103.3	91.4	67.4	42.8
FSR-ADMM	104.3	94.1	58.7	38.0	25.7
<i>House</i>					
Proposed	<b>219.2</b>	<b>217.9</b>	<b>204.8</b>	<b>172.0</b>	<b>116.3</b>
PPD	213.9	205.5	167.7	96.0	40.0
FSR-ADMM	147.5	140.1	121.7	84.8	29.3
<i>Average</i>					
Proposed	<b>114.2</b>	<b>113.6</b>	<b>108.6</b>	<b>93.6</b>	<b>66.6</b>
PPD	110.4	107.6	91.6	58.4	30.0
FSR-ADMM	91.3	86.8	66.8	46.4	24.3

**Table 3** Comparison of average PSNR values in dB of the SISR methods for a magnification factor  $s = 3$ . Noise level  $\sigma$  is the standard deviation of noise present in LR image

Method	$\sigma = 0$	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$
Proposed	22.460	22.230	<b>21.750</b>	<b>20.801</b>	<b>19.903</b>
PPD	<b>22.733</b>	<b>22.407</b>	21.635	20.571	19.452
FSR-ADMM	21.483	20.953	20.224	19.399	18.546
Dn + SRCNN	21.018	20.470	19.727	18.889	18.025
Bicubic	19.396	19.934	19.289	18.369	17.325



**Fig. 4** Effect of the choice of  $\alpha$  on the performance of the proposed method. The chart on left shows the variation of PSNR for various values of  $\alpha$ . The chart in the right shows the variance of PSNR for various values of  $\alpha$

challenges in plug-and-play algorithms. Study on the effect of  $\alpha$  in the algorithm shows that by choosing a slowly increasing  $\rho$ , it is ensured that denoising algorithm does not cause oversmoothing and forces convergence.

## References

1. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. *IEEE Signal Process. Mag.* **20**(3), 21–36 (2003)
2. Yue, L., Shen, H., Li, J., Yuan, Q., HongyanZhang, Zhang, L.: Image super-resolution: the techniques, applications and future. *Signal Process.* **14**, 389–408 (2016)
3. Nasrollahi, K., Moeslund, T.B.: Super-resolution: a comprehensive survey. *Mach. Vis. Appl.* **25**(6), 1423–1468 (2014)
4. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **19**(11), 2861–2873 (2010)
5. Rahiman, V.A., George, S.N.: Single image super resolution using neighbor embedding and statistical prediction model. *Comput. Electr. Eng.* **62**, 281–292 (2017)
6. Rahiman, V.A., George, S.N.: Single image super resolution using neighbor embedding and fusion in wavelet domain. *Comput. Electr. Eng.* 1–16 (2017)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: *Proceedings of European Conference on Computer Vision (ECCV)* (2014)
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), 1–122 (2011)
9. Zhao, N., Wei, Q., Basarab, A., Kouame, D., Tourneret, J.Y.: Fast single image super-resolution using a new analytical solution for  $\ell_2 - \ell_2$  problems. *IEEE Trans. Image Process.* **25**(8), 3683–3697 (2016)
10. Venkatakrishnan, S., Bouman, C., Wohlberg, B.: Plug and play priors for model based reconstruction. In: *Proceedings of IEEE Global Conference on Signal and Information Processing*, pp. 945–948 (2013)
11. Brifman, A., Romano, Y., Elad, M.: Turning a denoiser into a super resolver using plug and play priors. In: *Proceedings in Proceedings of the IEEE International Conference on Image Processing*, pp. 1404–1408 (2016)
12. Chan, S.H., Wang, X., Elgendi, O.A.: Plug-and-play admm for image restoration: fixed point convergence and applications. *IEEE Trans. Comput. Imaging* **3**(1), 84–98 (2017)
13. Liu, X., Tanaka, M., Okutomi, M.: Single-image noise level estimation for blind denoising. *IEEE Trans. Image Process.* **22**(12), 5226–5237 (2013)
14. Blanchet, G., Moisan, L.: An explicit sharpness index related to global phase coherence. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 1065–1068 (2012)
15. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. on Image Process.* **16**(8) (2007)

# Indoor–Outdoor Scene Classification with Residual Convolutional Neural Network



Seema Kumari, Ranjeet Ranjan Jha, Arnav Bhavsar and Aditya Nigam

**Abstract** In this paper, we demonstrate the effectiveness of a customized ResNet to address the problem of indoor–outdoor scene classification both for color images as well as depth images. Such an approach can serve as an initial step in a scene classification/retrieval pipeline or a single-image depth estimation task. The classification framework is developed based on Residual Convolutional Neural Network (ResNet-18) to classify any random scene as indoor or outdoor. We also demonstrate the invariance of the classification performance with respect to different weather conditions of outdoor scenes (which one can commonly encounter). The performance of our classification strategy is analyzed on different varieties of publicly available datasets of indoor and outdoor scenes that also have corresponding depth maps. The suggested approach achieves almost an ideal performance in many scenarios, for both color and depth images, across datasets. We also show positive comparisons with other state-of-the-art methods.

**Keywords** Scene classification · RCNN · Scene analysis · Depth information · Image retrieval

## 1 Introduction

Scene analysis plays an important role in different applications such as image retrieval, personal photo tagging, color constancy, and robotics [1]. Retrieving a specific image from an accumulation of ever-increasing visual data often needs bet-

---

S. Kumari · R. R. Jha · A. Bhavsar (✉) · A. Nigam

MAS Lab, School of Computing and Electrical Engineering, IIT Mandi, India

e-mail: [arnav@iitmandi.ac.in](mailto:arnav@iitmandi.ac.in)

S. Kumari

e-mail: [seema\\_kumari@students.iitmandi.ac.in](mailto:seema_kumari@students.iitmandi.ac.in)

R. R. Jha

e-mail: [d16044@students.iitmandi.ac.in](mailto:d16044@students.iitmandi.ac.in)

A. Nigam

e-mail: [aditya@iitmandi.ac.in](mailto:aditya@iitmandi.ac.in)

ter organization [2], which can be facilitated by scene analysis [2]. However, with the increasing trend of data complexity being considered, the inter-class similarity and intra-class variability may pose an issue in scene analysis. In addition to developing more sophisticated approaches, which directly operate on all classes considered together, hierarchical approaches can be considered that first address a coarse classification problem (e.g., indoor–outdoor scenes) followed by a fine-grained classification problems. Further, with the easy availability of range cameras, the scene analysis can now also consider depth information.

Indeed, in general, scene analysis can be linked with 3D vision related applications in different ways (e.g., coupling of object detection and segmentation with 3D reconstruction), where depth information plays a pivotal role. In applications which do not involve depth cameras, estimating depth from intensity images, traditionally, involves multiple views [3]. However, this entails the baggage of establishing correspondences, handling occlusions, etc. Thus, in recent years, the task of learning-based depth estimation has been considered which enables estimation depth from single intensity image [4–6]. However, such learning-based methods often focus on specific type of scene (e.g., indoor–outdoor, specific outdoor content, indoor scenes with similar structures, etc.), as the characteristics of multiple types of scenes is difficult to learn. This domain specific nature often restricts the learning-based algorithm in estimating depth for any arbitrary data. To address a general case, one approach can be to use separate models learned from indoor and outdoor scenes along with a scene analysis preclassifier that can classify the scene type. This is arguably a simple and useful alternative to developing sophisticated learning algorithms.

Thus, considering the usefulness of such a coarse level scene classification across (at least) the above important application domains, in this paper, we suggest an approach to address the task of indoor or outdoor scene classes. In addition to the image based scene analysis, which is a popular research area [1], we have explored such a task for depth maps, which do not have rich contents as intensity images. Interestingly, even with low visual content, we demonstrate that the depth maps alone can produce good classification performance for this task; a useful implication for a scene analysis pipeline. Also, since the conditions for outdoor scenes particularly vary significantly, we also consider such variations (e.g., sunset, fog, morning, overcast, evening).

Our approach involves a customized residual convolutional neural network framework ResNet-18, which is much smaller than the standard ResNet-152 network for both intensity and depth image classification. We demonstrate that such a simpler network architecture can yield almost an ideal classification performance across various scenarios. Moreover, we use a relatively small training dataset (in hundreds) to fine-tune a pretrained ResNet. We also compare the proposed method, with some state-of-the-art methods, and show a visibly superior performance.

Hence, the contributions of this paper can be summarized as: (1) An approach to perform indoor versus outdoor scene classification for images as well as for depth maps, with a relatively simple ResNet architecture, and a small training dataset. (2) Considering various atmospheric conditions for the outdoor scenes, thus showing an effective invariance of the approach across conditions. (3) Validating the

classification approach across various datasets, and demonstrating a consistent high performance, and positive comparisons with state of the art.

The rest of the paper is organized as follows: Sect. 2 reviews some of the related works of scene classification. In Sect. 3, the proposed method for scene analysis is described. Section 4 shows the experimental results with qualitative and quantitative comparisons. Section 5 summarizes the paper.

## 2 Related Work

The traditional methods for scene classification have used different hand-crafted features to capture such aspects, e.g., color histogram, texture, edges, and shape properties [7]. These types of features used for region-based image annotation and retrieval. Different types of classifiers such as K-nearest neighbor, Bayesian, SVM, ICA, PCA, ANN are also used for scene classification [8, 9]. There has been some earlier work on indoor–outdoor scene classification. In [10], a technique based on straightness of edges is reported, where the hypothesis is that the proportion of straight edges in indoor images is greater as compared to outdoor images. However, this is generally a weak assumption. A probabilistic neural network technique is reported in [11] for classification of indoor/outdoor images. In this approach, an unsupervised method based on fuzzy c-means clustering (FCM) [12] for segmentation is proposed. Color, texture, and shape information are used as features that are extracted from each segmented image, which are then given to a PNN (Probabilistic Neural Network) for classification [11]. Furthermore, a neural network learning based technique is also proposed in indoor–outdoor scene classification [13], which works with features based on image color, entropy, DCT coefficients, and edge orientation. Depth information has also been used as additional information to discriminate between indoor and outdoor scenes in [1]. The approach is further improved by combining the depth ( $3D_B$ ) feature with the feature of an existing method [14].

More recently, convolutional neural network (CNN) became very popular in different fields of computer vision and image processing. The advancement of deep convolutional neural networks allows us to replace traditional features and related methods by CNN-based features with linear kernel based SVM classifier to achieve better performance as compared to the traditional methods [15]. Further, effectiveness and robustness of learned CNN features has also helped in achieving better results in the fields of image classification [16, 17], object detection [18], image segmentation [19], image retrieval [20], etc.

The general scene classification task has also been addressed with CNN-based approaches [21–26]. For instance, CNN-based features have been encoded into a new bag-of-semantics (BoS) representation [21]. However, extracting feature from pretrained CNN often requires fixed size input image. Hence, one may have to resize an image, which may lead to loss of information or addition of artifacts [22]. However, in some CNN-based approaches [23, 24, 27], this fixed size limitation has been addressed. In these approaches [23, 27, 28], images of arbitrary size can be

given as input. The convolutional layers can operate on differently sized images, and fixed sized feature representations is computed for classification. Furthermore, state-of-the-art convolutional neural networks have been used to report a baseline scene classification CNNs, which is used to perform place recognition [25]. The intermediate representation in CNNs can also behave as a useful feature representation. This intermediate representation helps in recognizing places [29]. Further, the behavior of the learned feature in a CNN for scene classification is investigated as generic features in other tasks of visual recognition [22].

Similar to some works mentioned earlier, the residual neural network [26] can handle different dimensional images; an aspect which we too appreciate in our work, where we use a ResNet.

### 3 Proposed Approach

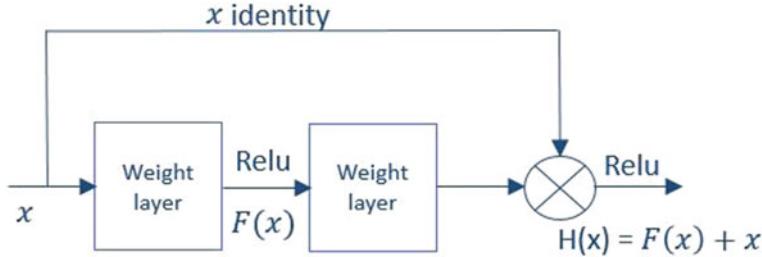
In the following subsections, we discuss various aspects of the proposed framework, which involves modified residual convolutional neural network (ResNet-18).

#### 3.1 *Residual Learning*

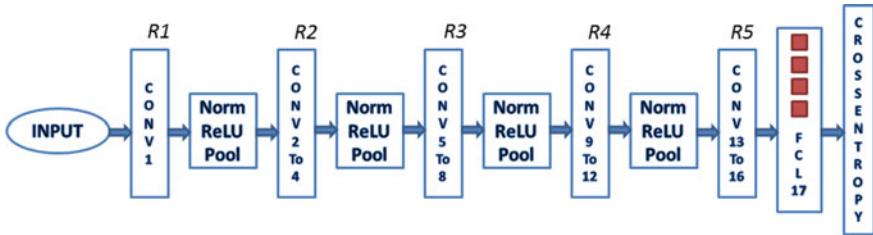
In this paper, we have used residual learning framework instead of traditional CNN models such as AlexNet [22], VGGNet-16 [30], GoogleNet [31], because of its improved performance over such models. The block diagram of residual neural network is shown in Fig. 1. In Fig. 1, the residual mapping is considered which can be expressed as  $H(x) = F(x) + x$ , where  $x$ ,  $H(x)$ , and  $F(x)$  represent the identity function (input = output), few stacked nonlinear layers and the residual function, respectively. The formulation of a residual network can be realized by feed-forward neural networks with some skip connections. This network introduces identity short-cut connection (or a skip connection) (Residual learning framework) [26], which is used to skip one or more layers. In our network, the skip connections are simply performed by identity mapping. The skip connections in the ResNet help in resolving the vanishing gradient problem, which is common to deep learning, and due to which the performance of standard CNN models can degrade. Moreover, it has been shown that it is easier to optimize the network with a residual mapping than the direct mapping, and hence it is more likely to achieve a more optimal convergent solution.

#### 3.2 *Network Architecture*

In this approach, we have used modified residual convolution neural network (ResNet) to classify the indoor-outdoor images. The architecture of ResNet-18 as



**Fig. 1** Block diagram of basic residual learning



**Fig. 2** Block diagram of residual convolutional neural network for scene classification

shown in Fig. 2 takes image or depth map as input and provide class label as output. The originally proposed ResNet is a 152-layer architecture. However, considering the problem of two-class classification, we employ a simpler architecture of 18 layers which consist of 16 convolution layers and 2 fully connected layers along with Max-Pool, Batch normalization, Rectified linear unit activation as represented by Pool, Norm, and ReLU in the block diagram. In the block diagram of ResNet,  $R_1$ ,  $R_2$ ,  $R_3$ ,  $R_4$  and  $R_5$  are residual blocks. In the ResNet-18 model, each residual block contains two convolution layers. Block 1 and 2 have 64 filters, 3rd, 4th, 5th, 6th blocks consist of 128 filters, whereas block 7 and 8 have 256 filters. Each filter is of size  $3 \times 3$ . Further, the final fully connected layer is used as the classification layer.

### 3.3 Implementation Detail

In this paper, we have not used the standard ResNet-18 network, but an adaptation to the existing model of ResNet-18. Initially, we have used the pretrained weights of ImageNet, which are learned using 1.2 million images with 1000 categories [32]. These pretrained weights are expected to be tuned to the new data for classification in better way than a random initialization. Note that, we have used 18 layer. The weights of the final fully connected layer are learned using our datasets of indoor and outdoor images. Since, we have not used any scaling factor to scale the images, these images are directly feeding to network. In the first stage, 64 filters (Conv 1 layer) are applied

to the input images. Further, this is processed by ReLU (Rectified Linear Unit) and a max pool layer. In the next step, we have grouped 4 consecutive convolution layers with 128 filter maps and this process is repeated till the last convolutional layer. Identity mapping connections are created between alternate convolutional layers. Further, the resultant data is delivered to fully connected layer and it gives the final output vector after mapping the data with particular class thus obtaining the final prediction probability by using cross-entropy loss.

The cross-entropy is used as a loss function in this method, it gives the output of the classifier as probability value between 0 and 1. Further, cross-entropy can be calculated by using Eq. 1.

$$L(y, p) = - \sum_{i=1}^N y_i \log(p_i). \quad (1)$$

Here,  $N$  is the number of classes,  $\log$  is the natural log,  $y$  is representing the binary (0 or 1) number. Classification of true class ( $i$ ) will lead to  $y = 0$ . The probability of  $y = 0$  is represented by  $p$ .

## 4 Experimentation

We have used four (two indoor and outdoor) publicly available datasets to train and validate our model such as KITTI [33], Make3D [4], NYU depth [34] and Matterport [35]. These datasets contain intensity images with corresponding depth maps. For experimentation, we have used both intensity images as well as depth maps. The brief description of dataset and experimental results are given below.

### 4.1 Outdoor Dataset

Virtual KITTI [33] dataset is contained 50 high-resolution monocular videos (21260 frames), it is generated from five different virtual worlds in urban settings under different imaging and weather conditions. Virtual worlds are created by a game engine through a real-to-virtual cloning method. This dataset is used in several video understanding tasks: depth estimation, scene-level and instance-level semantic segmentation, object detection, optical flow, and multi-object tracking [33]. The data has 10 different variations in form of whether variation or imaging such as: clone, 15-deg-right, 15-deg-left, 30-deg-right, 30-deg-left, morning, sunset, overcast, fog, rain. In this paper, we have used 21,260 images with corresponding depth maps for scene analysis. Here, the resolution of images is  $1242 \times 375$ .

The Make3D [4] dataset consists of intensity images with small resolution ground truth depth map. In our method, we have taken different numbers of images to train

and test our model from both (Kitti and Make3D) the datasets as shown in first column of Tables 1 and 2.

## 4.2 Indoor Dataset

The NYUV1 depth [34] dataset is composed of 64 indoor scenes, which are recorded as video sequences by using both intensity and depth cameras of the Microsoft Kinect. In this paper, we have used labeled dataset that consists of 2229 images with corresponding depth map. The resolution of images is  $640 \times 480$ .

Matterport [35] dataset contains 194,400 intensity and corresponding depth maps of 90 building-scale scenes. All of the images were captured by using matterport’s Pro 3D camera. Here, we have taken only single building scene images with the corresponding depth map to our model. In our method, we have used different numbers of images to train and test our model from the both (Matterport and NYUV1) datasets as shown in first column of Tables 1 and 2.

## 4.3 Experimental Results

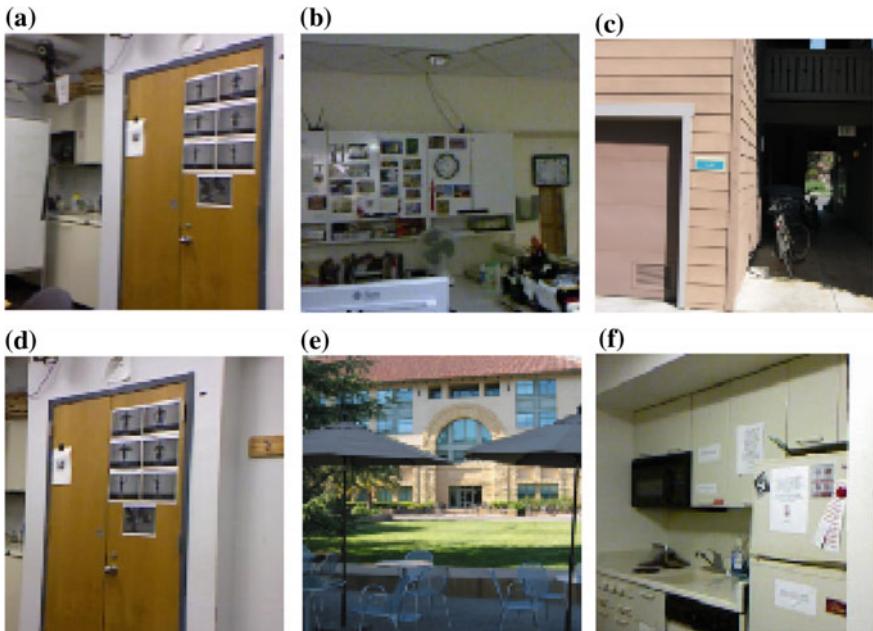
For the implementation purpose, we train our model by considering 600 images. 100 images are considered to test the performance of the proposed approach. The quantitative results are presented in Tables 1 and 2, whereas qualitative results are shown in Figs. 3 and 4. Table 1 shows the results for indoor versus outdoor scenes along with different weather conditions for intensity images. Whereas the scene classification using depth maps are quantified in Table 2. The network is trained using NVIDIA Geforce GTX 1080 GPU and it takes about 4 min to train the model.

**Table 1** Quantitative results of indoor/outdoor scene classification for intensity images by using RCNN

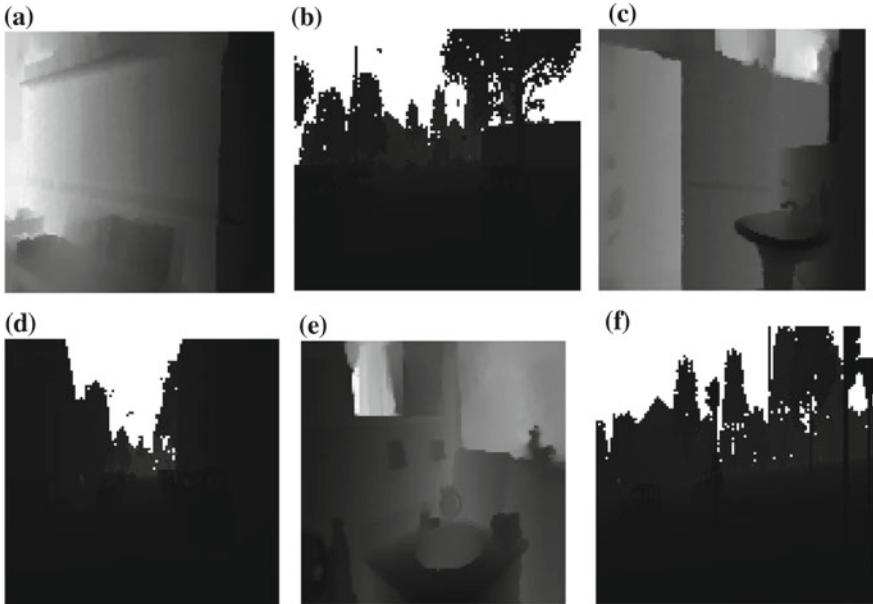
Images (train/test)	Datasets	Testing	Accuracy
600/100	NYUV1 and KITTI	Indoor/clone	1
600/100	NYUV1 and KITTI	Indoor/overcast	1
600/100	NYUV1 and KITTI	Indoor/morning	1
600/100	NYUV1 and KITTI	Indoor/sunset	1
600/100	NYUV1 and KITTI	Indoor/fog	1
600/100	NYUV1 and KITTI	Indoor/rain	0.93
740/60	NYUV1 and Make3D	Indoor/outdoor	0.92
600/100	NYUV1 and Make3D	Indoor/outdoor	0.97
600/100	Matterport and KITTI	Indoor/outdoor	1
660/400	NYUV1 and KITTI	Indoor/all outdoor weather condition	0.99

**Table 2** Quantitative results of indoor/outdoor scene classification for depth maps by using RCNN

Images (train/test)	Datasets	Testing	Accuracy
600/100	NYUV1 and KITTI	Indoor/clone	1
600/100	NYUV1 and KITTI	Indoor/overcast	1
600/100	NYUV1 and KITTI	Indoor/morning	1
600/100	NYUV1 and KITTI	Indoor/sunset	1
600/100	NYUV1 and KITTI	Indoor/fog	1
600/100	NYUV1 and KITTI	Indoor/rain	1
740/60	NYUV1 and Make3D	Indoor/outdoor	0.92
600/100	NYUV1 and Make3D	Indoor/outdoor	0.97
600/100	Matterport and KITTI	Indoor/outdoor	1
660/400	NYUV1 and KITTI	Indoor/all outdoor different weather condition	1

**Fig. 3** Qualitative results of classification for intensity images: **a, b, d, f** indoor scene of NYUV1 dataset; **c, e** outdoor scene of Make3D dataset

One can observe that our model is able to discriminate the indoor and outdoor scenes quite accurately for intensity images in Table 1. Here, one can note that our model is behaving as an ideal classifier in the task of classification between virtual outdoor (KITTI dataset) and real indoor (NYU dataset) images. The reason behind this



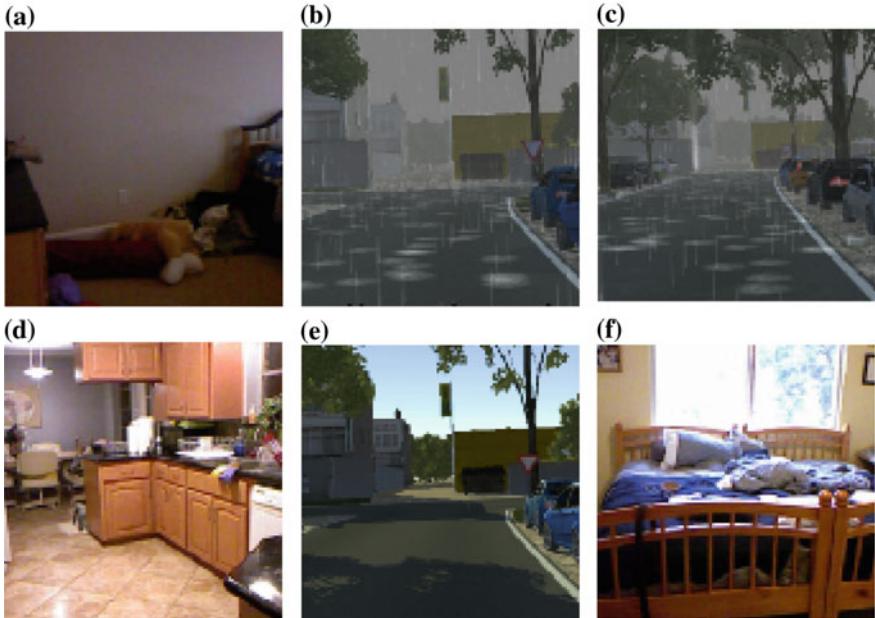
**Fig. 4** Qualitative results of classification for depth images: **a, c, e** indoor scene of NYUV1 dataset; **b, d, f** outdoor scene of KITTI dataset

**Table 3** Quantitative results for comparison of our approach with previously proposed approaches

Method	Indoor	Outdoor	Average
Proposed	0.94	0.98	0.96
Baseline CNN	0.54	0.81	0.68
Centrist [14] + $3D_B$ [1]	0.94	0.91	0.93
NN [13]	0.95	0.92	0.94
PNN [11]	0.94	0.91	0.93
Edge straightness (rule based) [10]	0.71	0.73	0.72
Edge straightness (k-NN) [10]	0.66	0.67	0.67

is that the texture of virtual (KITTI) dataset is different as compared to real (Make3D) dataset. In absence of intensity images, we have used only depth maps to train our model. Interestingly, the performance is quite similar to the intensity images. Though depth maps are not rich in color, texture, edge information, the proposed framework still is able to discriminate between different scenes. This demonstrates the effectiveness of our approach in scene analysis using depth maps, which involve only a high-level appearance of the scene content.

Furthermore, we have conducted more studies on Virtual Kitti dataset, which consists of different weather conditions such as fog, rain, overcast, sunset, morning, and clone. These weather conditions are used for the scene classification as shown



**Fig. 5** Qualitative results of classification for intensity images: **a, d, f** indoor scene of NYUV1 dataset; **b, c, e** outdoor scene with weather variation of KITTI dataset

in Tables 1 and 2. Tables 1 and 2 considers only two-class classification, i.e., indoor versus a particular weather condition, whereas last row of Table 2 shows the results for indoor versus all outdoor weather condition classification. It can be seen that the approach is clearly invariant to the appearance changes due to weather conditions.

The accuracy of the quantitative results are consistent in the qualitative results as can be observed in Figs. 3, 4 and 5. The high accuracy in classification indicates that such an approach can be used as an initial step in the pipelines of depth estimation as well as image retrieval.

#### 4.4 Comparison

In this paper, we have also compared our results with existing approaches [1, 10, 11, 13, 14] and a baseline CNN model as shown in Table 3. To compare our results, we have used IITM-SCID2 dataset, which is used in [1, 10, 11], and it consists of 907 indoor and outdoor images. For experimentation, we have subdivided this dataset (IITM-SCID2) into 193 indoor and 200 outdoor images for training the model and 249 indoor and 260 outdoor images for testing. The baseline CNN model is used for comparison based on 13 convolution, 5 max-pooling, and 5 fully connected layers along with rectified linear unit activations.

One can observe from Table 3 that on an average, our framework is able to produce better classification accuracy as compared to the other approaches involving conventional approaches as well as CNN- based approach. The improved performance of our result as compared to the conventional approaches [10] is due to the deep learning framework that is able to learn an appropriate model. Further, the baseline CNN model typically needs large number of training images to learn meaningful model. Further, it also requires the same dimensional training images, a requirement not followed in IITM-SCID2 dataset, wherein the images need to be resized. On the other hand, our framework based on modified ResNet18 can handle different dimensional images. Our framework also outperforms the approach in [1], which uses depth map of scenes along with their intensity images.

## 5 Summary and Future Work

In this paper, we have developed a framework in order to address the indoor–outdoor scene classification task based on residual convolution neural network (ResNet18). Further, we have demonstrated the ability of our framework to classify images with different weather conditions, and is able to classify images as well as depth maps with ideal accuracies in many cases. Our high-quality results indicate that such a framework can be useful in the pipelines for single-image-based depth estimation, as well as scene image analysis and retrieval based on color as well as depth maps.

## References

- Pillai, I., Satta, R., Fumera, G., Roli, F.: Exploiting depth information for indoor-outdoor scene classification. In: Maino, G., Foresti, G.L. (eds.) *Image Analysis and Processing—ICIAP 2011*, pp. 130–139. Springer, Berlin (2011)
- Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor-outdoor image classification. *IEEE Trans. Image Process.* **17**(12), 2381–2392 (2008)
- Das, S., Ahuja, N.: Performance analysis of stereo, vergence, and focus as depth cues for active vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(12), 1213–1219 (1995)
- Saxena, A., Sun, M., Ng, A.Y.: Make3d: learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(5), 824–840 (2009)
- Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: *Advances in Neural Information Processing Systems*, pp. 2366–2374 (2014)
- Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5162–5170 (2015)
- Saber, E., Tekalp, A.M.: Integration of color, shape, and texture for image annotation and retrieval. In: *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 3, pp. 851–854 (1996)
- Devroye, L., Györfi, L., Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*. Springer (1996)

9. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification* (2nd edn). Wiley-Interscience (2000)
10. Payne, A., Singh, S.: Indoor vs. outdoor scene classification in digital photographs. *Pattern Recogn.* **38**(10), 1533–1545 (2005)
11. Gupta, L., Pathangay, V., Patra, A., Dyana, A., Das, S.: Indoor versus outdoor scene classification using probabilistic neural network. *EURASIP J. Adv. Signal Process.* **2007**(1), 094298 (2006)
12. Havens, T.C., Bezdek, J.C., Leckie, C., Hall, L.O., Palaniswami, M.: Fuzzy c-means algorithms for very large data. *IEEE Trans. Fuzzy Syst.* **20**(6), 1130–1146 (2012)
13. Tao, L., Kim, Y.H., Kim, Y.T.: An efficient neural network based indoor-outdoor scene classification algorithm. In: 2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE), pp. 317–318 (2010)
14. Wu, J., Rehg, J.M.: Centrist: a visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(8), 1489–1501 (2011)
15. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 1. NIPS’14, pp. 487–495. MIT Press, Cambridge, MA, USA (2014)
16. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision—ECCV 2014*, pp. 346–361. Springer International Publishing, Cham (2014)
17. Yoo, D., Park, S., Lee, J., Kweon, I.: Fisher kernel for deep neural activations. *CoRR abs/1412.1628* (2014)
18. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR abs/1311.2524* (2013)
19. Kang, K., Wang, X.: Fully convolutional neural networks for crowd segmentation. *CoRR abs/1411.4464* (2014)
20. Zhao, F., Huang, Y., Wang, L., Tan, T.: Deep semantic ranking based hashing for multi-label image retrieval. *CoRR abs/1501.06272* (2015)
21. Dixit, M., Chen, S., Gao, D., Rasiwasia, N., Vasconcelos, N.: Scene classification with semantic fisher vectors. In: CVPR, IEEE Computer Society, pp. 2974–2983 (2015)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105. Curran Associates Inc. (2012)
23. Yoo, D., Park, S., Lee, J.Y., Kweon, I.S.: Multi-scale pyramid pooling for deep convolutional representation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 71–80 (2015)
24. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR abs/1406.4729* (2014)
25. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* pp. 1–1 (2018)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, IEEE Computer Society, pp. 770–778 (2016)
27. Gao, B., Wei, X., Wu, J., Lin, W.: Deep spatial pyramid: The devil is once again in the details. *CoRR abs/1504.05277* (2015)
28. Gupta, S., Pradhan, D., Dileep, A.D., Thenkanadiyoor, V.: Deep spatial pyramid match kernel for scene classification. In: ICPRAM, pp. 141–148 (2018)
29. Zhu, C.: Place recognition: an overview of vision perspective. *CoRR abs/1707.03470* (2017)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014)
31. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015)

32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
33. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: CVPR (2016)
34. Silberman, N., Fergus, R.: Indoor scene segmentation using a structured light sensor. In: Proceedings of the International Conference on Computer Vision—Workshop on 3D Representation and Recognition (2011)
35. Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: learning from RGB-D data in indoor environments. CoRR **abs/1709.06158** (2017)

# Comparison Between LGBP and DCLBP for Non-frontal Emotion Recognition



Hardik Dosi, Rahul Keshri, Pravin Srivastav and Anupam Agrawal

**Abstract** Emotion recognition has captured the attention of many researchers these days. However most of the researches have happened in emotion recognition from frontal faces. In real-world conditions, we may not always be able to capture the frontal faces. Hence, emotion recognition from non-frontal faces is the new research area. The Local Binary Patterns (LBP) is an important feature extraction technique. In our effort to find better variant of LBP for non-frontal emotion detection, we have used two possible variants namely Local Gabor Binary Pattern (LGBP) and Diagonal Crisscross Local Binary Pattern (DCLBP). The LGBP is further implemented with and without Angle Classification; which leads to total three methods of feature extraction. These three methods are used to classify the images based on facial emotion expressions. An image is divided into number of blocks. Feature vectors are created by concatenating histograms computed from each sub-block. Multi-class SVMs are used to classify angles and expressions. A comparative analysis of the three methods for non-frontal emotion recognition has been carried out and is presented in this paper. By analyzing different variants of a LBP, we can understand the importance of feature representation in non-frontal emotion recognition. Our experimental studies show that the LGBP with angle classification outperforms other two variants.

**Keywords** Emotion recognition · Non-frontal · LBP · LGBP · DCLBP

---

H. Dosi · R. Keshri · P. Srivastav (✉) · A. Agrawal

Interactive Technologies and Multimedia Research Lab, Indian Institute of Information Technology, Allahabad, India  
e-mail: [pis2017001@iiita.ac.in](mailto:pis2017001@iiita.ac.in)

A. Agrawal  
e-mail: [anupam@iiita.ac.in](mailto:anupam@iiita.ac.in)

## 1 Introduction

Since the last decade of the twentieth century, many efforts have been made to improve the interaction between human beings and computers. According to a study, it has been observed that when two human beings interact with each other's verbal cues provide only 7% of the meaning of the message and vocal cues about 38%. However, facial expressions provide 55% of the meaning of the message [1]. Therefore, we can say that facial expressions provide a significant amount of information during the interaction between human beings.

Emotion recognition depends on the feature being used to represent facial emotions. In general, two categories of feature extraction methods exist which are used for facial expression recognition: geometric-based methods and appearance-based methods [2]. The features extracted using geometric-based methods contain information about the location and shape of facial features whereas the features extracted using appearance-based methods contain information about the changes in appearance of the face which includes furrows, wrinkles, and bulges.

The features extracted using geometric-based methods are sensitive to noise whereas the features extracted using appearance-based methods are less dependent on initialization and are capable of encoding micro-patterns present in the skin texture which is important for facial expression recognition. The Local Binary Patterns (LBP) which represents textures have therefore widely been used to extract emotions from the non-frontal faces [3]. Among all the variants, LGBP has been the most efficient one for emotion recognition in case of non-frontal faces [4–6]. In our effort to find better variants of LBP for emotion recognition in non-frontal faces, we are going to compare the results of emotion recognition using DCLBP from that obtained using LGBP.

Two different methods have been applied to detect emotion. This has been done to see the behavior of LBP-based features in case of different methods [7–9]. In one case, the emotion is being recognized after angular classification while in the other case emotion recognition is being done without angular classification.

Datasets such as KDEF [10] are now available for non-frontal images. We have compared the results of LGLBP and DCLBP in case of non-frontal images by applying different methods. Multi-class Support Vector Machines (SVMs) have been used for classification.

This paper is structured as follows: Sect. 2 describes the work done in the field of non-frontal emotion recognition. Section 3 describes the methodology for non-frontal emotion recognition. Section 4 gives details of the result analysis. Section 5 consists of conclusions and future scope.

## 2 Related Work

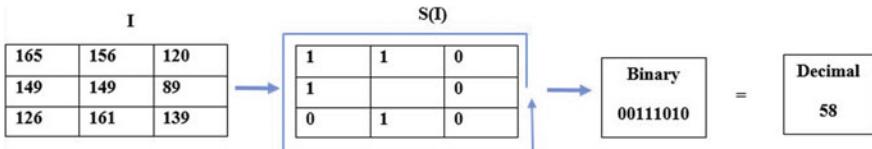
Work is being done in the field non-frontal emotion recognition. Zeng et al. [11] in his approach extracted features from the images by using manually labeled feature points in the image. However manual labeling of points in the image is a limitation for non-frontal images. Hu et al. [12] classified expressions across five different angles of the head by applying LBPs, HOGs, and SIFT techniques. The authors then used SVMs for classification. However, the authors also tried to classify expressions by extracting features with manually labeled facial points. Hu et al. [12] concluded that non-frontal views perform better than frontal views for expression recognition. On the other hand, Moore et al. [13] suggested that frontal view is better than non-frontal views for facial expression recognition. Moore et al. analyzed appearance-based LBP features and LBP variants on BU3DFE [14] and multi-pie dataset [15]. Seyedehsamaneh et al. [16] detected emotion without going for pose detection. He used the movement of facial landmarks. Tian et al. [17] used a neural network based technique to classify expressions on the PETS dataset. Tang et al. [9] used BU3DFE dataset [14] and applied Markov Model Super vectors and K- Nearest Neighbors to make the classification process rotation and scaling invariant but the performance was not satisfactory on all the expressions other than happy.

## 3 Proposed Approach

In this paper, we have used three different approaches to compare the efficacy of LBP-based features for non-frontal emotion recognition. The first approach involves extracting features from the images and forming the feature vectors. These feature vectors are passed into SVM to find the final emotion. Unlike the first approach, the second approach involves angle classification before emotion recognition. The second approach has been adapted from one proposed by Moore and Bowden [7] with slight optimization. The third approach involves DCLBP, which is a texture descriptor. DCLBP was proposed by Hossain and Shahera [8]. This descriptor has never been used before to extract features from images to perform expression classification. We have used Support Vector Machines (SVM) for classification because of its ability to classify high dimensional data quickly.

### 3.1 Feature Extraction

We have explored and compared two variants of LBP, namely Local Gabor Binary Patterns (LGBP) and Diagonal Crisscross Local Binary Pattern (DCLBP) for feature extraction. The basics of LBP, LGBP, and DCLBP are as follows:



**Fig. 1** Calculation of LBP for a pixel with a neighborhood of  $3 \times 3$  [7]

**Local Binary Pattern (LBP).** LBP is an operator which is applied to an image pixel-by-pixel to extract features from it. In many computer vision applications LBPs have been successfully applied as a texture descriptor [6]. LBP is computationally very simple and is tolerant to illumination change. The steps to compute LBP of an image are as follows. A  $3 \times 3$  window of pixels is considered around a center pixel  $f_c$ . The neighboring pixels of the center pixel are labeled as  $f_p$  where  $p \in \{0, \dots, 7\}$ . These neighboring pixels are compared against the threshold (the center pixel) to obtain a binary string. An example of the calculation of LBP operator is shown in Fig. 1. The threshold function  $S(f_p - f_c)$  is given by Eq. (1).

$$S(f_p - f_c) = \begin{cases} 1 & \text{if } f_p \geq f_c \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The value of LBP for the central pixel  $f_c$  is computed by multiplying each bit of binary string by  $2^p$ , as given in Eq. (2)

$$LBP = \sum_{p=0}^7 S(f_p - f_c) 2^p \quad (2)$$

**Uniform Local Binary Pattern.** Uniform Local Binary Pattern is a particular case of Local Binary Pattern. A local binary pattern which has at most two bitwise transitions from 0 to 1 or 1–0 for a circular binary string is called as uniform local binary pattern. 00000110 and 1011111 are examples of uniform local binary pattern whereas 01010110 is an example of non-uniform local binary pattern. By using uniform local binary patterns the histogram size gets reduced from 256 bins to 59 bins.

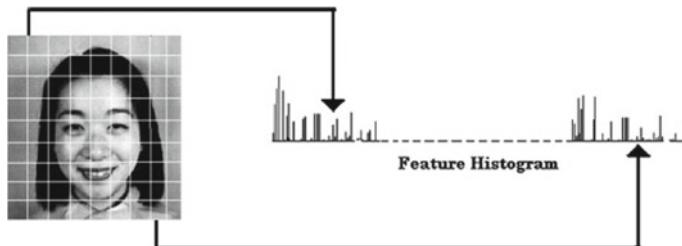
The histogram of the image, or any sub-region of it, obtained after application of LBP operator is used as a feature vector. This feature vector contains the necessary texture information about the image.

**Local Gabor Binary Pattern (LGBP).** Before applying local binary pattern (LBP) to an image, if Gabor filter is applied to the image, then the obtained LBP map is known as Local Gabor Binary Pattern (LGBP) [7]. Gabor filter is very useful in detecting edges from the image. It has been already used successfully for frontal expression recognition [18]. This is why it is an extremely good technique for feature extraction, texture representation, and texture discrimination. Application of Gabor filter along with local binary pattern enhances the power of the histogram obtained

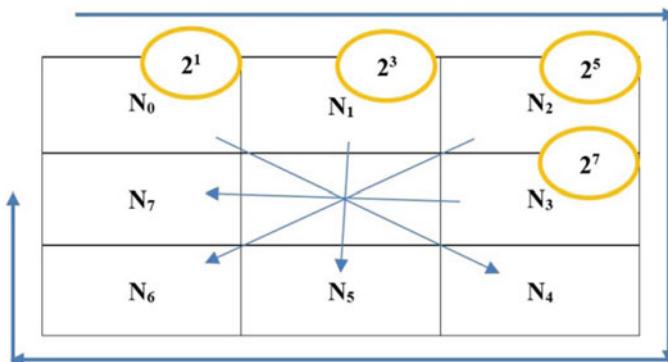
from the image. In order to extract the LGBP feature vector of the image the following procedure is followed. By using eight different orientations  $\vartheta \in \{0, \dots, 7\}$ , eight Gabor filters are obtained. Each of these eight filters is applied to the original image. This gives us a set of eight Gabor kernels. Uniform local binary pattern is calculated for each kernel to obtain eight LGBP maps. Feature vector is obtained from these maps by dividing each LGBP map into  $9 \times 9 = 81$  blocks. A 59-bin histogram is computed for each block. All 81 histograms obtained from all the blocks of the maps are concatenated to form the feature vector as shown in Fig. 2.

**Diagonal Crisscross Local Binary Pattern (DCLBP).** The diagonal crisscross LBP operator is a variant of the original LBP operator. DCLBP [8] considers the diagonal variations in the intensity of the pixel values as shown in Fig. 3. The steps to calculate DCLBP [8] as shown in Fig. 3 are as follows:

1. Calculate the differences  $N_0-N_4$ ,  $N_1-N_5$ ,  $N_2-N_6$ , and  $N_3-N_7$  as shown in Fig. 3.
2. For each positive difference assign the respective weights as shown in Fig. 3 and add the results.
3. Finally, replace the center pixel value with the calculated value of step 2.



**Fig. 2** Dividing the image into 81 blocks with histogram creation and con-catenation [7]

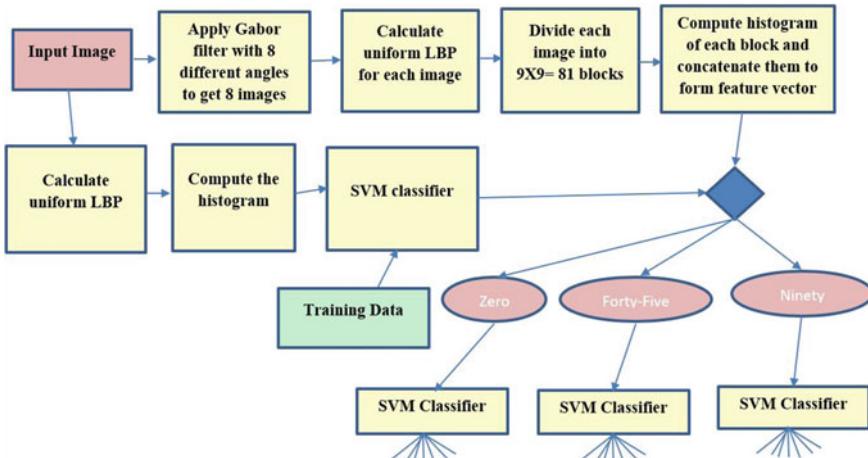


**Fig. 3** The notations associated and calculation of DCLBP [8]

### 3.2 Method

**LGBP without angle classification.** In this method, eight Gabor filters were applied to the input image which resulted in eight different Gabor kernels. Now, for each of these Gabor Kernels, a corresponding LBP image was created using the LBP operator. Each of these obtained LBP images was divided into  $9 \times 9 = 81$  blocks to capture local information from the image. The reason for dividing the image into 81 blocks and not in 49 or 64 blocks is because we found better accuracy with 81 blocks after trying to vary this value. Now for each of these blocks a 59-bin histogram was computed and finally concatenated to form a feature vector [7] as shown in Fig. 4. The number of dimensions of this feature vector are  $38,232 (=59 \times 81 \times 8)$ . Finally, this feature vector was passed to the SVM classifier. The SVM classifier was previously trained by using similar feature vectors using labeled data.

**LGBP with angle classification.** This method is quite similar to the one in which LGBP was used to detect emotion without angle classification. The method is shown in Fig. 4. The only difference here is an additional step where the image is being classified for knowing the angle from which the image has been taken, before classifying the emotion for the image. There is also an increase in the number of SVM classifiers because here different SVM is trained for detecting emotions for different angle. Since the KDEF [10] dataset has images taken from 0, 45 to 90°, three different SVM classifiers were used to detect emotions for the given three different angles. For example, one SVM was trained to detect emotions for all the images which contains image at 0° angle. Similarly we have different SVMs to detect emotions for different angles. Histogram from the Uniform LBP of the image was used to form feature vector for SVM with angle output for the given image. Based on the output of SVM



**Fig. 4** Flowchart showing the procedure of Expression classification with angle classification

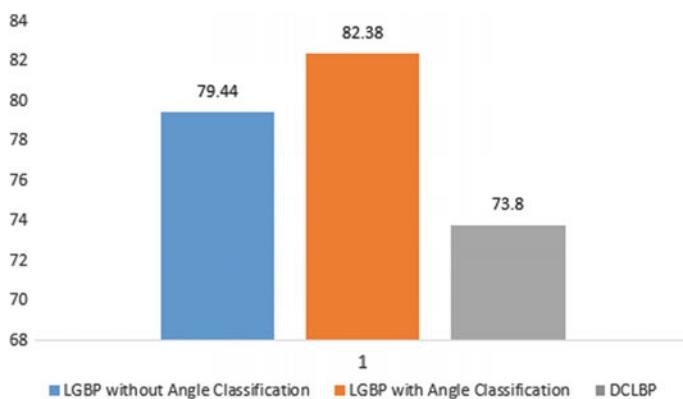
with angle output, the LGBP is transferred to the SVM for the respective angle. The SVM for the respective angle then classifies the given image to detect the correct emotion.

**Diagonal Crisscross LBP without angle classification.** In this method, DCLBP of the input image was computed. After this, the complete image was converted into a one-dimensional vector to form the feature vector. This feature vector was then passed on to the SVM to detect the final emotion for the given image.

## 4 Result Analysis

A 10-fold cross validation was performed on KDEF [10] dataset to determine the accuracy of all the three approaches. The KDEF dataset is divided into 10 equal sets with each set containing images of 14 persons with all their expressions from all the three angles ( $14 \times 6 \times 3 = 252$  images in each set). The SVMs were trained on nine sets and tested on one set. This procedure was applied 10 times with a different test set every time. Figure 5 shows the overall accuracy of the three methods, i.e., LGBP without angle classification, LGBP with angle classification and Diagonal Crisscross LBP.

Figure 5 analyses the result method wise. Highest accuracy was obtained by using LGBP with angle classification with 82.38% accuracy followed by LGBP without angle classification having 79.44% accuracy. The method involving DCLBP had an accuracy of 73.80%. The approach of using DCLBP as a feature extraction method performed the poorest among the three and therefore we can conclude that DCLBP is not a better descriptor for facial expression recognition than LGBP. The LGBP with angle classification performed better than LGBP without angle classification due to the added step of classification of angle for an image. The confusion matrices obtained from the three methods are shown in Tables 1, 2, 3 and 4.



**Fig. 5** The accuracies of the three methods

**Table 1** Confusion matrix of angle classification for LGBP with angle classification

	Zero	Forty-five	Ninety
Zero	94.76	3.33	0.95
Forty-five	4.76	93.09	5.95
Ninety	0.48	3.75	92.86

**Table 2** Confusion matrix of expression classification for LGBP without angle classification

	Happy	Sad	Surprise	Anger	Disgust	Fear
Happy	92.86	1.43	0.48	1.9	1.9	2.86
Sad	1.43	74.76	0.95	6.67	13.33	13.81
Surprise	1.43	3.81	89.52	0.48	0.95	9.52
Anger	0.95	6.19	0.95	84.29	4.76	2.38
Disgust	0.95	2.38	0.48	3.33	74.23	6.19
Fear	3.81	10.95	7.62	2.86	4.76	65.71

**Table 3** Confusion matrix of expression classification for LGBP with angle classification

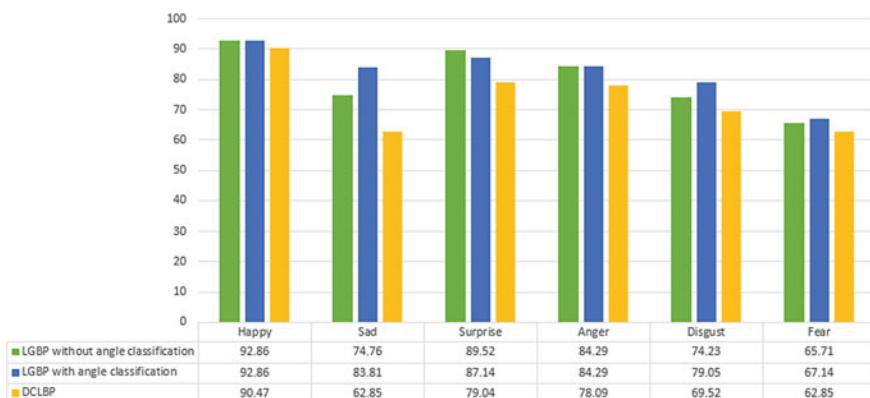
	Happy	Sad	Surprise	Anger	Disgust	Fear
Happy	92.86	0.95	1.43	0	2.38	2.38
Sad	0.48	83.81	0.95	5.71	7.14	10.48
Surprise	0	0.95	87.14	0.48	0.48	10.48
Anger	1.43	5.24	0.48	84.29	7.14	5.24
Disgust	3.33	3.81	1.9	5.71	79.05	4.29
Fear	1.9	4.76	8.09	3.81	3.81	67.14

**Table 4** Confusion matrix of expression classification for DCLBP classification

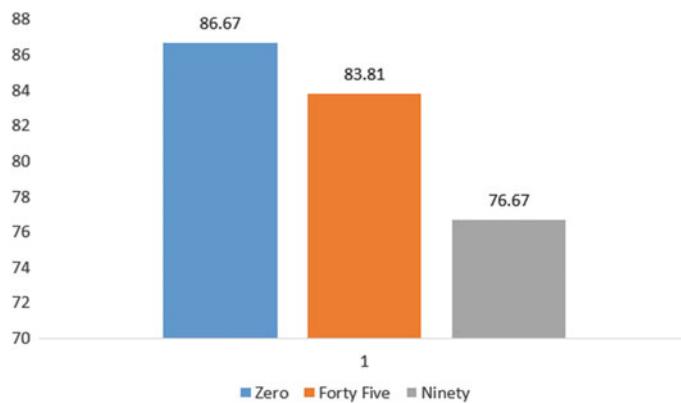
	Happy	Sad	Surprise	Anger	Disgust	Fear
Happy	90.47	1.9	1.9	4.76	7.61	6.66
Sad	1.9	62.85	0.95	6.66	5.71	9.52
Surprise	1.9	3.8	79.04	1.9	1.9	17.14
Anger	0.95	13.33	0.95	78.09	11.42	2.85
Disgust	2.85	4.76	0.95	6.66	69.52	1.9
Fear	2.85	12.38	18.09	1.9	3.8	62.85

From Table 1 we can see that the performance of the method LGBP with angle classification is highest because of good accuracy for given by SVM for angle output. The overall accuracy for angle prediction is pretty high (93.57%). The best performing angle is “Zero” followed by “Forty-Five”. This is because of the fact that, with increase in the angle, less portion of the face is available for feature extraction and therefore poorer is the extracted feature.

Figure 6 shows a comparison between expression-wise accuracy observed for the three methods. The best performing expression is “Happy” followed by “Surprise”. This is because the changes in the appearance of the face are quite distinctive and unique than in case of other expressions. From Fig. 6 we can see that the expression “Happy” has performed well in all the three approaches. “Fear” expression has performed the poorest in all these approaches. Figure 7 analyses the result angle-wise LGBP with angle classification. It can be seen from Fig. 7 that images taken from



**Fig. 6** Expression wise comparison of accuracies obtained using the three techniques



**Fig. 7** Angle-wise comparison of accuracy of LGBP with angle classification

ninety degrees have least accuracy of 76.67% for emotion detection. It is because only a small portion of the face is visible from that angle. Extracting sufficiently distinctive features from facial images with 90° angle is a difficult task. Images taken from 0° have the highest accuracy of 86.67% among all the angles for emotion detection. The reason is that most of the facial part is visible when the face is at 0° angle.

## 5 Conclusion and Future Work

In this paper, we are trying to find the efficacy of a new variant of LBP called Diagonal Crisscross local binary pattern (DCLBP) for non-frontal emotion detection. The performance of DCLBP is compared to LGBP for non-frontal emotion detection using in one case without angle classification and in the other case with angle classification. The KDEF [10] dataset has been used for analysis. The result clearly shows the superiority of LGBP over the new variant of LBP called the DCLBP. The result also shows that the emotion can be detected best when the face is at angle zero degrees because most of the facial features are visible. The emotion detection becomes very poor when the face is at angle 90° when the least facial features are visible.

The results of emotion detection are best when the emotion classification is done after knowing the facial angle. Thus LGBP with angle classification gives better results than LGBP without angle classification.

In future, the techniques used in this paper could be extended with some other datasets which have more number of angle variations. In this paper, we considered angle variations of face about y-axis only. Angle variations about x-axis can be considered for further extension of this work. Unsupervised classification techniques can be explored.

## References

1. Mehrabian, A.: *Silent Messages*. Wadsworth Publishing Company, Inc., Belmont, CA, (1971)
2. Tian, Y., Kanade, T., Cohn, J.: Facial expression analysis. In: *Handbook of Face Recognition*. Springer, (Chapter 11) (2005)
3. Lyons, J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* **21**, 1357–1362 (1999)
4. Gong, S., McOwan, P.W., Shan, C.: Dynamic facial expression recognition using a bayesian temporal manifold model. In: *Proceedings of the British Machine Vision Conference*, vol. 1, pp. 297–306 (2006)
5. Liao, S., Fan, W., Chung, A.C.S., Yeung, D.Y.: Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features, In: *ICIP*, pp. 665–668 (2006)
6. Shan, C., Gritti, T.: Learning discriminative lbp-histogrambins for facial expression recognition. In: *Proceedings of the British Machine Vision Conference* (2008)
7. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. *Comput. Vis. Image Underst.* **115**(4), 541–558 (2011)

8. Hossain, S.: Study on Co-occurrence-based image feature analysis and texture recognition employing diagonal-crisscross local binary pattern. Technical Report. Kyushu Institute of Technology, Japan (2013). <http://hdl.handle.net/10228/5282>
9. Tang, H., Hasegawa-Johnson, M., Huang, T.: Non-frontal view facial expression recognition based on ergodic hidden Markov model supervectors. In: 2010 IEEE International Conference on Multimedia and Expo (ICME) (2010)
10. Lundqvist, D., Flykt, A., Öhrman, A.: The Karolinska Directed Emotional Faces—KDEF. In: CD ROM from Department of Clinical Neuroscience, Psychology Section. Karolinska Institute (1998). ISBN 91-630-7164-9
11. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.S.: Multi-view facial expression recognition. In: FG2008, 2008 ICPR 2008. 8th IEEE International Conference on Automatic Face and gesture Recognition, Automatic Face & Gesture Recognition (2008)
12. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.S.: A study of non-frontal-view facial expressions recognition. In: 19th International Conference on Pattern Recognition. ICPR 2008, pp. 1–4 (2008)
13. Moore, S., Bowden, R.: The effects of pose on facial expression recognition. In: Proceedings of British Machine Vision Conference (BMVC2009), London, UK (2009)
14. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, pp. 211–216. <https://doi.org/10.1109/fgr.2006.6>. 6 Apr 2006
15. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. *Image Vis. Comput.* **28**(5), 807–813 (2010)
16. Shoailelangari, S., Yau, W.-Y., Teoh, E.-K.: Pose-invariant descriptor for facial emotion recognition. *Mach. Vis. Appl.* **27**(7), 1063–1070 (2016)
17. Tian, Y., Brown, L., Hampapur, A., Pankanti, S., Senior, A., Bolle, R.: Real world real-time automatic recognition of facial expressions. In: Proceedings of IEEE Workshop (2003)
18. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J.: Dynamics of facial expression extracted automatically from video. *J. Image Vis. Comput.* 615–625 (2004)

# Co-Detection in Images Using Saliency and Siamese Networks



Milan Zinzuvadiya, Vatsalkumar Dhameliya, Sanjay Vaghela, Sahil Patki, Nirali Nanavati and Arnav Bhavsar

**Abstract** Co-Detection is an important problem in computer vision, which involves detecting common objects from multiple images. In this paper, we address the co-detection problem and propose an integrated deep learning model involving two networks for co-detection. Our proposed model detects the objects of individual images using a convolutional neural network by generating the saliency maps, which are passed as input in a Siamese neural network to ascertain whether the salient objects in both the images are similar or different. We have tested our model on the iCoseg dataset achieving high-quality results.

**Keywords** Co-detection · Deep learning · Siamese network · Saliency network

## 1 Introduction

Co-Detection/Co-segmentation is a task of detecting common or similar objects from multiple images. Such a task can have useful implications for organization, indexing, and retrieval of images. Generally, the task of locating common objects across a set

---

M. Zinzuvadiya · V. Dhameliya · S. Vaghela · S. Patki  
Sarvajanik College of Engineering and Technology, Surat 395001, Gujarat, India  
e-mail: [milanzinzuvadiya99@gmail.com](mailto:milanzinzuvadiya99@gmail.com)

V. Dhameliya  
e-mail: [vatsal17137@gmail.com](mailto:vatsal17137@gmail.com)

S. Vaghela  
e-mail: [vaghelasanjay96@gmail.com](mailto:vaghelasanjay96@gmail.com)

S. Patki ·  
e-mail: [sahilpatki239@gmail.com](mailto:sahilpatki239@gmail.com)

N. Nanavati · A. Bhavsar (✉)  
Indian Institute of Technology Mandi, Mandi, India  
e-mail: [arnav@iitmandi.ac.in](mailto:arnav@iitmandi.ac.in)

N. Nanavati  
e-mail: [nirali.nanavati@scet.ac.in](mailto:nirali.nanavati@scet.ac.in)

of images is considered as a pixel-level localization (i.e., the co-segmentation task). Some existing approaches of co-segmentation include the use of Markov Random Field models (MRF) [1, 2], co-saliency-based methods [3, 4], discriminative clustering [2], etc. Some of these approaches consider that different regions in images are largely homogeneous, i.e., pixels have the same properties like color, contrast, intensity, and texture. This helps in employing a smoothness constraint-based energy, which can then be minimized using graph-cut-based methods. The GrabCut-based methods [5–7] refine a coarse segmentation estimate. Considering foreground and background dissimilarity, it involves an energy term of inter-image region matching in MRF to explore co-detection. These methods work efficiently when subjected to images having similar backgrounds and co-segments the foreground object. However, when subjected to images with different backgrounds and variation in pose, size, and viewpoints, these methods do not generate satisfactory results. The co-saliency methods assume that in most of the images, the detected salient areas contain at least parts of the foreground object. While this is a fair assumption, some approaches also consider the notion of repetitiveness, i.e., they concentrate on only those parts of saliency maps that are frequently repeated in most images.

Contrary to traditional methods, in this work, we take a different route considering two aspects. First, instead of considering a co-segmentation problem (involving pixel-level labeling), we consider a co-detection task, which involve detection of similar objects at a bounding box level. With a renewed interest in object detection in recent years (with approaches such as YOLO, Region CNN, and their faster variants becoming popular), we believe that a co-detection problem is sufficient and suitable for many applications. Second, we make use of recent deep learning methods, and integrate these to provide a co-saliency based alternate solution for the co-detection problem, which does not involve imposing smoothness or repetition constraints.

Our proposed approach consists of two steps: For achieving a coarse segmentation of objects, we use a saliency network proposed known as Non-Local Deep Features for Salient Object Detection (NLDF) [8]. We then use the images based on the saliency outputs, in a Siamese network to ascertain whether the objects are similar or different. We demonstrate that our co-detection approach yields high-quality results using a dataset containing different backgrounds, poses, and contrast in images.

## 2 Related Work

As mentioned above, traditional methods of co-segmentation use Markov Random Field-based energy functions and optimization techniques [2], and GrabCut MRF [7]-based methods. A positive aspect about such methods is their unsupervised nature. Such methods have tested simple features such as color and grayscale, edges, or texture, as well as more complex features such as object, focus, and background.

In [9], the authors use an energy minimization approach. The cost function proposed in this paper combines spectral and discriminative clustering. The optimization of this model is carried out using an efficient EM method.

The approach in [10] consists of three steps: First, the large regions which are shared across the set of images are detected that induce affinity between those images. Second, these detected shared regions are used to score multiple overlapping segment candidates. Finally, in the third step, accuracy of co-segmentation is improved by propagating the likelihood between different images.

In [11], the authors cast image co-detection and co-localization into a single optimization problem that integrates information from low-level appearance cues with that of high-level localization cues in a very weakly supervised manner. The optimization problem is over bounding box and superpixel level. This method leverages two representations at different levels and simultaneously discriminates between foreground and background at the bounding box and superpixel level using discriminative clustering.

All the abovementioned methods mainly work on smoothness-based energy minimization approach, GrabCuts, discriminative clustering, affinity between large co-objects, etc. These methods essentially try to differentiate between foreground and background by using different methods, and related costs or scores for pixels are used in order to co-segment similar objects.

Closely related to our approach is the work in [12], which uses Deep CNN for solving image similarity. While the task is not exactly that of co-detection that we consider, it involves a similar notion of computing image similarity. Here, SimNet mentioned in [12], which is a deep Siamese network is trained on pairs of positive and negative images using a pair mining strategy inspired by Curriculum learning. The authors also use a multi-scale CNN, where the final image embedding is a joint representation of top as well as lower layer embeddings. They use contrastive loss function in their implementation to minimize the loss.

Unlike the above methods, the foreground object in our approach is detected using a recently proposed saliency network [8], which involves learning local features from training images, through convolutional neural network. Once we get the salient map for images, we process these images to obtain selective regions, and pass them into a Siamese network, which gives us similarity between objects. While this is similar to the image similarity task in [12], our overall approach is arguably, simpler. Moreover, in our proposed method, instead of contrastive loss, we make use of triplet loss for Siamese training which has been shown to be an effective loss for Siamese networks.

### 3 Proposed Approach

We now discuss the key concepts used in our proposed model, which includes the saliency and Siamese networks, and the overall approach.

### 3.1 Saliency

In the context of images, saliency means the most noticeable object present in an image. Traditional methods detect salient objects by extracting local pixel-wise or region-wise features and compare them with global features to get saliency score [13]. Using this Saliency score, saliency maps are generated consisting of salient objects. Recently, deep learning has taken center stage in computer vision tasks and even saliency detection problem has been widely attempted using deep learning models. Methods like [4, 8, 14, 15] which use CNNs (Convolutional neural network) have been found very effective in such salient object detection tasks. We use NLDF method [8] to generate saliency map in our approach.

### 3.2 Siamese Network

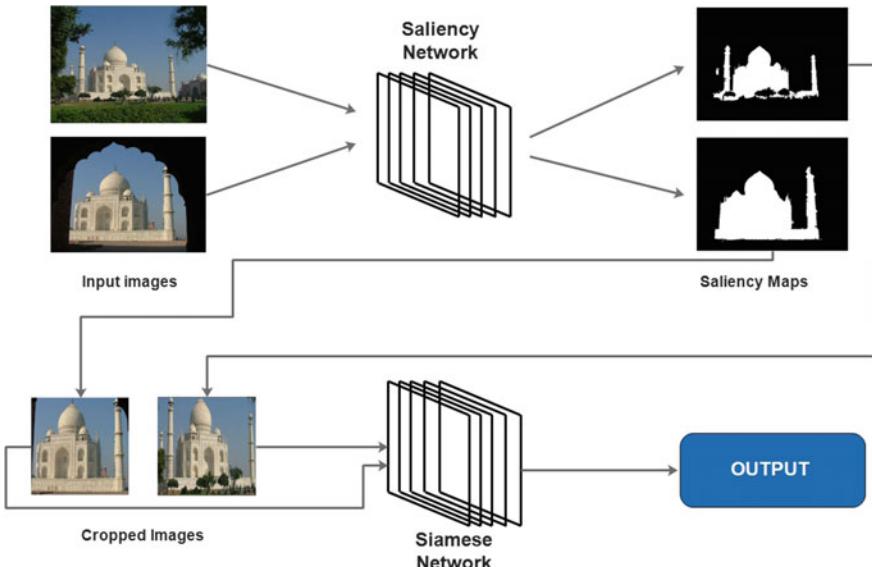
Siamese neural network is used to find similarity between two comparable things. For example, images, signatures [16], and sentences [17]. A Siamese network involves two or more identical subnetworks having the same parameters. Subnetworks are designed in such a way that they share the same weights and during the training process, these weights are updated simultaneously across each subnetwork. We use Siamese network to ascertain whether the salient objects present in the two images are similar or not.

**Triplet Loss** Traditional Siamese Network uses contrastive loss, which is based on only two inputs. The value of loss is dependent on whether they are similar or dissimilar. In our proposed approach, during training of the Siamese network, we use relative distance loss function which is known as Triplet loss function [18], wherein we give three images at the time of training: Anchor image, Positive image (which has the same object as Anchor image) and Negative image (which has different object than Anchor image). During training, the network is expected to differentiate between the anchor-positive image pair and the anchor-negative image pair.

### 3.3 Overall Methodology

We use Saliency and Siamese network to perform co-detection in images. In Fig. 1, we show the overall flow of our proposed model during the testing phase, where we process the pairs of images from a set of images over which the co-detection task is to be carried out. The pair-wise processing of images in this manner, helps in identifying the images which contain similar salient object.

**Salient object detection and processing:** Given multiple images for which the co-detection task has to be addressed, these are passed into a pretrained saliency network to generate saliency maps for these images. We use these saliency maps to

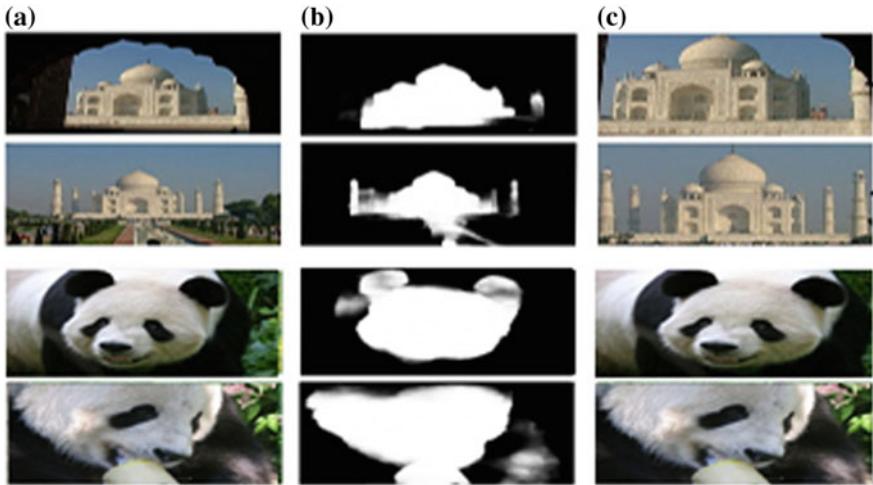


**Fig. 1** Proposed model using Saliency and Siamese networks

extract the salient object from the images. Note that the object extraction process essentially feeds into the detection process, where a bounding box can be decided based on the extent of the saliency map, once the object passes the similarity test of the Siamese network. Usually in the case of input images having the same object with different sizes, it is highly possible that Siamese network may give wrong output. This happens because of network failing to learn object features due to size differences (especially with relatively small datasets). In our proposed approach, we eliminate this variance in object size by preprocessing RGB input images before feeding it into the Siamese network. They are cropped in such a way that the input image to Siamese network contains only scaled salient objects as shown in Fig. 2.

**Siamese network:** As shown in Table 1, the Siamese subnetwork contains a total of six convolution blocks from CONV-1 to CONV-6 with each block containing one convolution layer followed by max-pooling layer. Here, we apply convolution in two dimensions, that is, the width and height of the image and apply rectified linear(ReLU) activation function on output of each convolution layer except block CONV-6. We use max-pooling operation of stride 2, which down samples every depth slice of the input by factor of 2 along with both width and height while depth dimension remains unchanged, e.g., feature map of dimension  $480 * 480 * 16$  converted to  $240 * 240 * 16$ , same as  $240 * 240 * 32$  converted to  $120 * 120 * 32$  and so on. As shown in Table 1, at the end of the CONV-6, the Siamese subnetwork gives us the feature vector of dimension  $1 \times 128$ .

We train our Siamese network with 10 k triplets images for 8 epochs with a batch size of 2. For each epoch, three input images are loaded into *anchorImageOutput*,



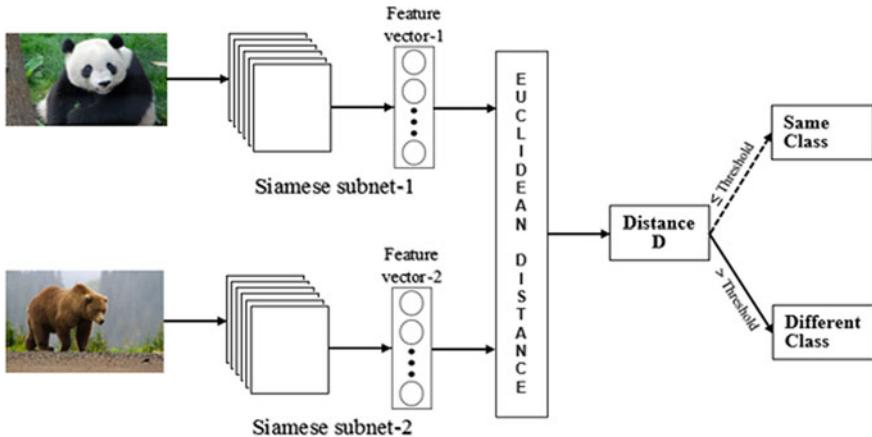
**Fig. 2** Saliency Output using [13] after preprocessing where, **a** shows original images, **b** are saliency maps and **c** shows cropped images

**Table 1** Siamese subnetwork architecture

Block	Layer	Kernel	Output
Conv1	Convolutional	10 * 10	480 * 480 * 16
	Max-pooling	2 * 2	240 * 240 * 16
Conv2	Convolutional	7 * 7	240 * 240 * 32
	Max-pooling	2 * 2	120 * 120 * 32
Conv3	Convolutional	5 * 5	120 * 120 * 64
	Max-pooling	2 * 2	60 * 60 * 64
Conv4	Convolutional	3 * 3	60 * 60 * 128
	Max-pooling	2 * 2	30 * 30 * 128
Conv5	Convolutional	1 * 1	30 * 30 * 256
	Max-pooling	2 * 2	15 * 15 * 256
Conv6	Convolutional	1 * 1	15 * 15 * 2
	Max-pooling	2 * 2	8 * 8 * 2

*positiveImageOutput*, and *negativeImageOutput*, respectively. These images are fed separately into each subnetwork of Siamese network. During training, the loss is calculated by using *TripletLoss* function. We have used *Gradient Descent Optimizer* to minimize the loss.

Figure 3 is the testing architecture for our Siamese network. While testing we have two Siamese subnetworks, whereas while training, we have three Siamese subnetworks. This is because in training process, the network tries to learn the difference between similar images, i.e., positive pairs and different images, i.e., negative pairs.



**Fig. 3** Siamese network testing Architecture

Triplet Loss function used in our network requires three input images and hence, there are three subnetworks during training. However, during testing, we do not require any loss function, hence we just calculate Euclidean distance between two images and based on that value, we can decide whether images have similar objects or different objects.

## 4 Experimental Results

We now describe various aspects of our experimentation. We then provide and discuss our results.

### 4.1 Dataset

To train and validate our approach, we use iCoseg dataset [19, 20]. This dataset mainly consists of 4 main categories, viz., Landmarks, Sports, Animals, and Miscellaneous. These 4 categories are further divided into a total of 38 classes. We preprocess classes having sparse data and obtain 35 classes. There are a total of 650 images in the dataset. We are using triplet loss [18] for minimizing the error during training process. Therefore, in order to train our Siamese network, we require triplets of images. Hence, we generate 10,000 training triplets from iCoseg dataset. Triplet contains three images from which two of the images are from the same class (anchor image and positive image) and one is from a different class (negative image). Images are resized to  $480 \times 480$  pixels for training. For testing our network, we generated

2000 testing pairs from iCoseg dataset, which contains 1000 positive pairs and 1000 negative pairs. Both training triplets and testing pairs contain different images.

## 4.2 Implementation Details

Our Siamese network is implemented using TensorFlow [21] library of Python, which contains three Siamese subnetworks and weights are shared between them in order to decrease learning parameters. All the weights and the biases of convolution layers are initialized by using Xavier initializer.

For our model, we have used Gradient Descent Optimizer with learning rate of 0.001. The preprocessing of images is done by using OpenCV [22] library of Python, saliency map are converted to RGB and bounding box are cropped containing only salient object. The margin value for triplet loss function is set to 1. We have used ReLU [23] as our activation function because it reduces the likelihood of vanishing gradient [24], the model is trained for 8 epochs with batch size of 2. We use the Euclidean distance. The threshold for identifying whether images are similar or not is set to 1.86375, which was empirically decided. We have experimented with various thresholds and provide these results.

## 4.3 Evaluation Metrics

For evaluating our proposed model, we calculate F-score and Jaccard index. F-score is the harmonic mean of precision and recall. Jaccard similarity index is the intersection of union between ground truth and segmented result. For evaluating Siamese network, we use F-score and segmentation results are evaluated using Jaccard index.

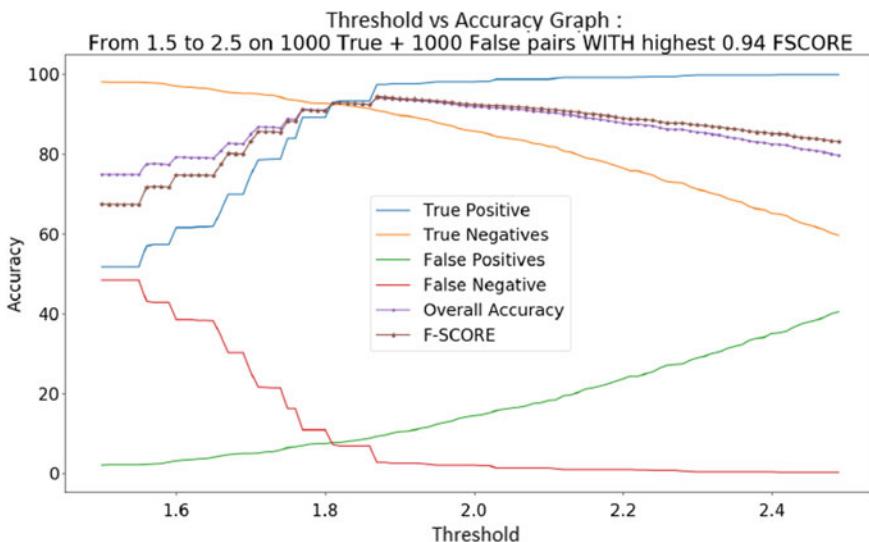
## 4.4 Results

Table 2 lists out epoch-wise F-score and Overall Accuracy values along with its corresponding threshold value. The threshold value is set in such a way that the F-score and overall accuracy maximizes. We observe that the F-score is quite high starting from epoch 5, and largely remains consistent. At epoch 8 we achieve the maximum F-score value of 0.9442 along with overall accuracy 94.25 at a threshold value of 1.86375. This indicates that the proposed approach is indeed performing effectively, in correctly matching the salient objects across all classes.

Figure 4 is a graph of Threshold versus Accuracy for the last (eighth) epoch. We have plotted Overall Accuracy, F-score, True Positive, True Negative, False Positive and False Negative for a range of threshold value from 1.5 to 2.5 on 1000 positive, i.e., pair of images with the same object and negative pairs, i.e., pair of images with

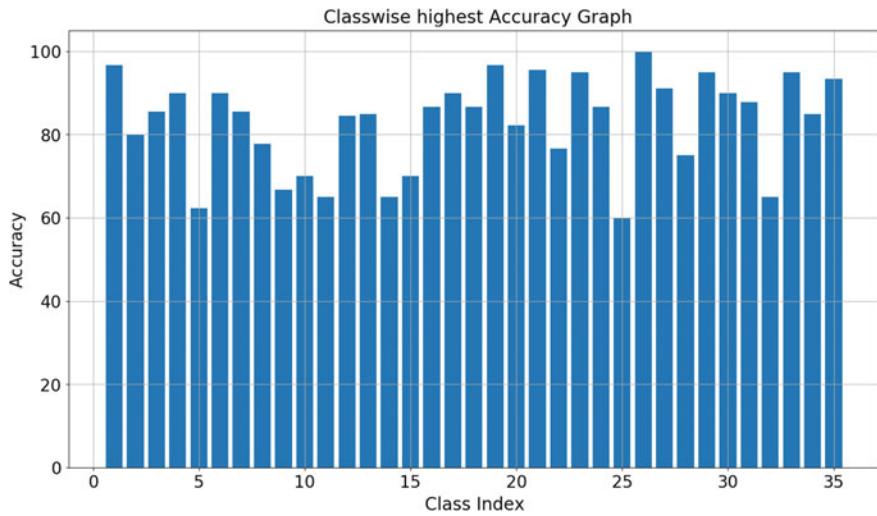
**Table 2** Epoch-wise F-score and overall accuracy

Epochs	F-Score	Threshold	Overall Accuracy
Epoch 1	0.85	1.72010	83.20
Epoch 2	0.81	1.62403	79.05
Epoch 3	0.86	1.51912	84.70
Epoch 4	0.92	1.35486	91.85
Epoch 5	0.91	1.74252	91.00
Epoch 6	0.93	1.44106	93.35
Epoch 7	0.89	1.91879	90.05
Epoch 8	0.94	1.86375	94.25

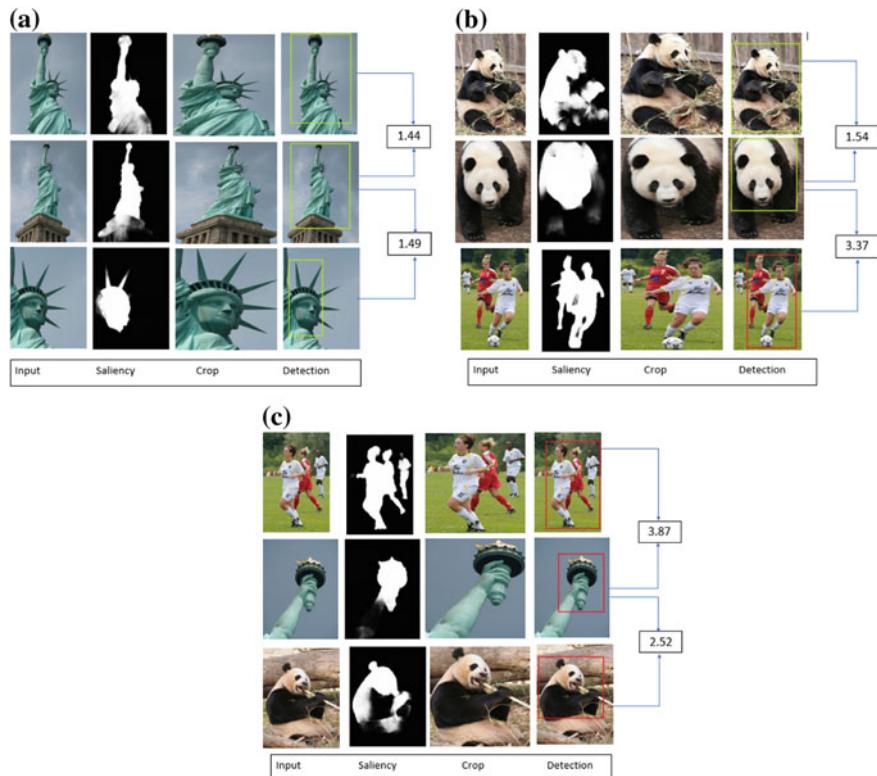
**Fig. 4** Result of 1000 positive + 1000 negative pairs

different objects. We can see that maximum F-score is obtained at threshold value of 1.86. Figure 5 shows the bar chart for class-wise accuracy obtained on iCoseg dataset. It is observed that the classes having more number of training images are showing better accuracy, which is quite natural. However, in general, one can observe that for most of the classes, the performance is quite high.

Finally, in Fig. 6, we show some visual results for co-detection. In Fig. 6a, an example with the same images in a set is shown. The green bounding boxes denote a correct co-detection, and the scores from the Siamese network are shown between two pairs of images. We note that in both the cases, the scores are similar and small, despite significant scale and view changes. In Fig. 6b, an example image set containing a different image is depicted. Here, we note that the scores from the Siamese network are small (and similar to the earlier case) for the images containing



**Fig. 5** Classwise accuracy for iCoseg dataset



**Fig. 6** Some visual results for co-detection in iCoseg dataset

the same salient entity (top-two), in spite of view variations. On the other hand, the score is large between images which contain different objects. Correspondingly, the correct co-detection (green bounding boxes) naturally occurs between images with the same entity. The red box denotes a different object in the set. Similarly, Fig. 6c depicts a case, where there are no common objects, and hence the relative scores are high. As a result, there are no co-detections (denoted by red boxes.)

## 5 Conclusion and Future Work

We have proposed an approach for addressing the co-detection problem using a deep learning framework which integrates a saliency network and a Siamese network. Saliency network is used to obtain salient objects (which are likely to be the common objects across images), and the Siamese network is used for comparing those objects to ascertain whether objects from given images are similar or not. Our approach is able to effectively differentiate between similar and dissimilar salient objects. Experimental analysis on iCoseg dataset proves that for a large variety of object classes, our approach provides high-quality results in ascertaining whether given images have a similar object or not. As future work, one can plan to employ our co-detection approach for the task of image retrieval or annotation. Also, while there are various co-segmentation methods, but not many co-detection approaches, such an approach may be reviewed and compared along with combination of object detection methods and similarity estimation methods.

## References

1. Yu, H., Xian, M., Qi, X.: Unsupervised co-segmentation based on a new global gmm constraint in mrf. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 4412–4416 (2014)
2. Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching—incorporating a global constraint into mrfs. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR '06, vol. 1, pp. 993–1000. IEEE Computer Society, Washington, DC, USA (2006)
3. Zhang, Q., Gong, S., Liu, C., Ying, W.: Similar background image co-segmentation with co-saliency. In: 2017 International Smart Cities Conference (ISC2), pp. 1–2 (2017)
4. Kim, J., Pavlovic, V.: A shape preserving approach for salient object detection using convolutional neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 609–614 (2016)
5. Lattari, L., Montenegro, A., Vasconcelos, C.: Unsupervised cosegmentation based on global clustering and saliency. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2890–2894 (2015)
6. Malmer, T.: Image segmentation using grabcut. **5**, 1–7 (2010)
7. Gao, Z., Shi, P., Reza Karimi, H., Pei, Z.: A mutual grabcut method to solve co-segmentation, **2013** (2013)

8. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6593–6601 (2017)
9. Joulin, A., Bach, F., Ponce, J.: Multi-class cosegmentation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 542–549 (2012)
10. Faktor, A., Irani, M.: Co-segmentation by composition. In: 2013 IEEE International Conference on Computer Vision, pp. 1297–1304 (2013)
11. Sharma, A.: One shot joint colocalization and cosegmentation. arXiv preprint [arXiv:1705.06000](https://arxiv.org/abs/1705.06000) (2017)
12. Appalaraju, S., Chaoji, V.: Image similarity using deep CNN and curriculum learning. arXiv preprint [arXiv:1709.08761](https://arxiv.org/abs/1709.08761) (2017)
13. Xi, X., Luo, Y., Li, F., Wang, P., Qiao, H.: A fast and compact salient score regression network based on fully convolutional network. arXiv preprint [arXiv:1702.00615](https://arxiv.org/abs/1702.00615) (2017)
14. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep cnn features. IEEE Trans. Image Process. **25**(11), 5012–5024 (2016)
15. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5300–5309. IEEE (2017)
16. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)
17. Neculoiu, P., Versteegh, M., Rotaru, M.: Learning text similarity with siamese recurrent networks. In: Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 148–157 (2016)
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
19. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: icoseg: Interactive co-segmentation with intelligent scribble guidance. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2010)
20. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: Interactively co-segmentating topically related images with intelligent scribble guidance. Int. J. Comput. Vis. **93**(3), 273–292 (2011)
21. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint [arXiv:1603.04467](https://arxiv.org/abs/1603.04467) (2016)
22. Culjak, I., Abram, D., Pribanic, T., Dzapo, H., Cifrek, M.: A brief introduction to OpenCV. In: 2012 Proceedings of the 35th International Convention MIPRO, pp. 1725–1730 (2012)
23. contributors, W.: Rectifier (neural networks)—wikipedia, the free encyclopedia (2018), Accessed 4 Apr 2018
24. contributors, W.: Vanishing gradient problem—wikipedia, the free encyclopedia (2018), Accessed 4 Apr 2018

# **Hand Down, Face Up: Innovative Mobile Attendance System Using Face Recognition Deep Learning**



**Aditi Agrawal, Mahak Garg, Surya Prakash, Piyush Joshi and Akhilesh M. Srivastava**

**Abstract** Computer Vision is considered as the science and technology of the machines that see. When paired with deep learning, it has limitless applications in various fields. Among various applications, face recognition is one of the most useful real-life problem-solving applications. We propose a technique that uses image enhancement and facial recognition technique to develop an innovative and time-saving class attendance system. The idea is to train a Convolutional Neural Network (CNN) using the enhanced images of the students in a certain course and then using that learned model, to recognize multiple students present in a lecture. We propose the use of deep learning model that is provided by *OpenFace* to train and recognize the images. This proposed solution can be easily installed in any organization, if the images of all persons to be marked this way are available with the administration. The proposed system marks attendance of students 100% accurately when captured images have faces in right pose and are not occluded.

**Keywords** Automated attendance management · Mobile biometrics · Face recognition · Image enhancement · OpenFace · Convolutional neural network (CNN)

---

A. Agrawal · M. Garg · S. Prakash (✉) · P. Joshi · A. M. Srivastava  
Discipline of Computer Science and Engineering, Indian Institute of Technology Indore, 453552  
Indore, India  
e-mail: [surya@iiti.ac.in](mailto:surya@iiti.ac.in)

A. Agrawal  
e-mail: [aditi.agarwal1695@gmail.com](mailto:aditi.agarwal1695@gmail.com)

M. Garg  
e-mail: [mahak2449@gmail.com](mailto:mahak2449@gmail.com)

P. Joshi  
e-mail: [phd1301101004@iiti.ac.in](mailto:phd1301101004@iiti.ac.in)

A. M. Srivastava  
e-mail: [phd1701101001@iiti.ac.in](mailto:phd1701101001@iiti.ac.in)

## 1 Introduction

Attendance in academics plays a huge role in determining the grades of a student, and also, is a useful measure for the teacher to assess the performance and growth of students. Attendance has traditionally been registered by calling out the name of each student, or by passing around a sheet of paper where each student writes their name in order to declare that they are present in class. Both the above two methods have their shortcomings. The calling out of the name of each student is very time-consuming and not immensely practical in a large classroom where a lot of precious class time gets wasted in marking attendances, leaving little time for the actual lesson. The second method, where all students write down their names on a sheet of paper for attendance is usually prone to students covering up for their friends absent from the class by fraudulently writing down their absent friend's name. Additionally, the exercise of each student writing down their name on a piece of paper one by one is also a time-consuming process, with a huge probability of inaccuracy.

To tackle such issues that arise in traditional attendance management of academic institutions, and to make the process of attendance-taking more swift and efficient, a lot of methods have been devised. Technology has been instrumental to bringing to life a wide range of novel solutions for attendance management, such as, fingerprint scanners, iris scans, RFID tags, face recognition, etc. These newer solutions make the exercise of registering attendances less manual and consequently, more robust, but, they also have some shortcomings. The fingerprint scanners installed in some academic institutions can be compared to the traditional attendance marking process, as they require students to press their finger on a fingerprint scanner for a few seconds, which marks them present after a few seconds of processing. The time taken is sufficiently large when we consider the fact that each student has to spend a few seconds, or sometimes even minutes (if their fingerprints does not match in a single trial), on the fingerprint scanner. Thus, fingerprint scanners are also time-consuming and might require more than one trial for a single student. Similar is the case of iris scanners. An additional issue is that the fingerprint scanners and iris scanners are installed in a particular corner of the classroom and every student has to leave his/her seat to reach the scanner and mark themselves present. This time taken also adds to the latency of the attendance management, eventually, wasting a lot of precious class time. RFID tags, on the other hand are not that time-consuming, but, pose the risk of getting lost or students carrying more than their own RFID tag to mark the attendance of not just themselves, but also of their friends who are absent from the class. The above-discussed technologies are also physically intrusive, as the person has to specifically position himself/herself in a particular place and then wait for some time to be correctly recognized by the system.

Most of the above issues are solved by a huge extent when we use facial recognition to mark the attendance of students in a classroom. It is not time-consuming, since the students do not have to declare themselves present in front of any system, rather, the teacher captures a photograph of all the students in a few frames, such that a union of all the faces in each image provides us with the set of all the students present in the classroom. It is also not physically intrusive like the other attendance systems.

Facial recognition is a biometric method of identifying an individual by comparing live capture or digital image data with the stored record for that person. The basic idea here is to use Facial Recognition in a classroom full of students to automatically mark their attendance for the corresponding course lecture. Using simple logic, that, if a student is not present in class, his/her face will not be present in the image that the teacher takes, hence, his/her attendance should not and will not be marked. Rest of the paper is organized as follows. In Sect. 2, review of the existing work is presented. In Sect. 3, few preliminary techniques are presented which are required in the proposed technique. In Sect. 4, the proposed technique is presented. Experimental analysis is discussed in Sect. 5. Finally, paper is concluded in Sect. 6 which is the last section.

## 2 Literature Review

The problem of marking attendance of any particular group correctly and efficiently, while not wasting a considerable amount of time of the group, has been discussed and researched upon in the past. Correctly marked attendance is a necessity when it comes to the scenario of a classroom, where attendance of a pupil can be loosely associated with their growth and personal improvement, while also providing the teacher with a means to keep track of each pupil's regularity and sincerity. An automated attendance system also eliminates the chances of false attendance being registered, which might happen due to negligence of the teacher or students registering attendance for their friends who might be absent from class.

There has been tremendous research on the topic of marking attendance, in order to improve efficiency, and also reduce the time taken. In [1], a smartphone-based student attendance system is discussed, which stores attendance data in a web server, to be easily accessible to the teacher, who is the only one authorized to mark and also modify attendance of students. An attendance vigilance system using wireless network with biometric fingerprint authentication can be found in [2]. A bimodal biometric student attendance system using face recognition in combination with fingerprint authentication is presented in [3], to make the algorithm more robust and less prone to spoofing. In [4], a stress-free non-intrusive attendance system is developed using face recognition, which uses a camera to acquire facial images that are made into templates using Fisherfaces algorithm. The Fisher Linear Discrimination algorithm is used while taking attendance to compare stored image with the acquired image and find a match. It also has the additional facility of providing information related to class attendance to handheld devices via available cellular networks. To reduce the hassle of entering attendance in logbooks manually, which are prone to easy manipulation, in [5], an automated attendance system called AUDACE has been presented, which marks attendance using face recognition, where faces are matched with the help of Principal Component Analysis. The absentee list is read aloud through a voice conversion system. In [6], a human face recognition based user authentication system for student attendance has been proposed which implements face recognition by combining Discrete Wavelet Transforms (DWT) and Discrete Cosine Transform (DCT) to extract the features of student's face, followed by applying Radial Basis

Function (RBF) for classifying the facial objects. In [7], an automated attendance system is built using Eigen Face database and PCA algorithm with MATLAB GUI. In [8], a student monitoring system using a camera mounted on the entrance of the classroom to capture frontal images of students for face recognition is used. Another attendance system using face recognition is presented in [9] which utilizes the eigen-face algorithm. An automated attendance system using video-based face recognition is proposed in [10], where a video of the students in a classroom is input to the system, which outputs a list of students present.

### 3 Preliminaries

We have developed an innovative attendance system using face recognition and deep learning. In this section, we state all the required preliminaries which are instrumental to developing our proposed technique.

#### 3.1 Convolutional Neural Network

Convolutional Neural Networks are neural networks made up of neurons that have certain weights and biases assigned to them, which are particularly well-adapted to classify images. Each neuron receives several inputs, performs a dot product over them, and optionally passes it through an activation function and responds with an output. CNN has a loss function on last layer to learn end-to-end mappings. Convolutional neural networks allow networks to have fewer weights and they are considered as a very effective tool in the field of computer vision which includes image recognition and classification.

#### 3.2 Technologies Used

The technologies we used to develop our attendance system were *OpenFace*, *Docker*, and *Django*. *OpenFace* [11] is a Python and Torch implementation of face recognition with deep neural networks and is based on FaceNet [12]. We used a *Docker* [13] container to create, deploy, and run our attendance system using *OpenFace*. *Docker* allowed us to package up all libraries and dependencies into one container, which gave us the great advantage that our application can run on any machine regardless of any difference it has from the machine we used to write and test our code. The final web application with a complete User Interface was created in *Django* [14].

## 4 Proposed Technique

In this section, we discuss the proposed technique for taking the attendance in a classroom using images of the class. We first capture images of students from our input device (a smartphone with a decent camera). We perform enhancements to augment its quality and then recognize faces in that image to mark students present in the class.

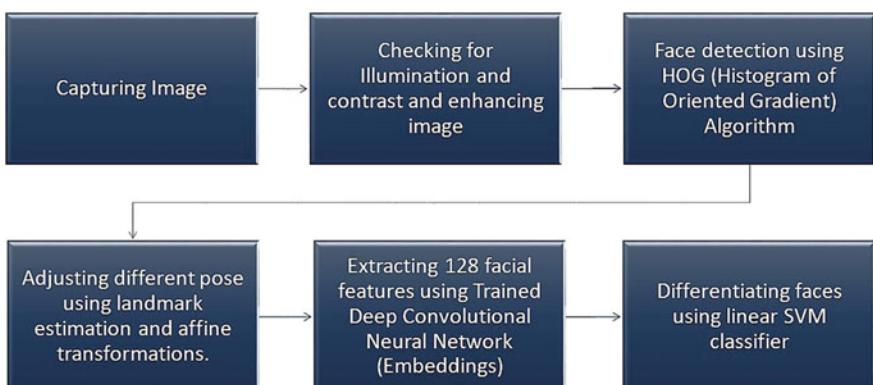
A flowchart of the proposed technique can also be seen in Fig. 1.

### 4.1 Data Acquisition

The Data Acquisition phase includes the registration of all students on the attendance system using frontal face images captured with the help of any camera with a satisfactory resolution that does not lead to the loss of features of their faces. This data is used to train a deep learning model that learns each person's face in the database. For real-time data acquisition and maintaining attendance, we built a Web application using *Django*.

### 4.2 Preprocessing

In the preprocessing step, the main goal is to enhance the quality of captured image while preserving the important features of each frontal face. For this purpose, the image is processed, and contrast-based and illumination-based quality estimations are made for the images. A single image is classified into four types: *Illuminated*, *Dull*, *Shadow*, and *Dark*. The purpose of image enhancement preprocessing step is to



**Fig. 1** Flowchart of the proposed system

check the illumination and contrast in the image and then apply the respective techniques (Homomorphic Filtering & Histogram Equalization) in order to enhance the prospects of obtaining better features. The two techniques described above namely Homomorphic Filtering and Histogram Equalization are applied on images depending on which class they fall (i.e., *Dull*, *Shadow*, or *Dark*). Illuminated images do not need any enhancement. If the images that fall under the class *Shadow*, Homomorphic Filtering is used. If the image falls under *Dull*, Histogram Equalization is used. Finally, if the image falls *Dark*, both Homomorphic Filtering and Histogram Equalization are applied on the image. [15].

### 4.3 Face Detection

After the enhanced image is obtained, the operations required to begin the process of marking attendance can be started. As our technique utilizes taking a picture of the class and subsequently, recognizing faces of pupils appearing in the picture to mark attendance, our first step is to detect the face of each person, which can later be fed into a machine learning algorithm that correctly identifies the image and then, another module marks the attendance for the person. For face detection, the algorithm used is Histogram of Oriented Gradients (HOG). Now for every pixel we draw an arrow showing the direction in which the image gets darker. Image is divided into small areas with  $16 \times 16$  pixels in each and this area is replaced by an arrow (Gradient for every pixel would be too much detail). Now, to find faces we need to find the part in the image that looks very similar to the known HOG pattern of a face (Extracted from lots of images). To deal with faces turned a little sideways or in a different pose than a full frontal face image, we use Face Landmark Estimation. The idea is to come up with 68 points that exist on every face. These are called landmarks. Example of landmarks is shown in Fig. 2 To have the 68 features in a face roughly at the same places, we use affine transformations.

### 4.4 Face Recognition

Now, we have a centered face image, which is processed by *OpenFace*. *OpenFace* makes use of a Deep Convolutional Neural Network that recognizes 128 different features (Embeddings) for a face. We save those measurements (features) for each face that we are training the model for. We train a linear SVM classifier with the help of each face's generated features. The input to the classifier is each face in the input sample image, which then tries to find a match for each sample face with a face in the database.



**Fig. 2** All the detected faces are shown enclosed in bounding boxes. Also, the face landmarks of each face have been demonstrated

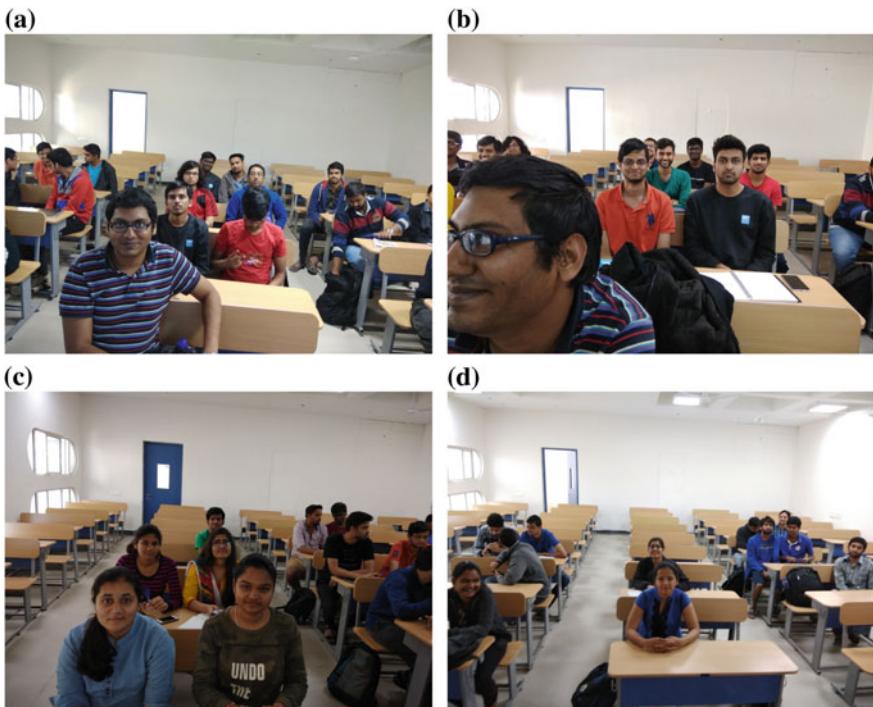
#### 4.5 Real-Time Data Acquisition

Our attendance system application enables each teacher to mark the attendance of each student with great accuracy and ease. The teacher captures and uploads a few images of the class such that a union of all faces present in each image gives us the faces of all the students present in class that day. These images are immediately processed using a classifier, which takes detected faces as inputs and tries to find a match among the database of student faces. The roll numbers of students for whom a match is found are saved in a text file, which is used to mark attendance for the students present. A flowchart of the proposed technique can also be seen in Fig. 1. Some examples of images are shown in Figs. 3 and 4.

### 5 Experimental Results

#### 5.1 Database

Our database consists of five to six images of 24 subjects that were captured using the camera from different smartphones. These images would be used to generate embeddings for each subject. In these images, some were taken with the subject facing front and some with the subject facing a little bit sideways at an angle. For testing purpose we captured 66 images for a course at different days in which these



**Fig. 3** Some example images that gave substandard accuracy. **a** Some of the faces are not facing the camera properly (bad pose). **b** Some faces are hidden due to occlusion. **c** A lot of faces are in the wrong pose. **d** Faces turned around, looking down (wrong pose)

students were enrolled. These images were taken keeping in mind that there is proper illumination inside the class.

## 5.2 Performance Analysis

In this section, the means employed to determine the performance quantifiers are discussed.

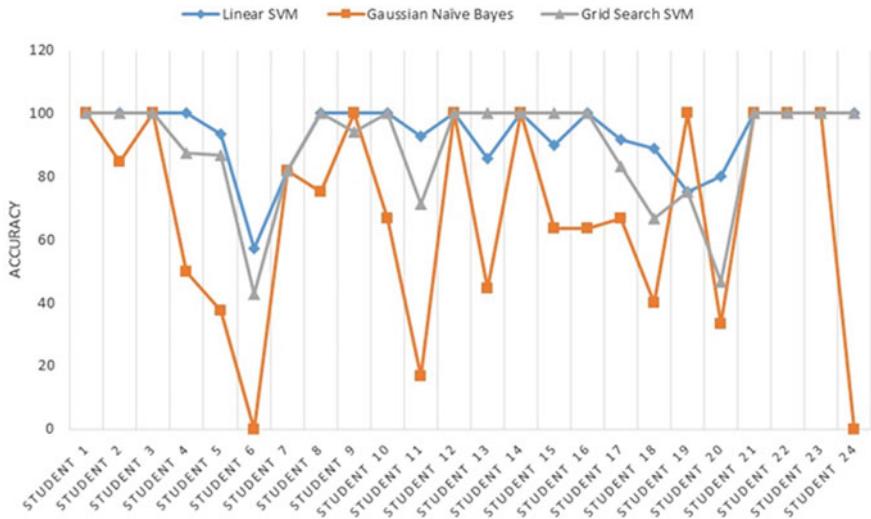
We experimented the classification part using different classifiers. The one with the best results was the linear SVM classifier. The table in Fig. 5 shows the output for linear SVM classifier. It shows the accuracy for each subject being present on different days. The accuracy can be calculated by

$$\text{Accuracy} = \frac{\text{Faces recognized correctly}}{\text{Total number of faces present in all lectures}}$$



**Fig. 4** Some example images that gave 100% accuracy

**Fig. 5** Table showing results for linear SVM classifier. Rows represent students, Columns represent days, “1” in the table shows that that the student was present and marked correctly, “0” indicates that the student was absent. Lastly, “-1” shows that the student was present but marked incorrectly



**Fig. 6** A plot of accuracy for all the three classification algorithms that we used to test our model. The three classification algorithms are linear support vector machine (Linear SVM), Gaussian Naive Bayes and Grid-Search Support Vector Machine (Grid-Search SVM), respectively

We had a total of 229 faces, and 212 were recognized correctly. The best average accuracy came out to be 92.5%. We calculated the accuracy for different classifiers, namely—Gaussian Naive Bayes, Grid-Search SVM, and Linear SVM classifier. The accuracy for Grid-Search SVM was 89.0% and the accuracy for Gaussian Naive Bayes came out to be 67.66%. The graph in Fig. 6 shows the accuracy subject-wise for different classifiers. We can see that the Linear SVM classifier and Grid-Search SVM classifier are comparable, but Linear SVM classifier has the best accuracy. The accuracy of the technique was hindered by some faulty images. Some faulty images are shown in Fig. 3. The reasons for substandard accuracy were bad posture and overlapping of faces (Occlusion) due to which the algorithm was not able to detect and recognize faces. Some corrective measures could be kept in mind while clicking the photograph. Every student that the photograph covers should have their face up and properly visible in a correct posture. Also none of the faces should be occluded by other faces. Few samples of good images that gave 100% accuracy are shown in Fig. 4.

### 5.3 User Interface

As a final step in our technique, we have built a Web application which we call as *Hand Down Face Up*. This application implements the front end of the proposed technique to produce a final product which can be used to mark as well as maintain

the attendance. We have described the working of the application in two parts below. The first one provides the details of the front end whereas the second part outlines the backend of the application.

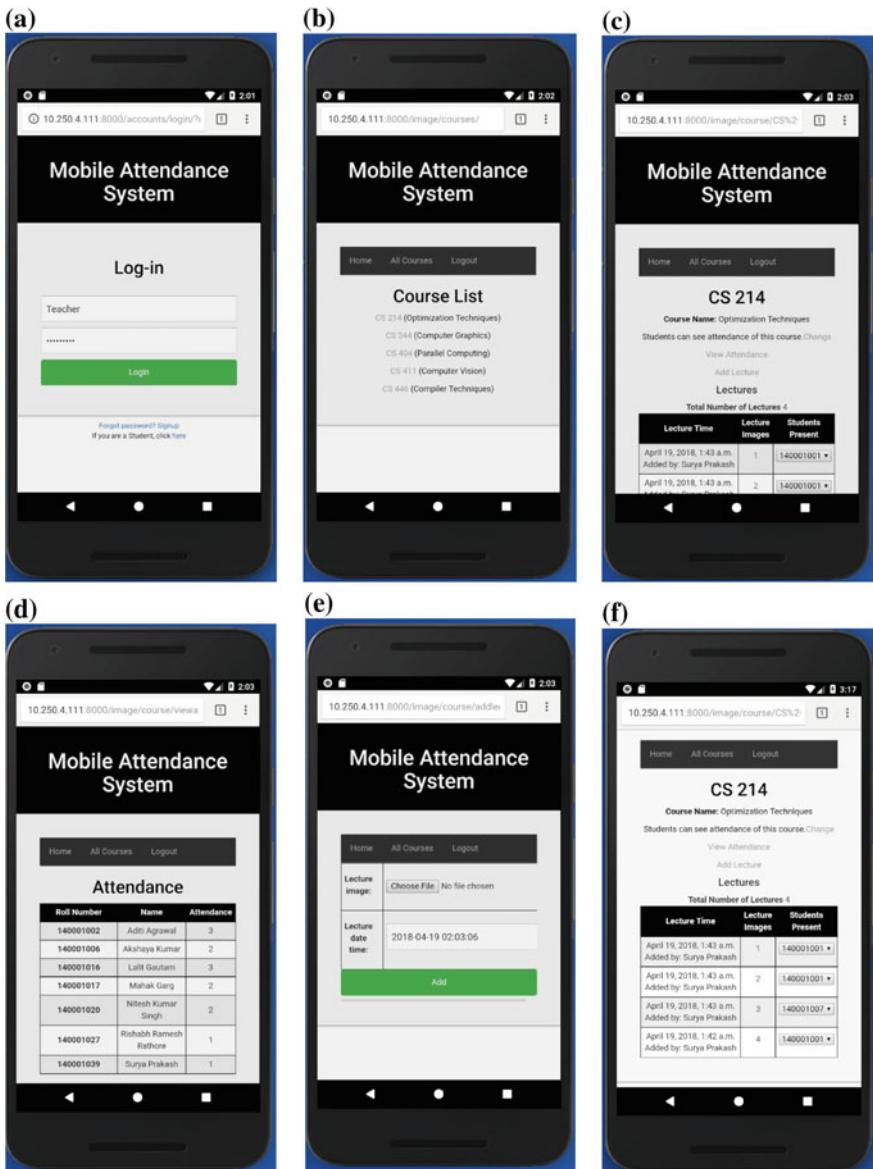
Home screen of the application is a Sign-In page, which allows teachers to access their personalized course data. A teacher's personal space allows him/her to add images of the current lecture using the camera from his/her smartphone. These images are then processed and then the attendance of the students is marked (automated). Also, the user interface allows teachers to view the complete attendance of a course, including the images of the lectures that they took. The application also has an option for students to view their attendance of the courses they are enrolled in.

The teacher captures and uploads a maximum of five images on the application. These images are then processed using the embeddings generated from training (using *OpenFace*) and a text file is generated with all the student's Roll Number who are present. The application then uses this text file and marks the attendance of students with respect to course and lecture. Figure 7 shows the User Interface for a teacher.

## 6 Conclusions

In this paper, we have proposed a fast, automated, and feasible attendance system. Our proposed technique is capable of identifying faces of students with the help of facial features, and thus, increases the convenience for marking the attendance of students. Initially, five to six images of each student are collected, enhanced, and a deep network is trained which learns all these faces using 128 features (Embeddings) of each face. In our developed web application, the teacher signs in to his/her personal account and adds images for the lecture. These images are preprocessed and faces are detected. Following it, these detected faces are recognized using a suitable classifier. The students whose faces are recognized are marked present. The proposed system is found to be marking the attendance of the students 100% accurately when linear SVM classifier is used and when images used for attendance marking contain properly posed and non-occluded faces.

**Acknowledgements** Authors are thankful to the volunteers who willingly gave their images for this study.



**Fig. 7** User interface of hands down, face up mobile attendance web application. **a** Log-in page for teacher. **b** Course list which is displayed after the teacher selects *all courses*. **c** View of a course when it is selected from the *all courses* menu, **(d)** and **(e)** show the screens that appear on choosing *view attendance* and *add lecture* options, respectively, from the course view. **f** Lecture timings and list of students present

## References

1. Islam, M.M., Hasan, M.K., Billah, M.M., Uddin, M.M.: Development of smartphone based student attendance system. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 230–233, Dhaka, (2017). <https://doi.org/10.1109/r10-htc.2017.8288945>
2. Adal, H., Promy, N., Srabant, S., Rahman, M.: Android based advanced attendance vigilance system using wireless network with fusion of bio-metric fingerprint authentication. In: 2018 20th International Conference on Advanced Communication Technology (ICACT), pp. 217–222, Chuncheon-si Gangwon-do, Korea (South), (2018). <https://doi.org/10.23919/icact.2018.8323702>
3. Charity, A., Okokpujie, K., Etinosa, N.O.: A bimodal biometric student attendance system. In: 2017 IEEE 3rd International Conference on Electro-Technology for National Development (NIGERCON), pp. 464–471, Owerri (2017). <https://doi.org/10.1109/nigercon.2017.8281916>
4. Okokpujie, K., Noma-Osaghae, E., John, S., Grace, K.A., Okokpujie, I.: A face recognition attendance system with GSM notification. In: 2017 IEEE 3rd International Conference on Electro-Technology for National Development (NIGERCON), pp. 239–244, Owerri (2017). <https://doi.org/10.1109/nigercon.2017.8281895>
5. Poornima, S., Sripriya, N., Vijayalakshmi, B., Vishnupriya, P.: Attendance monitoring system using facial recognition with audio output and gender classification. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp. 1–5, Chennai (2017). <https://doi.org/10.1109/icccsp.2017.7944103>
6. Lukas, S., Mitra, A.R., Desanti, R.I., Krisnadi, D.: Student attendance system in classroom using face recognition technique. In: 2016 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1032–1035, Jeju (2016). <https://doi.org/10.1109/ictc.2016.7763360>
7. Rekha, E., Ramaprasad, P.: An efficient automated attendance management system based on Eigen face recognition. In: 2017 7th International Conference on Cloud Computing, Data Science Engineering—Confluence, pp. 605–608, Noida, (2017). <https://doi.org/10.1109/confluence.2017.7943223>
8. Chintalapati, S., Raghunadh, M.V.: Automated attendance management system based on face recognition algorithms. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–5, Enathi (2013). <https://doi.org/10.1109/iccic.2013.6724266>
9. Kurniawan, V., Wicaksana, A., Prasetyowati, M.L.: The implementation of eigenface algorithm for face recognition in attendance system. In: 2017 4th International Conference on New Media Studies (CONMEDIA), pp. 118–124, Yogyakarta (2017). <https://doi.org/10.1109/conmedia.2017.8266042>
10. Raghuwanshi, A., Swami, P.D.: An automated classroom attendance system using video based face recognition. In: 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), pp. 719–724, Bangalore (2017). <https://doi.org/10.1109/rteict.2017.8256691>
11. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: a generalpurpose face recognition library with mobile applications. CMU-CS-16-118, School of Computer Science, Carnegie Mellon University (2016)
12. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823, Boston, MA (2015). <https://doi.org/10.1109/cvpr.2015.7298682>
13. Merkel, D.: Docker: lightweight linux containers for consistent development and deployment. Linux J. (239) (2014)
14. Django (Version 2.0.1) [Computer Software] (2018). [https://django-project.com](https://.djangoproject.com)
15. Subramanyam, B., Joshi, P., Meena, M.K., Prakash, S.: Quality based classification of images for illumination invariant face recognition. In: 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), pp. 1–6, Sendai (2016). <https://doi.org/10.1109/isba.2016.7477245>

# Trajectory Classification Using Feature Selection by Genetic Algorithm



Rajkumar Saini, Pradeep Kumar, Partha Pratim Roy  
and Umapada Pal

**Abstract** Trajectory classification helps in understanding the behavior of objects being monitored. The raw trajectories may not yield satisfactory classification results. Therefore, features are extracted from raw trajectories to improve classification results. All the extracted features may not be useful for classification. Hence, an automatic selection scheme is essential to find optimal features from the pool of handcrafted features. This paper uses a genetic framework to choose the optimal set of features for trajectory classification purpose. Seven features costing 18 dimensions have been extracted from raw trajectories. Next, Genetic Algorithm (GA) has been used to find the optimal set of features from them. The binary encoding scheme has been used in GA. The 7-bit long chromosomes have been coded in this work. Bits of chromosomes represent trajectory features to be used in classification. Finally, trajectories have been classified using optimal features. Trajectory classification has been done using Random Forest (RF) based classifier and compared with Support Vector Machine (SVM). The results are evaluated using three trajectory datasets, namely I5, LabOmni2, and T15. The classification rates of 99.87%, 93.32%, and 90.58% have been recorded for datasets I5, LabOmni2, and T15, respectively.

**Keywords** Trajectory · Surveillance · Classification · Genetic algorithm · Random forest · Support vector machine

---

R. Saini (✉) · P. Kumar · P. P. Roy  
IIT Roorkee, Roorkee, India  
e-mail: [rajkumar.saini@ltu.se](mailto:rajkumar.saini@ltu.se); [rajkumarsaini.rs@gmail.com](mailto:rajkumarsaini.rs@gmail.com)

P. Kumar  
e-mail: [pradeep.iitr7@gmail.com](mailto:pradeep.iitr7@gmail.com)

P. P. Roy  
e-mail: [proy.fcs@iitr.ac.in](mailto:proy.fcs@iitr.ac.in)

U. Pal  
ISI Kolkata, Kolkata, India  
e-mail: [umapada@isical.ac.in](mailto:umapada@isical.ac.in)

## 1 Introduction

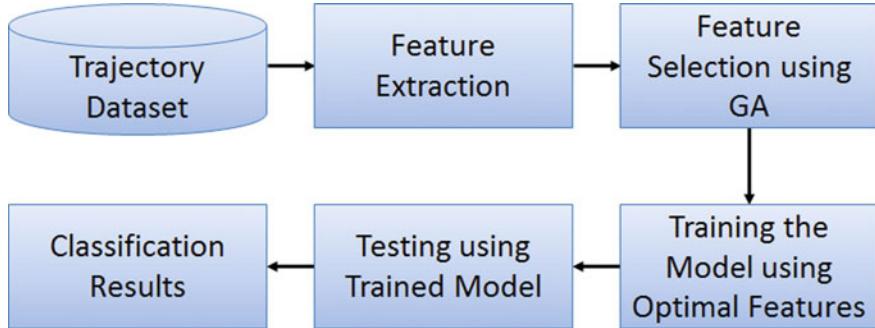
Trajectory learning enables us to understand the nature of motion patterns occurring in the area being monitored. Trajectories can be tracked using sensors such as GPS or video camera. Extracted trajectory patterns may be classified into groups based on how similar are they. The grouping of trajectories could differ and depends upon the objective of classification. It could be human behavior analysis, scene semantic analysis, anomalous activity detection, scene segmentation, etc.

The authors of [16] highlighted the issues that are faced in building automated vehicles. Authors surveyed the research works focusing on the role of humans in three different scenarios, i.e., inside the vehicle cabin, around the vehicle, and inside surrounding vehicles. Such a survey could help the research community toward trajectory modeling and classification for automatic analysis.

Lao et al.[13] have proposed a semantic analysis framework for analyzing human behavior using monocular camera video. They proposed a 3D reconstruction scheme to understand the human actions from different views. The technique was applied to classify the human posture and events occurring in the surveillance video. The technique was based on the usage of a single camera and could not handle the occlusions. A multi-camera based trajectory tracking has been proposed in [14]. It uses Social Force Model (SFM) to reidentify the movements in non-observed areas by analyzing the surveillance scene.

Authors in [9] proposed a trajectory analysis framework using Long Short-Term Memory–Recurrent neural Network (LSTM-RNN). The authors predicted the behavior of vehicles surrounding the ego-vehicle using cues extracted from their 3D trajectories and compared the performance on public trajectory datasets. A vehicle trajectory prediction framework based on Hidden Markov Model (HMM) [11] was proposed in [23]. The authors used double-layered architecture and Viterbi algorithm for trajectory prediction. Xun et al. [22] proposed a trajectory tracking and prediction framework for traffic videos. Authors considered the characteristics of traffic video, road structure profile, and trajectory movement patterns in their framework. They used a multi-target data correlation and particle filter for improved tracking and HMM was used for prediction.

However, in spite of very good advancement in object detection and tracking; trajectory classification could be poor without the use of proper features. The features could be selected manually by researchers or it could be made automatic using optimization techniques such as GA [6]. GA has been in a variety of applications such as parameter optimization [21], hybrid classification, drug design, image classification, ozone concentration forecast, etc. Wu et al. [21] proposed a hybrid GA kernel function and parameter optimization for Support Vector Regression (HGA-SVR). It tries to search for the kernel function and kernel parameters to improve the accuracy of SVR for electric load forecasting network. Mehmet et al. [6] proposed a hybrid technique based on GA, Bayesian, and k-NN focusing improved classification by eliminating data that make difficult to learn. The technique was tested on Iris, Breast



**Fig. 1** Description of the proposed work

Cancer, Glass, Yeast, and Wine out of which it improved the classification on the first four datasets.

In this paper, we exploit the potential of GA to select the optimal set of features extracted from the raw trajectories. We have also compared the performance of SVM- (Linear), SVM- (RBF), and RF-based classifiers in the context of trajectory classification [12]. The rest of the paper is organized as follows. In Sect. 2, we discuss the proposed framework. Feature extraction process is given in Sect. 2.1. Section 2.2 describes the details of how GA has been incorporated. In Sect. 2.3, description of classifiers is presented. In Sect. 3, experimental results obtained using publicly trajectory datasets, are presented. We conclude in Sect. 4 by highlighting some of the possible future work.

## 2 Proposed Work

The proposed framework efficiently uses GA to find the optimal set of features that have the highest trajectory classification rate. Figure 1 depicts the process flow of the proposed framework. Given a trajectory dataset, first, features are extracted using raw trajectories. Next, an optimal set of features are selected using GA. A model is trained using the optimal features and then testing is performed using the trained model. Finally, the test results are recorded.

### 2.1 Feature Extraction

Features have the capability to represent trajectories that can be distinguished easily as compared to raw ones. In this paper, we have used seven features out of which we are aimed to find the best set of features using GA that can classify trajectories with

higher classification rates. The extraction of features from raw trajectories is given below.

A trajectory can be defined as a finite sequence of  $(x^t, y^t)$  pairs followed by an object. The pair  $(x_t, y_t)$  represents the position of the object at time  $t$  in two-dimensional Euclidean space  $R^2$ . Trajectories may vary in length, i.e., the number of  $(x^t, y^t)$  pair may not be the same for all the trajectories. Given a set of such object trajectories, the following features have been extracted from raw trajectories.

**Trajectory Position (TP):** The trajectories can be extracted from videos or maybe recorded through sensors such as Global Positioning System (GPS). Such trajectories can be represented by the ordered sequence of  $(x, y)$  positions that objects follow. Formally, a trajectory  $T$  can be defined as given in (1).

$$T = \{(x^1, y^1), (x^2, y^2), \dots (x^t, y^t), \dots (x^{|T|}, y^{|T|})\} \quad (1)$$

**Velocity (V) and Acceleration (A):** Velocity gives very useful information that can distinguish among trajectories of different classes. Velocity of a trajectory over time can be computed using (2). Similarly, acceleration can be calculated using (2) over  $\Delta T$ .

$$\Delta T = \{(x^{t+1} - x^t, y^{t+1} - y^t)\} \quad \forall t = 1, 2, \dots (|T| - 1) \quad (2)$$

**Movement Direction (MD):** Movement direction [5, 8, 10, 19] has been calculated using every two alternate trajectory points as shown in Fig. 2a. Given points  $P^1(x^1, y^1)$  and  $P^3(x^3, y^3)$ , movement direction can be calculated using (3) and (4).

$$P(P^x, P^y) = P^3(x^3, y^3) - P^1(x^1, y^1), \quad (3)$$

$$\alpha = \cos^{-1}\left(\frac{P^x}{|P|}\right), \quad \beta = \cos^{-1}\left(\frac{P^y}{|P|}\right) \quad (4)$$

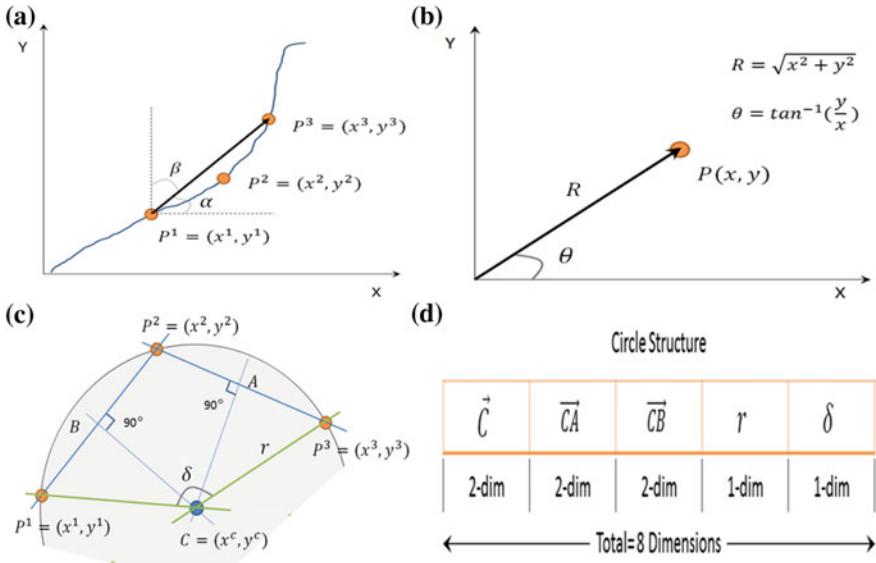
**Radial Distance (R) and Polar Angle ( $\theta$ ):**  $R$  and  $\theta$  work as polar coordinates. Given a trajectory point  $P(x, y)$ ,  $R$  and  $\theta$  can be computed using (5) as shown in Fig. 2b.

$$R = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}\left(\frac{y}{x}\right) \quad (5)$$

**Circle Structure (CS):** CS defines curvature [5, 8, 18] by fitting circle using three consecutive trajectory points. The points can also be selected after a fixed interval. Given such three points  $P^1$ ,  $P^2$  and  $P^3$ , a circle is drawn using them as shown in Fig. 2c. The center of the circle is calculated using (6).

$$\vec{C} = \frac{\sin 2P^1 \vec{P^1} + \sin 2P^2 \vec{P^2} + \sin 2P^3 \vec{P^3}}{\sin 2P^1 + \sin 2P^2 + \sin 2P^3} \quad (6)$$

Therefore, all the features contribute to 18-dimensional feature vectors in total.



**Fig. 2** Computation of **a** movement direction in terms of  $\alpha$  and  $\beta$  **b**  $R$  and  $\theta$  and **c** Circle Structure (CS). **d** Representation of CS

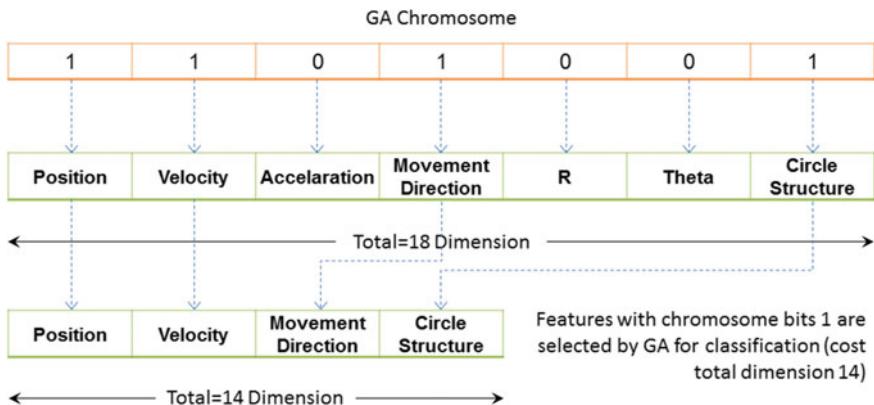
## 2.2 Genetic Framework

Genetic Algorithm (GA) [4, 6] is an evolutionary process that is inspired by the natural selection of chromosomes for mating them and generating new off-springs from parents. It involves three genetic operators, namely, Selection, Crossover, and Mutation. GA randomly initializes the possible set of solutions as initial population and uses these operators to generate new off-springs. All these operators have associated fitness criteria to work. During iterations, GA selects the chromosomes from the population to perform crossover. Two parent chromosomes produce children (possible solutions) after crossover. These are next evaluated against fitness function and if they fit the conditions they are added to the population for further testing. Mutation comes into play when no optimal solution is found by crossover. GA try to find the optimal solution by changing a few bits of chromosomes. The process is carried out until the convergence criteria are met. In our work, we have used GA to find the optimal set of features to classify surveillance trajectory datasets.

**Definition of Chromosomes:** Given trajectories with  $N$  features, we define chromosomes as binary strings of length  $N$ , i.e.,  $N$ -bit binary chromosomes composed of bits either 0 or 1. Features are selected for classification if the corresponding bits of a chromosome is 1. Features with chromosome bits 0 are rejected. Figure 3a shows the definition of chromosomes used in this work. 7-bit long chromosomes have been defined to select the features shown in Fig. 3b. Figure 4 shows an example of features

GA Chromosome						
0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	0 or 1
1-bit	1-bit	1-bit	1-bit	1-bit	1-bit	1-bit
Total=7 bits						
Trajectory Features						
Position	Velocity	Acceleration	Movement Direction	R	Theta	Circle Structure
2-dim	2-dim	2-dim	2-dim	1-dim	1-dim	8-dim
Total=18 Dimension						

**Fig. 3** Chromosome Definition: **a** 7-bit long chromosome for each feature in **b** store either 0 or 1. **b** Seven features with their dimensions having a total of 18 dimensions



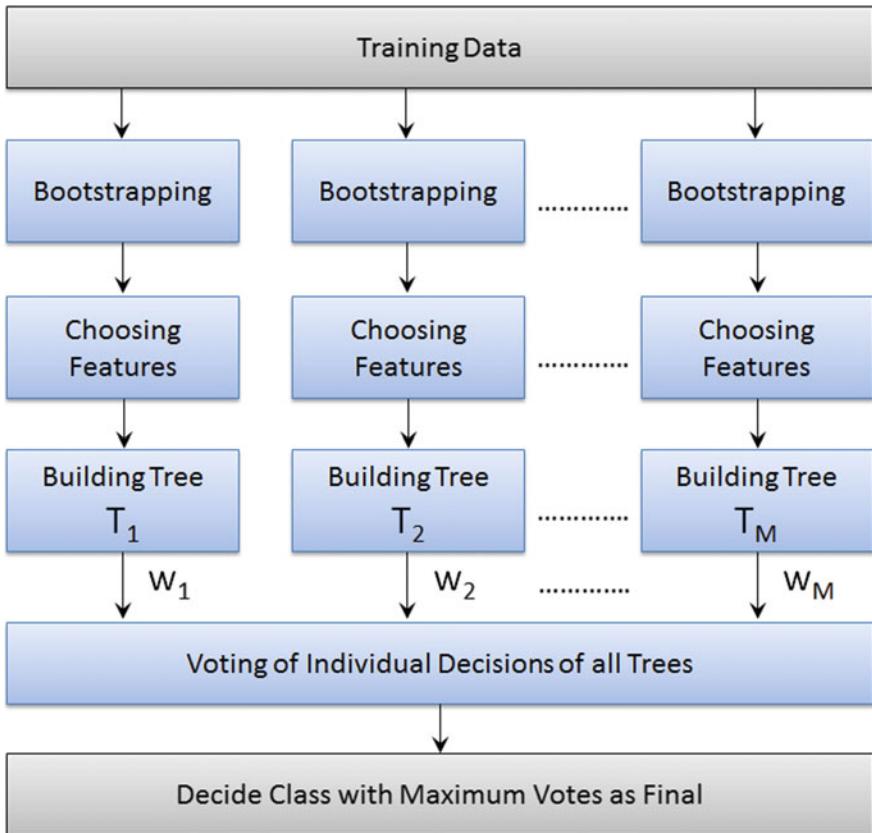
**Fig. 4** Example of feature selection: 4 features having corresponding chromosome bits 1 have been selected for trajectory classification

selected as chromosome bits are 1. It shows that four features corresponding to four 1s in the chromosome have been selected for classification.

The Accuracy has been used as the optimization function in the genetic framework. The classifiers RF and SVM have been tested using the features selected by the genetic algorithm.

### 2.3 Trajectory Classification Using RF and SVM

In this paper, we have compared the performances of two classifiers, namely RF and SVM using the features selected by GA. These classifiers are discussed below.



**Fig. 5** RF classification process

**Random Forest (RF):** RF [1, 3] is a supervised classifier that uses bootstrapping of Features into multiple training subsets. Next, it builds classification trees for each training subset. The final classification is made by collecting decisions from all the trees and choosing the final class having maximum votes. The voting can be done by assigning equal shares to the decisions of all trees or a weighting scheme can be adopted to assign unequal weights to the decisions of all trees. Figure 5 shows the procedure of the RF classifier to decide the final class through voting using decisions of  $M$  number of trees.

The selection of root nodes and splitting of features are done on the basis of information gain ( $I_G$ ) and entropy of features. The nodes are split only if there is a positive  $I_G$ . The  $I_G$  of splitting training data ( $S$ ) into subsets ( $S_j$ ) could be done using (7).

$$I_G = - \sum_j \frac{|S_j|}{|S|} E(S_j) \quad (7)$$

where  $|S_j|$  and  $|S|$  are the size of the sets  $S_j$  and  $S$ , respectively.  $E(S_j)$  is the entropy of set  $S_j$ .

**Support Vector Machine (SVM):** SVM [2] has been used by researchers to solve classification problems such as iris recognition [17], cancer classification [7]. It seeks to find boundaries (hyperplanes in case of Linear kernel) that can distinguish among data classes. It maps data points in space such that data points from different classes are as far as possible so that there is a clear gap among them. There are two most common kernel that have been widely used, namely, Linear and Radial Basis Function (RBF). Given  $l$  training samples-class pair as  $(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)$ , where  $x_i \in R^n$ ,  $i = 1, 2, \dots, l$  and  $y_i \in \{1, 2, \dots, k\}$ , where,  $y_i$  is the class  $x_i$  belong to.

A test sample is assigned a class associated with decision function having the maximum value. SVM uses kernels to reduce the search space of parameters. The Linear and RBF kernels are shown in (8) and (9), respectively.

$$K(x_p, x_q) = x_p^T x_q \quad (8)$$

$$K(x_p, x_q) = \exp(-\gamma ||x_p - x_q||^2), \gamma > 0 \quad (9)$$

where  $\gamma$  is a parameter in RBF kernel that helps in adjusting decision boundaries to reduce error.

### 3 Experimental Results

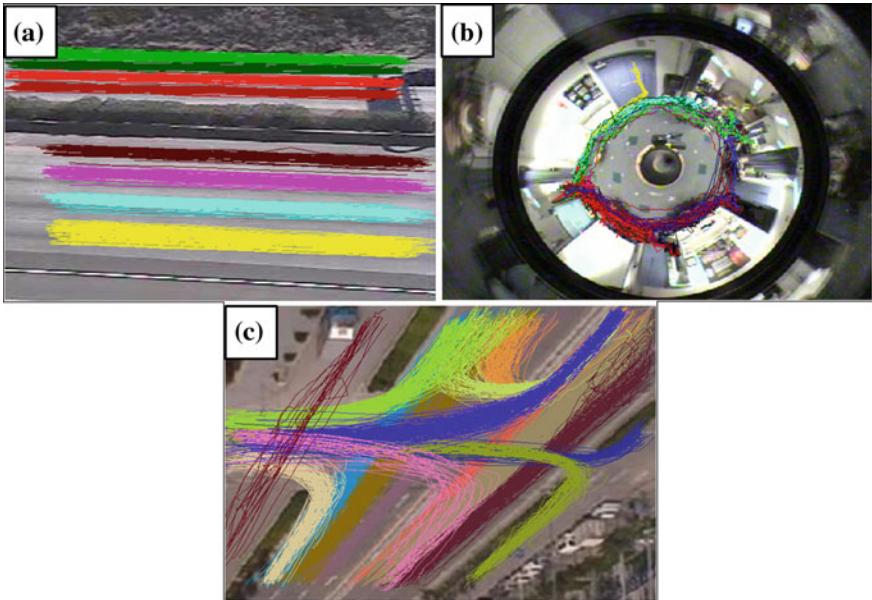
In this section, we present the results obtained using our algorithm applied on public datasets and present comparative performance evaluation done against benchmark methods of clustering.

#### 3.1 Datasets and Ground Truths

Through experiments, we have tested our proposed methodology applied on *I5* [15], *LabOmni2* [15] and *T15* [20] datasets. The datasets are discussed below.

**I5 Dataset:** *I5* dataset contains highway trajectories of vehicles in both direction of *I5* outside of UCSD. Trajectories are obtained by a simple visual tracker. Units are pixels. The trajectories belong to eight lanes. Hence, *I5* dataset has been labeled into eight classes. A total of 806 trajectories are there in the dataset. The number of trajectories in all the classes are not the same and varies from 75 to 137. *I5* dataset is shown in Fig. 6a.

**LabOmni2 Dataset:** *LabOmni2* dataset consists of human trajectories walking through a lab captured using an omnidirectional camera. The dataset consists of 209



**Fig. 6** Samples of trajectory dataset. **a** I5 **b** LabOmni2 **c** T15. Classes are printed in color

trajectories from 15 classes. The number of trajectories in a class varies from 3 to 36. Trajectory points are the pixels of captured video. The LabOmni2 dataset is shown in Fig. 6b.

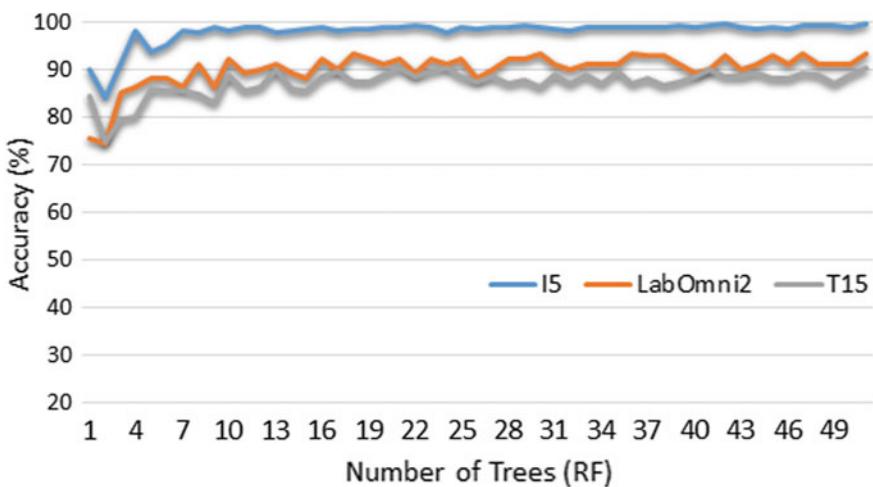
**T15 Dataset:** The trajectories of *T15* dataset [20] are labeled into 15 different classes. It consists of 1500 noisy trajectories that have been collected over a span of time. The size of each class is not fixed and it varies from 19 to 278. The *T15* dataset is shown in Fig. 6c.

### 3.2 Classification Results

Here, we present the results of the proposed classification framework using *I5*, *LabOmni2*, and *T15* datasets. Twofold cross-validation strategy has been used to test the performance of the framework. GA generated feature combinations has been tested using RF and SVM classifiers. Some of the features combinations selected by GA and their accuracies using the RF classifier has been shown in Table 1. The accuracy of 99.87% is recorded on the *I5* dataset using features {TP, V, A}. The accuracies of 93.32% and 90.58% are obtained using features {TP, V, R,  $\theta$ , CS} on *LabOmni2* and *T15* datasets, respectively. The last row of the table shows the performance when all features are considered in classification. It is clear from the table that classification rates are less with all features as compared to other feature com-

**Table 1** Accuracies on the datasets using features from GA and RF

Features	I5 (%)	LabOmni2 (%)	T15 (%)
TP, V, A	<b>99.87</b>	92.83	89.40
TP, V, A, MD	99.87	92.85	89.40
TP, V, MD	<b>99.87</b>	91.36	86.21
TP, V, MD, CS	99.50	90.40	86.21
TP, V, R, $\theta$ , CS	99.63	<b>93.32</b>	<b>90.58</b>
TP, V, A, R, $\theta$	99.75	92.83	89.40
TP, V, A, MD, R, $\theta$	99.75	92.36	87.65
TP, A, MD, R, $\theta$ , CS	99.63	91.90	90.26
TP, V, A, MD, R, $\theta$ , CS	99.63	92.83	90.26

**Fig. 7** Performance using RF by varying number of trees

binations. Thus, there is an improvement in the classification accuracies if optimal features are used.

The RF classifier has been tested using a number of trees varying from 1 to 50. The variation in the performance using the different number of trees is shown in Fig. 7. The accuracies of 99.87%, 93.32%, and 90.58% has been achieved at (36, 37, 42), (18, 30, 36, 47), and (21, 24) number of trees on *I5*, *LabOmni2*, and *T15* datasets, respectively.

The proposed GA framework has also been tested using the SVM classifier. Table 2 shows the performance of the SVM and RF classifiers. The performance of the SVM classifier has been tested using Linear and RBF Kernel (column 2 and 3). It could be noted from Table 2 that RF performs well over all the datasets as compared to the SVM classifier.

**Table 2** Comparison of performance on the datasets using SVM and RF

Dataset	SVM (Linear) (%)	SVM (RBF) (%)	RF (%)
I5	99.25	88.75	99.87
LabOmni2	86.14	85.19	93.32
T15	87.43	83.43	90.58

## 4 Conclusion

In this paper, a trajectory classification framework using GA and RF has been presented. Features have been extracted from raw trajectories and GA has been used to select an optimal set of optimal features and RF has been used for classification. Binary chromosomes of 7-bit length have been used in GA. The framework has been tested using three public datasets, namely *LabOmni2*, *I5* and *T15*. The best features and corresponding performance have been reported. The accuracies as high as 99.87, 93.32, and 90.58% has been recorded on *LabOmni2*, *I5*, and *T15* datasets, respectively. The performance of RF has been compared with SVM where RF outperforms SVM-based classification. We have used GA as it has the potential to find optimal features. Also, it converges fast as compared to grid search-based techniques. In future, more robust features and other optimization techniques could be used.

## References

1. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
2. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
3. Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., Dickhaus, H.: Automated sleep stage identification system based on time-frequency analysis of a single EEG channel and random forest classifier. *Comput. Methods Programs Biomed.* **108**(1), 10–19 (2012)
4. Ghamisi, P., Benediktsson, J.A.: Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *IEEE Geosci. Remote Sens. Lett.* **12**(2), 309–313 (2015)
5. Ghods, V., Kabir, E., Razzazi, F.: Decision fusion of horizontal and vertical trajectories for recognition of online farsi subwords. *Eng. Appl. Artif. Intell.* **26**(1), 544–550 (2013)
6. Goldberg, D.E., Deb, K., Kargupta, H., Harik, G.R.: Rapidaccurate optimization of difficult problems using fast messy genetic algorithms. *Int. Conf. Genet. Algorithms* **5**, 56–64 (1993)
7. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**(1), 389–422 (2002)
8. Jaeger, S., Manke, S., Waibel, A.: Online handwriting recognition: the NPen++ recognizer. *IJDAR* **3**, 169–180 (2001)
9. Khosroshahi, A., Ohn-Bar, E., Trivedi, M.M.: Surround vehicles trajectory analysis with recurrent neural networks. In: 19th IEEE International Conference on Intelligent Transportation Systems, pp. 2267–2272 (2016)
10. Kumar, P., Saini, R., Roy, P.P., Dogra, D.P.: Study of text segmentation and recognition using leap motion sensor. *IEEE Sens. J.* **17**(5), 1293–1301 (2017)

11. Kumar, P., Gauba, H., Roy, P.P., Dogra, D.P.: Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognit. Lett.* **86**, 1–8 (2017)
12. Kumar, P., Saini, R., Behera, S.K., Dogra, D.P., Roy, P.P.: Real-time recognition of sign language gestures and air-writing using leap motion. In: Fifteenth IAPR International Conference on Machine Vision Applications, pp. 157–160 (2017)
13. Lao, W., Han, J., De With, P.H.: Automatic video-based human motion analyzer for consumer surveillance system. *IEEE Trans. Consum. Electron.* **55**(2), 591–598 (2009)
14. Mazzon, R., Cavallaro, A.: Multi-camera tracking using a multi-goal social force model. *Neurocomputing* **100**, 41–50 (2013)
15. Morris, B., Trivedi, M.: Learning trajectory patterns by clustering: experimental studies and comparative evaluation. In: Computer Vision and Pattern Recognition, pp. 312–319 (2009)
16. Ohn-Bar, E., Trivedi, M.M.: Looking at humans in the age of self-driving and highly automated vehicles. *IEEE Trans. Intell. Veh.* **1**(1), 90–104 (2016)
17. Rai, H., Yadav, A.: Iris recognition using combined support vector machine and hamming distance approach. *Expert. Syst. Appl.* **41**(2), 588–593 (2014)
18. Tagougui, N., Kherallah, M., Alimi, A.M.: Online Arabic handwriting recognition: a survey. *IJDAR* **16**(3), 209–226 (2013)
19. Wang, D.H., Liu, C.L., Zhou, X.D.: An approach for real-time recognition of online Chinese handwritten sentences. *Pattern Recognit.* **45**(10), 3661–3675 (2012)
20. Weiming, H., Xi, L., Guodong, T., Maybank, S., Zhongfei, Z.: An incremental DPMM-based method for trajectory clustering, modeling, and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(5), 1051–1065 (2013)
21. Wu, C.H., Tzeng, G.H., Lin, R.H.: A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Syst. Appl.* **36**(3), 4725–4735 (2009)
22. Xun, L., Lei, H., Li, L., Liang, H.: A method of vehicle trajectory tracking and prediction based on traffic video. In: 2nd IEEE International Conference on Computer and Communications, pp. 449–453 (2016)
23. Ye, N., Zhang, Y., Wang, R., Malekian, R.: Vehicle trajectory prediction based on hidden Markov model. *KSII Trans. Internet Inf. Syst.* **10**(7) (2016)

# Action Recognition from Egocentric Videos Using Random Walks



**Abhimanyu Sahu, Rajit Bhattacharya, Pallabh Bhura  
and Ananda S. Chowdhury**

**Abstract** In recent years, action recognition from egocentric videos has emerged as an important research problem. Availability of several wearable camera devices at affordable costs has resulted in a huge amount of first- person/egocentric videos. Recognizing actions from this extensive unstructured data in the presence of camera motion becomes extremely difficult. Existing solutions to this problem are mostly supervised in nature, which require a large number of training samples. In sharp contrast, we propose a weakly supervised solution to this problem using random walk. Our solution requires only a few training samples (seeds). Overall, the proposed method consists of three major components, namely, feature extraction using PHOG (Pyramidal HOG) and a Center-Surround model, construction of a Video Similarity Graph (VSG), and execution of random walk on the VSG. Experimental results on five standard ADL egocentric video datasets clearly indicate the advantage of our solution.

**Keywords** Action recognition · Egocentric video · Center-surround model · Video similarity graph · Random walks

---

A. Sahu · R. Bhattacharya · P. Bhura · A. S. Chowdhury (✉)

Department of Electronics and Telecommunication Engineering, Jadavpur University,  
Kolkata 700032, India

e-mail: [as.chowdhury@jadavpuruniversity.in](mailto:as.chowdhury@jadavpuruniversity.in)

A. Sahu

e-mail: [abhimanyusahu009@gmail.com](mailto:abhimanyusahu009@gmail.com)

R. Bhattacharya

e-mail: [rajitbh@gmail.com](mailto:rajitbh@gmail.com)

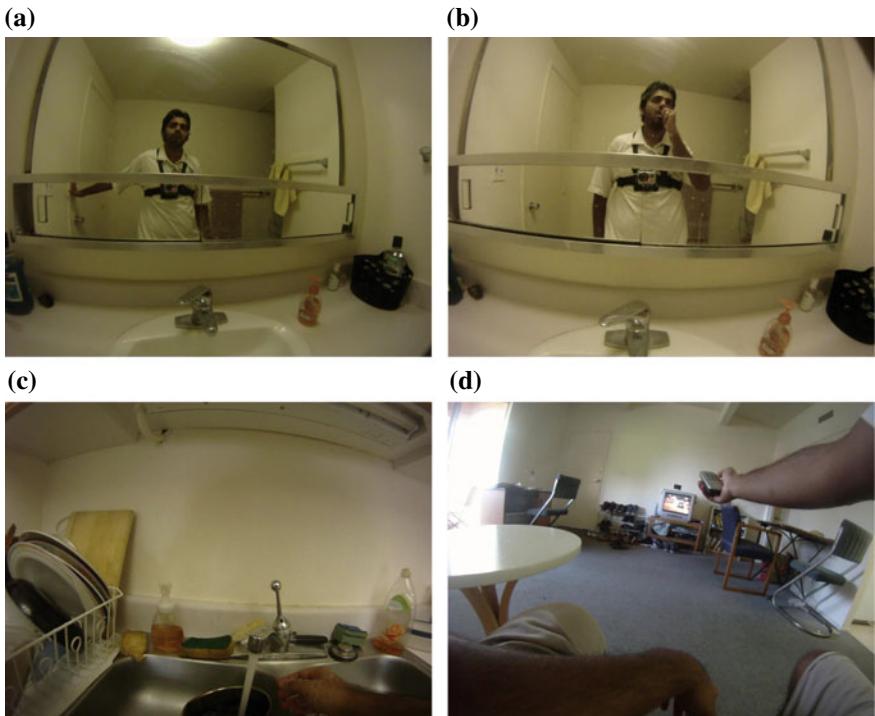
P. Bhura

e-mail: [pallabhbhura1997@gmail.com](mailto:pallabhbhura1997@gmail.com)

## 1 Introduction

Several lightweight wearable devices like GoPro, Google Glass, Autographer, Microsoft Sense-Cam [1] with high-performance processors and substantive memory are available nowadays at affordable costs. These cameras have enabled users to generate a large number of egocentric (first-person) videos, which can be employed to unravel new applications in diverse fields like social media, daily lives, and sports. Many such applications require recognizing actions such as running, walking, watching TV, using cell phones, and the likes where analysing visual information from the wearers camera becomes an important step [2]. In Fig. 1a, we show how a person has used “chest-mounted” wearable camera for recording an egocentric video in the ADL dataset [3]. In Fig. 1b, c, d, respectively show typical actions like *brushing teeth*, *washing dishes* and *watching tv*. Action recognition from egocentric video is extremely challenging because the camera fields of view are very different and also there is considerable camera motion and shakes.

In this work, we proposed a novel approach to address the existing issues of action recognition in egocentric videos [3]. In recent times, many approaches, both



**Fig. 1** **a** Chest-mounted wearable camera used for recording egocentric video and **b–d** typical actions in the constituent sample video frames

supervised or unsupervised [4, 5], have been employed in activity recognition from egocentric wearable cameras. Work on activity recognition, as reported in [6], discusses an object-centric approach to recognize activities in daily lives. Singh et al. [7] introduced a deep learning descriptor, similarly Ercolano et al. proposed [8] two deep learning architectures based on Long Short-Term Memory (LSTM) networks for first-person action recognition. However, unlike [3, 6, 8], here we proposed a random walk based approach that uses small training samples (seeds). Hence, our method can notionally be considered to be weakly supervised.

The main contribution of our proposed approach is classified into three fields. First, Pyramidal Histogram of Oriented Gradients (PHOG) features [9] and features derived from a Center-Surround model (CSM, [10]) are used to represent each frame. From the center-surround model, we obtain the differences in the entropy and optic flow values between the central and the surrounding regions of a frame. Second, we derive a video similarity graph (VSG) to capture the similarity between different video frames, modelled as the vertices of the graph. We then mark only 5% of the total vertices as seeds in the VSG with the help of available ground-truth. Finally, we obtain the labels of the remaining 95% unmarked vertices using random walks.

## 2 Related Work

Action recognition from egocentric videos has become an open research area for many researchers. We mention here some representative works on action and activity (composition of many short-term actions) recognition in egocentric video. Pirsiavash and Ramanan [3] presented an object-centric representation for recognizing daily activities from first-person camera views. Fathi et al. [11] analysed egocentric activities to ascertain different activities including hands, actions and objects. In another work, Fathi et al. discussed how different object models can be learned from egocentric video of household activities [12]. Some methods have also focused on recognizing daily activities by making use of a kitchen or cooking dataset [13, 14]. Lately, deep learning based approaches became popular in solving the action recognition problem. In, [15] Ma et al. presented a multi-stream deep architecture for joint action, activity and object recognition in egocentric videos. Ercolano et al. [8] used two deep learning architectures based on Long Short-Term Memory (LSTM) networks for first-person actions recognition. For a sparse coding-based solution to the action recognition problem, please see [16]. In [17], the authors employed stacked auto-encoders to solve the same problem. By and large, most of the existing solutions for recognizing actions/activities are supervised in nature, hence a large number of training samples are necessary to get better results.

In sharp contrast, our random walk based model captures more interesting actions with only a few training samples (seeds). So, our method can be designated as a weakly supervised approach. No random walk based solution has thus far been reported for this problem. The main contributions of our work are the following:

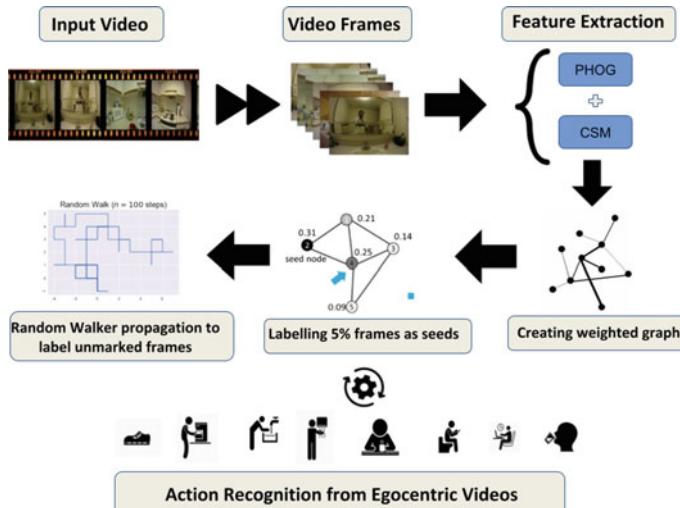
- (1) We propose a Random walker algorithm to recognize the important action from egocentric video.
- (2) We have also developed a center-surround model to better capture the characteristics of an egocentric video.

### 3 Proposed Framework

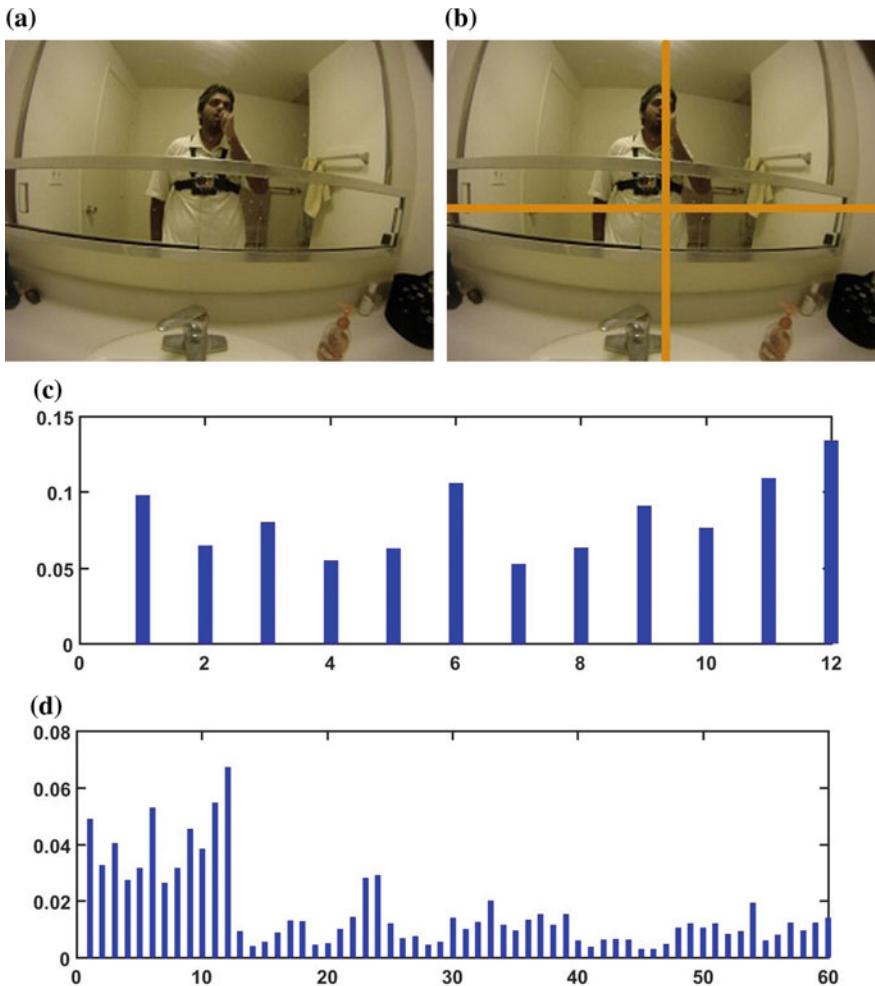
Our solution pipeline has three main steps, namely, (1) Feature extraction, (2) Construction of the VSG Graph and (3) Random walker algorithm. An overview of our proposed framework is described in the form of a block diagram in Fig. 2. A detailed description of each of the components is now provided below.

#### 3.1 PHOG Feature Extraction

In this work, we have used (PHOG) [9] for feature extraction. We choose PHOG because it represents the local shape and spatial information of the shape. We used two levels of pyramid for building the PHOG features. At level 0, the entire frame is considered as one single region and the histogram of edge orientation is calculated for that region. For level 1, the frame is partitioned into four cells as shown in



**Fig. 2** Our activity recognition framework



**Fig. 3** Extraction of PHOG features from a video frame: partition at different pyramid resolution for different level **a**  $L = 0$ , **b**  $L = 1$  and **c-d** Concatenation of all the HOG feature vectors of all sub-frame in two pyramid resolutions to obtain the PHOG features descriptor

Fig. 3b. Then HOG is determined for 5 ( $= (1 \text{ at level } 0 + 4 \text{ at level } 1)$ ) regions. We have also used  $\kappa = 12$  bins of histogram. Hence, the final PHOG descriptor of the entire video frame is a vector of size  $\kappa \times \sum_{l \in L} 4^l = 60$ , which is illustrated in Fig. 3. The number of levels and bins are chosen to balance accuracy and execution time.

### 3.2 Features from the Center-Surround Model

As the wearable camera is constantly moving and shaken, the same objects may disappear and re-appear in consecutive frames [18]. So, we have used center-surround model [10] to properly capture the motion state of a moving camera (ego-motion). This approach helps in properly recognizing important actions.

The center-surround model is used to capture the importance of a frame. A frame  $f$  of dimension  $W \times H$  is divided into a center region  $c$  of dimension  $aW \times bH$  and a surrounding region  $s$  of dimension  $(1-a)W \times (1-b)H$ . We separately find for each frame the optimal set of  $(a, b)$  for which the difference in entropy (as shown in Fig. 4b) and the difference in optical flow (as shown in Fig. 4c) between the center and the surrounding are maximized. The entropy of the center ( $E_c$ ) and that of the surrounding region ( $E_s$ ) are defined below:

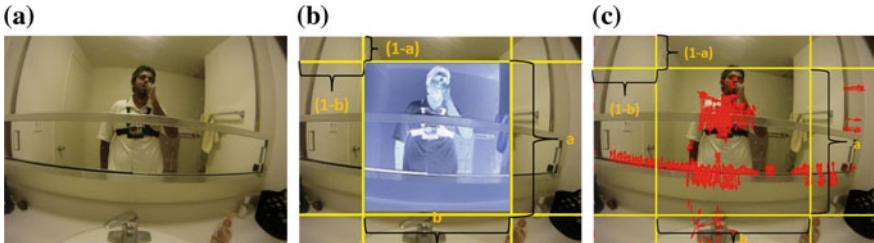
$$\begin{aligned} E_c &= - \sum_{k=1}^{m_c} p_{kc} \log_2(p_{kc}) \\ E_s &= - \sum_{k=1}^{m_s} p_{ks} \log_2(p_{ks}) \end{aligned} \quad (1)$$

where  $p_{kc}$  denotes the rate of occurrence of the  $k$ th grey level in  $c$ ,  $p_{ks}$  denotes the rate of occurrence of the  $k$ th grey level in  $s$ , and  $m_c$  and  $m_s$  respectively denote the total number of grey levels in  $c$  and  $s$ . To emphasize the absolute difference of the entropy values, we use an exponentiation and write:

$$\Delta E_{cs} = \exp(|E_c - E_s|) \quad (2)$$

The optimal  $(a, b)$  is obtained corresponding to the maximum value of  $\Delta E_{cs}$  (i.e.,  $(\Delta E_{cs}^*)$ ) in the following manner:

$$(a_E^*, b_E^*) = \underset{a, b \in (0, 1)}{\operatorname{argmax}} \{\Delta E_{cs}\} \quad (3)$$



**Fig. 4** Center-Surround Model: **a** Original Image, **b** High entropy values at the center, **c** High optical flow values at the center

Egocentric videos are often found to be unstable as they are typically captured with wearable cameras, which have constant motion. An egocentric video frame may contain fast or slow motion areas corresponding to fast or slow head turns. Motion blur can also occur when the camera wearer is moving on a vehicle. We use a weighted optical flow based on center-surround formulation to properly capture the motion (shown in Fig. 4c). Lucas\_Kanade method is applied for finding the flow vectors [19]. We give more emphasis (weight) to the motion of a pixel near the center. Let the coordinates of the center pixel in a frame be  $(x_c, y_c)$ . Euclidean distance of any pixel  $p$  with coordinates  $(x, y)$  from the center is  $d(x, y) = \sqrt{(x - x_c)^2 + (y - y_c)^2}$ . We define min-max normalization to normalize distance  $D_{norm}(x, y)$  as:

$$D_{norm}(x, y) = \frac{d(x, y) - \min(d(x, y))}{\max(d(x, y)) - \min(d(x, y))} \quad (4)$$

The distance based weight for any pixel  $w(x, y)$  is derived as follows:

$$w(x, y) = 1 - d_{norm}(x, y) \quad (5)$$

Let  $(u(x, y), v(x, y))$  be the flow vector for the pixel  $p$  with coordinates  $(x, y)$ . Then, the motion of the center ( $M_c$ ) and that of the surrounding region ( $M_s$ ) can be written as:

$$\begin{aligned} M_c &= \sum_{p(x,y) \in c} (\sqrt{u(x, y)^2 + v(x, y)^2})(w(x, y)) \\ M_s &= \sum_{p(x,y) \in s} (\sqrt{u(x, y)^2 + v(x, y)^2})(w(x, y)) \end{aligned} \quad (6)$$

So, the absolute difference of the motion values of the center and the surrounding region is given by:

$$\Delta M_{cs} = (|M_c - M_s|) \quad (7)$$

The above difference is already quite high ( $\sim 10^4$ ) and hence we did not use any exponentiation in this case. The optimal  $(a, b)$  is obtained corresponding to the maximum value of  $\Delta M_{cs}$  (i.e.,  $(\Delta M_{cs}^*)$ ) in the following manner:

$$(a_M^*, b_M^*) = \underset{a,b \in (0,1)}{\operatorname{argmax}} \{\Delta M_{cs}\} \quad (8)$$

Finally, we combine the 60-dimensional PHOG vectors with CSE (Center-Surround Entropy) and CSM (Center-Surround Motion) to represent each egocentric video frame. So, in our approach, each video frame is represented using a 62-dimensional vector (with 60-dimensional PHOG, 1-dimensional CSE, and 1-dimensional CSM).

### 3.3 Construction of the VSG Graph

A weighted video similarity graph  $VSG = (V, E)$  is built in the 62-dimensional feature space. Here, each frame  $f$  is a vertex/node of the VSG graph. Let us denote the total number of vertices to be  $n$ . The edge set is denoted by  $E = \{e_{mn}\}$ , where  $e_{mn}$  denotes the edge connecting the vertices  $v_m$  and  $v_n$ . The edge weight between two vertices is a measure of their similarity. We compute the edge weight  $w(e_{mn})$  between the vertices  $v_m$  and  $v_n$  in the following manner:

$$w(e_{mn}) = \exp(-d_{mn}^2/\sigma^2) \quad (9)$$

where  $d_{mn}$  is the Euclidean distance between the frames  $m$  and  $n$  and  $\sigma$  is a normalization parameter. As proposed in [20, 21],  $\sigma = \beta * \max(d)$ , where  $\beta \leq 0.2$  and  $d$  is the set of all pair-wise distances. If  $d_{mn} \leq \sigma$ , then the edge between them is preserved; else it is removed. In this manner, we scale down the size of the VSG graph.

### 3.4 Random Walk on VSG

We mark 5% of the vertices in VSG as seeds with one of the  $k$  action labels from the available ground-truth and denote the marked set as  $(V_M)$ . The remaining 95% unmarked vertex set is denoted by  $(V_N)$ . Using the random walker algorithm, we mark each  $v_n \in V_N$  with one of the  $k$  available action labels.

Let us assume a random walker to be at an unmarked vertex  $v_n \in V_N$ . Then, the probability with which he can move to a marked vertex  $v_m \in V_M$  is:  $p_{nm} = \frac{w(e_{nm})}{d_n}$ , where  $d_n$  is the degree of the vertex  $v_n$ . The goal is to obtain the probabilities of each random walker reaching first which (among  $k$ ) marked vertices. The label of that particular marked vertex, where a random walker reaches first, will be deemed as the label of that unmarked vertex, from which the random walker has started moving. Following [22], we cast the above problem as a discrete Dirichlet problem. The problem is solved using anisotropic interpolation on graphs [23]. The discrete graph Laplacian matrix is defined as:

$$L_{nm} = \begin{cases} d_n, & \text{if } n = m \\ -w(e_{nm}) & \text{if } v_n \text{ and } v_m \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

The above matrix  $L$  can be re-structured as:

$$L = \begin{bmatrix} L_M & B \\ B^T & L_N \end{bmatrix} \quad (11)$$

Here,  $L_M$ ,  $L_N$ , and  $B$  respectively represent the Laplacian matrices within marked vertices, within unmarked vertices, and between marked and unmarked vertices. Further, let  $x_n^s$  denote the probability of each cluster label  $s$ ,  $0 \leq s \leq k$  at any unmarked vertex  $v_n$ . We also define a labeled vector with length  $|V_M|$  for each cluster label  $s$  at a marked vertex  $v_m$  as:

$$m_j^s = \begin{cases} 1, & \text{if } \text{label}(v_j) = s \\ 0, & \text{if } \text{label}(v_j) \neq s \end{cases} \quad (12)$$

The solution to the system of equations  $L_N X = -BM$  obtained using conjugate gradient algorithm provides the potentials for all cluster labels [24]. The cluster labels for each  $v_n$  is deemed as the cluster label equivalent to  $\max(x_n^s)$ .

## 4 Experimental Results

In this work, experiments are implemented on a desktop PC with Intel Xeon(R) CPU E5-2690 v4 @ 2.60GHz, 16 Core and 128GB of DDR2-memory. The table illustrates the comparison of our results with one baseline method and three well-known existing approach [3, 6, 7].

### 4.1 Dataset

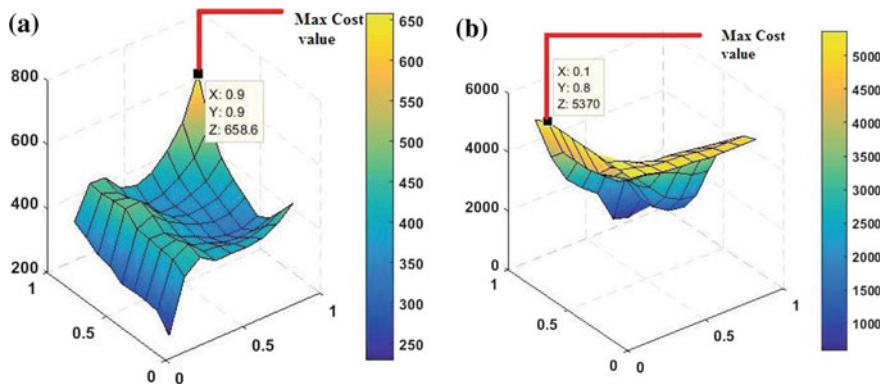
We use ADL dataset [3] for experimentation. It contains 20 videos with 10 different activities recorded in a lab kitchen with a counter facing camera (chest-mounted) involving 20 people. That dataset has been annotated with activity label and object label. Pirsavash and Ramanan [3] have used that dataset to perform a thorough investigation of state-of-the-art algorithms in both action and object recognition. All videos are in MPEG-1 form with each frame of dimension  $960 \times 1280$ . The video is captured at the rate of 30 fps and with  $170^\circ$  viewing angle of camera. The entire duration of the video is around 10h. We have experimented with 5 videos. Detailed information of these videos are given in Table 1.

### 4.2 Tuning of the Parameters

In this section, we discuss evaluation of the parameters  $a$  and  $b$  with the help of Fig. 5. In Fig. 5, x- and y-axes represent the variations in the parameters  $a$  and  $b$  respectively. In the z-axis we show how the center-surround Entropy ( $\Delta E_{cs}^*$ ) varies with  $a$  and  $b$ . The optimal values are those for which the difference in the center-surround entropy

**Table 1** Details of ADL dataset

Video no.	Duration (hr:mm:ss)	Frames	Action name
P_04	00:26:19	47328	Brushing teeth, washing hands/face, make up, using cell
P_09	00:21:28	38631	Drying hands/face, combing hair, adjusting thermostat
P_11	00:08:12	14775	Washing dishes, making coffee, grabbing water from tap
P_16	00:14:00	25197	Drinking water/bottle, drinking coffee/tea, watching tv
P_17	00:14:45	26547	Taking pills, using computer, laundry, vacuuming, writing

**Fig. 5** A frame from P\_11 video showing optimum  $a$  and  $b$  for maximizing differences in center-surround **a** entropy and **b** optical flow

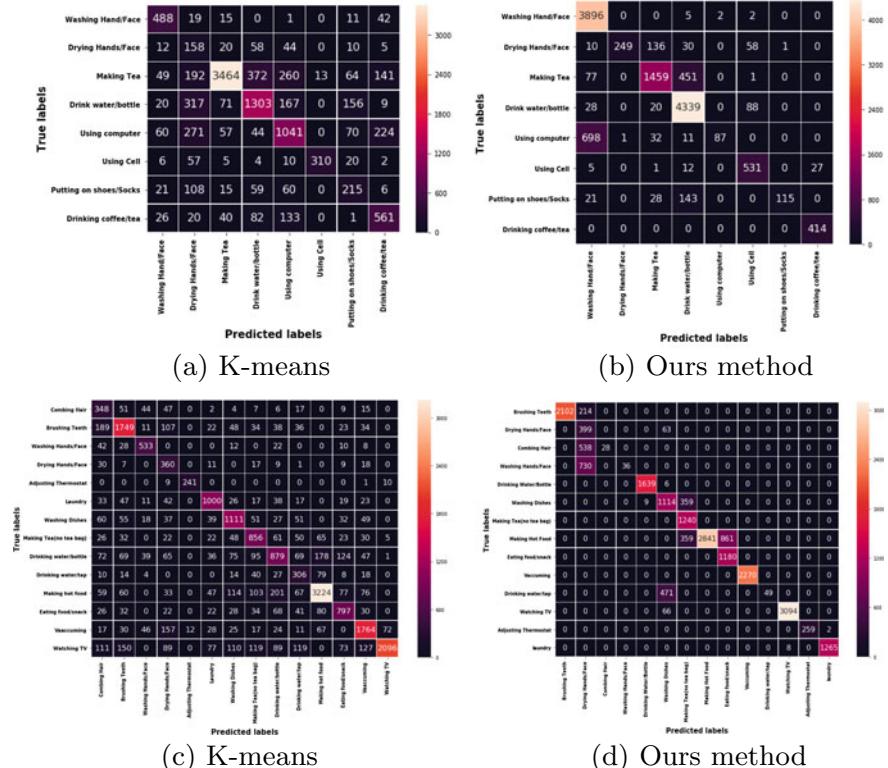
is maximized. From the plots, we determine  $a = 0.9$  and  $b = 0.9$  for a frame from P\_11 video in ADL dataset (Fig. 5a). Similarly, we determine,  $a = 0.1$  and  $b = 0.8$  for a frame from P\_11 video in ADL dataset (Fig. 5b) to maximize the difference in the center-surround motion ( $\Delta M_{cs}^*$ ).

### 4.3 Results on ADL Dataset

We compare our method with three recent approaches, namely, [3, 6, 7] and one baseline method (K-means clustering) on five ADL datasets. In the K-means based approach, we have clustered the frames with  $k$  action labels on the same three features (PHOG,  $\Delta C_{cs}$  and  $\Delta M_{cs}$ ). In Table 2, we show the mean accuracy values for different actions using the K-means and our proposed method. To show statistical significance of the betterment, we have carried out a t-test. We observe that the proposed approach

**Table 2** Mean accuracy values on five videos in ADL datasets

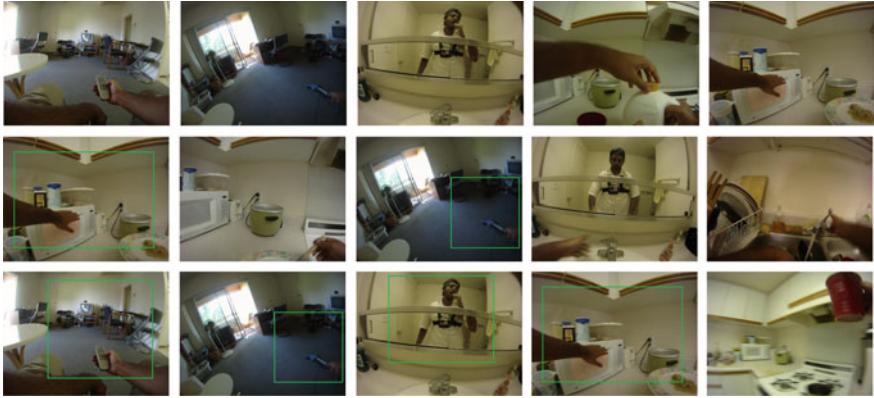
Video no.	K-means	Ours
P_04	0.5394	0.6026
P_09	0.6573	0.7492
P_11	0.6849	0.8545
P_16	0.7199	0.8261
P_17	0.6870	0.7645
Means	0.6637	<b>0.7538</b>



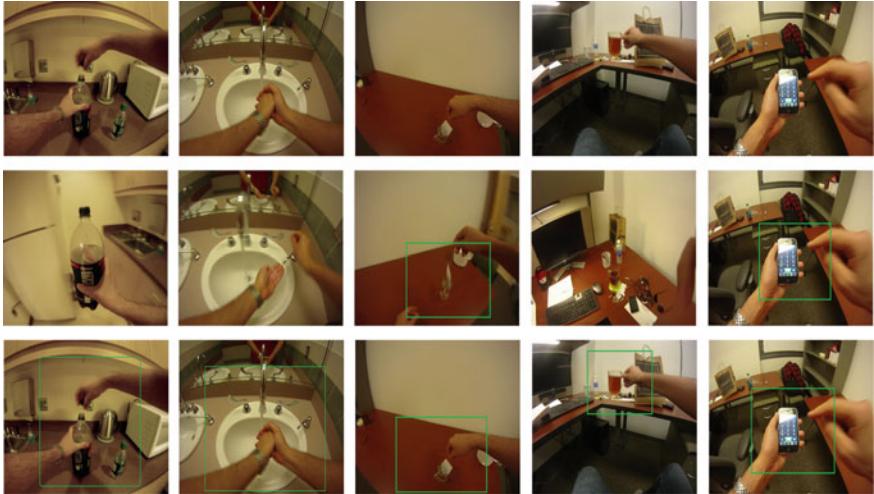
**Fig. 6** Confusion matrix for P\_11 (first row) and P\_16 (second row) videos ADL datasets

outperforms K-means in a statistically significant way (at  $p < 0.05$ ) with a p-value of 0.047455 at 95% confidence level.

We further show a confusion matrix for K-means and our approach in Fig. 6 for videos P\_11 and P\_16. In Fig. 7, we respectively show some important actions like “watching tv”, “vacuuming”, “brushing teeth”, and “making hot food” in P\_16 video recognized from the proposed method as well as the K-means algorithm. Similarly, in Fig. 8, we show important actions like “Drinking water”, “washing hand/face”,



**Fig. 7** Five important actions on P\_16 Video: First\_row: ground truth, Second row: K-means and Third row: Our proposed method. Important actions shown in green boxes. Our method captures four out of five such actions as compared to two out of five in K-means



**Fig. 8** Five important actions on P\_11 Video: First\_row: ground truth, Second row: K-means and Third row: Our proposed method. Important actions shown in green boxes. Our method captures five out of five such actions as compared to two out of five in K-means

“making tea”, and “drinking coffee” in P\_11 video. Both the figure clearly illustrates that we are able to detect more important actions based on matches with the ground-truth.

In Table 3, we compare the mean accuracy values of our solution with three state-of-the-art works. The results clearly reveal our approach having a mean accuracy value of 75.38% outperforms K-means, [6, 7] and looses only marginally to [3].

**Table 3** Comparison of mean accuracy values

CVPR 2012 [3]	BMVC 2013 [6]	CVPR 2016 [7]	K-means	Ours
77.00	38.70	37.58	66.37	75.38

## 5 Conclusion

A novel approach for action recognition from egocentric videos using random walks has been proposed in this work. Our solution requires only a handful of training samples in the form of seeds. Experimental comparisons indicate that the mean accuracy values obtained from the proposed approach are extremely competitive. In future, we plan to focus on extending the current framework to identify composite actions and activities.

## References

1. Bolanos, M., Dimiccoli, M., Radeva, P.: Toward storytelling from visual lifelogging: an overview. *IEEE Trans. Hum.-Mach. Syst.* **47**(1), 77–90 (2017)
2. Yan, Y., Ricci, E., Liu, G., Sebe, N.: Egocentric daily activity recognition via multitask clustering. *IEEE Trans. Image Process.* **24**(10), 2984–2995 (2015)
3. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR, pp. 2847–2854 (2012)
4. Koohzadi, M., Charkari, N.M.: Survey on deep learning methods in human action recognition. *IET Comput. Vis.* **11**(8), 623–632 (2017)
5. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **15**(3), 1192–1209 (2013)
6. McCandless, T., Grauman, K.: Object-centric spatio-temporal pyramids for egocentric activity recognition. In: BMVC (2013)
7. Singh, S., Arora, C., Jawahar, C.V.: First person action recognition using deep learned descriptors. In: CVPR, pp. 2620–2628 (2016)
8. Ercolano, G., Riccio, D., Rossi, S.: Two deep approaches for ADL recognition: a multi-scale LSTM and a CNN-LSTM with a 3D matrix skeleton representation. In: RO-MAN, pp. 877–882 (2017)
9. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR, pp. 401–408 (2007)
10. Sahu, A., Chowdhury, A.S.: Shot level egocentric video co-summarization. In: ICPR, Beijing, China (2018) (accepted)
11. Fathi, A., Farhadi, A., Rehg, J.M.: Understanding egocentric activities. In: ICCV, pp. 407–414 (2011)
12. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: CVPR, pp. 3281–3288 (2011)
13. Lei, J., Ren, X., Fox, D.: Fine-grained kitchen activity recognition using RGB-D. In: ACM Conference on Ubiquitous Computing, pp. 208–211 (2012)
14. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: CVPR, pp. 1194–1201 (2012)
15. Ma, M., Fan, H., Kitani, K.M.: Going deeper into first-person activity recognition. arXiv preprint [arXiv:1605.03688](https://arxiv.org/abs/1605.03688) (2016)

16. Alfaro, A., Mery, D., Soto, A.: Action recognition in video using sparse coding and relative features. In: CVPR, pp. 2688–2697 (2016)
17. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS, vol. 2011, no. 2, p. 5 (2011)
18. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR, pp. 1346–1353 (2012)
19. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. IJCA **12**(1), 43–77 (1994)
20. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 888–905 (2000)
21. Panda, R., Kuanar, S.K., Chowdhury, A.S.: Scalable video summarization using skeleton graph and random walk. In: ICPR, pp. 3481–3486 (2014)
22. Paragios, N., Chen, Y., Faugeras, O.D.: Handbook of Mathematical Models in Computer Vision. Springer Science and Business Media (2006)
23. Grady, L., Schwartz, E.: Anisotropic interpolation on graphs: the combinatorial Dirichlet problem. Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems (2003)
24. Grady, L.: Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1768–1783 (2006)

# A Bag of Constrained Visual Words Model for Image Representation



Anindita Mukherjee, Jaya Sil and Ananda S. Chowdhury

**Abstract** We propose a bag of constrained visual words model for image representation. Each image under this model is considered to be an aggregation of patches. SURF features are used to describe each patch. Two sets of constraints, namely, the must-link and the cannot-link, are developed for each patch in a completely unsupervised manner. The constraints are formulated using the distance information among different patches as well as statistical analysis of the entire patch data. All the patches from the image set under consideration are then quantized using the Linear-time-Constrained Vector Quantization Error (LCVQE), a fast yet accurate constrained k-means algorithm. The resulting clusters, which we term as constrained visual words, are then used to label the patches in the images. In this way, we model an image as a bag (histogram) of constrained visual words and then show its utility for image retrieval. Clustering as well as initial retrieval results on COIL-100 dataset indicate the merit of our approach.

**Keywords** Image representation · Constrained visual words · LCVQE · Image retrieval

---

A. Mukherjee  
Dream Institute of Technology, Kolkata, India  
e-mail: [anin1201@gmail.com](mailto:anin1201@gmail.com)

J. Sil  
IEST Sibpur, Howrah, India  
e-mail: [js@cs.iests.ac.in](mailto:js@cs.iests.ac.in)

A. S. Chowdhury (✉)  
Jadavpur University, Kolkata 700032, India  
e-mail: [as.chowdhury@jadavpuruniversity.in](mailto:as.chowdhury@jadavpuruniversity.in)

## 1 Introduction

Patch-based models of image representation play a significant role in diverse applications like image retrieval [1], image classification, and object recognition [2]. Bag of Visual Words (BoVW) has evolved as an image patch-based model [1]. The constituent patches in BoVW are first represented by elementary local features like SURF [3] or SIFT [4] and are then quantized by the K-means algorithm [5]. Finally, each image patch is assigned the label of the nearest cluster (visual words) and an image is represented by a bag (histogram) of visual words.

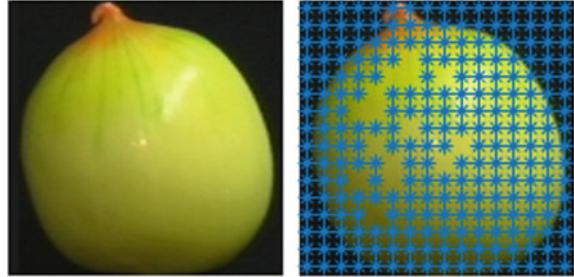
Some works are reported on improved BoVW models in connection with image retrieval. For example, in [6], Bouachir et al. applied fuzzy weighting. The authors in [7] used affinity based assignment to achieve a better model. Dimitrovski et al. have applied BoVW model along with predictive clustering tree for improving image retrieval [8]. In some works, random forest has been used in conjunction with BoVW to improve image retrieval. For example, see [9, 10].

In this paper, we propose a new patch-based image representation model by improving the traditional BoVW. We apply the Linear-time-Constrained Vector Quantization Error (LCVQE) [11], a fast yet accurate constrained k-means algorithm to quantize the image patches represented by SURF features along with *must-link* (ML) and *cannot-link* (CL) constraints. In particular, based on distance and statistical information, we present a new approach for deriving the constraints among image patches in a completely unsupervised manner. We term the resulting image representation model as Bag of Constrained Visual Words (BoCVW). The utility of this image representation model is shown in image retrieval. Experiments on COIL-100 [12] dataset demonstrate the advantage of our formulation. In [13], the authors have presented a spatially constrained BoVW model for hyperspectral image classification. However, in that paper, the authors have used constrained feature extraction rather than constrained clustering.

## 2 Proposed Model

We first briefly describe SURF feature extraction. We then show how the ML and CL constraints can be obtained. We then present the basics of Linear-time-Constrained Vector Quantization Error (LCVQE), a fast yet accurate constrained K-means algorithm. In the last section, we elaborate BoCVW model and its application in image retrieval.

**Fig. 1** SURF feature extraction: Locations for the strongest SURF features shown in blue crosses



## 2.1 SURF Feature Extraction

Here, we provide a basic description of SURF features for detecting interest points following [3]. SURF uses Hessian Matrix. The Hessian Matrix  $H(\mathbf{x}, \sigma)$  for any point  $\mathbf{x} = (x, y)$  in an image  $I$  at a scale  $\sigma$  is mathematically expressed as

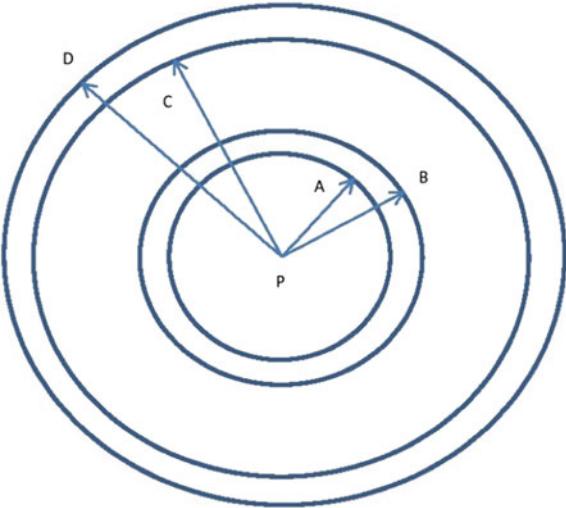
$$H(\mathbf{x}, \sigma) = \begin{bmatrix} L_{xx}(\mathbf{x}, \sigma) & L_{xy}(\mathbf{x}, \sigma) \\ L_{yx}(\mathbf{x}, \sigma) & L_{yy}(\mathbf{x}, \sigma) \end{bmatrix} \quad (1)$$

In Eq. (1),  $L_{xx}(\mathbf{x}, \sigma)$  marks the convolution of the Gaussian second- order derivative  $\frac{\delta^2}{\delta x^2}g(\sigma)$  with the image  $I$  at point  $\mathbf{x}$  and so on. In this work, we have adopted grid-based point selection for SURF feature extraction. Further, we have used two scales and kept 80% of the strongest SURF features from an image. Variance of the SURF descriptors is used in determining their strength. Choices for the number of scales, as well as the threshold for strength, are governed by a trade-off between accuracy and computational time. Please see Fig. 1 in this connection. Each image patch is described by a 64-dimensional SURF vector.

## 2.2 Formulation of the Constraints

We now describe the formulation of the ML and CL constraints [11, 14]. In most cases, the constraints are extracted with some prior domain knowledge. Note that the patch labels are not known. So, the ML and the CL constraints in our model are extracted in a completely unsupervised manner. If an image patch is deemed as a must-link constraint of another image patch, then these two patches should share the same cluster. Conversely, if two image patches are marked by a cannot-link constraint, then they cannot be put in the same cluster. In this work, we employ Cityblock distance over that of Euclidean to improve the overall computational time. The Cityblock distance between two patches  $P$  and  $Q$  is denoted as  $d_{CityBlock}(P, Q)$  and is given by

**Fig. 2** Extraction of ML and CL constraints for any patch  $P$ . A marks the nearest neighbor of  $P$  and must be linked to it. B is at a distance  $\bar{\lambda}$  from A and the hyperspherical shell AB marks the region for picking up additional patches that must be linked to  $P$ . D marks the farthest neighbor of  $P$  and cannot be linked to it. C is at a distance  $\bar{\lambda}$  from D and the annular shell CD marks the region for picking up additional patches which cannot be linked to  $P$



$$d_{CityBlock}(P, Q) = \sum_{i=1}^{64} ||P_i - Q_i|| \quad (2)$$

In Fig. 2, we schematically show how the ML and CL constraints are determined. The nearest neighbor of any patch is the patch with minimum distance from it. Naturally, the two patches must be linked. Similarly, the farthest neighbor of any patch is the patch at the maximum distance from it. Hence, these two patches cannot be linked. Let  $A$  and  $D$  be the nearest neighbor and farthest neighbor of patch  $P$  respectively. Further let  $ML(P)$  be the must-link constraint set of  $P$  and  $CL(P)$  be the cannot-link constraint set of  $P$ . So, we can write the following:

$$A = \arg \min_Q d_{CityBlock}(P, Q) \quad (3)$$

$$D = \arg \max_Q d_{CityBlock}(P, Q) \quad (4)$$

$$ML(P) = \{A\} \quad (5)$$

$$CL(P) = \{D\} \quad (6)$$

We now discuss a mechanism to augment the two sets  $ML(P)$  and  $CL(P)$ . In order to achieve that we need to construct a search space. Let  $\Lambda$  be the variance–covariance matrix of all the patches under consideration. Since each patch is a 64-dimensional vector,  $\Lambda$  has 64 eigenvalues. The largest eigenvalue represents variance magnitude in the direction of the largest spread of the patch data and likewise. Let  $\bar{\lambda}$  be the average of these 64 eigenvalues ( $\lambda_i$ ,  $i = 1 \dots 64$ ), i.e.,



**Fig. 3** Sample images in the COIL-100 database. Each image has considerable background portion (black in color) surrounding an object of interest

$$\bar{\lambda} = \left| \frac{\sum_{i=1}^{64} \lambda_i}{64} \right| \quad (7)$$

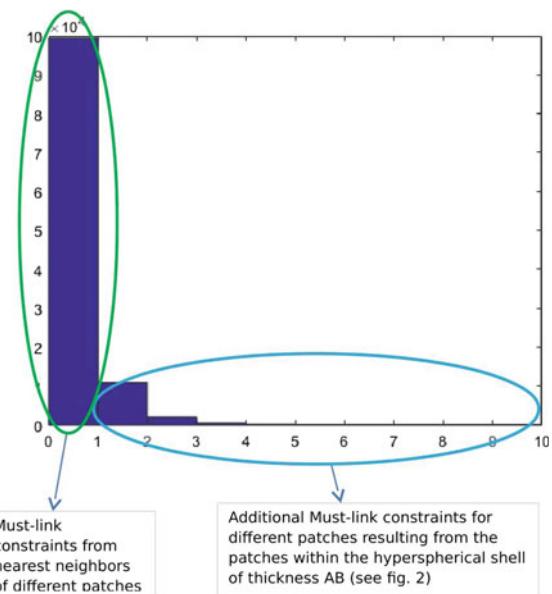
So,  $\bar{\lambda}$  indicates the average spread of the patch data. We use  $\bar{\lambda}$  to build a search space for obtaining additional ML and CL constraints. Four concentric hyperspheres of radius  $PA$ ,  $PB (= PA + \bar{\lambda})$ ,  $PC (= PD - \bar{\lambda})$  and  $PD$ , each centered at patch  $P$  are constructed in the 64-dimensional space (please see Fig. 2). If a patch  $Q_i$  falls within the hyperspherical shell  $AB$ , then it is a close neighbor of  $P$  and is deemed as a must-link constraint of  $P$ . Similarly, if a patch  $Q_i$  falls within the hyperspherical shell  $CD$ , then it is a distant neighbor of  $P$  and is hence deemed as a cannot-link constraint of  $P$ . So, we can write:

$$ML(P) = \{A\} \cup_i Q_i \text{ if } (d(P, A) \leq d(P, Q_i) \leq d(P, B)) \quad (8)$$

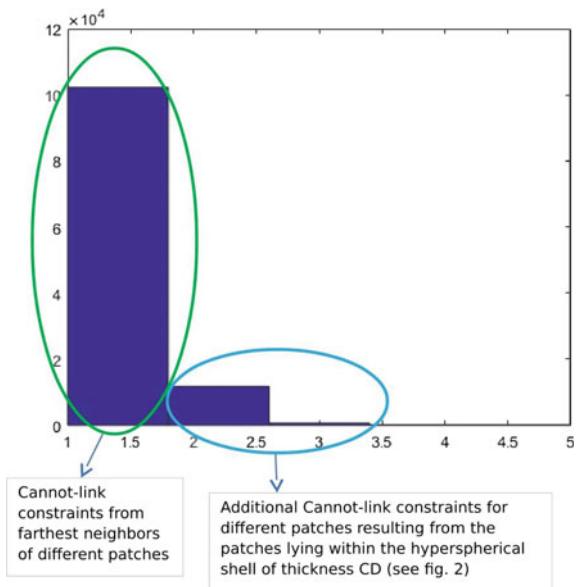
$$CL(P) = \{D\} \cup_i Q_i \text{ if } (d(P, C) \leq d(P, Q_i) \leq d(P, D)) \quad (9)$$

Here,  $d(P, A)$  denotes the Cityblock distance between patches  $P$  and  $A$  and so on. The patches, which lie within the hyperspherical shell of thickness  $BC$ , are neither must-linked nor cannot-linked to  $P$ . In the representation model, we focus more on objects in the image and not on the associated backgrounds. In an image, there can typically be patches from the background regions as well. For example, see the images in Fig. 3. For each such patch in a background, there can be many close background patches which may be must-linked to that patch. Likewise, for each such patch, there can be many distant object patches which may be cannot-linked to that patch. In order to restrict the number of ML and CL constraints, we decide to set a threshold  $\tau$ . In the experimental results section, we show how an optimal  $\tau$  has been set. We want to emphasize that the proposed framework for constraint evaluations is general in nature. The parameter  $\tau$  may not be necessary in a different dataset with images having very little background. We now show the distributions of ML and CL constraints in Figs. 4 and 5 respectively.

**Fig. 4** Histogram of ML constraints: All patches have at least one ML constraint in form of the nearest neighbor. However, some patches have more than one ML constraint



**Fig. 5** Histogram of CL constraints: All patches have at least one CL constraint in form of the farthest neighbor. However, some patches have more than one CL constraint



### 2.3 *Constrained K-Means Algorithm*

Visual vocabulary is constructed by quantizing the local descriptors using k-means algorithm. Constrained Vector Quantization Error(CVQE) algorithm [14] generalizes the k-means algorithm. In order to handle the must-link and cannot-link constraints, CVQE modifies the error function to penalize the violated constraints.

Linear-time-Constrained Vector Quantization Error (LCVQE) algorithm [11] minimizes a function with two components, namely, the vector quantization error and a penalty for violated constraints. The algorithm is very fast as it considers at most two clusters at a time. Violated must-link constraints update each centroid toward the opposite instance. For the violated cannot-link constraints, the patch that is farther from the cluster centroid is first obtained and the closest centroid to that patch is then brought toward it. For detailed mathematical formulations, please see [11].

### 2.4 *BoCVW Model and Its Application in Image Retrieval*

Each cluster, obtained from the LCVQE algorithm, forms a constrained visual word. We assign the label of the nearest cluster to individual image patches. So, each image is denoted by a histogram (bag) of constrained visual words (BoCVW). A  $K$ -dimensional BoCVW vector, where  $K$  is the number of clusters, represent an image. In the results section, we show the value of  $K$  is fixed.

We next apply the BoCVW model for the problem of image retrieval. All the training images are first represented by respective  $K$ -dimensional BoCVW vectors. Similarly, a query image is also represented by a  $K$ -dimensional BoCVW vector. During retrieval, we simply compute the Cityblock distance in the  $K$ -dimensional space between the test image and training images. The training images are then ranked based on the ascending order of these Cityblock distance values and retrieved accordingly.

## 3 Experimental Results

We first discuss the performance measures both for clustering as well as for retrieval. Then, detailed results on clustering are presented. We then show some results on image retrieval. Experiments on clustering as well as retrieval are shown on images from the well-known COIL-100 database. All the implementations are done in MATLAB. SURF feature extraction, LCVQE algorithm, and BoCVW-based retrieval runs in few seconds. Constraint extraction, which in turn needs distance computation between all patches, needs few hours to complete. All experiments are done on a desktop PC with Intel(R) Core(TM) i5-2400 @3.10GHz and 16 GB of DDR2-memory.

### 3.1 Performance Measures

Since we propose an improved clustering solution, we include two cluster validation measures, namely **Davies–Bouldin Index** [15] and **Calinski–Harabasz Index** [16] to demonstrate that the constrained k-means algorithm [5] with the proposed way of generating constraints is indeed better than k-means from purely a clustering perspective. The definitions of the above indices are now given below.

**Davies–Bouldin Index (DBI):** Davies–Bouldin Index uses average scatters ( $S_i, S_j$ ) within two clusters ( $C_i, C_j$ ) and the separation ( $M_{ij}$ ) between them to evaluate cluster quality. For  $K$  clusters, DBI is given by

$$DBI = \frac{\sum_{i=1}^K D_i}{K} \quad (10)$$

where the expression of  $D_i$  is as follows:

$$D_i = \max_{j \neq i} \frac{S_i + S_j}{M_{ij}} \quad (11)$$

For a good clustering, DBI value should be low.

**Calinski–Harabasz index (CHI):** Calinski–Harabasz index evaluates the cluster validity based on the average between and within cluster sum of square distances. CHI is defined as

$$CHI = \frac{\sum_i n_i d^2(c_i, c) / (K - 1)}{\sum_i \sum_{x \in C_i} d^2(x, c_i) / (n - K)} \quad (12)$$

where  $c_i$  center of the cluster  $C_i$ ,  $K$  is total the number of clusters, and  $n$  is the total number of data points (patches). For a good clustering, CHI value should be high.

For analyzing retrieval results for different competing methods, we have shown precision versus recall curves [17]. In addition, we also include recognition rates (mean precision without recall) [6].

### 3.2 Clustering Results

Here, we first show experimental results for the selection of best  $\tau$  (threshold for the number of ML and CL constraints) and best  $K$  (number of clusters). From Table 1, it is clear that the best clustering results for the LCVQE algorithm in the BoVW model are obtained with  $\tau = 0.001$  and  $K = 100$ . We now compare the clustering results with that of the k-means clustering using  $K = 100$ . Table 2 clearly shows that LCVQE algorithm yields better results over that k-means. This, in turn, makes BoCVW model of image representation a better one as compared to that of BoVW model. We finally show an ablation study where we show the results of LCVQE

**Table 1** DBI (the lower the better) and CHI (the higher the better) values for different combinations of  $\tau$  and  $K$ . The best combination is shown in bold

$\tau$	$K$	DBI	CHI
<b>0.001</b>	<b>100</b>	<b>1.6836</b>	<b><math>5.331 \times 10^3</math></b>
0.001	200	1.6941	$3.389 \times 10^3$
0.001	300	1.7177	$2.525 \times 10^3$
0.001	500	1.7345	$1.751 \times 10^3$
0.003	100	1.6852	$5.271 \times 10^3$
0.003	200	1.6992	$3.316 \times 10^3$
0.003	300	1.7528	$2.436 \times 10^3$
0.003	500	1.8070	$1.669 \times 10^3$
0.005	100	1.6839	$5.251 \times 10^3$
0.005	200	1.6900	$3.388 \times 10^3$
0.005	300	1.6737	$2.563 \times 10^3$
0.005	500	1.7490	$1.752 \times 10^3$

**Table 2** DBI (the lower the better) and CHI (the higher the better) values for constrained K-means and K-means with  $K = 100$ . The best results are shown in bold

Algorithm	DBI	CHI
LCVQE	<b>1.6836</b>	<b><math>5.331 \times 10^3</math></b>
K-Means [5]	1.6998	$3.312 \times 10^3$

**Table 3** DBI (the lower the better) and CHI (the higher the better) values for constrained k-means and k-means with  $K = 100$ . The best results are shown in bold

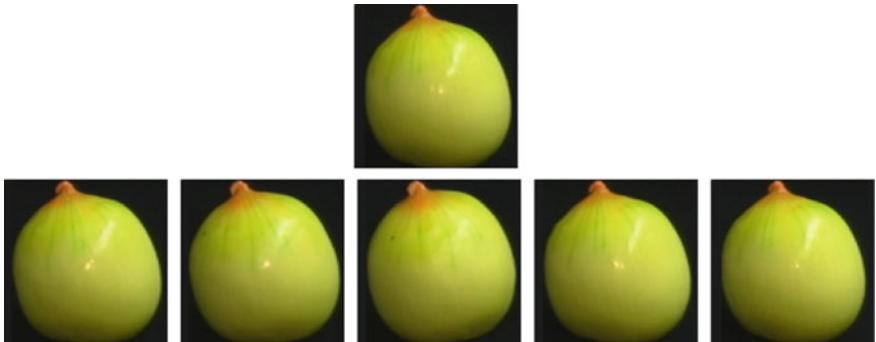
Algorithm	DBI	CHI
LCVQE	<b>1.6836</b>	<b><math>5.331 \times 10^3</math></b>
LCVQE1	<b>1.6836</b>	$5.276 \times 10^3$

clustering with all the ML and CL constraints (as shown in Eqs. (8) and (9)) versus LCVQE algorithm (denoted as LCVQE1) with only the nearest neighbor and the farthest neighbor as the respective must-link and cannot-link constraint (as shown in Eqs. (5) and (6)). Table 3 shows that though DBI values for LCVQE and LCVQE1 are same, CHI value for LCVQE is better. This, in turn, justifies the construction of the search space to capture additional ML and CL constraints.

### 3.3 Retrieval Results

We now show some initial results on image retrieval. For illustration, three query images, and the corresponding top five retrieved images using the proposed BoCVW model for Coil-2, Coil-5, and Coil-3 are presented in Figs. 6, 7 and 8 respectively. The results demonstrate that for Coil-2 and Coil-5, we have successfully retrieved all five images. For Coil-3, four out of five retrieved images are correct. The respective ROC curves for these three query images are presented next in Figs. 9, 10 and 11. Two of the three ROC curves (for Coil-2 and Coil-5) show excellent retrieval performance with AUC (area under the curve) values in both cases to be 1.0.

For retrieval, we compare our BoCVW method with four different approaches. The first approach used fuzzy weighting [6]. The second algorithm is based on term frequency and inverse document frequency (tfx) [2]. The third method is based on



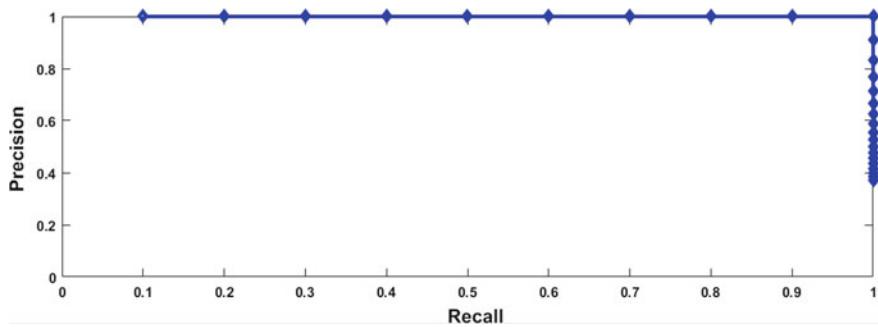
**Fig. 6** Results for Coil-2: Top row shows the query image. Bottom row shows the best five retrieved images



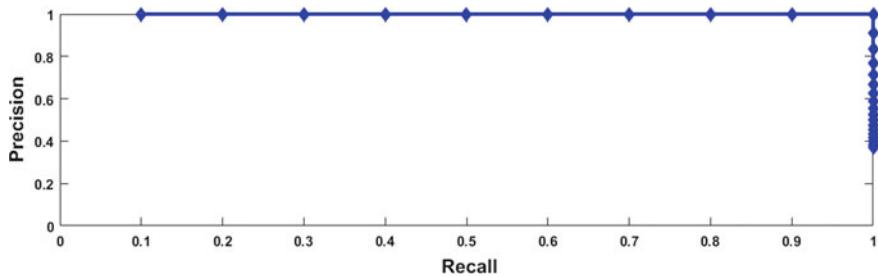
**Fig. 7** Results for Coil-5: Top row shows the query image. Bottom row shows the best five retrieved images



**Fig. 8** Results for Coil-3: Top row shows the query image. Bottom row shows the best five retrieved images

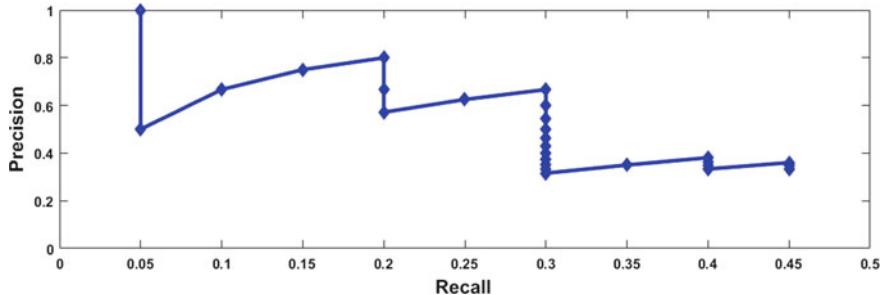


**Fig. 9** ROC curve for Coil-2



**Fig. 10** ROC curve for Coil-5

only term frequency (txx) [18]. The last approach used an affinity-based visual-visual word assignment model (*vwa*) [7]. The table shows that our method beats three out of the four competing methods and loses marginally to one method (Table 4).



**Fig. 11** ROC curve for Coil-3

**Table 4** Recognition Rates for five algorithms: txx [18], tfx [2], fuzzy weighting [6], vwa [7], and BoCVW

Image	txx	tfx	Fuzzy weighting	vwa	BoCVW
Coil 1	0.5	0.4	0.65	0.8	0.6
Coil 2	0.4	0.1	0.45	0.6	1.0
Coil 3	0.9	0.95	1.0	1.0	0.8
Coil 4	1.0	0.9	1.0	1.0	0.8
Coil 5	0.25	0.1	0.75	0.75	1.0
Average	0.715	0.615	0.8	0.86	0.84

## 4 Conclusion

We proposed Bag of Constrained Visual Words (BoCVW), a new model for patch-based image representation. We have also applied the model for the problem of image retrieval. Results on COIL-100 dataset show the benefits of our approach. In future, we will perform more comparisons. We also plan to use concepts from information theory and deep learning [19] in the BoCVW framework to achieve better performance.

## References

1. Sivic, J., Zisserman, A.: Video Google: efficient visual search of videos. In: Toward Category-Level Object Recognition, pp. 127–144 (2006)
2. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the ICCV, pp. 470–477 (2003)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
5. Hartigan, J.A., Wong, M.A.: A k-means clustering algorithm. Appl. Stat. **28**, 100–108 (1979)

6. Bouachir, W., Kardouchi, M., Belacel, N.: Improving bag of visual words image retrieval: a fuzzy weighting scheme for efficient indexation. In: Proceedings of the SITIS, pp. 215–220 (2009)
7. Mukherjee, A., Chakraborty, S., Sil, J., Chowdhury, A.S.: A novel visual word assignment model for content based image retrieval. In: Balasubramanian, R., et al. (eds.) Proceedings of the CVIP, Springer AISC, vol. 459, pp. 79–87 (2016)
8. Dimitrovski, I., Kocev, D., Loskovska, S., Dzeroski, S.: Improving bag-of-visual-words image retrieval with predictive clustering trees. *Inf. Sci.* **329**(2), 851–865 (2016)
9. Fu, H., Qiu, G.: Fast semantic image retrieval based on random forest. In: Proceedings of the ACM MM, pp. 909–912 (2012)
10. Mukherjee, A., Sil, J., Chowdhury, A.S.: Image retrieval using random forest based semantic similarity measures and SURF based visual words. In: Chaudhuri, B.B., et al. (eds.) Proceedings of the CVIP, Springer AISC, vol. 703, pp. 79–90 (2017)
11. Pelleg, D., Baras, D.: K-means with large and noisy constraint sets. In: Proceedings of the ECML, pp. 674–682 (2007)
12. Nene, S.A., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-100), Tech. Report, Department of Computer Science, Columbia University CUCS-006-96 (1996)
13. Zhang, X., et al.: Spatially constrained bag-of-visual-words for hyperspectral image classification. In: Proceedings of the IEEE IGARSS, pp. 501–504 (2016)
14. Davidson, I., Ravi, S.S.: Clustering with constraints: feasibility issues and the k-means algorithm. In: 5th SIAM Data Mining Conference (2005)
15. Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2**, 224–227 (1979)
16. Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat.* **3**(1), 1–27 (1974)
17. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: ideas, influences, and trends of the new age. *ACM Comput. Surv.* **40**(2), 1–60 (2008)
18. Newsam, S., Yang Y.: Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery. In: Proceedings of the ACM GIS, Article No. 9 (2007)
19. Wan, J., et al.: Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the ACM MM, pp. 157–166 (2014)

# Activity Recognition for Indoor Fall Detection in 360-Degree Videos Using Deep Learning Techniques



Dhiraj, Raunak Manekar, Sumeet Saurav, Somsukla Maiti, Sanjay Singh, Santanu Chaudhury, Neeraj, Ravi Kumar and Kamal Chaudhary

**Abstract** Human activity recognition (HAR) targets the methodologies to recognize the different actions from a sequence of observations. Vision-based activity recognition is among the most popular unobtrusive technique for activity recognition. Caring for the elderly who are living alone from a remote location is one of the biggest challenges of modern human society and is an area of active research. The usage of smart homes with an increasing number of cameras in our daily environment provides the platform to use that technology for activity recognition also. The omnidirectional cameras can be utilized for fall detection activity which minimizes the requirement of multiple cameras for fall detection in an indoor living scenario. Consequently, two vision-based solutions have been proposed: one using convolutional neural networks in 3D-mode and another using a hybrid approach by combining convolutional neural networks and long short-term memory networks using 360-degree videos for human fall detection. An omnidirectional video dataset has been generated by recording a set of activities performed by different people as no such 360-degree video dataset is available in the public domain for human activity recognition. Both, the models provide fall detection accuracy of more than 90% for omnidirectional videos and can be used for developing a fall detection system for indoor health care.

**Keywords** Activity recognition · Omnidirectional video · 3D convolutional networks · Long short-term memory networks · Fall activity · Daily activity

## 1 Introduction

Human fall is one of the major health risks in the modern lifestyle particularly for old people who are living alone, which may cause death in some situations if proper medication is not followed after the event. Along with this, it may result in post-fall

---

Dhiraj (✉) · R. Manekar · S. Saurav · S. Maiti · S. Singh · S. Chaudhury  
CSIR-Central Electronics Engineering Research Institute, Pilani, Pilani, India  
e-mail: [dhiraj@ceeri.res.in](mailto:dhiraj@ceeri.res.in)

Neeraj · R. Kumar · K. Chaudhary  
Samsung Research India, New Delhi, India

syndrome such as permanent immobilization, depression, etc., which further restricts the movement. As majority of deaths due to injury happen because of fall, so, early detection of fall is an important step so as to timely support the elderly by warning or informing their family members. The fall accidents cannot be completely prevented but fall detection system can save lives if it can identify a fall event and an alert can then be generated instantaneously.

In this paper, we focus on vision-based approaches for fall detection. The cameras have now become omnipresent as they provide very rich information about persons and their environment. The complete view of the living room is provided using a single omnidirectional camera installed in room ceilings with the downward-facing arrangement. So, vision-based fall detection systems prove to be reliable and play a major role in future healthcare and assistance systems. Due to a rapid rise in neural network architectures, deep learning techniques based on convolution operation have improved the obtained results in many techniques such as image classification, object detection, segmentation, and image captions. In this paper, we present two approaches, one uses a 3D Convolutional Neural Networks (3DCNN) for complete end-to-end learning in the form of feature extraction and classification and another uses a combination of CNN for feature extraction and LSTM for fall detection. More specifically, we propose architectures which detect a fall in 360-degree videos which take advantage of the capacity of CNNs to be sequentially trained on the generated 360-degree a dataset. First of all, we have generated an omnidirectional dataset in a lab environment to simulate indoor living scenario. It is then subsequently used to acquire the relevant features for activity recognition. As an outcome of this research carried out, this paper presents the following main contributions:

1. To the best of our knowledge, this is the first attempt that fall activity has been targeted on 360-degree videos using deep learning approach. In that respect, it is crucial to address this problem as no such 360-degree video database exists in the public domain which boosts up the activity recognition.
2. The preprocessing techniques such as Frame differencing (FD), Dense optical flow (OF), and normal RGB have also been tested as alternatives to Raw RGB as input video and are compared for their suitability for fall detection in 360-degree videos.
3. As fall detection is a sequential activity, techniques which can model the temporal relation between frames such as CNN and LSTM have been used and hyper tuned for their best utilization in this activity.

## 2 Related Work

The fall detection activity has been majorly targeted by three main approaches, namely ambient device-based, wearable device-based, and video-based. The ambient device-based approaches majorly use pressure sensors due to its cost-effective and less intrusive nature, but it has its inherent disadvantages in the form of false

alarms and volume quantity usage for better signal acquisition [1]. The wearable sensor devices rely on sensors such as accelerometers [2] or posture sensors [3] to detect the motion and location of a person but are very inconvenient for elderly fall detection due to its intrusive nature and burden of time for sensor placement. Video-based methods are nowadays the most popular detection strategies due to their unobtrusive nature, easy availability, and wide environment suitability. In those techniques, frames obtained from the video provides useful features to detect fault activity. In some techniques, such as Gaussian Mixture Model (GMM), Support Vector Machine (SVM) or Multilayer Perceptron (MLP), those features are used as input to the classifier to trigger the automatic detection of fall activity and can be further extended to tracking also as proposed by Lee and Mihailidis [4]. The silhouette information can also be used to detect positions and in combination with a matching system, the deformation in the shape of the human body can be utilized thereafter for activity detection [5]. Kwolek and Kepski [6] used the depth maps derived from Kinect in conjunction with the inertial measurement unit and classifier in the form of an SVM. Foroughi et al. [7] use human shape variation to detect a fall. Features extracted in the form of the best fit ellipse around the human body, histograms of silhouette, and temporal changes in head pose, are fed to a traditional classifier such as SVM [8] or an MLP ANN [9] for classification of motion and detection of fall action. People have targeted fall detection using combinations of time motion images and also using Eigenspace methods [10]. The bounding boxes of the objects have also been computed to find out if they contain a human and the fall event has been detected by means of features extracted from it [11].

Fall activity has also been targeted by using supervised learning methods in which features are extracted from raw data and using them as input to a classifier to learn class decision, for example, Charfi et al. [12] derived features from raw data and applied some transformations to them and used SVM as a classifier. Vision-based fall detection techniques also include using the 3D information about the environment in which fall event takes place by using depth cameras such as Microsoft Kinect or time-of-flight cameras. People such as Mastorakis and Makris [13] have adapted the 3D bounding box strategy to harness the 3D information. The major drawback of 3D-based approaches is in terms of the system deployment due to their narrow field-of-view and limited depth, and requirement of multiple synchronized cameras focused on the same area.

So, out of three feasible approaches, video-based activity recognition is most widely used by the researchers. However, instead of doing feature engineering ourselves, the deep learning-based approaches provide end-to-end learning in terms of feature extraction and classification to be handled on their own.

### 3 Database and Techniques

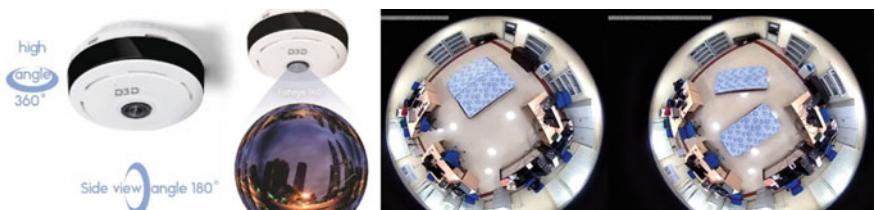
#### 3.1 Dataset

To facilitate the learning and testing of any model and classifier, we need a large amount of data. As fall event is multi-frame-based activity, a set of frames are required for learning by a model. Based on the current literature, it was found that out of the vision-based databases which are publically available, none of those belong to a 360-degree category. 360-degree fisheye cameras are a popular topic in the security industry today and offer a 360-degree view using a single lens, and give users the ability to digitally Pan, Tilt, and Zoom (PTZ) in the live video. The availability of the widest field-of-view without any blind spots makes them ideal for wide areas. Since 360-degree cameras have no moving parts, they avoid the lag (latency) associated with PTZ devices.

The database was recorded in a controlled environment which simulates indoor living as shown in Fig. 1.

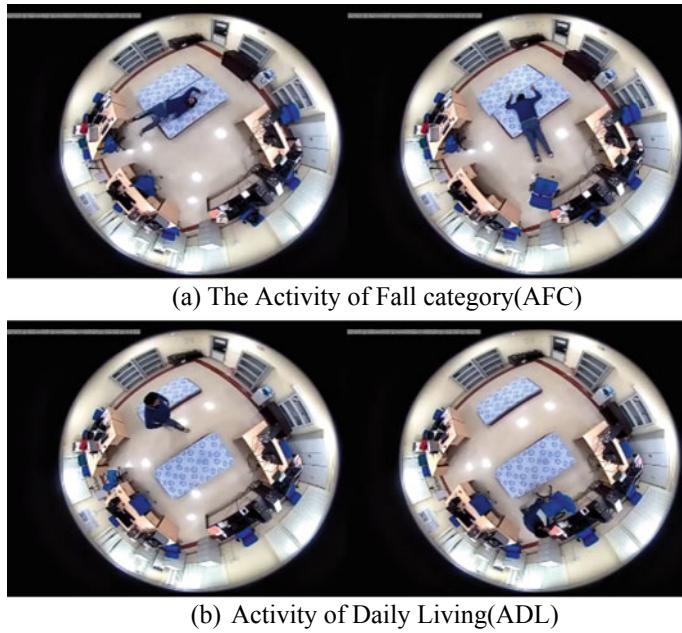
In addition, fisheye cameras tend to have a longer life, since all of their parts are stationary and do not wear out quickly. The specifications of the camera used in database preparation are Camera Model: D3D 360 Degree, Frame height: 960, Frame width: 1280, and Frame rate: 20 fps. The sample clips of the recording of AFC and ADL are shown in Fig. 2.

We prefer 360-degree video-based fall activity analysis due to its inherent advantages in terms of the very wide angle of view thus providing complete living space view possible to be framed using a single camera. The total number of video sequences thus generated is 2718 with 1327 sequences of Activity of Fall Category (AFC) and 1391 of Activity of Daily Living (ADL). From the videos captured during database preparation, it has been found that on average, a person takes around 3–5 s during the fall event. So it has been decided that video clips should be of 5-s duration for uniformity across all instances. All composite videos of each subject have been segmented in frames and then as per the Activity of Daily Living (ADL) and Activity of Fall category (AFC), instances clips were generated for 5 s duration.



(a) Camera installation    (b) Periphery data collection    (c) Centralized data collection

**Fig. 1** Database preparation environment



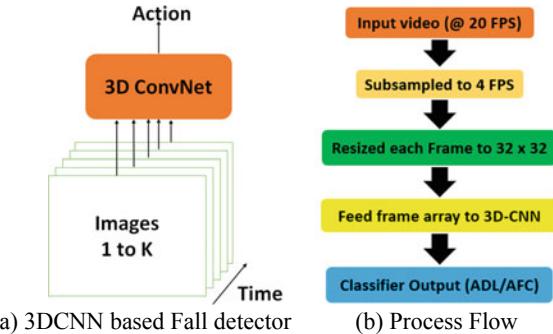
**Fig. 2** Human actions representing activity of fall category and activity of daily living

### 3.2 Techniques Adopted

The design of our fall detection architecture was driven by objectives such as minimizing the hand-engineered image processing steps and making the system generic and deployable on a low computational power machine. The first requirement was met by developing a system which works on human motion without correlating the image appearance such as the fall activity being represented by a stack of frames in which each frame represents a pose or an instance in the fall action. Another requirement of minimizing the hand-engineered image processing steps was met by using models which prove to be versatile in automatic feature extraction such as CNN [14] and LSTM [15].

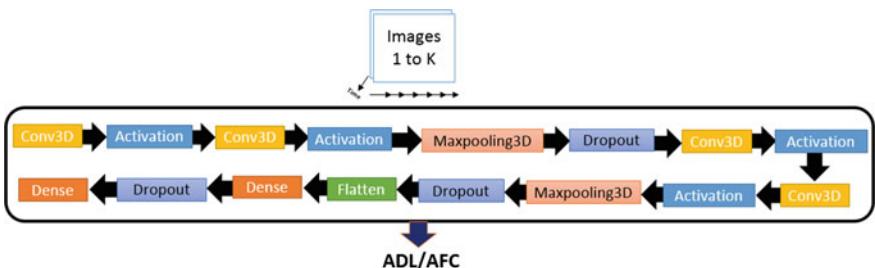
#### **3D Convolutional Neural Networks (3DCNN)**

The convolution-based neural networks have proved to be convenient tools to achieve generic features. The convolutional operation can be armed with temporal information such as 3D convolutional networks (3D ConvNet) [16, 17]. Since fall detection is a time-based event, time management is crucial and a way to cope with time and motion has to be added to CNNs. To incorporate the motion information in video analysis, 3D convolution is performed in the convolutional layers of CNN so that discriminative features along both the spatial and temporal dimensions are captured as shown in Fig. 3, where (a) shows the block-level details and (b) shows the process flow details.

**Fig. 3** 3D-ConvNet

The initial set of 100 frames in a 5 s duration will be down sampled to 20 frames which are resized to the desired input frame size of  $32 \times 32$ . This array of 20 frames now intermittently represents the complete action which was earlier represented in 5 s clips. The volume of frames is processed in a batch using a set of convolutional filters which extracts the features representing both spatial and timing information. The feature extraction is followed by activation using ReLU to introduce nonlinearity in the network. This feature extraction is repeated again followed by max pooling to reduce the computation requirements of the model. The drop out of 0.25 is used to ensure the prevention of overfitting of the model as the number of training data is not quite huge. The process is repeated again twice for feature merging followed by flattening of neuron connections and dense layers in the form of a fully connected layer. The details about the layers of the proposed 3DCNN architecture are shown in Fig. 4.

It consists of four convolution layers with characteristics such as kernel size = (3, 3, 3), padding = same, and activation = relu/softmax. The max pooling layer is having a pool size of (3, 3, 3) and dropout = 0.25. The output of the fourth convolution layer is further flattened to a 1D array. The final dense layer produces two class outputs: Fall or Not Fall. It can be strengthened by the fact that by applying multiple distinct convolutional operations at the same location on the input, multiple types of features can be extracted. The training methodology has to be adopted by keeping in mind the generality of the learned features for different fall scenarios. The hyperparameters

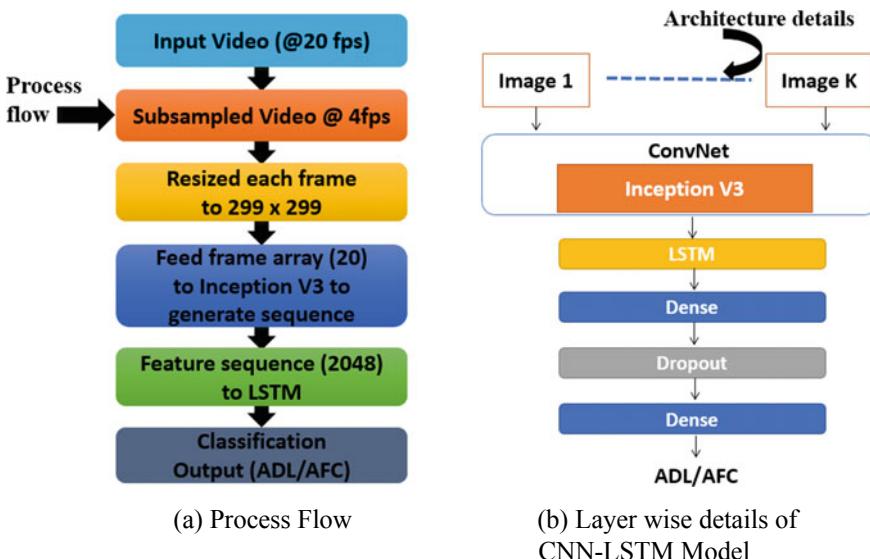
**Fig. 4** 3DCNN architecture: layer-wise

used in training are Learning rate:  $10e-4$ , loss function = categorical cross entropy and train-val-test split with 60-20-20 proportion for training, validation, and testing.

### **Long Short-Term Memory Networks (LSTM)**

CNN has demonstrated a great ability to produce a rich representation of input and achieving promising results on object detection. They are able to learn multiple layers of feature hierarchies by themselves which is termed as representational learning. Recurrent neural networks (RNNs) are a class of artificial neural networks, where the recurrent architecture allows the networks to demonstrate the dynamic temporal behavior. So, many researchers [18, 19] have combined RNN with CNN for action recognition with promising results. In particular, LSTM has been proved by Donahue et al. [20], Shi et al. [21] and Ng et al. [22] to be useful for activity recognition specifically from videos. LSTMs include a memory to model temporal dependencies in time series problems. Since the video is a sequence of frames, a model with RNN in the form of LSTM is established and after extracting the features in videos, all features  $X_t$  are fed as input in LSTMs. Experiments have been conducted with Feature lengths of 40 and 20 sequences but superior results are obtained with sequences of length 20. The block-level composition of CNN-LSTM is shown in Fig. 5, where (a) shows the process flow details and (b) shows the layer-wise composition of the proposed model.

This unified framework is able to capture time dependencies on features extracted by convolutional operators. The proposed model consists of two parts: convolution-based feature extraction and long short-term memory (LSTM)-based sequence mod-

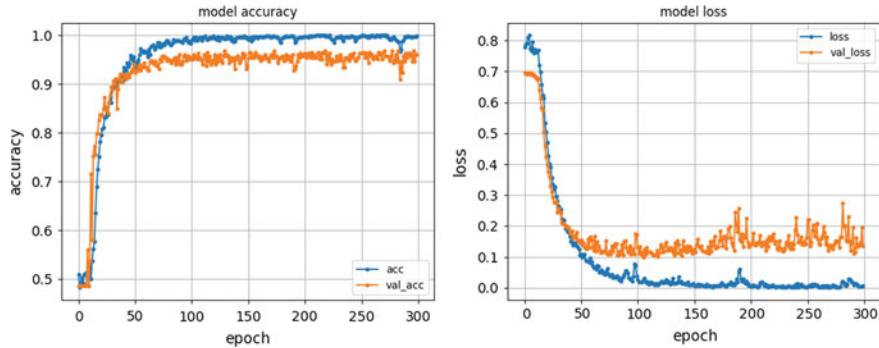


**Fig. 5** CNN-LSTM model for action recognition

eling. CNN provides a rich representation of the input image by embedding it into a fixed feature vector. For feature extraction, the Google Inception v3 network with pretrained weights of ImageNet is used to extract the key features called as input sequences from the training videos. Sequences of different frame lengths varying from 5 to 40 frames have been generated and tested for finding the best sequence length. With experiments, it was decided to use 20 frame sequences for training the LSTM network. The LSTM learns the temporal information in the provided sequences. The LSTM has been trained with 2048 features followed by a dense layer of size 512 with ReLU as the activation function. Further, Softmax classifier is employed to the layer output in two classes, i.e., Fall (ADL) or Not Fall (AFC). The forget parameter is skipped for LSTM and Adam is used as an optimizer along with category cross entropy as a loss function.

## 4 Results and Discussion

In this paper, deep learning-based models such as 3DCNN and CNN-LSTM have been proposed which can process the frame sequence-based video events for fall event detection. In 3DCNN, the model constructs feature set from both spatial and temporal dimensions by performing the 3D convolutions on a stack of frames. The developed deep architecture constructs multiple channels of information from neighboring frames and performs convolution and subsampling separately in each channel. The final feature representation is obtained by combining information from all channels. This feature set is used to train the model and subsequent classification using Softmax classifier. The hyperparameters of both the models in the form of the learning rate, number of epochs, batch size, and random state have been tried to find the best optimum values. Both models need input in the form of a stack of frames for their training. To find the effect of input format on model training and its convergence, different preprocessing techniques have been tried to find their effect on overall accuracy. It has been found that if input frames are processed in the form of their optical flow images, the model provides very low accuracy in the range of 50–55% due to lack of relevant information present after optical flow operation. The frame differencing is difficult to apply in this scenario as it requires the reference frame to be free from the subject which is not always possible as different scenarios require different arrangements in room structure. Overall, on using the processed frames using frame differencing, the model provides reasonable good accuracy in the range of 85–90%, but the best results have been obtained on using the Raw RGB frames with an accuracy of more than 95%.



**Fig. 6** 3DCNN model performance

#### 4.1 3DCNN Model Performance on 360-Degree Videos

The parameter details of the model are represented below

Input parameter: Input video Size:  $640 \times 480$ , FPS: 20, X shape:  $1804 \times 32 \times 32 \times 5 \times 1$ , Y shape:  $1804 \times 2$ , Random state = 46, 44.

Training parameter: Learning rate: 10e-4, Batch size: 128.

The plots representing the model accuracy and loss are shown in Fig. 6.

Here, acc denotes Model accuracy, Val\_acc denotes Model validation accuracy, loss denotes Model loss, and val\_loss denotes Model validation loss. The mean accuracy obtained with RGB videos is 95.3%. A dropout of 0.25 is used to prevent the overfitting of the model by randomly dropping the weights of the neurons and forcing it to learn the generality of the activity. The 10 fold cross validation approach is used to ensure that the model is not moving toward the overfitting. In general, the model will achieve a relatively stable accuracy of more than 90% in approximately 150 epochs but to ensure the reliability it is allowed to run for 300 epochs to report the validation accuracy of more than 95%. The model loss reports an all low of around 0.15.

The classification results of 3DCNN for unseen test cases are shown in Fig. 7 for the ADL and AFC category.

#### 4.2 CNN-LSTM Model Performance on 360-Degree Videos

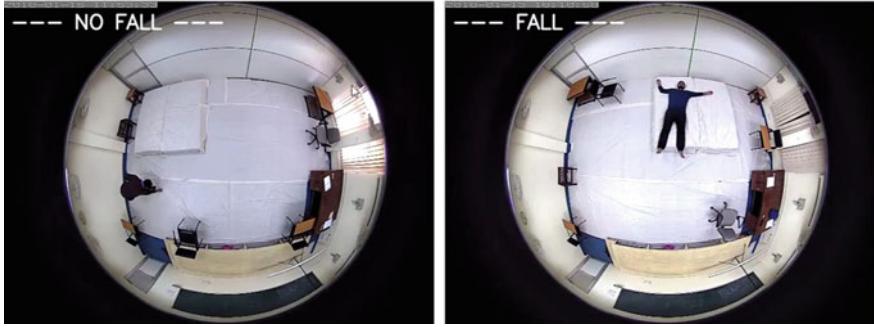
The parameter details of the model are represented below

Input parameters: Input video size:  $640 \times 480$ , FPS: 20

Feature extraction of size 20 by using Inception-v3

Input Shape  $756 \times 20 \times 2048$  for LSTM

Early stopper: Patience = 5

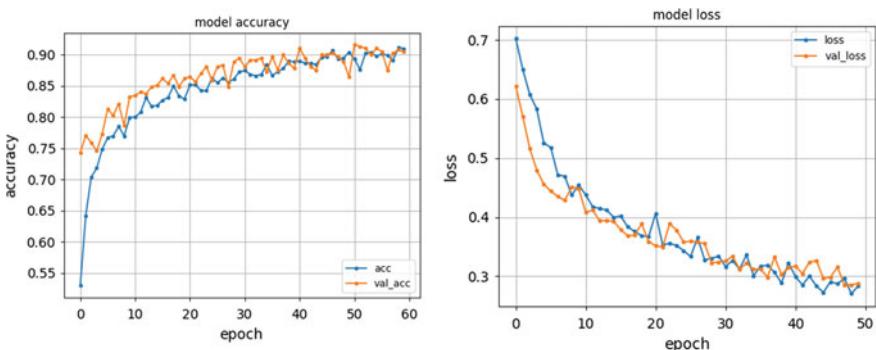


**Fig. 7** Activity classification performance of 3DCNN

Training parameter: Learning rate =  $10e-5$ , epoch = 60, Random state = 841.

The training model accuracy and loss are shown in Fig. 8. Here, acc denotes Model accuracy, Val\_acc denotes Model validation accuracy, loss denotes Model loss, and val\_loss denotes Model validation loss. The mean accuracy obtained with RGB videos is 91.62%. To prevent the overfitting of the model, the multifold cross validation approach has been used on the model by validating with 20% number of samples. It has been observed that the model achieves the accuracy of more than 90% in approximately 50 epochs and on further run the model performance does not improve considerably. The classification results of CNN-LSTM model for unseen test cases are shown in Fig. 9 for the ADL and AFC category.

The comparative analysis of the two approaches for activity classification is presented in Table 1.



**Fig. 8** CNN-LSTM model performance



**Fig. 9** Activity classification performance of CNN-LSTM

**Table 1** Performance comparison between 3DCNN and CNN-LSTM

Model	Depth	Learning rate	Number of epochs	Validation accuracy (%)
3DCNN	20	10e-4	300	95.3
CNN-LSTM	20	10e-5	60	91.62

## 5 Conclusion

In this paper, we presented a successful application of convolutional networks and long short-term memory networks to 360-degree videos for activity detection to create vision-based fall detector system which obtained state-of-the-art results on a benchmark database created in-house with 22 participants. To the best of our knowledge, this is the first effort to generate such an omnidirectional video database of fall activity. We have used indoor scenario illustrating activities in real life of elderly people. The CNN is applicable to a wide range of tasks because of their relatively uniform architecture. ConvNets are ideal models for hardware implementation and embedded applications. The ability for real-time processing can be acceptable. The proposed 3DCNN and LSTM models have been tested with different scenarios of fall events such as front fall, backward fall, side fall, fall during standing up or sitting on chair, imbalance resulting in fall and daily activities, events such as lying down, sitting on chair, picking up object, and squatting to evaluate its performance. The experiment results showed that the 3DCNN model trained and validated on both ADL and AFC activity videos provide 95.3% accuracy on the test set. The LSTM model also suits well to fall event detection with 91.62% accuracy for RGB omnidirectional video samples. Both models are not affected by the position of the person and can face any side, angle and can also walk around as opposed to the normal RGB camera where better accuracies are mainly claimed in situations when the camera faces the person from the front. Future research will be targeted to explore the utilization of unsupervised training of 3DCNN and LSTM models.

## References

1. Alwan, M., Rajendran, P.J., Kell, S., Mack, D., Dalal, S., Wolfe, M., Felder, R.: A smart and passive floor-vibration based fall detector for elderly. In: IEEE International Conference on Information & Communication Technologies (ICITA), pp. 1003–1007 (2006)
2. Estudillo-Valderrama, M.A., Roa, L.M., Reina-Tosina, J., Naranjo-Hernandez, D.: Design and implementation of a distributed fall detection system—personal server. *IEEE Trans. Inf Technol. Biomed.* **13**, 874–881 (2009)
3. Kang, J.M., Yoo, T., Kim, H.C.: A wrist-worn integrated health monitoring instrument with a tele-reporting device for telemedicine and telecare. *IEEE Trans. Instrum. Meas.* **55**, 1655–1661 (2006)
4. Lee, T., Mihailidis, A.: An intelligent emergency response system: preliminary development and testing of automated fall detection. *J. Telemed. Telecare* **11**(4), 194–198 (2005)
5. Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J.: Robust video surveillance for fall detection based on human shape deformation. *IEEE Trans. Circuits Syst. Video Technol.* **21**(5), 611–622 (2011)
6. Kwolek, B., Kepski, M.: Human fall detection on the embedded platform using depth maps and wireless accelerometer. *Comput. Methods Programs Biomed.* **117**(3), 489–501 (2014)
7. Foroughi, H., Rezvanian A., Paziraei A.: Robust fall detection using human shape a multi-class support vector machine. In: IEEE 6th Indian conference on Computer Vision, Graphics & Image Processing (ICVGIP), pp. 413–420 (2008)
8. Foroughi, H., Yazdi, H.S., Pourreza, H., Javidi, M.: An eigenspace-based approach for human fall detection using integrated time motion image and multi-class support vector machine. In: IEEE 4th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 83–90 (2008)
9. Foroughi, H., Aski, B.S., Pourreza, H.: Intelligent video surveillance for monitoring fall detection of elderly in home environments. In: IEEE 11th International Conference on Computer and Information Technology (ICCIT), pp. 24–27 (2008)
10. Foroughi, H., Naseri, A., Saberi, A., Yazdi, H.S.: An eigenspace-based approach for human fall detection using integrated time motion image and neural network. In: IEEE 9th International Conference on Signal Processing (ICSP), pp. 1499–1503 (2008)
11. Miaou, S.-G., Sung, P.-H., Huang, C.-Y.: A customized human fall detection system using omni-camera images and personal information. In: Proceedings of the 1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare, D2H2 2006, pp. 39–42, USA, April (2006)
12. Charfi, I., Miteran, J., Dubois, J., Atri, M., Tourki, R.: Definition and performance evaluation of a robust SVM based fall detection solution. In: Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012, pp. 218–224, Italy, November 2012
13. Mastorakis, G., Makris, D.: Fall detection system using Kinect’s infrared sensor. *J. Real-Time Image Proc.* **9**(4), 635–646 (2012)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
15. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.M.: Video LSTM convolves, attends and flows for action recognition. *Comput. Vis. Image Underst.* **166**, 41–50 (2018)
16. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013)
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei, L.F.: Large-scale video classification with convolutional neural networks. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1725–1732, Columbus, OH, USA, 23–28 June 2014
18. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention (2015). [arXiv:1511.04119](https://arxiv.org/abs/1511.04119)

19. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep Bi-directional LSTM with CNN features. *IEEE Access.* **6**, 1155–1166 (2018)
20. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2625–2634, Boston, MA, USA, 8–10 June 2015
21. Shi, Y., Tian, Y., Wang, Y., Huang, T.: Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Trans. Multimedia* **19**, 1510–1520 (2017)
22. Ng, J.Y., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4694–4702, Boston, MA, USA, 8–10 June 2015

# A Robust Watermarking Scheme for Copyright Protection



Satendra Pal Singh and Gaurav Bhatnagar

**Abstract** In this paper, a robust image watermarking scheme based on all phase sine biorthogonal transform (APSBT), singular value decomposition (SVD) and dynamic stochastic resonance (DSR) is presented. Firstly, the cover image is transformed by APSBT and then a gray-scale logo is embedded through the singular value decomposition. After the authentication process which essentially resolves the false-positive extraction of SVD in watermarking, a phenomenon based on dynamic stochastic resonance is deployed for the logo extraction from the watermarked image. The simulation results demonstrate that the proposed scheme has better performance in the aspect of robustness and invisibility.

**Keywords** Image watermarking · Singular value decomposition · Dynamic stochastic resonance (DSR) · All phase sine biorthogonal transform (APSBT)

## 1 Introduction

In digital era, advance development in wireless and communication technologies increases the role of social networks in everyday lives, where sharing of multi-media data such as images, audios and videos become a widespread practice due to wide availability of digital devices like smart-phone and digital cameras. However, sharing of the digital images may expose to some serious threat such as duplication, modification and illegal manipulation. Therefore, the protection of intellectual property right (IPR) is one of major challenge in information security [1]. Digital watermarking [2, 3] is most commonly adopted method, which effectively addresses these issue by protecting ownership of digital data. The digital watermarking process consists of two phase namely watermark embedding and extraction. In watermark embedding, an imperceptible signal called watermark representing the legitimate ownership is embedded into the host signal in a way such that the quality of water-

---

S. P. Singh (✉) · G. Bhatnagar

Department of Mathematics, Indian Institute of Technology Karwar, Karwar, India  
e-mail: [pg201383504@iitj.ac.in](mailto:pg201383504@iitj.ac.in)

marked signal has no perceptual degradation, while in watermark extraction, the presence of watermark identifies the legitimate owner of the original signal.

In last two decades, a number of invisible watermarking schemes have been proposed in the literature. However, plenty of watermarking schemes fail to resist against strong noise and geometric attacks [4]. Therefore, to address these issue a DSR based extraction methodology is employed in this paper to increase the robustness. Generally, digital watermarking algorithm are classified into spatial domain [5] and frequency domain [6–8] techniques. In frequency domain techniques, a watermark is embedded into the coefficients generated using the transformation such as discrete cosine transform (DCT) [6], discrete Fourier transform (DFT) [7] and discrete wavelet transform (DWT) [8]. In [9], authors have proposed a scheme using the application of stochastic resonance, where the ‘stochastic resonance’ has been used to study the optimal response in the system [10]. In [11], authors have reported a semi-blind watermarking scheme using SVD and DSR. They have embedded a gray-scale logo into the host image by extracting the singular values (SVs) and verify the presence of the watermark signal using the dynamic stochastic resonance (DSR). This application has been extended to a watermarking technique based on discrete cosine transform (DCT) [6]. In [12], authors have proposed a robust watermarking system, in which embedding of the watermark is carried out through singular value decomposition. However, false-positive detection problems are founds in the watermark extraction. In order to address this issue, several algorithms have been presented in last decade but most of them are not able to remove the false-positive detection completely. Therefore, our main motive is to design a robust watermarking system which is free from false-positive detection.

In this framework, a novel and robust image watermarking technique using SVD and DSR have been proposed. The input image is divided into non-overlapping blocks and then each block is transformed using all phase biorthogonal transform. These coefficients are used in the watermark embedding by extracting the significant coefficients of the watermark using singular value decomposition. In addition, a new authentication casting step is introduced to verify the authenticity of the watermark image. In this process, an authentication key is generated by extracting the features of the watermark image. In watermark extraction, firstly the authenticity of watermark image is verified using the original key to remove the falsification problem. This authenticated watermark and watermarked image are used in DSR based extraction process. The main advantage of DSR is that it improves the degree of robustness of the extracted watermark by utilizing the noise present in the system due to different intentional or unintentional attacks. The main reason is that it incorporates with SVD by the enhancing singular values (SVs) using the iterative process which reduces the effect of noise. The robustness of the proposed scheme is validated with different kind of distortions.

The rest of this paper is organized as follows: Sect. 2, provides a brief description of the all phase sine biorthogonal transform and dynamic stochastic resonance. Section 3 presents the proposed image watermarking scheme and an authentication casting process followed by the experimental results in Sect. 4. Finally, Sect. 5 summarises the concluding remarks.

## 2 Preliminaries

### 2.1 All Phase Sine Biorthogonal Transform

All Phase Sine Biorthogonal Transform (APSBT) can deduce from the discrete sine transform (DST) using sequence filtering in time domain. Discrete sine transform (DST) matrix can be described [13] as follows:

$$S_{i,j} = \frac{2}{\sqrt{2P+1}} \sin \left[ \frac{(2i+1)(j+1)\pi}{(2P+1)} \right]; \quad i, j = 1 \dots P-1 \quad (1)$$

For DST based all phase digital filtering can be design using a digital sequence. For this purpose, consider a digital sequence  $z(q)$  where each member of sequence correspond to  $P$  different values. The average of these values can be assigned as the all phase filtering. The output response can be expressed as follows:

$$y(q) = \sum_{i=0}^{P-1} \sum_{j=0}^{P-1} [H_{i,j} z(q - i + j)] \quad (2)$$

where

$$H_{i,j} = \frac{1}{P} \sum_{m=0}^{P-1} F_P(m) S(i, m) S(j, m) \quad (3)$$

Substituting Eq. (3) into Eq. (2) and after solving, the output is

$$y(q) = \sum_{\tau=-(P-1)}^{P-1} h(\tau) z(q - \tau) = h(q) * z(q) \quad (4)$$

where  $h(\tau)$  represent the unit impulse response and can be express as follows:

$$h(\tau) = \begin{cases} \sum_{i=\tau}^{P-1} H_{(i,i-\tau)} & \tau = 0, 1, \dots, P-1 \\ \sum_{i=0}^{\tau+P-1} H_{(i,i-\tau)} & \tau = -1, -2, \dots, -P+1 \end{cases} \quad (5)$$

Also,  $H_{i,j} = H_{j,i}$ . From Eqs. (3) and (5), we have

$$h(\tau) = \sum_{m=0}^{P-1} V(\tau, m) F_P(m) \quad \tau = 0, 1, \dots, P-1 \quad (6)$$

Matrix representation of Eq. (6) can be express as  $h = VF$ , where the transformation corresponding to matrix  $V$  is used to describe relationship between unit-pulse time response in time domain and sequence response in transform domain. This matrix  $V$  is know as APBST matrix and elements of  $V$  can be represented as:

$$V_{i,j} = \frac{1}{P} \sum_{i=0}^{P-1-i} S(j, \ell) S(j, \ell + i) \quad (7)$$

From Eqs. (1) and (7), APBST matrix can be given as follows:

$$V_{i,j} = \begin{cases} \frac{1}{P} & i = 0, j = 0, 1, \dots, P - 1 \\ \frac{4}{P(2P+1)} * \beta & i = 1, \dots, P - 1 \\ & j = 0, 1, \dots, P - 1 \end{cases} \quad (8)$$

where

$$\beta = \sum_{i=0}^{P-1-i} \sin \left[ \frac{(2j+1)(\ell+1)\pi}{(2P+1)} \right] \sin \left[ \frac{(2j+1)(\ell+i+1)\pi}{(2P+1)} \right] \quad (9)$$

## 2.2 Dynamic Stochastic Resonance

Stochastic Resonance (SR) is a remarkable approach which optimizes the performance of a non-linear system in presence of the noise at certain level. The general framework of SR is simple and can be described with some basic structural features like: (i) a potential activation barrier, (ii) an input signal, (iii) a source of noise. The noise can be considered as a free source of energy and as a result, this can affect a signal by enhancing it or degrade it. In contrast, if a non-linear interaction takes place between signal and noise exist at particular level then the signal may be benefited by their combined effect. Therefore, performance of the system may be enhanced in the presence of noise during transmission or signal detection.

A classic non-linear system that quantify the stochastic resonance (SR) can be describe with Langevin equation of motion [10] in the presence of periodic forcing as follows:

$$\dot{q}(t) = -V'(q) + K_0 \cos(\Omega t + \phi) + \omega(t) \quad (10)$$

where  $K_0$ ,  $\phi$  and  $\Omega$  represent signal amplitude, phase and frequency respectively. Also,  $V(q)$  denotes the bi-stable potential as given below:

$$V(q) = a \frac{q^2}{2} + b \frac{q^4}{4} \quad (11)$$

The potential  $V(q)$  is bi-stable and attains its minima at  $q = \pm\sqrt{\frac{a}{b}}$  with bi-stable parameter  $a$  and  $b$ . From Eq.(11), the autocorrelation of zero mean Gaussian noise  $\omega(t)$  [10] is given as

$$\langle \omega(t), \omega(0) \rangle = 2D\delta(t) \quad (12)$$

where  $D$  is the intensity with respect to stochastic fluctuation  $\omega(t)$ . If the input signal has weak amplitude then particle is unable to jump the potential barrier between two local states in the absence of noise and it remains confined near about one of the local states without transitions. But in the presence of periodic force and small noise, the input signal may be able to capitalize the noise and their combined effect enables the noise driven switching between the states. This switching happen at the Kramers frequency [14] as describe below:

$$r_k = \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{\Delta V}{D}\right) \quad (13)$$

The periodicity of transition between the states is given at the rate  $T_k(D) = 1/r_k$  and the maximum signal to noise ratio (SNR) can be observed with bi-stable parameter  $a = 2\sigma^2$ . From Eqs.(11), (12) and using the Eluer-Maryama's iterative method [15], the system can be modelled as:

$$q(n+1) = q(n) + \Delta t[a * q(n) - b * q^3(n) + \text{input}] \quad (14)$$

where  $\Delta t$  is a bistable parameter represent the sampling time. The term ‘input’ is used for sequence of input signal and noise. Further details can be found in [11].

### 3 Proposed Watermarking Technique

Let  $H$  and  $W$  be the original host and watermark images respectively of sizes  $M \times N$ . The watermark embedding can be formulated as follows:

#### 3.1 Embedding Process

1. The Host image  $H$  is divided into  $k \times k$  non-overlapping blocks and each block ( $B_i | i = 1, \dots, M \times N / k^2$ ) is transformed using all phase sine biorthogonal transform.
2. Obtained SVs of transformed image  $H_{SBT}$  and watermark  $W$ .

$$H_{SBT} = U_{SBT} * S_{SBT} * V_{SBT}^T \quad (15)$$

$$W = U_w * S_w * V_w^T \quad (16)$$

3. The SVs of  $S_{SBT}$  is modified using the SVs of watermark as follows:

$$S_{SBT}^w = S_{SBT} + \alpha * S_w \quad (17)$$

where  $\alpha$  is payload factor.

4. Obtain the modify transformed image by applying the inverse SVD operation.

$$H_{SBT}^{new} = U_{SBT} * S_{SBT}^w * V_{SBT}^T \quad (18)$$

5. Obtain the watermarked image by employing the inverse all phase discrete sine biorthogonal transform (IAPSBT) on  $H_{SBT}^{new}$ .

### 3.2 Authentication Casting Process

The main motivation of this step to design a process to rectify the false-positive problem of SVD in watermarking. The central idea is to cast an authentication process to verify the singular vectors of the watermark, which mainly causing the false-positive detection. The main steps involved in the process is summarized as follows:

1. The matrix  $U_w$  and matrix  $V_w$  is divided into non-overlapping blocks and each block is transformed into using all phase sine discrete biorthogonal transform.

$$U_{SBT}^w = \text{APSBT}\{U_w\}, \quad V_{SBT}^w = \text{APSBT}\{V_w\} \quad (19)$$

2. Select  $p$  random blocks from  $U_{SBT}^w$  and  $V_{SBT}^w$  using the secret key  $s$ . Let us denote them by  $\{U_i\}$  and  $\{V_i\}|i = 1 \dots p$ .
3. Choose  $L$  frequency coefficients from each selected block using zig-zag scan and stack into a vector. Let the vectors be  $S_U$  and  $S_V$ .
4. Construct binary sequences  $B_U$  and  $B_V$  from  $S_U$  and  $S_V$  as follows.

$$B_t = \begin{cases} 0 & (S_t < 0) \\ 1 & (S_t \geq 0) \end{cases}, \quad \text{where } t \in \{U, V\} \quad (20)$$

5. Estimate the authentication key using the XOR operation between  $B_U$  and  $B_V$  as follows.

$$A_{key} = \text{XOR}(B_U, B_V) \quad (21)$$

### 3.3 Extraction Process

Firstly, authenticity of watermark image is tested and if, it is found to be authentic then proceed for the watermark detection. The DSR phenomenon is employed in the extraction process to improve the robustness of the scheme.

#### Authentication Process

1. Let  $\bar{U}$  and  $\bar{V}$  be the received left and right singular vectors of the watermark image.
2. The authentication key  $A'_{key}$  is generated from  $\bar{U}$  and  $\bar{V}$  using the same secret key  $k$  as described in Sect. 2.
3. The authentication is successful if the normalized hamming distance [NHD] between the keys is less than from the threshold value.

$$Ham(A_{key}, A'_{key}) \begin{cases} \leq T, & \text{Successful Authentication} \\ > T, & \text{Unsuccessful Authentication} \end{cases} \quad (22)$$

In case of successful authentication, the extraction is performed using authentic singular vectors. In contrast, whereas extraction will be debarred for unsuccessful authentication. Let  $U_{aut} = \bar{U}$  and  $V_{aut} = \bar{V}$  denote the left and right singular vectors.

#### Watermark Extraction

1. The watermarked image  $Hw$  is divided into  $k \times k$  blocks and each block is transformed using all phase discrete biorthogonal transform into APSBT domain.

$$Hw_{SBT} = \text{APSBT}\{H_w\} \quad (23)$$

2. Obtained SVs of the transformed image  $Hw_{SBT}$ .

$$Hw_{SBT} = Uw_{SBT} * Sw_{SBT} * Vw_{SBT}^T \quad (24)$$

3. The process of DSR is initialized using the SVs and the corresponding bi-stable parameters as follows

$$q(0) = 0; b = \bar{m} * \frac{4a^3}{27}; a = k * 2 * \sigma_0^2; \quad (25)$$

where  $\bar{m} < 1$  and  $k < 1$  is the entity to ensure the sub-threshold condition in DSR and  $\sigma_0$  is standard deviation of SVs of intentional or unintentional attacked watermarked image in APSBT domain.

4. Obtained the tuned coefficients of input vector in the iterative process as provided Eq. (14).
5. Estimate SVs of the watermark as given below:

$$\lambda_{ext} = \frac{y_i - \lambda_{S_N}}{\alpha} \quad (26)$$

where  $\lambda_{ext}$  are extracted bits of watermark and  $\lambda_{S_N}$  are singular values of input image.  $y = \{y_i | i = 1 \dots n\}$  are the DSR tuned singular values.

6. Reconstruct watermark image  $W_R$  using inverse SVD transform and the authenticated singular vector as.

$$W_R = U_{aut} \lambda_{ext} V_{aut}^T \quad (27)$$

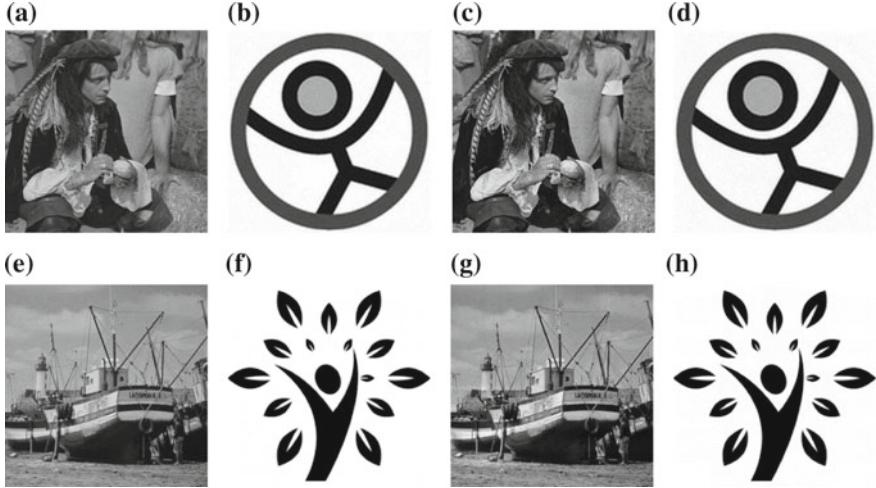
7. Compare the extracted watermark with original one by computing the correlation coefficients  $\rho$ , in which the value of  $\rho$  become smaller and smaller. In contrast, watermark corresponding to maximum  $\rho$  is considered to be optimal and final watermark.

## 4 Experimental Results

The efficiency of proposed watermarking scheme is measured by different experiments using the MATLAB platform. Standard gray scale images namely Pirate and Boat are used as the host images whereas ‘Circle’ and ‘Tree’ logos are considered as the corresponding watermarks. The size of host and watermarks images are  $256 \times 256$  and shown in Fig. 1a, e, b, f. The watermarked images and extracted watermarks are shown in Fig. 1c, g, d, h. The host images are partitioned into  $k \times k$  blocks with  $k = 8$  in the experiments. For perceptual analysis, the watermarked image is compared with the original one by estimating the peak signal to noise ratio (PSNR). The PSNR value for boat and pirate image are 31 db and 30 db respectively. The value of energy or payload factor for embedding strength is set to 0.10.

### 4.1 Performance of Authentication Process

The main objective of authentication process is to identify the authenticity of watermark image in the extraction process. For this purpose, the authenticity of watermark image is examined by generating an authentication key  $A'_{key}$  for extracted watermark and compute the normalized hamming distance with original key  $A_{key}$ . If this NHD is less than some prefixed threshold value then watermark image is considered to be authentic otherwise watermark image is not authentic. The threshold  $T$  is set to 0.450. The estimated NHD can be determined as follows:



**Fig. 1** Host images: **a** Pirate image, **e** Boat image; Watermark images: **b** Circle, **f** Tree; Watermarked images: **c** Watermarked pirate image, **g** Watermarked boat image; Extracted watermarks: **d** Circle, **h** Tree

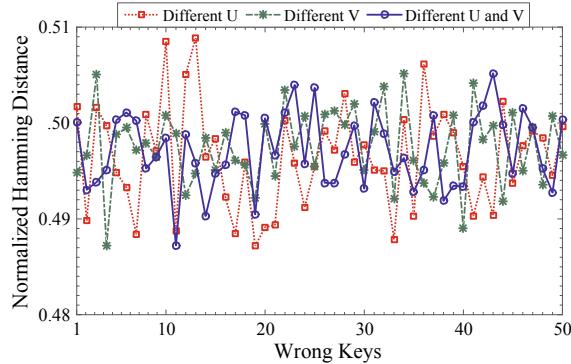
$$d(A_{key}, A'_{key}) = \frac{1}{L} \sum_{j=1}^L |A_{key} - A'_{key}| \quad (28)$$

here  $L$  represent the length of underlying keys. For experimental purpose, 50 different gray-scale images have been taken from SIPI image database and considered as the watermark. In order to check the authenticity of watermarks, SVD is performed on all watermarks including the original watermark. The estimated  $U$  matrix along with true  $V$  component is used to produce an authentication key and evaluate the hamming distance between the estimated keys and the original one. If NHD less than some threshold value, the watermark is considered as the authenticate watermark otherwise non-authentic. The similar process is repeated with  $U$  matrix of original watermark and  $V$  matrices of all 50 considered watermark. The process is also extended by passing the wrong  $U$  and  $V$  matrices of considered watermark. The estimated normalized hamming distances in all cases are depicted in Fig. 2. The maximum and minimum NHD are 0.4883 and 0.5093 respectively. The threshold  $T$  is set as 0.45. From the figure, it can be observed all watermarks are identified as the in-authenticate. For the true watermark, hamming distance is found to be zero and verify as the authenticate watermark.

## 4.2 Robustness Analysis

The robustness of the proposed algorithm is investigate by the underlying attacks such as noise addition (Gaussian, salt & pepper, speckle), Filtering operation (aver-

**Fig. 2** Normalized hamming distance analysis between the existent (original) and non-existent singular vectors



age, median), geometric attacks (cropping, rotation, re-sizing), JPEG compression and other attacks (sharp, histogram equalization (HE), Gaussian bluing, contrast adjustment). A DSR based approach is applied for watermark extraction in extraction process. The authenticity of the extracted logo image is validated by computing the normalized correlation coefficients (NCC). The normalized correlation (NC) coefficients can be described as follows:

$$\rho(w_1, w_2) = \frac{\sum_{i,j} (w_1 - \mu_{w_1})(w_2 - \mu_{w_2})}{\sqrt{\sum_{i,j} (w_1 - \mu_{w_1})^2} \sqrt{\sum_{i,j} (w_2 - \mu_{w_2})^2}} \quad (29)$$

where  $w_1$  and  $w_2$  represent the original and the extracted watermark with their mean values  $\mu_{w_1}$  and  $\mu_{w_2}$ . The extracted watermark circle and tree have the correlation coefficient 0.9999 and 1 respectively.

The degree of the robustness of proposed algorithm is measured by a series of different kind of attacks. The watermark images have been extracted from the attacked watermarked images by Gaussian noise addition (40%), salt & pepper noise (40%) and speckle noise (40%). The extracted watermarks are shown in Fig. 3i, ii, iii and corresponding correlation coefficients are listed in Table 1. Also, efficiency of the proposed scheme is also tested against some common geometric attacks like cropping (30%), resizing and rotation (40°). For resizing, firstly the watermarked image size is reduced from 256 to 64 and then scale up to the original watermarked image. The extracted watermarks after these attacks are shown in Fig. 3ix, x, xi. The performance of proposed scheme is also analyzed with JPEG compression (90%), SPIHT compression (90%) image sharpening (90%), histogram equalization (HE) and by increasing the contrast of the watermarked image (90%). The visual quality extracted watermarks demonstrate the robustness of the algorithm and depicted in Fig. 3vii, viii, xii-xiv.

The computational complexity of the proposed scheme has been minimized by the optimization of bi-stable parameters  $a$ ,  $b$ ,  $m$  and  $\Delta t$ . The variation in NC of

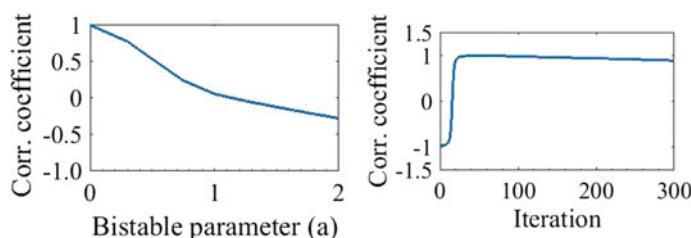


**Fig. 3** **a** Gaussian noise (40%), **b** Salt & Pepper noise (40%), **c** Speckle noise (40%), **d** Gaussian blur ( $9 \times 9$ ), **e** Average filter ( $9 \times 9$ ), **f** Median filter ( $9 \times 9$ ), **g** JPEG compression (90%), **h** SPIHT compression (90%), **i** Cropping (40%), **j** Resizing ( $256 \rightarrow 64 \rightarrow 256$ ) **k** Rotation (40°) **l** Histogram equalization, **m** Sharpen (100%), **n** Contrast adjustment (100%), **o** Gamma correction ( $\gamma = 3$ ) **p** Wrapping (75%) **q** Row deletion **r** Image tempering. The corresponding extracted watermark can be seen in Second, Fourth and Sixth row

watermarked image suffered from Gaussian noise with bistable parameter  $a$  after  $n$  iteration has been illustrated in Fig. 4. It can be observed form the figure, that maximum value of NC is achieved at  $a = 0.05$  and  $n = 45$  which is considered as optimal number of iteration.

**Table 1** NC after different distortions on experimental images

	Pirate		Boat	
Attacks	NC.	Iterations	NC	Iterations
Gaussian noise (40%)	0.9912	54	0.9590	68
Salt & Pepper (40%)	0.9243	300	0.9826	300
Speckle (40%)	0.9761	270	0.9901	268
Gaussian Blur (9 × 9)	0.9886	300	0.9769	69
Average Filter (9 × 9)	0.9841	300	0.9802	300
Median Filter (9 × 9)	0.9914	300	0.9823	290
Resizing (256 → 64 → 256)	0.9929	300	0.9901	255
Cropping (30%)	0.9886	300	0.9790	283
Rotation (40°)	0.9899	223	0.9952	269
Histogram equalization	0.9989	75	0.9880	43
Sharpen (100%)	0.9666	250	0.9541	226
Contrast adjustment (100%)	0.9782	149	0.9890	300
Gamma correction ( $\gamma = 3$ )	0.9458	300	0.9516	174
JPEG comp. (90%)	0.9953	300	0.9955	222
SPIHT comp. (90%)	0.9967	294	0.9980	300
Row deletion	0.9988	47	0.9911	82
Wrap (75%)	0.9897	300	0.9825	294
Image tempering	0.9907	155	0.9925	123

**Fig. 4** Variation of correlation coefficient with: **a** Bi-stable parameter  $a$ , **b** Number of iterations

## 5 Conclusion

In this paper, a robust image watermarking algorithm based on APSBT, DSR and SVD has been presented. In proposed method, A DSR based approach is applied for watermark extraction, which shows the good robustness against a variety of attacks. In DSR mechanism, the distribution of noisy coefficient is changed during the state transition from weak state to strong state by iterative process. This is due to the noise that was introduced during the attack on watermarked image that is utilized to suppress the noise influence in watermark extraction. This increases the robustness of the proposed algorithm. In addition, a verification step is also introduced to check the authenticity of the watermark image. The statistical results demonstrate that the performance of the proposed scheme is robust against different image distortions.

**Acknowledgements** The authors gratefully acknowledges the support of SERB, DST, India for this research work.

## References

1. Barni, M., Bartolini, F., Piva, A.: Improved wavelet-based watermarking through pixel-wise masking. *IEEE Trans. Image Process.* **10**(5), 783–791 (2001)
2. Katzenbeisser, S., Petitcolas, F.A.P.: *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech House, Boston (2002)
3. Singh, S.P., Bhatnagar, G.: A new robust watermarking system in integer DCT domain. *J. Vis. Commun. Image Represent.* **53**, 86–101 (2018)
4. Dong, P., Brankov, J.G., Galatsanos, N.P., Yang, Y., Davoine, F.: Digital watermarking robust to geometric distortions. *IEEE Trans. Image Process.* **14**(12), 2140–2150 (2005)
5. Singh, S.P., Bhatnagar G.: A novel chaos based robust watermarking framework. In: Proceedings of the International Conference Computer Vision and Image Processing, CVIP(2), pp. 439–447 (2017)
6. Jha, R.K., Chouhan, R.: Dynamic stochastic resonance-based grayscale logo extraction in hybrid SVD-DCT domain. *J. Frankl. Inst.* **351**(5), 2938–2965 (2014)
7. Ganic, E., Ahmet, M.E.: A DFT-based Semi-blind Multiple Watermarking Scheme for Images. CUNY Brooklyn College, 2900 (2004)
8. Bhatnagar, G., Wu, Q.M.J.: Biometrics inspired watermarking based on a fractional dual tree complex wavelet transform. *Future Gener. Comput. Syst.* **29**(1), 182–195 (2013)
9. Wu, G., Qiu, Z.: A novel watermarking scheme based on stochastic resonance. In: IEEE International Conference on Signal Processing, vol. 2, pp. 1–4 (2006)
10. Gammaitoni, L., Hanggi, P., Jung, P., Marchesoni, F.: Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–287 (1998)
11. Chouhan, R., Jha, R.K., Chaturvedi, A., Yamasaki, T., Aizawa, K.: Robust watermark extraction using SVD-based dynamic stochastic resonance. In: IEEE International Conference on Image Processing, pp. 2745–2748 (2011)
12. Liu, R., Tan, T.: An SVD-based watermarking scheme for protecting rightful ownership. *IEEE Trans. Multimed.* **4**(1), 121–128 (2002)
13. Hou, Z.X., Wang, C.Y., Yang, A.P.: All phase biorthogonal transform and its application in JPEG-like image compression. *Signal Process. Image Commun.* **24**(10), 791–802 (2009)
14. Hannes, R.: *The Fokker-Planck Equation*. Berlin, Heidelberg (1996)
15. Gard, T.C.: *Introduction to Stochastic Differential Equations*. Marcel-Dekker, New York (1998)

# Multi-scale Image Fusion Scheme Based on Gray-Level Edge Maps and Adaptive Weight Maps



Jitesh Pradhan, Ankesh Raj, Arup Kumar Pal and Haider Banka

**Abstract** Digital cameras and other digital devices cannot focus on all significant objects within the single frame due to their limited depth of focus. Consequently, few objects get attention in the captured image while the rest of the objects becomes background information. This problem can be overcome using different multi-focus image fusion techniques because it combines all the partially focused objects of different parent images into a single fully focused fused image. Hence, the final fused image focuses on each and every object of the parent images. In this paper, a novel multi-focus image fusion technique has been proposed which uses different edge-finding operators (like Sobel, Prewitt, Roberts, and Scharr) on preprocessed images. These different edge-finding operators have great pixel discrimination property which helps us to locate all the vital textural information of different partially focused images. Subsequently, an adaptive weight calculation approach has been introduced to generate weight maps of different parent images. Finally, all these parent image weight maps have been deployed into the winner-take-all scheme to integrate all parent images into a single fused image. Further, we have also considered different sets of partially focused images for experimental analysis. The experimental outcomes reveal that the proposed scheme is outperforming as compared to recent state of the arts.

**Keywords** Adaptive weight maps · Depth of focus · Edge maps · Image fusion · Multi-focus

---

J. Pradhan (✉) · A. Raj · A. Kumar Pal · H. Banka

Department of Computer Science and Engineering, Indian Institute of Technology (ISM),  
Dhanbad 826004, India

e-mail: [jitpradhan02@gmail.com](mailto:jitpradhan02@gmail.com)

A. Raj

e-mail: [ankesh.1200@gmail.com](mailto:ankesh.1200@gmail.com)

A. Kumar Pal

e-mail: [arupkpal@gmail.com](mailto:arupkpal@gmail.com)

H. Banka

e-mail: [haider.bank@iitm.ac.in](mailto:haider.bank@iitm.ac.in)

## 1 Introduction

### 1.1 Background

The human brain has capability to simulate very complex calculations within a fraction of second which helps our visual system to focus different depth of field (DoF) objects simultaneously. Consequently, the discrimination power of human visual system is highly enhanced to see all salient objects of a single scene with very high clarity. But in digital devices, the different DoF can be achieved by adjusting different camera settings to focus on different objects of the same scene. As a result, only a single object can be focused by a typical single lens digital camera. The object which has similar DoF with respect to the digital camera will be focused and the remaining part of the image will be considered as background information. Hence, it is required to capture multiple images with different DoF settings of digital camera to capture all prominent objects of the same scene. The major overhead with this approach is the processing of large number of images which exceedingly increases the storage space along with the transmission bandwidth. To address all these issues, multi-focus image fusion scheme [9] has been deployed. In this multi-focus image fusion process, multiple partially focused parent images of a single scene have combined together to generate a highly enhanced fused image which replicates the effect of human visual system. Thus, the final fused image binds all the salient focused regions of different parent images into a single frame. The multi-focus image fusion techniques are very useful in the field of medical imaging, robotics, military, machine vision, photography, and remote sensing. Hence, in today's scenario, a highly precise and efficient multi-focus image fusion scheme is required to address all these key challenges.

### 1.2 Literature Review

At the present time, the different multi-focus image fusion schemes can be categorized into two classes, which are spatial and transformed domain-based schemes [1]. Here, spatial domain schemes directly use linear combination image fusion approaches into the source parent images. Generally, all these spatial domain schemes use either pixel-based or block-based fusion techniques. Pixel-based techniques consider each and every pixel of the parent image and perform averaging operation for image fusion process. Consequently, these methods often fall under misregistration and noise problems. All these overheads can be controlled by considering block-based approaches rather than pixel-based approaches because the importance of every pixel were defined with the help of its neighborhood pixels. Initially, Goshtasby et al. [3] have addressed the image registration problem in the multi-focus image fusion scheme. They have registered all partially focused images into similar frame size and employed uniform nonoverlapping block division fusion scheme. Further, they have integrated the blocks of different parent images, which have higher gradi-

ent average into the final fused image. In 2007, Huang et al. [5] have also addressed image registration problem by considering the performance of sharpness matrices. Piella et al. [12] have used saliency factors of different image regions for image fusion. They have performed uniform block division and considered the degree of saliency of each block for image fusion process. Luo et al. [10] have introduced a new multi-focus image fusion scheme which follows the region partition tactics. This scheme removes the redundant regions of the parent images by considering the region homogeneity components. Some contemporary researchers have also proposed several transform domain-based image fusion schemes. Here, Li et al. [6] have introduced a transform domain based multi-focus image fusion scheme which uses the combination of curvelet and wavelet transforms. Further, in 2009, Looney et al. [8] have introduced a novel image fusion scheme which was inspired from the data-driven strategies. This data driven scheme decomposes the image signals into original scale components. In particular, this technique has been called as empirical mode decomposition (EMD). Later, Nejati et al. [11] have proposed a new focus creation scheme based on the surface area of the different intersection regions of parent images. In this scheme, all input images have been segmented based on the extracted intersection points. Afterward, they have performed the images fusion process based on the surface area of different regions. In 2017, Luo et al. [9] have used edge intensity (EDI) values along with higher order singular value decomposition (HOSVD) for multi-scale image fusion. In their scheme, they have used EDI to measure the edge sharpness and HOSVD to decompose the parent images. Finally, they have employed a sigmoid function to perform the image fusion task.

### **1.3 Motivation and Contribution**

Generally, all natural images possess rich textural information, clear and sharp edges along with definite boundaries. Thus the focused region of any image will have more detailed and sharp edges and boundaries. All these localized image information can be easily extracted by applying different transformation tools along with image decomposition approaches. But, this combination of image decomposition and transformation increases the time overhead and complexity of the image fusion process. These additional overheads can be controlled by applying different spatial domain techniques. In this paper, we have used different edge detection operators to enhance and extract the local geometrical structures of the image. In this paper, our main contributions are as follows:

- Point matching image registration technique has been adopted to produce identical frame images of all partially focused parent images.
- Nonlinear anisotropic diffusion method has been employed on parent images to eliminate all possible noise.
- Different edge-finding operators have been used to emphasize the local geometrical structure of the partially focused parent images.

- An adaptive weight calculation scheme has been introduced to calculate the importance of each parent image.
- Weight map based winner-take-all approach has been proposed for final image fusion process.

## 2 Basic Building Blocks

### 2.1 Nonlinear Anisotropic Diffusion (NLAD)

In general, most of the images possess homogeneous and nonhomogeneous pixel regions. The image regions where a group of pixels shows similar behaviors are treated as homogeneous pixel region. Subsequently, those pixel regions which show huge diversity in their characteristics are treated as nonhomogeneous pixel regions. Here, a partial differential equation (PDE) has been employed in anisotropic diffusion to enhance the nonhomogeneous pixel regions. This approach not only preserves the nonhomogeneous pixel regions but also perform region smoothening operation. In this work, we have adopted the nonlinear anisotropic diffusion [4] scheme to eliminate the noise factors from the partially focused images. This NLAD scheme is efficient and robust since it considers the nonhomogeneous pixel regions. This scheme computes a flux function to control the diffusion of parent images. For any image  $P$ , NALD flux function has been defined as follows:

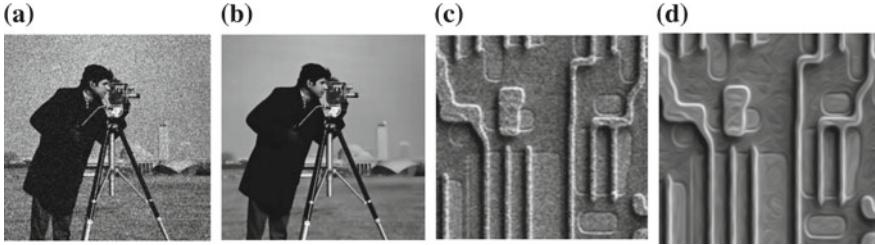
$$f(i, j, t) = df(\delta|P(i, j, t)|) \quad (1)$$

Here, flux function or rate of diffusion has been represented by  $f$ . Further, gradient operator and number of iteration/scale/time have been represented by  $\delta$  and  $t$ , respectively. At the same time, diffusion function  $df$  is defined as follows:

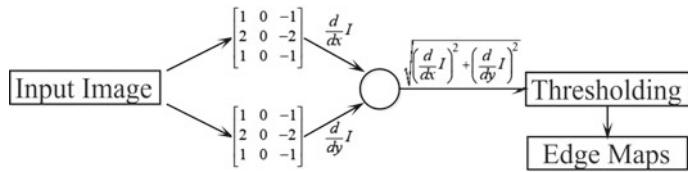
$$df(|X|) = e^{-\{(|X|/C)\}^2} \quad (2)$$

$$df(|X|) = \left( \frac{1}{1 + (|X|/C)^2} \right) \quad (3)$$

Diffusion function has been defined in two different ways and these functions have the capability to create trade-off between homogeneous region smoothening and edge preservation. If any image possess more sharp boundaries than first function, it will be very useful, whereas the second function will be more useful for images with wide regions. Here,  $C$  is a free parameter which helps to validate the clear boundaries of different image regions. Figure 1 shows the output of NLAD process and in this Fig. 1, we can see that the noise has been efficiently removed.



**Fig. 1** **a** Noisy cameraman image, **b** Output noise free cameraman image **c** Noisy texture image, and **d** Output noise free texture image



**Fig. 2** Schematic diagram of simple edge detection process

## 2.2 Edge Detection

In different vision systems, edge detection [2] plays a critical role since it gives local information of different geometrical structures, T junctions, straight lines, and textural patterns. Edge detection is very useful in image fusion, segmentation, object detection, and object tracking. In general, different edges can be detected by analyzing the variations along x-axis and y-axis gradients. These gradients can be computed with the help of different edge kernels. In this paper, we have used Sobel, Prewitt, Robert, and Scharr edge kernels to detect edges. Later, Fig. 2 demonstrates the simple schematic diagram of edge detection process.

## 3 Proposed Multi-focused Image Fusion Scheme

A new multi-focused image fusion scheme has been introduced in this paper, which uses edge maps and adaptive weight maps of different parent image for image integration. In this scheme,  $n$  different partially focused parent images have been selected. All these parent images have been captured at different DoF for a single scene. Suppose,  $P_1, P_2, P_3, \dots, P_n$  represent the  $n$  different partially focused parent images and all these images have been used as input images. First, we have employed Point Matching Image Registration technique on all input images. This technique will transform all these images into a similar image frame or coordinate system. In the next step, we have generated the grayscale images of all registered parent images,

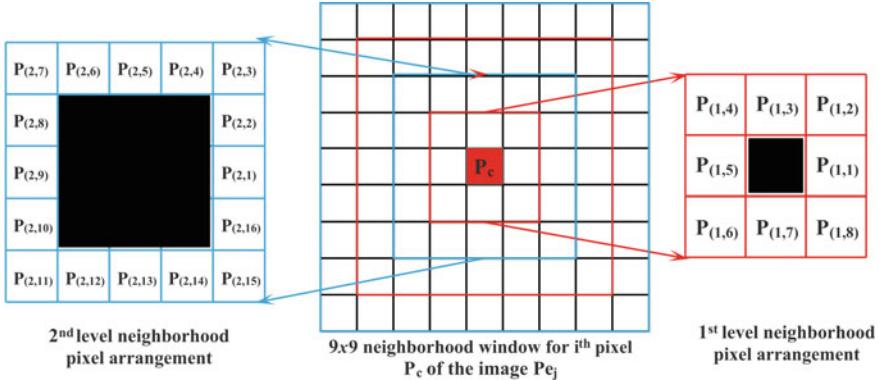
i.e.,  $Pg_1, Pg_2, Pg_3, \dots, Pg_n$ . Many times it has been observed that these grayscale images possess high noise factor which ultimately affects the quality of image integration. Hence, to avoid this problem, we have used nonlinear anisotropic diffusion (NLAD) scheme in grayscale parent images. This NLAD scheme not only removes the noise but also elevates the visually salient features of any image. All these noise-free images have been represented as  $Pf_1, Pf_2, Pf_3, \dots, Pf_n$ . Next, we have used different edge detection kernels to enhance and detect all local geometrical structures and boundaries of the parent images. This process will generate a detailed grayscale edge map for every noise-free image, i.e.,  $Pe_1, Pe_2, Pe_3, \dots, Pe_n$ . Finally, we have used the proposed weight map calculation scheme along with winner-take-all approach to integrate all parent images into a single fully focused image. In the following subsection, we have discussed this proposed image integration scheme in detail.

### 3.1 Adaptive Weight Map Calculation

Adaptive weight map  $Pw$  of any parent image grayscale edge map  $Pe$  shows the significance of each and every pixel in its surroundings. Those pixels which lie in the focused area will have higher weights in the weight map. These weight maps will be used for the final image fusion process. We have selected different neighborhood window size for every pixel of the grayscale edge map to calculate their weights. Here, the weight of any pixel reflects the uniqueness of the pixel in its surroundings. In the experiments, we have considered  $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ ,  $9 \times 9$ , and  $11 \times 11$  neighborhood window size for weight map calculation and for image integration. Here, we will discuss how an  $m \times m$  neighborhood window size-based weight map for  $n \times n$  grayscale edge map has been calculated where  $m < n$ .

First, we have considered a grayscale edge map  $Pe_j$  of the  $j$ th parent image  $P_j$ . For any pixel  $Pe_j(x, y)$ , we have extracted the  $m \times m$  size neighborhood window. In this window, the center pixel will be the  $Pe_j(x, y)$ . Let  $Nw_i$  represent the  $m \times m$  size neighborhood window for the  $i$ th pixel  $P_c$  of a grayscale edge map  $Pe_j$ . Further, we have extracted the different neighborhood levels with respect to the center pixel  $P_c$  of  $Nw_i$ . For an  $m \times m$  window, there will  $(m - 1)/2$  neighborhood levels. The immediate neighborhood pixels of the center pixel will fall under level 1 and the last pixels of the window will fall under the  $(m - 1)/2$  level. Figure 3, shows the  $9 \times 9$  neighborhood representation for the  $i$ th pixel  $P_c$  of a given edge map  $Pe_j$ . In the above figure, the center pixel  $P_c$  is the  $i$ th pixel of the image and its weight  $W_c$  in  $m \times m$  size neighborhood window will be calculated as follows:

$$W_c = \frac{1}{2} \times P_c + \frac{1}{2} \sum_{k=1}^{(m-1)/2} w_k \quad (4)$$



**Fig. 3** 9 × 9 neighborhood representation for the  $i$ th pixel  $P_c$  of a given edge map  $Pe_j$

Here,  $w_k$  represents the relative weight value of the  $k$ th neighborhood level which is defined as follows:

$$\mu_k = \frac{1}{N} \times \sum_{l=1}^N P_{(k,l)} \quad (5)$$

$$d_k = P_c - \mu_k \quad (6)$$

$$w_k = \frac{d_k}{\sum_{r=1}^{(m-1)/2} d_r} \quad (7)$$

Here,  $N$  is the number of pixels falls under  $k$ th neighborhood level and  $P_c$  gives the intensity value of the center pixel.  $P_{(k,l)}$  is the intensity value of the  $l$ th pixel of  $k$ th neighborhood level. In this way, an adaptive weight map  $Pw_j$  will be generated for given grayscale edge map  $Pe_j$ .

### 3.2 Winner-Take-All Scheme

In this step, we will select all the partially focused parent images  $P_1, P_2, P_3, \dots, P_n$  as well as their weight maps  $Pw_1, Pw_2, Pw_3, \dots, Pw_n$ . These original maps and weight maps will be used for the final image integration process. First, we will select the  $i$ th pixel from each partially focused parent image with their corresponding weight values from the weight maps. The  $i$ th pixel of any image which has maximum weight value will be the winner pixel and will be integrated in the final fully focused image. The final fully focused fused image  $F_f$  is calculated as follows:

$$F_f^C(x, y) = \begin{cases} F_f^C(x, y) = P_1^C(x, y), & \text{If } fw(Pw_i(x, y)) = Pw_1(x, y) \\ F_f^C(x, y) = P_2^C(x, y), & \text{If } fw(Pw_i(x, y)) = Pw_2(x, y) \\ F_f^C(x, y) = P_3^C(x, y), & \text{If } fw(Pw_i(x, y)) = Pw_3(x, y) \\ \dots \\ F_f^C(x, y) = P_i^C(x, y), & \text{If } fw(Pw_i(x, y)) = Pw_i(x, y) \\ \dots \\ F_f^C(x, y) = P_n^C(x, y), & \text{If } fw(Pw_i(x, y)) = Pw_n(x, y) \end{cases} \quad (8)$$

Here,  $F_f^C(x, y)$  represents the  $i$ th pixel of the different color components of the final image  $F_f$  where,  $C \in \{\text{Red}, \text{Green}, \text{Blue}\}$  color components. So,  $i$ th pixel of different color components will be written as  $F_f^{\text{Red}}(x, y)$ ,  $F_f^{\text{Green}}(x, y)$  and  $F_f^{\text{Blue}}(x, y)$ . Similarly,  $P_i^C(x, y)$  represents the  $i$ th pixel of the different color components of any parent image  $P_i$ . Simultaneously,  $fw(\cdot)$  is the winner function and it is defined as follows:

$$fw(Pw_i(x, y)) = \max \{Pw_1(x, y), Pw_2(x, y), Pw_3(x, y), \dots, Pw_n(x, y)\} \quad (9)$$

### 3.3 Proposed Multi-focused Image Fusion Architecture

Algorithm 1 shows the different steps of the proposed multi-focus image integration scheme. Subsequently, the schematic block diagram of the suggested image integration scheme has been shown in Fig. 4.

---

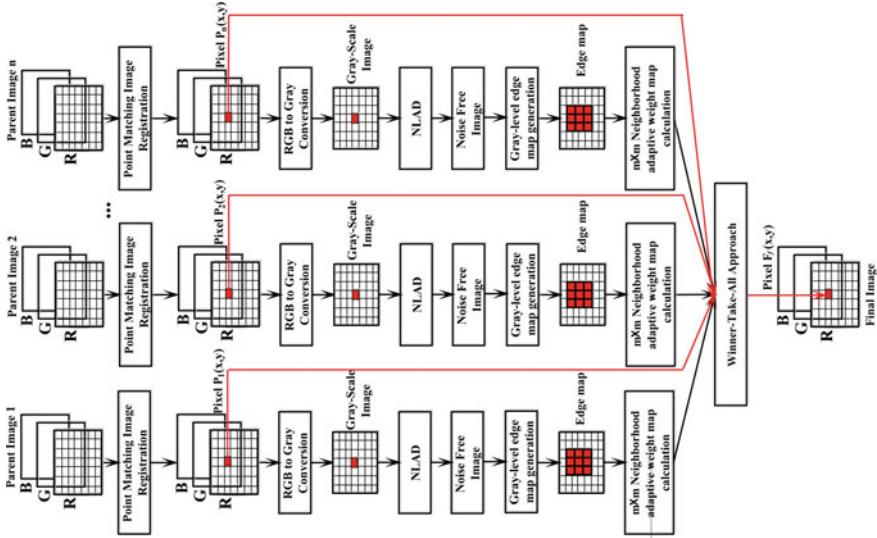
#### Algorithm 1 Multi-scale Image Fusion.

---

**Input:**  $n$  partially focused parent images, i.e.,  $P_1, P_2, P_3, \dots, P_n$ .

**Output:** Final fully focused integrated image  $F_f$ .

- 1: Select  $n$  partially focused parent images  $P_1, P_2, P_3, \dots, P_n$  of the same scene with different DOF.
  - 2: Use Point Matching Image Registration technique to transform all parent images to achieve identical frame structure..
  - 3: Compute grayscale images  $Pg_1, Pg_2, Pg_3, \dots, Pg_n$  of all registered images.
  - 4: Generate the noise free images  $Pf_1, Pf_2, Pf_3, \dots, Pf_n$  by employing nonlinear anisotropic diffusion (NLAD) in all grayscale parent images.
  - 5: Apply Sobel, Prewitt, Robert, or Scharr edge kernels in all noise free images to generate grayscale parent edge maps  $Pe_1, Pe_2, Pe_3, \dots, Pe_n$ .
  - 6: Apply  $m \times m$  neighborhood adaptive weight map calculation technique to generate weight maps  $Pw_1, Pw_2, Pw_3, \dots, Pw_n$  of all grayscale edge maps.
  - 7: Select all  $n$  weight maps  $Pw_1, Pw_2, Pw_3, \dots, Pw_n$  and original parent images  $P_1, P_2, P_3, \dots, P_n$  for winner-take-all scheme to integrate all important pixels of different parent images into single fully focused image  $F_f$ .
-



**Fig. 4** Schematic block diagram of the proposed multi-scale image integration scheme

## 4 Experimental Results and Discussion

In this paper, we have conducted several image fusion experiments on two different partially focused image datasets. These image datasets are standard image sets which are widely used in image fusion experiments. These datasets are Book and Clock image datasets. In these datasets, two different focused images with their resultant fused image are present for comparative analysis. We have employed Sobel, Pre-witt, Robert, and Scharr edge kernels in all image sets in the experiments. Figure 5 shows the output of grayscale edge maps for different Book images. We have carried out our first image fusion experiment on Book image dataset, which contains two partially focused and one reference image of size  $1024 \times 768$ . We have used  $3 \times 3$  neighborhood window for adaptive weight map generation along with all four edge detection kernels. The output of the image fusion process on this Book dataset has been shown in Fig. 6.

In this Book image dataset fusion experiments, we have also considered different neighborhood windows for adaptive weight map calculation. Further, we have used these different adaptive weight maps with different edge kernels for fusion experiments. Here, we have calculated peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and root mean square error (RMSE) to check the performance of image fusion. Tables 1, 2, 3, and 4 show the experimental results of all these image fusion experiments on Book image set. Next, we have carried out our second on Clock image dataset which contains two partially focused and one reference image of size  $512 \times 512$ . The output integrated images generated from these experiments



**Fig. 5** Output different grayscale edge maps for different Book images



**Fig. 6** Image fusion results on Book image set **a** Parent image 1 **b** Parent image 2 **c** Reference image **d** Sobel output **e** Prewitt output **f** Roberts output **g** Scharr output

have been presented in Fig. 7. Later, Tables 5, 6, 7, and 8 show the experimental results of all these image fusion experiments on Clock image dataset.

Further, we have also compared our proposed multi-scale image fusion scheme with other standard image fusion schemes. These are curvelet transform, morpho-

**Table 1** Image fusion performance on Book image dataset using Sobel kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	27.7930	26.6628	25.6663	24.8803	24.2065
RMSE	0.0411	0.0464	0.0521	0.0570	0.0616
SSIM	0.9466	0.9453	0.9441	0.9438	0.9436

**Table 2** Image fusion performance on Book image dataset using Prewitt kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	27.7436	26.6666	25.6654	24.8698	24.2098
RMSE	0.0410	0.0464	0.0521	0.0571	0.0612
SSIM	0.9471	0.9454	0.9442	0.9437	0.9435

**Table 3** Image fusion performance on Book image dataset using Robert's kernels

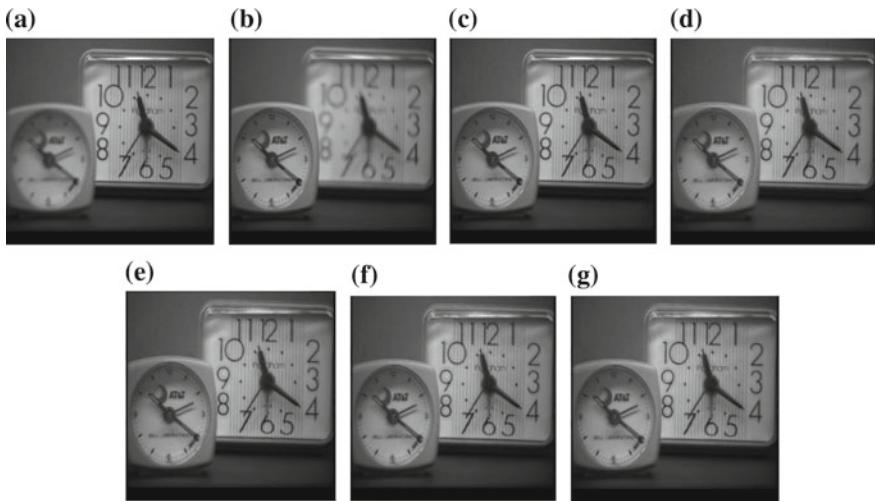
Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	27.7597	26.7356	25.7679	24.9607	24.2779
RMSE	0.0409	0.0460	0.0515	0.0565	0.0611
SSIM	0.9473	0.9463	0.9463	0.9461	0.9457

**Table 4** Image fusion performance on Book image dataset using Scharr kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	27.7044	26.6684	25.6735	24.8862	24.2094
RMSE	0.0412	0.0464	0.0520	0.0570	0.0616
SSIM	0.9465	0.9453	0.09441	0.09441	0.09441

logical component analysis (MCA), discrete cosine transform (DCT), TV-11 model, and block-level DCT-based image fusion schemes. Here, Table 9 demonstrates the comparative analysis in terms of PSNR and execution time for Book image dataset. We have calculated these results by comparing the reference Book image with the output fused Book image.

Later, Table 10 also demonstrates the comparisons between other standard methods [7] and our proposed image fusion method. In this table, we have shown comparisons in terms of Average Gradient (AG), Edge Retention, Figure definition (FD),



**Fig. 7** Image fusion results on Clock image set **a** Parent image 1 **b** Parent image 2 **c** Reference image **d** Sobel output **e** Prewitt output **f** Roberts output **d** Scharr output

**Table 5** Image fusion performance on Clock image dataset using Sobel kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	30.9533	32.8023	33.6181	33.9233	33.2724
RMSE	0.0283	0.0229	0.0208	0.0201	0.0202
SSIM	0.9366	0.9395	0.9419	0.9423	0.9419

**Table 6** Image fusion performance on Clock image dataset using Prewitt kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	30.9500	32.7623	33.6412	33.8807	33.8861
RMSE	0.0283	0.0230	0.0208	0.0202	0.0202
SSIM	0.9368	0.9396	0.9421	0.9424	0.9423

QAB/F, Cross-Entropy (CE), Mutual Information (MI), Cross-Entropy (CE), Relatively Warp (RW), Edge Intensity, (EI), and Structural Similarity (SSIM). These other standard image fusion methods are Morphological Pyramid method (MP), Ratio Pyramid method (RP), Laplacian Pyramid method (LP), DWT-based method

**Table 7** Image fusion performance on Clock image dataset using Roberts kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	30.5863	32.1903	33.0665	33.7695	33.8801
RMSE	0.0296	0.0246	0.0222	0.0205	0.0202
SSIM	0.9277	0.9315	0.9353	0.9372	0.9386

**Table 8** Image fusion performance on Clock image dataset using Scharr kernels

Performance parameter	Different neighborhood window size for adaptive weight map calculation				
	$3 \times 3$	$5 \times 5$	$7 \times 7$	$9 \times 9$	$11 \times 11$
PSNR	30.9394	32.7662	33.6152	33.9159	33.8922
RMSE	0.0284	0.0230	0.0209	0.0201	0.0202
SSIM	0.9362	0.9388	0.9414	0.9423	0.9418

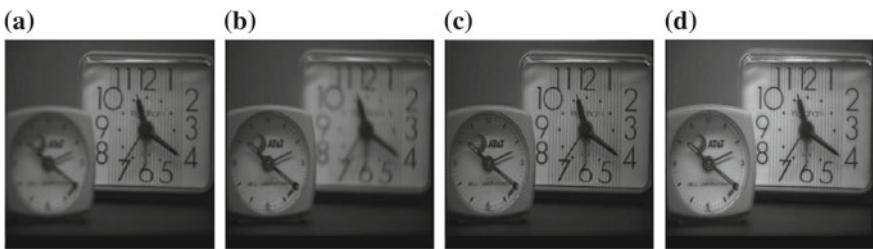
**Table 9** PSNR and Execution time performance analysis between different multi-scale image fusion techniques and the proposed image fusion technique

S. No.	Schemes	PSNR	Execution time (s)
1.	MCA	27.7943	6.8715
2.	DCT	22.4146	1.9588
3.	Block DCT	25.5751	5.3015
4.	TV	21.7943	6.5781
5.	Curvelet	25.2725	1.6655
6.	Proposed	27.7930	6.7821

(DWT), average-based method (AV), Select maximum based method (Max), PCA-based method (PCA), Contrast Pyramid method (CP), FSD Pyramid method (FSD), and Gradient Pyramid method (GP). After analyzing the performance of our proposed method in Tables 9 and 10, we can say that the proposed methods in producing satisfactory results. Simultaneously, in Table 9, we can see that the PSNR value achieved through the proposed method is better than other methods. Further, we have also compared our proposed image integration technique with the Singh et al. [4] and our method has shown comparable performances as compared to their method. Here, Fig. 7 shows the visual comparison between our suggested technique and Sing's technique (Fig. 8).

**Table 10** Different evolution parameter-based performance comparisons between current multi-scale image fusion techniques and the proposed image fusion technique

S. No.	Method	RW	SSIM	FD	MI	QAB/F
1.	MCA	0.0408	0.8977	7.1332	4.3807	0.6329
2.	TV-II	0.1928	0.7931	12.5483	3.8879	0.6386
3.	LP	0.0062	0.9211	11.6845	5.2433	0.7746
4.	RP	0.0097	0.9436	10.0131	5.6569	0.7625
5.	MP	0.0177	0.09048	12.1859	4.4438	0.7296
6.	GP	0.0862	0.8954	10.4233	4.0176	0.7416
7.	FSD	0.0861	0.8927	10.4930	4.0011	0.7310
8.	DWT	0.0033	0.9136	12.3349	4.8076	0.7621
9.	CP	0.0087	0.9168	11.9958	5.3669	0.7712
10.	PCA	0.0266	0.9587	8.8619	5.4402	0.7990
11.	MAX	0.0205	0.9547	9.2181	5.8521	0.7666
12.	AV	0.0270	0.9626	8.8379	5.3814	0.7978
13.	Proposed	0.0401	0.9466	11.0998	6.3146	0.7996



**Fig. 8** Image fusion result comparison on Clock image set **a** Parent image 1 **b** Parent image 2 **c** Singh et al. **d** Proposed technique

## 5 Conclusions

In this work, the authors have suggested a novel multi-scale image fusion scheme which uses the different edge detection kernels to generate detailed grayscale edge map. Initially, we have used robust point matching based image registration scheme followed by NLAD process on a registered grayscale parent image. Later, we have used four different edge detection kernels to preserve the local structure of the noise-free image and to enhance the edges, boundaries, and texture of the image. This approach will generate a grayscale detailed edge map. Later, we have proposed an adaptive weight map generation scheme based on the neighboring pixels values. This weight map will give the uniqueness and importance of each and every pixel of any parent image. Finally, we have employed the winner-take-all approach which used parent image weight maps to generate the final fully focused fused image. We have

also carried out different image fusion experiments on two standard image datasets and the results are satisfactory. These experimental results have also shown better PSNR values as compared to the other state-of-the-arts methods.

## References

1. Bai, X., Zhang, Y., Zhou, F., Xue, B.: Quadtree-based multi-focus image fusion using a weighted focus-measure. *Inf. Fusion* **22**, 105–118 (2015). <https://doi.org/10.1016/j.inffus.2014.05.003>, <http://www.sciencedirect.com/science/article/pii/S1566253514000669>
2. Davis, L.S.: A survey of edge detection techniques. *Comput. Graph. Image Process.* **4**(3), 248–270 (1975). [https://doi.org/10.1016/0146-664X\(75\)90012-X](https://doi.org/10.1016/0146-664X(75)90012-X), <http://www.sciencedirect.com/science/article/pii/0146664X7590012X>
3. Goshtasby, A.A.: Fusion of multifocus images to maximize image information (2006). <https://doi.org/10.1117/12.663706>
4. Harbinder Singh, V.K., Bhooshan, S.: Anisotropic diffusion for details enhancement in multiexposure image fusion. *ISRN Signal Process.* **2013**, 1–18 (2013). <https://doi.org/10.1155/2013/928971>
5. Huang, W., Jing, Z.: Evaluation of focus measures in multi-focus image fusion. *Pattern Recognit. Lett.* **28**(4), 493–500 (2007). <https://doi.org/10.1016/j.patrec.2006.09.005>, <http://www.sciencedirect.com/science/article/pii/S0167865506002352>
6. Li, S., Yang, B.: Multifocus image fusion by combining curvelet and wavelet transform. *Pattern Recognit. Lett.* **29**(9), 1295–1301 (2008). <https://doi.org/10.1016/j.patrec.2008.02.002>, <http://www.sciencedirect.com/science/article/pii/S0167865508000561>
7. Liu, Z., Chai, Y., Yin, H., Zhou, J., Zhu, Z.: A novel multi-focus image fusion approach based on image decomposition. *Inf. Fusion* **35**, 102–116 (2017). <https://doi.org/10.1016/j.inffus.2016.09.007>, <http://www.sciencedirect.com/science/article/pii/S1566253516300781>
8. Looney, D., Mandic, D.P.: Multiscale image fusion using complex extensions of emd. *IEEE Trans. Signal Process.* **57**(4), 1626–1630 (2009). <https://doi.org/10.1109/TSP.2008.2011836>
9. Luo, X., Zhang, Z., Zhang, C., Wu, X.: Multi-focus image fusion using hosvd and edge intensity. *J. Vis. Commun. Image Represent.* **45**, 46–61 (2017). <https://doi.org/10.1016/j.jvcir.2017.02.006>, <http://www.sciencedirect.com/science/article/pii/S1047320317300433>
10. Luo, X., Zhang, J., Dai, Q.: A regional image fusion based on similarity characteristics. *Signal Process.* **92**(5), 1268–1280 (2012). <https://doi.org/10.1016/j.sigpro.2011.11.021>, <http://www.sciencedirect.com/science/article/pii/S0165168411004063>
11. Nejati, M., Samavi, S., Karimi, N., Soroushmehr, S.R., Shirani, S., Roosta, I., Najarian, K.: Surface area-based focus criterion for multi-focus image fusion. *Inf. Fusion* **36**, 284–295 (2017). <https://doi.org/10.1016/j.inffus.2016.12.009>, <http://www.sciencedirect.com/science/article/pii/S156625351630255X>
12. Piella, G.: A general framework for multiresolution image fusion: from pixels to regions. *Inf. Fusion* **4**(4), 259–280 (2003). [https://doi.org/10.1016/S1566-2535\(03\)00046-0](https://doi.org/10.1016/S1566-2535(03)00046-0), <http://www.sciencedirect.com/science/article/pii/S1566253503000460>

# Comparison of Reconstruction Methods for Multi-compartmental Model in Diffusion Tensor Imaging



Snehlata Shakya and Sanjeev Kumar

**Abstract** Diffusion tensor imaging (DTI) is one of the magnetic resonance techniques to describe the anisotropic diffusion in terms of its orientation. DTI gives the direction of white matter fibers in a single direction. However, multi-fiber heterogeneity can be present at several places of the human brain. Recently, a multi-compartmental model (which uses noncentral Wishart distributions) was proposed to improve the state of the art of solving this multi-fiber heterogeneity. In this model, nonnegative least square (*NNLS*) method was used for solving the inverse problem which is based on  $L_2$  norm minimization. In this paper, results are obtained with the least absolute shrinkage and selection operator ( $L_1$  regularization). In particular, we study the performance of *NNLS* and nonnegative *lasso* methods and shown that the later method outperforms for several cases.

**Keywords** Crossing fibres · DTI · Lasso · Least absolute shrinkage · MRI

## 1 Introduction

Diffusion tensor imaging (DTI) is a magnetic resonance imaging (MRI) technique, which was first introduced by Basser et al. [1] in 1994. It plays an important role in estimating the direction of diffusivity, location, and anisotropy of the human brain's white matter tracts. Technical details and applications of this technique can be found in [2–5]. A second-order diffusion tensor, associated with a voxel, describes the shape and orientation of fiber. Multi-compartmental model gives a mixture of second-order diffusion tensors, which better solves the directionality of multi-fibers. A voxel is considered to have several compartments and the MR signal can be seen as a weighted sum of signals from those compartments. Several methods have been proposed ear-

---

S. Shakya (✉) · S. Kumar  
Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India  
e-mail: [sneh022@gmail.com](mailto:sneh022@gmail.com)

S. Kumar  
e-mail: [malikfma@iitr.ac.in](mailto:malikfma@iitr.ac.in)

lier to solve this fiber heterogeneity within a voxel. Bi-exponential diffusion MRI [6], mixture of Gaussian distributions [6, 7], mixture of central Wishart distributions [8], and mixture of hyper-spherical von Mises–Fisher Distributions [9] are some models that work to resolve multiple fiber case. Recently, Shakya et al. [10] proposed a mixture of noncentral Wishart distribution model for solving multi-fiber heterogeneity more accurately. They have used  $L_2$  norm-based  $NNLS$  method [11] for solving the linear system of equations.  $NNLS$  is a constrained optimization problem with the constraint of nonnegative coefficients. We are using non-negativity constraint because the solutions (here weights of fibers in different directions within a voxel) can never be negative. This choice reduces the computation cost and physically it does not give false results. We revisited the model [10] by using  $L_1$  regularization-based method, namely, nonnegative least square *lasso* [12] and compared the results with previously used  $NNLS$  method. We present some simulations for two fibers crossing at different separation angles. We also did analysis by introducing Rician noise in the data.

## 2 Theory

In Diffusion MRI, gradient pulse field is also applied along with the homogeneous magnetic field. Signals in the absence and presence of these pulse gradient gives the diffusion of water molecules:

$$\frac{S}{S_0} = \exp[-\gamma^2 G^2 \delta^2 (\Delta - \frac{\delta}{3}) D] \quad (1)$$

where,

$S$ : MR signal in presence of diffusion weighting gradients,

$S_0$ : MR signal in absence of any diffusion weighting gradients,

$G$ : Strength of gradient pulse,

$\gamma$ : Gyromagnetic ratio,

$\delta$ : Duration of diffusion gradient,

$\Delta$ : Time between two gradient pulse.

The MR signal decay equation in probabilistic framework can be given as follows [8]:

$$S = S_0 \int_{\mathbf{P}_n} f(\mathbf{D}) \exp(-b\mathbf{g}^T \mathbf{D}\mathbf{g}) d\mathbf{D}, \quad (2)$$

$$= S_0 \int_{\mathbf{P}_n} f(\mathbf{D}) \exp(-\text{trace}(\mathbf{B}\mathbf{D})) d\mathbf{D} = S_0 \mathbf{L}_f(\mathbf{B}), \quad (3)$$

where

$\mathbf{P}_n$ : space defined on  $n \times n$  symmetric positive-definite (SPD) matrices,

$\mathbf{D}$ : diffusion tensor,

$\mathbf{B} = \mathbf{b}\mathbf{g}\mathbf{g}^T$ ;  $\mathbf{B} \in \mathbf{P}_n$  with  $\mathbf{b} = (\gamma\delta\mathbf{G})^2t$  is the “b-value”,

$t$ : Effective diffusion time,

$\mathbf{G}$  and  $\mathbf{g}$ : Magnitude and direction of the diffusion sensitizing gradient  $\mathbf{G}$ ,

$f(\mathbf{D})$ : Density function on the space of  $\mathbf{P}_n$  with respect to some measure  $d\mathbf{D}$ ,

$\mathbf{L}_f$ : Standard Laplace transform of a function  $f$ .

## 2.1 Mixture of Noncentral Wishart Distributions Model

Detailed description about the model is available in [10], here we present a brief summary of the model. Laplace transform of noncentral Wishart distribution [13] gives the following relation:

$$\int \exp(-\text{trace}(\Theta\mathbf{u})) W_n(p, \Sigma, \Omega) d\mathbf{u} = |\mathbf{I}_n + \Theta\Sigma|^{-p} \times \\ \exp[-\text{trace}(\Theta(\mathbf{I}_n + \Theta\Sigma)^{-1}\Omega)], \quad (4)$$

where  $(\Theta + \Sigma^{-1}) \in \mathbf{P}_n$ .  $W_n(p, \Sigma, \Omega)$  is the notation for noncentral Wishart distribution with  $p$  as shape parameter,  $n$  as dimension of space,  $\Sigma$  as scale parameter and  $\Omega$  as the noncentrality parameter. By making use of Eqs. 3 and 4, we have:

$$\frac{S}{S_0} = \sum_{i=1}^N w_i (1 + \text{trace}(\mathbf{B}\Sigma_i))^{-p} \exp[-\text{trace}(\mathbf{B}(\mathbf{I}_n + \mathbf{B}\Sigma_i)^{-1}\Omega_i)], \quad (5)$$

where  $w_i$  are the mixture weights. If we fix  $p_i$ 's,  $\Sigma_i$ 's and  $\Omega_i$ 's then above system of equations will become linear,  $\mathbf{Aw} = \mathbf{s}$ , with  $\mathbf{s} = S/S_0$ . We will need to solve this system for unknown  $\mathbf{w}$ . Parameter estimation part is explained in [10]. Previously, weights  $\mathbf{w}$  were computed from (i) pseudo-inverse, (ii) Tikhonov regularization, (iii)  $L_1$  minimization with equality constraints, (iv)  $L_1$  minimization with quadratic constraint, and (v) Nonnegative least square (*NNLS*) methods [8]. The *NNLS* method was performing best among all abovementioned methods. We now solve the linear system of equations with *NNLS* [11] and nonnegative squared *lasso* [12] methods. The *lasso* is a  $L_1$  minimization and most fundamental technique but it is less used. MATLAB™ built-in function *lsqlnonneg* is used for *NNLS* and online available code is used for implementing *lasso* [14]. NURBS-Snakes based Energy Minimization [15] and smeared entropy maximization [16] are other optimization approaches that can further be implemented.

## 2.2 NNLS and Lasso Methods

The *NNLS* is a convex optimization problem which mostly performs good. The problem is defined as follows:

$$\min_{\beta \geq 0} \frac{1}{n} \|y - \mathbf{A}\beta\|_2^2, \quad (6)$$

Here,  $y$  are known observations,  $\mathbf{A} \in \mathbb{R}^{n \times p}$  is known as design matrix and  $\beta$  are the unknowns to be estimated.

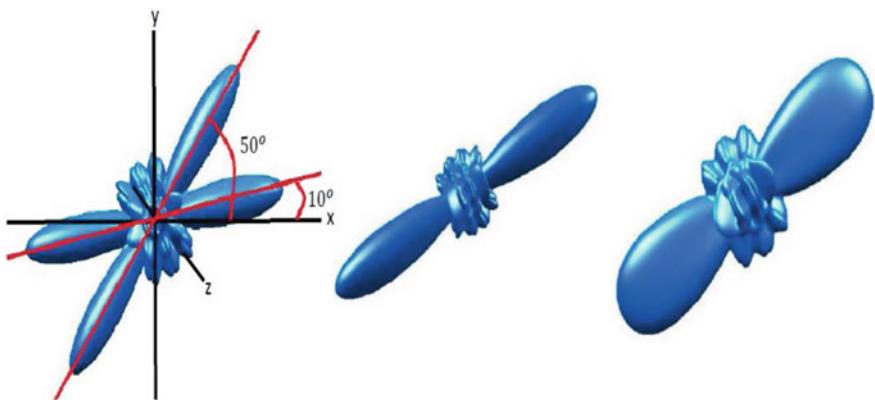
The nonnegative *lasso* method is not much different from *NNLS*, it incorporates a tuning parameter  $\lambda$ ,

$$\min_{\beta \geq 0} \frac{1}{n} \|y - \mathbf{A}\beta\|_2^2 + \lambda \|\beta\|_1, \lambda > 0 \quad (7)$$

The tuning parameter needs to be specified and it may vary for different applications. Once this parameter is optimized, the algorithm gives significantly improved results. In the present study, we defined this parameter from mean and standard deviation of the observations.

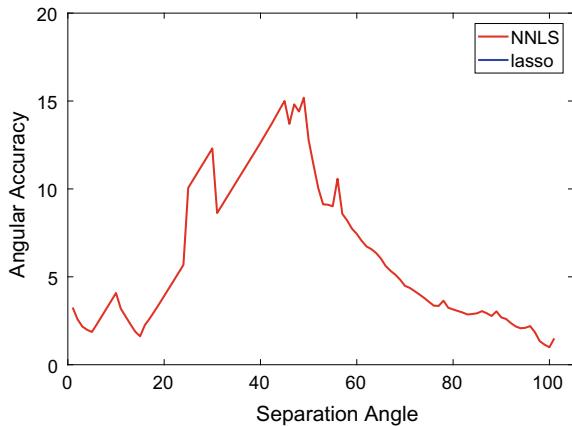
### 3 Results and Discussion

Crossing fibers may not be distinguished for small separation angles. We have done a simulation for crossing fiber case where the separation angle is  $40^\circ$ . Polar angle is  $90^\circ$ , i.e., fibers are in xy-plane making  $90^\circ$  angle with z-axis. Azimuthal angles (angle from x-axis in xy-plane) are chosen as  $10^\circ$  and  $50^\circ$ . Figure 1 displays the case with the left image as original, right with *lasso* and the middle with *NNLS*. Solving with *NNLS* method gives a single fiber orientation which is close to the mean of two fiber directions. The other method, *lasso*, predicts the presence of two fibers. Therefore, it is necessary to choose the algorithm carefully. Spherical harmonics



**Fig. 1** Visualization of two crossing fibers. Left is the original fiber orientation ( $10^\circ, 50^\circ$ ), middle is with *NNLS*, and the right is with *lasso*

**Fig. 2** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ$ ,  $\phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = \theta_2 = 90^\circ$



expansions with order 10 are used for visualization. Increasing the order, two fibers will be distinguishable visually for *lasso* method. But, then it will be computationally costly.

Next, we have done several simulations with changing the separation angles between two crossing fibers. First, the results are produced in the absence of any noise in the data and second, the noise is introduced to the simulated data. Daducci et al. [17] have made a quantitative comparison of 20 different reconstruction methods. They made a comparison of results in terms of angular accuracy,  $\bar{\theta}$ , which is defined below. We will also use the same error estimate for comparing the results.

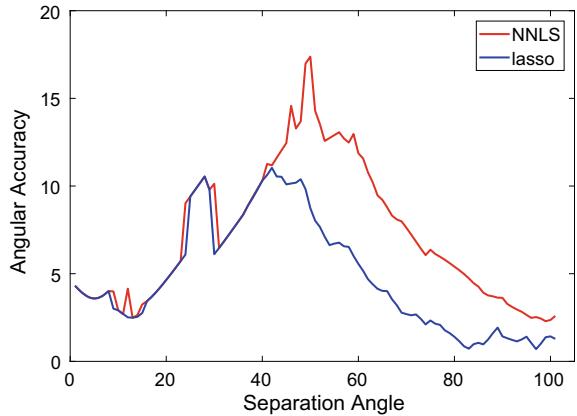
$$\text{Angular Accuracy } \bar{\theta} = \frac{180}{\pi} \arccos(|\mathbf{d}_{true} \cdot \mathbf{d}_{estimated}|) \quad (8)$$

where  $\mathbf{d}_{true}$  and  $\mathbf{d}_{estimated}$  represents the true and estimated fiber direction, respectively, and  $(\cdot)$  shows dot product.

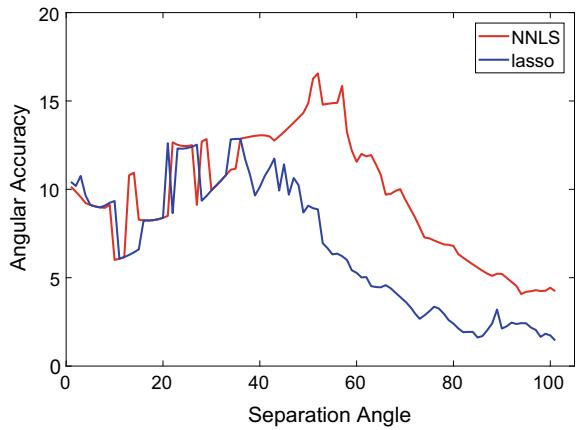
Figure 2 shows the angular accuracy for two crossing fiber cases. One fiber is having an angle of  $\phi_1 = 10^\circ$  and it is kept fixed. For other fiber angle  $\phi_2$ , we took 101 values starting from  $10^\circ$  to  $100^\circ$  with an equal interval of  $0.9^\circ$ . Polar angle was fixed at  $\theta_1 = \theta_2 = 90^\circ$ , which means fibers are in xy-plane. Angular accuracy with *lasso* method is comparatively less to *NNLS* when separation angles are more than  $40^\circ$  and less than  $80^\circ$ . We repeated the simulation for polar angles of  $\theta_1 = \theta_2 = 60^\circ$  and  $\theta_1 = 40^\circ, \theta_2 = 60^\circ$ . Results are displayed in Fig. 3 and Fig. 4, respectively. Here also the error behavior is similar to the previous case of  $\theta_1 = \theta_2 = 90^\circ$ . The performance of *lasso* algorithm is better for separation angles larger than  $40^\circ$ . Both the algorithms show similar results for separation angles less than  $40^\circ$  but for most of the cases, angular accuracy is not more than  $10^\circ$ .

We further did simulations in presence of Rician noise (explained in [10]). Results are presented for three different standard deviations of noise,  $\sigma = 0.01, 0.03$ , and  $0.05$ . The experiments are repeated 100 times for each case, mean and standard

**Fig. 3** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ, \phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = \theta_2 = 60^\circ$



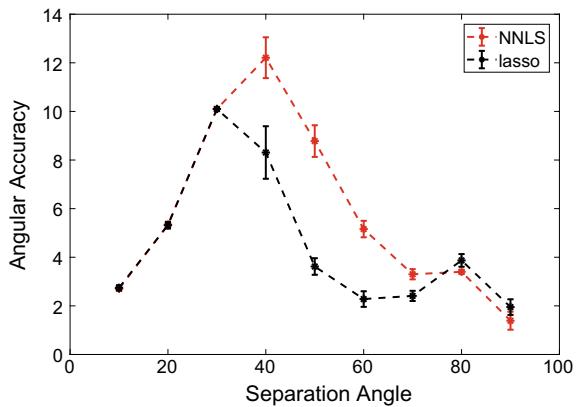
**Fig. 4** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ, \phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = 40^\circ, \theta_2 = 60^\circ$



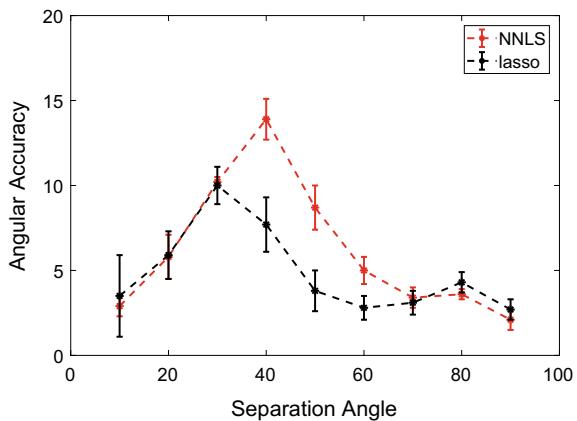
deviation of the angular accuracy are plotted. Figures 5, 6, and 7 show results for  $\sigma = 0.01$ ,  $\sigma = 0.03$ , and  $\sigma = 0.05$ , respectively. The azimuthal angles were chosen as  $\phi_1 = 10^\circ$  for all cases and  $\phi_2 = 20^\circ, 30^\circ, 40^\circ, 50^\circ, 60^\circ, 70^\circ, 80^\circ, 90^\circ, 100^\circ$ . Error is maximum when the separation angle is  $40^\circ$  for all the cases and the pattern is similar for  $\sigma = 0.01, 0.03$ . Both the algorithms give approximately same error when separation angle is less than  $40^\circ$  and greater than  $60^\circ$  while a significant difference is observed in angular accuracy for  $40^\circ \leq \bar{\theta} \leq 60^\circ$ . When we further increase the noise level in the data,  $\sigma = 0.05$ , it is observed that the angular accuracy is slightly more for some cases with *lasso* algorithm. However, the overall performance is better than *NNLS* method. Tuning the regularization parameter may further improve the results.

We are working to optimize the regularization parameter for its best performance for the given application. In future, we will also be applying this method on real data.

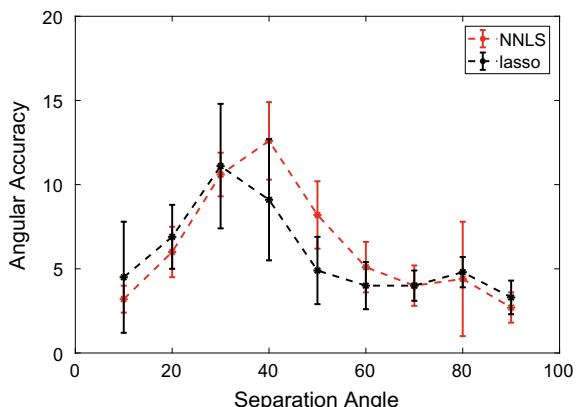
**Fig. 5** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ, \phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = \theta_2 = 90^\circ, \sigma = 0.01$



**Fig. 6** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ, \phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = \theta_2 = 90^\circ, \sigma = 0.03$



**Fig. 7** Angular accuracy (in degrees) with changing the separation angles of two crossing fibers,  $\phi_1 = 10^\circ, \phi_2 = 20^\circ, \dots, 100^\circ$  and  $\theta_1 = \theta_2 = 90^\circ, \sigma = 0.05$



## 4 Conclusion

Multi-fiber orientations are reconstructed with multi-compartmental Wishart distributed model using *NNLS* and *lasso* methods. Comparison of these two methods was done in terms of angular accuracy and it is observed that the *lasso* method outperforms the *NNLS* method. However, the computational time is increased with *lasso* but it was not the primary concern in the present analysis. We mainly worked to improve the angular accuracy.

**Acknowledgements** One of the authors Snehlata Shakya acknowledge the financial support as an Institute Postdoctoral Fellowship for carrying out this work.

## References

1. Bassier, P.J., Mattiello, J., LeBihan, D.: MR diffusion tensor spectroscopy and imaging. *Biophys. J.* **66**, 259–267 (1994)
2. Jones, D.K., Leemans, A.: Diffusion tensor imaging. *Methods Mol. Biol.* **11**, 127–144 (2011)
3. Mori, S., Zhang, J.: Principles of diffusion tensor imaging and its applications to basic neuroscience research. *Neuron* **51**, 527–539 (2006)
4. Mori, S., Barker, P.B.: Diffusion magnetic resonance imaging: its principle and applications. *Anat. Rec.* **257**, 102–109 (1999)
5. Luypaert, R., Boujraf, S., Sourbron, S., Osteaux, M.: Diffusion and perfusion MRI: basic physics. *Eur. J. Radiol.* **38**, 19–27 (2001)
6. Inglis, B.A., Bossart, E.L., Buckley, D.L., Wirth, E.D., Mareci, T.H.: Visualization of neural tissue water compartments using biexponential diffusion tensor MRI. *Magn. Reson. Med.* **45**(4), 580–587 (2001)
7. Tuch, D.S., Reese, T.G., Wiegell, M.R., Makris, N., Belliveau, J.W., Wedeen, V.J.: High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Mag. Reson. Med.* **48**, 577–582 (2002)
8. Jian, B., Vemuri, B.C., Özarslan, E., Carney, P.R., Mareci, T.H.: A novel tensor distribution model for the diffusion-weighted MR signal. *NeuroImage* **37**, 164–176 (2007)
9. Kumar, R., Vemuri, B.C., Wang, F., Syeda-Mahmood, T., Carney, P.R., Mareci, T.H.: Multi-fiber reconstruction from DW-MRI using a continuous mixture of hyperspherical von Mises-Fisher distributions. *Inf. Process. Med. Imaging* **5636**, 139–150 (2009)
10. Shakya, S., Batool, N., Özarslan, E., Knutsson, H.: Multi-fiber reconstruction using probabilistic mixture models for diffusion MRI examinations of the brain. *Modeling, Analysis, and Visualization of Anisotropy. Mathematics and Visualization*, pp. 283–308. Springer, Cham (2017)
11. Lawson, C.L., Hanson, R.J.: Solving Least-Squares Problems, p. 161. Prentice Hall, Upper Saddle River, NJ (1974)
12. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–88 (1996)
13. Mayerhofer, E.: On the existence of non-central Wishart distributions. *J. Multivar. Anal.* **114**, 448–456 (2013)
14. <http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>
15. Saini, D., Kumar, S., Gulati, T.R.: Reconstruction of freeform space curves using NURBS-snakes based energy minimization approach. *Comput. Aided Geom. Des.* **33**, 30–45 (2015)
16. Shakya, S., Saxena, A., Munshi, P., Goswami, M.: Adaptive discretization for computerized tomography. *Res. Nondestr. Eval.* **29**(2), 78–94 (2018)

17. Daducci, A., Canales-Rodriguez, E.J., Descoteaux, M., Garyfallidis, E., Gur, Y., Lin, Y.-C., Mani, M., Merlet, S., Paquette, M., Ramirez-Manzanares, A., Reisert, M., Rodrigues, P., Sepehrband, F., Caruyer, E., Choupan, J., Deriche, R., Jacob, M., Menegaz, G., Prckovska, V., Rivera, M., Wiaux, Y., Thiran, J.-P.: Quantitative comparison of reconstruction methods for intra-voxel fiber recovery from diffusion MRI. *IEEE Trans. Med. Imaging* **33**(2), 384–399 (2014)

# Design of Finite Impulse Response Filter with Controlled Ripple Using Cuckoo Search Algorithm



Anil Kumar, N. Agrawal and I. Sharma

**Abstract** In this paper, an efficient design of finite impulse response (FIR) filter is presented with improved fitness function using cuckoo search algorithm (CSA). CSA is recently proposed evolutionary technique (ET), which has efficient ability of exploration and therefore used in FIR filter design. The fitness function is constructed in the frequency domain as a mean squared error (MSE) between the designed and desired response. In this fitness function, tolerable limits for magnitude response in passband and stopband region have been embedded, which helps in gaining the controlled ripple in irrespective bands. The designed filters are realized on general-purpose microcontroller using Arduino platform and filter performance is tested using fidelity parameters, which are; passband error ( $Er_{pb}$ ), stopband error ( $Er_{sb}$ ), and minimum stopband attenuation ( $A_s$ ). The exhaustive experimental analysis confirms that proposed methodology is statically stable and obtains improved fidelity parameters when compared with previous state of the art.

**Keywords** Finite impulse response (FIR) · Evolutionary techniques (ETs) · Controlled ripple

## 1 Introduction

Finite impulse response (FIR) filters are exhaustively used in digital signal processing application due to their absolute stability and linear phase property [1]. FIR filters are prominently designed using windowing techniques due to their ease of implementation with suitable transfer function. However, the obtained solution is

---

A. Kumar (✉) · N. Agrawal · I. Sharma  
PDPM Indian Institute of Information Technology,  
Design and Manufacturing, Jabalpur, Madhya Pradesh 482005, India  
e-mail: [anilkdee@gmail.com](mailto:anilkdee@gmail.com)

N. Agrawal  
e-mail: [nikhil.agrawal@iiitdmj.com](mailto:nikhil.agrawal@iiitdmj.com)

I. Sharma  
e-mail: [ila.sharma23@gmail.com](mailto:ila.sharma23@gmail.com)

suboptimal and suitable for higher order filter only. Another approach is gradient-based methods in which a differentiable error function with respect to filter tape is constructed and various constraints are employed for the desired filter requirements [2]. Since the frequency response of FIR filter involves the polynomial of trigonometric function, which makes the error function as multimodal error function. Thus, gradient-based methods too result in suboptimal solution due to quick convergence. Therefore, several evolutionary techniques (ETs) are used for FIR filter design.

In recent scenario, optimization methods that are inspired by some nature mechanism have been abruptly used in solving multidimensional, non-differential, and multimodal functions [3, 4]. Therefore, these methods are also used in digital filter design, where an error function is formed and minimized [5–8]. Initially, the genetic algorithm (GA) was used in filter design, later on, many variants has been invented with improved results, but, due to so many intermediated functions and large evaluation, GA suffers from poor convergence [9–11]. Another ET, known as particle swarm optimization (PSO) has been extensively used due to its simple mechanism of exploration [12, 13]. While, artificial bee colony (ABC) algorithm has been also used in filter design due to its deep search mechanism [14]. After that, the concept of hybrid techniques using two ET has been practiced such as differential evolution particle swarm optimization (DEPSO), Hybrid-PSO and Hybrid-QPSO for digital filter, and filterbank design [7, 15, 16].

It is evident from the literature that various ETs emerge as an effective tool for designing a digital filter. Recently, another ET named as cuckoo search algorithm (CSA), inspired by egg-laying mechanism of cuckoo bird has been developed and found effective, when compared with the above-discussed techniques [17, 18]. The depth of the exploration is controlled by a parameter known as probability of laying cuckoo egg ( $P_\alpha$ ), and its value also controls the rate of convergence. Therefore, the efficient use of search space is possible in CSA and that's why it has been used in numerous engineering applications of signal processing such as image enhancement, electroencephalography (EEG) signal filtering, filter design [17–20]. Therefore, in this paper, efficient digital FIR filters are designed by using CSA with controlled ripple. For this purpose, a new fitness function is adopted, where margin for acceptable ripple is introduced, so that the stopband attenuation is improved.

## 2 Overview of Evolutionary Techniques (ETs)

ETs are the optimization methods, which have been formulated by the distinct behavior of different species or nature-inspired biological activities. In all of the ETs, the common principle is the formulation of a search space matrix ( $X$ ), whose either row or column vector is chosen as a possible solution. The feasibility of solution is measured on the basis of an error function known as fitness function. This search

space matrix is continuously updated by different principle equations of respective ET. The solutions of  $X$ , which gained improved fitness, are considered as local best solutions ( $\mathbf{PB}$ ), while the solution from  $\mathbf{PB}$ , which has best fitness values till current time is considered as global best solution ( $\mathbf{gb}$ ). For example; the Darwin's theory has been modeled for optimization and called as differential evolution (DE) optimization algorithm [21], genetic activities of crossover and mutation occurred in the formulation of improved cell has adopted for the formulation of genetic algorithm (GA) [3]. Similarly, the communication behavior of insects, birds, and fishes has been studied and formulated as particle swarm optimization (PSO) algorithm [22]. On the other side, the colonial work strategy of honeybee has been used for artificial bee colony (ABC) algorithm [15]. Moreover, numerous ETs have been developed using various nature-inspired phenomena like bacterial foraging optimization (BFO), flower poly optimization (FPO), and firefly optimization (FFO). Besides, the hybrid techniques, which are formed by merging the concept of two ETs have resulted in getting improved solution [7, 23, 24]. An impressive and detailed description of ETs is reported in [25–27].

Recently, the cuckoo search algorithm (CSA) has been attracted by many researcher's interest due to their fast exploration capabilities. CSA is based on the study on the egg-laying tactic of certain birds in different nests in order to capture more space, identification of other bird egg and its disposal, and proved as ground-breaking in optimization method among different ETs.

### 3 Problem Formulation

This section describes the design mechanism of FIR filter in frequency domain. The transfer function of FIR filter is expressed as [13]

$$H(e^{j\omega}) = \sum_{n=0}^N h(n) \cdot e^{-nj\omega}, \quad (1)$$

where  $h(n)$  is the impulse response of FIR filter,  $N$  is the order of filter and therefore, and  $N + 1$  is the length of its impulse response. Based on the impulse response symmetry and its length, these FIR filters are categorized into four types. In this paper, odd length and symmetric impulse response FIR filter are designed, which is known as Type-I filter. Due to symmetric impulse response that is  $h(n) = h(N - n)$ , FIR filter transfer function becomes [13]

$$H(e^{j\omega}) = e^{\frac{-jN\omega}{2}} \cdot G(e^{j\omega}) \quad (2)$$

and

$$G(e^{j\omega}) = h\left(\frac{N}{2}\right) + 2 \sum_{k=1}^{\frac{N}{2}} h\left(\frac{N}{2} - k\right) \cdot \cos(\omega k). \quad (3)$$

The above expression can be expressed in vector form as

$$G(e^{j\omega}) = \mathbf{a}^T \cdot \mathbf{c}(\omega), \quad (4)$$

where

$$\mathbf{a} = [h(N/2) \ h(N/2 - 1) \ h(N/2 - 2) \ \dots \ h(0)] \quad (5)$$

and

$$\mathbf{c}(\omega) = [\cos(0\omega) \ 2 \cdot \cos(\omega) \ \dots \ 2 \cdot \cos\left(\frac{N}{2}\omega\right)] \quad (6)$$

The task is to find the appropriate coefficient vector ( $\mathbf{a}$ ) so that the response of FIR filter should be close to the desired magnitude response defined as

$$H_o(e^{j\omega}) = \begin{cases} 1 \pm \delta_p, & \omega \in \text{passband} \\ \delta_s, & \omega \in \text{stopband} \end{cases}, \quad (7)$$

where  $\delta_p$  is the passband ripple and  $\delta_s$  is the stopband ripple. To achieve this objective, fitness function is minimized using CSA.

### 3.1 Formulation of an Objective Function

In this paper, a fitness function is constituted, which allows the magnitude to swing between the tolerable limits. From the literature, it has been observed that fitness functions are developed with tolerable limits in passband (*pb*) and stopband (*sb*) region, however, magnitude response is restricted to swing between  $\delta_p$  and 1 [9]. The authors in [16, 18] have proposed the fitness function in which magnitude response is allowed to swing in the limits of  $1 \pm \delta_p$  and  $\delta_s$  in passband and stopband region, respectively. The fitness employed for FIR filter design is defined as [16, 18]

$$e_{pb}(\omega) = \begin{cases} (|H(e^{j\omega})| - (1 + \delta_p))^2, & \text{if } |H(e^{j\omega})| > 1 + \delta_p \\ (|H(e^{j\omega})| - (1 - \delta_p))^2, & \text{if } |H(e^{j\omega})| < 1 - \delta_p, \omega \in pb \\ 0, & \text{if } 1 - \delta_p \leq |H(e^{j\omega})| \leq 1 + \delta_p \end{cases} \quad (8)$$

Now, the error computed in *sb* region as

$$e_{sb}(\omega) = \begin{cases} (|H(e^{j\omega})| - \delta_s)^2, & \text{if } |H(e^{j\omega})| > \delta_s, \omega \in sb \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The fitness function is computed as

$$fit = \frac{1}{NS_{pb}} \sum e_{pb}(\omega) + \frac{1}{NS_{sb}} \sum e_{sb}(\omega), \quad (10)$$

where  $NS_{pb}$  and  $NS_{sb}$  are the number of frequency samples in passband and stopband region, respectively.

### 3.2 Design Procedure of FIR Filter Using CSA

The design of FIR using CSA and proposed fitness function is conducted in the following steps:

1. Initialization of  $X$  with uniformly distributed random number in the range of –1 to 1.
2. Each solution vector from  $X$  is used for frequency response evaluation with  $NS$  sample frequency response using Eq. (4), and the computation of fitness is done using Eqs. (8)–(10).
3. The solution vector having the best *fit* value is considered as  $gb$  and then perform  $PB = X$ .
4. Update  $X$  using updated Eq. of CSA [27]:

$$X^{k+1} = X^k + \alpha \oplus Lévy, \quad (11)$$

where  $\alpha$  is a scaling factor and *Lévy* represents Lévy flight mechanism.

5. Compute the fitness of updated  $X$  using Step 2.
6. Greedy-based selection is conducted for finding the improved solution vectors from updated  $X$ . If any solution vector *fit* value is improved then it is included in  $PB$ .
7. Elimination of elements of  $X$  is conducted, which is an equivalent operation of detection of egg by host bird. The identification of elements is governed by a control factor  $P_\alpha$ .

8. The solution from ***PB***, which has the best *fit* value, is sort out and its *fit* value is better than existing ***gb*** solution, then update the current ***gb*** solution.
9. Repeat from Steps 4–8 until iteration count reaches to termination or desired *fit* values is acquired.

## 4 Results and Discussion

Some benchmark design examples of FIR filter are considered from earlier state-of-the-art methods. The specifications of filters are mentioned in Table 1, which are taken from the previous state of the art and used for testing of proposed design approach. In CSA, tuning of control parameter is an essential task, which is step-size ( $\alpha$ ) and the probability of dropping of element ( $P_\alpha$ ). Parameter  $P_\alpha$  controls the depth of exploration, which also affects the convergence rate. If its value is too small then exploration is deeper; however, convergence will be slow and vice versa for too high value. Therefore, the authors in [8], have carried exhaustive experimental analysis and reported that the value of  $P_\alpha = 0.05$  is a suitable choice for digital filter design and the same value is considered in this paper. The effect of  $\alpha$  is not varied and therefore, it is taken as 1. The other design parameters such as population size = 30, maximum number of iterations = 2500, upper limit of elements = 1, and upper limit of elements = -1.

In the first experiment, the statically stability of the proposed method is studied. Each design example of Table 1 is designed using the proposed method for 30 times. The efficiency of the designed filter is tested on the basis of the following fidelity parameters:

$$\text{Passband error: } Er_{pb} = \frac{1}{N_{pb}} \sum_{\omega \in pb} e_{pb}(\omega), \quad (12)$$

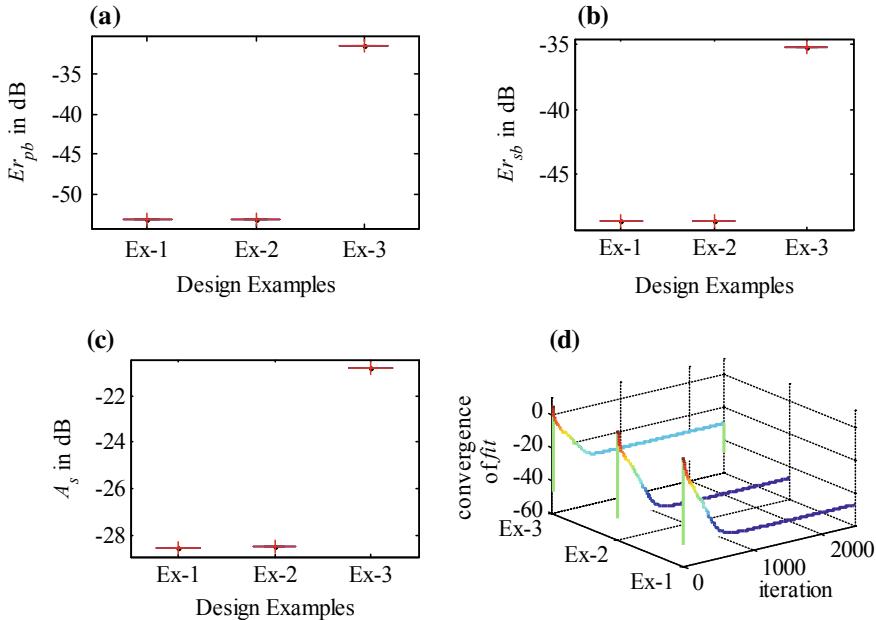
$$\text{Stopband error: } Er_{sb} = \frac{1}{N_{sb}} \sum_{\omega \in sb} e_{sb}(\omega), \quad (13)$$

and

$$\text{Minimum Stopband attenuation: } A_s = 20 \log 10(|H(e^{j\omega})|)|_{\omega=\omega_s}. \quad (14)$$

**Table 1** FIR filter design specifications

Filter type	$N$	$pb$	$sb$	$\delta_p$	$\delta_s$
LPF [28, 29]	20	[0.00, 0.45]	[0.55, 1.00]	0.10	0.01
HPF [28]	20	[0.55, 1.00]	[0.00, 0.45]	0.10	0.01
HPF [6]	20	[0.48, 1.00]	[0.00, 0.38]	–	–



**Fig. 1** **a** Variation in  $Er_{pb}$ , **b** variation in  $Er_{sb}$ , and **c** variation in  $A_s$ , obtained for 30 experimental trials for design examples specification mentioned in Table 1, **d** mean of the convergence of fit function obtained for 30 experimental trials for design examples specification mentioned in Table 1

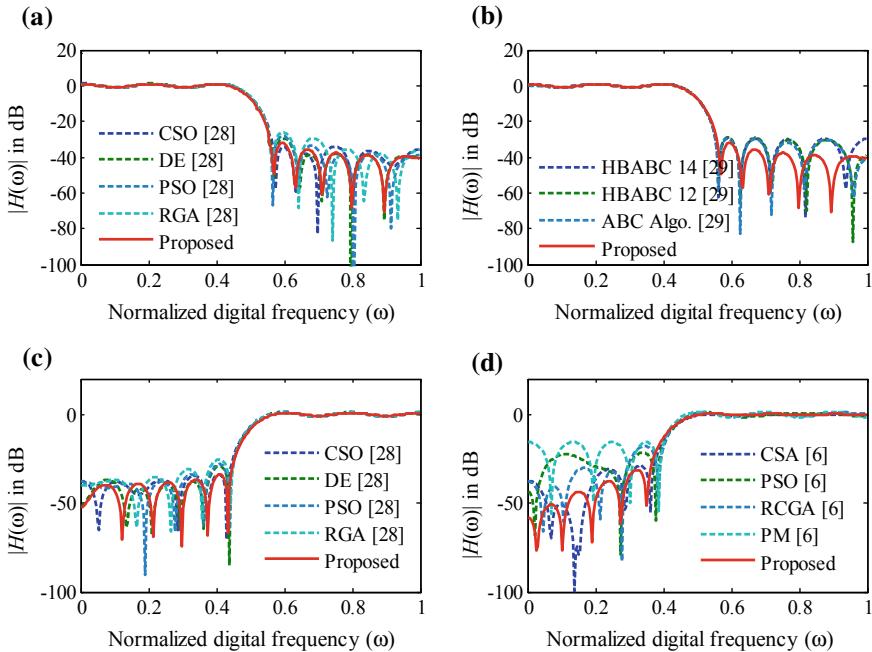
The variation in fidelity parameter is shown in Fig. 1, in which Fig. 1a depicts the variation in values of  $Er_{pb}$  obtained for 30 trials of design. Similarly, the variations of  $Er_{sb}$  and  $A_s$  have been shown in Fig. 1b and c, respectively. The convergence of all three examples is shown in Fig. 1d. The best values of fidelity parameters obtained are mentioned in Table 2 along with the fidelity parameters of the previously designed FIR filters. FIR filters designed using the proposed methodology acquires improved values, however, there is a slight reduction in values of  $A_s$ . Whereas, the frequency responses of the designed FIR filter have been plotted along with the frequency responses as in earlier reported filters and is shown in Fig. 2.

## 5 Implementation of Filter on Arduino

The designed filter is realized on an Arduino MEGA 256 board. The obtained filter coefficients are embedded in FIR direct form structure as shown in Fig. 3a. Arduino supports only positive analog values ranging from 0 to 5 volts, which can be converted into digital value with 10-bit precision. Therefore, for testing purpose, digital sinusoidal swept signal swinging between 1 and -1 is applied at input of the filter and a constant value is added for biasing so that the range out will vary from 0 to

**Table 2** Comparative analysis fidelity parameters of the proposed method with previous techniques

Technique	Filter type	Parameters						$A_s$
		$er_{pb}$ (dB)	$er_{sb}$ (dB)	$er_{pb}^{\max}$	$er_{sb}^{\max}$	$\hat{\delta}_{pb}$	$\hat{\delta}_{sb}$	
RGA [28]	LPF	-45.872	-37.267	0.014	0.039	0.114	0.049	-26.716
DE [28]	LPF	-39.152	-42.121	0.036	0.043	0.136	0.053	-25.457
PSO [28]	LPF	-42.274	-40.741	0.023	0.035	0.123	0.045	-26.963
CSO [28]	LPF	-35.971	-42.275	0.064	0.054	0.164	0.064	-23.936
HBABC14 [29]	LPF	-83.538	-36.727	$6.7 \times 10^{-4}$	0.025	0.100	0.0353	-30.984
HBABC12 [29]	LPF	-69.015	-37.630	0.002	0.0242	0.102	0.0342	-30.601
ABC algorithm [29]	LPF	-60.806	-38.259	0.004	0.0232	0.103	0.0332	-31.130
CSO [28]	HPF	-40.414	-42.614	0.032	0.003	0.132	0.063	-23.991
DE [28]	HPF	-38.215	-42.624	0.037	0.001	0.137	0.042	-27.462
PSO [28]	HPF	-41.099	-40.862	0.025	0.001	0.124	0.046	-26.719
RGA [28]	HPF	-45.513	-36.950	0.019	0.002	0.119	0.055	-28.404
CSA [6]	HPF	-25.799	-34.848	0.152	0.004	0.010	0.060	-24.458
PSO [6]	HPF	-21.512	-26.232	0.158	0.007	0.119	0.083	-60.152
RCGA [6]	HPF	-22.089	-26.290	0.116	0.016	0.113	0.127	-27.078
PM [6]	HPF	-18.534	-18.510	0.166	0.027	0.166	0.166	-23.691
CSA proposed	LPF	-65.742	-45.628	0.005	0.032	0.100	0.042	-27.361
HPF		-48.858	-47.060	0.034	0.001	0.093	0.043	-27.263
HPF		-31.441	-35.213	0.132	0.008	0.047	0.091	-20.839

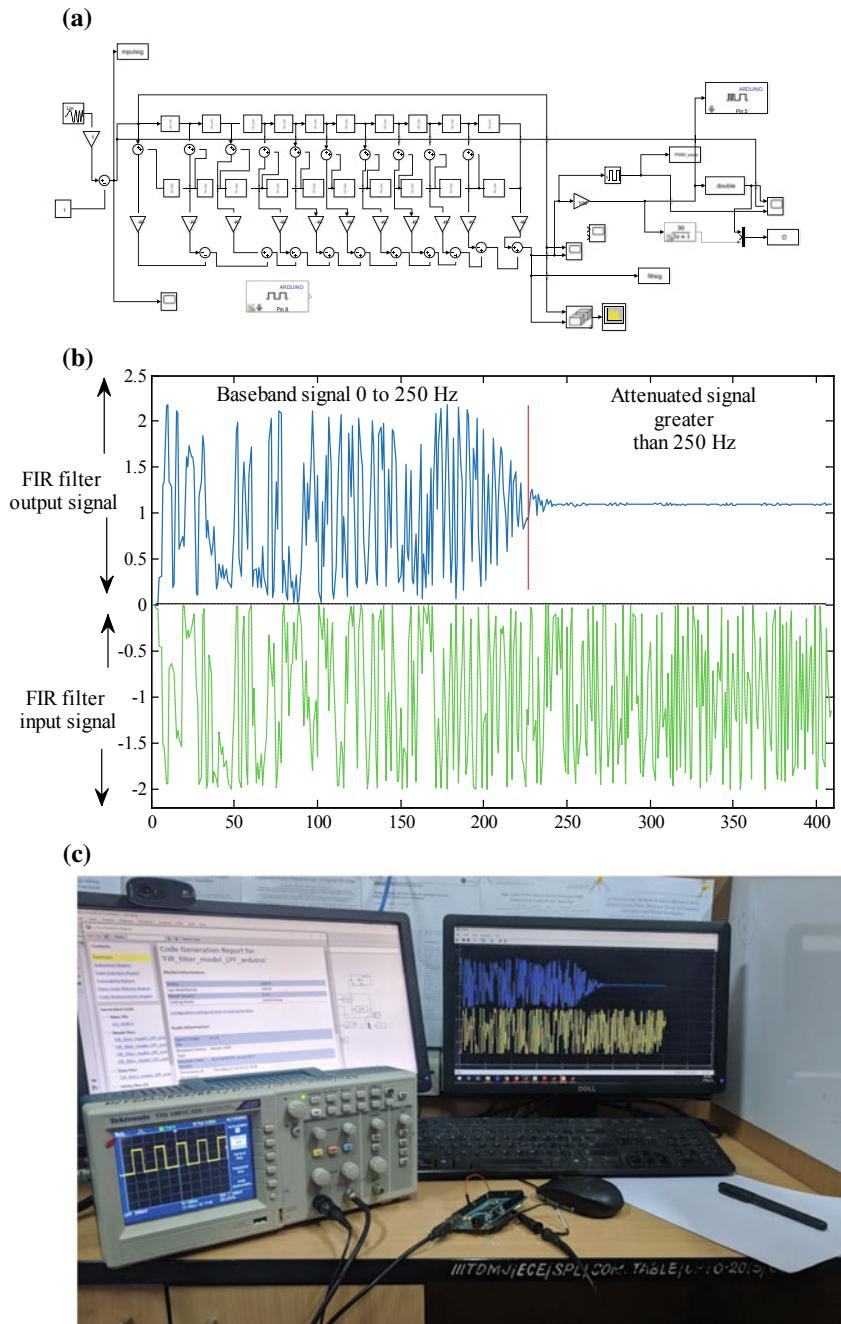


**Fig. 2** **a** Magnitude response of LPF designed by fitness function using various ETs used in [28] and proposed technique, **b** Magnitude response of LPF designed by fitness function given in [28] with ABC algorithm, HBABC algorithm, and proposed technique, **c** magnitude response of HPF designed by fitness function proposed in [28] with various ETs and proposed technique, and **d** magnitude response of HPF designed by fitness function proposed in [6], proposed technique

2. The filter response is verified by observing Fig. 3b. Arduino MEGA 256 does not have digital to analog converter; however, there is an arrangement provided to get pulse width modulated (PWM) signal for appropriated filter signal. Either this PWM can be used for getting analog output or the use of digital to analog converter will provide the appropriate analog out; however, this is not yet considered as this paper presents a possible implementation of the designed optimal FIR filter. The experimental setup is depicted in Fig. 3c.

## 6 Conclusion

In this paper, an efficient design of FIR filter using CSA with improved fitness functions is presented. The fitness function, which is the measure of the closeness of the achieved response, is constructed in the frequency domain as mean squared error of designed and desired response. In desired response, there is a margin of tolerance in passband and stopband region. The fitness function is formed such that the designed



**Fig. 3** **a** Simulink model of FIR filter used for realization using Arduino MEGA 256, **b** FIR filter response in discrete-time domain, and **c** experimental setup

magnitude response swings in the limits of the desired region, which helps in getting controlled ripple. It has found that CSA with tuned value of controlled parameters with improved fitness function has obtained better fidelity parameter values, when compared with previous techniques. Finally, the designed FIR filter is realized on the general-purpose microcontroller using Arduino platform.

**Acknowledgements** This work is supported in part by the Department of Science and Technology, Govt. of India under Grant No. SB/S3/EECE/0249/2016.

## References

1. Schlichthärle, D.: Digital Filters—Basics and Design, 2nd edn. Springer, Berlin Heidelberg (2000)
2. Çiloğlu, T.: An efficient local search method guided by gradient information for discrete coefficient FIR filter design. *Sig. Process.* **82**(10), 1337–1350 (2002)
3. Man, K.F., Tang, K.S., Kwong, S.: Genetic algorithms: concepts and applications in engineering design. *IEEE Trans. Ind. Electron.* **43**(5), 519–534 (1996)
4. N. Agrawal, A. Kumar, V. Bajaj, G.K. Singh, Design of bandpass and bandstop infinite impulse response filters using fractional derivative. *IEEE Trans. Ind. Electron.* 1–11 (2018). <https://doi.org/10.1109/tie.2018.2831184>
5. Reddy, K.S., Sahoo, S.K.: An approach for FIR filter coefficient optimization using differential evolution algorithm. *AEU – Int. J. Electron. Commun.* **69**(1), 101–108 (2015)
6. Aggarwal, A., Rawat, T.K., Upadhyay, D.K.: Design of optimal digital FIR filters using evolutionary and swarm optimization techniques. *AEU – Int. J. Electron. Commun.* **70**(4), 373–385 (2016)
7. Agrawal, N., Kumar, A., Bajaj, V.: Design of digital IIR filter with low quantization error using hybrid optimization technique. *Soft. Comput.* **22**(9), 2953–2971 (2017)
8. Agrawal, N., Kumar, A., Bajaj, V., Singh, G.K.: High order stable infinite impulse response filter design using cuckoo search algorithm. *Int. J. Autom. Comput.* **14**(5), 589–602 (2017)
9. Tang, K.-S., Man, K.-F., Kwong, S., Liu, Z.-F.: Design and optimization of IIR filter structure using hierarchical genetic algorithms. *IEEE Trans. Ind. Electron.* **45**(3), 481–487 (1998)
10. N. Karaboga, B. Cetinkaya, Design of minimum phase digital IIR filters by using genetic algorithm, in *6th Proceedings of the Nordic Signal Processing Symposium, NORSIG 2004* (IEEE, Espoo, 2004), pp. 29–32
11. Yu, Y., Xinjie, Y.: Cooperative coevolutionary genetic algorithm for digital IIR filter design. *IEEE Trans. Ind. Electron.* **54**(3), 1311–1318 (2007)
12. Ababneh, J.I., Bataineh, M.H.: Linear phase FIR filter design using particle swarm optimization and genetic algorithms. *Digit. Signal Proc.* **18**(4), 657–668 (2008)
13. Sharma, I., Kuldeep, B., Kumar, A., Singh, V.K.: Performance of swarm based optimization techniques for designing digital FIR filter: a comparative study. *Eng. Sci. Technol. Int. J.* **19**(3), 1564–1572 (2016)
14. Karaboga, N., Kalinli, A., Karaboga, D.: Designing digital IIR filters using ant colony optimisation algorithm. *Eng. Appl. Artif. Intell.* **17**(3), 301–309 (2004)
15. B. Luitel, G.K. Venayagamoorthy, Differential evolution particle swarm optimization for digital filter design, in *2008 IEEE Congress on Evolutionary Computation* (IEEE, Hong Kong, 2008), pp. 3954–3961
16. Agrawal, N., Kumar, A., Bajaj, V., Lee, H.-N.: Controlled ripple based design of digital IIR filter, in *21st IEEE International Conference on Digital Signal Processing (DSP)* (IEEE, Beijing, 2016), pp. 627–631

17. Ahirwal, M.K., Kumar, A., Singh, G.K.: EEG/ERP adaptive noise canceller design with controlled search space (CSS) approach in cuckoo and other optimization algorithms. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **10**(6), 1491–1504 (2013)
18. Agrawal, N., Kumar, A., Bajaj, V.: Digital IIR filter design with controlled ripple using cuckoo search algorithm, in *2016 International Conference on Signal and Information Processing (IConSIP)* (IEEE, Vishnupuri, 2016), pp. 1–5
19. Kumar, M., Rawat, T.K.: Optimal fractional delay-IIR filter design using cuckoo search algorithm. *ISA Trans.* **59**, 39–54 (2015)
20. Gotmare, A., Patidar, R., George, N.V.: Nonlinear system identification using a cuckoo search optimized adaptive Hammerstein model. *Expert Syst. Appl.* **42**(5), 2538–2546 (2015)
21. Liu, G., Li, Y., He, G.: Design of digital FIR filters using differential evolution algorithm based on reserved genes, in *IEEE Congress on Evolutionary Computation* (IEEE, Barcelona, 2010), pp. 1–7
22. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization: an overview. *Swarm Intell.* **1**(1), 33–57 (2007)
23. Rafi, S.M., Kumar, A., Singh, G.K.: An improved particle swarm optimization method for multirate filter bank design. *J. Franklin Inst.* **350**(4), 757–769 (2013)
24. Sidhu, D.S., Dhillon, J.S., Kaur, D.: Hybrid heuristic search method for design of digital IIR filter with conflicting objectives. *Soft. Comput.* **21**(12), 3461–3476 (2016)
25. Gotmare, A., Bhattacharjee, S.S., Patidar, R., George, N.V.: Swarm and evolutionary computing algorithms for system identification and filter design: a comprehensive review. *Swarm Evol. Comput.* **32**, 68–84 (2017)
26. Agrawal, N., Kumar, A., Bajaj, V.: A new design method for stable IIR filters with nearly linear-phase response based on fractional derivative and swarm intelligence. *IEEE Trans. Emerg. Top. Comput. Intell.* **1**(6), 464–477 (2017)
27. Baderia, K., Kumar, A., Singh, G.K.: Hybrid method for designing digital FIR filters based on fractional derivative constraints. *ISA Trans.* **58**, 493–508 (2015)
28. Saha, S.K., Ghoshal, S.P., Kar, R., Mandal, D.: Cat swarm optimization algorithm for optimal linear phase FIR filter design. *ISA Trans.* **52**(6), 781–794 (2013)
29. Dwivedi, A.K., Ghosh, S., Londhe, N.D.: Low-power FIR filter design using hybrid artificial bee colony algorithm with experimental validation over FPGA. *Circuits Syst. Signal Process* **36**(1), 156–180 (2017)

# Robustly Clipped Sub-equalized Histogram-Based Cosine-Transformed Energy-Redistributed Gamma Correction for Satellite Image Enhancement



Himanshu Singh, Anil Kumar and L. K. Balyan

**Abstract** In this paper, a new proposal is reported for image quality enhancement. Here, statistically clipped, bi-histogram equalization-based adaptive gamma correction along with its cosine-transformed energy redistribution is introduced for improvement of low-contrast dark images. This approach computes the clipping limit adaptively by observing stretched histogram bins, mean, and median values for each sub-histogram. Considering the clipping limit as the lowest of these three values, this limit ensures the conservation of information content of the image to a great extent. This adaptive clipping limit selection also resolves the issue of over-emphasization of high-frequency bins during sub-histogram equalization. For harvesting more information and better illumination, gamma correction is imparted by using the adaptive gamma value-set. The corresponding gamma value-set is itself derived from the previously sub-equalized interim intensity channel. In addition to this, two-dimensional (2D) discrete cosine transformation (DCT) for gamma-corrected channel is also employed for incorporating the energy-based textural enhancement framework. Validation is performed here by analyzing the enhancement for various remotely sensed dark satellite images by evaluating standard performance indices.

**Keywords** Adaptive gamma correction · Bi-histogram equalization · Image quality enhancement · Optimal histogram clipping · Remotely sensed images

## 1 Introduction

Digital satellite imagery is growing as an obligatory basis for information gathering in various contemporary applications nowadays [1]. In all kinds of hand-held electronic devices at the consumer end having image capturing feature and auto-focusing property are highly desired. In robotics, object tracking purpose can be resolved by image processing. Imaging and corresponding quality enhancing are highly indispensable for human-computer interfacing as well as for gesture recognition during

---

H. Singh (✉) · A. Kumar · L. K. Balyan

Indian Institute of Information Technology Design and Manufacturing, Jabalpur, India  
e-mail: [himanshu.iiitj@gmail.com](mailto:himanshu.iiitj@gmail.com)

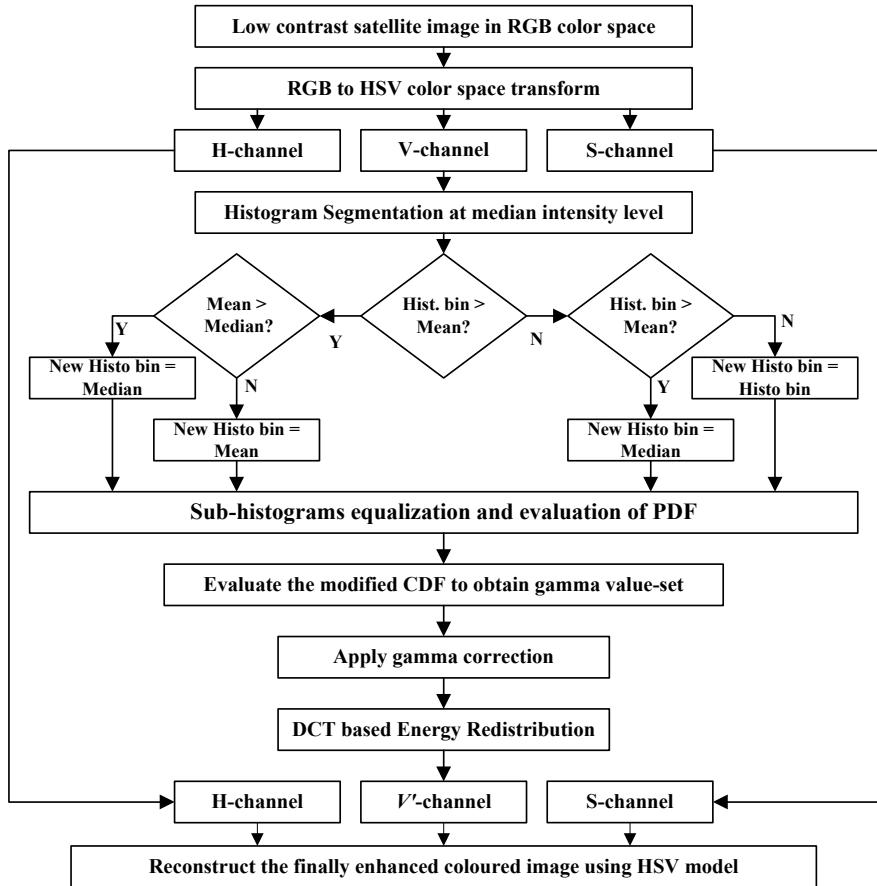
automated operation of the machines. It is highly desired to restore the captured low-quality image from its degradation; and when it comes to remotely sensed, poorly illuminated dark satellite images, the each and every fine detail is of prime concern. Especially, as the remotely captured satellite images usually cover a very large geographical area hence, information content per patch is very high and, a very adaptive and highly robust quality enhancement is desired. Also, textural details of these satellite images are of very high concern, so along with intensity-based enhancement, textural roughness/smoothness-based processing is also required [2, 3]. For enhancement of such kind of images, researchers shifted their orientation toward general histogram equalization (GHE) [4] mainly because no more complexities are associated [5–10] with in its understanding. It is easy to conclude that GHE is quite unable to preserve the local spatial features of image and hence, inefficient for imparting appreciable quality improvement of the image. Due to this, the sole attention of the researchers got shifted to histogram distribution along with local histogram modifications and also toward their corresponding advantages. In the same run, along with fuzzy inspired smoothening for histogram which is followed by peaks-based histogram subdivision is also introduced which is termed as brightness preserving dynamic fuzzy HE (BPDFHE) [6]. Later on, exposure-based sub-image HE (ESIHE) [7] was introduced, where exposure of the images is calculated, so that on the basis of it, histogram subdivision is imparted along with histogram sub-equalization. Next to it, median–mean-dependent sub-image-clipped HE (MMSICHE) [8] is suggested for image enhancement. Here, along with further bisecting both the sub-histograms on the basis of median count of the corresponding pixels in both sub-histograms, clipping is also performed before sub-equalization. In the same sequence, recursive-ESIHE (R-ESIHE) [9] was suggested as sequential iterative implications of the EISHE until the exposure of the image is going to attain a predefined threshold value. Similarly, a multilevel histogram separation is also suggested, which is termed as recursively separated-ESIHE (RS-ESIHE) [9]. In the same sequence, various other histogram-based approaches are also proposed by associating various intelligent pipelined blocks like averaging histogram equalization (AVGHEQ) [10], HE-based optimal profile compression (HEOPC) [11] and finally, the HE with maximum intensity coverage (MAXCOVER) [12] was also proposed for overall quality enhancement and corresponding better image visualization. Gamma correction (GC), in one form or the other, is one of the most popular and powerful mechanisms for image illumination correction and its efficient variations revolutionize the entire work spirit among most of the image processing researchers. By obtaining a more adaptive gamma value-set containing individual gamma values for all intensity levels, adaptive GC with weighting distribution (AGCWD) [13] is also proposed. Further, various efficient variants of the GC [14–19] are also suggested for imparting more robustness in this context. By taking the advantage of the unsharp masking filters (UMF), the intensity and edge-based adaptive UMF (IEAUMF) [20] is also proposed in the same context for efficient incorporation of the edge augmentation. Sigmoidal curve-based cosine-transformed Regularized-HE [21] is also proposed for highlighting the texture. In the same sequence, although several kinds of enhancement methodologies are proposed till date for widely diverse characteristics

of images from various domains, (contextual literature survey is explicitly presented in [1, 2]), still most of them are lagging when it comes to the matter of enhancement of different domain images through a single approach. The rest of the manuscript is organized as follows: Sect. 2 deals with the problem formulation and the proposed methodology. Performance evaluation and comparison-based experimental results are presented in Sect. 3 and finally, conclusions are drawn in Sect. 4.

## 2 Proposed Methodology

While processing multiband or multispectral satellite images, individual bands need similar and parallel quality enhancement of each channel. Here, for simplicity, RGB color satellite images are taken into account. For basic color image processing, usually chromatic as well as non-chromatic information content is isolated so that only intensity channel can be processed for imparting overall image enhancement. Here, the prime objective is to enhance the V-channel for which first of all histogram is derived, and then it is divided on the basis of its median value. Thus obtained sub-histograms are comprised of equal number of bin values; and hence, such type of division is highly appreciated. For proper avoidance of unbalanced bin-value distribution or resolving the issue of over-emphasization of high-frequency bins during sub-equalization, optimal clipping limit is derived by taking the lowest of all three values (i.e., mean count, median count, and corresponding bin value). Afterward, proper subsequent sub-equalization can be imposed, and this adaptively equalized histogram itself is used to derive the desired gamma value-set using its corresponding CDF value-set (just by subtracting it from unity). Hence, highly adaptive gamma value-set has been derived for subsequent gamma correction, and thus overall quality enhancement for the V-channel can be done. The entire methodology is comprised of some basic operations in an organized manner that leads to contrast and entropy content enhancement specifically for dark satellite images. For dark images, where most of the pixels are having very low-intensity values; some amount of intensity-level boosting is also required. If the gamma value-set is derived directly from them, it may lead to saturation and unnatural artifacts. Keeping this in mind, first optimal clipping, and then sub-equalization has been imposed here before applying the gamma correction, due to which the derived gamma value-set is highly adaptive, and leads to proper enhancement. In case of dark image enhancement, some amount of mean brightness enhancement is also desired for properly increased contrast evaluation. The entire methodology is a single-step process (lacking any iterative step; and hence less complex), and can be explained using process flow diagram as shown in Fig. 1. Considering an L-bit image, where intensity values of the pixels are varying from 0 to  $2L - 1$ , the complete approach hereby employed can be understood stepwise as follows:

**Step 1:** Linear stretching for all bands/channels of the image can be done for exploring more and more intensity span as (symbolizing  $\Upsilon \in \{R, G, B\}$ )



**Fig. 1** Process flow visualization for the proposed approach

$$\tilde{\Upsilon}(u, v) \leftarrow \frac{\Upsilon(u, v) - \Upsilon_{\min}}{\Upsilon_{\max} - \Upsilon_{\min}}, \quad (1)$$

Here,  $\Upsilon_{\max} = \max\{\Upsilon(u, v)\}$  and  $\Upsilon_{\min} = \min\{\Upsilon(u, v)\}$  symbolize for corresponding channel-wise maximum and minimum intensity values.

**Step 2:** Intensity channel is isolated for further processing by following the conversion from RGB to HSV domain.

**Step 3:** Keeping  $H$  and  $S$  channels as such; histogram for intensity ( $V$ ) channel is identified as

$$h(i) = \{n(i)\}. \quad (2)$$

Here,  $n(i)$  stands for pixel count having  $i$ th intensity value and  $h(i)$  for intensity channel histogram.

**Step 4:** Considering median intensity value ( $I_m$ ) as the basis for histogram division; both sub-histograms can be identified individually as

$$h_{lo}(i) = \{h(i)|0 \leq i \leq I_m\}; \quad h_{hi}(i) = \{h(i)|I_m < i \leq L - 1\}. \quad (3)$$

**Step 5:** Evaluate the mean  $\mu_j(\bar{i})$  pixel count as well as median  $m_j(\bar{i})$  pixel count for  $j$ th sub-histogram. Here,  $j = 1, 2$  for each sub-histogram.

**Step 6:** Evaluate the modified histogram by considering minimum among bin value, mean pixel count, and median pixel count for each sub-histogram individually, as

$$\begin{aligned} \tilde{h}(i) &= \{\min(h_{lo}(i), \mu_1(\bar{i}), m_1(\bar{i}))|0 \leq i \leq I_m\}, \\ \tilde{h}(i) &= \{\min(h_{hi}(i), \mu_2(\bar{i}), m_2(\bar{i}))|I_m < i \leq L - 1\}. \end{aligned} \quad (4)$$

**Step 7:** Considering  $N_j$  = pixel-count corresponding to  $j$ th sub-histogram, CDFs can be evaluated for both the sub-histograms independently as

$$c_j(i) = \frac{1}{N_j} \sum_{k=0}^i \tilde{h}(k). \quad (5)$$

**Step 8:** Sub-histograms are individually equalized by the following generalized equalization behavior, as governed by the following equation (here,  $j = 1$  and 2):

$$\tilde{I}_j = I_{j\_min} + (I_{j\_max} - I_{j\_min}) * c_j(i). \quad (6)$$

**Step 9:** Both sub-equalized images collectively leads to overall equalization, as

$$\tilde{I} = \tilde{I}_1 \cup \tilde{I}_2. \quad (7)$$

Here, union is just a collective association for the individually equalized sub-images.

**Step 10:** Evaluate the modified CDF (which is also weighting distributed due to RGB color contrast stretching as in Step 1) by considering histogram and its PDF for this overall equalized image and termed it as  $cdf_m(i)$  that can be calculated eventually by modified histogram  $(\tilde{h}_m(i))$ .

**Step 11:** Obtain the gamma value-set using this modified CDF as follows:

$$\gamma(i) = 1 - cdf_m(i). \quad (8)$$

**Step 12:** Consequently, gamma correction is directly imparted using Eq. (8), as

$$I_{enh}(i) = [\tilde{I}(i)]^{\gamma(i)}. \quad (9)$$

**Step 13:** As obtained in Eq. (9), thus obtained gamma-corrected channel is identified as  $I_{enh}(i)$ . It can be assumed that thus obtained image is illumination-wise corrected and enhanced. Textural enhancement can be achieved by evaluating the 2D DCT coefficients from the cumulative distribution based gamma-corrected image as

$$DM(a, b) = c_a c_b \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \hat{I}(m, n) \cdot \cos\left(\frac{\pi(2m+1)h}{2M}\right) \cdot \cos\left(\frac{\pi(2n+1)w}{2N}\right). \quad (10)$$

where  $0 \leq m, 0 \leq n, a \leq M - 1$ , and  $b \leq N - 1$ . Here,  $c_a$  and  $c_b$  are evaluated as follows:

$$c_a = \begin{cases} \sqrt{1/M}, & a = 0 \\ \sqrt{2/M}, & 1 \leq a \leq M - 1 \end{cases} \quad (11)$$

$$c_b = \begin{cases} \sqrt{1/N}, & b = 0 \\ \sqrt{2/N}, & 1 \leq b \leq N - 1 \end{cases} \quad (12)$$

**Step 14:** Now, the idea is to suggest the textural improvement through the energy redistribution in the cosine-transformed domain. Here, the core idea is just to employ the reframing of the DCT coefficients matrix ( $DM$ ). The objective is to impart the emphasis over the lower energy components among the matrix elements and hence those matrix elements are having the magnitude less than 15% of the maximum energy components. This value is empirically derived through experimentation to make the approach entirely non-iterative; otherwise, this thresholding value can be derived optimally. Thus, through proper thresholding for isolating the lower energy transformed coefficients and for imparting proper scaling through a scaling parameter ( $\xi$ ), is given in a relative manner as

$$\xi = \sqrt{\frac{\text{var}(I_{CDGC})}{\text{var}(I_{IN})}} = \sqrt{\frac{\sigma_{CDGC}^2}{\sigma_{IN}^2}} = \frac{\sigma_{CDGC}}{\sigma_{IN}} = \frac{\text{std}(I_{CDGC})}{\text{std}(I_{IN})}, \quad (13)$$

$$DM'(a, b) = \begin{cases} DM(a, b), & \text{if } |DM(a, b)| > 15\% \text{ of } DM_{\max} \\ \xi * DM(a, b), & \text{otherwise} \end{cases} \quad (14)$$

**Step 15:** Later on, the inverse discrete cosine transformation is imparted over the DM matrix and hence, through proper restoration, the enhanced image can be obtained back in the RGB domain through proper conversion scheme

$$\left[ \hat{R}(m, n), \hat{G}(m, n), \hat{B}(m, n) \right]^T = T_{HSI}^{RGB} \left[ H(m, n), S(m, n), \hat{I}(m, n) \right]^T. \quad (15)$$

Here,  $T_{HSI}^{RGB}$  is  $HSI$  to  $RGB$  transformation.

### 3 Experimentation: Performance, Evaluation, and Comparison

#### 3.1 Assessment Criterion

Performance evaluation and comparison is presented by highly relevant metrics like brightness (B), contrast (V), entropy (H), sharpness (S), and colorfulness (C), here.

#### 3.2 Qualitative Assessment

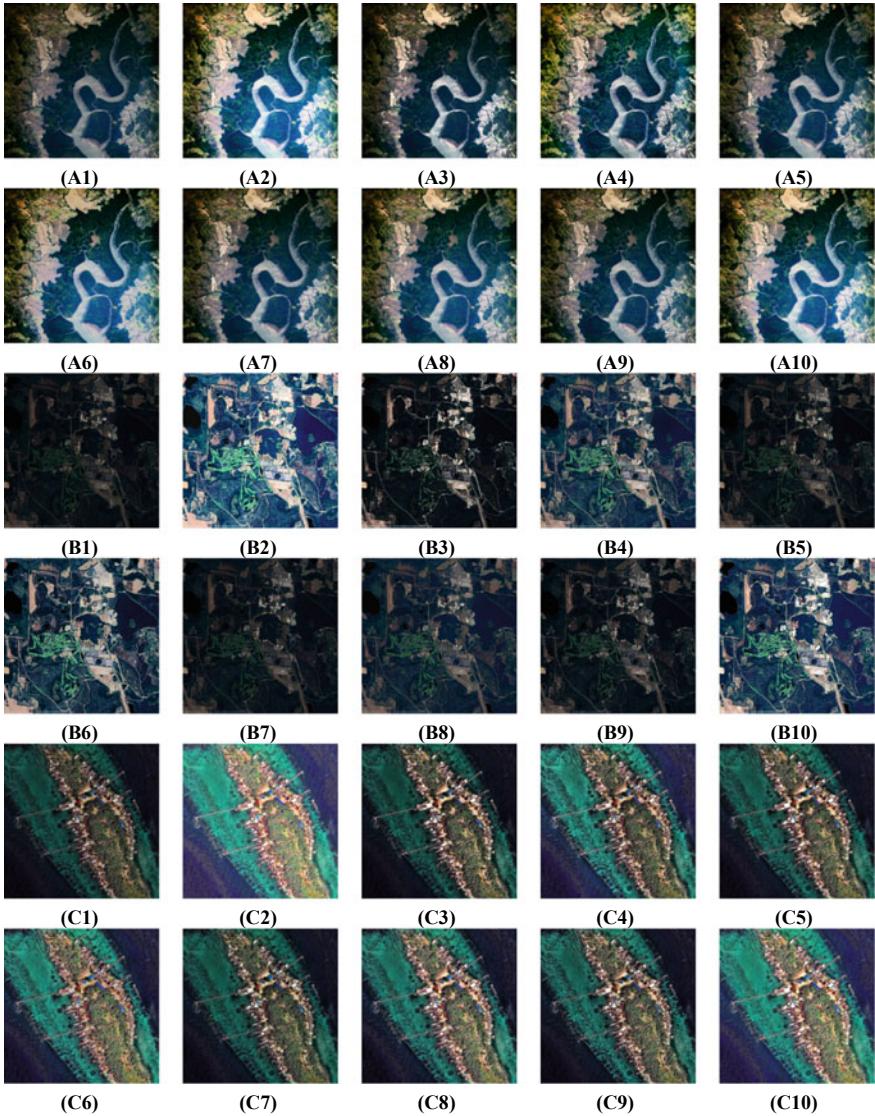
Comparative qualitative evaluation with recently published state-of-the-art methodologies (namely, GHE, MMSICHE, ADAPHE, AVHEQ, AGCWD, HEOPC, HEMIC, and IEAUMF) is presented in Fig. 2, for highlighting the significant contribution.

#### 3.3 Quantitative Assessment

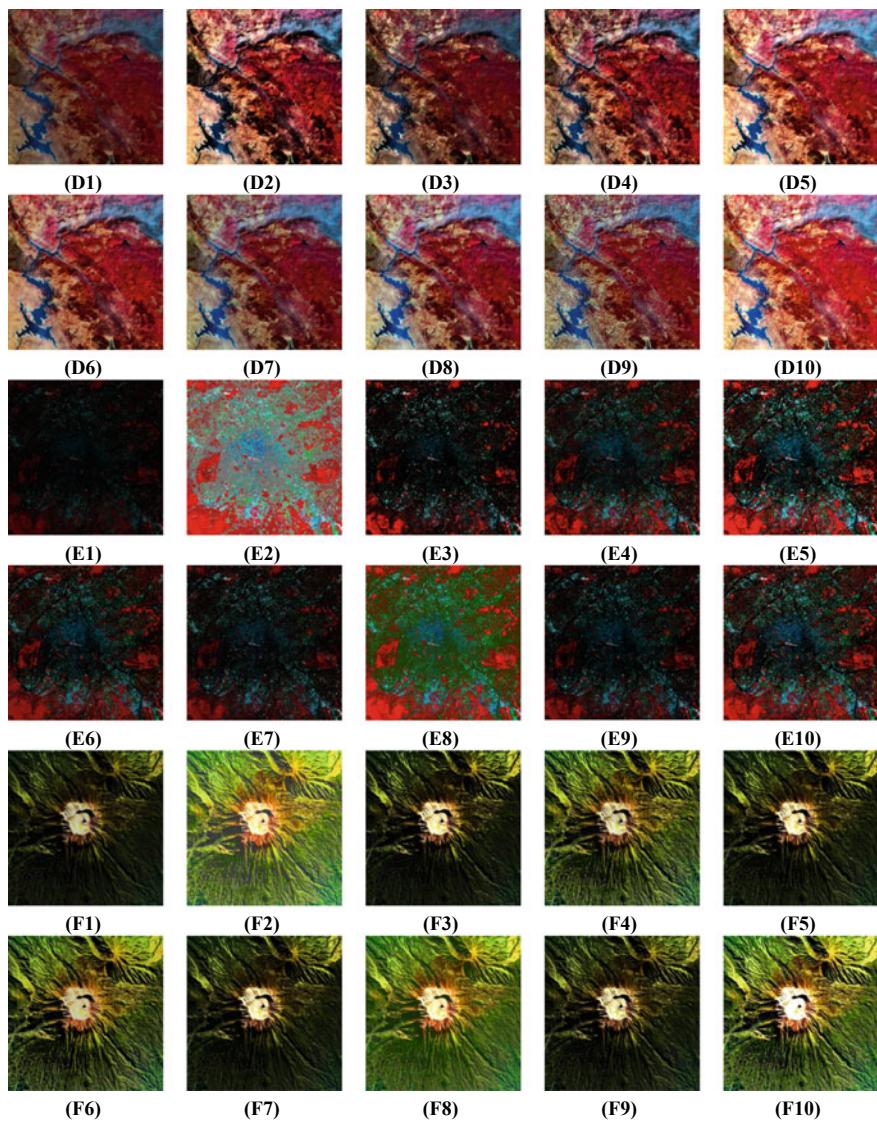
For explicit comparative numerical assessments, relevant indices are listed in Table 1. It is quite obvious for dark images that some amount of intensity level boosting is also required for imparting sufficient contrast enhancement. Hence, mean and variance both must be improved. Shannon entropy and variation of gradient must be also increased and hence, sharpness of the image must also be high along with the colorfulness of the image. Considering the abovementioned performance metrics, the values listed in Table 1, clearly show the efficacy of the proposed approach.

## 4 Conclusion

It can be easily concluded that this approach is highly adaptive to the image under consideration. Here, the beauty resides in achieving adaptive behavior along with its domain-independent effectiveness. Any kind of dark and low contrast, monochrome as well as color image can be enhanced or improved through it, irrespective of its domain. This methodology clearly outperforms the other state-of-the-art methodologies in terms of quantitative analysis, qualitative analysis, and complexity. The adaptive behavior resides in obtaining the clipping limit selection for each sub-histogram independently along with the evaluation for gamma value-set itself by optimally clipped and equalized histogram. The first step leads to the restoration of information content, and later on for entropy enhancement through adaptive gamma correction. Finally, DCT-based energy redistribution leads to textural improvement; as here,



**Fig. 2** Qualitative/visual assessment and comparison: A1–C1: input images [22–24]; A2–C2: GHE [4]; A3–C3: MMSICHE [8]; A4–C4: ADAPHE [5]; A5–C5: AVHEQ [10]; A6–C6: AGCWD [13]; A7–C7: HEOPC [11]; A8–C8: HEMIC [12]; A9–C9: IEAUMF [20]; and A10–C10: the proposed approach; D1–F1: input images [22–24]; D2–F2: GHE [4]; D3–F3: MMSICHE [8]; D4–F4: ADAPHE [5]; D5–F5: AVHEQ [10]; D6–F6: AGCWD [13]; D7–F7: HEOPC [11]; D8–F8: HEMIC [12]; D9–F9: IEAUMF [20]; and D10–F10: the proposed approach



**Fig. 2** (continued)

**Table 1** Quantitative/numerical evaluation and comparison among input images [22–24], GHE [4], MMSICHE [8], ADAPHE [5], AVHEQ [10], AGCWDF [13], HEOPC [11], HEMIC [12], IEAUMF [20], and the proposed approach using various metrics are termed as brightness (B), contrast (V), entropy (H), sharpness (S), and colorfulness (C)

S.No	Indices	Input	GHE	MMSICHE	ADAPHE	AVHEQ	AGCWDF	HEOPC	HEMIC	IEAUMF	Proposed
1.	<b>B</b>	0.2573	0.5004	0.2955	0.4104	0.3193	0.4243	0.322	0.3849	0.3237	<b>0.3952</b>
	<b>V</b>	0.0304	0.0859	0.0596	0.0623	0.0492	0.063	0.0461	0.052	0.0501	<b>0.0744</b>
	<b>H</b>	6.8359	7.2603	7.0938	7.4255	7.0964	7.1282	6.9724	7.2701	7.0812	<b>7.1438</b>
2.	<b>S</b>	0.299	0.513	0.3982	0.6406	0.3803	0.4353	0.3694	0.4034	0.462	<b>0.5084</b>
	<b>C</b>	0.1264	0.267	0.1352	0.2081	0.156	0.2267	0.16	0.2016	0.1592	<b>0.2152</b>
	<b>B</b>	0.106	0.5019	0.1573	0.3226	0.1253	0.3337	0.1294	0.2186	0.14	<b>0.374</b>
3.	<b>V</b>	0.0076	0.0846	0.0408	0.0431	0.012	0.0655	0.0102	0.015	0.0191	<b>0.1136</b>
	<b>H</b>	5.5645	6.7583	6.0061	6.8852	5.749	6.547	5.7736	6.3054	5.8876	<b>6.6864</b>
	<b>S</b>	0.2513	1.0145	0.4821	0.6922	0.318	0.8437	0.2956	0.406	0.4206	<b>0.96</b>
4.	<b>C</b>	0.0489	0.2549	0.0652	0.1576	0.0572	0.1586	0.0602	0.1113	0.0619	<b>0.2208</b>
	<b>B</b>	0.2232	0.526	0.2581	0.3636	0.2893	0.3759	0.2651	0.3477	0.3006	<b>0.3956</b>
	<b>V</b>	0.044	0.0669	0.0723	0.0881	0.0421	0.0855	0.0355	0.0485	0.0621	<b>0.1204</b>
	<b>H</b>	5.9635	6.3469	6.3378	6.5588	7.2381	6.1279	7.1482	7.3685	7.2859	<b>7.295</b>
	<b>S</b>	0.6581	0.8567	0.8033	1.0538	0.5041	0.9919	0.5082	0.5637	0.7700	<b>1.2143</b>
	<b>C</b>	0.1328	0.3077	0.1481	0.2312	0.1615	0.2414	0.1497	0.1984	0.1827	<b>0.3164</b>
	<b>B</b>	0.3519	0.501	0.3787	0.4778	0.6367	0.5218	0.4485	0.528	0.4671	<b>0.7138</b>
	<b>V</b>	0.0094	0.0859	0.0339	0.0531	0.0544	0.0345	0.0183	0.0294	0.0573	<b>0.0544</b>
	<b>H</b>	6.88	7.2077	7.0756	7.4709	7.6275	7.2984	7.2257	7.4517	7.5047	<b>7.6057</b>
	<b>S</b>	0.2415	0.7182	0.4474	0.6778	0.5638	0.4524	0.337	0.4254	0.8688	<b>0.9816</b>
	<b>C</b>	0.2126	0.3204	0.2288	0.3076	0.3684	0.3169	0.2687	0.3135	0.295	<b>0.652</b>

(continued)

**Table 1** (continued)

S.No	Indices	Input	GHE	MMSICHE	ADAPHE	AVHEQ	AGCWD	HEOPC	HEMIC	IEAUMF	Proposed
5.	<b>B</b>	0.0593	0.606	0.124	0.1823	0.2673	0.1801	0.1216	0.4108	0.1215	<b>0.2221</b>
	<b>V</b>	0.0052	0.0303	0.0523	0.0444	0.1169	0.0396	0.0262	0.0414	0.0286	<b>0.0529</b>
	<b>H</b>	3.0088	3.8787	3.2479	4.0184	3.3521	3.2883	3.827	4.4696	3.3381	<b>4.5044</b>
	<b>S</b>	0.2793	0.6183	0.6961	0.7219	1.1687	0.7111	0.5675	0.6452	0.5926	<b>0.8388</b>
	<b>C</b>	0.0856	0.5084	0.2119	0.2458	0.3749	0.2406	0.1639	0.4531	0.1684	<b>0.3657</b>
6.	<b>B</b>	0.1427	0.5241	0.1733	0.3193	0.192	0.3167	0.1792	0.4229	0.1852	<b>0.4136</b>
	<b>V</b>	0.0301	0.0649	0.055	0.0716	0.0551	0.0746	0.047	0.0482	0.055	<b>0.1709</b>
	<b>H</b>	5.2945	6.105	5.6796	6.2385	5.4903	5.7127	5.5841	6.3608	5.4861	<b>6.199</b>
	<b>S</b>	0.4507	0.8664	0.5591	0.9714	0.6184	0.9177	0.5539	0.6086	0.6641	<b>1.2098</b>
	<b>C</b>	0.1209	0.3464	0.1583	0.2326	0.1653	0.2494	0.1502	0.2762	0.1599	<b>0.3646</b>

it is employed for imparting overemphasis of the lower energy cosine-transformed coefficients which in turn leads to textural enhancement of the image under consideration. For the desired cosine-transformed emphasis for the energy redistribution and corresponding quality improvement, two parameters, namely, thresholding level and scaling parameter is decided empirically here through a rigorous experimentation, so that approach can remain non-iterative. Although more improvement can be imparted by deriving both of these parameters though quantitative optimization, but it can lead to iterative behavior of the employed method and hence, not suggested in this work. Still, the outperformance of this non-iterative approach (with empirically concluded values of scaling and thresholding parameters) can be explicitly identified when compared to qualitatively and quantitatively with the state-of-the-art methodologies.

## References

1. Singh, H., Kumar, A., Balyan L.K., Singh, G.K.: A novel optimally weighted framework of piecewise gamma corrected fractional order masking for satellite image enhancement. *Comput. Electr. Eng.*, 1–17 (2017). <https://doi.org/10.1016/j.compeleceng.2017.11.014>
2. Singh, H., Kumar, A., Balyan L.K., Singh, G.K.: Swarm intelligence optimized piecewise gamma corrected histogram equalization for dark image enhancement. *Comput. Electr. Eng.*, 1–14 (2017). <https://doi.org/10.1016/j.compeleceng.2017.06.029>
3. Demirel, H., Ozcinar, C., Anbarjafari, G.: Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition. *IEEE Geosci. Remote. Sens. Lett.* **7**(2), 333–337 (2010)
4. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2006)
5. Zuiderveld, K.: VIII.5. Contrast limited adaptive histogram equalization. *Graph. Gems*, 474–485 (1994). <https://doi.org/10.1016/b978-0-12-336156-1.50061-6>
6. Sheet, D., Garud, H., Suveer, A., Mahadevappa, M., Chatterjee, J.: Brightness preserving dynamic fuzzy histogram equalization. *IEEE Trans. Consum. Electron.* **56**(4), 2475–2480 (2010)
7. Singh, K., Kapoor, R.: Image enhancement using exposure based sub image histogram equalization. *Pattern Recogn. Lett.* **36**, 10–14 (2014)
8. Singh, K., Kapoor, R.: Image enhancement via median-mean based sub-image-clipped histogram equalization. *Optik-Int. J. Light Electr. Opt.* **125**, 4646–4651 (2014)
9. Singh, K., Kapoor, R., Sinha, S.K.: Enhancement of low exposure images via recursive histogram equalization algorithms. *Opt.: Int. J. Light. Electron Opt.* **126**(20), 2619–2625 (2015)
10. Lin, S.C.F., Wong, C.Y., Rahman, M.A., et al.: Image enhancement using the averaging histogram equalization approach for contrast improvement and brightness preservation. *Comput. Electr. Eng.* **46**, 356–370 (2014)
11. Wong, C.Y., Jiang, G., Rahman, M.A., Liu, S., Lin, S.C.F., Kwok, N., Shi, H., Yu, Y.H., Wu, T.: Histogram equalization and optimal profile compression based approach for colour image enhancement. *J. Vis. Communun. Image Represent.* **38**, 802–813 (2016)
12. Wong, C.Y., Liu, S., Liu, S.C., Rahman, A., Lin, C., Jiang, G., Kwok, N., Shi, H.: Image contrast enhancement using histogram equalization with maximum intensity coverage. *J. Mod. Opt.* **63**(16), 1618–1629 (2016)
13. Huang, S.C., Cheng, F.C., Chiu, Y.S.: Efficient contrast enhancement using adaptive gamma correction with weighting distribution. *IEEE Trans. Image Process.* **22**, 1032–1041 (2013)

14. Singh, H., Kumar, A., Balyan, L.K., Singh, G.K.: Slantlet filter-bank-based satellite image enhancement using gamma-corrected knee transformation. *Taylor & Fr.-Int. J. Electron.* **105**(10), 1695–1715 (2018). <https://doi.org/10.1080/00207217.2018.1477199>
15. Singh, H., Kumar, A.: Satellite image enhancement using beta wavelet based gamma corrected adaptive knee transformation. In: 5th IEEE International Conference on Communication and Signal Processing (ICCSP), India, pp. 128–132 (2016). <https://doi.org/10.1109/j.compeleceng.2017.06.029>
16. Singh, H., Agrawal, N., Kumar, A., Singh, G.K., Lee, H.N.: A novel gamma correction approach using optimally clipped sub-equalization for dark image enhancement. In: IEEE International Conference on Digital Signal Processing (DSP), China, pp. 497–501 (2016). <https://doi.org/10.1109/icdsp.2016.7868607>
17. Singh, H., Kumar, A., Balyan, L.K., Singh, G.K.: Regionally equalized and contextually clipped gamma correction approach for dark image enhancement. In: 4th IEEE International Conference on Signal Processing and Integrated Networks (SPIN), pp. 431–436 (2017). <https://doi.org/10.1109/spin.2017.8049988>
18. Singh, H., Kumar, A., Balyan, L.K., Singh, G.K.: Dark image enhancement using optimally compressed and equalized profile based parallel gamma correction. In: 6th IEEE International Conference on Communication and Signal Processing (ICCSP), pp. 1299–1303 (2017). <https://doi.org/10.1109/iccsp.2017.8286592>
19. Singh, H., Kumar, A., Balyan, L.K., Singh, G.K.: A novel optimally gamma corrected intensity span maximization approach for dark image enhancement. In: 22nd IEEE International Conference on Digital Signal Processing (DSP), London, United Kingdom, pp. 1–5 (2017). <https://doi.org/10.1109/icdsp.2017.8096035>
20. Lin, S.C.F., Wong, C.Y., Jiang, G., Rahman, M.A., Ren, T.R., Kwok, N., Shi, H., Yu, Y.H., Wu, T.: Intensity and edge based adaptive unsharp masking filter for color image enhancement. *Opt.: Int. J. Light. Electron Opt.* **127**, 407–414 (2016)
21. Fu, X., Wang, J., Zeng, D., Huang, Y., Ding, X.: Remote sensing image enhancement using regularized-histogram equalization and DCT. *IEEE Geosci. Remote Sens. Lett.* **12**(11), 2301–2305 (2015)
22. NASA Visible Earth: Home. <http://visibleearth.nasa.gov/>. Last accessed 2 June 2017
23. SATPALDA. <http://www.satpalda.com/gallery/>. Last accessed 2 June 2017
24. CRISP. <http://www.crisp.nus.edu.sg/> Last accessed 2 June 2017

# Author Index

## A

- Abdu Rahiman, V., 313  
Aggarwal, Ridhi, 213  
Agrawal, Aditi, 363  
Agrawal, Anupam, 339  
Agrawal, N., 471  
Agrawal, Palash, 27  
Anoop, B.N., 115  
Aravamuthan, G., 173

## B

- Babu, Suresh, 173  
Bag, Soumen, 75, 183  
Balyan, L.K., 483  
Banka, Haider, 445  
Bhatnagar, Gaurav, 431  
Bhattacharya, Rajit, 389  
Bhavsar, Arnav, 197, 325, 351  
Bhura, Pallabh, 389  
Bhuyan, M.K., 101  
Biswas, Prabir Kumar, 139, 243

## C

- Chanda, Bhabatosh, 271  
Chaudhary, Kamal, 417  
Chaudhuri, Bidyut B., 271  
Chaudhuri, Debasis, 257  
Chaudhury, Santanu, 417  
Chowdhury, Ananda S., 389, 403  
Chowdhury, Kuntal, 257

## D

- Dagnew, Guesh, 227  
Das, Apurba, 127  
De, Kanjar, 27

## Dhakrey, Puran, 301

- Dhameliya, Vatsalkumar, 351  
Dhiraj, 417  
Dixit, Anuja, 75  
Dosi, Hardik, 339  
Dubey, Shiv Ram, 271  
Dwivedi, Kshitij, 65

## E

- Enan, Sadman Sakib, 87

## G

- Gadde, Prathik, 101  
Garg, Mahak, 363  
Gautam, Anjali, 149  
George, Sudhish N., 313  
Ghosh, Dipak Kumar, 271

## H

- Haque, Samiul, 87  
Harit, Gaurav, 53, 161, 213  
Harshalatha, Y., 139  
Hatzinakos, Dimitrios, 87  
Howlader, Tamanna, 87

## J

- Jat, Dinesh, 101  
Jha, Ranjeet Ranjan, 325  
Joshi, Piyush, 363

## K

- Kalambe, Shrijay S., 1  
Kandpal, Neeta, 301  
Karar, Vinod, 1  
Kar, S., 173

Kashyap, Suraj Kumar, 101

Kaur, Harkeerat, 13

Keshri, Rahul, 339

Khanna, Pritee, 13

Kumar, Anil, 471, 483

Kumari, Seema, 325

Kumar, Manoj, 65

Kumar, Nikhil, 301

Kumar, Pradeep, 377

Kumar, Ravi, 417

Kumar, Sanjeev, 461

Kushwaha, Riti, 285

## M

Mahbubur Rahman, S.M., 87

Maiti, Somsukla, 417

Manekar, Raunak, 417

Menon, Sandeep N., 115

Mukherjee, Anindita, 403

## N

Nain, Neeta, 285

Nanavati, Nirali, 351

Neeraj, 417

Nigam, Aditya, 325

## P

Pai, Shashidhar, 127

Pal, Arup Kumar, 257, 445

Pal, Umapada, 377

Pankajakshan, Arjun, 197

Patki, Sahil, 351

Poddar, Shashi, 1

Pradhan, Jitesh, 445

Prakash, Surya, 363

Priyanka, Roy, 183

## R

Rajan, Jeny, 115

Raj, Ankesh, 445

Rajasekhar, P., 173

Raman, Balasubramanian, 149

Roy, Partha Pratim, 27, 377

Roy, Swalpa Kumar, 271

Rufus, Elizabeth, 1

## S

Sadhyा, Debanjan, 149

Sahu, Abhimanyu, 389

Saini, Rajkumar, 377

Saini, Ravi, 39

Saurav, Sumeet, 39, 417

Shakya, Snehlata, 461

Shanmugham, Sabari R., 65

Sharma, I., 471

Sharma, M.K., 243

Sheet, D, 243

Shekar, B.H., 227

Shenoy, Vinayak S., 127

Shrikhande, S.V., 173

Shylaja, S.S., 127

Sil, Jaya, 403

Singh, Himanshu, 483

Singh, Nitin, 65

Singh, Sanjay, 39, 417

Singh, Satendra Pal, 431

Srivastava, Akhilesh M., 363

Srivastava, Divya, 53, 161

Srivastav, Pravin, 339

## T

Tiwari, Anil Kumar, 213

## V

Vaghela, Sanjay, 351

Verma, R.K., 173

Vinay, Tanush, 127

Vineeth Reddy, V.B., 115

Vishwakarma, Amit, 101

## Y

Yadav, Madhulika, 39

Yadav, Rahul, 27

Yadav, Vikas, 27

Yeshwanth, A., 115

## Z

Zinzuvadiya, Milan, 351