

Balasubramanian Raman
Sanjeev Kumar
Partha Pratim Roy
Debashis Sen *Editors*

Proceedings of International Conference on Computer Vision and Image Processing

CVIP 2016, Volume 1

Advances in Intelligent Systems and Computing

Volume 459

Series editor

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

About this Series

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.

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

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India
e-mail: nikhil@isical.ac.in

Members

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

Emilio S. Corchado, University of Salamanca, Salamanca, Spain
e-mail: escorchedo@usal.es

Hani Hagras, University of Essex, Colchester, UK
e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary
e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA
e-mail: vladik@utep.edu

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

Jie Lu, University of Technology, Sydney, Australia
e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico
e-mail: epmelin@hafsamx.org

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

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland
e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

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

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

Balasubramanian Raman
Sanjeev Kumar · Partha Pratim Roy
Debashis Sen
Editors

Proceedings of International Conference on Computer Vision and Image Processing

CVIP 2016, Volume 1



Springer

Editors

Balasubramanian Raman
Department of Computer Science
and Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand
India

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

Partha Pratim Roy
Department of Computer Science
and Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand
India

Debashis Sen
Department of Computer Science
and Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand
India

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-10-2103-9

ISBN 978-981-10-2104-6 (eBook)

DOI 10.1007/978-981-10-2104-6

Library of Congress Control Number: 2016952824

© Springer Science+Business Media Singapore 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #22-06/08 Gateway East, Singapore 189721, Singapore

Preface

The first International Conference on Computer Vision and Image Processing (CVIP 2016) was organized at Indian Institute of Technology Roorkee (IITR) during February 26 to 28, 2016. The conference was endorsed by International Association of Pattern Recognition (IAPR) and Indian Unit for Pattern Recognition and Artificial Intelligence (IUPRAI), and was primarily sponsored by the Department of Science and Technology (DST) and Defense Research and Development Organization (DRDO) of the Government of India.

CVIP 2016 brought together delegates from around the globe in the focused area of computer vision and image processing, facilitating exchange of ideas and initiation of collaborations. Among 253 paper submissions, 119 (47 %) were accepted based on multiple high-quality reviews provided by the members of our technical program committee from 10 different countries. We, the organizers of the conference, were ably guided by its advisory committee comprising distinguished researchers in the field of computer vision and image processing from seven different countries.

A rich and diverse technical program was designed for CVIP 2016 comprising five plenary talks, and paper presentations in eight oral and three poster sessions. Emphasis was given to the latest advances in vision technology such as deep learning in vision, non-continuous long-term tracking, security in multimedia systems, egocentric object perception, sparse representations in vision, and 3D content generation. The papers for the technical sessions were divided based on their theme relating to low-, mid-, and high-level computer vision and image/video processing and their applications. This edited volume contains the papers presented in the technical sessions of the conference, organized session-wise.

Organizing CVIP 2016, which culminates with the compilation of these two volumes of proceedings, has been a gratifying and enjoyable experience for us.

The success of the conference was due to synergistic contributions of various individuals and groups including the international advisory committee members with their invaluable suggestions, the technical program committee members with their timely high-quality reviews, the keynote speakers with informative lectures,

the local organizing committee members with their unconditional help, and our sponsors and endorsers with their timely support.

Finally, we would like to thank Springer for agreeing to publish the proceedings in their prestigious Advances in Intelligent Systems and Computing (AISC) series. Hope the technical contributions made by the authors in these volumes presenting the proceedings of CVIP 2016 will be appreciated by one and all.

Roorkee, India

Balasubramanian Raman
Sanjeev Kumar
Partha Pratim Roy
Debashis Sen

Contents

Background Modeling Using Temporal-Local Sample Density Outlier Detection	1
Wei Zeng, Mingqiang Yang, Feng Wang and Zhenxing Cui	
Analysis of Framelets for the Microcalcification	11
K.S. Thivya and P. Sakthivel	
Reconfigurable Architecture-Based Implementation of Non-uniformity Correction for Long Wave IR Sensors	23
Sudhir Khare, Brajesh Kumar Kaushik, Manvendra Singh, Manoj Purohit and Himanshu Singh	
Finger Knuckle Print Recognition Based on Wavelet and Gabor Filtering	35
Gaurav Verma and Aloka Sinha	
Design of Advanced Correlation Filters for Finger Knuckle Print Authentication Systems	47
Gaurav Verma and Aloka Sinha	
A Nonlinear Modified CONVEF-AD Based Approach for Low-Dose Sinogram Restoration	57
Shailendra Tiwari, Rajeev Srivastava and K.V. Arya	
System Design for Tackling Blind Curves	69
Sowndarya Lakshmi Sadasivam and J. Amudha	
A Novel Visual Word Assignment Model for Content-Based Image Retrieval	79
Anindita Mukherjee, Soman Chakraborty, Jaya Sil and Ananda S. Chowdhury	
Online Support Vector Machine Based on Minimum Euclidean Distance	89
Kalpana Dahiya, Vinod Kumar Chauhan and Anuj Sharma	

Design and Development of 3-D Urban Geographical Information Retrieval Application Employing Only Open Source Instruments	101
Ajaze Parvez Khan, Sudhir Porwal and Sangeeta Khare	
A Textural Characterization of Coal SEM Images Using Functional Link Artificial Neural Network	109
Alpana and Subrajeet Mohapatra	
Template-Based Automatic High-Speed Relighting of Faces	119
Ankit Jalan, Mynepalli Siva Chaitanya, Arko Sabui, Abhijeet Singh, Viswanath Veera and Shankar M. Venkatesan	
An Improved Contextual Information Based Approach for Anomaly Detection via Adaptive Inference for Surveillance Application	133
T.J. Narendra Rao, G.N. Girish and Jeny Rajan	
A Novel Approach of an (n, n) Multi-Secret Image Sharing Scheme Using Additive Modulo	149
Maroti Deshmukh, Neeta Nain and Mushtaq Ahmed	
Scheimpflug Camera Calibration Using Lens Distortion Model	159
Peter Fasogbon, Luc Duvieubourg and Ludovic Macaire	
Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL)	171
Sonali Mishra, Lokesh Sharma, Bansidhar Majhi and Pankaj Kumar Sa	
Robust Image Hashing Technique for Content Authentication based on DWT	181
Lokanadham Naidu Vadlamudi, Rama Prasad V. Vaddella and Vasumathi Devara	
Robust Parametric Twin Support Vector Machine and Its Application in Human Activity Recognition	193
Reshma Khemchandani and Sweta Sharma	
Separating Indic Scripts with ‘matra’—A Precursor to Script Identification in Multi-script Documents	205
Sk.Md. Obaidullah, Chirita Goswami, K.C. Santosh, Chayan Halder, Nibaran Das and Kaushik Roy	
Efficient Multimodal Biometric Feature Fusion Using Block Sum and Minutiae Techniques	215
Ujwalla Gawande, Kamal Hajari and Yogesh Golhar	
Video Synopsis for IR Imagery Considering Video as a 3D Data Cuboid	227
Nikhil Kumar, Ashish Kumar and Neeta Kandpal	

Performance Analysis of Texture Image Retrieval in Curvelet, Contourlet, and Local Ternary Pattern Using DNN and ELM Classifiers for MRI Brain Tumor Images	239
A. Anbarasa Pandian and R. Balasubramanian	
ROI Segmentation from Brain MR Images with a Fast Multilevel Thresholding	249
Subhashis Banerjee, Sushmita Mitra and B. Uma Shankar	
Surveillance Scene Segmentation Based on Trajectory Classification Using Supervised Learning.	261
Rajkumar Saini, Arif Ahmed, Debi Prosad Dogra and Partha Pratim Roy	
Classification of Object Trajectories Represented by High-Level Features Using Unsupervised Learning.	273
Rajkumar Saini, Arif Ahmed, Debi Prosad Dogra and Partha Pratim Roy	
A Hybrid Method for Image Categorization Using Shape Descriptors and Histogram of Oriented Gradients.	285
Subhash Chand Agrawal, Anand Singh Jalal and Rajesh Kumar Tripathi	
Local Binary Pattern and Its Variants for Target Recognition in Infrared Imagery	297
Aparna Akula, Ripul Ghosh, Satish Kumar and H.K. Sardana	
Applicability of Self-Organizing Maps in Content-Based Image Classification	309
Kumar Rohit, R.K. Sai Subrahmanyam Gorthi and Deepak Mishra	
Road Surface Classification Using Texture Synthesis Based on Gray-Level Co-occurrence Matrix	323
Somnath Mukherjee and Saurabh Pandey	
Electroencephalography-Based Emotion Recognition Using Gray-Level Co-occurrence Matrix Features	335
Narendra Jadhav, Ramchandra Manthalkar and Yashwant Joshi	
Quick Reaction Target Acquisition and Tracking System	345
Zahir Ahmed Ansari, M.J. Nigam and Avnish Kumar	
Low-Complexity Nonrigid Image Registration Using Feature-Based Diffeomorphic Log-Demons	357
Md. Azim Ullah and S.M. Mahbubur Rahman	
Spotting of Keyword Directly in Run-Length Compressed Documents	367
Mohammed Javed, P. Nagabhushan and Bidyut Baran Chaudhuri	

Design and Implementation of a Real-Time Autofocus Algorithm for Thermal Imagers.....	377
Anurag Kumar Srivastava and Neeta Kandpal	
Parameter Free Clustering Approach for Event Summarization in Videos	389
Deepak Kumar Mishra and Navjot Singh	
Connected Operators for Non-text Object Segmentation in Grayscale Document Images	399
Sheshera Mysore, Manish Kumar Gupta and Swapnil Belhe	
Non-regularized State Preserving Extreme Learning Machine for Natural Scene Classification.....	409
Paheding Sidike, Md. Zahangir Alom, Vijayan K. Asari and Tarek M. Taha	
A Local Correlation and Directive Contrast Based Image Fusion.....	419
Sonam and Manoj Kumar	
Multi-exposure Image Fusion Using Propagated Image Filtering	431
Diptiben Patel, Bhoomika Sonane and Shanmuganathan Raman	
Tone Mapping HDR Images Using Local Texture and Brightness Measures	443
Akshay Gadi Patil and Shanmuganathan Raman	
Pre- and Post-fingerprint Skeleton Enhancement for Minutiae Extraction	453
Geevar C. Zacharias, Madhu S. Nair and P. Sojan Lal	
Content Aware Image Size Reduction Using Low Energy Maps for Reduced Distortion	467
Pooja Solanki, Charul Bhatnagar, Anand Singh Jalal and Manoj Kumar	
Artificial Immune Hybrid Photo Album Classifier	475
Vandna Bhalla and Santanu Chaudhury	
Crowd Disaster Avoidance System (CDAS) by Deep Learning Using eXtended Center Symmetric Local Binary Pattern (XCS-LBP) Texture Features	487
C. Nagananthini and B. Yogameena	
A Novel Visualization and Tracking Framework for Analyzing the Inter/Intra Cloud Pattern Formation to Study Their Impact on Climate	499
Bibin Johnson, J. Sheeba Rani and Gorthi R.K.S.S. Manyam	

Cancelable Biometric Template Security Using Segment-Based Visual Cryptography	511
P. Punithavathi and S. Geetha	
PCB Defect Classification Using Logical Combination of Segmented Copper and Non-copper Part	523
Shashi Kumar, Yuji Iwahori and M.K. Bhuyan	
Gait Recognition-Based Human Identification and Gender Classification	533
S. Arivazhagan and P. Induja	
Corner Detection Using Random Forests	545
Shubham Pachori, Kshitij Singh and Shanmuganathan Raman	
Symbolic Representation and Classification of Logos	555
D.S. Guru and N. Vinay Kumar	
A Hybrid Method Based CT Image Denoising Using Nonsubsampled Contourlet and Curvelet Transforms	571
Manoj Diwakar and Manoj Kumar	
Using Musical Beats to Segment Videos of <i>Bharatanatyam Adavus</i>	581
Tanwi Mallick, Akash Anuj, Partha Pratim Das and Arun Kumar Majumdar	
Parallel Implementation of RSA 2D-DCT Steganography and Chaotic 2D-DCT Steganography	593
G. Savithri, Vinupriya, Sayali Mane and J. Saira Banu	
Thermal Face Recognition Using Face Localized Scale-Invariant Feature Transform	607
Shruti R. Uke and Abhijeet V. Nandedkar	
Integrating Geometric and Textural Features for Facial Emotion Classification Using SVM Frameworks	619
Samyak Datta, Debasish Sen and R. Balasubramanian	
Fast Non-blind Image Deblurring with Sparse Priors	629
Rajshekhar Das, Anurag Bajpai and Shankar M. Venkatesan	
Author Index	643

About the Editors

Balasubramanian Raman is Associate Professor in the Department of Computer Science and Engineering at Indian Institute of Technology Roorkee from 2013. He has obtained M.Sc degree in Mathematics from Madras Christian College (University of Madras) in 1996 and Ph.D. from Indian Institute of Technology Madras in 2001. He was a postdoctoral fellow at University of Missouri Columbia, USA in 2001–2002 and a postdoctoral associate at Rutgers, the State University of New Jersey, USA in 2002–2003. He joined Department of Mathematics at Indian Institute of Technology Roorkee as Lecturer in 2004 and became Assistant Professor in 2006 and Associate Professor in 2012. He was a Visiting Professor and a member of Computer Vision and Sensing Systems Laboratory at the Department of Electrical and Computer Engineering in University of Windsor, Canada during May August 2009. So far he has published more than 190 papers in reputed journals and conferences. His area of research includes vision geometry, digital watermarking using mathematical transformations, image fusion, biometrics and secure image transmission over wireless channel, content-based image retrieval and hyperspectral imaging.

Sanjeev Kumar is working as Assistant Professor with Department of Mathematics, Indian Institute of Technology Roorkee from November 2010. Earlier, he worked as a postdoctoral fellow with Department of Mathematics and Computer Science, University of Udine, Italy from March 2008 to November 2010. He has completed his Ph.D. in Mathematics from IIT Roorkee, India in 2008. His areas of research include image processing, inverse problems and machine learning. He has co-convened the first international conference on computer vision and image processing in 2016, and has served as a reviewer and program committee member of more than 20 international journals and conferences. He has conducted two workshops on image processing at IIT Roorkee in recent years. He has published more than 55 papers in various international journals and reputed conferences. He has completed a couple of sponsored research projects.

Partha Pratim Roy received his Ph.D. degree in Computer Science in 2010 from Universitat Autònoma de Barcelona, Spain. He worked as postdoctoral research fellow in the Computer Science Laboratory (LI, RFAI group), France and in Synchromedia Lab, Canada. He also worked as Visiting Scientist at Indian Statistical Institute, Kolkata, India in 2012 and 2014. Presently, Dr. Roy is working as Assistant Professor at Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Roorkee. His main research area is Pattern Recognition. He has published more than 60 research papers in various international journals, conference proceedings. Dr. Roy has participated in several national and international projects funded by the Spanish and French government. In 2009, he won the best student paper award in International Conference on Document Analysis and Recognition (ICDAR). He has gathered industrial experience while working as an Assistant System Engineer in TATA Consultancy Services (India) from 2003 to 2005 and as Chief Engineer in Samsung, Noida from 2013 to 2014.

Debashis Sen is Assistant Professor at the Department of Electronics and Electrical Communication Engineering in Indian Institute of Technology (IIT) Kharagpur. Earlier, from September 2014 to May 2015, he was Assistant Professor at the Department of Computer Science and Engineering in Indian Institute of Technology (IIT) Roorkee. Before joining Indian Institute of Technology, he worked as a postdoctoral research fellow at School of Computing, National University of Singapore for about 3 years. He received his PhD degree from the Faculty of Engineering, Jadavpur University, Kolkata, India in 2011 and his M.A.Sc. degree from the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada in 2005. He has worked at the Center for Soft Computing Research of Indian Statistical Institute from 2005 to 2011 as a research scholar, and at the Center for Signal Processing and Communications and Video Processing and Communications group of Concordia University as a research assistant from 2003 to 2005. He is currently an associate editor of IET Image Processing journal. He has co-convened the first international conference on computer vision and image processing in 2016, and has served as a reviewer and program committee member of more than 30 international journals and conferences. Over the last decade, he has published in high-impact international journals, which are well cited, and has received two best paper awards. He heads the Vision, Image and Perception group in IIT Kharagpur. He is a member of Institute of Electrical and Electronics Engineers (IEEE), IEEE Signal Processing Society and Vision Science Society (VSS). His research interests include vision, image and video processing, uncertainty handling, bio-inspired computation, eye movement analysis, computational visual perception and multimedia signal processing.

Background Modeling Using Temporal-Local Sample Density Outlier Detection

Wei Zeng, Mingqiang Yang, Feng Wang and Zhenxing Cui

Abstract Although researchers have proposed different kinds of techniques for background subtraction, we still need to produce more efficient algorithms in terms of adaptability to multimodal environments. We present a new background modeling algorithm based on temporal-local sample density outlier detection. We use the temporal-local densities of pixel samples as the decision measurement for background classification, with which we can deal with the dynamic backgrounds more efficiently and accurately. Experiment results have shown the outstanding performance of our proposed algorithm with multimodal environments.

Keywords Background modeling • Temporal-local sample density • Outlier detection

1 Introduction

Background modeling methods are very important and essential for many applications in the fields like computer vision and pattern recognition. More and more researchers have put their interests in this particular topic, and numerous background modeling methods have appeared in recent literature. Many of them utilize parametric models to describe pixel locations and we classify them as parametric methods. Others we call them samples-based methods which establish their models

W. Zeng (✉) · M. Yang · Z. Cui

School of Information Science and Engineering, Shandong University, Jinan, China
e-mail: wei.baldwin.zeng@gmail.com

M. Yang
e-mail: yangmq@sdu.edu.cn

Z. Cui
e-mail: 478715631@qq.com

F. Wang
Shandong Institute for Product Quality Inspection, Jinan, China
e-mail: wfzj0531@sina.com

using the set of the previously observed values. Most existing methods can output a relatively good result when the background only includes a static scene, but when dealing with scenes containing dynamic background, they often struggled.

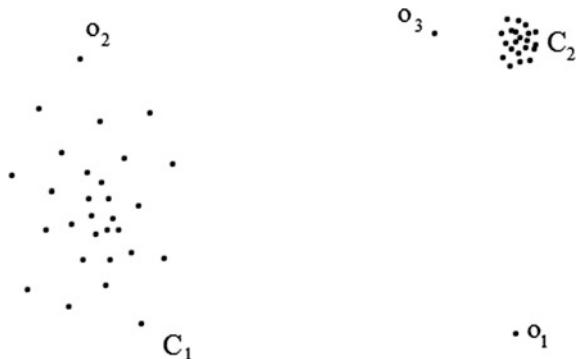
A novel background modeling method based on sample density outlier detection is proposed in this paper. Specifically, we do not estimate the probability density function of a pixel, instead we utilize the set of previously observed sample values as the initial background model of the pixel. Then the classification process utilizes the local background factor (LBF) of a sample to describe the degree of being background instead of a binary property. Finally, we compare the LBFs of newly observed values to their nearest neighbors and update the background model. Our proposed algorithm considers the background classification from a more temporal-local perspective and experiments have shown that our method outperforms other existing algorithms when the scenes contain dynamic backgrounds and in some color-similar situations.

2 Pixel Model

For each pixel location, we use N to represent the number of the background samples. In the well-known algorithm, visual background extractor method (ViBE) [1], to classify a newly observed value, the algorithm compares it with its nearest values among the background samples by defining a fixed radius and a constant threshold, which in general, considers background classification from a more global perspective. Nevertheless, for many stimulating real-world situations, we must classify the pixels relatively to their local neighborhoods.

To demonstrate, think about the following example given in Fig. 1. This kind of distribution of background samples may happen very often in scenes with dynamic backgrounds (tree leafs, water wave, and so on). It should be noted that the distribution in real world is always in three-dimensional color space while here we demonstrate to you in two-dimensional space just for perceptual intuition.

Fig. 1 A general sample distribution of dynamic backgrounds



In our paper, o , p , and q are used to denote the background samples in a model. The notation $d(p, o)$ is used to represent the distance between samples p and o . As to the set of samples, we use C to denote. In this example, both o_1 and o_3 should be classified as outliers, which is a terminology in data mining means considerably different from other samples, whereas o_2 should be considered to belong to cluster C_1 even though C_1 is much sparser. If measured by the methods using distance-based classification, o_1 can be classified easily as an outlier, o_2 may be misclassified as an outlier depending on the given global parameters. Meanwhile, o_3 , which may stand for a color-similar foreground sample, cannot be classified unless some of C_1 are misclassified.

Aiming to solve this notorious and universal problem, we utilize the temporal-local sample density to describe the degree of a sample being background, with which we can significantly change the current struggling situation.

3 Framework of Our Proposed Method

First we initialize the background model by the N most recently gained pixel values, as did in [2]. For each newly observed pixel, we compare the LBF to its neighborhood to decide whether it belongs to the background or not. That is the core of our proposed method. The LBF is a description of the possibility of being background, by comparing the local density of the pixel to its nearest neighborhoods. The LBF measures the degree of possibility instead of a binary property, which combines the advantage of parametric techniques to the samples-based methods. At the updating stage, after comparing the LBFs of the newly observed pixels to the average value of their nearest neighborhoods' LBFs and determining whether they belong to background or not, we only update those pixels that are currently background, and perform this update with the probability $t = 1/T$, in which T is the update rate. In order to ensure the spatial consistency of our background modeling algorithm, we also update (with the probability $t = 1/T$) one of the neighboring pixel with its own current pixel value.

4 Local Background Factor

The local background factor is based on the outlier detection model [3] in data mining, with which we describe the degree of being background for each pixel in our proposed algorithm. Based on the distance between sample p and an arbitrary sample $o \in C$, first we define the k -distance of sample p as

- At least there are k samples $o' \in C$ which satisfy the constraint that $d(p, o') \leq d(p, o)$
- At most there are $k - 1$ samples $o' \in C \setminus \{p\}$ which satisfy the constrain that $d(p, o') < d(p, o)$

Having the k -distance of p , we define the k -distance neighborhood of p as the region including all the samples whose distances from p are smaller than its k -distance

$$N_k(p) = \{q \in C \setminus \{p\} | d(p, q) \leq k - \text{distance}(p)\} \quad (1)$$

In order to reduce the statistical fluctuations of $d(p, o)$ when we calculate the local density, we define the reachability distance of sample p as

$$\text{reach-dist}_k(p, o) = \max\{k - \text{distance}(o), d(p, o)\} \quad (2)$$

Our goal is to compare the k -distance neighborhood samples' temporal-local densities with the newly observed value, which means that we have to calculate the temporal-local densities of samples dynamically. So we define the temporal-local reachability density (TLRD) of p as

$$\text{TLRD}_k(p) = 1 / \left[\frac{\sum_{o \in N(p)} \text{reach-dist}_k(p, o)}{|N_k(p)|} \right] \quad (3)$$

Note that the TLRD can be ∞ if the summation of all the distances is 0. If this situation happens, we can definitely label this sample p as a background pixel.

Finally, the LBF of p is defined as

$$\text{LBF}_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\text{TLRD}_k(o)}{\text{TLRD}_k(p)}}{|N_k(p)|} \quad (4)$$

The LBF of sample p describes to what extent we call p a background sample. It is the mean of the ratio of the TLRD of newly observed value and those of its k -distance-closest neighbors. We can easily see that if the TLRD of p is lower, and the TLRDs of p 's k -closest neighbors are higher, the LBF value of p is definitely higher.

Getting the LBF of a sample offers many advantages. First and the most important thing is we consider the background classification process from a more local perspective. We can handle multimodal situations more easily and accurately with dynamic thresholds according to the neighborhoods instead of a fixed global parameter. At the updating stage, since we get the LBFs of the samples, we can purposefully choose the pixels with largest LBFs to be replaced by the new confirmed background pixels to ensure the correct update trend, rather than using a first-in first-out strategy or a random scheme like most other samples-based techniques do.

5 Experiments and Results

A series of experiments based on data set Wallflower sequences [4] are conducted to evaluate our proposed method. The algorithm can deal with different categories, especially scenes with dynamic backgrounds, more accurately and efficiently comparing to most other existing methods. First we determine our parameters k and N using the percentage of correct classification (PCC), which is defined as

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

where TN is short for the true negatives, TP represents the number of true positives, FN stands for the false negatives, and FP is defined as the false positives.

To choose the best value for k , the evolution of the performance of our proposed method is computed for k ranging from 1 to 20. Figure 2 shows that when determined $k = 3$ and $k = 4$, we can obtain the best PCCs. As we know, the larger the value of k is, the slower the computational speed of LBF is. So we determine the optimal number of k to $k = 3$. Figure 3 shows PCCs obtained for N ranging from 2 to 100. It is clear from the results that a bigger N provides a better PCC. But it tends to saturate when numbers are bigger than 20. In our algorithm we select $N = 20$ as our optimal value of N to induce the computational cost.

Figure 4 shows some intuitive situations with dynamic backgrounds and the output of our proposed method. Figure 4a shows a man walks in front of the camera with shaking trees behind him and Fig. 4b shows the output we get through our

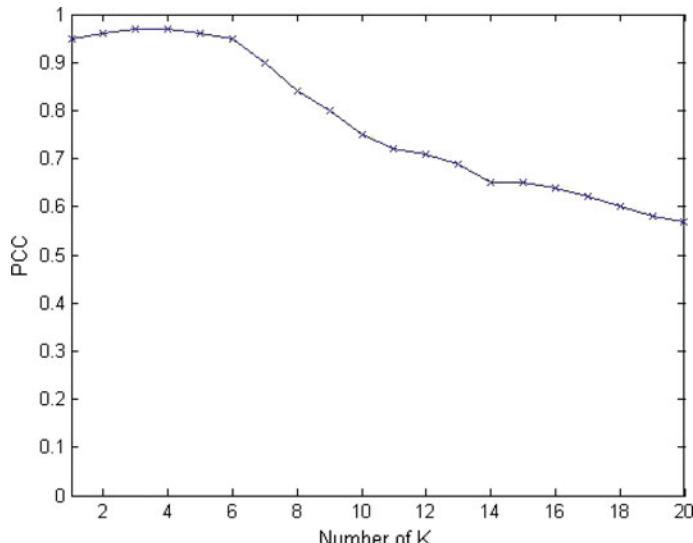


Fig. 2 Determination of parameter k

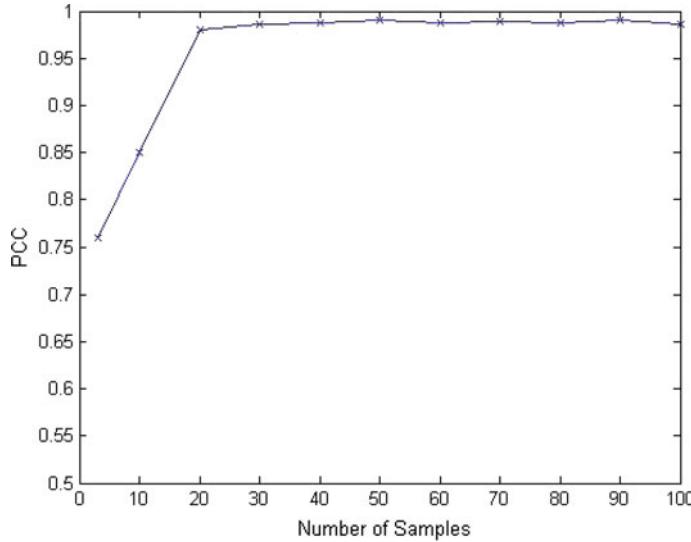


Fig. 3 Determination of parameter N

method. Figure 4c deals with the similar situation with a man standing in front of a waving water surface and Fig. 4d shows the result of our method. We find that, using our proposed method we could get relatively accurate binary maps of foreground objects. Figure 4e demonstrates one kind of situation that the foreground object has the similar color with the background, for example, the silvery car on the cement road. Using parametric techniques or distance-based algorithms usually cannot distinguish the background from the foreground well, but we can get relatively good results using our proposed method because using the local density as the measurement of decision, we can magnify the difference between the samples. Figure 4f is the output of the state-of-art algorithm ViBe [1], which shows the representative output of most existing methods, and Fig. 4g shows the result of our method. We can see that our algorithm can get more accurate output.

In Table 1, we also give values of average recall, precision, PCC, and *F*-measure, of some state-of-the-art methods, using the public dynamic background data set [5] provided on the website of CDNET (ChangeDetection.NET). The metrics used are as follows:

$$\text{Recall} = \frac{TP}{FN + TP} \quad (6)$$

$$\text{Precision} = \frac{TP}{FP + TP} \quad (7)$$

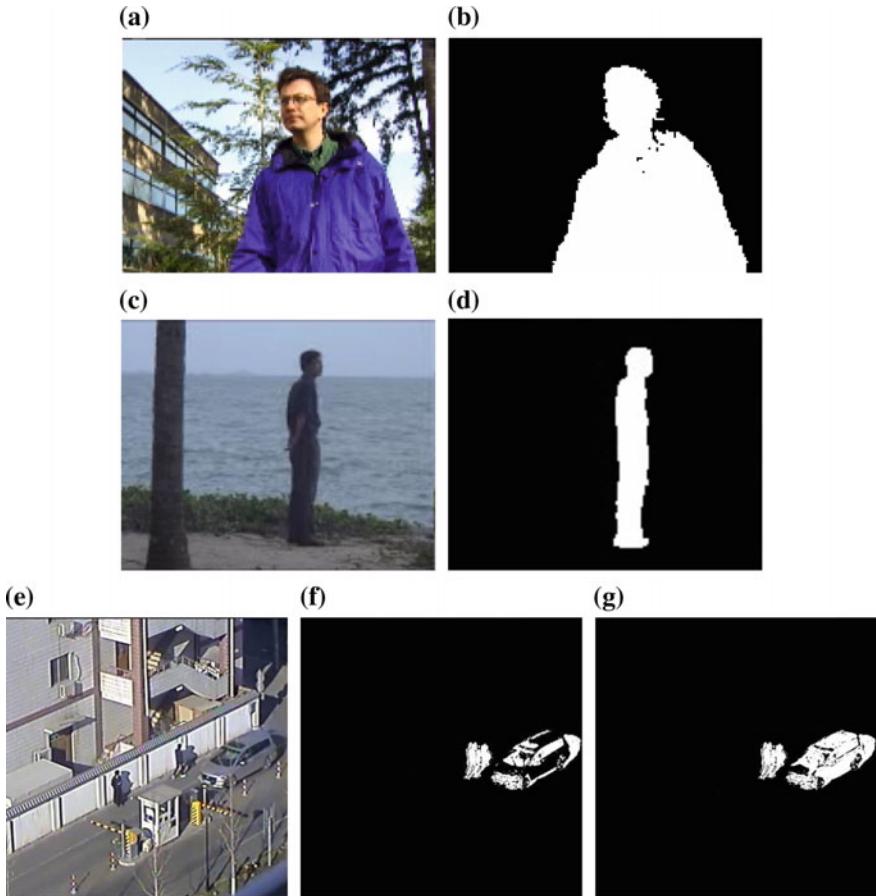


Fig. 4 Results of our proposed method dealing with scenes with dynamic backgrounds and color-similar situation

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Table 1 shows that our proposed method outperforms other methods when dealing with dynamic backgrounds. The average precision has a great improvement against ViBe. Although the PCC is at the average level, the average *F*-Measure of our method gains a significant performance improvement.

Table 1 Comparison to other existing methods when dealing with dynamic backgrounds

Method	Average recall	Average precision	Average PCC	Average <i>F</i> -measure
GMM [6]	0.8019	0.6213	0.9883	0.6328
SOBS [7]	0.8798	0.5856	0.9836	0.6439
EFIC [8]	0.6667	0.6849	0.9908	0.5779
KNN [9]	0.8047	0.6931	0.9919	0.6865
SC SOBS [10]	0.8918	0.6283	0.9831	0.6686
CP3-online [11]	0.7260	0.6122	0.9934	0.6111
PBAS [2]	0.6955	0.8326	0.9946	0.6829
ViBe [1]	0.8021	0.5346	0.9872	0.6217
Our LBF	0.8292	0.7788	0.9899	0.7652

6 Conclusion

In this paper, we propose a novel background modeling based on temporal-local sample density outlier detection. We calculate the LBF of each sample and classify the newly observed values by comparing their LBFs with their nearest neighborhoods. The proposed method can deal with scenes containing dynamic backgrounds more efficiently and get more accurate results compared to other existing methods. Future work will be focused on the modification of the determination rules and the improvement of the computational speed.

Acknowledgments The work is supported by the Natural Science Foundation of Shandong Province under Grant No. ZR2014FM030, No. ZR2013FM032 and No. ZR2014FM010.

References

1. O. Barnich and M. Van Droogenbroeck.: ‘ViBe: A Universal Background Subtraction Algorithm for Video Sequences’, *IEEE Transactions on Image Processing*, 2011, 20(6), p 1709–1724
2. M. Hofmann, P. Tiefenbacher: ‘Background Segmentation with Feedback: The Pixel-Based Adaptive Segmenter’, IEEE Workshop on Change Detection, 2012, doi:[10.1109/CVPRW.2012.6238925](https://doi.org/10.1109/CVPRW.2012.6238925)
3. M. M. Breunig, H.P. Kriegel, R.T. Ng, and J.Sander,: ‘LOF: Identifying Density-based Local Outliers’, SIGMOD RECORD, 2000, 29(2), p 93–104
4. K. Toyama, et al.: “Wallflower: Principles and Practice of Background Maintenance”, Seventh International Conference on Computer Vision, September 1999, Kerkyra, Greece, pp. 255–261
5. N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar, changedetection.net: A new change detection benchmark dataset, in Proc. IEEE Workshop on Change Detection (CDW-2012) at CVPR-2012, Providence, RI, 16–21 Jun., 2012
6. Z. Zivkovic.: ‘Improved adaptive gaussian mixture model for background subtraction’, Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, England, 2004, pages 28–31, doi:[10.1109/ICPR.2004.1333992](https://doi.org/10.1109/ICPR.2004.1333992)

7. L. Maddalena, A. Petrosino,: ‘A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications’, *IEEE Transactions on Image Processing*, 2008, 17(8), p 1168–1177
8. G. Allebosch, F. Deboeverie, ‘EFIC: Edge based Foreground background segmentation and Interior Classification for dynamic camera viewpoints’, In Advanced Concepts for Intelligent Vision Systems (ACIVS), Catania, Italy, pp. Accepted, 2015
9. Z. Zivkovic, F. van der Heijden,: ‘Efficient adaptive density estimation per image pixel for the task of background subtraction’, *Pattern Recognition Letters*, 2006, 27(7), p 773–780
10. L. Maddalena, A. Petrosino, “The SOBS algorithm: what are the limits?”, in proc of IEEE Workshop on Change Detection, CVPR 2012
11. Dong Liang, Shun'ichi Kaneko, “Improvements and Experiments of a Compact Statistical Background Model”, [arXiv:1405.6275](https://arxiv.org/abs/1405.6275)

Analysis of Framelets for the Microcalcification

K.S. Thivya and P. Sakthivel

Abstract Mammography is used commonly to detect the cancer in breast at the early stage. The early stage of breast cancer detection helps in avoiding the removal of breast in women and even the death caused due to breast cancer. There are many computer aided softwares that are designed to detect breast cancer but still only biopsy method is effective in predicting the exact scenario. This biopsy technique is a painful one. To avoid this, a novel classification approach for classifying microcalcification clusters based on framelet transform is proposed. The real-time mammography images were collected from Sri Ramachandra Medical Centre, Chennai, India in order to evaluate the performance of the proposed system. The GLCM features (contrast, energy and homogeneity) are extracted from the framelet decomposed mammograms with different resolution levels and support vector machine classifier is used to classify the unknown mammograms into normal or abnormal initially and then further classifies it as benign or malignant if detected as abnormal. The result shows that framelet transform-based classification provides considerable classification accuracy.

Keywords Framelet transform • Support vector machine • GLCM features (contrast, energy and homogeneity) • Microcalcifications and Mammography

1 Introduction

As per the medical research reports, one of the important criterion for death in women, is due to breast cancer. Microcalcification which is a major indicator of breast malignancy is classified by many methods. A matrix consisting of wavelet

K.S. Thivya (✉)

Department of Electronics and Communication Engineering,
Easwari Enginnering College, Chennai 600 089, India
e-mail: grss_dew@yahoo.co.in

P. Sakthivel

Department of Electronics and Communication Engineering,
Anna University, Chennai, India

coefficients of each image is used; then by selecting the threshold the Euclidian distance is maximized by the columns and the classification is done by using the features which are nothing but the selected columns [1]. DT CWT is used for decomposing the mammograms for different levels and SVM is used for the classification of breast cancer [2]. Two-stage wavelet transforms is applied for detecting and segmenting the granular microcalcifications [3]. Skewness and kurtosis property is used for detecting the microcalcification [4]. Texture features were extracted and the SVM is used for classification purpose [5].

Curvelet transform is employed in extracting the features of texture [6]. The Split Bregman method is used to sort out the resulting minimization problem [7]. Linearized Bregman iteration is proposed and analyzed for image deblurring in tight frame domains [8]. For image fusion, framelet transform is used [9]. Segmentation and surface reconstruction is done by using wavelet frame models [10].

Neural network is used as a classifier and the equivalent regularization properties has been discussed [11]. For detecting microcalcification automatically several state-of-the-art machine learning methods are discussed [12]. An automatic detection method for both the microcalcification and masscalcification is explained by using a feed forward neural network [13]. Different textural features are extracted and neural network is employed to classify the mass automatically [14]. Gabor wavelet features of textural analysis are discussed [15].

The scheme proposed for the classification of mammogram as normal/abnormal and benign/malignant for microcalcification is shown in Fig. 1. The GLCM features from framelet decomposed mammogram are extracted and fed to SVM classifier linear kernel. The flow of the paper is as follows. The explanation of the flow

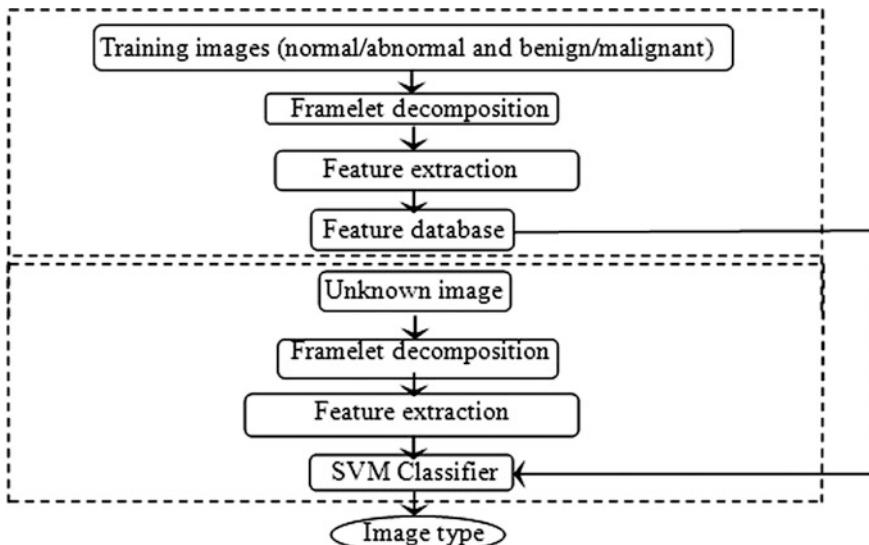


Fig. 1 Flow chart of analysis of framelets for the microcalcification

diagram of the proposed system is discussed in Sect. 2. An explanation of material used in the propose system and the ROI extraction from the mammograms is discussed in Sect. 2.1. The details of feature extraction using framelet transform are given in Sect. 2.2. Section 3 deals with the demonstration of the proposed system and Sect. 5 gives the classification rate results and concludes the proposed system.

2 Methodology

The proposed method is used to classify whether the given mammogram is a normal or abnormal one. If any abnormality is detected, then the second stage of classification is done to identify whether it is benign or malignant. To do this, a two-phase operation is involved. In the first phase, the system is trained with the images to detect whether it is normal or abnormal, for processing this, a framelet transform is applied. Framelet transform is used because it can be much more directionally selective for image processing and has shift invariant property. It is also less sensitive to corruption.

At various levels of decomposed framelets the GLCM features are extracted. The feature that shows the similarity is grouped to a same class and is used as reference for classification. In the second phase the similar steps are repeated to construct another database for classifying benign and malignant. For both the stages, separate SVM linear kernel is used to classify by making use of the two different databases created for the two phases mentioned above.

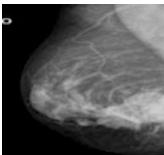
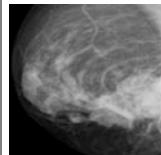
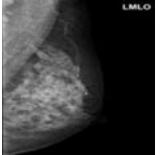
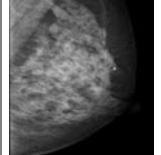
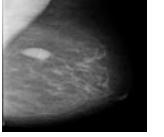
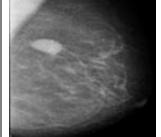
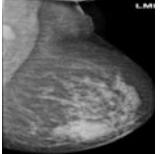
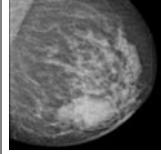
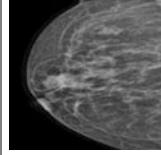
2.1 Material Used

The real-time digital mammogram images were collected from Sri Ramachandra Medical Centre, Chennai. The size of the image is 1024*1024. Half of the image contains only the background information; therefore, a manual cropping is done to remove the background information. This is done by aligning the centre of ROI with the given centre of abnormality by the database provided. The cropped ROI size is 256*256. Generally, a mammogram can be filmed in four different angles. In the proposed system, the MLO view images are only considered because it covers most of the upper quadrant breast and also the arm pit area. The training set images for benign and malignant classification is presented as a tabular column below for an overview of how the database was constructed from the collected data (Table 1).

2.2 Framelet Feature Extraction

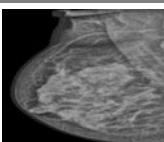
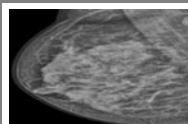
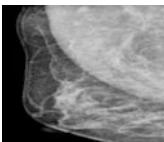
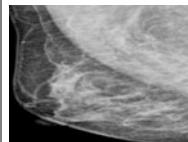
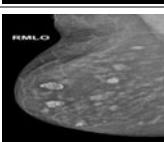
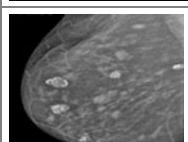
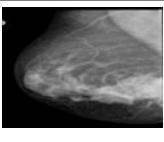
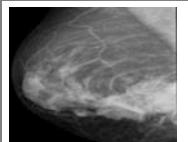
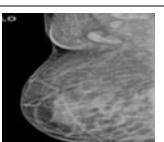
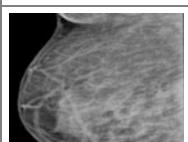
The framelet transform is applied for the images which will in turn provide us the framelet decomposition. For different levels of decomposed framelets feature extraction is done. In this proposed method, GLCM features are classified.

Table 1 Tabular column of the collected database to train the proposed system

No	MLO	Zoomed MLO	Us report	Histopathology report
1			Spiculated, microlobulated masses with nipple retraction and probable malignant right axillary lymphadenopathy—right breast (BIRADS VI)	Positive for malignant cells
2			Focal macro calcification left breast {BIRADS II}	Cytology is suggestive of a benign breast lesion
3			Fibroadenoma with cystic degeneration and calcific foci in the left breast—BIRADS III	Features are that of benign lesion
4			Hypo echoic lesion in left breast subareolar region at 8–9 o'clock position. The lateral aspect of the lesion appears well defined, while the medial aspect appears ill-defined with speculated margins. Calcific specks noted within.-f/s/o malignant mass	Cytology is positive for malignant cells suggestive of invasive ductal carcinoma
5			Irregular spiculated mass lesion with group of microcalcification within the mass—right breast with multiple right axillary lymph nodes showing loss of fatty hilum. (birads v)	Positive for malignant cells

(continued)

Table 1 (continued)

No	MLO	Zoomed MLO	Us report	Histopathology report
6			Features are suggestive of multicentric breast carcinoma—right breast (birads 5)	Positive for malignant cells
7			Cyst with thin septa—left breast (u -3)	Benign breast tissue
8			Hypoechoic lesion at 12'o clock position in the right breast—birads iii	Histology shows features of a benign breast lesion
9			Spiculated, microlobulated masses with nipple retraction and probable malignant right axillary lymphadenopathy—right breast (birads vi)	Positive for malignant cells
10			Simple cyst in right breast-birads ii	Features are in favour of benign proliferative breast disease

The GLCM features are contrast, energy and homogeneity. These features are calculated by the formulae given below in [16].

$$(i) \text{Homogeneity (angular second moment)} - \text{ASM} = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} \{P(i,j)\}^2 \quad (1)$$

$$(ii) \text{Contrast} = \sum_{n=0}^{G-1} n^2 \left\{ \sum_{i=1}^G \sum_{j=1}^G P(i, j) \right\}, |i-j|=n \quad (2)$$

$$(iii) \text{Energy} = \sum_{i=0}^{2G-2} i P_{x+y}(i) \quad (3)$$

The extracted features are used for reference and are created as a database. Two sets of databases are created, one for normal and abnormal and the second one is for benign and malignant. Using the database, the two-stage SVM classifier classifies the image type at each level.

3 Demonstration of the Proposed System

The prototype of proposed breast cancer diagnosis system discussed in the previous section is established in this section. It allows the radiologists to have easy access to the system and helps them in diagnosing breast cancer using digital mammograms. A PC platform is required to run the proposed system. Figure 2 shows the complete generalized frame of the proposed system.

In order to display the acquired digital mammogram, a larger window is designed inside the frame. Also, a small window is designed for displaying the ROI region. The right side of the frame contains a toolbar for accessing the various functions of the proposed system such as selection of input images, ROI extraction, saving the extracted ROI and classification processes. Initially, only the “Input Image” pushbutton is enabled for selecting the mammogram for diagnosis. All other functions are disabled. The “Input Image” pushbutton allows the user to select the digital mammogram for diagnosis. As soon as the mammogram is selected, it will

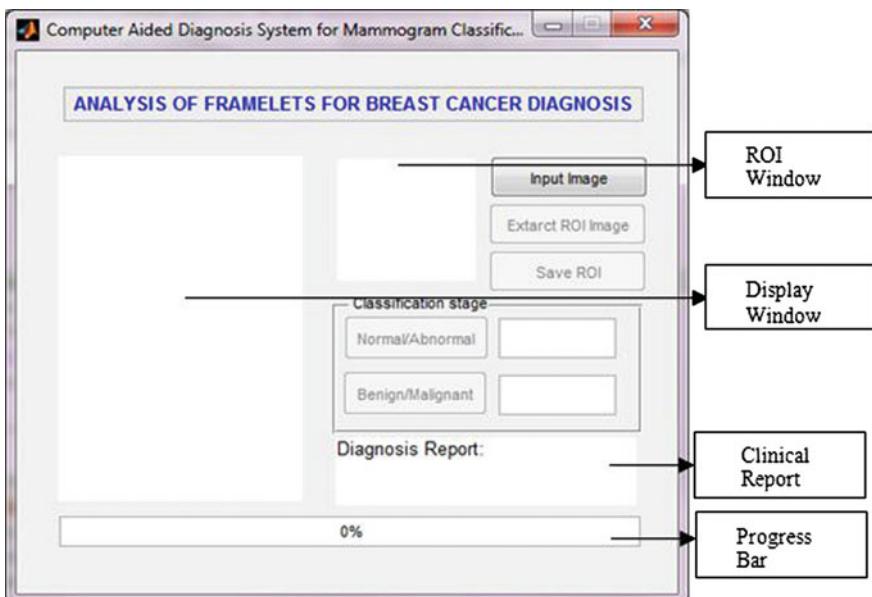


Fig. 2 A complete generalized frame window of the proposed system

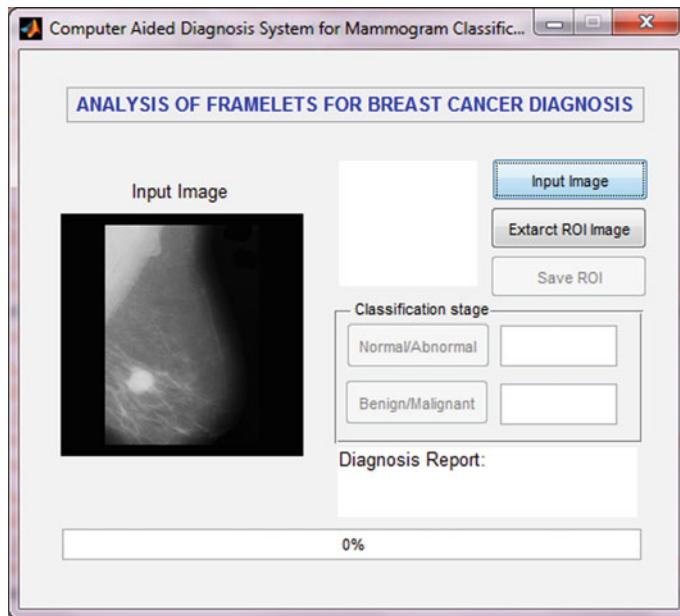


Fig. 3 A selected mammogram in the larger window

be displayed in the larger window of the frame which is shown in Fig. 3. Also the “Extract ROI Image” pushbutton is enabled.

When the “Extract ROI Image” pushbutton is clicked, the first module of the system, i.e. ROI extraction is executed that allows the user for selecting the suspicious region from original mammogram image. Figure 4 shows the screenshot of the ROI extraction module. The user has to click the approximate centre of the abnormality or the ROI that has to be diagnosed. This produces an ROI of size 256×256 pixels around the centre of abnormality as shown in Fig. 5. The “save ROI” pushbutton enables the user to save the extracted ROI for further analysis. After ROI extraction, the system is ready to analyze the abnormality in the ROI. To ease the analysis, the other modules of the proposed system such as feature extraction and classification stages are integrated into a single step.

To execute the stage I classification ‘Normal/Abnormal’ pushbutton and for stage II classification, “Benign/Malignant” pushbutton is designed. If the diagnosis of the given ROI is abnormal then only the stage II classification is enabled and the user has to click the “Benign/Malignant” pushbutton to find the abnormality of the ROI. Finally, the diagnosis report is also generated in the “Diagnosis Report” page. The progress bar in the frame visualizes the progression of the proposed cancer diagnosis system. Figures 6 and 7 show the entire system output for normal and abnormal cases, respectively.

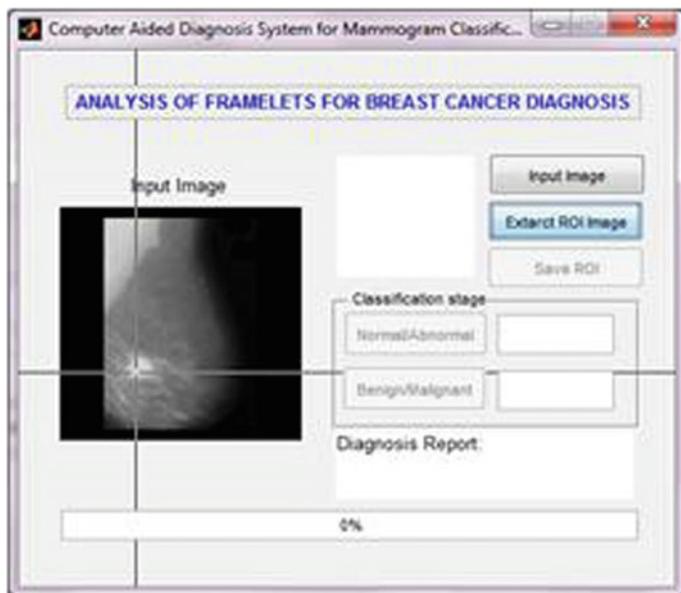


Fig. 4 Screenshot of ROI extraction process

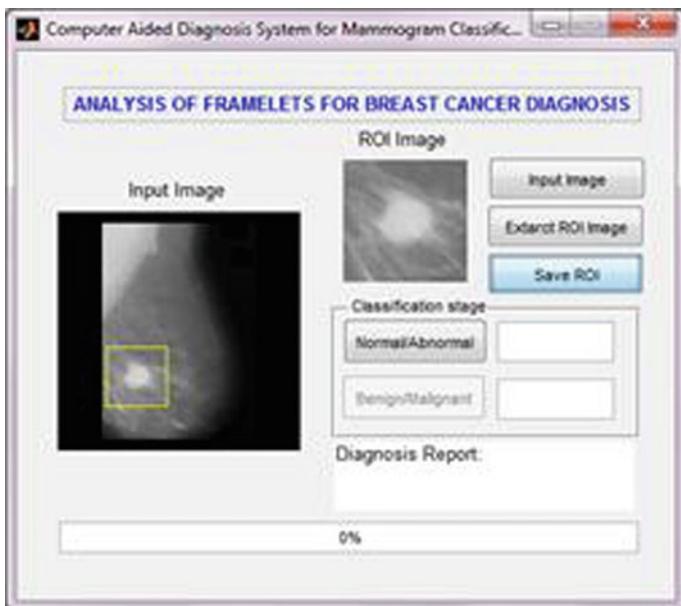


Fig. 5 Acquired digital mammogram and the selected ROI in the respective windows

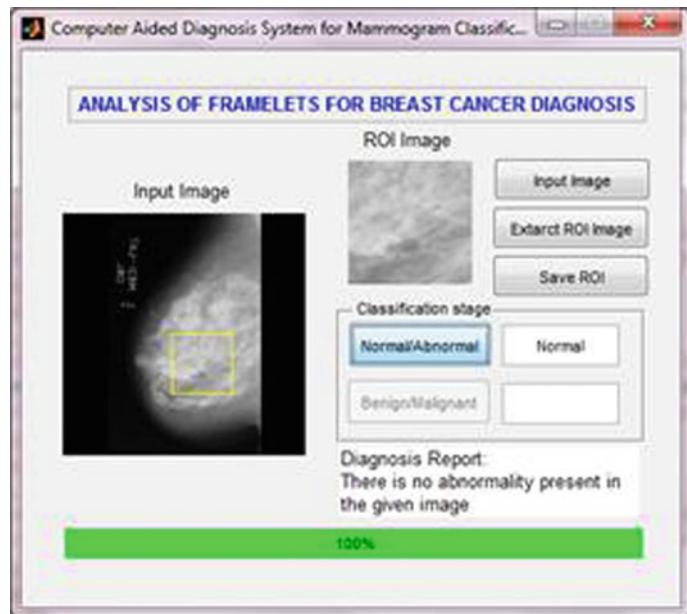


Fig. 6 Entire system output: normal mammogram

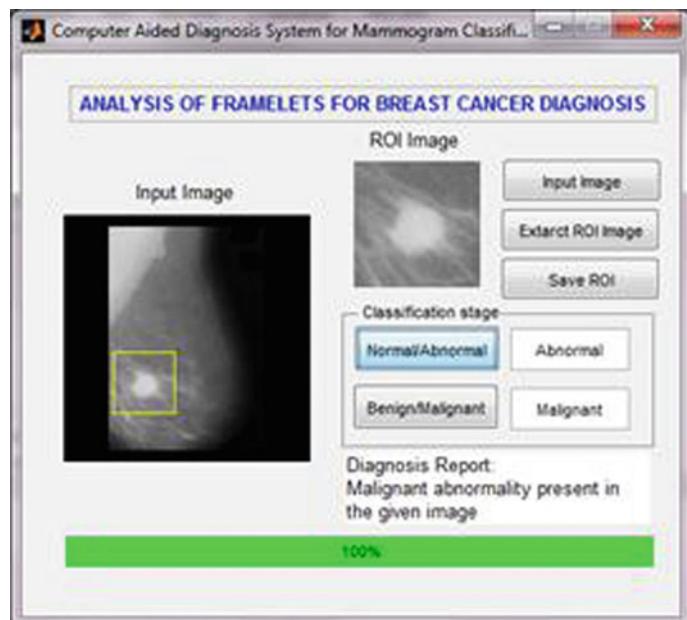


Fig. 7 Entire system output: abnormal mammogram

Table 2 Classification accuracy of the proposed microcalcification system using ROI

Framelet filter	Stage	Classification rate (%)							
		Decomposition level							
		1	2	3	4	5	6	7	8
Haar	I	81	81	85.13	85.54	87	88.53	86	93
	II	85	85	90.12	91	93	93	95.16	96.17
Piecewise linear	I	93	90	89.76	93	96	94	90	90
	II	96.83	100	100	100	97	97	94.31	94.31
Piecewise cubic	I	93	95.84	98	98	97	98	93.41	91
	II	100	100	100	95	94	93.24	91	91.32

4 Classification Accuracy Based on Framelet Transform

The framelet transform is used to represent the mammogram in multi-scale representation. The size of ROI extracted from the original mammogram image is 256*256. Up to eighth level the decomposition is available for the given image. In each level, the GLCM features extracted from each subbands are used as the feature vector of the corresponding mammogram (Table 2).

5 Conclusion

The proposed system performs well with the real time data collected from the hospital. In each fold, two-third of the total images are randomly selected to train the classifier while the remaining images are employed for testing. It can be concluded from the table that the extracted features based on piecewise cubic framelet produces higher accuracies than Haar and piecewise linear. The satisfying highest classification accuracies obtained are 98 % and 100 % for stage 1 and 2 classifier, respectively.

In order to take the work to next level, the application of the proposed system on all the four views of mammographic images can be done. In the feature extraction stage, other multiresolution and multidirectional analyses such as Shearlet transform, Contourlet transform can be used. In addition to GLCM feature extraction, other features such as statistical features, geometry features, etc., can also be calculated to create a more accurate database for reference thereby increasing the classification accuracy. By employing the above-mentioned changes, the biopsy method can be fully eliminated through helping the patients from pain and mental stress.

References

1. Ibrahima Faye and Brahim Belhaouari Samir, "Digital mammograms Classification Using a Wavelet Based Feature Extraction Method", IEEE conference on Computer and Electrical Engineering, 2009, pp 318–322.
2. Andy Tirtajaya and Diaz D. Santika, "Classification of Micro calcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine", IEEE International Conference on Advances in Computing, Control and Telecommunication Technologies, December 2010, pp 164–166.
3. R.N. Strickland & H.I. Hahn, "Wavelet Transform for detecting micro calcification in Mammograms", IEEE Transactions on Medical Imaging, Vol. 15, April 1996, pp 218–229.
4. K. Prabhu Shetty, Dr. V. R. Udupi, "Wavelet Based Microcalcification Detection on Mammographic Images", IJCSNS International Journal of Computer Science and Network Security, vol. 9 No. 7, July 2009, pp. 213–217.
5. Sara Dehghani and Mashallah Abbasi Dezfooli, "Breast Cancer Diagnosis System Based on Contourlet Analysis and Support Vector Machine", World Applied Sciences Journal, vol. 13 (5), 2011, pp 1067–1076.
6. Mohamed Meselhy Eltoukhy and Ibrahima Faye, "Curvelet Based Feature Extraction Method for Breast Cancer Diagnosis in Digital Mammogram", IEEE International Conference on Intelligent and Advanced Systems, June 2010, pp 1–5.
7. Jian-Feng Cai, Hui Ji, Chaoqiang Liu and Zuowei Shen, "Framelet-Based Blind Motion Deblurring From a Single Image", IEEE Transactions on Image Processing, vol. 21, no. 2, February 2012.
8. Jian-Feng Cai, Stanley Osher and Zuowei Shen, "Linearized Bregman Iterations for Frame Based Image Deblurring", SIAM Journal on Imaging Sciences vol. 2, no. 1, January 2009.
9. M. J. Choi, D. H. Lee and H. S. Lim, "Framelet-Based Multi Resolution Image Fusion with an Improved Intensity-Hue-Saturation Transform", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 37, Part B7, 2008.
10. Bin Dong and Zuowei Shen, "MRA Based Wavelet Frames and Applications: Image Segmentation and Surface Reconstruction", Proceedings of the International Society of Optics and Photonics", 2012.
11. Smola A. J., Scholkopf B., and Muller K. R., "The connection between regularization operators and support vector kernels", Neural Networks New York, vol. 11, November 1998, pg 637–649.
12. Liyang Wei, Yongyi Yang, Robert M. Nishikawa and Yulei Jiang," A Study on Several Machine –Learning Methods for Classification of Malignant and Benign Clustered Microcalcifications", IEEE Transactions on Medical Imaging, vol 24, No. 3, March 2005.
13. De Melo, C. L., Costa Filho, C. F., Costa, M. G & Pereira, W. C, "Matching input variables sets and feed forward neural network architectures in automatic classification of microcalcifications and microcalcification clusters", 3rd International Conference on Biomedical Engineering and Informatics (BMEI), Vol. 1, pp. 358–362, 2010.
14. Cascio, D. O. N. A. T. O., Fauci, F., Magro, R., Raso, G., Bellotti, R., De Carlo, F & Torres, E. L, "Mammogram segmentation by contour searching and mass lesions classification with neural network", IEEE Transactions on Nuclear Science, Vol. 53, No. 5, pp. 2827–2833, 2006.
15. B. S. Manjunath & W. Y. Ma (1996), "Texture feature for browsing and retrieval of image data", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. No. 18(8), pp: 837–842.
16. Steve R. Gunn,"Support Vector Machines for Classification and Regression", Technical report, University of Southampton, 10 May 1998.

Reconfigurable Architecture-Based Implementation of Non-uniformity Correction for Long Wave IR Sensors

**Sudhir Khare, Brajesh Kumar Kaushik, Manvendra Singh,
Manoj Purohit and Himanshu Singh**

Abstract Infra Red (IR) imaging systems have various applications in military and civilian sectors. Most of the modern imaging systems are based on Infra Red Focal Plane Arrays (IRFPAs), which consists of an array of detector element placed at focal plane of optics module. Performance of IRFPAs operating in Medium Wave Infra Red (MWIR) and Long Wave Infra Red (LWIR) spectral bands are strongly affected by spatial and temporal Non-Uniformity (NU). Due to difference in the photo response of detector elements within the array, Fixed-Pattern Noise (FPN) becomes severe. To exploit the potential of current generation infrared focal plane arrays, it is crucial to correct IRFPA for fixed-pattern noise. Different Non-Uniformity Correction (NUC) techniques have been discussed and real-time performance of two-point non-uniformity correction related to IR band is presented in this paper. The proposed scheme corrects both gain and offset non-uniformities. The techniques have been implemented in reconfigurable hardware (FPGA) and exploits BlockRAM memories to store the gain and offset coefficients in order to achieve real-time performance. NUC results for long-range LWIR imaging system are also presented.

1 Introduction

Infrared (IR) imaging systems are widely used in a variety of applications like remote sensing, surveillance, medical, fire and mine detection, etc. Largely, Infrared imaging systems are based on the Infra Red Focal Plane Array (IRFPA) [1], which consists of an array of infrared detector elements aligned at focal plane of the imaging system [2, 3]. Recently, there has been an increasing research in IR detector technologies that resulted in realization of large detector formats like

S. Khare (✉) · M. Singh · M. Purohit · H. Singh
Instruments Research and Development Establishment,
Dehradun 248008, India
e-mail: sudhir_khare@hotmail.com

B.K. Kaushik
Indian Institute of Technology, Roorkee 247667, India

640×512 , 1024×768 , etc., having smaller pitch and better thermal sensitivity. The performance of IR detector is strongly affected by several degrading factors such as the lens diameter causing blurred image, detector's photo response resulting in intensity loss, under sampling because of limited active area of each detector (limited pixel size), Poisson (shot) noise and the additive Johnson noise generated by the electrons. One of the most challenging and degrading effect is caused by random spatial and temporal photo response non-uniformity of photodetectors. Since each individual detector in the array has a different photo response under identical irradiance, due to mismatch of fabrication process, it results in fixed-pattern noise (FPN) or non-uniformity [4] superimposed on the true image. These fluctuations between pixels leads to degradations such as $1/f$ noise associated with detectors, corresponding readout input devices and the nonlinear dependence of the detector gain. This non-uniformity changes slowly in time with change in the FPA temperature, bias voltages, and scene irradiance. These changes reflect in the acquired image in the form of a slowly varying pattern superimposed on the image resulting in reduced resolving capability. NUC techniques normally assume a linear model for the detectors, characterizing thus the non-uniformity response problem as a gain (responsivity) and offset (detector-to-detector dark current) estimation problem.

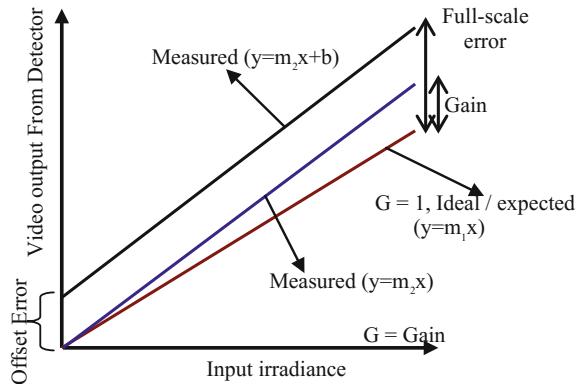
2 Non-Uniformity Correction (NUC): Concepts

The non-uniformity arises due to number of factors, prominent among which are the large variation in responsivity (gain) and detector-to-detector dark current (offset). The magnitude of the offset and gain variation depends on the two things: (i) the active material of IRFPA and (ii) the technology used to fabricate the FPA detector. The responsivity variations is the least ($\sim 1\%$) in case of PtSi Schottky barriers, but may be quite large ($\sim 10\%$) in case of MCT-based detectors.

Each detector element is associated with a fixed offset which is different for the different elements which is known as Fixed-Pattern Noise (FPN) offset. Gain of each detector element is not ideal ($\text{Gain} = 1$). It differs from pixel to pixel, due to which there will be variation in output of the particular element. These variation needs to be compensated. Non-uniformity between pixels values is represented by equation of line shown in Fig. 1.

Gain of detector element is represented as slope of the line (i.e., m_1) passing through origin (assuming zero FPN offset) thus for $m_1 = 1$ no correction is required to the detector output. For gain other than unity the detector output will have to be multiplied by the reciprocal of the gain value for that element to get correct pixel value. Similarly FPN offset in the detector output represented as b , i.e., offset of the line with reference to origin on y-axis. Hence, to get the correct pixel value, the FPN offset will be subtracted through the detector output. Thus, the NUC includes both offset and gain compensation.

Fig. 1 Offset, gain, and full-scale errors



Several NUC techniques have been tried out to overcome the problem of fixed-pattern noise in IR detector array. Keeping vast application area of infrared imaging system, a continuous development is in progress to improve the NUC techniques. Mainly there are two types of NUC techniques: (i) Calibration method-based [5] NUC techniques (ii) Scene-based [6–8] NUC techniques.

2.1 Calibration-Based Techniques

To correct for non-uniformities, simplest and most accurate methods is calibration-based method. Single point correction (SPC), two-point correction (TPC) and multiple point correction (MPC) methods are common method which fall under calibration-based techniques. Parameters like gain and the offset are estimated by exposing FPA to a uniform IR radiation source at one or more temperatures. The response of the individual pixels is recorded simultaneously to calculate gain and the offset coefficients. In case of TPC and MPC, two and more temperatures are used, respectively, to compute gain and offset coefficient. These coefficients are stored in suitable format and then used to compensate for the non-uniformity using associated sensor on-board electronics. The performance of the present method is optimal when the detector response varies linearly and is time invariant between the calibration temperatures.

2.2 Scene-Based Non-uniformity Compensation

The scene-based non-uniformities compensation [9] uses different image processing algorithms by exploiting change in the actual scene-related features or the motion in order to compute coefficients of scene temperature per detector. The true image from the fixed-pattern noise-affected scene is generated by compensating these scene-based coefficients. Statistically, the temperature diversity provides a reference

point which is common to all detectors. The detectors response can be normalized for the non-uniformity based upon this reference point calculation. These algorithms are difficult to implement in real time and they do not provide the required radiometric accuracy. Since the scene-based NUC algorithms normally use motion as one of the criteria for separating the true image from the FPN, these algorithms usually leave artifacts in the image due to presence of non-moving objects, which are required to be corrected algorithmically.

3 The Mathematical Model of Calibration-Based Techniques

In order to provide completeness, calibration-based NUC been presented. Computational complexity and implementation strategy of these models have also been presented.

3.1 Single Point Correction (SPC)

The SPC method is used to correct the offset of every pixel in the IRFPA. This is performed by placing a uniform IR radiation source in front of the imager lens. Using one point correction, the fixed-pattern noise will be minimum at the reference temperature and with perfect correction [3]; there will be no fixed-pattern noise at reference temperature (Fig. 2). The residual FPN is produced due to the different spectral response of detectors along with the truncation errors in the normalization algorithm. Fixed-pattern noise tends to increase as the background temperature deviates from the reference calibration temperature. This increase depends upon how far the detector responsivity curves deviates from linearity. This method is used to update an existing NUC and can be performed in the field environment easily.

Fig. 2 FPN after single point correction

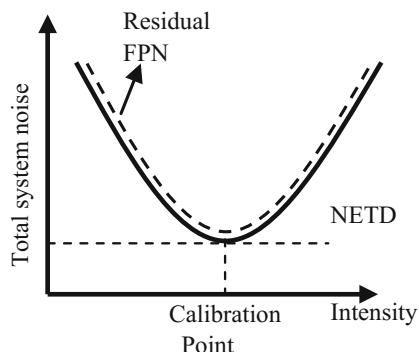
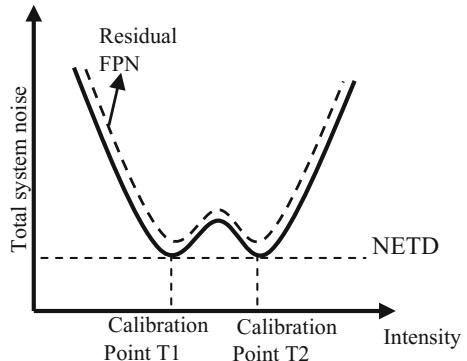


Fig. 3 FPN after two-point correction



3.2 Two-Point Correction (TPC)

The most common and widely used method to correct for non-uniformity of IRFPAs is the TPC method. In two-point correction method [10, 11], the spatial noise will be minimum at two reference intensities and increases for other intensities. There is a curve known as W-curve (Fig. 3) where, in the region between two references, the spatial noise is less compared to the spatial noise outside two references. This method uses two uniform IR sources at two different temperatures (T_1 and T_2) to estimate the gain and offset of each detector element to compensate the non-uniformity.

3.3 Multiple Point Correction (MPC)

To cater for wide operating temperature ranges, multi point correction technique [12] is most suitable method for non-uniformity correction. MPC also known as piecewise-linear correction method is an extension of the two-point method where, a number of different temperature points are used to divide the nonlinear response curve into several piecewise linear sectors to correct for non-uniformity. The fixed-pattern noise may be additive or multiplicative, for arrays with dark currents, the noise powers are additive and arrays with different responsivities produce multiplicative noise. The response of detector element in an FPA is nonlinear in nature, but it is modeled as a linear response having a multiplicative gain and an additive offset.

A two-point non-uniformity correction assumes that the value of each pixel can be corrected by multiplying it by gain and adding an offset to it. The measured signal Y_{ij} for (ij) th detector element in the FPA at given time t can be expressed as:

$$Y_{ij}(t) = \alpha_{ij}(t) \cdot X_{ij}(t) + \beta_{ij}(t) \quad (1)$$

where, $\alpha_{ij}(t)$ and $\beta_{ij}(t)$ are the gain and offset of the (ij) th detector element, and $X_{ij}(t)$ is real irradiance received by the detector element.

From Eq (1), the real incident radiation (irradiance) is given by

$$X_{ij}(t) = \frac{Y_{ij}(t) - \beta_{ij}(t)}{\alpha_{ij}(t)} \quad (2)$$

Now to perform 2 point calibration, IR imaging system captures the images corresponding to lower and higher temperature from uniform radiation source (blackbody).

Defining $\alpha_{ij}(t)$ [11]

$$\alpha_{ij}(t) = \frac{T_{2ij} - T_{1ij}}{M_2 - M_1} \quad (3)$$

$$\beta_{ij}(t) = M_1 - \alpha_{ij}(t) \cdot T_{1ij} \quad (4)$$

$$M = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n T_{ij} \quad (5)$$

where $T_{1ij}(t)$ & $T_{2ij}(t)$ are (ij) th detector element intensities at lower and higher temperatures at time t [4]. M_1 and M_2 are mean intensities (mean signal output) of the all detector elements in one frame (76,800 values in 320×256 FPA) at lower and higher temperatures. Corrected output of (ij) th detector element can be obtained from Eq. (1) using values of $\alpha_{ij}(t)$ and $\beta_{ij}(t)$ calculated from Eqs. (3) and (4).

In staring IR focal plane arrays, each detector will have different gain and offset coefficient and this variation produces fixed-pattern noise. Figure 4 shows signal outputs of different detectors for same input intensities.

Fig. 4 Effect of fixed-pattern noise before correction

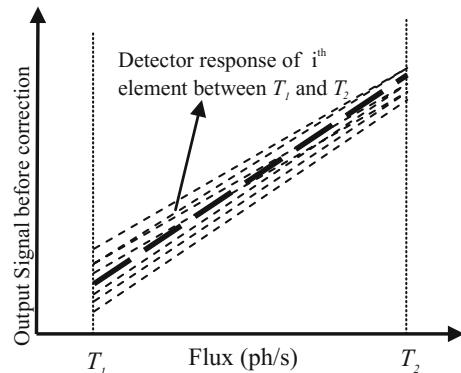


Fig. 5 Effect of fixed-pattern noise after correction

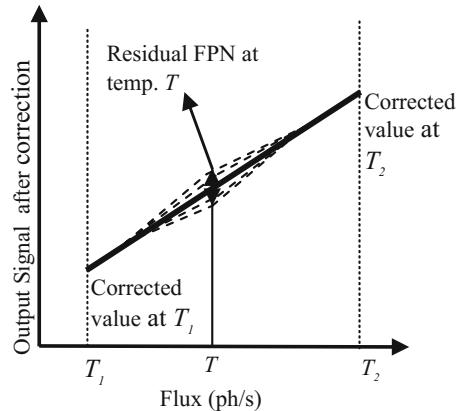


Figure 5 illustrates the normalized output after correction at two points. Fixed-pattern noise will be minimum at two reference temperatures T_1 and T_2 ; it increases for any other reference temperature. If all detectors had linear responsivities, then all the curves would coincide (as shown in Fig. 5), spatial noise is minimum between T_1 and T_2 . The residual spatial noise is present at temperature T .

4 Application to LWIR Imaging System

The two-point NUC scheme is implemented under the present scope of work and tested on LWIR cooled Imaging system based on 320×256 MCT-based IR focal plane array.

Design parameters of LWIR Imaging System

- Spectral band: 8–12 μm (LWIR)
- Detector: 320×240 MCT IRFPA (cooled)
- F-number: 2
- Aperture Dia: 130 mm
- Spatial Resolution: 115 μrad
- Video output: CCIR-B, 50 Hz

The LWIR imaging system electronics board generates detector interface signals, performs two-point non-uniformity correction, image processing tasks and finally generates CCIR-B compatible video output. Two sets of image data are captured at 10°C and higher temperature 35 °C, respectively, with integration time of 20 μs . Twelve image frames at each lower and higher temperature are acquired to correct the temporal noise and used to compute gain and offset coefficient. These gain and offset coefficients are used to correct the uncorrected image data. Figure 6a shows the raw IR image and Fig. 6b shows the image after NUC. Figure 7a, b shows the 3-Dimensional representation (histogram) of image data before and after NUC. Figure 8a, b illustrates the IR image from LWIR imaging system before and

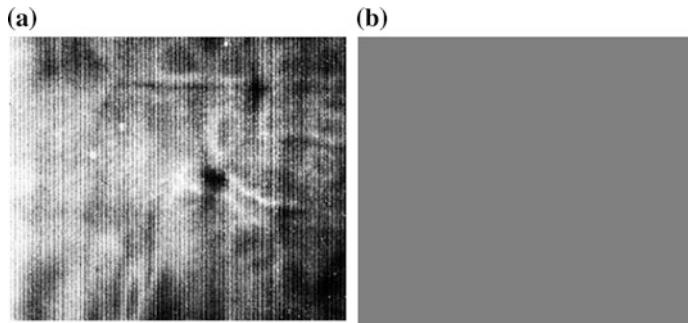


Fig. 6 Image frame **a** before and **b** after two-point NUC

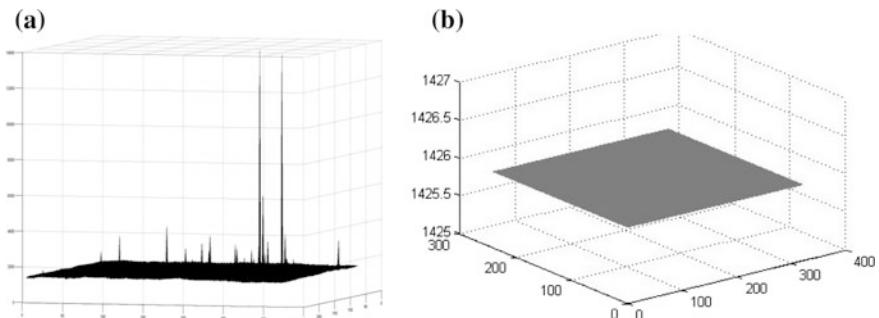


Fig. 7 3-Dimensional plot **a** raw data and **b** data after NUC

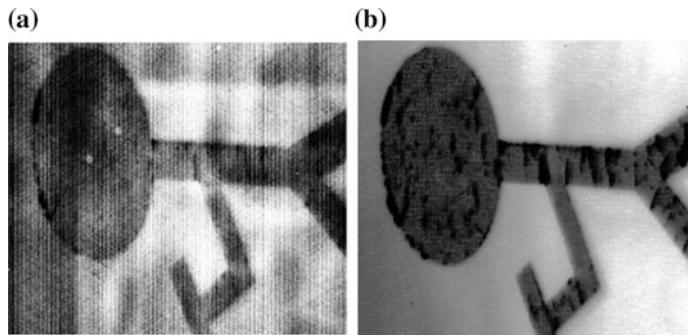


Fig. 8 Image frame **a** before and **b** after two-point NUC

after two-point NUC. To determine the effectiveness of given NUC method, the Image quality measurements which are performed on the corrections (after NUC) are crucial. This method is offline calibration at factory level, thus, different tables for different temperature ranges are stored in Look up Tables (LUTs) and user can select the desired table as per field environmental conditions.

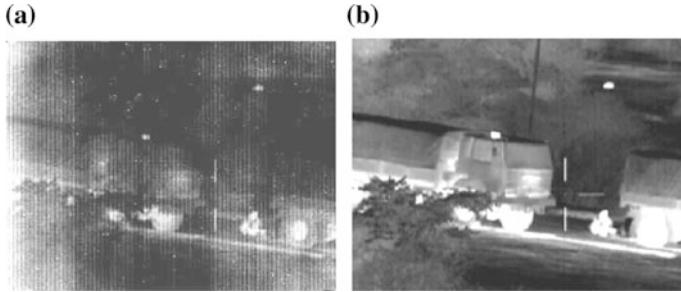


Fig. 9 Image frame **a** before and **b** after two-point NUC

Figure 9a, b illustrates the result of another IR image before and after two-point NUC.

In the present work, Residual Fixed-Pattern Noise (RFPN) [13, 14] is used to measure the NUC capability of the proposed method.

$$\text{RFPN} = \frac{\text{Standard Deviation (Output level)}}{\text{Mean (Output level)}} \quad (6)$$

$$\text{RFPN} = \frac{SD}{M} = \frac{1}{m.n} \sqrt{\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - y_{ij})^2} \quad (7)$$

where x_{ij} is corrected image, y_{ij} is reference two-point calibrated image, and (m, n) is total number of pixels in the image frame.

After NUC correction, Measured Standard Deviation (σ) = 64.1784

Mean (M) = 1452

So, measured Residual Fixed-Pattern Noise $RFPN = \frac{\sigma}{M} = 0.0442$

5 FPGA-Based Hardware Implementation

The NUC correction implemented in present work is based upon classical two-point NUC correction algorithms. FPGA-based Architecture of prototype hardware is shown in Fig. 10. IR detector having array of size 320×240 producing 14-bit digital data is exposed to a uniform IR source, i.e., blackbody at lower and higher temperature. Raw video digital data at different temperatures is stored in the SRAM through a serial link, this data is used to calculate the offset and gain coefficients. This is a time consuming and complex task to perform, where around 76,800 pixels with 14 bit data are processed. Since, these values are only valid for a given ambient temperature, the gain and offsets coefficients are stored in two flash

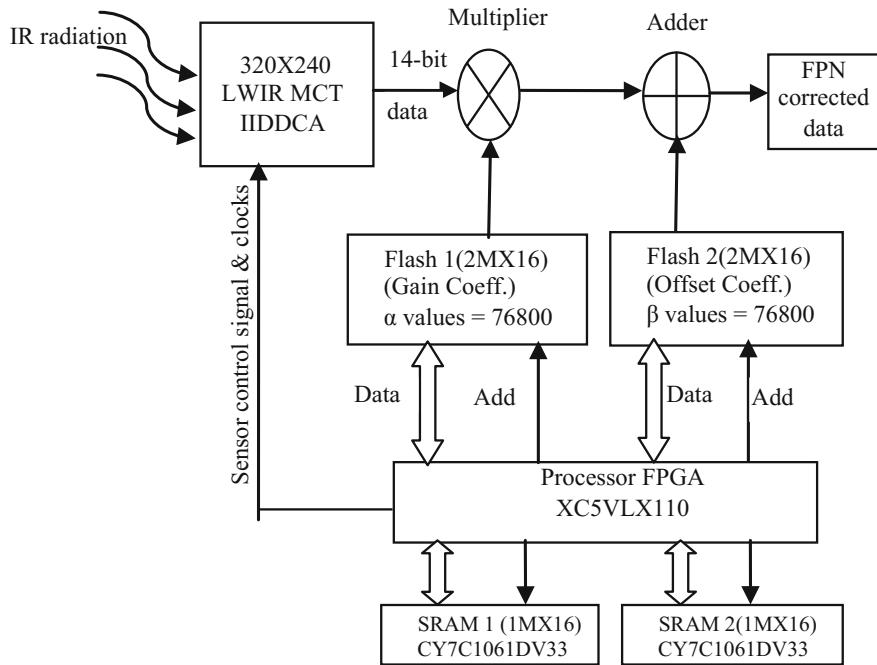


Fig. 10 FPGA-based hardware implementation of two-point NUC

memories having the capacity of at least one image frame. The implementation is carried out in an FPGA (Xilinx XC5VLX110)-based electronics board [15, 16]. Flash memory controller detector interfaces are designed in RTL using VHDL [17] and processing module to apply these coefficients on incoming data has also been designed in VHDL (employing Xilinx ISE tool). The data path and control path is designed to exploit the parallelism within the processing block to achieve real-time performance. Resource utilization of the targeted device is given in Table 1.

Table 1 Device utilization summary

Device utilization summary			
Logic utilization	Used	Available	Utilization (%)
Number of slice flip flops	4,578	69,120	6
Number of occupied slices	2,159	17,280	12
Number of slices LUTs	5,006	69,120	7
Number of BUFG/BUFGCTRLs	7	128	28
Number of BlockRAM/FIFO	20	200	5
Number of DSP48Es	11	64	17
Number of DCM_ADVs	2	12	16

6 Conclusion

In the present paper, an approach namely blackbody calibration to perform NUC using two-point method is implemented on LWIR 320×240 IRFPA-based imaging system. There is significant reduction in FPN in the output images obtained after NUC. The experimental results confirm that the two-point calibration-based methods using uniform IR radiation source are effective and efficient for real-time correction of fixed-pattern noise visible due to non-uniformity. The real-time implementation of FPGA-based hardware architecture and realization using the 2-point NUC algorithm is also given for LWIR imaging system.

Acknowledgments The authors would like to thank Director, IRDE Dehradun and Head, Department of Electronics & Communication Engg, IIT Roorkee for their support in carrying out this work.

References

1. D. Scribner, M. Kruer, and J. Killiany. Infrared focal plane array technology. Proceedings of the IEEE, 79(1) (1991) 66–85.
2. J. M. Lloyd, Thermal Imaging Systems, Plenum Press, New York (1975).
3. G. Holst, Electro-Optical Imaging System Performances, SPIE Press, Bellingham (1995).
4. D. Scribner, M. Kruer, and C. Gridley. Physical limitations to nonuniformity correction in focal plane arrays. Proceedings of SPIE, 865 (1987) 185–201.
5. Daniel Pipa, Eduardo A. B. da Silva, Carla Pagliari, and Marcelo M. Perez, Joint Bias And Gain Nonuniformity Correction Of Infrared Videos, IEEE, (2009) 3897–3900.
6. Yang Weiping, Zhang Zhilong, Zhang Yan, and Chen Zengping, A Novel Non-Uniformity Correction Algorithm Based On Non-Linear Fit, International Scholarly and Scientific Research & Innovation Vol: 6 2012-12-23 (2012).
7. J. Harris and Y. Chiang. nonuniformity correction of infrared image sequences using the constant-statistics constraint. IEEE Transactions on Image Processing, 8(8), August (1999) 1148–1151.
8. Lixiang Geng, Qian Chen, Weixian Qian, and Yuzhen Zhang, Scene-based Nonuniformity Correction Algorithm Based on Temporal Median Filter, Journal of the Optical Society of Korea Vol. 17, No. 3, (2013) 255–261.
9. Lixiang Geng *et al.*, “Scene-based nonuniformity correction algorithm based on temporal median filter,” *Journal of the Optical Society of Korea*, vol. 17, No. 3, pp. 255–261, 2013.
10. A. Friedenberg, & Goldblatt, I. Non uniformity two point linear correction errors in infrared focal plane arrays. Opt. Eng., 3(4), (1998) 1251–1253.
11. A. Kumar, S.Sarkar, R.P. Agarwal. A novel algorithm and FPGA based adaptable architecture for correcting sensor non-uniformities in infrared system Elsevier, Microprocessor and Microsystems, 31 (2007) 402–407.
12. Yang Weiping *et al.*, “A novel non-uniformity correction algorithm based on non-linear fit,” *International Scholarly and Scientific Research and Innovation*, vol. 6, pp. 404–407, 2012.
13. V. N. Borovytsky, Residual error after non uniformity correction, Semiconductor physics, quantum electronics and opto-electronics 3 (1), (2000) 102–105.
14. Liu Huitong, Wang Qi, Chen Sihai, Yi Xinjian, Analysis of the residual errors after non uniformity correction for Infrared focal plane array, IEEE Conferences (2000) 213–214.

15. J. A. Kalomiros and J. Lygorus, Design and Evaluation of a hardware/software FPGA based systems for fast image processing, Elsevier, Microprocessor and Microsystems, Vol 32, (2008) 95–106.
16. Domingo Benitez, Performance of reconfigurable architectures for image-processing applications, Journal of Systems Architecture, Vol. 49, (2003) 193–210.
17. Skahill, K. VHDL for programmable logic (Book), Addison Wesley, CA, USA, (2006).

Finger Knuckle Print Recognition Based on Wavelet and Gabor Filtering

Gaurav Verma and Aloka Sinha

Abstract Biometric systems are used for identification- and verification-based applications such as e-commerce, physical access control, banking, and forensic. Among several kinds of biometric identifiers, finger knuckle print (FKP) is a promising biometric trait in the present scenario because of its textural features. In this paper, wavelet transform (WT) and Gabor filters are used to extract features for FKP. The WT approach decomposes the FKP feature into different frequency subbands, whereas Gabor filters are used to capture the orientation and frequency from the FKP. The information of horizontal subbands and content information of Gabor representations are both utilized to make the FKP template, and are stored for verification systems. The experimental results show that wavelet families along with Gabor filtering give a best FKP recognition rate of 96.60 %.

Keywords Biometric recognition • Finger knuckle print • Wavelet transform • Gabor filter

1 Introduction

In real-life applications, biometrics-based information security systems are much more advanced and reliable over knowledge-based systems because of its ability to access information with authenticity of a person [1]. Biometric features have been widely used for recognition systems. In the field of biometric research, researchers have investigated new biometric trait known as finger knuckle print [2]. The FKP refers to the morphological features around the phalangeal joints when the fingers

G. Verma (✉) · A. Sinha

Department of Physics, Indian Institute of Technology Delhi, New Delhi, India
e-mail: gaurav.sgs85@gmail.com

A. Sinha
e-mail: aloka@physics.iitd.ac.in

are slightly bent. These textural features are unique and highly distinctive and have the potential to discriminate between individuals. In comparison to other biometrics, it has several advantages such as easy accessibility, requires contact less image acquisition systems, does not involve any kind of emotions or expressions, and provides stable recording of features. In addition, it also has no stigma associated with criminal investigation, and the user acceptance is high because of easier implementation in real-time applications. Recently, some researchers have investigated the utility of the FKP features for person identification and verification [3–5]. It is found from reported work that multiresolution analysis of the FKP has not been explored for the design of the recognition system.

Wavelet analysis is one of the useful techniques for multiresolution analysis [6, 7]. This is used to decompose the image into several frequency subbands at different resolutions. Many researchers have investigated biometric-based recognition system using the wavelet [6–8]. The attractiveness of wavelet analysis also includes low computational cost but it does not extract directional information. In order to extract directional feature, Gabor filters is an important tool that shows features at different orientations and frequencies of the image [9, 10]. The inclusion of directional features can help in improving the recognition performance.

In this paper, the WT and Gabor filtering are used for FKP verification system. The features of subband using WT carries useful and distinct features and Gabor representation give different orientation and frequency information, which is utilized for FKP recognition. The proposed scheme exhibits improved verification performance with a low value of EER.

2 Methodology

In this section, wavelet and Gabor filtering-based schemes are briefly described.

2.1 Wavelet Scheme

Discrete wavelet transform (DWT) decomposes the image into approximation and details subbands using filter banks [7]. These filter banks consist of low pass filters and high pass filters, which is followed by subsampling operation by a factor of two ($\downarrow 2$) as shown in Fig. 1.

Here, the ‘LL1’ subbands are obtained through a series of low pass filtering operation and show coarser representation of the original FKP image, which possesses lower frequency components [6]. The ‘LH1’ subbands are obtained by a sequence of low pass and high pass filtering which provides extraction of horizontal information of FKP. The ‘HL1’ subband is the result of applying high pass filtering followed by low pass filtering. It provides vertical direction details of FKP images. The ‘HH1’ subband is identified through sequentially high pass filtering operation

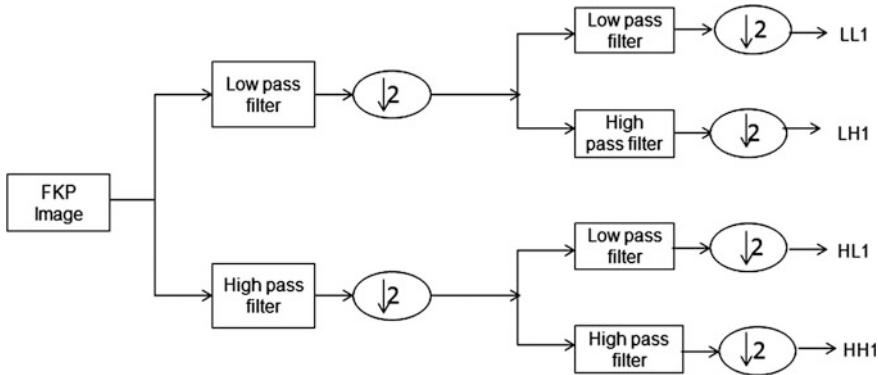


Fig. 1 Representation of one level 2D-DWT decomposition of FKP features

and gives high frequency characteristics of FKP. This region is known as diagonal detail. This process is implemented iteratively to represent more decomposition levels of desired resolution.

2.2 Gabor Filter

Gabor filtering is a useful tool for feature extraction and texture classification in image processing and computer vision application [9–11]. The Gabor filters are designed by combining complex sinusoidal wave and Gaussian functions as:

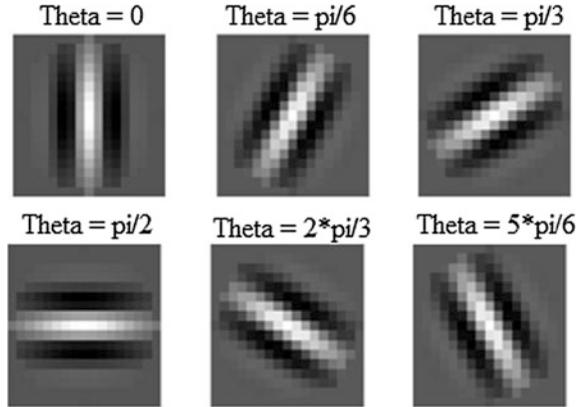
$$G(x, y, \sigma, \Theta, \gamma) = \exp(-x'^2 + \gamma^2 y'^2 / 2\sigma^2) \cdot \exp(i(2\pi x' / \lambda + \psi)) \quad (1)$$

where, $x' = x\cos\Theta + y\sin\Theta$, $y' = -x\sin\Theta + y\cos\Theta$, λ is the wavelength of sinusoidal wave, σ represents the width of the Gaussian function, ψ represents the phase offset, and γ is known as spatial aspect ratio of the Gabor function. The orientation of an image can be obtained as $\Theta_i = i\pi/6$, ($i = 0\text{--}5$). Therefore, the design of Gabor filters requires a set of filter parameters to capture the orientation and spatial frequencies, which are listed in Table 1 and the designed filters are shown in Fig. 2.

Table 1 The Gabor filters parameter

S. no.	Size	Θ_i	λ	σ	ψ
1	17×17	0	6	2.5	0
2	17×17	$\pi/6$	6	2.5	0
3	17×17	$\pi/3$	6	2.5	0
4	17×17	$\pi/2$	6	2.5	0
5	17×17	$2\pi/3$	6	2.5	0
6	17×17	$5\pi/6$	6	2.5	0

Fig. 2 The six orientations of Gabor representation



3 Proposed Scheme

The block diagram of the proposed system is shown in Fig. 3. The FKP images after capturing undergo preprocessing then the feature extraction is done by using wavelet families and Gabor filtering. The feature vector is designed based on the energy of the wavelet subband and orientation of the Gabor filter. Finally, decisions are made based on the matching parameter.

3.1 Preprocessing

In order to extract region of interest (ROI) of the FKP images, the preprocessing steps are used, which is described in [5]. The original size of FKP is 384×288 and the size of obtained ROI of FKP image is set to be 110×221 along the X-axis and the Y-axis, as shown in Fig. 4a, b, respectively.

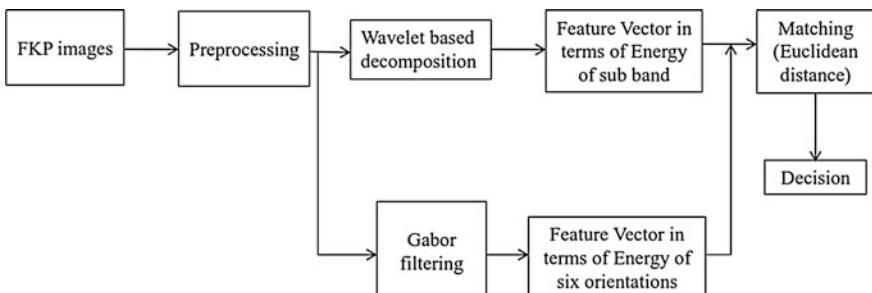
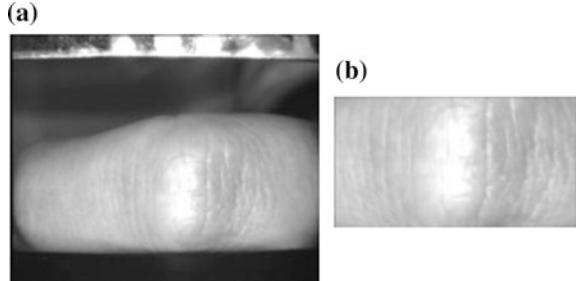


Fig. 3 Block diagram of proposed FKP verification system

Fig. 4 **a** Original FKP image. **b** Extracted ROI of FKP image



3.2 Feature Extraction of FKP Using Wavelet

Opted wavelets such as Haar, Daubechies, Symlets, Coiflet, discrete Meyer, and Biorthogonal are applied to decompose the FKP images into four subbands with different frequency contents till the third decomposition level. If the resolution of an input FKP image is 110×221 , the subbands LH1, HL1, HH1, and LL1 at the first decomposition level are of size 55×111 , the subbands LH2, HL2, and HH2 are of size 28×56 at the second decomposition level and the subbands LH3, HL3, and HH3 are of size 14×28 at the third decomposition level. The representation of four decomposed subbands of FKP for the Haar wavelet till the third level is shown in Fig. 5. Similar, subband decomposition is carried out by using the other opted wavelets. As can be seen from Fig. 5, the different subbands of FKP wavelet images carry different textures and feature details, which are further utilized for the design of the feature vector.

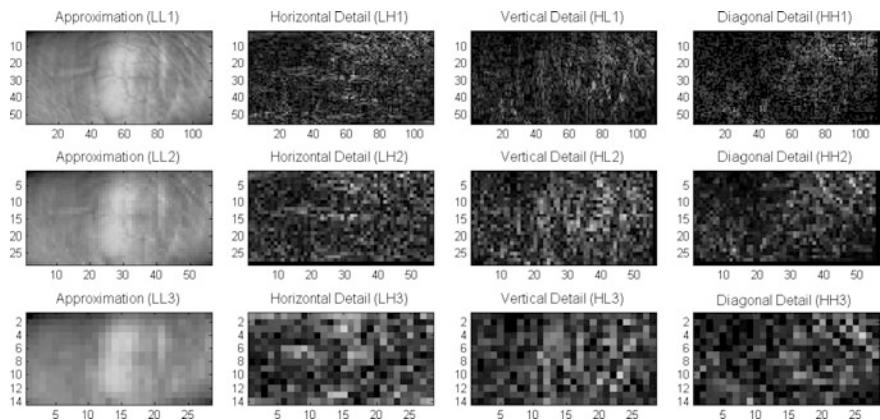


Fig. 5 Subbands of FKP images at all decomposition level for the Haar wavelet

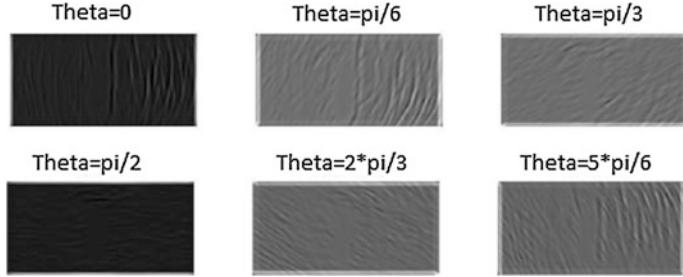


Fig. 6 Six FKP orientations and frequencies of Gabor filtered FKP image

3.3 Feature Extraction of FKP Using Gabor Filter

Each FKP image is convolved with the designed 2D Gabor. As a result, this captures features at different orientations and frequencies of FKP image. Figure 6 shows six Gabor representation of one FKP image and constitutes the texture features of FKP images and are further coded into the feature vector.

3.4 Performance Evaluation

To investigate the performance, preliminary studies have been carried out by choosing several parameters such as mean, variance, entropy, and energy to characterize the subband information [8]. However, energy is found to be a very efficient parameter to characterize and classify the features of the FKP. It is analyzed that horizontal subband provides the most distinguishing features as compared to other subbands, which are utilized to encode. The energy of the subband at the selected level is expressed as:

$$E_{\text{subband level}(k)} = \sum \sum |X_{\text{subband}}(i,j)|^2 \quad (2)$$

where, X_{subband} represents the decomposed subbands and ‘ k ’ denotes the number of decomposition level. Thus, the energy of horizontal subband of FKP at the first, second, and third sub-levels are calculated as $E_{WT1} = [E_{LH1}]$, $E_{WT2} = [E_{LH1}; E_{LH2}]$, and $E_{WT3} = [E_{LH1}; E_{LH2}; E_{LH3}]$. In addition, we have also calculated energy of six Gabor texture features of FKP image using (2).

$$E_{\text{Gabor}} = [E_0; E_{\pi/6}; E_{\pi/3}; E_{\pi/2}; E_{2\pi/3}; E_{5\pi/6}] \quad (3)$$

where E_{Gabor} represents the feature vector of FKP for six different orientations.

Finally, we have combined the resulting feature of WT subband at the third decomposition level and Gabor filter as:

$$E_{\text{Feature vector}} = [E_{\text{WT3}}; E_{\text{Gabor}}] \quad (4)$$

The obtained feature vectors are processed for matching through the Euclidian distance (ED) [7], which is used to measure the similarity between the feature vectors of the test image with the stored template. If $d_i(x)$ is the feature vector for the test FKP image ‘x’ and $d_i(y)$ is also the feature vector of FKP class ‘y’ from the database, then ED is expressed as:

$$ED = \sqrt{\sum_{i=1}^N (d_i(x) - d_i(y))^2} \quad (5)$$

where N is the number of features in the feature vector set.

4 Results and Discussion

Numerical experiments have been carried out on the Poly U FKP database to demonstrate the performance of the proposed systems. This database contains FKP samples of 165 subjects. In this analysis, 500 FKP images of the 50 subject are chosen for the proposed system [12]. We conducted experiments on Intel Core (TM) i7-3770 processor with 3.40 GHz and 8 GB RAM, Windows7, and MATLAB R2013a platform. The performance of the system is evaluated in terms of the equal error rate (EER) and receiver operating characteristics (ROC) curve. The ROC curve is plotted in between the false acceptance ratio (FAR) and the false rejection ratio (FRR) [1]. The values of FAR and FRR are determined by varying the value of threshold of ED.

4.1 Experiment 1

In this experiment, an extensive analysis has been carried out for the FKP verification by using wavelets. Each FKP image of the subject is matched against all the FKP images of the concerned subject. Therefore, the number of matching score is generated as 500 (50×10) for genuine and 24500 (50×490) for imposter. A ROC curve has been plotted for the varying threshold of matching parameter and is shown in Fig. 7.

The experimental results are summarized in Table 2 in terms of recognition rate for the different wavelets. The obtained results show that different wavelet families provide similar recognition rate because all the wavelets are extracting similar features due to band pass filtering. It can be seen that the best recognition rate (92.89 %) at third level decomposition is achieved for Haar wavelet. This recognition rate is similar to Bior6.8, Bior2.8, and Coif4 wavelet families at the same

Fig. 7 ROC curve for Haar wavelet

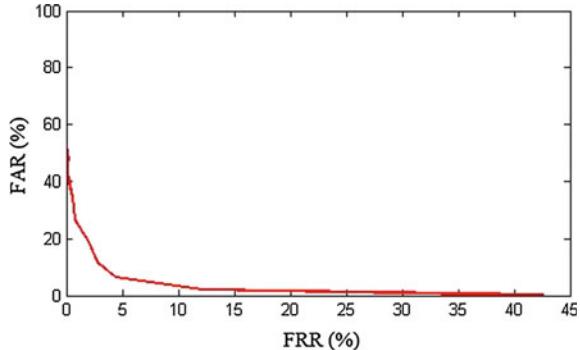


Table 2 Recognition rate for the FKP verification at different decomposition levels

S. no.	Wavelet	First level (%)	Second level (%)	Third level (%)
1	Harr	90.07	91.76	92.89
2	Db2	89.37	91.82	92.61
3	Db4	88.29	90.81	91.15
4	Dmey	86.66	88.67	90.75
5	Sym7	87.95	88.44	91.51
6	Sym8	88.45	88.85	90.90
7	Coif4	87.81	88.74	91.37
8	Coif5	87.76	87.88	91.49
9	Bior2.8	89.66	86.41	92.12
10	Bior6.8	88.51	86.97	91.53

decomposition level. It can also be seen that the FKP recognition rate of the first decomposition level is lower than the second level, and it continuously increases till the third decomposition level. The maximum recognition rate is achieved at third decomposition level for all the wavelet families considered as shown in Table 2.

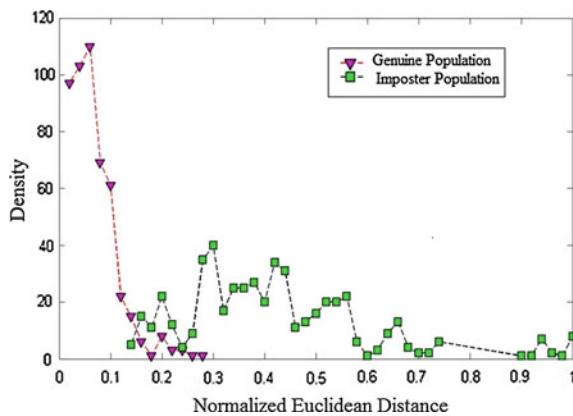
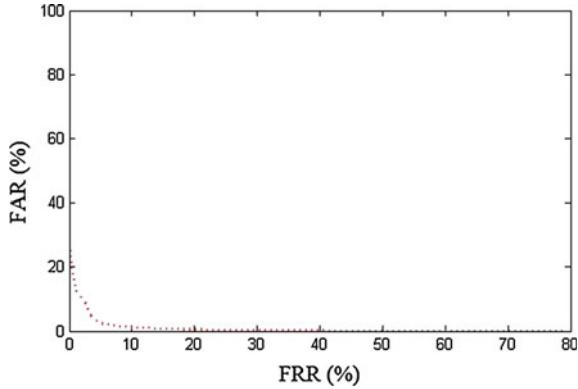
4.2 Experiment 2

The aim of this experiment is to analyze FKP recognition performance by using features from wavelet and Gabor filters at the third decomposition level. The normalized ED is plotted against population density for genuine and imposter as shown in Fig. 8.

It can be observed that ED distribution for the genuine and imposter are separated and most of the genuine are well discriminated from the imposter population. For EER calculation, a ROC curve has also been plotted for Haar wavelet as shown in Fig. 9.

Table 3 Recognition rate of combined scheme for the FKP verification

S. no.	Wavelet filter	WT at third level (%)	Gabor + WT at third level (%)
1	Harr	92.89	95.26
2	Db2	92.61	95.37
3	Db4	91.15	96.26
4	Dmey	90.75	95.14
5	Sym7	91.51	96.30
6	Sym8	90.90	95.55
7	Coif4	91.37	95.74
8	Coif5	91.49	95.91
9	Bior2.8	92.12	95.77
10	Bior6.8	91.53	95.58

Fig. 8 Graph between population density and normalized Euclidean distance**Fig. 9** ROC curve of combined scheme for Haar wavelet

Experimental results for combined scheme are presented in Table 3 in terms of recognition rate in comparison to WT subbands at the third level. It is observed from Table 3 that the recognition rates have considerably improved for each

wavelet families by inclusion of orientational features. The best-achieved recognition rate is 96.30 %.

As compared to the previous method [2], the proposed method has improved recognition rates. Although, the technique reported in [3] shows good results in comparison to the proposed method but the computational complexity increases due to a series of transforms that are used to extract the features of the FKP. In the proposed scheme, wavelet method has been used to extract the few subband features while six Gabor filters are used to investigate the FKP features at six different orientations. As a result, the recognition rate is considerably improved by inclusion of orientation features. So, the possibility of utilizing the hidden information in a variety of directions can be explored for FKP recognition. This will lead to achieve higher recognition rates accompanied with low computational complexities.

5 Conclusion

In this paper, a wavelet and Gabor-based method is proposed for feature extraction of FKP. The wavelet transform is employed for FKP decomposition into different subbands such as horizontal, vertical, and diagonal subbands. It is observed from numerical experiments that the horizontal subbands of FKP image at each decomposition level provide significant amount of textural features to discriminate between interclass and the intraclass images. Therefore, horizontal subband information of FKP are chosen till the optimal decomposition level while six Gabor representation of FKP is selected for analysis. The experimental results demonstrate the effectiveness of the proposed scheme and analyze the performance of the proposed techniques for the FKP recognition systems. The Haar wavelet decomposition achieves the highest recognition rate of 92.89 % at third level while the combined scheme (Gabor + WT) gives a best recognition rate 96.30 % for the FKP identity verification.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Cir. syst. video technol.* 14, 4–20 (2004).
2. Woodard, D.L., Flynn, P.J.: Finger surface as a biometric identifier. *Comput. Vis. Image Underst.* 100 (3), 357–384 (2005).
3. Kumar, A., Ravikanth, C.: Personal authentication using finger knuckle surface. *IEEE Trans Inf. Forens. Secur.* 4, 98–109 (2009).
4. Zhang, L., Zhang, L., Zhang, D., Zhu, H.: Online finger knuckle print verification for personal authentication. *Pattern. Recognit.* 43, 2560–2571 (2010).
5. Verma, G., Sinha, A.: Finger knuckle print verification using minimum average correlation energy filter. *IJECS.* 5, 233–246 (2014).
6. Mallat, S. : A theory of multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Machine Intell.* 11, 674–693 (1989).

7. Vetterli, M., Kovačević, J.: *Wavelets and subband coding*. Prentice Hall Englewood Cliffs New Jersey (1995).
8. Kim, J., Cho, S., Choi, J., Marks, R. J.: Iris Recognition using wavelet features. *J. VLSI Signal Process.* 38, 147–156 (2004).
9. Zhang, B. L., Zhang, H. H., Ge, S. S.: Face recognition by applying wavelet subband representation and kernel associative memory. *IEEE Trans. Neural Networks.* 15, 166–177 (2004).
10. Kong, W. K., Zhang, D., Li, W.: Palmprint feature extraction using 2-D Gabor filters. *Pattern Recognit.* 36, 2339–2347 (2003).
11. Wang, R., Wang, G., Chen, Z., Zang, Z., Wang, Y.: palm vein identification system based on Gabor wavelet features. *Neural Comput & Applic.* 24, 161–168 (2013).
12. http://www.comp.polyu.edu.hk/_biometrics/FKP.htm.

Design of Advanced Correlation Filters for Finger Knuckle Print Authentication Systems

Gaurav Verma and Aloka Sinha

Abstract Biometric recognition systems use automated processes in which the stored templates are used to match with the live biometrics traits. Correlation filter is a promising technique for such type of matching in which the correlation output is obtained at the location of the target in the correlation plane. Among all kinds of available biometrics, finger knuckle print (FKP) has usable significant feature, which can be utilized for person identification and verification. Motivated by the utility of the FKP, this paper presents a new scheme for person authentication using FKP based on advanced correlation filters (ACF). The numerical experiments have been carried out on the Poly U FKP database to evaluate the performance of the designed filters. The obtained results demonstrate the effectiveness and show better discrimination ability between the genuine and the imposter population for the proposed scheme.

Keywords Advanced correlation filter • Biometric recognition • Finger knuckle print

1 Introduction

Biometrics is the field of research in which physiological and behavioral characteristics of humans are utilized for recognition-based applications [1]. Such types of systems are superior and improve the security performance in terms of user authentication over knowledge-based scheme. Researchers are continuously trying to explore new approaches for biometrics-based recognition systems. Recently, a new biometric trait known as finger knuckle print (FKP) has been investigated by researchers for ongoing research [2]. When a finger is bent, a FKP is formed which

G. Verma (✉) · A. Sinha

Department of Physics, Indian Institute of Technology Delhi, New Delhi, India
e-mail: gaurav.sgs85@gmail.com

A. Sinha
e-mail: aloka@physics.iitd.ac.in

possesses a unique and a highly distinctive skin pattern. This is rich in texture features [2]. Among various kinds of biometric modalities, FKP offers several advantages for identification- and verification-based applications. It is easily accessible and requires contact less image acquisition systems, and does not involve any kind of emotions or expressions. Thus, it provides stable features for recording [3]. Due to the distinguishing features of FKP, it has attracted more attention in the field of biometric research. Numerous research works based on FKP have been reported with different techniques for recognition [2–4]. Kumar et al. have used features of finger knuckle surface for personal identification [3]. Zhang et al. have developed an experimental device to record FKP images and also proposed a FKP verification system using band limited phase only correlation filters for personal authentication [4]. The correlation filters have important features such as shift invariance and distortion tolerance. Correlation filter is a useful technique to evaluate the performance of biometrics-based recognition system [5].

In this paper, ACF such as synthetic discriminant function (SDF), minimum average correlation energy (MACE) filter, and optimal tradeoff synthetic discriminant function (OTSDF) have been designed for subject verification using FKP. A study has been carried out to show the effectiveness of the SDF filter to discriminate between the genuine and the imposter population. The performance of MACE filter is investigated in the presence of noise. The OTSDF filter is analyzed in terms of noise robustness and peak sharpness which shows a performance in between the minimum variance synthetic discriminant function (MVSDF) and the MACE filter.

2 Advanced Correlation Filters

Correlation filters are widely designed and implemented in the area of biometrics authentication- and verification-based applications [5, 7]. Different approaches to design composite correlation filters for pattern recognition have been discussed [6]. Multi-criteria optimization techniques for the design of filter have been investigated [7]. The basic correlation filter, the matched spatial filter (MSF), is applied for detection of reference subject in the presence of additive white noise [8]. Its performance degrades due to distortion and geometrical changes (scale or rotation) in the reference subject. To overcome the limitations of the MSF, more advanced correlation filter design schemes are reported in Refs. [9–12]. In these schemes, a single filter or template from a set of training images is constructed in frequency domain and stored for verification. The basic process of recognition based on correlation filter has been shown in Fig. 1. The motivation of the proposed work is to explore potential utility of ACF by combining the advantages of the correlation process for FKP verification. This leads to the idea of the proposed FKP verification system.

The first advanced correlation filter, the synthetic discriminant function filter (h_{SDF}) is the weighted sum of MSFs [9]. In this filter, a linear combination of training set of images is used to design h_{SDF} and can be expressed as:

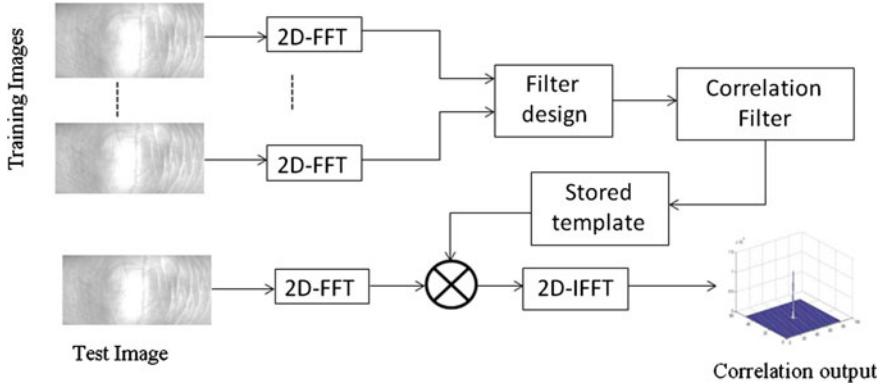


Fig. 1 Block diagram of correlation process

$$h_{SDF} = X(X^+ X)^{-1} u \quad (1)$$

where the superscript ‘+’ denotes the conjugate transpose and u is the column vector of prespecified peak constraints. The set of training images is represented by a column vector (X) of size $d \times N$, where d is the total number of pixels and N is the number of training images.

The MACE filter (h_{MACE}) is one of the most attractive filters because of their high discrimination performance [11]. This filter is used to minimize the average correlation energy (ACE) of the correlation outputs due to the training images [10]. The MACE filter has been proposed for finger knuckle print verification [11]. Due to the minimization of the ACE, sharp correlation peaks are obtained at the location of a trained object. The MACE filter (h_{MACE}) is defined as:

$$h_{MACE} = D^{-1} X (X^+ D^{-1} X)^{-1} u \quad (2)$$

where D is the diagonal matrix of the average of the power spectrum of training image set. As reported in literature, the SDF filter produces correlation output peak with large sidelobes, whereas the MACE filter provides a sharp correlation peak but it is highly sensitive to noise. Thus, the performance will be optimized by a trade-off in between output noise variance (ONV) and ACE. This type of advanced correlation filter is known as optimal tradeoff SDF filter.

In the OTSDF (h_{OTSDF}), the weighted sum of the output noise variance (ONV) and ACE are minimized [12]. This is expressed as:

$$h_{OTSDF} = (\alpha D + \beta C)^{-1} X \left[X + (\alpha D + \beta C)^{-1} X \right]^{-1} u \quad (3)$$

where C is the input noise covariance matrix. The relation between α and β is given as:

$$\alpha^2 + \beta^2 = 1 \quad (4)$$

$$\beta = \sqrt{1 - \alpha^2} \quad (5)$$

where ‘ α ’ and ‘ β ’ are nonnegative constants. The performances of OTSDF correlation filter have been studied in terms of noise robustness and peak sharpness.

3 Numerical Experiment

The advanced correlation filters have been synthesized for the verification of FKPs. The Poly U FKP [13] image database of the Hong Kong Polytechnic University is used for this. This database includes 7920 images of 165 subjects in which there are 125 males and 25 females. Each person has 12 images of each finger.

In the present study, twelve hundred FKP images are used in this analysis. In order to design the correlation filter, the region of interest of a FKP is extracted by using the proposed method [4]. Three FKP training images are selected to synthesize the filter using Eqs. (1), (2), and (3). The designed filters are stored as a template for matching at verification stage. The correlation processes are performed by cross-correlating the test images with the stored template. The correlation output can be expressed as:

$$c(p, q) = x(p, q) \odot h(p, q) \quad (6)$$

$$c(p, q) = F^{-1} \{ (X(m, n) \cdot H^*(m, n)) \} \quad (7)$$

where $c(p, q)$ is the correlation output by taking the inverse Fourier transform (F^{-1}) of $X(m, n)$ and $H(m, n)$. $X(m, n)$ and $H(m, n)$ are the test image and the designed filter in frequency domain while $x(p, q)$ and $h(p, q)$ are the test image and the designed filter in space domain, respectively. The obtained correlation output is used to determine the genuine or the imposter person. The block diagram of the correlation process for FKP verification is shown in Fig. (1).

4 Results and Discussion

In order to design the SDF correlation filter, three FKP training images for a particular subject are used to synthesize the filter using Eq. (1) as shown in Fig. 2a. The representation of the h_{SDF} filter in space domain is shown in Fig. 2b.

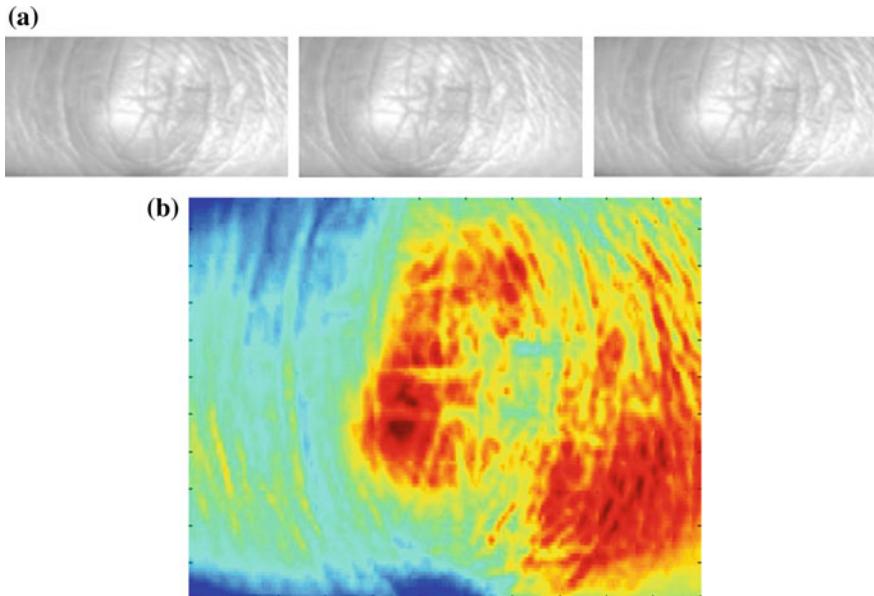


Fig. 2 **a** FKP training images. **b** Representation of SDF filter in space domain

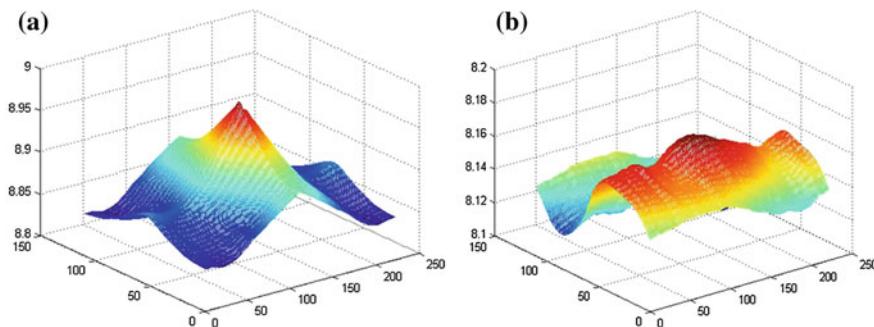


Fig. 3 The SDF correlation output. **a** Genuine person. **b** Imposter person

The obtained correlation peak of the h_{SDF} filter for the genuine and imposter person are shown in Fig. 3a, b respectively. A large sidelobe and higher peak values of correlation peak are observed for the genuine person while it has a flattened side lobe and lower values of peak in the case of imposter person.

For each of the filter, the average of maximum peak values of correlation output for genuine as well as imposter are obtained, and plotted against the number of subjects along the Y-axis and X-axis, respectively. The points in both the graphs correspond to a particular subject for which the filter is synthesized. As shown in Fig. 4, the average maximum peak values of correlation output for each genuine

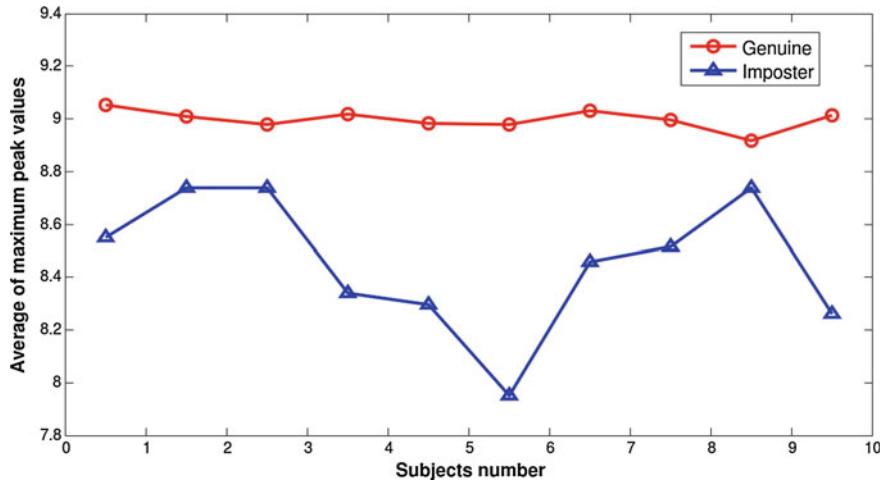


Fig. 4 The graph plotted between number of subjects and average of maximum peak values for SDF filter

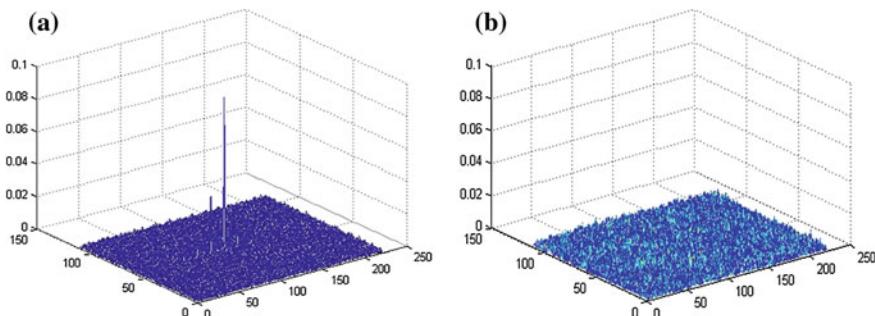


Fig. 5 The MACE filter correlation output. **a** Genuine person. **b** Imposter person

population are higher than that of corresponding imposter populations for a filter designed for a particular person. The upper curve line shows the intraclass average of maximum peak values and the lower curve line shows interclass average of maximum peak values. This shows a clear separation between the averages of maximum peak values for different populations. Thus, FKP-based verification system using SDF filter can discriminate the genuine population from the imposter population.

The limitation of large sidelobes of the SDF filter is overcome using the MACE filter. As explained earlier, the MACE filter has been designed for FKP verification [11]. The cross-correlation outputs reported in [11] for the genuine and imposter person are shown in Fig. 5a, b respectively. As shown in Fig. 5, a sharp correlation peak is obtained for the genuine person while multiple peaks of low peak values are obtained for the imposter person.

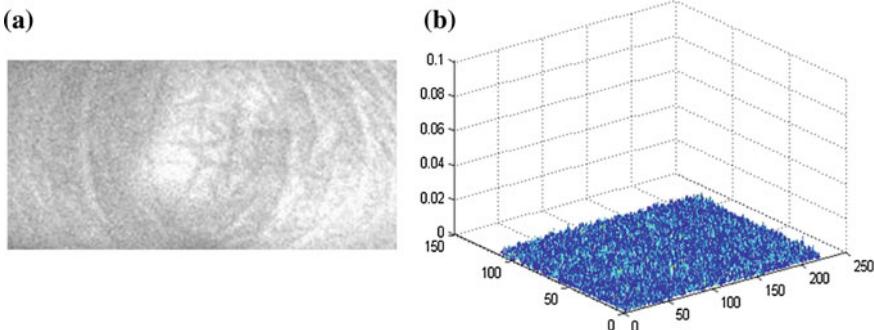


Fig. 6 The MACE filter correlation output in presence of noise. **a** Noisy image. **b** Correlation output

The MACE filter behavior in the presence of noise in FKP images is also studied. In order to calculate the correlation of noisy genuine image, noise of zero mean and 0.05 variance are added to the FKP images. The noisy image and its correlation output with MACE filter are shown in Fig. 6a, b, respectively.

It can be seen from Fig. 6b that the correlation output for noisy genuine is similar to the correlation output for the imposter person as shown in Fig. 5b. This shows that the MACE filter is unable to verify the genuine person in the presence of noise. Thus, the MACE filter is highly sensitive to noise for FKP recognition.

The OTSDF correlation output peaks are obtained by the correlation process as explained earlier in section III. The cross-correlation outputs of OTSDF filter for different values of α using Eq. (3) are used to evaluate the performance of the designed OTSDF filter. When α is set to zero, then it produces a sharp correlation output with a large sidelobe and when α is set to one, then it produces sharp correlation peak as shown in Figs. 7 and 8, respectively, for genuine and imposter person.

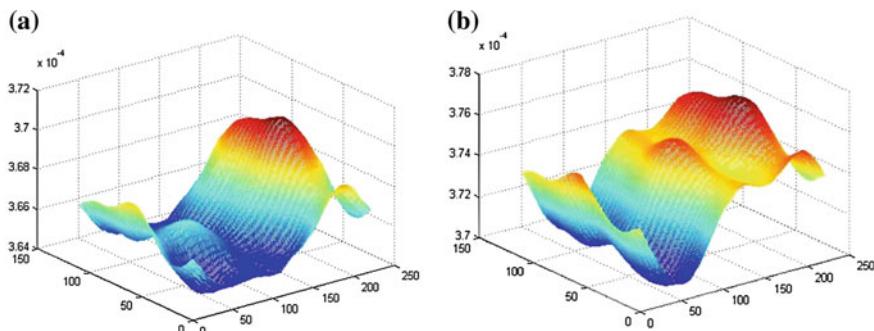


Fig. 7 The OTSDF correlation output for $\alpha = 0$. **a** Genuine person. **b** Imposter person

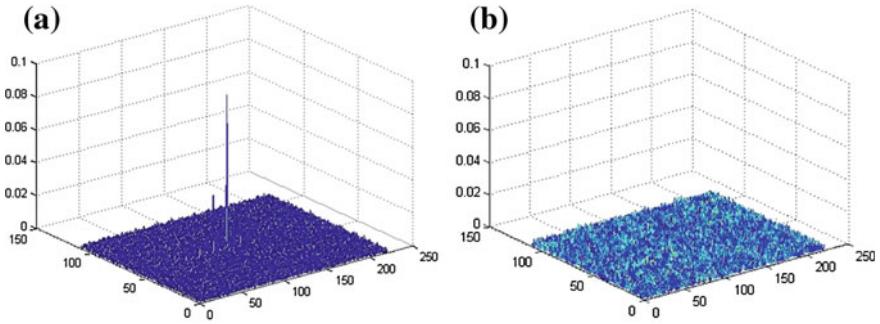


Fig. 8 The OTSDF correlation output for $\alpha = 1$. **a** Genuine person. **b** Imposter person

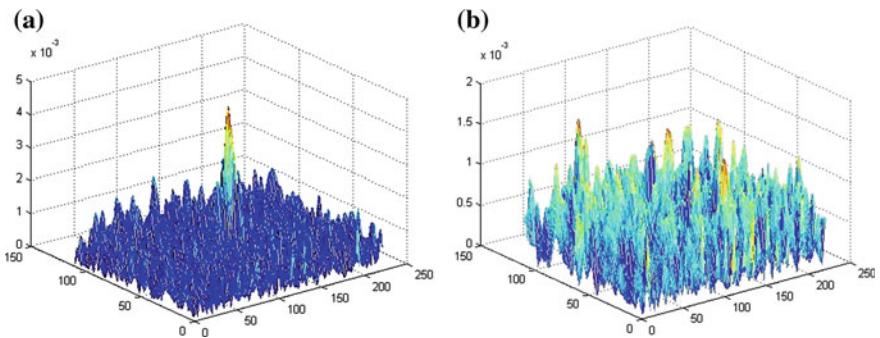


Fig. 9 The OTSDF correlation output for $\alpha = 0.33$. **a** Genuine person. **b** Imposter person

The behavior of the designed OTSDF filter for a value of α in between zero to one is studied in terms of noise tolerance with reduction in sidelobes. The cross-correlation output peaks for the genuine and imposter person for $\alpha = 0.33$ and $\alpha = 0.67$ are shown in Figs. 9 and 10, respectively.

The performances of the OTSDF filter have also been measured by calculating the peak to side lobe ratio (PSR) for the FKP verification. The PSR can be defined as:

$$\text{PSR} = \text{Peak} - \text{Mean} / (\text{Standard deviation}) \quad (8)$$

The details of PSR calculation using Eq. (8) are explained earlier [10]. The obtained PSR values are given in Table 1 for the genuine and imposter persons.

It can be seen from Table 1 that as ' α ' increases the value of PSR also increases for genuine person. It means a reduction in the sidelobe and an increase in peak sharpness. The large sidelobes are found for the zero values of ' α ' while the maximum peak sharpness is obtained when ' α ' reaches a value of one. The change in the values of ' α ' produces a noticeable change in the strength of sidelobes in the correlation plane and can be utilized to discriminate between genuine and imposter person.

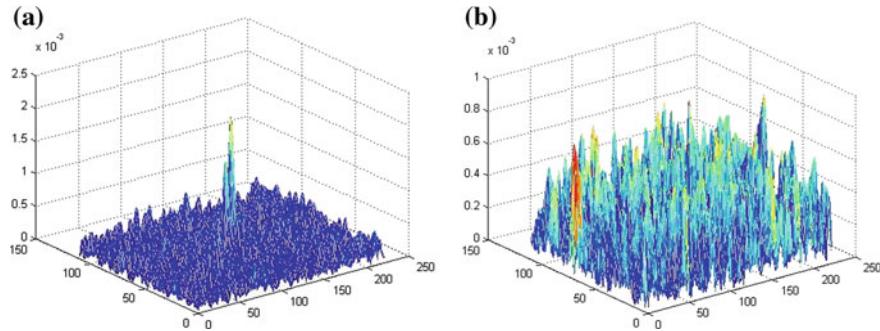


Fig. 10 The OTSDF correlation output for $\alpha = 0.67$. **a** Genuine person. **b** Imposter person

Table 1 PSR values for different values of ' α '

S.No.	Subject	$\alpha = 0.0$	$\alpha = 0.33$	$\alpha = 0.67$	$\alpha = 1.0$
1	Genuine	4.8664	6.8862	9.3738	74.7570
	Imposter	3.2595	3.1187	2.2463	4.6794
2	Genuine	4.5895	5.0838	7.3367	60.5058
	Imposter	3.4274	3.3411	2.9624	3.9371
3	Genuine	4.3847	5.5309	8.6515	54.8092
	Imposter	3.2111	3.0236	3.0705	4.6528
4	Genuine	4.7895	5.2449	8.3819	55.1727
	Imposter	3.2769	3.8452	2.3745	4.8891

The performance of proposed scheme is compared with the other existing methods [2–4] for FKP verification. These methods are very much accurate even when they involve higher number of computation and complexity compared to the proposed scheme and employ a number of transforms to extract FKP feature. In the proposed scheme, ACF are explored for FKP recognition which has simple implementation process with low computational complexities and need less feature storage for FKP verification.

5 Conclusion

The performance of advanced correlation filters has been presented for FKP verification. The SDF filter is able to discriminate between the genuine and imposter person for FKP recognition. The filter performances are also studied in the presence of noise in the FKP images. It is found that the MACE filter is very sensitive to noise and is unable to discriminate between subjects. In the optimal tradeoff SDF filter, which uses a tradeoff between the ONV and ACE, the performance of the designed filters is in between MVSDF and MACE filters which is seen by a reasonable noise tolerance and sharpness in peak.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Cir. syst. video technol.* 14, 4–20 (2004).
2. Woodard, D.L., Flynn, P. J.: Finger surface as a biometric identifier. *Comput. Vis. Image Underst.* 100 (3), 357–384 (2005).
3. Kumar, A., Ravikanth, C.: Personal authentication using finger knuckle surface. *IEEE Trans Inf. Forens. Secur.* 4, 98–109 (2009).
4. Zhang, L., Zhang, L., Zhang, D., Zhu, H.: Online finger knuckle print verification for personal authentication. *Pattern. Recognit.* 43, 2560–2571 (2010).
5. Vijaya Kumar B.V.K., Savvides M., Venkataramani K., Xie C.: Spatial Frequency Domain Image Processing for Biometric Recognition. *IEEE Int'l Conf. on Image Processing.* 53–56 (2002).
6. Vijaya Kumar B.V.K.: Tutorial survey of composite filter designs for optical correlators. *Appl. Opt.* 31, 4773–4801 (1992).
7. Refregier P.: Filter Design for optical pattern recognition: Multi-criteria optimization approach. *Opt. Lett.* 15, 854–856 (1990).
8. Vanderlugt A.: Signal detection by complex spatial filtering. *IEEE Trans. Inf. Theory.* 10, 139–145 (1964).
9. Hester C.F. Casasent D.: Multivariant technique for multiclass pattern recognition. *Appl. Opt.* 19, 1758–1761 (1980).
10. Mahalanobis A., Vijaya Kumar B.V.K., Casasent D.: Minimum average correlation energy filters. *Appl. Opt.* 26, 3633–3630 (1987).
11. Verma G., Sinha A.: Finger knuckle print verification using minimum average correlation energy filter. *IJECS.* 5, 233–246 (2014).
12. Vijaya Kumar B.V.K., Carlson D.W., Mahalanobis A.: Optimal tradeoff synthetic discriminant function filters for arbitrary devices. *Opt. Lett.* 19, 1556–1558 (1994).
13. http://www.comp.polyu.edu.hk/_biometrics/FKP.htm.

A Nonlinear Modified CONVEF-AD Based Approach for Low-Dose Sinogram Restoration

Shailendra Tiwari, Rajeev Srivastava and K.V. Arya

Abstract Preprocessing the noisy sinogram before reconstruction is an effective and efficient way to solve the low-dose X-ray computed tomography (CT) problem. The objective of this paper is to develop a low-dose CT image reconstruction method based on statistical sonogram smoothing approach. The proposed method is casted into a variational framework and the solution of the method is based on minimization of energy functional. The solution of the method consists of two terms, viz., data fidelity term and a regularization term. The data fidelity term is obtained by minimizing the negative log likelihood of the signal-dependent Gaussian probability distribution which depicts the noise distribution in low-dose X-ray CT. The second term, i.e., regularization term is a nonlinear CONvolutional Virtual Electric Field Anisotropic Diffusion (CONVEF-AD) based filter which is an extension of Perona–Malik (P–M) anisotropic diffusion filter. The main task of regularization function is to address the issue of ill-posedness of the solution of the first term. The proposed method is capable of dealing with both signal-dependent and signal-independent Gaussian noise, i.e., mixed noise. For experimentation purpose, two different sinograms generated from test phantom images are used. The performance of the proposed method is compared with that of existing methods. The obtained results show that the proposed method outperforms many recent approaches and is capable of removing the mixed noise in low-dose X-ray CT.

Keywords X-ray computed tomography • Statistical sinogram smoothing • Image reconstruction algorithm • Noise reduction • Anisotropic diffusion

S. Tiwari (✉) • R. Srivastava

Department of Computer Science & Engineering,
Indian Institute of Technology (BHU), Varanasi 221005, India
e-mail: stiwari.rs.cse@iitbhu.ac.in

R. Srivastava
e-mail: rajeev.cse@iitbhu.ac.in

K.V. Arya
Multimedia & Information Security Lab, ABV-Indian Institute of Information
Technology & Management, Gwalior 474015, India
e-mail: kvarya@iitm.ac.in

1 Introduction

Nowadays X-ray computed tomography (CT) is one of the most widely used medical imaging modalities for various clinical applications such as diagnosis and image-guided interventions. Recent discoveries in medical imaging have certainly improved the physician's perspectives for better understanding of diseases and treatment of the patients. Unfortunately, on the other hand it may also show some potential harmful effects of X-ray radiation including lifetime risk of genetic, cancerous, and other diseases due to overuse of imaging diagnostic [1]. Therefore, the effort should be made to reduce the radiation in medical applications. To realize this objective, many algorithms have been developed during the last two decades for CT radiation dose reduction [1–3]. From these algorithms, preprocessing the noisy and under sampled sinogram by statistical iterative methods has shown great potential to reduce the radiation dose while maintaining the image quality in X-ray CT as compared with the FBP reconstruction algorithm. For the sinogram smoothing purpose, it is very important to consider a widely studied effective regularization terms or priors [4–6]. The main aim of using the regularization priors during reconstruction is to lower the noise effects and preserve the edges consequently maximizing the important diagnostic information. However, one of the drawbacks associated with using regularization term is over penalization of the image or its gradient which may lead to loss of basic fine structure and detailed information. To address these drawbacks, still several priors which include smoothing, edge-preserving regularization terms, and iterative algorithms with varying degrees of success have been already studied and used to obtain high-quality CT reconstruction images from low-dose projection data. In view of the above-discussed problems, an anisotropic diffusion nonlinear partial differential equation (PDE) based diffusion process was developed by Perona and Malik (P-M) [8] for image smoothing while preserving the small structures. In this process, the diffusion strength is controlled by a gradient magnitude parameter to preserve edges. The over smoothing problem associated with P-M model was addressed by Ghita [9], by proposing the concept of Gradient Vector Flow (GVF) field for the implementation of the anisotropic diffusion models. But it has the disadvantage to produce undesirable staircase effect around smooth edges. Due to this effect, this method could not remove the isolated noise accurately and falsely recognize the boundaries of different blocks that actually belong to the same smooth area as edges. However, due to the presence of mixed noise in the sinogram data, i.e., signal-dependent and signal-independent Gaussian noise; these methods cannot be applied directly. Even, GVF fields have no ability to find edge when images are corrupted by extraneous or Gaussian noise, and thus the denoising effect of mixed noisy images remains unsatisfactory.

In this work, the low-dose CT image reconstruction has been improved by modifying the CONVEF-based P-M approach which is used as a prior in the denoising process. The proposed modified CONVEF-AD serves as a regularization or smoothing term for low-dose sinogram restoration to deal with the problem of

mixed (Poisson + Gaussian) noise as well as ill-posedness issue. The proposed reconstruction model provides many desirable properties like better noise removal, less computational time, preserving the edges, and other structure. It can also overcome the staircase effect effectively. The proposed model performs well in low-dose X-ray CT image reconstruction. Also, the proposed results are compared with some recent state-of-the-art methods [5, 6].

Rest of the paper is divided into the following sections. Section 2 presents the methods and materials of the work. The proposed variational framework for sinogram restoration using CONVEF-AD regularized statistical image reconstruction method is presented in Sect. 3. Section 4 presents the discussion on simulation experiments results achieved by the proposed CONVEF-AD reconstruction method in both the simulated data and CT data. The conclusions are given in Sect. 5.

2 Methods and Models

Noise modeling of the projection (or sinogram) data, specifically for low-dose CT, is essential for the statistics-based sinogram restoration algorithms. Low-dose (or mAs) CT sinogram data usually contained serious staircase artifacts. It also follows a Gaussian distribution with a nonlinear signal independent as well as Poisson distribution with signal-dependent noise model between the sample mean and variance. To address this issue, several approaches are available in the literature for statistical sinogram smoothing methods like penalized weighted least square (PWLS) [7], Poisson likelihood (PL) [3] methods, etc. [5]. However, these existing methods often suffer from noticeable resolution loss especially in the case of constant noise variance over all sinogram data [1–3].

The formation of X-ray CT images can be modeled approximately by a discrete linear system as follows:

$$g = Af, \quad (1)$$

where $f = (f_1, f_2, \dots, f_N)^T$, is the original image vector to be reconstructed, N is the number of voxels, the superscript T is the transpose operator, $g = (g_1, g_2, \dots, g_M)^T$, is the measured projection vector data, M is the total number of sampling points in the projection data, $A = \{a_{ij}\}$, $i = 1, 2, \dots, M$ and $j = 1, 2, \dots, N$ is the system matrix of size $I \times J$ and relates with f and g . The value of a_{ij} is commonly calculated by using the intersectional length of projection ray i with pixel j .

It has been shown in [3, 10] that there are two principal sources of noise occurred during CT data acquisition, X-ray quanta noise (signal-dependent compound Poisson distribution), and system electronic background noise (signal-independent Gaussian or normal distribution with zero mean). However, it is numerically difficult to directly implement these models for data noise simulation. Several reports have discussed the approximation of this model by the Poisson

model [4]. Practically, the measured transmission data N_i can be assumed to statistically follow the Poisson distribution upon a Gaussian distributed electronic background noise [2]:

$$N_i \approx \text{Poisson}(\tilde{N}_i) + \text{Gaussian}(m_e, \sigma_e^2) \quad (2)$$

where m_e and σ_e^2 are the mean and variance of the Gaussian distribution from the electronic background noise, \tilde{N}_i is the mean of Poisson distribution. In reality, the mean m_e of the electronic noise is generally calibrated to zero (i.e., ‘dark current correction’) and the associative variance slightly changes due to different settings of tube current, voltage, and durations in a same CT scanner [5]. Hence, in a single scan, the variance of electronic background noise can be considered as uniform distribution.

Based on the noise model in Eq. (2), the calibrated and log-transformed projection data may contain some isolated noise points which follow approximately a Gaussian distribution. After removing this noisy points from the projection data, there is a relationship between the data sample mean and variance, which is described as [6]:

$$\sigma_i^2 = \frac{1}{N_{0i}} \exp(\tilde{g}_i) \left(1 + \frac{1}{N_{0i}} \exp(\tilde{g}_i) (\sigma_e^2 - 1.25) \right) \quad (3)$$

where N_{0i} is the incident X-ray intensity, \tilde{g}_i is the mean of the log-transformed ideal sinogram datum g_i on path i , and σ_e^2 means background Gaussian noise variance. During implementation, the sampling mean \tilde{y}_i could be calculated approximately by taking the average of neighboring 3×3 window. The parameters N_{0i} and σ_e^2 could be measured as a part of the standard process in CT routine calibration systems [3].

3 The Proposed Model

The energy minimization function is used to obtain sinogram smoothing in variational framework can be defined as:

$$f^* = \arg \min_{f \geq 0} E(f) \quad (4)$$

where the energy functional is described as follows:

$$E(f) = E_1(f) + \lambda E_2(f) \quad (5)$$

In the Eq. (5), $E_1(f)$ is used to represents data fidelity (or equivalently, data fitting, data mismatch, and data energy) term which ensures the modeling of statistics of projection data f and the measurement g . $E_2(f)$ is a regularization (or

equivalently, prior, penalty, and smoothness energy) term. The parameter λ is called balancing parameter that controls the degree of prior's influence between the estimated and the measured projection data.

By taking the (negative) log likelihood of the estimated data and ignoring the constant and irrelevant term, the joint probability distribution function (pdf) can be expressed as [10]:

$$P(g|f) = \prod_{i=1}^M P(g_i|f_i) = \prod_{i=1}^M \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(g_i - f_i)^2}{2\sigma_i^2}\right) \quad (6)$$

where $g = (g_1, g_2, \dots, g_M)^T$, is the measured projection vector data. Then, ignoring the constant, the negative logarithm function can be written as:

$$E_1(f) = \ln P(g|f) = \sum_{i=1}^M \left\{ \frac{(g_i - f_i)^2}{2\sigma_i^2} \right\} \quad (7)$$

According to the MRF framework, the smoothness energy is calculated by the negative log likelihood of the priori [4]. The focus in this paper is to use nonlinear CONVEF-based P-M anisotropic diffusion regularization term for two reasons. (i) The integral nature of the edge-preserving MRF prior does not suit well for high continuity of the sparse measured data. (ii) The PDE-based AD model is capable of giving the optimal solution in less computational time. Therefore, the smoothness function can be defined as follows [10]:

$$E_2(f) = \arg \min \left(\lambda \int_{\Omega} \phi(\|\nabla f\|^2) d\Omega \right) \quad (8)$$

where $\phi(\|\nabla f\|^2)$ represents gradient norm of the image of corresponding energy function. The solution of the proposed sinogram smoothing of Eq. (4) can be described by substituting Eqs. (7) and (8) to Eq. (4):

$$E(f) = \arg \min_{f \geq 0} E(f) = \sum_{i=1}^M \left\{ \frac{(g_i - f_i)^2}{2\sigma_i^2} \right\} + \int_{\Omega} (\lambda \|\nabla f\|^2) d\Omega \quad (9)$$

The functional $E(f)$ is defined on the set of $f \in BV(\Omega)$ such that $\log f \in L^1(\Omega)$ and f must be positive. After minimizing Eq. (9) using combined approach of Euler–Lagrange minimization technique with gradient descent, the solution to above Eq. (9) can be written as:

$$f_t = \frac{\partial f}{\partial t} = \sum_{i=1}^M \left\{ \frac{(g_i - f_i)^2}{\sigma_i^2} \right\} + \lambda \operatorname{div}(c(|\nabla f|) \nabla f), \quad \text{with} \quad \frac{\partial f}{\partial \vec{n}} = 0 \text{ on } \partial\Omega \quad (10)$$

where div and ∇ are known as divergence and gradient operator, respectively, $f_{(t=0)} = f_0$ is the initial condition for noisy image. The value of diffusion coefficient $c(\cdot)$ represents nonnegative monotonically decreasing function of the image or its gradient. Generally $c(\cdot)$ takes as:

$$c(|\nabla f|) = 1/l + \left((\nabla f)^2/k \right)^2 \quad (11)$$

where k is a conductance parameter also known as gradient magnitude threshold parameter that controls the rate of diffusion and the level of contrast at the boundaries. Since the scale-space generated by these function is different. The diffusion coefficient c favors high-contrast edges over low-contrast ones. The small value of k is well capable of preserving small edges and other fine details but the smoothing effect on results is poor and weak. Conversely, on the large value of k , denoising effects on results is better but it will lose small edges and fine details. Since, it is reported in literature [3, 12] that second term in Eq. (9) is nonlinear AD prior to detect edges in multi-scale space, where diffusion process is controlled by the gradient coefficient of image intensity in order to preserve edges. However, P-M diffusion model can remove isolated noise and preserve the edges to some extent but it cannot preserve the edge details effectively and accurately which leads to blocking staircase effect. Moreover, it also gives poor results for very noisy images and the values of high gradient areas of the images get smoothen out that affects the fine edges. Therefore more diffusion cannot be allowed to remove the noise along edges.

To address these limitations of AD method, CONVEF-based P-M anisotropic diffusion process is introduced as a second term in Eq. (8). The second term is an expanded form of the P-M model [11] which is defined as:

$$\text{div}(c'(|\nabla f|)\nabla|\nabla f| \cdot \nabla f + c(f)\nabla^2 f) \quad (12)$$

where the first term in Eq. (12) is an inverse diffusion term used to enhance or sharpen the boundaries while the second term is a Laplacian term used for smoothing the regions that are relatively flat, ∇f is used to displace the inverse diffusion term to improve the denoising effect, as the GVF field basically implements a weighted gradient diffusion process. Thus, we used here a new version of CONVEF-AD based P-M equation defined as:

$$f_t = (1 - IN(f_0))(med(f_0) - f_{t-1}) + IN(f_0) \cdot (-E_{\text{CONVEF}} \cdot \nabla f + c\nabla^2 f) \quad (13)$$

where f_0 is the input noisy image, f_{t-1} is the updated f at iteration $t - 1$. $IN(f_0)$ is the Poisson or Gaussian estimator defined in [4, 11], med is the median filter, and E_{CONVEF} denotes the Convolutional Virtual Electric Field defined as follows [12]:

$$E_{\text{CONVEF}} = \left(-\frac{a}{r_h^{n+2}} \otimes q, -\frac{b}{r_h^{n+2}} \otimes q \right) \quad (14)$$

where $r_h = \sqrt{a^2 + b^2 + h}$, is a desired kernel that modifies the distance metrics, the variable h plays a significant role in multi-scale space filtering, (a, b) is used as the virtual electric field in the image pixel, and q denotes the magnitude of the image edge maps. The signal-to-noise (SNR) ratio, the smoothness, and quality of an image are directly dependent on the parameter h . Finally the proposed model is introduced by incorporating the new version of modified CONVEF-AD based P-M Eq. (13) within the variational framework of sinogram smoothing discussed in Eq. (10). By applying Euler–Lagrange combined with gradient descent minimization technique, the Eq. (10) reads as:

$$f_t = \frac{(g_i - f_i)}{\sigma_i^2} + \lambda((1 - IN(f_0))(med(f_0) - f_{t-1}) + IN(f_0) \cdot (-E \cdot \nabla f + c \nabla^2 f)), \quad (15)$$

The proposed model in Eq. (15) is capable of dealing with the case of mixed noise sinogram data by introducing the CONVEF-AD regularization term. The value of k is set to σ_e which represents the minimum absolute deviation (MAD) of the image gradient. To make the adaptive or automatized nature of value of k by using the following formula:

$$k = \sigma_e = 1.4826 \times \text{median}_f[\|\nabla f - \text{median}_f(\|\nabla f\|)\|], \quad (16)$$

For digital implementations, the Eq. (15) can be discretized using finite differences schemes [4]. After discretization of the proposed CONVEF-AD model it reads as:

$$f_{i,j}^{n+1} = f_{i,j}^n + \Delta t \left[\frac{(g_{i,j}^n - f_{i,j}^n)}{\left(\sigma_{i,j}^2 \right)^n} + \lambda \left((1 - IN(f_0)) \left(med(f_0) - f_{t-1}^n \right) + IN(f_0) \cdot (-E^n \cdot \nabla f^n + c \nabla^2 f^n) \right) \right], \quad (17)$$

where the index n represents the number of iteration. The Von Neumann analysis [4] shows that for the stability of discretized versions of Eq. (17), the following condition should be satisfied as $\Delta t / (\Delta f)^2 < 1/4$. If the grid size is set to $\Delta f = 1$ then $\Delta t \leq 1/4$.

4 Results and Discussion

In this work, two test cases are used as shown in Figs. 1a and 2a, both are computer-generated mathematical simulated phantom one is modified Shepp–Logan head phantom and another is CT phantom used to validate the result performance of the proposed CONVEF-AD based sinogram smoothing method for low-dose CT reconstruction. For simulation purposes, MATLAB v2013b has been

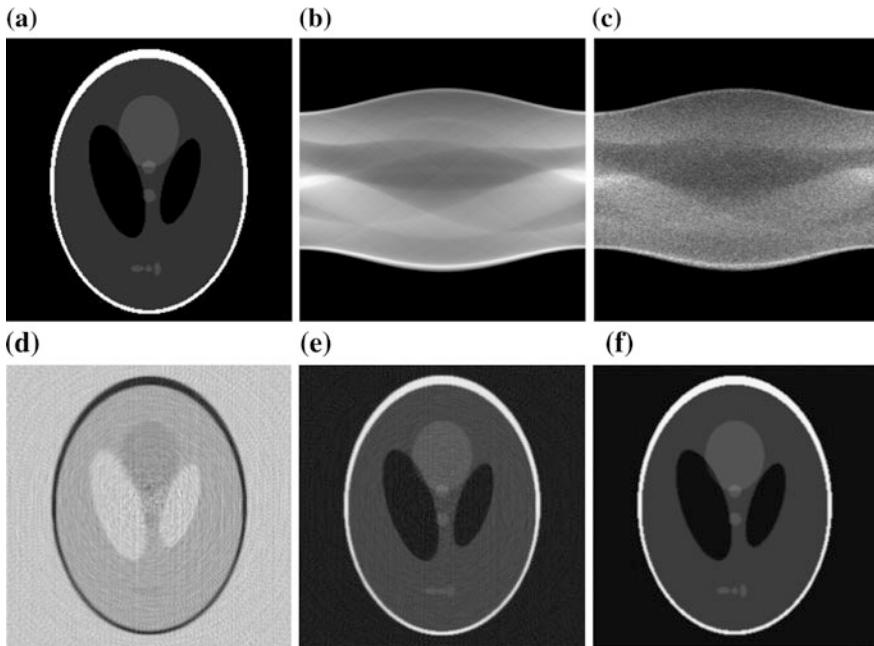


Fig. 1 The reconstructed results of modified Shepp–Logan phantom with similar standard methods from the noisy sinogram data. **a** Original Shepp–Logan phantom, **b** noise free sinogram, **c** noisy sinogram, **d** reconstructed image by TV+FBP, **e** reconstructed result by AD+FBP, and **f** reconstructed result by CONVEF_AD+FBP

used on a PC with Intel (R) Core (TM) i5 CPU 650 @ 3.20 GHz, 4.00 GB RAM, and 64-bit operating system. The brief description of the various parameters used for generation and reconstruction of the two test cases are as follows: Both test cases are of size 128×128 pixels and 120 projection angles were used. To simulate the noisy low-dose sinogram data, Eq. (2) was used, which is in mixed noise nature, i.e., both signal-dependent and signal-independent Gaussian distributed, and all are assumed to be 128 radial bins and 128 angular views evenly spaced over 1800. A mixed noise of magnitude 10 % is added to sinogram data. In simulation purposes fan-beam imaging geometry were used. By applying radon transform noise free sinogram were generated which is shown in Figs. 1b and 2b. After that isolated data from the noisy sinogram which are shown in Figs. 1c and 2c, were extracted by applying a 3×3 median filter and choose the output as an initial guess estimator. The proposed model consists of many advantages over Gradient Vector Flow (GVF) and Inverse-GVF method, like improvised numerical stability and efficient estimation of high order derivatives. Therefore, the proposed model introduced in Eq. (13) may effectively remove mixed noise properties, i.e., signal dependent and signal independent present in low-dose sinogram data and acceptable computational cost. Also compute the mean and variance by using Eq. (3) and then calculate the gradient coefficient by applying Eq. (11). In this

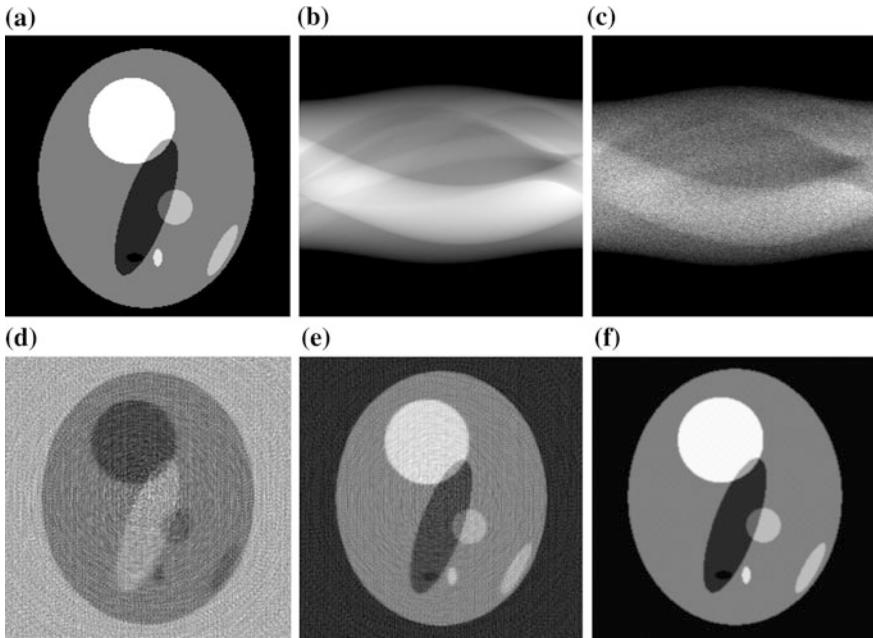


Fig. 2 The reconstructed results of CT phantom with different reconstruction methods from the noisy sinogram data. **a** original Shepp–Logan phantom, **b** noise free sinogram, **c** noisy sinogram, **d** reconstructed image by TV+FBP, **e** reconstructed result by AD+FBP, and **f** reconstructed result by CONVEF_{AD}+FBP

experiment, the parameters of the proposed models are as follows: The INGVF field parameters are set to 0.2. In our study, the whole algorithm is run for 100 iterations because visual result hardly changes for further iterations and within each iterations CONVEF_{AD} is run for three iterations. The CONVEF parameters: $n = 5$, $h = 15$ and the size of the kernel are defined to be one fourth of the test size images. In Eq. (14), E is calculated before the evolution of image, only ∇f has to be computed directly from the estimated image while the original Eq. (12) needs to compute the second-order derivative. The value of balancing parameter λ was set to 1 for each test case and the value of diffusion coefficient (Kappa) used by proposed CONVEF-AD based prior was calculated by using Eq. (16) for different test cases, within each iteration during sinogram smoothing. Update the value of the estimated image pixel-by-pixel using Eq. (17) until it reaches to a relative convergence. The reconstructed images produced by different algorithms have shown in Figs. 1d–f and 3d–f, respectively. From the figures, it can be observed that proposed method performs better in terms of noise removal as well as preservation of weak edges and structural fine detailed information. Also notice that, proposed method reduces the streaking artifacts and the results are close to the original phantom test cases. The graphs are plotted for different quantitative measures like SNR, RMSE, CP, and MSSIM for different algorithms as shown in Fig. 3a–d for both test cases. From

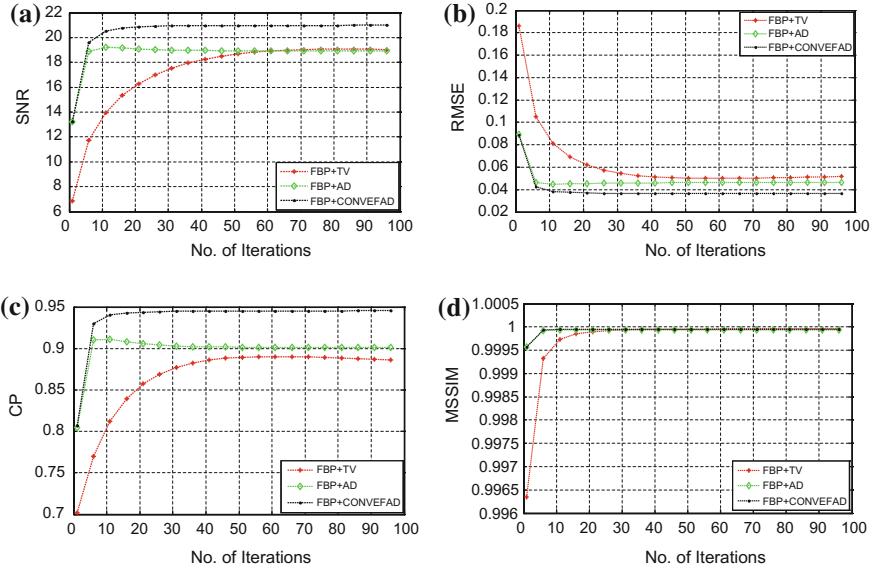


Fig. 3 The plots of **a** SNR, **b** RMSE, **c** CP, and **d** MSSIM of proposed and other models for Test case 1

Fig. 3a, it is observed that the SNR values associated with the proposed method are always higher than that produced by other algorithms such as Total Variation (TV) [5] and Anisotropic Diffusion (AD) [6] priors with traditional filtered back-projection (FBP), which indicates that the CONVEF-AD with FBP framework significantly improves the quality of reconstruction in terms of different quantitative measures like SNR, RMSE, CP, and MSSIM values. Figure 3b, shows that the RMSE values of proposed method are higher in comparison to other methods which indicate CONVEF-AD with FBP performs better than other methods. Figure 3c shows that the CP values of CONVEF-AD with FBP method are higher and close to unity in comparison to other methods which indicate that the CONVEF-AD with FBP framework is also well capable of preserving the fine edges and detailed structures during the reconstruction process. Figure 3d, shows that the MSSIM values of proposed method is higher which indicate better reconstruction; it also preserves the luminance, contrast, and other details of the image during the reconstruction processes. Table 1 shows that the quantification values of SNRs, RMSEs, CPs, and MSSIMs for both the test cases, respectively. The comparison tables indicate that proposed method produces images with prefect quality than other reconstruction methods in consideration.

Table 1 Different performance measures for the reconstructed images

Performance measures	For Fig. 1			For Fig. 2		
	FBP+TV [5]	FBP+AD [6]	FBP +CONVEF_AD (Proposed method)	FBP+TV [5]	FBP+AD [6]	FBP +CONVEF_AD (Proposed method)
SNR	19.0684	19.2164	20.9984	10.0959	13.9103	15.8961
RMSE	0.0455	0.0447	0.0364	0.0774	0.0499	0.0016
PSNR	75.0046	75.1526	76.9346	70.3851	74.1995	76.1854
CP	0.9077	0.9124	0.9456	0.8038	0.9341	0.9662
MSSIM	0.9889	0.9897	0.9933	0.9999	0.9999	1.0000

5 Conclusions

This paper proposes an efficient method for low-dose X-ray CT reconstruction using statistical sinogram restoration techniques. The proposed method has been modeled into a variational framework. The solution of the method, based on minimization of an energy functional, consists of two terms, viz., data fidelity term and a regularization function. The data fidelity term was obtained by minimizing the negative log likelihood of the noise distribution modeled as Gaussian probability distribution as well as Poisson distribution which depicts the noise distribution in low-dose X-ray CT. The regularization term is nonlinear CONVEF-AD based filter which is a version of Perona–Malik (P–M) anisotropic diffusion filter. The proposed method was capable of dealing with mixed noise. The comparative study and performance evaluation of the proposed method exhibit better mixed noise removal capability than other methods in low-dose X-ray CT.

References

1. Lifeng Yu, Xin Liu, Shuai Leng, James M Kofler, Juan C Ramirez-Giraldo, Mingliang Qu, Jodie Christner, Joel G Fletcher, and Cynthia H McCollough: Radiation Dose Reduction in Computed Tomography: Techniques and Future Perspective. *Imaging in medicine, Future Medicine*, 1(1), 65–84, (2009).
2. Hao Zhang, Jianhua Ma, Jing Wang, Yan Liu, Hao Han, Hongbing Lu, William Moore, Zhengrong Liang: Statistical Image Reconstruction for Low-Dose CT Using Nonlocal Means-Based Regularization. *Computerized medical imaging and graphics, Computerized Medical Imaging Society*, 8(6), 423–435, (2014).
3. Yang Gao, Zhaoying Bian, Jing Huang, Yunwan Zhang, Shanzhou Niu, Qianjin Feng, Wufan Chen, Zhengrong Liang, and Jianhua Ma: Low-Dose X-Ray Computed Tomography Image Reconstruction with a Combined Low-mAs and Sparse-View Protocol. *Optics Express*, 2(12), 15190–15210, (2014).
4. Rajeev Srivastava, Subodh Srivastava, Restoration of Poisson noise corrupted digital images with nonlinear PDE based filters along with the choice of regularization parameter estimation. *Pattern Recognition Letters*, 34(10), 1175–1185, (2013).

5. Xueying Cui, Zhiguo Gui, Quan Zhang, Yi Liu, Ruifen Ma: The statistical sinogram smoothing via adaptive-weighted total variation regularization for low-dose X-ray CT. *Optik - International Journal for Light and Electron Optics*, 125(18), 5352–5356, (2014).
6. Mendrik, A.M., Vonken, E.-J., Rutten A., Viergever M.A., van Ginneken B.,: Noise Reduction in Computed Tomography Scans Using 3-D Anisotropic Hybrid Diffusion With Continuous Switch. *Medical Imaging, IEEE Transactions*, 28(10), 1585–1594, (2009).
7. Jing Wang; Hongbing Lu, Junhai Wen, Zhengrong Liang: Multiscale Penalized Weighted Least-Squares Sinogram Restoration for Low-Dose X-Ray Computed Tomography. *Biomedical Engineering, IEEE Transactions*, 55(3),1022–1031, (2008).
8. Perona, P.; Malik, J.: Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 12(7), 629–639, (1990).
9. Ovidiu Ghita, Paul F. Whelan: A new GVF-based image enhancement formulation for use in the presence of mixed noise. *Pattern Recognition*, 43(8), 2646–2658, (2010).
10. Niu, Shanzhou et al.,: Sparse-View X-Ray CT Reconstruction via Total Generalized Variation Regularization. *Physics in medicine and biology*, 59(12), 2997–3017, (2014).
11. Hongchuan Yu,: Image Anisotropic Diffusion based on Gradient Vector Flow Fields. *Computer Vision - ECCV 2004, Lecture notes in Computer Science 3023*, 288–301, (2004).
12. Y, Zhu C, Zhang J, Jian Y,: Convolutional Virtual Electric Field for Image Segmentation Using Active Contours. *PLoS ONE*, 9(10), 1–10, (2014).

System Design for Tackling Blind Curves

Sowndarya Lakshmi Sadasivam and J. Amudha

Abstract Driving through blind curves, especially in mountainous regions or through roads that have blocked visibility due to the presence of natural vegetation or buildings or other structures is a challenge because of limited visibility. This paper aims to address this problem by the use of surveillance mechanism to capture the images from the blind side of the road through stationary cameras installed on the road and provide necessary information to drivers approaching the blind curve on the other side of the road, thus cautioning about possible collision. This paper proposes a method to tackle blind curves and has been studied for various cases.

Keywords Blind curves • Collision avoidance • Surveillance • Optical flow

1 Introduction

Mountainous regions have a number of hairpin turns or blind curves which are dangerous for drivers as they have limited or no information about the vehicles approaching them on the other side of the curve. Usually these regions may not be profound to have roads with good infrastructure due to heavy rains, landslides etc. In such cases, installation of a stand alone system at the blind curves that monitors and updates the drivers about the condition of the road ahead or about vehicles approaching from the other side of the blind curve would be useful to prevent accidents. This system can identify vehicles approaching the curve on the other side, or to identify the presence of some stationary object such as a broken down vehicle on the road, or to identify stray cattle or pedestrians crossing the road at the curve. This information would prove to be useful by keeping the driver informed

S.L. Sadasivam (✉) · J. Amudha
Department of Computer Science Engineering,
Amrita Vishwa Vidyapeetham, Bangalore, India
e-mail: sowndarya.lak07@gmail.com

J. Amudha
e-mail: j_amudha@blr.amrita.edu

well in advance and thus avoiding accidents. The proposed design in this paper makes use of surveillance mechanism to capture data from the curve and is analyzed to get the information about approaching vehicles and provides this data to the drivers on the other side of the curve. The proposed system makes use of optical flow method to identify vehicles and pedestrians around the blind curve and this data is processed to determine the necessary action. The paper is structured as follows—the second section provides an overview of related work in the area of improving visibility at blind curves, the third section describes the proposed system for tackling this problem and fourth section explains the implementation in brief, the fifth section provides experimental results, the sixth section provides the conclusion and the final section explains the scope for improvement and further analysis of this proposed model.

2 Related Work

There have been some researches done in the past few years about traffic prediction at the blind curves.

Inter-vehicular communication (IVC) is the approach that has been researched widely for communicating between vehicles to get the information about the vehicles in the vicinity which can be used in this scenario of blind curves. One study [1] about the connectivity of vehicles at different road settings shows that curved roads with vegetation and steep crest has connectivity issues and cannot assess the traffic situation correctly. Another study [2] explains the implementation of multihop connectivity. But this method will be useful only if there is sufficient number of vehicles for multi-hop communication and they are all interconnected. Yet this approach will not receive information about other objects like stray cattle in the path of the vehicles.

The flow of traffic at mountainous regions and cause for accidents has been investigated [3]. This study proposes a method of using pressure sensors near the curve to detect the type of the vehicle and to determine the speed of the vehicle. These sensors are also used to provide information about the vehicles approaching the curve.

Therefore it would be useful if there is a standalone system that can detect and provide information about possible hazards near the blind curves. For this purpose, this paper suggests the use of surveillance mechanism to detect and analyze the area around the blind curve.

3 Surveillance Based Tracking System

3.1 Assumptions

Proposed system has been built with an assumption that the vehicles around the blind curve move with a speed that is the general set speed limit at hilly regions in most of the countries. Also, the traffic density is assumed to be low to medium.

3.2 System Set up

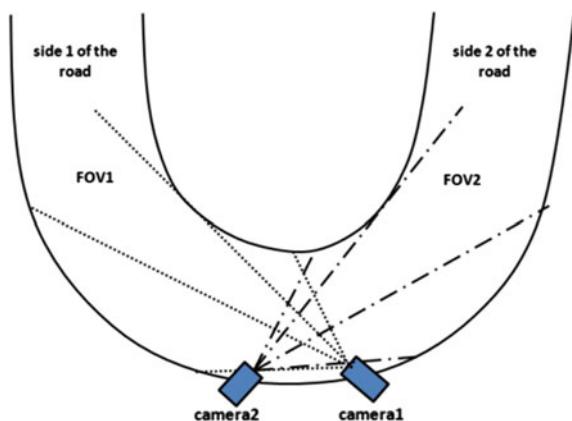
The proposed system makes use of two cameras at the blind curve with overlapping field of view to capture the complete information about the road. An example curve used for study is shown in Fig. 1. Camera 1 captures information on one side of the blind curve and Camera 2 captures information on the other side of the blind curve.

3.3 System Architecture

The system has been designed using a input unit which has an image capture unit with 2 cameras to capture information from both the sides of the blind curve. This information is fed to the image processing unit (DSP) which processes the images to extract useful information about the vehicles and pedestrians on the road. Finally alert messages are sent to the output unit through which the communication is established with the vehicles to transmit the necessary information.

The flow chart explaining the detection of vehicles and sending an alert is explained in Fig. 2. The image processing unit captures the images from both the

Fig. 1 Sample curve under study



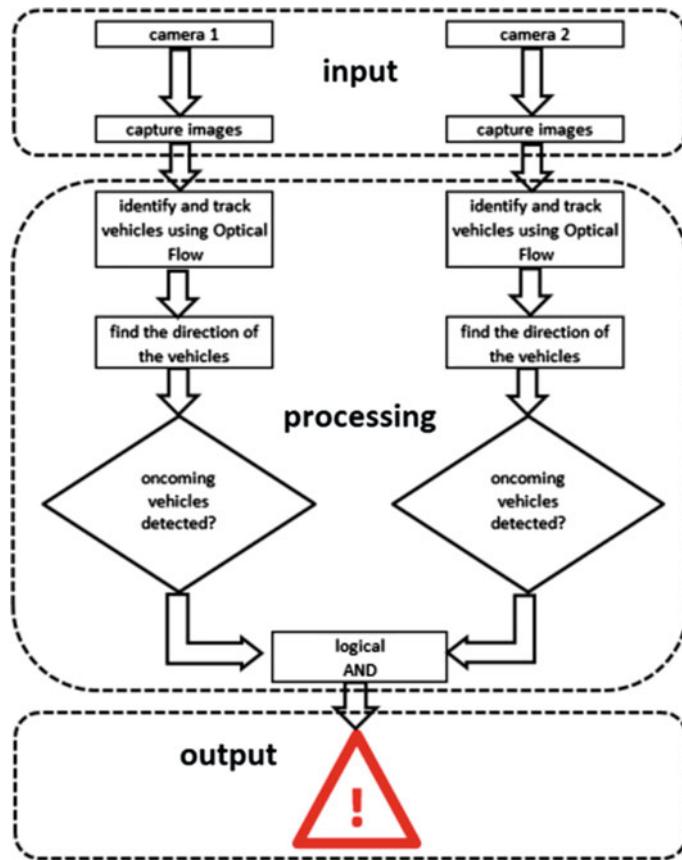


Fig. 2 Flow chart for detecting vehicles at the blind curve

sides of the curve simultaneously and processes them to identify the objects in motion. Based on the size of the object, vehicles are identified and the direction of motion of the vehicles is determined. If the vehicles are found to be approaching the curve on both the sides, an alert signal is raised to keep the drivers informed about oncoming vehicles.

4 Implementation

The system has been modeled using MATLAB Simulink. The optimum frame rate is studied in order to find the maximum distance vehicle would have moved between two frames. The frames are processed using Optical Flow for identifying vehicles (or objects) in motion.

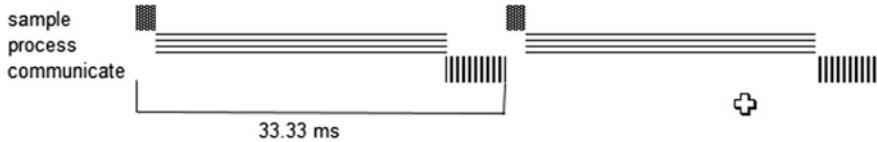


Fig. 3 Time chart of the operation

4.1 Frame Capture Rate

It is important to know the distance that the vehicle covers between two frames. Let S be the speed of the vehicle in km/hr. If the image capture rate is x fps, then the distance moved by the vehicle between two frames would be calculated as below-

$$\text{Speed} = S \text{ km/h} = S * (5/18) \text{ m/s.} \quad (1)$$

$$\text{Distance covered } (D) = S * (5/18) * (1/x) \text{ meters.} \quad (2)$$

Time available for capturing the images, processing them and responding would be $1/x$ seconds (Fig. 3).

4.2 Detection and Tracking

This study makes use of Optical Flow method in order to determine the flow of traffic at the blind curve region. The Horn-Schunk method has been applied to detect vehicles and track its motion [4] (Fig. 4). Two frames taken at t and $t + \delta t$ are compared to check the relative motion of the pixels. Using this information, the mean velocity is calculated which acts as the threshold for segmenting the image to extract the objects in motion from the images. Noise filtering is done to suppress random impulses. Morphological operation is then done to extract significant features from the image that are useful in representing the object.

All of these operations are performed on the Region of Interest selected to extract information from the portion of the road that is required for tracking.

Next, blob analysis is done on this pre-processed image to extract the blobs based on sizes and areas covered to classify the blobs into vehicles of different categories.

4.3 Direction Analysis

The optical flow vectors from the Optical Flow Model can be used for determining the direction of motion of the vehicle [5]. The angle of motion vector of the blob can

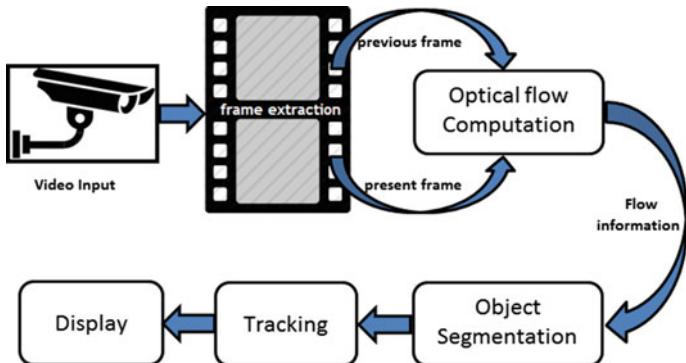
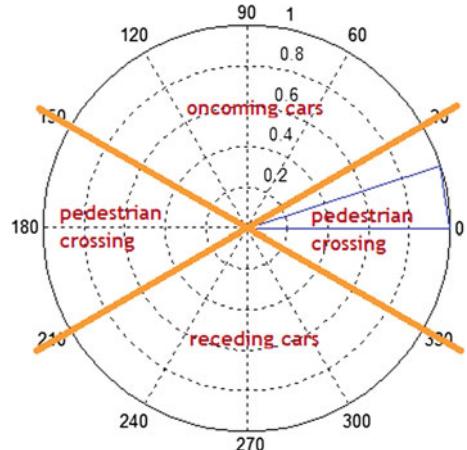


Fig. 4 Optical flow based moving object detection and tracking method

Fig. 5 Angle split for determining flow of vehicles



be used to determine if the vehicle is approaching the curve or not. If (u,v) represents the velocity components in the horizontal and vertical directions of a pixel at (x,y) , then the direction can be determined as

$$\delta(x,y) = \tan^{-1}(u/v) \quad (3)$$

The angle is determined by averaging the flow vectors of the blob.

The angle split shown in Fig. 5 has been considered for the model developed for the study of the system. If the derived angle is between 30° and 150° , the motion of the vehicle is towards the camera. If the derived angle is between 210° and 330° , the motion of the vehicle is away from the camera. The angles between 30° – 330° and 150° – 210° are used for pedestrians crossing the road.

5 Experimental Results

The proposed system has been designed and verified using MATLAB Simulink. Dataset from i-LIDS Image Library has been used for verifying the results. The vehicles could be identified using the optical flow method and the information could be calculated and displayed in real time based on the direction of flow of the vehicles (Fig. 6). Analysis has also been done for pedestrians crossing the road.

The various cases that have been studied with this model have been tabulated in Tables 1 and 2. Alert message is displayed based on the direction of motion of the vehicles and the pedestrians crossing the road.

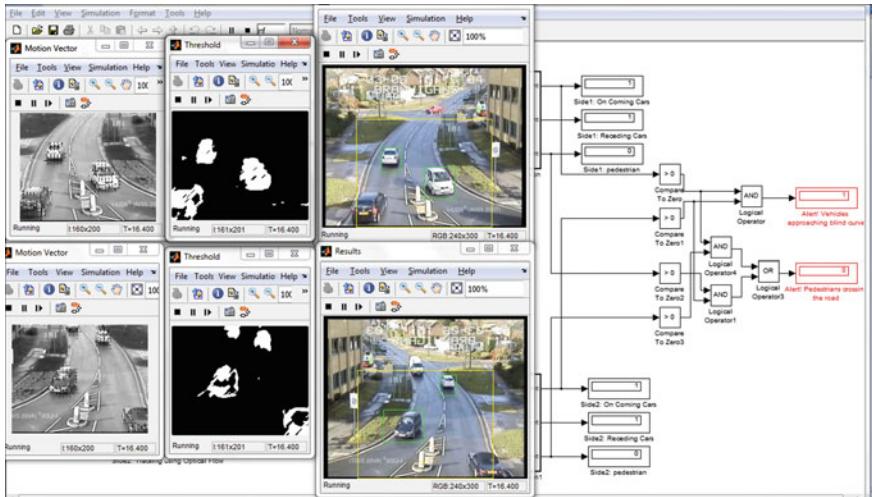


Fig. 6 Alert display when vehicles approach the curve from both sides

Table 1 Cases studied based on vehicle direction

Cases	Side 1			Side 2			Alert! Vehicles approaching on the other side of the road
	Vehicles	Direction	Pedestrian crossing	Vehicles	Direction	Pedestrian crossing	
Case1	✓	→	x	✓	←	x	✓
Case2	✓	→	x	✓	→	x	x
Case3	✓	←	x	✓	←	x	x
Case4	✓	←	x	✓	→	x	x

Table 2 Cases studied based on pedestrian movement

Cases	Side 1			Side 2			Alert! Vehicles approaching on the other side of the road
	Vehicles	Direction	Pedestrian crossing	Vehicles	Direction	Pedestrian crossing	
Case1 Present/not present	-	✓	✓	✓	←	x	✓
Case2 Present/not present	-	✓	✓	✓	→	x	x
Case3 ✓	→	x		Present/not present	-	✓	✓
Case4 ✓	↓	x		Present/not present	-	✓	x
Case5 x	-	✓	x		-	x	x
Case6 x	-	x	x		-	✓	x

6 Conclusion

The proposed system thus helps in tackling the problems faced by drivers while driving around the blind curves by providing information about vehicles approaching the curve. This prototype system has been built using Optical Flow method, but this method has some drawbacks. This method cannot be used for detecting still objects. A method like back ground subtraction would help to detect still objects. Simulation results show that system performance has been found to be good for implementation in real time.

7 Future Scope

Further study on this proposed prototype model can be continued by considering different lighting conditions such as night, foggy, rainy (use of IR cameras), different traffic densities and vehicle speeds, improving the processing time and efficiency of the system, method to detect stationary objects. Lastly, incorporation of this functionality as an augmented reality application can be researched.

References

1. Annette Böhm, Kristoffer Lidström, Magnus Jonsson and Tony Larsson, “Evaluating CALM M5-based vehicle-to-vehicle communication in various road settings through field trials”, 4th IEEE LCN Workshop On User MObility and VEhicular Networks(ON-MOVE), Denver, CO, USA, Oct. 2010.
2. Wen-Long Jin and Wilfred W.Recker, “An Analytical Model of Multihop Connectivity of Inter-Vehicular Communication System”, IEEE Transactions on wireless communications, Vol 9, No 1, January 2010.
3. Pranay D. Saraf and Nekita A. Chavan, “Pre-crash Sensing and Warning on Curves: A Review”, International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol 2, Issue 1, Jan 2013.
4. Ms. Shamshad Shirgeri, Ms. Pallavi Umesh Naik, Dr. G.R. Udupi and Prof. G.A. Bidkar, “Design and development of Optical flow based Moving Object Detection and Tracking (OMODT) System”, International Journal of Computational Engineering Research, Vol, 03, Issue, 4, April 2013.
5. Parul Matur and Amudha. J, “Anomalous Event Detection In Surveillance Video Using Visual Attention”, IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.

A Novel Visual Word Assignment Model for Content-Based Image Retrieval

Anindita Mukherjee, Soman Chakraborty, Jaya Sil
and Ananda S. Chowdhury

Abstract Visual bag of words model have been applied in the recent past for the purpose of content-based image retrieval. In this paper, we propose a novel assignment model of visual words for representing an image patch. In particular, a vector is used to represent an image patch with its elements denoting the affinities of the patch to belong to a set of closest/most influential visual words. We also introduce a dissimilarity measure, consisting of two terms, for comparing a pair of image patches. The first term captures the difference in affinities of the patches to belong to the common set of influential visual words. The second term checks the number of visual words which influences only one of the two patches and penalizes the measure accordingly. Experimental results on the publicly available COIL-100 image database clearly demonstrates the superior performance of the proposed content-based image retrieval (CBIR) method over some similar existing approaches.

Keywords Visual words · Assignment model · Dissimilarity measure · Image retrieval

1 Introduction

Content-based image retrieval (CBIR) has become a popular area of research for both computer vision and multimedia communities. It aims at organizing digital picture archives by analyzing their visual content [1]. The success of a CBIR system critically depends on the extraction of proper visual features and design of an appropriate

A. Mukherjee
Dream Institute of Technology, Kolkata, India

S. Chakraborty · A.S. Chowdhury (✉)
Jadavpur University, Kolkata, India
e-mail: aschowdhury@etce.jdvu.ac.in

J. Sil
IEST Sibpur, Howrah, India

dissimilarity measure [2]. Color and texture remain the most popular image features used for retrieval [3, 4]. Similarly, there exist several measures of image dissimilarity which can be employed for image and video retrieval [5, 6].

In recent past, bag of visual words (BoVW) has evolved as a popular framework for CBIR [7]. In this model, an image is represented as a collection of elementary local features. As a first step, distinctive local features from scale invariant keypoints, known as SIFT, are obtained after breaking an image into a set of patches [8]. SIFT features are found to be very useful for tasks like retrieval and classification because of their high discriminative power. In the second step, the SIFT descriptors are quantized by k-means algorithm to build a bag of visual words [9]. Then an image can potentially be represented as a vector of words like a single document. Consequently, document/text retrieval measures are extended for image retrieval and classification [10, 11]. However, BoVW-based representations suffer from some important limitations. The most serious drawback is that it fails to support spatial properties of local descriptors of images [12]. This is because a BoVW representation of an image is really a histogram of closest/most similar visual words for its constituent patches. Note that in the BoVW model, only the most similar visual word is allotted to an image patch using the nearest neighbor search. This type of assignment adversely affects the fidelity of the visual content. In recent past, some works have considered the impact of more than one visual word on an image patch. For example, Bouachir et al. have used a fuzzy c-means based approach to improve the retrieval performance of classical BoVW-based method [13]. However, fuzzy c-means based approach may fail to properly handle uncertainty for large and complex data like that of an image database. Optimal parameter setting for fuzzy c-means also becomes a challenge in such cases which in turn would adversely affect the retrieval performance [14].

In the present work, we propose a novel visual word assignment model for an image which can preserve its visual content. We select M most similar visual words instead of a single most similar visual word for representing each of the constituent patches in an image. So, each patch in our model is represented by a M -element vector, where each element of the vector, denotes the affinity of a patch towards a similar visual word. The affinity term is probabilistically computed first and then modulated using a orientation similarity factor. Our second contribution is the introduction of a measure of dissimilarity. The measure consists of two terms. Note that set of most similar M visual words will in general be different for a given pair of patches. The first term of our measure is the Frobenius norm of the affinity difference vector for the set of most similar visual words, which are common, to both the patches. The second term penalizes the dissimilarity measure by capturing the number of most similar visual words which influence only one of the two patches but not both.

The rest of the paper is organized in the following manner: in Sect. 2, we discuss in detail the proposed method including our measure of dissimilarity. In Sect. 3, we present and analyze the experimental results. The paper is concluded in Sect. 4 with an outline for directions of future research.

2 Proposed Method

In this section, we describe in details the proposed method. The section contains two parts. In the first part, we briefly discuss the construction of visual words. In the second subsection, we describe our stochastic assignment model.

2.1 Construction of Visual Words

An image consists of several salient patches. Stable key points are extracted from the image as centers of these salient patches using a difference-of-Gaussian function. Then, with a 4×4 surrounding, histogram gradient orientations are computed with 8 bins. In this way, every patch is represented by a $4 \times 4 \times 8$, a 128-dimensional SIFT feature vector [8]. For CBIR, images in the database are broken into a set of patches. So, we have a collection of image patches where each patch is a point in 128-dimensional (SIFT) feature space.

The local SIFT descriptors need to be quantized to build the visual vocabulary. We have applied k-means algorithm to cluster the image patches in the 128-dimensional feature space. Each cluster is treated as a unique visual word and the collection of such visual words form the visual vocabulary [9]. The center of the visual clusters are used to find the affinity of the patches in a given image.

2.2 Assignment Model for Visual Words

In the BoVW model, an image patch is assigned only the most similar visual word using the nearest neighbor strategy. Distances between a patch and the centers of the above clusters (visual words) are computed for that purpose. A patch is assigned the visual word for which the above distance is minimum [15]. In this work, we consider M most similar visual words to represent an image patch by choosing the M closest cluster centers. Let $d(\mathbf{p}, \mathbf{v})$ be the Euclidean distance between a patch (SIFT) vector $\mathbf{p} = [p_1, p_2, \dots, p_{128}]$ and a visual word \mathbf{v} with the cluster center vector $\mathbf{v} = [v_1, v_2, \dots, v_{128}]$. Then, $d(\mathbf{p}, \mathbf{v})$ is given by

$$d(\mathbf{p}, \mathbf{v}) = \sum_{j=1}^{128} \|(p_j - v_j)\| \quad (1)$$

We now compute the affinity $a(\mathbf{p}, \mathbf{v})$ of the patch p for the visual word V in a probabilistic manner, as shown below

$$a(\mathbf{p}, \mathbf{v}) = \frac{\frac{1}{d(\mathbf{p}, \mathbf{v})}}{\sum_{i=1}^M \frac{1}{d(\mathbf{p}, i)}} \quad (2)$$

So, a patch is initially represented in our model by a affinity vector $A(\mathbf{p}) = [a_{pv}]$, $v = 1, \dots, M$ where a_{pv} is actually $a(\mathbf{p}, \mathbf{v})$ as given in Eq. 2. Since the SIFT features take into consideration the orientation aspect, we further modify Eq. 2 by incorporating a measure of orientation. Let $o(\mathbf{p}, \mathbf{v})$ be the cosine similarity distance between the patch vector \mathbf{p} and the cluster center vector \mathbf{v} . Then $o(\mathbf{p}, \mathbf{v})$ can be written as

$$o(\mathbf{p}, \mathbf{v}) = \arccos \left(\frac{\mathbf{p} \cdot \mathbf{v}}{|\mathbf{p}| |\mathbf{v}|} \right) \quad (3)$$

Since, $a(\mathbf{p}, \mathbf{v}) \in [0, 1]$ and $o(\mathbf{p}, \mathbf{v}) \in [0, \pi]$, we normalize $o(\mathbf{p}, \mathbf{v})$ (denoted as $\bar{o}(\mathbf{p}, \mathbf{v})$) to change its range to $[0, 1]$. Note a higher value of $A(\mathbf{p}, \mathbf{v})$ indicates higher similarity between the patch p and the cluster center v . In contrast, a lower value of $o(\mathbf{p}, \mathbf{v})$ means the patch p and the cluster center v have similar orientation. So, we use an orientation adaptive affinity $oa(\mathbf{p}, \mathbf{v})$ as a measure of similarity between the patch p and the cluster center v which is given by

$$oa(\mathbf{p}, \mathbf{v}) = \frac{a(\mathbf{p}, \mathbf{v})}{\bar{o}(\mathbf{p}, \mathbf{v})} \quad (4)$$

Finally, a patch is represented by an orientation adaptive affinity vector $OA(\mathbf{p}) = [oa_{pv}]$, $v = 1, \dots, M$ where oa_{pv} is actually $oa(\mathbf{p}, \mathbf{v})$ as given in Eq. 4.

2.3 A Dissimilarity Measure-Based on Assignment Model

In this section, we introduce a new measure of dissimilarity between a pair of image patches/images. The measure takes into consideration the fact that the set of most similar visual words for one patch can be different from the set of most similar visual words for the second patch. Let us consider two patches p and q and denote the corresponding sets of M most similar visual words by M_p and M_q , respectively. So, $M_p = \{p_1, p_2, \dots, p_M\}$ and $M_q = \{q_1, q_2, \dots, q_M\}$. Let us further assume that C out of M are common most similar visual words, i.e., $C = |M_p \cap M_q|$. We, accordingly construct C -dimensional affinity vectors $OA(p_C)$ and $OA(q_C)$ for the patches p and q from the previously obtained M -dimensional vectors $OA(\mathbf{p})$ and $OA(\mathbf{q})$. So, there exist $(M - C)$ visual words which influence either the patch p or the patch q but not both. The proposed dissimilarity between the patches p and q is given by $ds(\mathbf{p}, \mathbf{q})$

$$ds(\mathbf{p}, \mathbf{q}) = \sqrt{(OA(p_C) - OA(q_C)) \cdot (OA(p_C) - OA(q_C))^T + (e^{(M-C)} - 1)} \quad (5)$$

The dissimilarity measure, as given in Eq. 5 consists of two terms. The first term is the Frobenius norm of the C -dimensional affinity difference vector of the two patches. The second term behaves like a penalty which accounts for the number of visual words influencing only one of the patches. Exponentiation is used to make the impact of the penalty high which improves the discriminative power of the visual words. Note that when $C = M$, the penalty term attains a zero value as expected. Finally, the dissimilarity measure between two images, say, one in the database and the other query, can be obtained by adding the pairwise dissimilarities of the individual patches. So, we can write

$$ds(I_{DB}, I_Q) = \sum_{i=1}^n ds(p_{I_{DB}}^i, p_{I_Q}^i) \quad (6)$$

In the above equation, $ds(I_{DB}, I_Q)$ denotes the dissimilarity between the database image I_{DB} and the query image I_Q . Further, $p_{I_{DB}}^i$ and $p_{I_Q}^i$ respectively denotes the i th patch in the database image I_{DB} and the i th patch in the query image I_Q .

3 Experimental Results

We use the publicly available COIL-100 image database [16] for experimentation. The database contains a total of 7200 images with 72 different images of 100 different objects having a viewpoint separation of 5° . The SIFT features are first extracted from the database images. Then k-means clustering algorithm is employed to obtain the visual words with k equal to 100. Since we propose an assignment model, where multiple visual words are used for describing a patch, we decide to compare our work with another assignment model. So, we chose [13] describing a fuzzy assignment model, and two other methods, namely, *term frequency-inverse document frequency* (*tfx*) based approach [15] and *term frequency* (*txx*) [17] based approach used within [13] for comparisons. The fuzzy weighting method also uses the value k as 100. Precision versus recall curves are drawn to compare the performances of the different methods [1].

We first show in Figs. 1 and 2 the precision versus recall curves for object 3 and object 10 of the COIL-100 database for three different values of M , namely, $M = 10$, $M = 30$ and $M = 70$. Both the figures clearly show that best results are obtained for $M = 10$. These two objects are chosen because precision-recall curves for them are available in [13]. In Figs. 3 and 4, we next present the precision-recall curves for all four competing methods. Figure 4 shows we clearly outperform [13, 15, 17] for object 10. Figure 3 demonstrates that our results are superior to [15, 17] and is marginally better than [13] for object 3.

We also present the recognition rate as an average precision for ten different objects, namely, Object 1 to Object 10 of the COIL-100 database [13]. In Table 1, we

Fig. 1 Precision versus recall curves for object 3 for three different values of visual words (M)

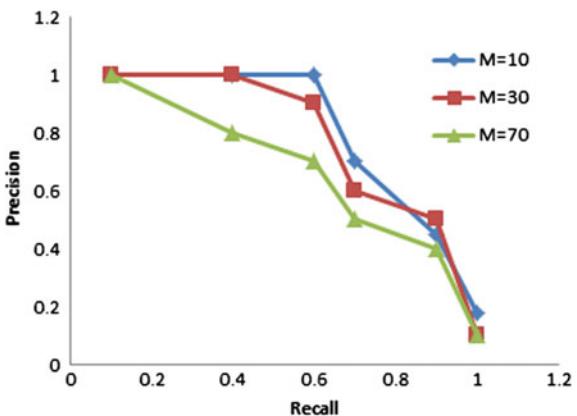


Fig. 2 Precision versus recall curves for object 10 for three different values of visual words (M)

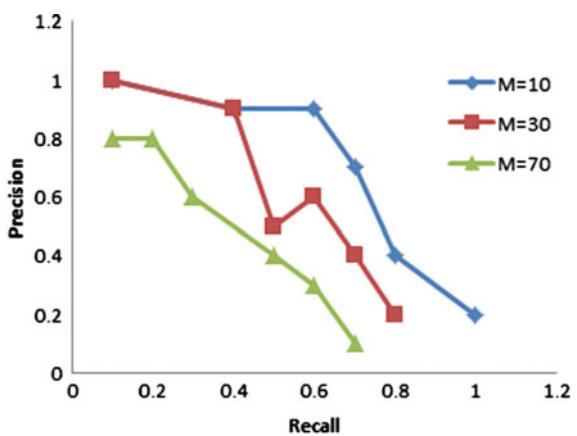
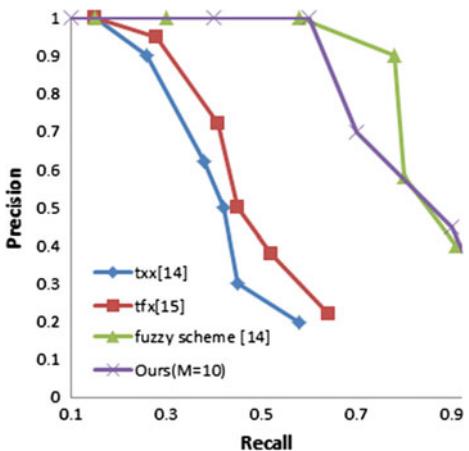


Fig. 3 Precision versus recall curves of four different methods: txx [17], tfx [15], fuzzy weighting [13], ours with M = 10 for object 3



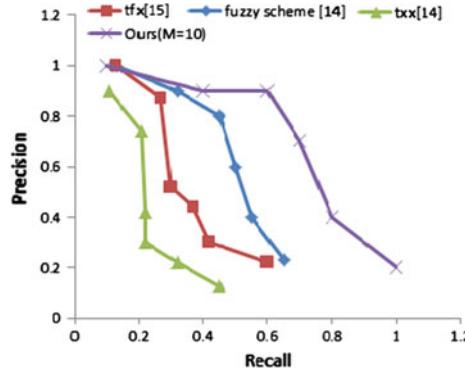


Fig. 4 Precision versus Recall curves of four different methods: txx [17], tfx [15], fuzzy weighting [13], ours with M = 10 for object 10

Table 1 Recognition rates for our method for M = 10: with and without the penalty term

Image	Ours (without penalty)	Ours
Object 1	0.5	0.8
Object 2	0.4	0.6
Object 3	0.8	1.0
Object 4	1.0	1.0
Object 5	0.6	0.75
Object 6	0.9	1.0
Object 7	0.8	1.0
Object 8	0.7	0.75
Object 9	0.5	0.7
Object 10	0.75	1.0
Average	0.695	0.86

show the results of our method with and without the penalty term. The results clearly demonstrate the utility of the penalty term as the recognition rate is always found to increase with the inclusion of the penalty term. Finally, in Table 2, we compare the recognition rates for all four methods. Once again, the values show the superiority of the proposed method where in all ten cases our method turns out to be the (single or joint) winner having achieved the highest recognition rate. Furthermore, our average recognition rate of 86 % is significantly better than all the other three competing methods. The second best is [13] with an average recognition rate of 80 % while [15, 17] are way behind with average recognition rates of 71.5 % and 61.5 % respectively.

Table 2 Recognition rate comparison among four different competing methods

Image	txx	tfx	Fuzzy weighting	Ours ($M = 10$)
Object 1	0.5	0.4	0.65	0.8
Object 2	0.4	0.1	0.45	0.6
Object 3	0.9	0.95	1.0	1.0
Object 4	1.0	0.9	1.0	1.0
Object 5	0.25	0.1	0.75	0.75
Object 6	1.0	1.0	1.0	1.0
Object 7	1.0	0.85	0.95	1.0
Object 8	0.55	0.5	0.7	0.75
Object 9	0.7	0.6	0.6	0.7
Object 10	0.85	0.75	0.9	1.0
Average	0.715	0.615	0.8	0.86

4 Conclusion

In this paper, we proposed a novel assignment model of visual words for the purpose of content-based image retrieval. Orientation adaptive probabilistic affinities of the patches are computed for M most similar visual words. We also introduce a novel patch/image dissimilarity measure with a penalty term, which is shown to improve the retrieval performance. A comparison of precision and recall values for experiments on COIL-100 database with three existing approaches clearly shows the superiority of the proposed scheme.

In future, we plan to address additional limitations of BoVW representations like lack of semantic knowledge [18] in the proposed model. Another direction of future work will be to explore tree-based clustering approaches for improving the visual words-based image retrieval [19].

References

1. Datta, R., Joshi, D., Li, J. and Wang, James Z.: Image retrieval: Ideas, influences, and trends of the new age, *ACM Comput. Surv.*, 40(2), 1–60, (2008).
2. Liu J.: Image Retrieval based on Bag-of-Words model, *CoRR*, [arXiv:1304.5168](https://arxiv.org/abs/1304.5168) (2013).
3. Liu G. and Yang J.: Content-based image retrieval using color difference histogram, *Pattern Recognition*, 46(1), 188–198, (2013).
4. Hejazi M. R. and Ho Y.: An efficient approach to texture-based image retrieval, *Int. J. Imaging Systems and Technology*, 17(5), 295–302, (2007).
5. Santini S. and Jain R.: Similarity Measures, *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(9), 871–883, (1999).
6. Goldberger J., Gordon S., Greenspan H.: An Efficient Image Similarity Measure Based on Approximations of KL-Divergence Between Two Gaussian Mixtures, *9th IEEE International Conference on Computer Vision (ICCV)*, 487–493, (2003).

7. Hu W., Xie N., Li, L., Zeng, X., Maybank, S. J.: A Survey on Visual Content-Based Video Indexing and Retrieval, *IEEE Trans. Systems, Man, and Cybernetics, Part C*, 41(6), 797–819, (2011).
8. Lowe D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *Int. J. Computer Vis.*, 60(2), 91–110, (2004).
9. Sivic, J., Zisserman, A.: Video Google: Efficient Visual Search of Videos, In *Toward Category-Level Object Recognition*, 127–144, (2006).
10. Squire, D. M., Wolfgang, M., Henning, M., Thierry P.: Content-based query of image databases: inspirations from text retrieval, *Pattern Recognition Lett.*, 21(13-14), 1193–1198, (2000).
11. Yang, J., Jiang, Y. G., Hauptmann, A. G., Ngo, C. W.: Evaluating bag-of-visual-words representations in scene classification, *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, 197–206, (2007).
12. Kato, H., Tatsuya H.: Image Reconstruction from Bag-of-Visual-Words, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 955–962, (2014).
13. Bouachir, W., Kardouchi, M., Belacel, N.: Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting Scheme for Efficient Indexation, *Fifth International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 215–220, (2009).
14. Kannan, S. R., Ramathilagam, S., Chung, P. C.: Effective fuzzy c-means clustering algorithms for data clustering problems, *Expert Syst. Appl.*, 39(7), 6292–6300, (2012).
15. Sivic, J., Zisserman A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *9th IEEE International Conference on Computer Vision (ICCV)*, 1470–1477, (2003).
16. Nene, S. A., Nayar, S. K., Murase, H.: Columbia Object Image Library (COIL-100), Technical Report, Department of Computer Science, Columbia University CUCS-006-96, (1996).
17. Newsam, S., Yang Y.: Comparing global and interest point descriptors for similarity retrieval in remote sensed imagery, *15th ACM International Symposium on Geographic Information Systems (ACM-GIS)*, 1–9, (2007).
18. Zhiwu, L., Wang L., Rong, W.: Image classification by visual bag-of-words refinement and reduction, *CoRR*, [arXiv:1501.04292](https://arxiv.org/abs/1501.04292), (2015).
19. Dimitrovski I., Kocev D., Loskovska S., Dzeroski S.: Improving bag-of-visual-words image retrieval with predictive clustering trees, *Information Sciences*, 329, 851–865, (2016).

Online Support Vector Machine Based on Minimum Euclidean Distance

Kalpana Dahiya, Vinod Kumar Chauhan and Anuj Sharma

Abstract The present study includes development of an online support vector machine (SVM) based on minimum euclidean distance (MED). We have proposed a MED support vector algorithm where SVM model is initialized with small amount of training data and test data is merged to SVM model for incorrect predictions only. This method provides a simpler and more computationally efficient implementation as it assign previously computed support vector coefficients. To merge test data in SVM model, we find the euclidean distance between test data and support vector of target class and the coefficients of MED of support vector of training class are assigned to test data. The proposed technique has been implemented on benchmark data set mnist where SVM model initialized with 20 K images and tested for 40 K data images. The proposed technique of online SVM results in overall error rate as 1.69 % and without using online SVM results in error rate as 7.70 %. The overall performance of the developed system is stable in nature and produce smaller error rate.

Keywords Support vector machines • Online support vector machines • Sequential minimal optimization • Euclidean distance • Classification

K. Dahiya

University Institute of Engineering and Technology, Panjab University,
Chandigarh, India
e-mail: kalpanas@pu.ac.in

V. Kumar Chauhan · A. Sharma(✉)

Computer Science and Applications, Panjab University, Chandigarh, India
e-mail: anujs@pu.ac.in
URL: <https://sites.google.com/site/anujsharma25>

V. Kumar Chauhan

e-mail: vkumar@pu.ac.in

1 Introduction

The Support Vector Machine (SVM) has been widely used in the real life applications meant for mainly classification purposes. The SVM technique is based on structural risk minimization principle [15]. The SVMs common advantages include: minimization of empirical risk and to prevent overfitting data; to find maximal margin hyperplane and convex quadratic form of problem; obtained classifier with support vectors and the kernel function-based training especially in the nonlinear cases [13]. The computational complexity of SVM is the main drawback as it requires lots of memory during training stage. Also, batch mode nature of SVM limit the use in real-life applications where training from unknown data is beneficial. The one popular solution to large memory use has been given by Platt where a sequential minimal optimization algorithm is proposed that handle two variables in one iteration [12]. This reduces the memory cost for the current process in execution and further advancements in SVMs to speed up these processes that help in less training time [4, 10]. The solution to other disadvantage as batch mode learning of SVM can be solved with online SVM where an incremental learning technique is adopted that include small amount of training data and use rest trained data with test data for training and prediction purposes. The online SVM learning is a challenging task as it includes training of SVM with test data where continuous adjustment of trained SVM parameters required. The re-train of SVM is one common answer to online SVM training. In this paper, we have proposed an online SVM using minimum euclidean distance technique. Our algorithm find support vectors closely related to misclassified test data and adopt their coefficients for future use. We apply offline step to adjust SVM parameters once training size exceeds a threshold level. The experimental results obtained on benchmark data prove the effectiveness of our proposed approach.

This section includes selected work in recent past for online SVM. In online SVM, a SVM trained through incremental algorithm where new samples and support vectors together re-train the SVM and incrementally update the SVM classifier. It works as an online support vector classifier that trains data in sequential order irrespective of traditional batch mode training. The data is trained with respect to KKT conditions and data is trained until no data points violates KKT conditions [8]. An online SVM as LASVM presented with useful features as fast training SVM. The LASVM include process and reprocess as two major processes. The process inserts a data point in current support vectors and search for other data point as violating pair with maximal gradient. The reprocess performs a direction search and removes blatant non-support vectors [2]. An active query-based incremental training SVM presented where subset of training data selected by K-means clustering to compute initial separated hyper planes. An active query is used to update the classifier at each iteration [5]. A larank algorithm to solve multiclass SVM that compute and update gradient in a random pattern [1]. Inspired from larank algorithm, an online svm proposed where misclassified examples are used to update model and weights are initialized for new example subject to importance of new class [18]. A $2v$ online granular SVM work for biometric classifier update algorithm that incrementally re-train the classi-

fier using online learning and establishes decision hyperplanes for improved classification. Their experiments were focused to face recognition [14]. An online independent SVM developed where approximation is controlled by user defined parameter. This procedure employs linearly independent observations and tries to project new observations [11]. An online core vector machine with adaptive minimum enclosing ball adjustment proposed where permanently training samples are deleted that do not influence final trained classifier and adjustment of classifier can be made online with new misclassified samples. These algorithm include steps as offline training of existing samples to get initial classifier coefficients; online adjustment with web and online adjustment of coefficients [17]. An online SVM based on convex hull vertices selection where small number of skeleton samples constitute an appropriate convex hull in each class of current training samples and classifier is updated with new samples. This procedure resulted in less training time and improved classification rate [16]. our approach include training of the test data items that are predicted incorrectly as used in literature work. The newness of our approach include MEDSV algorithm where we have not observed this technique in previous work.

This paper include four sections including this section as Introduction. The selected literature related to online SVM has been discussed in this section. The system overview has been discussed in section two. The section three presents details for computation of support vectors and proposed algorithm as Minimum Euclidean Distance Support Vectors (MEDSV). The experimental results and concluding remarks are discussed in section four.

2 System Overview

The present system include SVM classifier where a kernel trick is applied that takes data from lower dimension to higher dimension. This involves data mapped to higher dimension feature space where decision surface is found. A kernel function is applied to map data in higher dimension. In order to implement proposed work for online SVM, the support vector coefficients are computed initially or at threshold levels. The flow chart presented in Fig. 1 presents the proposed system. The common stages in the developed system include:

1. The model is initialized with some amount of small training data where support vectors are calculated. Repeat step 2 for every test data and check threshold condition after step 2.
2. The new test data is predicted w.r.t model and if prediction is wrong, the test data euclidean distance is measured w.r.t support vectors of model of target class. The support vector coefficients of MED target class support vector data are assigned to test data and merge to SVM model. The model is updated now.
3. If sufficient number of test data are predicted, the model is further updated with calculations of support vectors as in stage 1 and model is updated.

In above step 1, the model is initialized referred to build model with small amount of training data collected from all classes. the model now checks test data as stated in step 2 and consider test data further for incorrect prediction of test data class. The incorrect prediction in Fig. 1 referred as misclassified. For incorrect prediction cases, compute the euclidean distance between test data and target classes support vectors data. The minimum distance support vector is selected and its coefficients are assigned to test data. the test data is now merged to model with coefficients of target class data of MED in the model. The coefficients are referred to the weights for respective classes of a support vector. For example, in m classes, a support vector of class i will have $m - 1$ coefficients, one each for other $m - 1$ classes. In step 3, if sufficient number of test data are predicted (threshold level), the model is updated as happened in initializing of model in step 1. This helps in regular updation of model after frequent intervals. The threshold level is subject to decisions based on experiments where new coefficients need to be computed. The model presented in Fig. 1 is iterative in nature and scope of addition of new data or removal is feasible in view to improve predictions of test data. This makes system better in real life use where test data is opted as training data on the basis of level of complexity of test data predictions.

3 Minimum Euclidean Distance-Based Support Vectors Algorithm

The SVM solves labeled training data (x_i, y_i) , $i = 1, 2, 3, \dots, n$ for the decision function $h(x) = \text{sign}(w.z + b)$, where $z = \phi_z(x) \in Z$ is a feature map of x [6]. The w and b are the solutions of:

$$\begin{aligned} & \text{minimize } \frac{1}{2}(w.w) + C \sum_{i=1}^n \xi_i \\ & \text{s.t. } y_i((w.z_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (1)$$

C is penalty parameter and ξ_i is slack variable. In view to see large computational complexity as $O(n^3)$ of SVM training in equation (1), the dual form of (1) having less complexity. The lagrange multipliers are used to derive dual form of (1) and results into:

$$\begin{aligned} & \text{minimize } f(\alpha) = - \sum_{i=1}^n \alpha_i + \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (z_i \cdot z_j) \\ & \text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^n \alpha_i y_i = 0, \text{ for } i = 1, 2, \dots, n \end{aligned} \quad (2)$$

As discussed in previous section one, platt [12] sequential minimal optimization technique is one solution to large quadratic SVM problem where large quadratic

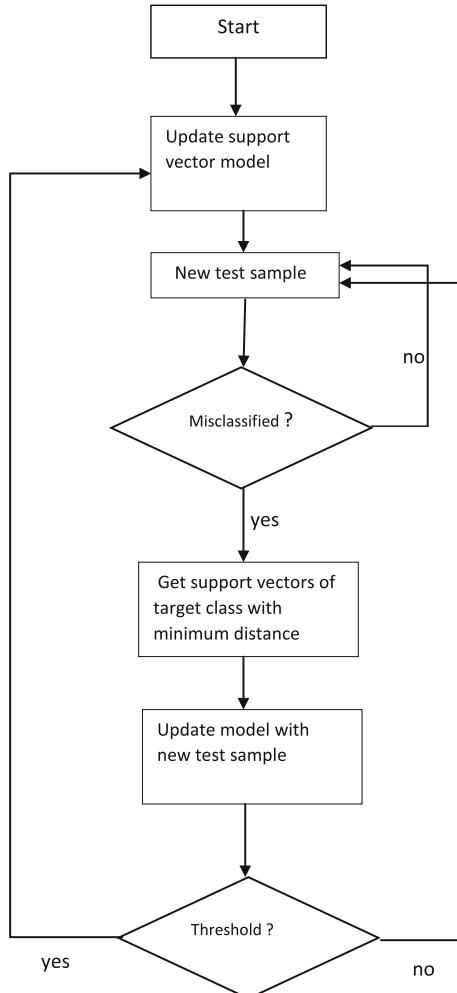


Fig. 1 Flow chart of proposed system workflow

problem is broken to subparts results into two variables at a time. The selection of current working set and updating lagrange multiplier for these two variables are the two important parts of sequential minimal optimization technique. These iterations update few components of problems until convergence level reached. We have used second order information to select current working set [7]. This method is suitable for large data sets and results in selection of few working sets that reduces overall execution time to compute support vectors. The current working sets are called maximal violating pairs where violation is subject to KKT conditions [3]. The maximal violating pair $B(i,j)$ is calculated as:

$$\begin{aligned} i &\in \arg \max_t \{-y_i f(\alpha^k)_t | t \in I_{up}(\alpha^k)\} \\ j &\in \arg \min_t \{Sub(\{i, t\} | t \in I_{low}(\alpha^k), -y_t \nabla f(\alpha^k)_t < -y_i \nabla f(\alpha^k)_i\} \quad (3) \\ Sub(\{i, t\}) &\equiv \nabla f(\alpha)^T \Delta \alpha + \frac{1}{2} \Delta \alpha^T \nabla^2 f(\alpha) \Delta \alpha, \Delta \alpha_l = 0 \text{ for all } l \neq i, j. \end{aligned}$$

The total number for selecting j in working sets checks only $O(n)$ possible pairs. Let $a_{ij} = (z_i, z_i) + (z_j, z_j) - 2(z_i, z_j)$ and $b_{ij} = -y_i \nabla f(\alpha^k)_i + y_j \nabla f(\alpha^k)_j$, this makes working set selection as:

$$\begin{aligned} i &\in \arg \max_t \{-y_i f(\alpha^k)_t | t \in I_{up}(\alpha^k)\} \quad (4) \\ j &\in \arg \min_t \frac{-b_{ij}^2}{a_{ij}} | t \in I_{low}(\alpha^k), -y_t \nabla f(\alpha^k)_t < -y_i \nabla f(\alpha^k)_i \} \end{aligned}$$

The updatation of lagranges multiplier for i, j is as follows:

$$\begin{aligned} \alpha_i^{new} &= \begin{cases} \alpha_i + \frac{y_j b_{ij}}{a_{ij}} & \text{if } a_{ij} > 0 \\ \alpha_i + \frac{y_j b_{ij}}{\tau} & \text{if } a_{ij} \leq 0 \end{cases} \\ \alpha_j^{new} &= \begin{cases} \alpha_j - \frac{y_i b_{ij}}{a_{ij}} & \text{if } a_{ij} > 0 \\ \alpha_j - \frac{y_i b_{ij}}{\tau} & \text{if } a_{ij} \leq 0 \end{cases} \end{aligned} \quad (5)$$

The τ is a constant. Then, α_i^{new} and α_j^{new} need to be clipped in range $[0, C]$ as discussed in SMO algorithm.

$$\alpha_i^{new,clipped} = \begin{cases} 0 & \text{if } \alpha_i^{new} < 0 \\ C & \text{if } \alpha_i^{new} > C \\ \alpha_i^{new} & \text{otherwise} \end{cases} \quad (6)$$

before $\alpha_j^{new,clipped}$, the updated α_i^{new} need to use for α_j^{new} as:

$$\alpha_j^{new} = y_i(y_i \alpha_i + y_j \alpha_j - y_i \alpha_i^{new,clipped}) \quad (7)$$

$$\alpha_j^{new,clipped} = \begin{cases} 0 & \text{if } \alpha_j^{new} < 0 \\ C & \text{if } \alpha_j^{new} > C \\ \alpha_j^{new} & \text{otherwise} \end{cases} \quad (8)$$

The α_i and α_j are the updated lagrange multipliers and lie on either or same side of hyperplanes as selected from from violating pairs i and j . These α_i and α_j are

multiplied to ‘−1’ or ‘+1’ as the case may be. the final value is group of coefficients (α_i and α_j) and the feature vector data of i and j violating pairs.

As discussed in section two, the model is initialized using small amount of training data or model is updated after frequent intervals. Only incorrect predicted test vector is merged to model with SV coefficients selected based on minimum distance. Therefore, two important stages as class prediction of a test vector and selection of SV coefficients happen in updated model. The first stage prediction of a class is based on the decision function as

$$h(x) = \text{sign}(\sum_{i=1}^n \alpha_i y_i z_i \cdot z + b) \quad (9)$$

Here, α_i 's are computed SV coefficients, y_i is target value, z_i is the i th SV and z is the test vector. The class prediction of a test vector is based on ‘one-versus-one’ strategy for multi classes problem with $c(\geq 2)$ classes. This makes $c(c - 1)/2$ training pairs of classes and the class that gains maximum votes for a test vector is the recognized class for test vector. In rare cases, if two or more classes shares same number of votes, the class with minimum index is opted. The second stage is to find minimum distance of test vector w.r.t SV of recognized class in model. The SV that results in minimum value of distance is selected and the corresponding SV coefficients are assigned to test vector. Finally, the selected SV coefficients and test vector together as a complete SV merged to recognized class in the model that results in new updated model for next test vector. The distance is computed using following formula as:

$$d = \sqrt{\sum_{i=1}^n (x_{t_i} - x_{j_i})^2} \quad (10)$$

In above equation to compute distance (d) between test vector and a support vector from model, the n is the number of feature vectors, (x_t) is the test vector and x_j is the current support vector in SVM model. As we have used threshold to check sufficient number of test vectors to merge in model, the model is again re-trained after threshold level is reached. This results in model with new and updated support vectors. The algorithm for online SVM using MED has been presented in MED Support Vectors (MEDSV) Algorithm.

In MEDSV algorithm, model is trained SVM with first time selected data. The algorithm goes to next steps if target class is not equal to recognized class. In step 3, support vectors belong to target class are selected from model, and from steps 5 to 11, selected support vectors euclidean distance is computed against test vector. The support vector with MED is selected and it is merged to model with its coefficients.

Algorithm 1 MEDSV Algorithm

```

1: Input: model is trained SVM structure,  $x_t$  is test vector,  $tc$  is target
   class of test vector and  $vc$  is recognized class of test vector.
2: return if  $tc == vc$  is true.
3: Initialize:  $med = +\infty$ ,  $flag = false$ .
4: Extract cSV from model of class  $tc$ . Let  $m$  is the size of cSV.
5: for  $j = 1 \rightarrow m$  do
6:    $ed = \sqrt{\sum_{i=1}^n (x_{t_i} - x_{j_i})^2}$ ,  $n$  is the size of feature vector.
7:   if  $ed \leq med$  then
8:      $med = ed$ .
9:      $flag = true$ .
10:  end if
11: end for
12: if  $flag == true$  then
13:    $model.length = model.length + 1$ .
14:    $model.SV = model.SV \cup x_t$ .
15: end if

```

4 Experimental Results

The performance of proposed online SVM using MED has been evaluated on benchmarked dataset mnist [9]. The mnist dataset include 60 K as training images for digit classes 0–9. As stated in Sect. 2, the developed system input small of training data, therefore, one third as 20 K training data images has been considered to train initial SVM model and rest two third as 40 K data images as test data. The test data is predicted class-by-class and SVM model is updated for every incorrect target class prediction for test data. We train our system under online environment for incorrect predicted test data images and do not include the correctly predicted test data images. The SVM training model include rbf kernel as:

$$K(z_i, z_j) = \exp(-\gamma \|z_i - z_j\|^2) \quad (11)$$

The feature vector length is 784 a row feature vector of an image with dimension as 28×28 . The initially trained model output in total support vectors as 9102 and these support vector for individual classes 0–9 are presented in Table 1. The evaluation has been performed in two separate experiments as SVM classification and online SVM classification. The SVM classification evaluate only a class classified

Table 1 Support vectors computed for classes 0–9 for initially trained SVM and online SVM

Class	SVM support vectors	Online SVM support vectors	Difference
0	576	579	3
1	569	576	7
2	911	929	18
3	1015	1062	47
4	971	1029	58
5	1229	1288	59
6	728	775	47
7	803	911	108
8	1099	1220	121
9	1201	1410	209
<i>Total →</i>	9102	9779	677

correctly or not and do not modify initially trained SVM. The online SVM classification refers to the proposed work in this study. In view to give clear performance of proposed online SVM, we have not included threshold level during testing of 40 K test data images. The error rate for 40 K test images shows that online SVM perform far better as compare to SVM. The online SVM overall error rate is 1.69 % and SVM error rate is 7.70 %. Table 2 presents the results for all classes 0–9 with SVM and online SVM. The output presented in Table 2 shows that classes 2, 3, 5, 7 and 8 results in improvement more than 6 % as 8.11 %, 10.89 %, 8.71 %, 6.19 % and 9.61 %, respectively. The improvements in all classes indicate that proposed online SVM capable to adapt complex pattern subject to similar type of pattern merging during online SVM using MED.

The overall results in Table 2 motivate to work with this approach as data initially trained with 20 K training images results in 9102 support vectors and test data with 40 K images using online SVM makes total support vectors as 9779 as shows in Table 1. From Table 1, we find that an increase of 677 support vectors observed for 40 K test images as its computed from Table 2 also. In Table 2, total correctly predicted test images are 39323 which are 677 less than 40 K images. This indicate that our algorithm MEDSV do not leave chance to add test data item that is incorrectly predicted. The MEDSV algorithm results shows that the proposed technique could be more helpful in real life applications where data is tested class wise. In addition, the training of a particular class should be a part of initial trained SVM so that it could assign proper support vector coefficients to incorrectly predicted test data items. In absence of particular class training during initial SVM training, the same class would try to adapt the pattern of closely related class patterns during online SVM training. The model presented in this study has been developed in view to consider training of all classes that will become part of prediction during test stages.

Our system is competitive, stable and produce smaller error rate. In addition, we have noticed two challenges of present system as first, it increases the error rate

Table 2 The incorrectly predicted test data images with SVM and online SVM

Class	Test images	Test images predicted using SVM	Test images predicted using Online SVM	Overall improvement (% upto two decimal places)
0	3923	3814	3920	2.70
1	4742	4640	4735	2
2	3958	3619	3940	8.11
3	4131	3634	4084	10.89
4	3842	3613	3784	4.45
5	3421	3064	3362	8.71
6	3918	3736	3871	3.45
7	4265	3893	4157	6.19
8	3851	3360	3730	9.61
9	3949	3545	3740	4.94
<i>Total →</i>	40000	36918 (7.70 % as error rate)	39323 (1.69 % as error rate)	6.01 % as average improvement

with increase in number of classes and second, it perform better in sequential data class-by-class. These challenges do not discourage the present findings as our online SVM overall performance is better and we have not used role of SMO technique during online training process as suggested in literature. The use of SMO technique could reduce the above mentioned challenges in return with very small increase in time complexity. We derived an efficient implementation of online SVM and pointed out that proposed MEDSV algorithm could be a promising technique. We showed that how assigning previously computed coefficients could help in extending SVM training data. We took advantage of previous work from literature where computation of coefficients has been done in online SVM. Our method compute coefficients initially or at threshold level. Our results indicate that proposed technique need to further work for large number of classes and large test data set in random order.

References

1. A. Bordes, L. Bottou, and J. Gallinari. Solving multiclass support vector machines with larank. *International Conference on Machine Learning*, 227:89–96, 2007.
2. A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning and Research*, 6:1579–1619, 2005.
3. C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
4. C. Chang, C. W. Hsu, C, and J. Lin, C. The analysis of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 11(4):1003–1008, 2000.
5. S. Cheng and F. Shih. An improved incremental training algorithm for support vector machines using active query. *Pattern Recognition*, 40(3):964–971, 2007.
6. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

7. R.E. Fan, P.H. Chen, and C.J. Lin. Working set selection using second order information for training support vector machines. *Journal of Machine learning Research*, 6:1889–1918, 2005.
8. W. Lau, K and H. Wu, Q. Online training of support vector classifier. *Pattern Recognition*, 36(8):1913–1920, 2003.
9. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pages 2278–2324, 1998.
10. J. Lin, C. A formal analysis of stopping criteria of decomposition methods for support vector machines. *IEEE Transactions on Neural Networks*, 13(5):1045–1052, 2002.
11. F. Orabona, C. Castellini, B. Caputo, L. Jie, and G. Sandini. On-line independent support vector machines. *Pattern Recognition*, 43:1402–1412, 2010.
12. J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Microsoft Research*, pages MSR-TR-98-14, 1998.
13. A. Shilton, M. Palaniswami, and T. A. Incremental training of support vector machines. *IEEE Transactions on Neural Networks*, 16(1):114–131, 2005.
14. R. Singh, M. Vatsa, A. Ross, and A. Noore. Biometric classifier update using online learning: A case study in near infrared face verification. *Image and Vision Computing*, 28:1098–1105, 2010.
15. V. Vapnik. The overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.
16. D. Wand, H. Qiao, and B. Zhang. Online support vector machine based on convex hull vertices selection. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):593–609, 2013.
17. D. Wang, B. Zhang, P. Zhang, and H. Qiao. An online core vector machine with adaptive meb adjustment. *Pattern Recognition*, 43:3468–3482, 2010.
18. M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King, and S. T. Wong. Novel cell segmentation and online svm for cell cycle phase identification in automated microscopy. *Bioinformatics*, 24(1):94–101, 2008.

Design and Development of 3-D Urban Geographical Information Retrieval Application Employing Only Open Source Instruments

Ajaze Parvez Khan, Sudhir Porwal and Sangeeta Khare

Abstract Numerous 3-D GIS are being developed today that are both commercially and freely available. A multicity web-based 3-D GIS (named as GLC3d) using open source and freely available software/packages exclusively has been developed. This paper presents the architecture, design, and overview of GLC3d. Open source tools and software's QGIS, Cesium, and MySQL are employed to develop this application. QGIS is utilized for data preparation such as raster, vector, and subsidiary data. MySQL is utilized as database store and data population. GLC3d is based on Cesium (a JavaScript library for creating 3-D globes and 2-D maps in a web browser) and WebGL (a JavaScript API for rendering interactive 3-D computer graphics and 2-D graphics) to display 3-D Globe on the web browser. 3-D visualization for the urban/city geodata is presented on an interactive 3-D Globe. Various city information are generated and incorporated in to the 3-D WebGIS for data representation and decision making.

Keywords GIS (Geographical Information System) • WebGL (Web Graphics Library) • Cesium • MySQL • KML (Key Hole markup Language) • GeoJSON • WMS (Web Map Service) • OGC (Open Geospatial Consortium) • DTM (Digital Terrain Model)

A.P. Khan (✉) · S. Porwal · S. Khare
DEAL, Ministry of Defence, DRDO, Dehradun, India
e-mail: ajazeparvez.khan@deal.drdo.in

S. Porwal
e-mail: sudhir.porwal@deal.drdo.in

S. Khare
e-mail: sangeetakhare@deal.drdo.in

1 Introduction

GIS has emerged as a powerful tool for organizing, storing, capturing, manipulating, and illustrating the information/details of geographic locations. 2-D GIS has grown into an important part of the world's information system with the major support from information technology, advanced hardware, and sophisticated softwares. It has ample applications in the field of research, health management, environmental sciences, transportation and geospatial industry, urban mapping, public health, sustainable development, climatology, and archaeology [1–6].

The 3-D web mapping is a step ahead of 2-D mapping, where users/planners can get a rational practical view for planning and decision making. 3-D rendering on web for GIS requires the user to possess WebGL-enabled web browser and an advanced Graphics Adapter. 3-D rendering software applications such as WebGLEarth, NASA World Wind, and Google Earth are available, which interactively display and render Geodata/DTM on 3-D Globe. Proprietary software like ArcGIS is available for creating the geodata, which may then be displayed on a 3- D Globe using the ArcGlobe software [7–9]. This paper presents a web-based 3-D GIS application developed on WebGL [10] and Cesium. This paper also discusses the advancement and advantages over existing 3-D GIS. Above-mentioned solutions are neither web enabled nor lack certain functionalities for Urban purposes for instance adding prominent user defined vector layers directly to the 3-D GIS, displaying 3-D terrain and feature details, House level Information [11]. A MultiCity 3-D WebGIS (GLC3d) has been developed capable of presenting the city information and geographical details to the minute level on a lightweight 3-D Globe.

WebGL provides 3-D graphic support for this application without the requirement of any additional ingredients [12]. WebGL is incorporated within Cesium's API to exhibit the 3-D geodata on the browser.

GIS technology primarily consists of raster and vector layers supported by additional data/information. The developed application presented in this paper provides the user with the capability to view and infer information from Multicity WebGIS (View Modes: 2-D and 3-D). Map image (and basemaps) is accessed over the Internet as OGC Web Map Service (WMS). Various applications like city planning, disaster management, and environmental monitoring may be benefited from this 3-D information [13].

2 Methodology

The approach devised to develop an open source web-based 3-D GIS is divided into four phases. In the first phase, the detailed information to be incorporated in the GIS is prepared, which primarily constitutes geocoding, data format conversion,

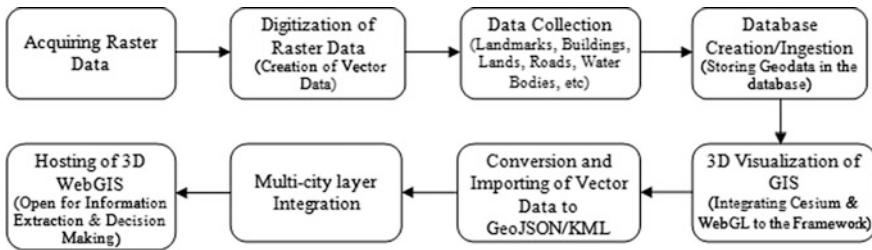


Fig. 1 Schematic procedure

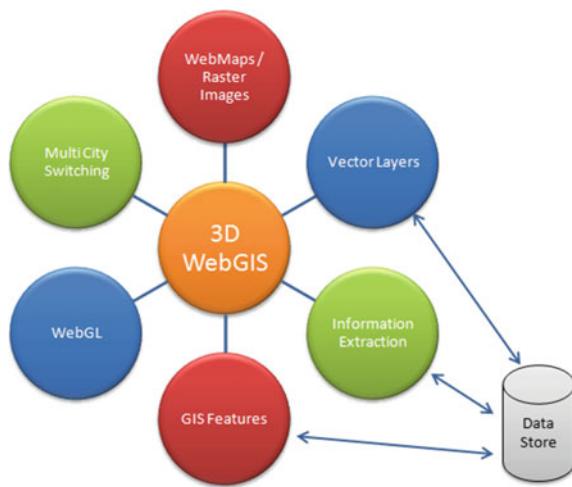
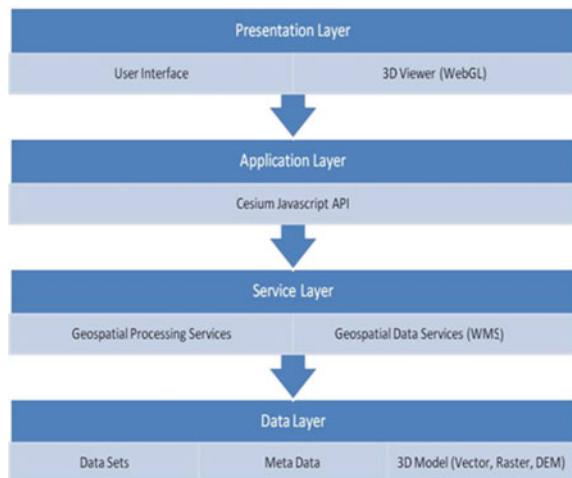
raster/vector layers designing, error detection, and rectification. To prepare the data for the 3-D WebGIS, QGIS software has been utilized [14].

In the second phase, database in MySQL is created to store the gathered information. This information includes the detailed urban characteristics of desired locations collected through contributory information available on assorted public domain or field survey.

For the third phase, 3-D visualization of the GIS is accomplished by incorporating Cesium and WebGL within the WebGIS application [15, 16] on HTML5 and supporting web technologies. Conglomerate city layers integration onto a single GIS facilitates moving between these layers. In the final phase, integrating all the above components within a single application GLC3d, which will be utilized as a web-based application (Fig. 1).

3 Design Approach

The designing of the WebGIS application includes remote sensing (to acquire the data), GIS (to derive the meaningful information from the data), DBMS (to store the data), and web technologies like Cesium and WebGL (to display the information interactively in 3-D on the web browser) in close alliance and synergy. The specialized design embraces the steps to include the best available open source software's/modules and accomplish a lightweight wide ranging urban web-based application. The design and quad-layered architecture of GLC3d is illustrated in Figs. 2 and 3.

Fig. 2 GLC3d design**Fig. 3** Layered architecture of GLC3d

4 Comparison

Several 3-D WebGIS application and solutions are available. Table 1 compares GLC3d with the prominent 3-D WebGIS applications, WebGL Earth being the closest.

GLC3d is a City/Urban level 3-D WebGIS that is free, open source, runs on WebGL with easy terrain customization and KML/GeoJSON support.

Table 1 GLC3d comparison with prominent 3-D WebGIS

Features	<i>GLC3d (developed)</i>	WebGL earth	Google earth	ArcGlobe	NASA's world wind
Platform	Any browser that supports WebGL	Any browser that supports WebGL	Windows, Linux, Mac OS	Windows, Linux	Cross-platform
Free commercial usage	Yes	Yes	No	No	Yes
Open source	Yes	Yes	No	No	Yes
3-D view	Yes	Yes	Yes	Yes	Yes
Client technology	WebGL	WebGL	Downloaded plug-in	OpenGL	OpenGL
Runs on web browser	Yes	Yes	Yes	Uses ArcGIS server online	No (desktop)
Native data format support	Yes	By importing tile using MapTiler	No	Yes	Yes
KML support	Yes	No	Yes	Yes	Yes
DTM display	Yes	Yes	Yes	Yes	Yes

5 Results

Snapshots of the developed 3-D GIS are illustrated in Fig. 4. It shows detailed information of every landmark, building, land area, road, water body, transport facility, and recreation facility.

The 3-D city model of an area along with the individual information is depicted in Fig. 5. The urban classification is further subdivided into different categories, for instance buildings have been classified as houses, schools, hospitals, fire stations recreational places, and Government offices.

The additional GIS operations in GLC3d are:

- Base map and feature/vector layers Switch (subdivided into different categories).
- Measurement of Distance, Area, and Perimeter.
- Searching for particular location on the Globe.
- Features Information with neighborhood details.
- Toggle and Switching between visualization modes of the data.
- Different cities' multi-ity switch.



Fig. 4 Detailed features on 3-D globe

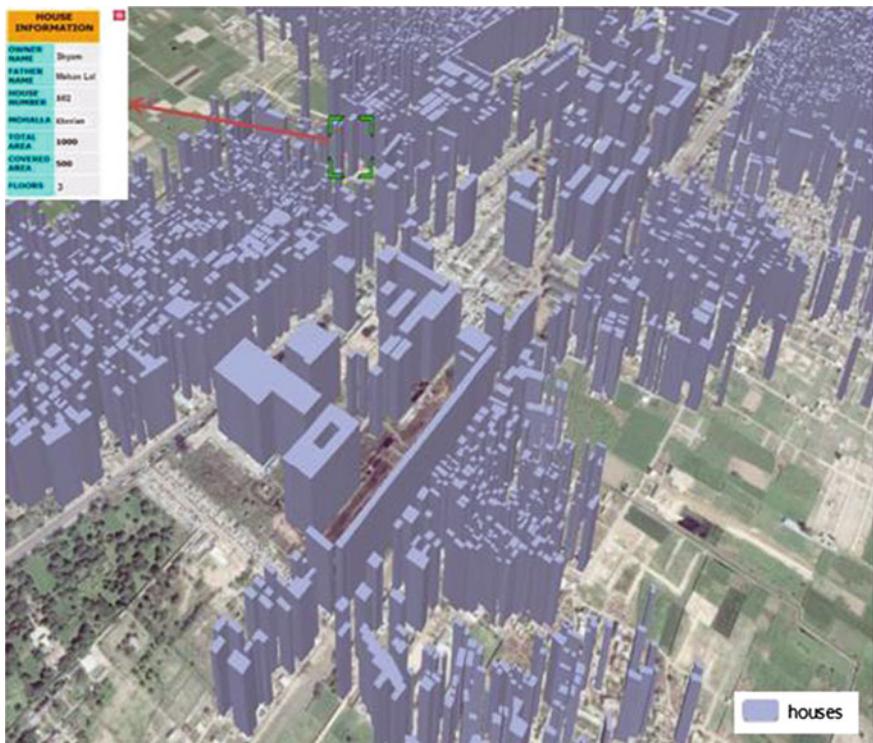


Fig. 5 3-D elevation model with individual information

6 Conclusion

A user-friendly high performance, economically viable [17] GIS application has been developed. Several Giant Web Maps like Google Maps, Bing OpenStreetMap, and others are also administering the GIS facilities on them [18–20]. Integration of GIS with Web Map and 3-D visualization has made GIS more useful and engaging in modern era information extraction and decision making.

3-D WebGIS application presented in this paper aims to retrieve the geographical information to the minutest degree possible, which is saved in spatial formats. The 3-D globe-based GIS application developed using only open source software/package/modules at back/front end authorizes the retrieval of significant urban and terrain information. The advantage of GLC3d is that it is a multicity 3-D WebGIS that is free, made only from open source tools, works on WebGL with superior support for KML/GeoJSON. Although this paper focuses on acquiring of urban/city information but this application may be extended to store and display any information/detail, and hence can be used in any GIS associated field, and decision making [1] for the benefit of mankind.

References

1. Paul A., Michael F. Goodchild, David J. Maguire and David W. Rhind: Geographical Information Systems, Second Edition, John Wiley and sons Inc. (1999).
2. Khan, A.P., Porwal, S., Khare, S.: Design and Development of Petite Urban Geographical Information Retrieval System Through Open Source Instruments, International Conference on Optics and Optoelectronics (ICOL), Dehradun (2014).
3. Zhen, Z.; Jing-min, J.; Fang, L. The application of geographic information system (GIS) in the field of public health, Second IITA International Conference on Geosciences and Remote Sensing (IITA-GRS), (2010) Vol. 2, 442–445.
4. Tirodkar, J.: Satellite Remote Sensing and GIS Based Sustainable Development of Coastal City, 4th International Conference on Advanced Computing & Communication Technologies (ACCT), (2014) 296–300.
5. Zou, S.; Feng, Z.D.; Yong L.; Xu, B.: GIS-assisted and climate-based modeling of spatial pattern of the potential ecological environments in the western part of the Chinese Loess Plateau, IEEE International Geosciences and Remote Sensing Symposium (IGARSS), (2014) Vol. 7, 4670–4673.
6. Khourami, E.H. and Santucci, J.F.: GIS in Archaeology, First International Symposium on Environment Identities and Mediterranean Area (ISEIMA), (2006) 441–446.
7. Qingquan, T.; Yongjun, Q.; Zhanying, W.; Qun, L.: Implementation on Google Maps-style WebGIS based on ArcGIS, 2nd International Conference on Advanced Computer Control (ICACC), (2010) Vol. 10, 60–64.
8. What is ArcGIS, support.esri.com, (accessed June 2015).
9. www.esri.com, (accessed Nov 2013).
10. cesiumjs.org, (accessed June 2015).
11. www.webglearth.org/feedback, (accessed July 2015).
12. Gregg Tavares: WebGL Fundamentals, www.html5rocks.com/en/tutorials/webgl/webgl_fundamentals, (accessed June 2015).

13. Krooks A., Kahkonen J., Lehto L., Latvala P., Karjalainen M., Honkavaara E.: WebGL Visualisation of 3D Environmental Models Based on Finnish Open Geospatial Data Sets, Volume XL-3, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, (2014) 163–169.
14. QGIS Documentation, docs.qgis.org, (accessed June 2015).
15. Cesium Developer Network, cesium.agi.com, (accessed June 2015).
16. Feng L.; Wang C.; Li C.; Li Z.: A Research for 3D-WebGIS based on WebGL, International Conference on Computer Science and Network Technology (ICCSNT), Vol. 1, (2011) 348–351.
17. Worrall, L. (eds): Geographic Information Systems: Developments and Applications, London: Belhaven Press.
18. <https://www.maps.google.com>, (accessed June 2015).
19. www.bing.com/maps, (accessed June 2015).
20. www.openstreetmap.org, (accessed June 2015).

A Textural Characterization of Coal SEM Images Using Functional Link Artificial Neural Network

Alpana and Subrajeet Mohapatra

Abstract In an absolute characterization trial, there are no substitutes for the final subtyping of coal quality independent of chemical analysis. Petrology is a specialty that deals with the understanding of the essential characteristics of the coal through appropriate chemical, morphological, or porosity analysis. Conventional analysis of coal by a petrologists is subjected to various shortcomings like inter-observer variations during screen analysis and due to different machine utilization, time consuming, highly skilled operator experience, and tiredness. In chemical analysis, use of conventional analyzers is expensive for characterization process. Thus, image analysis serves as an impressive automated characterization procedure of subtyping the coal, according to their textural, morphological, color, etc., features. Coal characterization is necessary for the proper utilization of coal in the power generation, steel, and several manufacturing industries. Thus, in this paper, attempts are made to devise the methodology for an automated characterization and sub-classification of different grades of coal samples using image processing and computational intelligence techniques.

Keywords Coal characterization • Image processing • Scanned electron microscopy • Computational intelligence

1 Introduction

Coal characterization is one of the major aspects for the proper utilization of coal. Prior to shipping, coal quality is commonly tracked by achieving random quality checks on finally finished coal block. The standard of a coal is mirrored within the variety of features like carbon content, ash content, and pores and edges present on its surface. Generally, good quality of coal has high carbon content, less ash

Alpana (✉) · S. Mohapatra

Department of Computer Science & Engineering, Birla Institute of Technology,
Mesra, Ranchi 835215, Jharkhand, India
e-mail: alpana.srk@gmail.com

content, and porosity. When the edges and pores are found to be a lot of on its surface with less carbon content and high ash content, the standard of coal is taken into account to be poor. The expert petrologists confirm the degree of quality of coal supported several methods like chemical analysis developed from existing expertise and by validating with standard samples. Chemical analysis like proximate analysis and ultimate analysis of coal are considered to be the standard procedure used for coal characterization. The coal is often categorized as per their carbon, oxygen, ash, and volatile matter content using this methodology. Characterization by these methods is time intense and needs highly skilled petrologists. Thus, an automated image-based characterizing system would be helpful. Image-based approach is considered to be an important tool for the characterization of coal that promotes the productivity and also raises the level of automation. In image-based approach, coal can be mainly characterized according to its textural, morphological, and color features.

In this paper, we propose how to characterize a coal according to its textural features using functional link neural network (FLANN). Instead of having advanced techniques like thermal imaging, infrared imaging, image analysis, etc., chemical analysis remains a gold characterization technique. Chemical analysis leads to be biased due to machine utilization, tiredness, etc., resulting in uncertain and conflicting results. Hence, there is a requirement for a robust and efficient automated system for a preliminary screening of coal, which has capability to increase the accurate outcomes without consuming more time and conflictiness.

Over the years, several image analysis techniques for characterization of coal images have been proposed. Typically, most of the techniques of image analysis are used to characterize according to the combustion potential of coal, particle size, and mass distribution in coal and ash content present in coal. Claudio et al. [1] proposed the interaction between oxygen and coal particles at a very low temperature. It might be utilized to calculate the utmost possible lane in the initial phase of oxidation that may then results to a voluntary combustion action of coal. Zhang et al. [2] discussed about an enhanced mass methodology for the prediction of particles of coarse coal by means of image analysis technique. This has been subsisted of two parameters, namely coal particles density and projected area. This paper concluded that the proposed method allowed the coarse coal particles to be sieved easily, accurately, and quickly. Jianguo et al. [3] estimated the grain size assessment of coal using image segmentation method. In this paper, the overlap difficulty of particles has been resolved by means of image enhancement technique as well as the edges on surfaces are detected by means of structural edge detection method. There are various identical researches on the characterization of coal in the literature survey. Due to composite features of the coal images and difference in the sample preparation techniques, a lot of task needs to be done to meet the real quality check demands. It is remarked from the history of survey that the automated process entirely depends upon the accurate feature extraction procedure.

In this research, we propose a methodology to automate the coal features from the images, which can complement the petrologists with an unbiased data for better characterization of coal. The method we introduce first separates the noises from the

image and then extracts the textural features from the given samples. These features can be used by FLANN classifier to classify the quality of coal accurately. The architecture of paper is organized as follows: the framework of the proposed method is described in Sect. 2. Section 3 presented the experimental results with detailed analysis on the results obtained. Finally, the concluding remarks are provided in the Sect. 4.

2 Methods

The method of characterization from microscopic images includes image acquisition, subimaging, feature extraction, and classification, respectively. These steps are briefly discussed as follows.

2.1 *Image Acquisition Using Scanned Electron Microscope (SEM)*

Coal samples are taken from CSIR-Central Institute of Mining and Fuel Research, Sambalpur, and Ranchi, India. Later, coal block is hammered and very small pieces are obtained. Afterwards, coal sample images are captured using scanned electron microscope under the setting of 5 kV accelerating voltage with an effective magnification of 500X. The captured coal image samples are shown in Fig. 1.

2.2 *Subimaging Using Bounding Box Method*

The captured coal sample images are generally larger in size, hence for accurate identification of features, the desired interested region has been determined using the bounding box techniques [4]. This method results subimages with no pattern should be overlapped. The bounding box method are shown in Fig. 2.

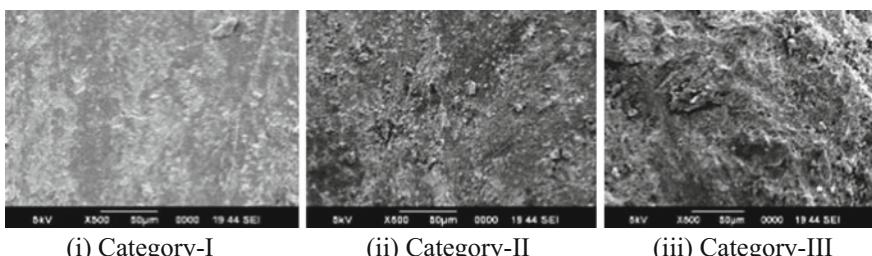


Fig. 1 Coal electron microscopic images

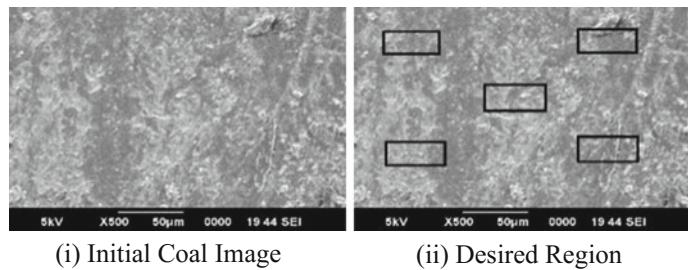


Fig. 2 Bounding box sub imaging

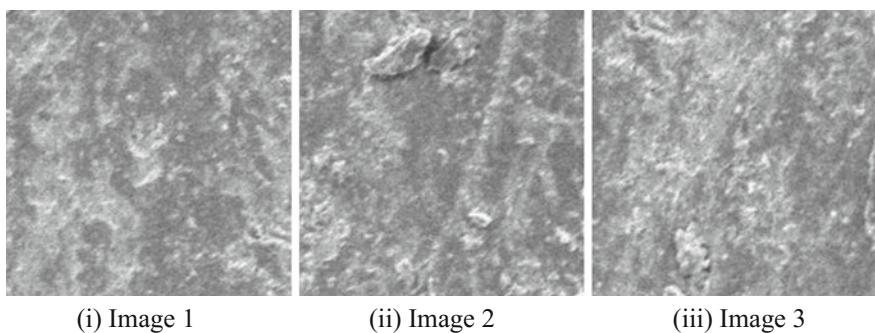


Fig. 3 Sample cropped sub images

The sample desired region of interest coal subimages is shown in Fig. 3. A gross of 15 subimages is taken using the bounding box method from the initial images of coal.

2.3 *GLCM Feature Extraction*

Gray level cooccurrence matrices (GLCM) are appropriate illustration of an image. It is used to characterize the textures of an image, where few contractions within the range of gray values have formerly been applied. The proportion of components of the cooccurrence matrix is exhibited to be decent texture descriptors, because of the ability to capture the analogous abundance of image features. The cooccurrence matrix resembles a joint probability density function and we use to characterize the probability density functions by calculating some statistical features from it. The most widely used parameters calculated from the gray level cooccurrence matrix are energy, entropy, contrast, homogeneity, and correlation. In every parameters case, gray level cooccurrence matrix has been standardized by subdividing its entire

elements by total amount of pairs of pixels taken, that results is a joint probability density function [5]:

$$P(x, y) = \frac{GLCM_d(x, y)}{\text{Total number of all pairs of pixels}} \quad (1)$$

where $GLCM_d(x, y)$ is a gray level cooccurrence matrix.

The parameters used here are formulized as follows:

i. Energy (e_1):

$$\sum_{x=0}^{GL-1} \sum_{y=0}^{GL-1} P(x, y)^2 \quad (2)$$

ii. Entropy (e_2):

$$\sum_{x=0}^{GL-1} \sum_{y=0}^{GL-1} P(x, y) \log P(x, y) \quad (3)$$

iii. Contrast (c_1):

$$\frac{1}{(GL-1)^2} \sum_{x=0}^{GL-1} \sum_{y=0}^{GL-1} (x - y)^2 P(x, y) \quad (4)$$

iv. Homogeneity (h_1):

$$\sum_{x=0}^{GL-1} \sum_{y=0}^{GL-1} \frac{P(x, y)}{1 + |x - y|} \quad (5)$$

v. Correlation (c_2):

$$\frac{\sum_{x=0}^{GL-1} \sum_{y=0}^{GL-1} x y p(x, y) - \mu_m \mu_n}{\sigma_m \sigma_n} \quad (6)$$

where

$$\begin{aligned} \mu_m &= \sum_{x=0}^{GL-1} x \sum_{y=0}^{GL-1} P(x, y) & \mu_n &= \sum_{y=0}^{GL-1} y \sum_{x=0}^{GL-1} P(x, y) \\ \sigma_m &= \sum_{x=0}^{GL-1} (x - \mu_m)^2 \sum_{y=0}^{GL-1} P(x, y) & \sigma_n &= \sum_{y=0}^{GL-1} (y - \mu_n)^2 \sum_{x=0}^{GL-1} P(x, y) \end{aligned}$$

In every cases, GL is the gross amount of gray levels that are used here.

2.4 Classification Using Functional Link Artificial Neural Network (FLANN)

Classifiers are mainly used for dividing the feature space into various classes depending on similar features in pattern recognition. In general, to figure out the classification work in pattern recognition, the standard algorithm is multilayer perceptron (MLP) [6]. Multilayer perceptron has one input layer, one output layer and multiple hidden layers. It takes much time to train the weight parameters vector and as the number of layers increases, the complexity of MLP neural network also increases. By considering above-mentioned issues, Functional link artificial neural networks (FLANN) have been used for the classification of coal into its sub categories. In contradiction to multiple layer perceptron networks, FLANN architecture signifies with feedforward network having single layer only. FLANN solves the nonlinear problems by means of the functionary broaden features, that is generally experienced in the single-layer neural network. The properties like simple architecture design and low network complexity inspires to use FLANN in characterization task [7]. A comprehensive study is represented that illustrates the effectiveness and performance of the FLANN classifier. The elements that are used as the training sets are texture features that are derived from the sub images and categories labels are target output. The five statistical texture inputs e_1, e_2, h_1, c_1 , and c_2 in the input layer are functionally expanded with the trigonometric basis functions as follows [8]:

$$\{1, e_1, \sin(\pi e_1), \dots, \sin(n\pi e_1), \cos(\pi e_1), \dots, \cos(n\pi e_1), e_2, \sin(\pi e_2), \dots, \sin(n\pi e_2), \cos(\pi e_2), \dots, \cos(n\pi e_2), c_1, \sin(\pi c_1), \dots, \sin(n\pi c_1), \cos(\pi c_1), \dots, \cos(n\pi c_1), h_1, \sin(\pi h_1), \dots, \sin(n\pi h_1), \cos(\pi h_1), \dots, \cos(n\pi h_1), c_2, \sin(\pi c_2), \dots, \sin(n\pi c_2), \cos(\pi c_2), \dots, \cos(n\pi c_2)\}$$

To determine the error, the actual output on the output layer is set to be compared with the target output provided by the petrologists. The backpropagation training method is used to update the weight matrix between the input–output layers depending on this error value. The proposed classifier (FLANN) is shown in Fig. 4.

The FLANN are trained with creation of the input–output patterns of different features for various images. The proposed FLANN classifier shows great enhancement in the classification accuracy in contrast to multilayer perceptron (MLP). The training convergence characteristics of the FLANN architecture for all the three desired individual output is depicted in Fig. 5.

2.5 K-fold Cross-Validation

Once the classification is done, the k-fold cross-validation [9] method is performed for training and testing of classifiers for the extraction of textural features of coal as the data set used in this study is too small. Here the datasets has been divided into

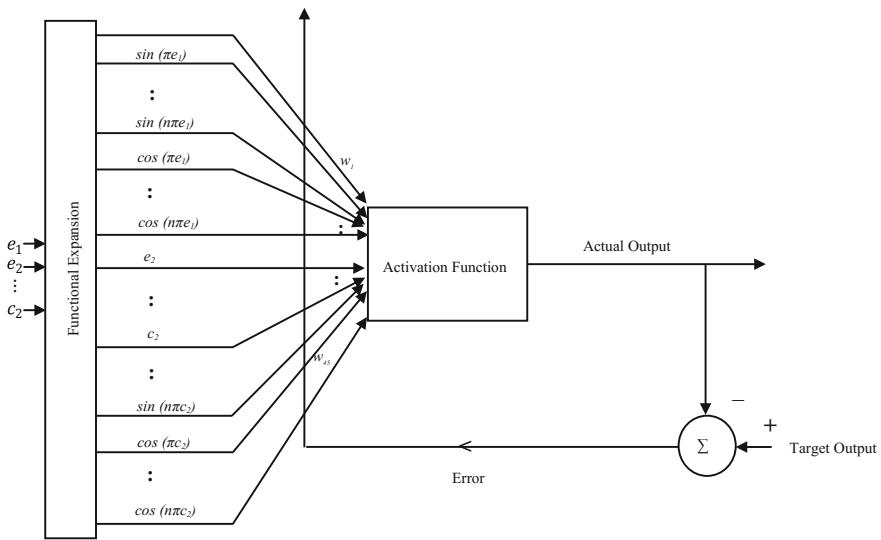
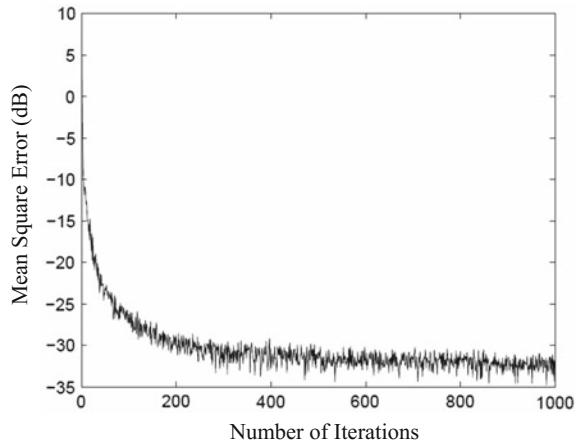


Fig. 4 Functional link artificial neural network architecture for coal characterization

Fig. 5 Training convergence characteristics of FLANN architecture



five parts as we consider $k = 5$ and each category represents approximately same division present in the original group. The three parts considered as training sets and two parts are considered as testing sets among the division of five parts fold. To make the several combinations of training and testing set, the process has been performed for fivefold. Then, the fivefold are taken to be averaged and overall characterization accuracy has been calculated.

3 Results and Analysis

The proposed method has been executed on Intel core i7, 2.40 GHz, 8 GB RAM and Windows 8.1 operating system. The programming simulation has been done on MATLAB 2012 version.

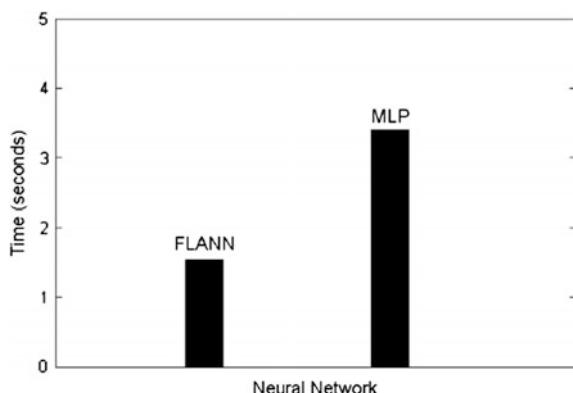
A total of 15 coal subimages that belongs to the testing coal image datasets has been utilized for the preparatory estimation of the proposed technique. Then a total of five textural features have been extracted using GLCM feature extraction method. Features like energy, entropy, contrast, homogeneity, and correlation has been extracted using GLCM method. Although, multilayer perceptron (MLP) is considered as the gold technique for classification of the nonlinear patterns. Instead of this, the use of multiple layers in MLP forms a complex network to train the dataset and consumes more time and memory. To overcome this problem, the proposed FLANN classifier is used to classify the different grades of coal. FLANN is able to solve the nonlinear pattern issue by means of single-layer perceptron. To check the efficiency of the proposed classifier, it has been compared with the characterization accuracy of the standard classifier, i.e., MLP used for the characterization of coal images. Once the coal has been classified accurately, then it is characterized as Category-I, Category-II, and Category-III classes that have been further indexed as best, good, and poor grades, respectively. The maximum characterization accuracy of 91.13 is found with the FLANN classifier after applying the fivefold cross-validation method and considered as a state of art among the two classifiers. Table 1 shows the average characterization accuracy with standard classifier.

It has been noticed that the FLANN takes less time to execute the classification in comparison to MLP. The difference in computation time taken by classifier is shown in Fig. 6.

Table 1 Average characterization accuracy with standard classifier

Neural network	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Characterization accuracy %
MLP	75.33	80.00	86.66	86.66	88.33	83.39
FLANN	86.66	92.33	86.66	92.33	97.67	91.13

Fig. 6 Computation time consumed for classification of coal images



Therefore, the proposed automated technique of coal characterization can be verified as robust and efficient method for the coal quality characterization.

4 Conclusion

The main aim of this paper is the use of functional link artificial neural network (FLANN) for the automated coal characterization. Acceptable results are achieved in comparison to standard method. The proposed method is comes under supervised scheme, still the results are much authenticate and accurate as required for automated characterization of coal. On the basis of these obtained results, future works include the morphological features extraction like edge detection, crack analysis, and porosity analysis using image processing and computational intelligence technique for developing the more robust method for an automated characterization of coal microscopic images.

References

1. Claudio Avila, Tao Wu, Edward Lester: Petrographic characterization of coals as a tool to detect spontaneous combustion potential. Elsevier, Fuel 125 (2014) 173–182.
2. Zelin Zhang, Jianguo Yang, Lihua Ding, Yuemin Zhao: An improved estimation of coal particle mass using image analysis. Elsevier, Powder Technology 229 (2012) 178–184.
3. Zelin Zhang, Jianguo Yang, Lihua Ding, Yuemin Zhao: Estimation of coal particle size distribution by image segmentation. Elsevier, International Journal of Mining Science and Technology 22 (2012) 739–744.
4. Mostafa S. Ibrahim, Amr A. Badr, Mostafa R. Abdallah, Ibrahim F. Eissa: Bounding Box Object Localization Based On Image Superpixelization. Procedia Computer Science 13 (2012) 108 – 119.
5. Maria Petrou, Pedro Gracia Sevilla: Image Processing- Dealing with texture, John Wiley & Sons Ltd, (2006).
6. Isik Yilmaz, Nazan Yalcin Erik and Oguz Kaynar: Different types of learning algorithms of artificial neural network (ANN) models for prediction of gross calorific (GVC) of coals. Scientific Research and Essays Vol. 5(16), (2010) 2242–2249.
7. B.B. Misra, S. Dehuri: Functional Link Artificial Neural Network for Classification Task in Data Mining. Journal of Computer Science 12 (2007) 948–955.
8. S. Mohapatra: Development of Impulsive Noise Detection Schemes for Selective Filtering in Images, Master's Thesis, National Institute of Technology, Rourkela, 2008.
9. R. Kohavi: A study of cross validation and bootstrap for accuracy estimation and model selection. 14th International Joint Conference on Artificial Intelligence 117 (1995) 1137–1143.

Template-Based Automatic High-Speed Relighting of Faces

Ankit Jalan, Mynepalli Siva Chaitanya, Arko Sabui, Abhijeet Singh,
Viswanath Veera and Shankar M. Venkatesan

Abstract This paper presents a framework which estimates the surface normals of human face, surface albedo using an average 3D face template and subsequently replaces the original lighting on the face with a novel lighting in real time. The system uses a facial feature tracking algorithm, which locates and estimates the orientation of face. The 2D facial landmarks thus obtained are used to morph the template model to resemble the input face. The lighting conditions are approximated as a linear combination of Spherical Harmonic bases. A photometric refinement is applied to accurately estimate the surface normal and thus surface albedo. A novel skin-mask construction algorithm is also used to restrict the processing to facial region of the input image. The face is relighted with novel illumination parameters. A novel, cost-effective, seamless blending operation is performed to achieve efficacious and realistic outputs. The process is fully automatic and is executed in real time.

Keywords Albedo • Spherical harmonics • Relighting • Face mask

A. Jalan (✉) · M.S. Chaitanya · A. Sabui · A. Singh · V. Veera · S.M. Venkatesan
Samsung R & D Institute, 2870, Phoenix Building,
Bagmane Constellation Business Park, Bangalore, India
e-mail: ankit.jalan@samsung.com

M.S. Chaitanya
e-mail: chaitanya.15@samsung.com

A. Sabui
e-mail: arko.sabui@samsung.com

A. Singh
e-mail: abhijeet.sh@samsung.com

V. Veera
e-mail: viswanath.v@samsung.com

S.M. Venkatesan
e-mail: s.venkatesan@samsung.com

1 Introduction

Face relighting in real time has a plethora of applications ranging from improvements in the accuracy of face detection and face recognition systems to previews for capturing selfies on mobile applications. With a real-time face relighting application, one could preview a relit face before actually taking a picture, resulting in enhanced user experience. Our face relighting approach could also be used to swap a face from one picture with a face of the same/another person with a different expression captured under different illumination conditions from another picture, or, the illumination itself can be transferred from face in one image to the face in other image. Consequently, this approach could also be extended to relight real or virtual objects of known geometry in a scene. The relighting of virtual objects in a VR scene can assist in generating realistic imagery.

Human face relighting has been an active area for research in the computer vision and graphics community. It presents multiple challenges as one has to cope with the diversity of faces (disparate shape, skin color, etc.), while their fundamental structure is mostly similar. The varying illumination conditions and artifacts like cast shadows, self-shadowing, and occlusions all add to the complexity. When the information regarding the scene is scarce, such as the absence of depth information or the availability of only single input image, this problem becomes particularly difficult. To overcome this problem, our method is supplemented with 3D sparse template of average face [1, 2]. In our framework, the model is warped to resemble the subject's face in order to achieve realistic and natural looking output.

Other approaches typically attempt face relighting by transferring the illumination from one image to another [3], estimate the 3D model of the subject's face using 3D Morphable models [4] or follow a Shape-from-Shading [5] approach and add new illumination condition. Many of these approaches generate good results but are time consuming. With an aim to be pragmatic we implemented a system which performs relighting in real time, while also being fully automatic. Our only assumptions are the availability of a single 3D face template, and that the input image is captured in sufficiently good lighting conditions, i.e., most of the face is within a well-lit environment. Our system achieves natural relighting by improving the surface normals using photometric approach, generating improved albedo and seamlessly blending the relit face-segment with the original image (Fig. 1).

This paper makes the following contributions:

- A novel, real-time template morphing driven by 2D landmarks in face images.
- A method to improve surface normals and albedo of the subject.
- A novel method of generating face-skin-mask.
- A novel seamless blending operation for realistic output.

The paper is organized as follows: Sect. 2 reviews the relevant work in this research area, Sect. 3 describes the procedure to localize face and obtain facial landmarks in real time followed by morphing the model, improving the surface normals and obtaining the albedo. Section 4 includes estimation of surface normal

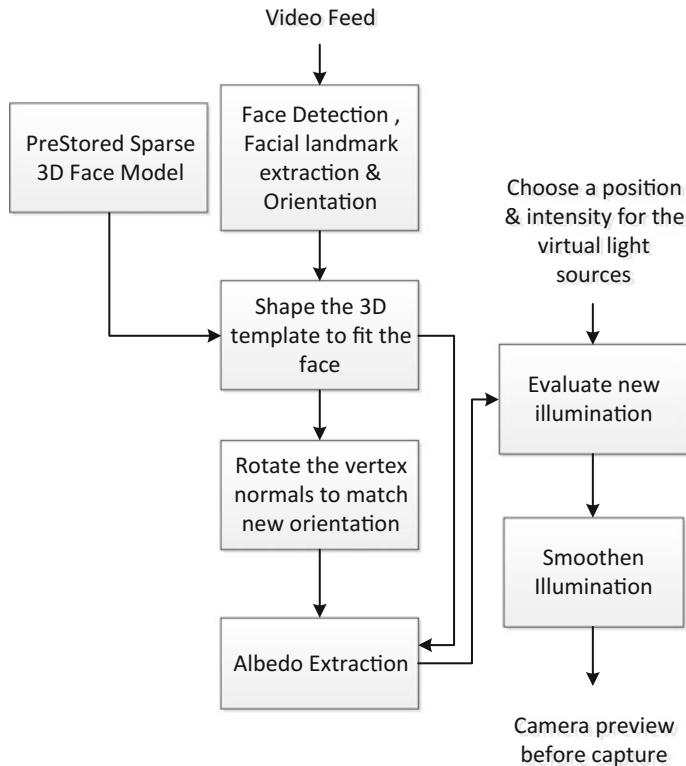


Fig. 1 Key steps of our relighting framework

and albedo, estimation of lighting coefficients. Section 5 includes relighting on various face databases. Section 6 concludes the paper with discussions on obtained results.

2 Related Work

Many approaches to face-shape estimation and face relighting [3, 6–12] lay emphasis on various aspects such as feature tracking, template morphing, acquiring illumination, generating albedo, and rendering the face with novel lighting condition.

Face feature tracking, deals with localizing and estimating 2D landmark points of a face in an image in real time. This paper is built on the work of Kazemi et al. [13], which involves training an ensemble of regression trees to estimate the positions of facial landmarks using pose-indexed features. This work forms the basis of this paper and the rest of the work is driven by it.

Modeling the geometry of face has been an active area of research. The problem is well tackled when it comes to the availability of multiple images. The work of Blanz et al. [4] develops the basis for 3D face morphable model, which is widely used in obtaining the geometry of a face by reducing the disparity between the observed and synthesized data. Our novel implementation of **template morphing** based on scaling the scan-lines differs as it uses only one template and it works very fast, in the order of a few (<10) milliseconds, with an expense of being less accurate but with visibly good results in relighting.

A well-known method of **acquiring illumination** condition based on the linear combination of Spherical Harmonic bases was implemented in the work of Wang et al. [7], and proposed by Ramamoorthi et al. [14]. They use the spatial coherence of face texture to make their algorithm robust toward harsh illumination conditions. Li et al. [3] proposed an illumination transferring technique from one face to another but this work restricts the variability in novel lighting parameters.

Acquiring skin-reflectance or albedo is a crucial step in the whole process of relighting. Removing the effect of original lighting from face image without affecting the texture is a well-researched topic. The work of Biswas et al. [5] estimated the true albedo by developing an image estimation framework. The initialized albedo is expressed as sum of true and unknown albedo and the LMMSE estimate of true albedo is computed. The work of Debevec et al. [15] is based on acquiring the reflectance of face by taking input data under various view-points and lighting conditions using a light stage.

A lot of work has been done to solve the problem of identifying the **face-skin mask**, which is critical for relighting and other beautification processing. [16, 17] is based on classification of skin region if pixel value falls within specified threshold in different color spaces or by training a Gaussian mixture model to classify pixels as skin or nonskin. Main drawback in these color-based methods is not being robust to varying skin-types or facial and optical artifacts like specularity. Our work is more robust and adaptive to skin type based on online clustering and Grab-cut-based foreground extraction algorithm [18].

A major contrast of our work with most of the previous related works with respect to surface normal refinement, i.e., face-shape estimation, albedo extraction, and skin-mask generation is that our method is fully automatic without any manual intervention. With the availability of just a single face image or video frame, our method can perform relighting with the user having full-control over the parameters of new lighting environment.

3 Face Detection and Model Fitting

3.1 Feature Detection

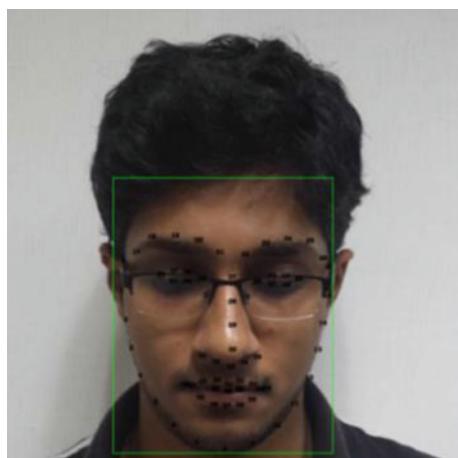
For the facial relighting effects to appear cogent and intuitive, not only does the face need to be detected and localized in an image or video frame, a 3D model template needs to be precisely fit to the facial image as well, with accurate scale and orientation. This is done with the aid of the facial feature extraction algorithms, which typically aim to estimate a high-dimensional vector representing the image-coordinates of a predefined set of facial-features in a consistent order.

For this paper, we have used the algorithm described in [13], which estimates face-shape using an ensemble of regression decision-trees in cascade and uses pose-indexed features efficiently to refine the shape in real time. The 68-points model (Fig. 2.) used was trained on the iBUG 300-W training dataset, and subsequently an n-point correspondence-based pose estimation algorithm [12] was used to estimate the rotation parameters of the face, detected in the image or video frame and these parameters were used to generate the rotated version of representative 3D facial model.

3.2 Face Template Morphing

We morph an average 3D face model using a simple piecewise linear process to approximate the subject's face-shape. An average 3D model is used for initialization because despite being diverse, human faces have the same underlying structure.

Fig. 2 The 68 feature points and the region of interest



In our approach, 68 facial landmarks, obtained by employing the face alignment algorithm discussed in [13], help morph the average 3D face model to represent the features of the subject's face. A correspondence map was built between the 68 feature points and matching points from the template of the average face. The vertices of the average face model were divided into 'Vertical Scan-lines' and 'Horizontal Scan-lines.' Refer to Fig. 3. Shaping the model to approximate the user's face is done in three phases: vertical scaling, horizontal scaling, and using the skin-mask to limit the set of vertices to fall into the facial region.

- **Vertical Scaling:** To morph the average face's point cloud to fit different features along the vertical direction, we subdivided the face vertically into segments. The correspondence map of the facial landmarks with points in the point cloud is used to find the scaling factors for each vertical scan line. The scaling factor, S_j , for the j th vertical scan line, within each segment, is evaluated as shown below.

$$S_j = \frac{d_{j,kazemi}}{d_{j,pc}} \quad (1)$$

$$v'_{i-1,j} \cdot y - v'_{i,j} \cdot y = S_j * (v_{i-1,j} \cdot y - v_{i,j} \cdot y) \quad (2)$$

where v' is the new set of vertices, $d_{j,kazemi}$ and $d_{j,pc}$ are the distance between the facial landmark positions and distance corresponding points in the point cloud, respectively, and $v_{i-1,j}$ and $v_{i,j}$ are neighboring vertices on the same vertical scan-line in the average model point cloud.

- **Horizontal scaling:** The face is subdivided into different horizontal segments to preserve the width of features such as nose. For each scan-line we estimate a target length by finding the corresponding x coordinate for a given y coordinate of the scan-line on both extremes using linear interpolation of the feature points. The scaling factors are then evaluated by dividing the target length with the



Fig. 3 3D template morphing. Facial landmarks, average template, morphed template and profile view of morphed template

distance between corresponding vertices on the face model. Within the nose region, scaling factors are similarly found using the width of the nose as target distance.

To determine the face-skin region, we devised an automatic method to construct a skin-mask, as shown in Fig. 4. Facial landmark points were used to construct a mask in which each pixel is classified as definite foreground, definite background, probable foreground, and probable background. Refer to Fig. 4. Since the feature vector has no data about the forehead, we clustered the skin-pixels to classify the unknown region of forehead as probable foreground or probable background. With this mask, we employed Grab-cut [18] algorithm to generate the skin-mask.

4 Estimating Surface Normal and Albedo

4.1 Lighting Coefficients

Appearance of face image depends on the illumination condition. It is desirable to remove the original illumination from the face to generate its surface albedo. The obtained albedo is used to add new lighting to the face.

We have used an approach similar to Wang et al. [7] to estimate the original illumination condition. Assuming that faces are Lambertian objects, we can approximate the intensity of a pixel in each channel as a linear combination of spherical harmonic bases [7, 14]. It is shown that second-order approximation in spherical harmonics captures a sufficiently high ($\sim 99\%$) portion of the illumination energy [19] in estimating the lighting condition. We obtained the nine coefficients corresponding to zeroth-, first-, and second-order spherical harmonic bases by solving the overdetermined system of linear equation (Eq. 3).

$$I(x, y) = \rho(x, y) * \sum_{k=1}^9 l_k * h_k(n(x, y)) \quad (3)$$

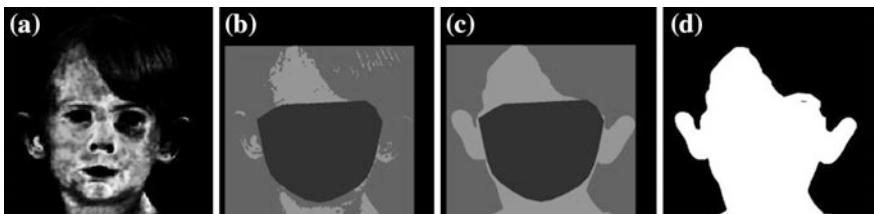


Fig. 4 **a** Gaussian clustering of skin pixel. **b** Interior polygon corresponds to definite foreground; Outer rectangular region corresponds to definite background, further based on clustering, probable foreground, and background respectively. **c** Output of Grab-cut on (b). **d** Generated face mask. Refer to Fig. 3 for original image

where $n(x, y)$ and $\rho(x, y)$ are the surface normal and the albedo, respectively, at (x, y) pixel location in the input image, l_k is the k th coefficient obtained as above and $h_k(n(x, y))$ is the spherical harmonic basis (See Appendix 8, 8.1) corresponding to surface normal $n(x, y)$.

4.2 Surface Normal Refinement

The template model deformation done previously (Sect. 3) provides a good initialization of surface normals and roughly resembles the input face. To obtain an efficacious relighting output, the true surface normals and albedo are required.

We modify the surface normals using photometric refinement algorithm by constructing an energy minimization framework [20].

- **Albedo Initialization:** In order to improve the surface normals, we need a good initialization of albedo for each pixel to solve Eq. 3. The albedo of any natural object comprises of a small color palette and piecewise smooth [21]. Therefore, the albedo of a human face is expected to be fairly constant. As an initialization, the intensity values are averaged over the face region in logarithmic space because of additive dependency in log-space. Using Eq. 3, we get,

$$\log(I(x, y)) = \log(\rho(x, y)) + \log\left(\sum_{k=1}^9 l_k h_k(n(x, y))\right) \quad (4)$$

We obtain the texture information as a ratio image of original input image and a blurred output of the same image. The albedo was initialized as a product of average of pixel intensity and the texture image.

- **Energy Minimization:** The estimated coefficients of spherical harmonic bases and the initialized albedo are used to estimate the surface normal by solving Eq. 3. We employed gradient descent technique with adaptive learning rate to minimize the difference between the observed intensity and the estimated intensity at each pixel.

$$E(x, y) = \sum_{R, G, B} \left((I(x, y) - \rho(x, y) * \sum_{k=1}^9 l_k h_k(n(x, y)))^2 \right) + \lambda_1 \left((\|n(x, y)\|)^2 - 1 \right)^2 + \lambda_2 \sum_{j, k \in \text{neighbour}} \|n(x, y) - n(j, k)\|^2 \quad (5)$$

where the first term in Eq. 5 represents the difference between the observed and the estimated intensity values, the second term encourages normalized solution and the third term ensures continuity among the neighboring pixels. λ_1 and λ_2 are empirical weights kept equal to 1.

- **Albedo Estimation:** The surface albedo component in Eq. 3 can now be obtained using the refined surface normals and the coefficients of spherical harmonics obtained from (Sect. 3.1). Using Eq. 3,

$$\rho(x, y) = \frac{I(x, y)}{\sum_{k=1}^9 l_k h_k(n(x, y))} \quad (6)$$

To improve the results, we can iteratively perform the estimation of coefficients, surface normal refinement, and the albedo estimation. It was observed that subsequent iterations do not improve the result significantly. We restricted ourselves to single iteration.

5 Experiments

5.1 Skin-Mask

To determine the region face-skin mask, we devised an automatic method to construct a skin-mask, as shown in Fig. 4. The 68-feature point vector of facial landmarks was used to construct a mask, in which, each pixel is classified as definite foreground, definite background, probable foreground, and probable background. Since the feature vector has no data about the forehead, we clustered the skin-pixels to classify the unknown region of forehead as a probable foreground or probable background. With this mask, we employed Grab-cut [18] algorithm to generate the skin-mask.

5.2 Relighting

The relighting process requires the albedo, the new illumination intensity and direction of light source (assuming point source at infinite distance) and surface normals of the face. The framework is implemented using OpenGL. The face region is divided into smaller regions called fragments and the relighting is performed on each of these fragments.

Let n_s be the normalized direction of illumination and n_j be the surface normal at an arbitrary point j . The new intensity is calculated by

$$I'(j) = \rho(j) * I_s * \langle n_s \cdot n_j \rangle \quad (7)$$

where $I'(j)$ is the new intensity, I_s is the intensity of novel light. Only those portions of the input image or video frame which fall inside the skin-mask are relit (Fig. 5).

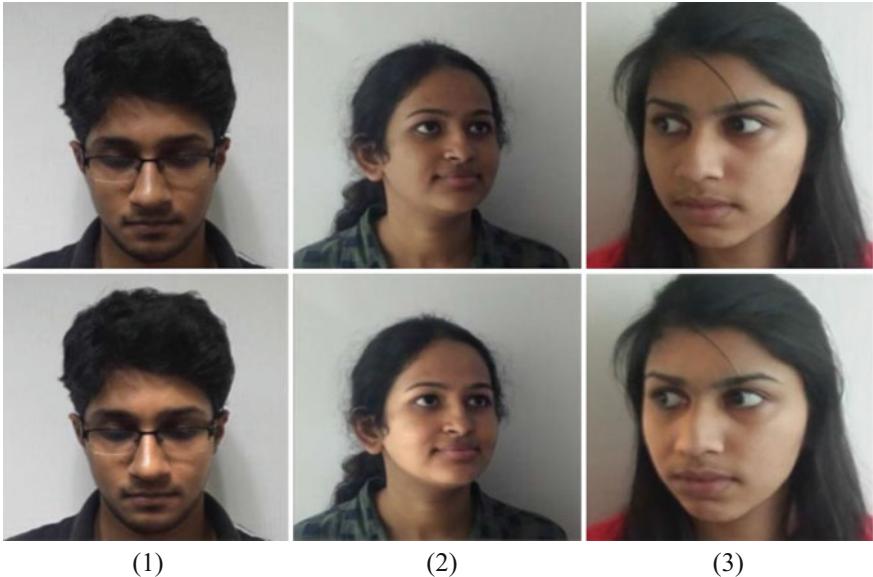


Fig. 5 Relighting still images. *Top row* consists of original input image and the *bottom row* includes relighted image

5.3 Blending

To obtain a good result, it is necessary to match the appearance of face in the image along the model's periphery. Applying a low-pass filter over the difference of relit image and the original image and then adding it back to the original image results in desired outputs while being extremely cost-effective.

$$I_{\text{final}} = I_{\text{original}} + G(I_{\text{relit}} - I_{\text{original}}) \quad (8)$$

where, $G(M)$ represents the Gaussian Blur of matrix, I_{final} , I_{original} , I_{relit} are the final obtained image, original input image, and the relighted image respectively.

6 Results and Conclusion

In Sects. 3 and 4, we had presented our approach and procedure to warp the face template model, estimate the surface normals, albedo, and skin-mask and use these to relight the face image with user having complete control over the illumination direction and intensity. In this section we evaluate the performance of our method.

The output of our method is reliably good and is generated faster than most of the previous works. Our method generates fast and reasonably accurate results even

Table 1 The time taken in milliseconds for the relighting process for different resolutions of input image. This timing was observed with an Intel i7 3630QM @ 2.4 GHz, NVIDIA GeForce GT 650 M GPU. The examples correspond to figures in Fig. 5. (in order)

Resolution	Example 1	Example 2	Example 3
600 × 600	20.96	25.04	26.91
400 × 400	15.52	14.46	14.99
256 × 256	10.78	10.45	10.43



Fig. 6 Albedo estimation

Table 2 The time taken in milliseconds for face-mask modeling

	Example 1	Example 2	Example 3
	6.025	5.69	7.07

when the face orientation is not front-facing. Since our framework achieves fast relighting (Ref. Table 1. For timings) for various poses and illumination conditions, it is feasible for many practical applications on mobile devices like Selfie-previews.

Although in the present work, model fitting is performed on each video frame, we can generate a 3D model specific to a subject's face by accumulating the model warps over a number of frames in different poses thus eliminating the need for fitting the model anymore as long as that particular subject's face is being tracked. We limit the number of iterations for albedo calculation to maximize the speed of the process. Even so, our method generates accurate albedo (Refer Fig. 6). Based on these experiments, we claim that our relighting method works fairly well and in real time for frontal as well as differently oriented faces (Table 2).

Appendix

Spherical Harmonics

Lighting conditions can be approximated by a linear combination of spherical harmonics. It is shown by [19] that with second order assumption in spherical harmonics, 99 % of the energy can be captured. In our method, we use spherical harmonics to estimate the lighting conditions which is further required to estimate the surface normals and the albedo of the face.

$$\begin{aligned}
 h_1(\vec{n}) &= \frac{1}{\sqrt{4\pi}}, & h_2(\vec{n}) &= \frac{2\pi}{3} \sqrt{\frac{3}{4\pi}} * n_z, \\
 h_3(\vec{n}) &= \frac{2\pi}{3} \sqrt{\frac{3}{4\pi}} * n_y, & h_4(\vec{n}) &= \frac{2\pi}{3} \sqrt{\frac{3}{4\pi}} * n_x, \\
 h_5(\vec{n}) &= \frac{\pi}{4} \sqrt{\frac{5}{4\pi}} * (3n_z^2 - 1), & h_6(\vec{n}) &= \frac{3\pi}{4} * \sqrt{\frac{5}{12\pi}} * n_y n_z, \\
 h_7(\vec{n}) &= \frac{3\pi}{4} * \sqrt{\frac{5}{12\pi}} * n_x n_z, & h_8(\vec{n}) &= \frac{3\pi}{4} * \sqrt{\frac{5}{12\pi}} * n_x n_y, \\
 h_9(\vec{n}) &= \frac{3\pi}{8} * \sqrt[3]{\frac{5}{12\pi}} * (n_x^2 - n_y^2)
 \end{aligned} \tag{9}$$

Specularity

Relighting assuming a Lambertian model captures only the diffuse component of illumination. To achieve realistic outputs, it is necessary to incorporate the specular component as well. For this purpose, we use the Torrance–Sparrow model with similar approximations as Shim et al. [9]. Instead of finding the specular coefficient, we stick to our template-based approach and create a generic face specular map using the Skin Reflectance Database of Weyrich et al. [22]. The specular component of illumination is defined in [9].

References

1. Bronstein AM, Bronstein MM, Kimmel R. Numerical geometry of non-rigid shapes. Springer Science & Business Media; 2008 Sep 18.
2. A. M. Bronstein, M. M. Bronstein, R. Kimmel, Calculus of non-rigid surfaces for geometry and texture manipulation, IEEE Trans. Visualization and Computer Graphics, Vol. 13/5, pp. 902–913, 2007.
3. Li, Q., Yin, W., & Deng, Z. (2010). Image-based face illumination transferring using logarithmic total variation models. The visual computer, 26(1), 41–49.
4. Blanz, V., & Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (pp. 187–194). ACM Press/Addison-Wesley Publishing Co.

5. Biswas, S., Aggarwal, G., & Chellappa, R. (2009). Robust estimation of albedo for illumination-invariant matching and shape recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5), 884–899.
6. Liu, Z., Shan, Y., Zhang, Z.: Expressive expression mapping with ratio images. In: SIGGRAPH'01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 271–276. ACM, New York (2001).
7. Wang, Y., Zhang, L., Liu, Z., Hua, G., Wen, Z., Zhang, Z., & Samaras, D. (2009). Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11), 1968–1984.
8. H. Shim, J. Luo and T. Chen, A Subspace Model-Based Approach to Face Relighting Under Unknown Lighting and Poses, IEEE TIP, pp. 1331–1341, 2008.
9. Shim H. Probabilistic Approach to Realistic Face Synthesis with a Single Uncalibrated Image. *Image Processing, IEEE Transactions on*. 2012 Aug;21(8):3784–93.
10. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. (2008). Face swapping: automatically replacing faces in photographs. *ACM Transactions on Graphics (TOG)*, 27(3), 39.
11. Paris, S., Sillion, F. X., & Quan, L. (2003, October). Lightweight face relighting. In Computer Graphics and Applications, 2003. Proceedings. 11th Pacific Conference on (pp. 41–50). IEEE.
12. Lepetit, V., Moreno-Noguer, F. and Fua, P., 2009. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2), pp. 155–166.
13. Kazemi, V., & Sullivan, J. (2014, June). One millisecond face alignment with an ensemble of regression trees. In 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 1867–1874). IEEE.
14. Ramamoorthi, Ravi, and Pat Hanrahan. An efficient representation for irradiance environment maps. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001.
15. Debevec, Paul, et al. Acquiring the reflectance field of a human face. *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000.
16. Phung, Son Lam, Abdesselam Bouzerdoum, and Douglas Chai. Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27.1 (2005): 148–154.
17. Phung, Son Lam, Abdesselam Bouzerdoum, and Douglas Chai. A novel skin color model in ycbcr color space and its application to human face detection. *Image Processing. 2002. Proceedings. 2002 International Conference on*. Vol. 1. IEEE, 2002.
18. Rother, Carsten, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)* 23.3 (2004): 309–314.
19. Basri R, Jacobs DW. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2003 Feb; 25(2):218–33.
20. Johnson, Micah K., and Edward H. Adelson. Shape estimation in natural illumination. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.
21. Barron, Jonathan T., and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012.
22. Weyrich, Tim, et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (TOG)*. Vol. 25. No. 3. ACM, 2006.

An Improved Contextual Information Based Approach for Anomaly Detection via Adaptive Inference for Surveillance Application

T.J. Narendra Rao, G.N. Girish and Jeny Rajan

Abstract Anomalous event detection is the foremost objective of a visual surveillance system. Using contextual information and probabilistic inference mechanisms is a recent trend in this direction. The proposed method is an improved version of the Spatio-Temporal Compositions (STC) concept, introduced earlier. Specific modifications are applied to STC method to reduce time complexity and improve the performance. The non-overlapping volume and ensemble formation employed reduce the iterations in codebook construction and probabilistic modeling steps. A simpler procedure for codebook construction has been proposed. A non-parametric probabilistic model and adaptive inference mechanisms to avoid the use of a single experimental threshold value are the other contributions. An additional feature such as event-driven high-resolution localization of unusual events is incorporated to aid in surveillance application. The proposed method produced promising results when compared to STC and other state-of-the-art approaches when experimented on seven standard datasets with simple/complex actions, in non-crowded/crowded environments.

Keywords Visual surveillance • Anomalous event detection • Spatio-temporal volume • Video processing • Bag of words • Object detection • Adaptive inference

T.J.N. Rao (✉) · G.N. Girish · J. Rajan

Department of Computer Science and Engineering, National Institute of Technology
Karnataka, Surathkal, Mangalore 575025, Karnataka, India
e-mail: raonaren25@gmail.com

G.N. Girish
e-mail: girishanit@gmail.com

J. Rajan
e-mail: jenyrajan@gmail.com

1 Introduction

In the last few years, there has been an accelerated expansion in the use of surveillance cameras, with a purpose of detecting crimes and to act as repellents for future threats. The reason behind this mass deployment is the steep rise in the crime rate, compromising the safety and security of the society in the present day global scenario.

An automatic visual surveillance system has the primary task of detecting an event which is anomalous in a particular context, avoiding the possible miss detections in a manual model [1]. Anomaly by definition is anything that departs from the normal, i.e., an anomalous event occurs rarely among the normal events. In other words such an event can be detected based on its very low probability of occurrence *in a particular context*. Context is crucial in decision-making because an activity which is unusual in one context may be normal in some other [2].

Adam et al. [3] developed an approach to capture motion features by several local monitors based on optical flow directions. If a measurement is unusual, the local monitor produces an alert and these alerts are combined to a final decision. The method has not considered contextual information for detection and suffers from the drawback that it cannot detect events as unusual which are composed of unusual sequences of short-term normal actions. The Mixture of Dynamic Textures (MDT) method [4] uses joint modeling of the appearance and dynamics of the scene along with the ability to detect temporal and spatial unusualness to determine anomalies in a crowded scene. The drawback is that the method is computationally intensive. The detection of anomalies in videos was addressed by Boiman et al. as a problem of detecting irregularities using Inference by Composition (IBC) method [5]. Contextual information around an event was used in this method for unusual event detection. The method tried reconstructing the query from the database using spatio-temporal patches. Because the method used all spatio-temporal patches to reconstruct, it is computationally intensive. This drawback of [5] was overcome in STC approach [2] of Roshthakhari et al. The method grouped similar volumes into a codebook and used probability framework for the inference mechanism. The use of codewords and their probability distribution functions instead of all the volumes reduces the time complexity by a great extent. The method is real-time, has unsupervised learning ability, and outperforms the state-of-the-art methods found in the literature. However, the method has to deal with a very large number of overlapping volumes. Also, a single experimental threshold is used for inference. It has been suggested that the performance of the method could be improved with an adaptive threshold. Inspired by this method, a similar probability-based anomalous event detection approach using contextual information is adopted in this paper. An attempt is made in the proposed work to develop an adaptive inference mechanism while further simplifying the method of [2], as will be discussed in the subsequent sections.

Further, video data captured from public spaces can often be ill-conditioned due to poor resolution [1]. But capturing frames in high resolution and their transmission consumes more energy and bandwidth. It is a fact that the lifetime of battery-operated camera nodes is limited by their energy consumption [6]. Hence,

this paper proposes an event-driven enhancement of only the frames containing the detected anomalous events.

Overall, the characteristic features and contributions of the proposed method include

- Construction of volumes and formation of ensembles in a *non-overlapped* manner at various scales. This is to simplify the codebook construction process and to reduce the complexity of the probabilistic modeling process.
- A new codebook construction process is proposed which is less complex and proved to be reliable as seen from the anomalous events detection rate in experiments section.
- Applying non-parametric approach to obtain the true distribution of codewords in the ensembles during probabilistic modeling.
- Proposal of three context-based adaptive inference mechanisms to detect the anomalous events in the query videos by adaptively determining the decision-making threshold.

The rest of the paper is organized as follows: Sect. 2 explains the proposed approach for anomalous event detection and localization. Section 3 shows the experimental results on the different datasets and finally Sect. 4 concludes the paper.

2 Proposed Method

The proposed method in principle is based on the method of STC [2]. However, the implementation of the concept has been varied, as will be explained in the following subsections. It is a fact that an anomalous event is considered as a rare event having very low likelihood of occurrence, among the majority of normal events. As mentioned earlier, the context of an event is also a deciding factor in identifying an anomalous event. Hence, the spatio-temporal region around an event (an ensemble) is used to obtain the required contextual information. For an observed query video, the probability of each ensemble composition is compared with those of the training sequence. If the probability of an ensemble of the query is less than a threshold, then that ensemble is inferred to have the *event(s) of interest*. The unusual event detection and localization process presented in this paper is carried out in two stages.

2.1 Initialization with a Training Video Sequence

A short video sequence describing the given context, including the normal activities involved in it is required for initialization of the proposed method. This initial training of the surveillance system is performed through the following three steps.

2.1.1 Formation of Spatio-Temporal Blocks and Construction of Volumes

The training video containing normal events is initially downsampled, and then divided into blocks called ensembles of equal size in both spatial and temporal domain. The downsampling of the video is done to reduce the number of volumes that can possibly be constructed, with minimal loss of information as seen from the experimental results in Sect. 3.1. These ensembles give the contextual information aiding in the anomaly detection. Each ensemble is then sampled to generate a set of non-overlapping spatio-temporal volumes at different spatial and temporal scales, for example, of size and in the increments of 10, up to the ensemble size as shown in Fig. 1. The algorithm for the same is given in Algorithm 1. The non-overlapping volume construction at each scale generates one volume for every 10 pixels approximately, thereby reducing many iterations when compared to volume construction around every pixel. It is known that human activities and any other natural spatial structures are not generally reproducible [2, 5]. Therefore, these local small variations in the activities in space and time are not to be missed. This is achieved by sampling at different scales. The proposed method differs from [2] where construction of volumes and ensembles is done around every pixel.

Since we need to know the activity(s) contained in the volumes, a descriptor describing the temporal changes (f_z) is defined for each volume v_i and the volumes

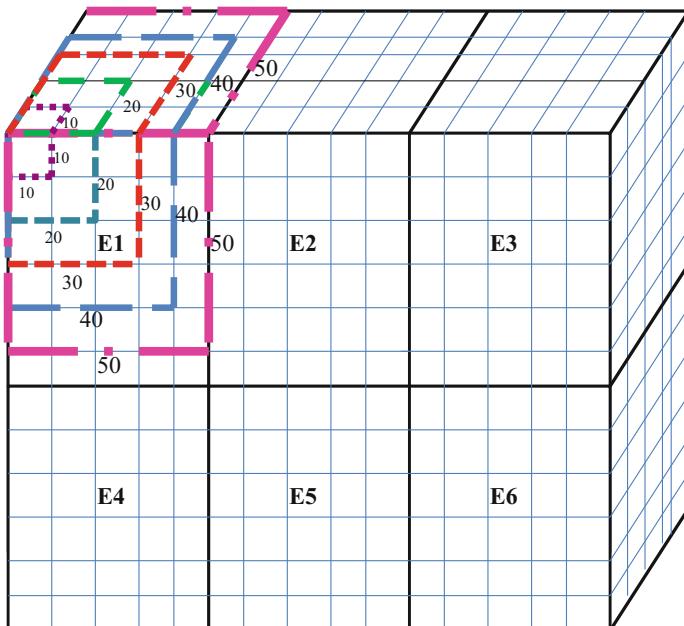


Fig. 1 Pictorial representation of ensembles (E1, E2, ..., E6) and volumes formed at different spatial and temporal scales within E1

are represented by their descriptors throughout the rest of the process. It is determined as the gradient ∇ along the temporal axis and is given by, [2]

$$f_z = \text{abs}(\nabla(v_i)), \text{ for all } v_i. \quad (1)$$

This captures all the changes that have occurred in the scene during the course of time. To reduce the computational complexity of the overall process, all these descriptors of volumes at various scales are converted into vectors, normalized to unit length, and stacked together to obtain a compact form.

2.1.2 Codebook Construction and Codeword Assignment to Volumes

The spatio-temporal volume construction process in the earlier step generates large number of volumes. For example, dense sampling of a one minute video with a frame rate of 30 frames/sec generates 4×10^4 volumes approximately. Correspondingly, storing them requires large memory space. Since the volumes are constructed in a non-overlapping manner, they are rarely identical, complicating the probability calculation of the arrangement of volumes, in the next step. However, since many volumes may be similar, grouping them reduces their number without considerable loss of information, while enabling the computation of probability. The single combined volume known as a codeword represents all the volumes of a group. All the codewords together form a codebook.

The new codebook construction procedure proposed is much simpler than the codebook procedure in [2] and also the pruning process of codewords is avoided here. The Euclidean distance is used as a measure of similarity to group the volumes. The volumes having the distance lesser than a threshold ϵ are grouped together and a codeword is formed. It is to be noted that the number of codewords formed is much less than the number of volumes created.

The similarity weight [2] between each volume and each of the codewords is then computed. The codeword with the highest degree of similarity for a volume is then assigned to it as shown in Fig. 2. The proposed codebook construction and codeword assignment algorithm is explained in Algorithm 1.

2.1.3 Non-parametric Probabilistic Modeling of Volume Configuration in an Ensemble

To understand the normal behaviour occurring in the context found in the training sequence, contextual information present in each ensemble is to be modeled. A probabilistic framework to model the spatio-temporal arrangement of volumes in an ensemble has been introduced in [2]. Applying probability theory to the configuration of volumes gives the probability of occurrence of that normal event arrangement. Such probability values of all the ensembles of the training sequence describe the likelihood of occurrence of normal events. An unusual ensemble will

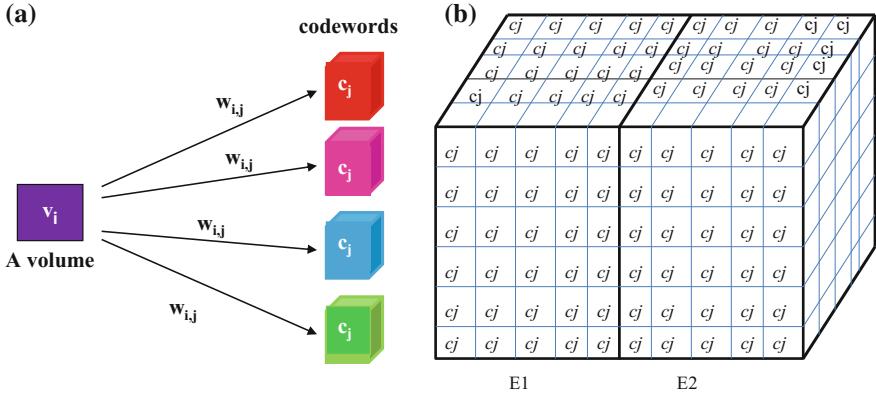


Fig. 2 **a** Labeling of each codeword c_j to each volume v_i with their similarity weight $w_{i,j}$. **b** Representation of the ensembles after the assignment of a codeword c_j with the highest weight to each of its volume v_i

be the one having very low probability for its composition. In [2] the arrangement of volumes in an ensemble has been considered relative to its central volume using the co-ordinate distances of volumes and relative position of codewords (representing the volumes) in codebook space. In the proposed method, it is assumed that every volume in an ensemble has a unique position as shown in Fig. 3a and these position numbers instead of coordinate differences are utilized to know the arrangement of volumes.

Since each volume is represented by a codeword assigned during codeword assignment procedure, every ensemble can be represented by a group of codewords. Let c be the codeword assigned to any video volume v_k and c' is the codeword for the central volume v_i and p be the position of the volume v_k . Thus, probability modeling of the volume arrangement becomes modeling of codeword arrangement around a central codeword. This probability of finding (c, c') given the observed video volumes (v_k, v_i, p) is termed as volume posterior probability [2] and is calculated using conditional probability principle as,

$$P(c, c' | v_k, v_i, p) = P(c|c', p) P(c|v_k) P(c' | v_i). \quad (2)$$

A non-parametric approach is used to find the first term of Eq. (2). It is obtained by considering all the ensembles belonging to the training sequence and finding the probability of observing the codeword c relative to the central codeword c' at every position p for an ensemble, as depicted in Fig. 3b. It is a non-parametric approach to estimate the true distributions, whereas in [2] this estimation is done parametrically using a mixture of Gaussians. The second and third terms of Eq. (2) are calculated by applying Baye's theorem of conditional probability. The product of posterior probabilities of all the volumes in an ensemble gives the posterior probability of that ensemble. The posterior probabilities of all the ensembles of the training sequence forms the database P_{ET} .

Initial Training

Let n be the number of frames of usual events for training.

Ensembles and spatio-temporal volumes construction:

- Downsample the n frames.
- Let $E^T = \{E_1^T, E_2^T, \dots, E_i^T\}$ be a set of ensembles formed, where $i=(n*6)/50$ (i.e., 6 ensembles per 50 frames).
- Divide each ensemble to obtain volumes $V = \{v_1, v_2, \dots, v_k\}$ of size $10 \times 10 \times 10$, up to $50 \times 50 \times 50$ making a total of N volumes for the training video.
- Smooth each volume with a Gaussian filter and obtain its gradient descriptor.

Codebook construction and assignment:

Let N be the number of volumes represented by their descriptor vectors, $V = \{v_i\}_{i=1}^N$

$\text{codes_set} = V$

$j=0$

while codes_set is not empty
for all volumes in codes_set
 if $\text{euclidean_distance}(v_i, v_j) < \epsilon$
 move v_i from codes_set to set_{j+1}
 end if
end for

$j=j+1$;

create new codeword, $c_j = \text{mean}(\text{set}_j)$
end while

Let M be the number of codewords in $C = \{c_j\}_{j=1}^M$

for all volumes in V

 for all codewords C
 compute the weight $w_{i,j}$
 end for

 assign c_j corresponding to $\max(w_{i,j})$ to v_i
end for

Training sequence probability database construction:

- Calculate volume posterior probability $P(c, c' | v_k, v_i, p)$ for each volume v_i , $i=1$ to k in each ensemble $E_i^T \in E^T$.
- Calculate ensemble posterior probabilities, $P_{E^T} = \{P_{E_1^T}, P_{E_2^T}, \dots, P_{E_i^T}\}$ of each ensemble $E_i^T \in E^T$.

where,
 ϵ = mean of the distances of all the volumes relative to v_i in codes_set

$$w_{i,j} = \frac{1}{\sum_j \frac{1}{\text{distance}(v_i, c_j)}} * \frac{1}{\text{distance}(v_i, c_j)}$$

Alg. 1. Algorithm for initialization, given a training video sequence.

2.2 Anomalous Events Detection in a Query Video

2.2.1 Event-Driven Processing

The query video sequence considered is sampled similar to training video to construct ensembles and volumes. If there exists a non-zero gradient descriptor, only then the codeword assignment and further processing is carried out. Else, it is ignored in order to avoid the unnecessary wastage of processing power. This event-driven processing of a query video is explained in Algorithm 2. At this point, the proposed method is modified to achieve the unsupervised learning of new normal events. For this purpose, the frequency of the codewords in different positions of the ensembles in the training is incremented, updated, and used to obtain the values of first term of Eq. (2) when a query is processed and the frequency values are carried forward for the next query, in order to learn. This feature is beneficial in contexts where the normal events pattern changes more often.

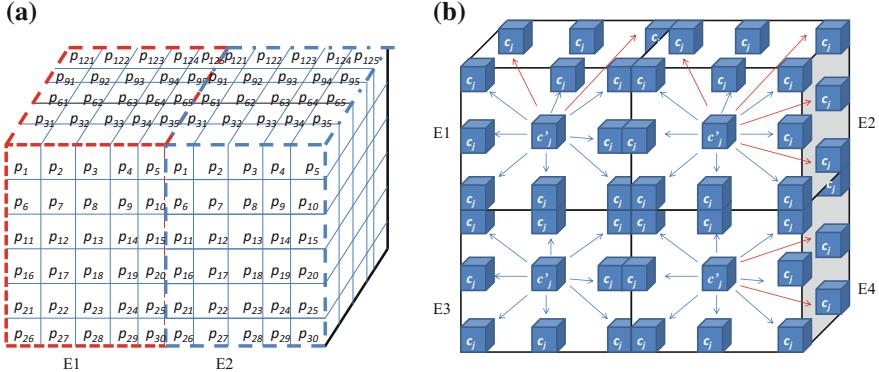


Fig. 3 Pictorial representation of the ensembles and the volumes within them. **a** Every volume in an ensemble has a unique position p_i and these position numbers are utilized to know the arrangement of volumes. **b** The arrangement of volumes is used to find the probability of observing the codeword c_j relative to the central codeword c'_j , given p for an ensemble, considering all the ensembles (E1, E2, E3, E4 in this representation)

2.2.2 Inference Mechanisms for Decision-Making

The decision-making regarding the presence of unusual events requires a threshold to be fixed. To avoid the use of a single experimentally obtained threshold value as in [2], context-based adaptive inference mechanisms have been developed in this work. Based on the normally encountered contexts under surveillance, three mechanisms suitable to those contexts have been proposed. Being adaptive gives these mechanisms the edge over single-valued inference mechanism as there is no guarantee that the single value will generate the expected detection rate for all contexts. This is because the probability values of both the training and query sequences change from one context to another and hence it may be expected that the decision-making threshold value also changes accordingly. Following are the inference mechanisms developed:

Inference Mechanism for Crowded/Non-crowded Real-World Contexts (IM 1). In a real-world context, it may be expected that an unusual event occurs rarely for a short duration amidst the normal events. *IM 1* is developed based on the fact that for the training sequence, the range of probabilities P_{ET} for the different ensembles obtained correspond to the normal events. Given a query, the probabilities of ensembles containing normal activities in it will be higher than the minimum value of P_{ET} . Hence, the minimum ensemble posterior probability of the training sequence added with a small constant deviation α is chosen as the decision-making threshold (T). Any ensemble posterior probability of the query less than this threshold is inferred to have anomalous events in it.

Inference Mechanism for Non-crowded Contexts Consisting of a Definite Activity Pattern (IM 2). The *IM 2* is applicable for a context where the objective of surveillance is to detect the activity in the query which deviates from that of the

training, e.g., when boxing activity occurs in place of usual walking activity, showing a pattern change. Thus, if the pattern of the ensemble posterior probabilities corresponding to the activity in a query does not match with that of the normal activity in the training, then the mismatched ensembles are anomalous.

Inference Mechanism for Contexts Involving Object Detection (IM 3). The similarity between the training and a usual sequence is computed using Euclidian distance between their ensemble probabilities P_{E^T} and P_{E^U} respectively. This forms the decision-making threshold (T). If the similarity between the training and any query is found to be less than the T (or distance greater than T), then the query is considered to have unusual event (object of interest). The ensemble of this query with minimum probability contains the anomalous object. This mechanism is found to be accurately working for the contexts having dynamic background (e.g., background of river).

The Algorithm 2 provides the anomalous event detection procedure.

2.2.3 Event-Driven High-Resolution Localization

Once any ensemble is inferred to be containing anomalous event(s), the frames constituting that ensemble are subjected to enhancement. This event-driven high-resolution localization enhances the visibility of the anomalous activity(s), thus aiding in surveillance application.

3 Experiments

The proposed method for unusual event detection was tested on seven different standard datasets namely, the *Subway Surveillance* dataset,¹ *Anomalous Walking Patterns* dataset [5],² the *UT-Interaction* dataset³ [7], the *MSR Action Dataset II*,⁴ *KTH* dataset [8],⁵ the *Weizmann* dataset,⁶ the *Spatio-temporal Anomaly Detection* dataset⁷ (*Train*, *Boat-river*, and *Canoe* video sequences) [9] with simple and complex activities, in both non-crowded and crowded environments and under varied illumination conditions.

¹from the authors of [3].

²<http://www.wisdom.weizmann.ac.il/~vision/Irregularities.html>.

³http://www.cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

⁴<http://research.microsoft.com/en-us/um/people/zliu/actionrecsrc/>.

⁵<http://www.nada.kth.se/cvap/actions/>.

⁶<http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html>.

⁷<http://www.cse.yorku.ca/vision/research/spatiotemporal-anomalous-behavior.shtml>.

Processing of a query video					
<ul style="list-style-type: none"> From the ensembles $E^Q = \{E_1^Q, E_2^Q, \dots, E_i^Q\}$, construct volumes $V = \{q_1, q_2, \dots, q_k\}$ and obtain their gradient descriptors as before. If there exists a non-zero descriptor, assign a codeword to each volume as before, from the codebook formed earlier for training. Compute the posterior probability of each ensemble $P_{EQ} = \{P_{E_1^Q}, P_{E_2^Q}, \dots, P_{E_i^Q}\}$ as in Section 2.1.3 					
Inference mechanisms:					
<table border="1"> <tr> <td style="padding: 5px;"><i>IM 1:</i> $T = \min(P_{ER}) + \alpha$ for all values in P_{EQ} if $P_{E_i^Q} < T$ $P_{E_i^Q}$ is anomalous end if end for</td><td style="padding: 5px;"><i>IM 2:</i> for all values in P_{EQ} if the pattern P_{EQ} ~ pattern of P_{ER} $P_{E_i^Q}$ deviating from pattern of P_{ER} is anomalous end if end for</td><td style="padding: 5px;"><i>IM 3:</i> $T = \text{distance}(P_{ER}, P_{EV})$ if $\text{distance}(P_{ER}, P_{EQ}) > T$ $\min(P_{EQ})$ is anomalous end if</td></tr> </table>			<i>IM 1:</i> $T = \min(P_{ER}) + \alpha$ for all values in P_{EQ} if $P_{E_i^Q} < T$ $P_{E_i^Q}$ is anomalous end if end for	<i>IM 2:</i> for all values in P_{EQ} if the pattern P_{EQ} ~ pattern of P_{ER} $P_{E_i^Q}$ deviating from pattern of P_{ER} is anomalous end if end for	<i>IM 3:</i> $T = \text{distance}(P_{ER}, P_{EV})$ if $\text{distance}(P_{ER}, P_{EQ}) > T$ $\min(P_{EQ})$ is anomalous end if
<i>IM 1:</i> $T = \min(P_{ER}) + \alpha$ for all values in P_{EQ} if $P_{E_i^Q} < T$ $P_{E_i^Q}$ is anomalous end if end for	<i>IM 2:</i> for all values in P_{EQ} if the pattern P_{EQ} ~ pattern of P_{ER} $P_{E_i^Q}$ deviating from pattern of P_{ER} is anomalous end if end for	<i>IM 3:</i> $T = \text{distance}(P_{ER}, P_{EV})$ if $\text{distance}(P_{ER}, P_{EQ}) > T$ $\min(P_{EQ})$ is anomalous end if			

Alg. 2. Proposed anomalous event detection procedure, given a query.

In UT-Interaction, MSR II, KTH, and Weizmann datasets, those activities which occur frequently or which appear normal in that context are classified to be usual events and those which deviate from these are considered to be unusual, as Ground Truth (GT). For the rest, the unusual events ground truth have been provided (refer Table 1 for the activity details). The smallest possible sequence containing the normal activities were chosen for initial training from all the above datasets. The training and query videos are downsampled to a dimension of 120×160 before processing, and then the ensembles of size $60 \times 53 \times 50$ are formed during experiments.

3.1 Performance Analysis

The experimental results given in Table 2 show the performance of the proposed method on the above datasets. It is to be noted that the training set used here for initialization for each of the datasets consists of a very small number of frames (relative to the length of the entire video) ranging from 50 to 600 frames so as to contain the desired length of normal actions.

Overall results show that the proposed method exhibits promising results, with an unusual event detection rate ((True positives/number of unusual queries tested) $\times 100$) of 91.35 % across all datasets, with only six false alarms. The likely reasons for the few missed detections observed could be due to the lack of noticeable movements by the participants at a distance from camera, poor illumination causing least gradient changes, and the close similarity of the events in queries to those of training. The results show that the inference mechanisms proposed worked as desired depending on the contexts.

Figure 4a shows the Receiver Operating Characteristics (ROC) curve of the proposed method for the *Anomalous Walking Patterns* dataset. It shows higher True Positive Rate at a lower False Positive Rate at the threshold obtained through *IM 1*,

Table 1 Details of the activities present in the datasets and the inference mechanism applied based on the context

Datasets	Activities classified as usual	Activities classified as anomalous	Inference mechanism applied
Subway Surveillance dataset (Exit video)	People exiting from the platform	Entering through the exit gate, moving in the wrong directions, loitering	<i>IM 1</i>
Anomalous Walking Patterns dataset	Walking	Stealth walking, crawling, falling, jumping over obstacles	<i>IM 1</i>
UT-Interaction dataset	Persons in the scene having no physical interactions	Punching, kicking, pushing	<i>IM 1</i>
MSR II dataset	Normal actions in a dining area	Hand waving, clapping, punching	<i>IM 1</i>
KTH dataset	Walking	Boxing, hand waving, clapping, running	<i>IM 2</i>
Weizmann dataset	Walking	Bending, jacking, jumping, hopping, sideways walking, skipping, one and two hands waving	<i>IM 2</i>
Spatio-temporal anomaly detection dataset (Train, Boat-river and Canoe videos)	People sitting inside the train cabin, no boat in the river	People entering and leaving the train cabin and, boat appearing in the scene	<i>IM 1</i> for train video sequence, <i>IM 3</i> for others

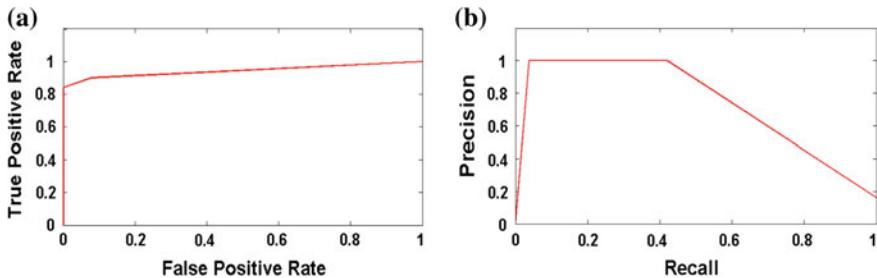
compared to the ROC of the state-of-the-art STC method [2]. Figure 4b shows the Precision-Recall (PR) curve of the proposed method for the *Train video* of *Spatiotemporal Anomaly Detection* dataset. It achieves higher Precision at a higher Recall at the threshold obtained through *IM 1*, compared to PR curves of STC [2] and Spatio-temporal Oriented Energy Filters [9]. Also, the detection rate of the proposed method in case of *Subway Surveillance* (Exit video) dataset is comparable to those of other methods [2, 3, 10, 11], as shown in Table 3.

As mentioned in [2], the method generates a huge number of volumes (10^6 volumes for 1 minute video) as a result of dense sampling around every pixel. If V is the total number of volumes generated for a video sequence, then the time complexity of the codebook construction is $O(V)$. If K is the number of volumes in an ensemble, and M is the number of codewords, the complexity of codeword assignment is $O(K \times M)$ and for the posterior probability calculation, it is $O(K \times M)$.

The proposed method generates relatively a very less number of volumes due to the non-overlapping (not around every pixel) volume construction procedure. For a 1 minute video with the frame dimension of 120×160 , only around 4×10^4 volumes are produced. Hence, codebook construction here requires far less iterations. The same is the case for codeword assignment as well. Though the

Table 2 Details of the experimental results for the different datasets using the proposed method

Datasets	No. of training frames	No. of unusual queries tested	True positives (False positives)	Detection rate
Subway Surveillance dataset (exit video)	600	29	27 (2)	27/29 (93.1 %)
Anomalous Walking Patterns dataset	100	08	08 (0)	8/8 (100 %)
UT-Interaction dataset	100	21	19 (0)	19/21 (90.47)
MSR II dataset	50	16	16 (2)	16/16 (100 %)
KTH dataset	100	80	77 (1)	77/80 (96.25 %)
Weizmann dataset	50	18	12 (0)	12/18 (66.6)
Spatio-temporal anomaly detection dataset	Train video	100	07	05 (0)
	Boat-river video	50	03	02 (1)
	Canoe video	100	03	03 (0)
Total		185	169(6)	169/185 (91.35 %)

**Fig. 4** **a** ROC curve of the proposed method for *Anomalous Walking Patterns* dataset. **b** PR curve of the proposed method for the *Train sequence*

complexity of probability calculation in this method is $O(K \times M \times K)$, it still requires much fewer iterations because of smaller K and M , leading to a reduction in the computational complexity when compared to that of [2].

Notwithstanding the involvement of non-overlapped volumes for anomalous event detection, the observed result may be possibly attributed to the use of non-parametric method to obtain the true distribution of the codewords in the ensembles, to the probability calculation for every volume of the ensembles and to

Table 3 Comparison of results of the different methods (results taken from corresponding references) for *Subway Surveillance* (exit video)

Dataset	Method	Unusual events detected	False positives
Subway Surveillance dataset (Exit video)	STC [2]	19/19	02
	ST-MRF [10]	19/19	03
	Dynamic Sparse Coding [11]	19/19	02
	Local Optical Flow [3]	09/09	02
	Proposed method	27/29	02

**Fig. 5** Sample snapshots of the detected anomalous events in the different dataset videos

the decision-making threshold derived from the training probability database. Figure 5 shows sample detections of anomalous events in experimented datasets by the proposed method.

4 Conclusion

This paper presents an improved approach for anomalous event detection based on the principle of STC method. An attempt has been made to reduce the complexity and improve the performance. The encouraging results observed justify the specific deviations incorporated. Avoiding the overlapping volumes and ensembles around every pixel aids in reducing the complexity of codebook construction, codeword assignment, and probabilistic modeling processes. The simple yet reliable codebook construction procedure proposed further simplifies the overall processing. The non-parametric probability estimation involving every volume may also have contributed toward obtaining encouraging results. The adaptive inference mechanisms proposed in this paper have proven to be effective in anomaly detection. The proposed method produced appreciable results in comparison to STC and other state-of-the-art methods.

Developing a single unified adaptive inference mechanism applicable to any given context, employing partially overlapped volumes/ensembles and using a more complex descriptor can further improve the performance and robustness of the method.

References

1. Gong, S., Loy, C.C., Xiang, T.: Security and Surveillance. In: Visual Analysis of Humans, pp. 455–472. Springer (2011).
2. Roshtkhari, M.J., Levine, M.D.: An On-line, Real-Time Learning Method for Detecting Anomalies in Videos Using Spatio-Temporal Compositions. Computer vision and Image Understanding 117(10), 1436–1452 (2013).
3. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust Real-Time Unusual Event Detection Using Multiple Fixed-Location Monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(3), 555–560 (2008).
4. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly Detection in Crowded Scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981. IEEE (2010).
5. Boiman, O., Irani, M.: Detecting Irregularities in Images and in Video. International Journal of Computer Vision 74(1), 17–31 (2007).
6. Marcus, A., Marques, O.: An Eye on Visual Sensor Networks. Potentials, IEEE 31(2), 38–43 (2012).
7. Ryoo, M.S., Aggarwal, J.: UT-Interaction Dataset, ICPR Contest on Semantic Description of Human Activities (SDHA). In: IEEE International Conference on Pattern Recognition Workshops. vol. 2, p. 4 (2010).
8. Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 32–36). IEEE.

9. Zaharescu, A., Wildes, R.: Anomalous Behaviour Detection Using Spatiotemporal Oriented Energies, Subset Inclusion Histogram Comparison and Event-Driven Processing. In: Computer Vision–ECCV 2010, pp. 563–576. Springer (2010).
10. Kim, J., Grauman, K.: Observe Locally, Infer Globally: A Space-Time MRF for Detecting Abnormal Activities with Incremental Updates. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. pp. 2921–2928. IEEE (2009).
11. Zhao, B., Fei-Fei, L., Xing, E.P.: Online Detection of Unusual Events in Videos via Dynamic Sparse Coding. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3313–3320. IEEE (2011).

A Novel Approach of an (n, n) Multi-Secret Image Sharing Scheme Using Additive Modulo

Maroti Deshmukh, Neeta Nain and Mushtaq Ahmed

Abstract Secret sharing scheme (SSS) is an efficient method of transmitting one or more secret images securely. The conventional SSS share one secret image to n participants. With the advancement of time, there arises a need for sharing multiple secret image. An (n, n) -Multi-Secret Image Sharing (MSIS) scheme is used to encrypt n secret images into n meaningless shared images and stored it in different database servers. For recovery of secrets all n shared images are required. In earlier work n secret images are shared among n or $n + 1$ shared images, which has a problem as one can recover fractional information from less than n shared images. Therefore, we need a more efficient and secure (n, n) -MSIS scheme so that less than n shared images do not reveal fractional information. In this paper, we propose an (n, n) -MSIS scheme using additive modulo and reverse bit operation for binary, grayscale, and colored images. The experimental results report that the proposed scheme requires minimal computation time for encryption and decryption. For quantitative analysis Peak Signal to Noise Ratio (PSNR), Correlation, and Root Mean Square Error (RMSE) techniques are used. The proposed (n, n) -MSIS scheme outperforms the existing state-of-the-art techniques.

Keywords Secret sharing scheme • Multi-secret image sharing scheme • Additive modulo • Correlation • RMSE • PSNR

1 Introduction

Nowadays, digital communication plays an important role in data transfer. There exists many applications for communicating secretly. As a result, the prime objective is to secure data from unauthorized access. So, a lot of algorithms and techniques

M. Deshmukh (✉) · N. Nain · M. Ahmed
Malviya National Institute of Technology, Jaipur, India
e-mail: maroti164@gmail.com

M. Deshmukh
National Institute of Technology, Srinagar, Uttarakhand, India

have been designed for data hiding. Watermarking, Steganography, and Cryptography are globally used to hide information. Watermarking is a technique which embeds information into an image. Steganography is used to conceal secret information within another object referred as cover work. Cryptography is used at sender's side to convert plaintext into ciphertext with the help of encryption key and on the other hand, receiver performs decryption operation to extract plaintext from ciphertext using the same key. Visual cryptography is a secret sharing scheme which was first proposed by Shamir [1] and Blakley [2] where a secret image is encrypted into multiple shares which independently disclose no information about the original secret image. The secret image is revealed only after superimposing a sufficient number of shares. The operation of separating the secret into n share is called encryption and the operation of recovery of secret by stacking of shares is called decryption. Visual cryptography schemes are of two types: (k, n) VCS and (n, n) VCS. In (k, n) secret sharing scheme the secret image is encrypted into n shares and at least k shares are required to recover the secret image, less than k shares are insufficient to decrypt the secret image. In (n, n) secret sharing schemes the secret image is encoded into n shares and all n shares are required to recover the secret image, less than n shared images will not reveal the secret. In VCS, the visual quality of the recovered image is poor due to using OR operation [3, 4].

In visual cryptography, the sharing of multiple secrets is a novel and useful application [5, 6]. In (n, n) Multi-Secret Image Sharing (MSIS) scheme n secret images are encrypted into n number of shares which independently disclose no information about the n secret images. For recovery of secret images, all n shares are required [7]. Currently, MSIS scheme have many applications in different areas such as missile launching codes, access control, opening safety deposit boxes, e-voting, or e-auction, etc. The rest of paper is organized as follows. In Sect. 2 discussed the previous work related to secret sharing schemes. The proposed method of an (n, n) -MSIS scheme is presented in Sect. 3. The experimental results and analysis are shown in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Literature Survey

Shyong and Jian [8] designed algorithms using random grids for the encryption of color images and secret gray level in such a way that no information is revealed by a single encrypted share whereas, the secret can be revealed when two shares are stacked. Wang et al. [9] proposed a random grids based incrementing visual cryptography scheme. The incremental revealing effect of the secret image and the size of each secret share is same as that of the original secret image without pixel expansion. The extended visual cryptography scheme based on random grid is first proposed by Gou and Liu et al. [10]. A secret image and k cover images are encoded into k share images. Where, the k share images shows k cover images individually.

The stacking of all the k share images reveals the secret image. Chen and Tsao et al. [11] proposed (k, n) RG-based VSS scheme for binary and color images. A secret image is encrypted into n meaningless random grids such that any random grid alone cannot reveal the secret information, while superimposing at least k random grids will reveal the original secret image. Deshmukh and Prasad [12] presents a comparative study of (k, n) visual secret sharing scheme for binary images and (n, n) VSS scheme for binary and grayscale images has been done for iris authentication. A correlation technique is used for matching the secret image and stacked image. Daoshun and Wang [13] proposed a deterministic (n, n) scheme for grayscale images and probabilistic $(2, n)$ scheme for binary images. Both scheme have no pixel expansion and it uses Boolean operations.

2.1 Tzung-Her-Chen's Method

Chen and Wu et al. [14] proposed an efficient $(n, n + 1)$ -MSIS scheme based on Boolean XOR operations. In this scheme n secret images are used to generate the $n + 1$ shares. For decryption all $n + 1$ shared images are required to reveal the n secret images. In this scheme the capacity of sharing multiple secret images are increased. It uses only XOR calculations for two meaningful images, so it cannot produce a randomized shared images.

This algorithm satisfies the $(n, n + 1)$ threshold criterion. The shortcomings of this method is as follows. The shared images S_i is acquired by $B_i \oplus B_{i-1} = (I_i \oplus R) \oplus (I_{i-1} \oplus R) = I_i \oplus I_{i-1}$, when $\{i = 2, 3, \dots, n - 1\}$. A scheme that uses only XOR operations for two secret images cannot produce the randomized image. Figure 1 shows the result of performing an XOR calculation on two secret images.

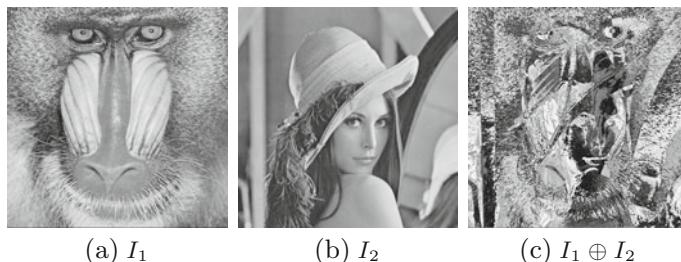


Fig. 1 XOR result: **a** First secret image (I_1). **b** Second secret image (I_2). **c** XOR result of two secret images ($I_1 \oplus I_2$)

2.2 Chien-Chang-Chen's Method

Chen and Wu et al. [15] presented a secure boolean based (n, n) -MSIS scheme. In this scheme, bit shift function is used to generate random image to meet the random requirement. The time required for encryption and decryption is almost same. The flaw of this scheme is that the CPU computation time is more.

In (n, n) -MSIS scheme, it should not disclose any secret information from less than n shared images. However, we can recover partial information from less than n shared images. The inaccuracy of Chien-Chang-Chen's threshold property is described. For odd n , $k = 2 \times \lfloor n/2 \rfloor = (n - 1)$, we get $I_1 \oplus I_2 \oplus \dots \oplus I_{n-1}$ from $N_1 \oplus N_2 \oplus \dots \oplus N_{n-1}$. $N_1 \oplus N_2 \oplus \dots \oplus N_{n-1} = (I_1 \oplus R) \oplus (I_2 \oplus R) \oplus \dots \oplus (I_{n-1} \oplus R)$. $N_1 \oplus N_2 \oplus \dots \oplus N_{n-1} = I_1 \oplus I_2 \oplus \dots \oplus I_{n-1} \oplus (R \oplus R \oplus \dots \oplus R)$. An XOR operation on even number of R is zero. Then, we have $R = F_2(I_1 \oplus I_2 \oplus \dots \oplus I_{n-1})$, and then can recover the secret images $G_1 = N_1 \oplus R, G_2 = N_2 \oplus R, \dots, G_{n-1} = N_{n-1} \oplus R$. Therefore, we get partial information from less than n shared images.

2.3 Ching-Nung Yang's Method

Yang et al. [3] proposed an enhanced boolean-based strong threshold (n, n) -MSIS scheme without revealing any fractional information from less than n shared images. This scheme uses the inherent property of image that a permuted image with a chaotic re-fixation of pixels without affecting their intensity levels.

3 Proposed Method

SSS scheme shares single secret image among n users. To recover the secret image, all n shared images are required. There is need arises for sharing more than one secret images. An (n, n) -MSIS scheme is an n-out-of-n scheme. Chen et al. [15] (n, n) -MSIS scheme should not disclose partial information of secret images from less than n shared images. One can recover partial information from less than n shared images. We overcome the inaccuracy in Chen et al. [15] scheme and propose an (n, n) -MSIS scheme. Proposed scheme do not disclose fractional information from less than n shared images. The use of only XOR operation on secret images is not an good idea because it reveals some information.

In earlier work, simple XOR operation is used for shares generation. The time required to perform XOR operation on secret images are more if the number of secrets increases. To overcome this problem we used additive modulo. In additive modulo simple addition and modulo operations are used which requires less computation time compare to boolean XOR operation. In modular arithmetic, we need to calculate the inverse of a number. Their are two types of modular arithmetic, first

one is an additive and other is multiplicative inverse. In Z_n , two numbers p and q are additive inverses of each other iff $p + q \equiv 0 \pmod{n}$. Here, each integer has a unique additive inverse. In Z_n , two numbers p and q are multiplicative inverse of each other iff $p \times q \equiv 1 \pmod{n}$ and an integer may not necessarily have a multiplicative inverse. Multiplicative inverse of p exists in Z_n iff $\gcd(n, p) = 1$, where p and n are relatively prime. The integer p in Z_n has a multiplicative inverse iff $\gcd(n, p) \equiv 1 \pmod{n}$.

Proposed scheme uses additive inverse because each number in Z_{256} has additive inverse but a number may or may not have a multiplicative inverse in Z_{256} . The grayscale and color images pixel value range from 0–255 so, modulus value is 256. The proposed scheme is applicable for binary, grayscale, and colored images. For binary images pixel value is either 0 or 1, so modulus value for binary images is 2 and we cannot apply reverse bit operation. In the proposed scheme, n secret images I_i , $i = 1, 2, \dots, n$, are encrypted into n shared images S_i , $i = 1, 2, \dots, n$. First the temporary shares T_i , $i = 1, 2, \dots, n$ are generated using additive modulo operation on secret images I_i , $i = 1, 2, \dots, n$. In second step the shares F_i , $i = 1, 2, \dots, n$ are generated using additive modulo operation on temporary shares T_i , $i = 1, 2, \dots, n$. Finally, shared images S_i , $i = 1, 2, \dots, n$ are generated using reverse bit operation on F_i , $i = 1, 2, \dots, n$. In reverse bit operation, it reverse the bits of the pixel value. If pixel value is 64 then binary value of 64 is 01000000. After applying reverse bit function on pixel value 64 then reverse value is 00000010 i.e. 2. The share generation algorithm of proposed (n, n) -MSIS scheme is given in Algorithm 1.

Algorithm 1 : Proposed Sharing Scheme

Input: n secret images $\{I_1, I_2 \dots I_n\}$.

Output: n shared images $\{S_1, S_2 \dots S_n\}$.

1. Generate temporary shares $\{T_1, T_2 \dots T_n\}$ using additive modulo

$$T_1 = (I_1) \bmod 256$$

$$T_i = (I_i + T_{i-1}) \bmod 256, \text{ where } \{i = 2, 3, \dots, n\}$$
 2. Generate n shares $\{F_1, F_2 \dots F_n\}$ using additive modulo

$$F_1 = (T_n) \bmod 256$$

$$F_i = (T_i + F_{i-1}) \bmod 256, \text{ where } \{i = 2, 3, \dots, n-1\}$$

$$F_n = (T_1 + F_{n-1}) \bmod 256$$
 3. Generate n shared images $\{S_1, S_2 \dots S_n\}$ using reverse bit operation

$$S_i = \text{ReverseBits}(F_i), \text{ where } \{i = 1, 2, \dots, n\}$$
-

The recovery procedure is just the reverse of the encryption algorithm. The time required to share n secret images is same as that of the time required to recover n secret images. The shares F_i , $i = 1, 2, \dots, n$ are obtained using reverse bit operation on shared images S_i , $i = 1, 2, \dots, n$. The temporary shares T_i , $i = 1, 2, \dots, n$ are obtained after performing additive inverse operation on F_i , $i = 1, 2, \dots, n$ shares. Finally recovered images are obtained after performing additive inverse operation on temporary shares, T_i , $i = 1, 2, \dots, n$ which is same as that of the secret images. The recovery procedure of proposed (n, n) -MSIS scheme is given in Algorithm 2.

Algorithm 2 : Proposed Recovery Scheme

Input: n shared images $\{S_1, S_2 \dots S_n\}$.

Output: n recovered images $\{R_1, R_2 \dots R_n\}$.

1. Recover $\{F_1, F_2 \dots F_n\}$ images using reverse bit operation
 $F_i = ReverseBits(S_i)$, where $\{i = 1, 2, \dots, n\}$
 2. Recover temporary shares $\{T_1, T_2 \dots T_n\}$ using additive inverse
 $T_1 = (F_n - F_{n-1})mod 256$
 $T_i = (F_i - F_{i-1})mod 256$, where $\{i = 2, 3, \dots, n - 1\}$
 $T_n = (F_1)mod 256$
 3. Recovered shares $\{R_1, R_2 \dots R_n\}$ using additive inverse
 $R_1 = (T_1)mod 256$
 $R_i = (T_i - T_{i-1})mod 256$, where $\{i = 2, 3, \dots, n\}$
-

4 Experimental Results

In this section, we demonstrate the experimental results and analysis of the proposed (n, n) -MSIS scheme. A $(4, 4)$ -MSIS experiment is selected to show the performance of the proposed method. The experiments for binary, grayscale, and colored images are performed on Intel(R) Pentium(R) CPU 2.10 GHz, 2 GB RAM machine using MATLAB 13. All the binary, grayscale, and colored images are of the size of 512×512 pixel.

Figure 2 shows the experimental results of proposed scheme of an (n, n) -MSIS scheme for grayscale images with $n = 4$. An input images I_1, I_2, I_3, I_4 are used as a secrets as shown in Fig. 2a–d. Figure 2e–h shows the four generated noise like a shadow image, S_1, S_2, S_3, S_4 respectively, no individual share reveals partial information of any secret images. Figure 2i–l shows four recovered images, R_1, R_2, R_3, R_4 respectively, which are same as the original images.

Figure 3 shows the experimental results of the proposed (n, n) -MSIS scheme for colored images with $n = 4$. The depth of the binary image is 1, for grayscale image it is 8 and the depth of colored image is 24. So, the computation time for colored images are more than binary and grayscale images and computation time for binary images are less than grayscale and colored images. Figure 3a–d shows the four secret images I_1, I_2, I_3, I_4 respectively. Figure 3e–h shows the four shared images, S_1, S_2, S_3, S_4 respectively, no share reveals partial information of any secret images. Figure 3i–l shows four recovered images, R_1, R_2, R_3, R_4 respectively, which are identical to the original secret images.

4.1 Quantitative Analysis

For quantitative analysis matching between the secret and recovered images is done using correlation, Root Mean Square Error and PSNR technique as shown in Table 1. The correlation value lies between +1 and -1. The correlation value +1 indicates

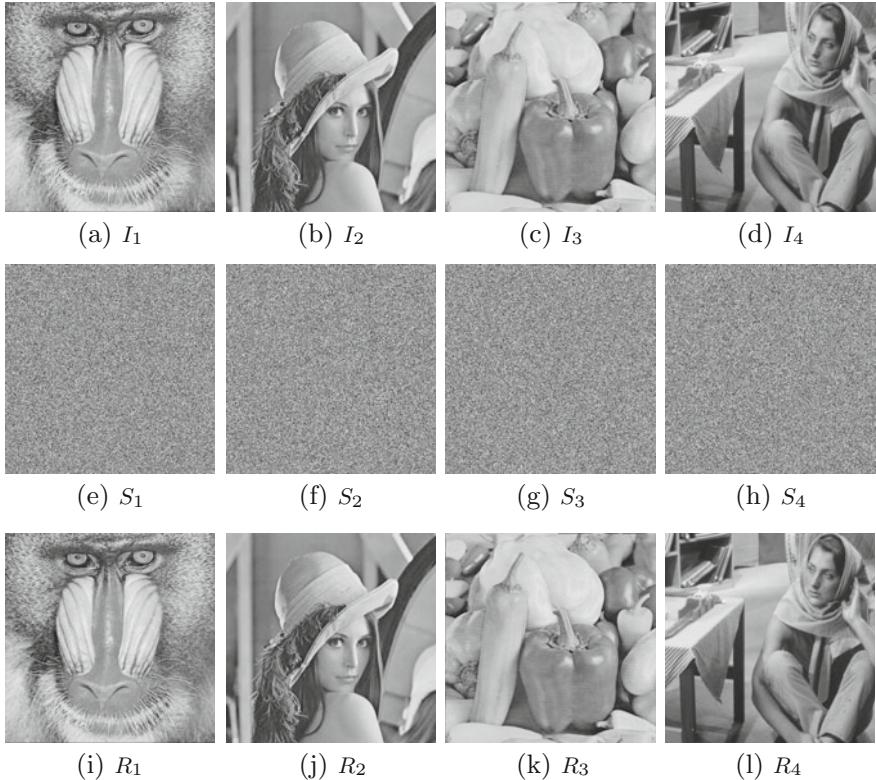


Fig. 2 Proposed scheme of (n, n) -MSIS scheme for grayscale images with $n = 4$: **a–d** Secret images (I_1, I_2, I_3, I_4). **e–h** Shared images (S_1, S_2, S_3, S_4). **i–l** Recovered images (R_1, R_2, R_3, R_4)

that the secret images and recovered images are the same and -1 indicates that both are opposite to each other, i.e., negatively correlated to each other and 0 represents both are uncorrelated. The correlation is given by

$$\text{Correlation} = \frac{n \sum PQ - (\sum P)(\sum Q)}{\sqrt{(n \sum P^2 - (\sum P)^2)(n \sum Q^2 - (\sum Q)^2)}} \quad (1)$$

where, n : number of pairs of scores, $\sum P$: sum of P scores, $\sum Q$: sum of Q scores.

The Peak Signal to Noise Ratio (PSNR) is applied to measure the quality of the recovered images. The PSNR value lies between 0 to ∞ . The higher PSNR indicates a better quality and lower PSNR denotes worse quality. PSNR is measured in decibels (dB). The PSNR of the proposed technique for binary, grayscale, and colored image is ∞ . To construct the RMSE, first determine the residuals. Residuals are the difference between the actual and the predicted values. It is denoted by $P(x, y) - Q(x, y)$, where $P(x, y)$ is the actual value and $Q(x, y)$ is the predicted value. It can be positive



Fig. 3 Proposed scheme of (n, n) -MSIS scheme for colored images with $n = 4$: **a–d** Secret images (I_1, I_2, I_3, I_4). **e–h** Shared images (S_1, S_2, S_3, S_4). **i–l** Recovered images (R_1, R_2, R_3, R_4)

or negative. RMSE value 0 represents secret image P and recovered image Q both the same (No error) otherwise both are different. The PSNR and RMSE are calculated as follows:

$$PSNR(dB) = 20 \log_{10} \frac{255}{RMSE} \quad (2)$$

where, 255 is the highest pixel value in grayscale and colored images. RMSE is the root mean squared error between the original secret P and the recovered image Q which is defined as

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{M \times N} \sum_{x=1}^M \sum_{y=1}^N (P(x, y) - Q(x, y))^2} \quad (3)$$

Comparison between the existing (n, n) -MSIS schemes and the proposed scheme is shown in Table 2. The time required for sharing and recovery of secret images are minimum using additive modulo compared to using XOR operation. The CPU

Table 1 Matching between secret and recovered images using correlation, RMSE, and PSNR technique

Secret and recovered images	Correlation	RMSE	PSNR
I_1, R_1	1.00	0	∞
I_2, R_2	1.00	0	∞
I_3, R_3	1.00	0	∞
I_4, R_4	1.00	0	∞

Table 2 Comparison of existing (n, n) -MSIS and proposed scheme

Parameters	Chen [14]	Wu [15]	Yang [3]	Proposed method
Time (s)	0.120	2.000	0.110	0.070
Secret images	n	n	n	n
Shared images	$n + 1$	n	n	n
Pixel expansion	No	No	No	No
Recovery type	Lossless	Lossless	Lossless	Lossless
Reveals secrets	Partial	Partial	Partial	No
Recovery strategy	XOR	XOR	XOR	Additive modulo
Sharing type	Rectangle	Rectangle	Rectangle	Rectangle
Color depth	Grayscale	Grayscale	Grayscale	Binary, Grayscale, Color
Sharing capacity	$n/n + 1$	n/n	n/n	n/n

computation time of encryption and decryption is same. In proposed (n, n) -MSIS schemes n secret images are used to produce shares and n shares are used to recover the secret images. The size of secret, shared, and reconstructed images are same, there is no pixel expansion. Lossless recovery, recovered images are same as that of secret images. To reveal secret images all n shared images are required. Additive modulo and reverse bit operations are used for sharing and recovery of secrets. Proposed schemes works well for binary, grayscale, and colored images. The sharing capacity of the proposed (n, n) -MSIS scheme is n/n .

5 Conclusion

In this paper, we overcome the inaccuracy in [3, 14, 15] (n, n) -MSIS methods, and proposed an (n, n) -MSIS scheme using additive modulo operation. The time required to perform XOR operation on secret images are more if the number of secret images increases. To overcome this problem we used additive modulo. The generated shares of proposed schemes are random and of same dimensions as that of the secret images.

An (n, n) -MSIS scheme using additive modulo takes less computation time compared to using XOR operation. The experimental results show that the proposed scheme performs effective compared to existing schemes.

References

1. Naor, Moni, and Adi Shamir. "Visual cryptography." *Advances in CryptologyEUROCRYPT'94*. Springer Berlin/Heidelberg, 1995.
2. Blakley, George Robert. "Safeguarding cryptographic keys." *Proceedings of the national computer conference*. Vol. 48. 1979.
3. Yang, Ching-Nung, Cheng-Hua Chen, and Song-Ruei Cai. "Enhanced Boolean-based multi secret image sharing scheme." *Journal of Systems and Software* (2015).
4. Wei, Shih-Chieh, Young-Chang Hou, and Yen-Chun Lu. "A technique for sharing a digital image." *Computer Standards and Interfaces* 40 (2015): 53–61.
5. Lin, Tsung-Lieh, et al. "A novel visual secret sharing scheme for multiple secrets without pixel expansion." *Expert systems with applications* 37.12 (2010): 7858–7869.
6. Lin, Kai-Siang, Chih-Hung Lin, and Tzung-Her Chen. "Distortionless visual multi-secret sharing based on random grid." *Information Sciences* 288 (2014): 330–346.
7. Blundo, Carlo, et al. "Multi-secret sharing schemes." *Advances in Cryptology-CRYPTO94*. Springer Berlin Heidelberg, 1994.
8. Shyu, Shyong Jian. "Image encryption by random grids." *Pattern Recognition* 40.3 (2007): 1014–1031.
9. Wang, Ran-Zan, et al. "Incrementing visual cryptography using random grids." *Optics Communications* 283.21 (2010): 4242–4249.
10. Guo, Teng, Feng Liu, and ChuanKun Wu. "k out of k extended visual cryptography scheme by random grids." *Signal Processing* 94 (2014): 90–101.
11. Chen, Tzung-Her, and Kai-Hsiang Tsao. "Threshold visual secret sharing by random grids." *Journal of Systems and Software* 84.7 (2011): 1197–1208.
12. Maroti Deshmukh, Munaga V.N.K. Prasad. "Comparative Study of Visual Secret Sharing Schemes to Protect Iris Image." *International Conference on Image and Signal Processing (ICISP)*, (2014): 91–98.
13. Wang, Daoshun, et al. "Two secret sharing schemes based on Boolean operations." *Pattern Recognition* 40.10 (2007): 2776–2785.
14. Chen, Tzung-Her, and Chang-Sian Wu. "Efficient multi-secret image sharing based on Boolean operations." *Signal Processing* 91.1 (2011): 90–97.
15. Chen, Chien-Chang, and Wei-Jie Wu. "A secure Boolean-based multi-secret image sharing scheme." *Journal of Systems and Software* 92 (2014): 107–114.

Scheimpflug Camera Calibration Using Lens Distortion Model

Peter Fasogbon, Luc Duvieubourg and Ludovic Macaire

Abstract Scheimpflug principle requires that the image sensor, lens, and object planes intersect at a single line known as “Scheimpflug line.” This principle has been employed in several applications to increase the depth of field needed for accurate dimensional measures. In order to provide a metric 3D reconstruction, we need to perform an accurate camera calibration. However, pin-hole model assumptions used by classical camera calibration techniques are not valid anymore for Scheimpflug setup. In this paper, we present a new intrinsic calibration technique using bundle adjustment technique. We elaborate Scheimpflug formation model, and show how we can deal with introduced Scheimpflug distortions, without the need to estimate their angles. The proposed method is based on the use of optical lens distortion models. An experimental comparison of the proposed approach with Scheimpflug model has been made on real industrial data sets under the presence of large distortions. We have shown a slight improvement brought by the proposed method.

Keywords Camera calibration · Optical distortions · Bundle adjustment · Scheimpflug

1 Introduction

Cameras provide focused image of an observed object when the sensor, lens, and object surface planes are parallel. This is possible for cases when the depth of field is large enough to view the whole object to be measured. However, acquiring the

P. Fasogbon (✉) · L. Duvieubourg · L. Macaire
Laboratoire CRISTAL, Université Lille, UMR CNRS 9189,
Cité Scientifique, Bat. P2, 59655 Villeneuve d'Ascq Cedex, France
e-mail: peter.fasogbon89@gmail.com

L. Duvieubourg
e-mail: luc.duvieubourg@univ-lille1.fr

L. Macaire
e-mail: ludovic.macaire@univ-lille1.fr

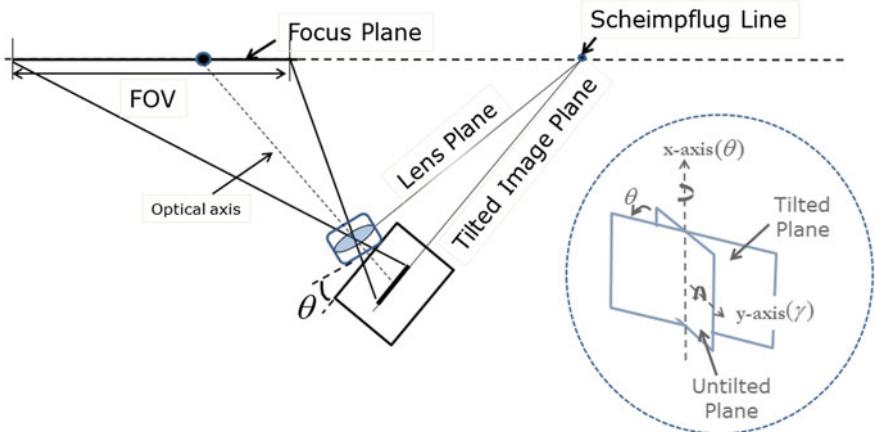


Fig. 1 Scheimpflug system set-up

image of the object surface at sharp focus is very difficult when the object of interest is very large, such as observing the facade of a tall building [1], or when the surface of interest is located in a plane that is oblique to the optical axis, such as Stereo particle image velocimetry (PIV) [2] in particle flow measurements.

However, when the surface of the observed object is not a plane, as for the case of a cylindrical object, only a small region of the object surface is in focus. The first solution consists in increasing the depth of field by decreasing the aperture number of the lens. This leads to decrease the gray level dynamic in the image which results in a poor image intensity. A better solution consists in arranging the optical setup so that it respects the Scheimpflug principle [1]. The idea is to tilt the lens plane so that the camera focus plane coincides with the oblique plane representing the object surface, as shown in Fig. 1. This allows us to obtain a sharp camera's field of view that is wide enough to perform 3D measurement.

Metric 3D reconstruction task requires intrinsic calibration of Scheimpflug cameras since the assumptions used for classical ones [3, 4] are not valid. There have been previous works to calibrate cameras under Scheimpflug principle [5, 6]. These methods use a calibration target that contains several feature points (dot/circle centers), whose locations are known.

The performance reached by camera calibration methods largely depends on the accuracy of the feature point detection. Due to perspective distortion as a result of the orientation of the calibration targets, circles are transformed to ellipses and other deformed shapes. Therefore, the centers of the dot-grids are not well determined. Fronto-parallel approach has been proposed in the past [7, 8] to tackle this problem under pin-hole camera model. A frontal image of the calibration target is created at each pose using the initial camera parameters. This frontal image is free from perspective distortion, which makes it easier to extract the dot-grid centers used as feature pixels. This approach provides good results, and the calibration error is

strongly reduced [8]. The use of fronto-parallel image transformation approach has been extended to Scheimpflug formation model by Legarda et al. [9], and in our previous work [6].

In this paper, we propose an original strategy that consists in fitting Scheimpflug model framework into the lens distortion models. Indeed, we can easily calibrate Scheimpflug cameras with small tilt angles by slightly altering the distortion model. Hence, this model takes into account the Scheimpflug angles without having to estimate these angles. This is possible thanks to iterative bundle adjustment technique and fronto-parallel transformation approach. In the next section, we review the Scheimpflug formation model, followed by presentation of the fronto-parallel approach. The new method based on distortion model, and its calibration procedure are explained in the fourth section. Finally, we compare the experimental results obtained by several Scheimpflug calibration methods using real acquired calibration images.

2 Scheimpflug Model

This section is directly based on our previous work about Scheimpflug formation model [6]. The image plane is tilted at given angles θ , and γ around the optical center for Scheimpflug setup (see Fig. 1). This adjustment ensures that the image plane, the lens plane, and the object plane intersect at a line denoted as “Scheimpflug line.” In that case, the sharp field of view for Scheimpflug setup with the oblique plane is larger than that of classical lens setup.

In this figure, $O_c(X_c, Y_c, Z_c)$ is the optical center in the camera coordinate, the untilted image plane is represented by (O_I, X_C, Y_C) in the camera coordinate, while the tilted image plane (O_t, x_t, y_t) results from its rotation with angle θ about the X_C -axis, and angle γ about the “ Y_c -axis” (see representation of γ in Fig. 1). Assuming no optical distortion, the figure illustrates how a 3D world point P_w is projected into the tilted image point $p_t(x_t, y_t, z_t = 0)$ instead of the untilted point $p(x, y)$.

The thin lens equations still hold for Scheimpflug setup so that $\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$, where f is the focal length, $d_i = \|\vec{O}_c p_t^a\|$ is the image distance, $d_o = \|\vec{O}_c P^a\|$ is the object distance from the center O_c of the lens along the optical axis, p_t^a and P^a are the projections of p_t and P_w to the optical axis.

The spatial pixel coordinates $p_t(u_t, v_t)$ in the tilted image are deduced from those of $P_w(X_w, Y_w, Z_w)$ in the world coordinate system under pin-hole model assumption (i.e., no distortion), thanks to the successive equations (1) [10], and (2–3) [6]. Equation (1) is based on classical pin-hole model while Eqs. (2–3) are deduced from our previous work about Scheimpflug formation model.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = [\mathcal{R} \mid \mathcal{T}] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \text{ then, } \lambda \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} d_i & 0 & 0 \\ 0 & d_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \quad (1)$$

$$\lambda \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \gamma & 0 & -\sin \gamma & d_i \sin \gamma \\ \sin \gamma \sin \theta & \cos \theta & \cos \gamma \sin \theta & -d_i \cos \gamma \sin \theta \\ \sin \gamma \cos \theta & -\sin \theta & \cos \gamma \cos \theta & -d_i \cos \gamma \cos \theta \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} d_i & 0 & 0 \\ 0 & d_i & 0 \\ 0 & 0 & d_i \\ \tan \gamma - \frac{\tan \theta}{\cos \gamma} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ d_i \end{bmatrix}, \quad (2)$$

$$\begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix} = K_b \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix}, \quad \text{where } K_b = \begin{bmatrix} 1/s_x & 0 & u_0 \\ 0 & 1/s_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

Equation (1) first converts the world coordinates $P_w(X_w, Y_w, Z_w)$ of a scene point into camera coordinate system $P_c(X_c, Y_c, Z_c)$ with extrinsic parameters, i.e., the rotation \mathcal{R} and translation \mathcal{T} . Then, from P_c , we compute the coordinates of its projection $p(x, y)$ on the “virtual” image plane denoted as the “untilted” image plane, thanks to the distance d_i . Thereafter, Eqs. (2)–(3) transforms the spatial coordinates (x, y) to pixel coordinates (u_t, v_t) in the tilted image plane, thanks to the Scheimpflug angles $\{\theta, \gamma\}$, the distance d_i , the pixel sizes s_x, s_y and the image center pixel coordinates u_0 , and v_0 .

Scheimpflug image formation model assumes no optical lens distortion. In reality, images are affected by distortions and the most common ones are radial and tangential ones. As these distortions have been caused by the lens, we propose to model them in the untilted image plane. This means that for any distorted pixel $\tilde{p}_t(\tilde{u}_t, \tilde{v}_t)$ on the tilted plane, we need to determine its location $\tilde{p}(\tilde{u}, \tilde{v})$ on the untilted one. Assuming that the Scheimpflug angles θ, γ are well estimated in Fig. 2, pixel $\tilde{p}_t(\tilde{u}_t, \tilde{v}_t)$ can then be projected back to its untilted one $\tilde{p}(\tilde{x}, \tilde{y})$. We can then transform point $\tilde{p}(\tilde{x}, \tilde{y})$ to its pixel coordinate $\tilde{p}(\tilde{u}, \tilde{v})$ using matrix K_b in Eq. (3). The relation between distorted pixel $\tilde{p}(\tilde{u}, \tilde{v})$, and undistorted $p(u, v)$ is shown below in Eq. (4),

$$\begin{aligned} \tilde{u} &= u + (u - u_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [2t_1 xy + t_2(r^2 + 2x^2)] \\ \tilde{v} &= v + (v - v_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [2t_2 xy + t_1(r^2 + 2y^2)] \end{aligned} \quad (4)$$

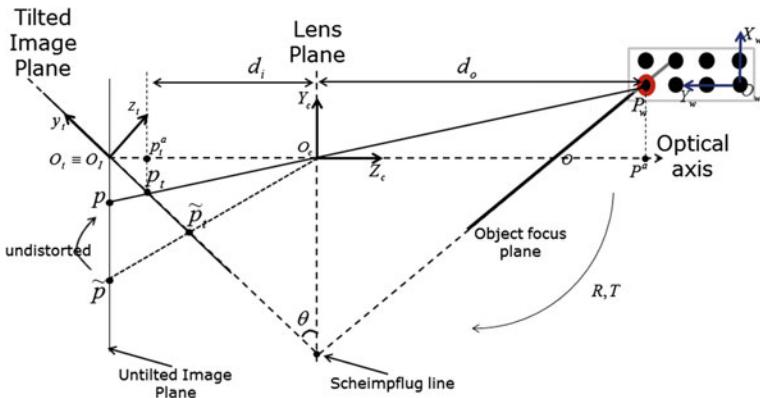


Fig. 2 Scheimpflug camera model, Scheimpflug angle γ on the y-axis is not visible on this figure

where $r = \sqrt{(u - u_0)^2 + (v - v_0)^2}$ is the radial distance, (k_1, k_2, k_3) and (t_1, t_2) are the radial and tangential distortion coefficients respectively.

The objective of calibrating a Scheimpflug camera is to determine the extrinsic parameters $[\mathcal{R}^{(j)} | \mathcal{T}^{(j)}]$ where $j = 1, \dots, m$, m being the number of calibration target poses, the constrained intrinsic parameters $(\mathcal{K} = f_x, f_y, u_0, v_0)$, the distortion coefficients (k_1, k_2, k_3) , (t_1, t_2) , and the Scheimpflug angles (θ, γ) . Note that (f_x, f_y) are the focal length values along axis parallel with the X_c and Y_c axes [3].

3 Fronto-Parallel Improvement

The performance of camera calibration method largely depends on the accurate extraction of the feature pixels from each pose of the calibration target. Due to perspective projection coupled with lens distortions, the shapes of the dot-grids are modified, which may lead to bad estimation of the spatial coordinates of the dot-grid centers (Fig. 3).

To avoid this problem, we apply a fronto-parallel image transformation that consists in creating a new image that looks as the calibration target plane is orthogonal to the optical axis of the camera. The resulting “frontal” image is free from any form of perspective distortion, which leads to accurate detection of the dot-grid centers.

The intrinsic calibration procedure for Scheimpflug camera with the fronto-parallel improvement is divided into the successive following steps:

1. Detect the n feature pixels in each of the m images corresponding to the target poses,
2. Project each feature pixel $\tilde{p}_t^{(i,j)}(\tilde{u}_t, \tilde{v}_t)$ to $\tilde{p}^{(i,j)}(\tilde{u}, \tilde{v})$ using Scheimpflug formation model with initial camera parameters (see Eqs. (1)–(4)),
3. Compute the intrinsic \mathcal{K} , and extrinsic $[\mathcal{R}^{(j)} | \mathcal{T}^{(j)}]$ using Zhang’s method [3],

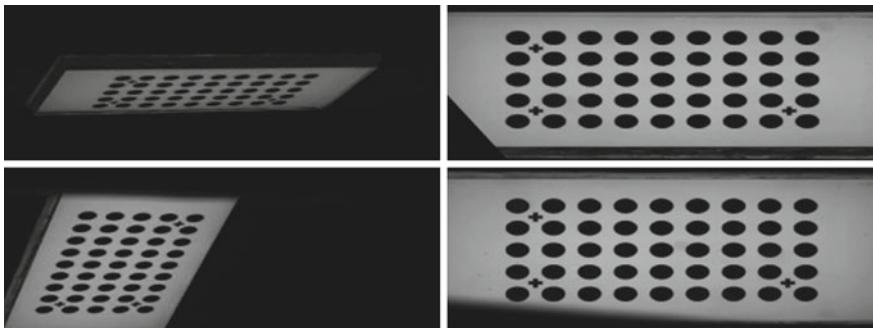


Fig. 3 The *left figure* are an example of the original tilted images and *right figure* are their created corresponding frontal images

4. Estimate optimal Scheimpflug angles $\{\hat{\theta}, \hat{\gamma}\}$ using the optimisation technique proposed in our previous work [6],
5. Project all feature pixels on the tilted plane finally into the untilted ones using the optimal angles $\{\hat{\theta}, \hat{\gamma}\}$, then estimate the parameters $(\hat{k}_1, \hat{k}_2, \hat{k}_3), (\hat{t}_1, \hat{t}_2)$ using bundle adjustment technique to refine the camera parameters,

Improvement using Fronto-Parallel

6. Create the frontal images using parameters $([\hat{\mathcal{R}}^{(j)}, \hat{\mathcal{T}}^{(j)}], \hat{\mathcal{K}}, (\hat{k}_1, \hat{k}_2, \hat{k}_3), (\hat{t}_1, \hat{t}_2), \hat{\theta}, \hat{\gamma})$, and extract the feature pixels in these images,
7. Re-project the detected feature pixels back into the “untilted” image plane, then perform bundle adjustment of all parameters as below,

$$\min \sum_{j=1}^m \sum_{i=1}^n \|\tilde{p}_{\{\hat{\theta}, \hat{\gamma}\}}^{(i,j)} - \hat{p}_{\{\hat{\theta}, \hat{\gamma}\}}^{(i,j)}(\hat{\mathcal{K}}, \hat{\mathcal{R}}^{(j)}, \hat{\mathcal{T}}^{(j)}, \hat{k}_1, \hat{k}_2, \hat{k}_3, \hat{t}_1, \hat{t}_2)\|^2 \quad (5)$$

where $\tilde{p}_{\{\hat{\theta}, \hat{\gamma}\}}^{(i,j)}$, $\hat{p}_{\{\hat{\theta}, \hat{\gamma}\}}^{(i,j)}$ are the real pixels on the “optimal” untilted plane, and estimated ones from bundle adjustment respectively.

8. Go back to step “6” until feature pixels mean variation is lower than a threshold.

We have tested three different methods for the extraction of the feature pixels (i.e., dot-grid centers) on the frontal image. The tested methods are center of gravity method, cross correlation using disc pattern with parabolic fit for sub-pixel detection, and image correlation [8]. All the tested methods tend to provide the same result in terms of calibration quality for our application.

4 Scheimpflug Lens Distortion Model

In this section, we describe our new approach to calibrate Scheimpflug camera effectively without the need to estimate the Scheimpflug tilt angles $\{\theta, \gamma\}$. We propose a simplified Scheimpflug formation based on classical lens distortion models. Thanks to previously explained fronto-parallel approach, it is possible to include this distortion model in the iterative bundle adjustment scheme.

4.1 Analogy of Scheimpflug to Distortion Model

Note that it is geometrically incorrect to write lens distortion on a tilted plane since it has been formulated for untilted ones. As the images provided by the camera result from the projection of world points to the tilted image plane, we propose to write a classical lens distortion expression for tilted plane points \tilde{p}_t as,

$$\begin{aligned} \tilde{u}_t &= u_t + (u_t - u_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [2t_2 x_t y_t + t_1(r^2 + 2x_t^2)] + [s_1 r^2] \\ \tilde{v}_t &= v_t + (v_t - v_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [2t_1 x_t y_t + t_2(r^2 + 2y_t^2)] + [s_2 r^2] \end{aligned} \quad (6)$$

The main difference between the classical distortion model of Eq. (4) and our model presented in Eq. (6) is the thin prism coefficients (s_1, s_2) in the latter equation. This is due to imperfection in the lens construction [8, 11].

From the Scheimpflug image formation in Eqs. (2) and (3) and by setting d_i to “1,” we can express the coordinates (x_t, y_t) of the point in the tilted image plane from the coordinates (x, y) in the untilted image plane as

$$\begin{aligned} x_t &= \frac{x - y \sin \gamma \tan \theta}{x \sin \gamma - y \tan \theta + \cos \gamma} + x_0 \\ y_t &= \frac{y \cos \gamma}{\cos \theta [x \sin \gamma - y \tan \theta + \cos \gamma]} + y_0 \end{aligned} \quad (7)$$

Our main goal here is to express the Scheimpflug formation model to a clearer form that allows better view of how the tilted point $p_t(x_t, y_t)$ is deduced from the untilted one $p(x, y)$. The required equation is shown in Eq. (8) below,

$$\begin{aligned} x_t &= x \frac{1}{\cos \gamma} - y \frac{\sin \gamma \tan \theta}{\cos \gamma} + x_0 x \frac{\sin \gamma}{\cos \gamma} - x_0 y \frac{\tan \theta}{\cos \gamma} \\ y_t &= y \frac{1}{\cos \theta} - y y_0 \frac{\sin \theta}{\cos \theta \cos \gamma} + x y_0 \tan \gamma \end{aligned} \quad (8)$$

Equation 8 shows that,

1. There is a scaling factor $\frac{1}{\cos \gamma}$ on both x and y coordinates which can be handled in the camera matrix by the focal length coefficients f_x and f_y ,
2. The y coordinate contributes individually to express x_t which will be taken into account by the camera matrix skew factor s ,
3. The $xy \equiv x_0 y$ or $y_0 x$ term contributes for both x_t and y_t , which is a nonlinear quantity (meaning that the camera matrix cannot handle it). This needs to be dealt with in a nonlinear optimisation technique.

Solution: We propose to introduce two new coefficients c_1 and c_2 in the distortion model so that Eq. (6) is expressed as

$$\begin{aligned} \tilde{u}_t &= u_t + (u_t - u_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [t_1(r^2 + 2x_t^2)] + [s_1 r^2] + c_1 x_t y_t \\ \tilde{v}_t &= v_t + (v_t - v_0)[k_1 r^2 + k_2 r^4 + k_3 r^6] + [t_2(r^2 + 2y_t^2)] + [s_2 r^2] + c_2 x_t y_t \end{aligned} \quad (9)$$

The difference between Eqs. (6) and (9) is that we have removed the second-order tangential ($2t_2, 2t_1$), and replaced it by coefficients (c_1, c_2) that are independent of the tangential distortion. However, we will see later that we can initialise the coefficients (c_1, c_2) using the tangential ones.

Therefore, the use of nonlinear optimisation in the form of an “iterative” bundle adjustment should be able to estimate the coefficients (c_1, c_2), and at the same time be able to refine the other system parameters.

4.2 Calibration Procedure

In summary, the calibration procedure for the new proposed method is as follows:

1. Detect the n feature pixels in each of the m images corresponding to the target poses,
2. Perform classical calibration procedure on the tilted image plane using Zhang's method [3]. The extracted parameters are $\hat{\mathcal{K}}$, $\hat{\mathcal{R}}^{(j)}$, and $\hat{\mathcal{T}}^{(j)}$, and the lens distortion coefficients $(\hat{k}_1, \hat{k}_2, \hat{k}_3, \hat{l}_1, \hat{l}_2)$
3. Initialize the “iterative” bundle adjustment scheme with the parameters estimated in step (2). Set the additional parameter $\hat{s} = 0$ [3], and coefficients $\hat{s}_1, \hat{s}_2 = 0$, $\hat{c}_1 = 2\hat{l}_2$, and $\hat{c}_2 = 2\hat{l}_1$,

Start Iterative bundle adjustment

4. Create the frontal images using $(\hat{\mathcal{R}}^{(j)}, \hat{\mathcal{T}}^{(j)}, \hat{\mathcal{K}}, \hat{k}_1, \hat{k}_2, \hat{k}_3, \hat{l}_1, \hat{l}_2, \hat{s}_1, \hat{s}_2, \hat{c}_1, \hat{c}_2)$,
5. Extract the feature pixels on the frontal image, then project them back to the tilted image plane,
6. Optimize all the parameters directly on the tilted image plane using bundle adjustment procedure as below,

$$\min \sum_{j=1}^m \sum_{i=1}^n \|\tilde{p}_t^{(i,j)} - \hat{p}_t^{(i,j)}(\hat{\mathcal{K}}, \hat{\mathcal{R}}^{(j)}, \hat{\mathcal{T}}^{(j)}, \hat{k}_1, \hat{k}_2, \hat{k}_3, \hat{l}_1, \hat{l}_2, \hat{s}_1, \hat{s}_2, \hat{c}_1, \hat{c}_2)\|^2 \quad (10)$$

7. Go back to step (4) until the feature pixel variation is lower than a threshold.

5 Experimental Results

For the experiment, we use a Basler camera with a resolution of 2040×1088 pixels, pixel size of $0.0055 \text{ mm} \times 0.0055 \text{ mm}$, and a Schneider-Kreuznach-f24mm Scheimpflug lens tilted to approximately $\theta_0 = 1.4^\circ$ with respect to the X-axis. The initial angle γ_0 has been set to 0° since we have no information about its adjustment on the Y-axis.

To compare the performances reached by the calibration procedures detailed in Sects. 3 and 4, we use two different sets of 25 and 75 calibration images of the proposed calibration target. The target is moved at several poses to cover the camera's field of view, with each image containing 45 feature pixels (see Fig. 4). The diameter of the circle pattern is 5 mm, and the distance between two centers is 7.5 mm. The camera used for the experiments includes a wavelength filter, and is placed in a thick “plexi” glass housing that causes strong distortion coupled with distortion introduced by Scheimpflug tilt effects.

The calibration parameters estimated by Legarda's method [5], Fasogbon [6] and our new approach are displayed in Table 1. The error evaluation is determined using the standard deviation and mean reprojection error between the true coordinates

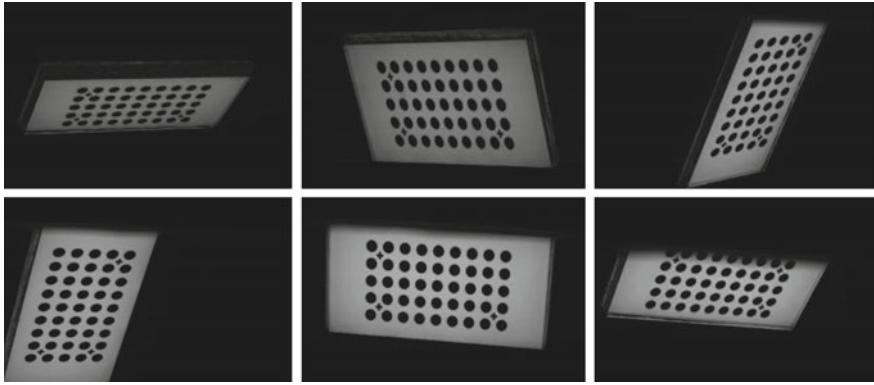


Fig. 4 Calibration images captured using Scheimpflug camera

Table 1 Calibration evaluation result on the three calibration procedures “Legarda [5], Fasogbon [6] and the proposed method” using fronto-parallel approach

Parameters	25 images			75 images		
	Legarda [5]	Fasogbon [6]	New Approach	Legarda [5]	Fasogbon [6]	New Approach
f_x, f_y (mm)	25.38, 25.39	25.54, 25.47	25.28, 25.36	25.49, 25.43	25.50, 25.43	25.61, 25.56
s	–	–	26.63	–	–	-21.55
u_0, v_0 (px)	1209, 676	1141, 629	1235, 647	1135, 631	1136, 616	1128, 657
θ, γ (°)	1.401, 0.010	1.399, 0.004	–	1.385, 0.015	1.387, 0.039	–
k_1	0.030	0.014	0.053	0.058	0.057	0.062
k_2	-4.347	-5.316	-4.120	-6.628	-6.602	-7.018
k_3	37.128	54.448	32.048	66.762	66.430	72.285
t_1	0.0099	0.0015	0.0133	0.0023	0.0001	-0.0040
t_2	0.0113	0.0029	0.0115	0.0036	0.0049	0.0137
s_1	-0.0101	–	-0.0067	-0.0065	–	0.0032
s_2	-0.0109	–	-0.0085	0.0021	–	-0.0104
c_1	–	–	0.0278	–	–	0.0013
c_2	–	–	0.0385	–	–	-0.0052
mean err (px)	0.0521	0.0525	0.0479	0.0724	0.0730	0.0716
std err (px)	0.0301	0.0309	0.0280	0.0499	0.0496	0.0495

of the feature pixels and the reprojected ones using the estimated camera parameters for each calibration method. We can conclude from the table that the new proposed method based on distortion model performs slightly better than the previous Scheimpflug model. We have used the same stopping criteria for the iterative bundle adjustment to avoid any bias between the three compared methods.

6 Conclusion

A new method for the calibration of cameras under Scheimpflug conditions has been presented. We showed in this paper that we can calibrate Scheimpflug cameras without the need to estimate the tilt angles θ , and γ . This is made possible thanks to the fronto-parallel transformation which avoids the perspective distortion problem, and the introduction of distortion coefficients in the iterative bundle adjustment scheme based on Levenberg Marquardt technique.

The goal of our project is to accurately measure the cross-section of tiny industrial cylindrical objects observed under large optical distortions. This means that slight improvement in the intrinsic calibration result is highly welcome. The proposed method performs much better than the past methods on small Scheimpflug tilt angles.

Our scheimpflug device is tilted with small angles, we should determine the validity space of the new method, i.e., determine the angle ranges with which this method provides satisfying results.

Acknowledgments This work has been financed by unique inter-ministry fund (FUI) of the Nord-Pas-de-Calais region in France.

References

1. Merklinger, H. M.: Focusing the view camera: A scientific way to focus the view camera and estimate depth of field. Internet edition 1.6.1, Canada (2010).
2. Astarita T.: A Scheimpflug camera model for stereoscopic and tomographic PIV. In: Proceedings of the 16th Int. Symp on Applications of Laser Techniques to Fluid Mechanics, pages 1–9, 2012.
3. Zhang Z.: A flexible new technique for camera calibration. In: IEEE Trans. Pattern Anal. Mach. Intell., pp. 1330–1334, 22(11), November (2000).
4. Armangue X., Salvi J., Batlle J.: A comparative review of camera calibrating methods with accuracy evaluation. In: Pattern Recognition, 35:1617–1635, 2002.
5. Legarda A., Izaguirre A., Arana N., Iturrospe A.: A new method for Scheimpflug camera calibration. In: proc. of the 10th IEEE International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics (ECMSM), pages 1–5, June 2011.
6. Fasogbon P., Duvieubourg L., Lacaze P.A. and Macaire L.: Intrinsic camera calibration equipped with Scheimpflug optical device. In: Proc. of the 12th International Conference on Quality Control by Artificial Vision (QCAV), June (2015).
7. Datta A., Kim J., Kanade T.: Accurate camera calibration using iterative refinement of control point. In: Proceedings of Workshop on Visual Surveillance (VS), October (2009).
8. Minh V., Zhaoyang W., Long L., Jun M.: Advanced geometric camera calibration for machine vision. In: Optical Engineering, 50(11): 110503-110503-3, 2011.
9. Legarda A., Izaguirre A., Arana N., Iturrospe A.: Comparison and error analysis of the standard pin-hole and Scheimpflug camera calibration models. In: proc. of the 11th IEEE International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics (ECMSM), pages 1–6, 2013.

10. Hall E., Tio L., James B.K., McPherson C. A., Sadjadi F. A.: Measuring curved surfaces for robot vision. In: Computer Journal. 15(12):42–54, 1982.
11. Wang J., Shi F., Zhang J., Liu Y.: A new calibration model of camera lens distortion. In: Pattern Recognition, 41(2): pp 607–615, 2008.

Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL)

Sonali Mishra, Lokesh Sharma, Bansidhar Majhi
and Pankaj Kumar Sa

Abstract Development of a computer-aided diagnosis (CAD) system for early detection of leukemia is very essential for the betterment of medical purpose. In recent years, a variety of CAD system has been proposed for the detection of leukemia. Acute leukemia is a malignant neoplastic disorder that influences a larger fraction of world population. In modern medical science, there are sufficient newly formulated methodologies for the early detection of leukemia. Such advanced technologies include medical image processing methods for the detection of the syndrome. This paper shows that use of a highly appropriate feature extraction technique is required for the classification of a disease. In the field of image processing and machine learning approach, Discrete Cosine Transform (DCT) is a well-known technique. Nucleus features are extracted from the RGB image. The proposed method provides an opportunity to fine-tune the accuracy for the detection of the disease. Experimental results using publicly available dataset like ALL-IDB shows the superiority of the proposed method with SVM classifier comparing it with some other standard classifiers.

Keywords Acute Lymphoblastic Leukemia • Discrete Cosine Transform • Water-shed segmentation • CAD system

S. Mishra (✉) · L. Sharma · B. Majhi · P.K. Sa
Pattern Recognition Research Lab, Department of Computer
Science and Engineering, National Institute of Technology, Rourkela 769008, India
e-mail: smishra.nitrkl@gmail.com

L. Sharma
e-mail: lksharma1064@gmail.com

B. Majhi
e-mail: bmajhi@nitrkl.ac.in

P.K. Sa
e-mail: pankajksa@nitrkl.ac.in

1 Introduction

Acute Leukemia is a rapidly increasing disease that affects mostly the cells that are not yet fully developed. Acute Lymphoblastic Leukemia (ALL) is a significant ailment caused by the unusual growth and expansion of white blood cells [1]. ALL begins with the abnormalities starting from the bone marrow, resulting in reducing the space for red blood cells. The ALL blasts become so numerous that they flood through the red blood cells and platelets. As the cells build up, they reduce the immunity to fight with the foreign material. Hence, it is essential to treat the disease within a short span of time after making a diagnosis. As per the survey done by American Cancer Society, it has approximated that, in 2015 a total of 1,658,370 has been diagnosed, out of which 589,430 died in the US. In India, the total number of individuals suffering from leukemia was estimated to be 1,45,067 in 2014. Furthermore, as per the Indian Association of blood cancer and allied diseases, among all the cancers which is dangerous and can cause death, leukemia constitute one-third of the cases. ALL is mostly seen in children below 14 years [2].

ALL is identified with the excessive production of immature lymphocytes that are commonly known as lymphoblasts. The uncontrolled manufacture of lymphoblasts puts a stop to the formation of blood in the marrow, which eventually leads to the cause of death. The recognition of the blast cells in the bone marrow of the patients suffering from acute leukemia is a crucial step in the recognition of the development stage of the illness and proper treatment of the patients. The percentage of blasts are an important factor to define the various stages of lymphoblastic leukemia. According to the French–American–British (FAB) standard, three different types of acute lymphoblastic leukemia are classified based on the morphology of blast cells [3].

The morphological identification of acute leukemia is mainly performed by the hematologists [4]. The process begins by taking the bone marrow sample from the patient's spine. Wright's staining method is applied to make the granules visible during analysis [5]. This process involves many drawbacks, such as slowness of the analysis, a very low accuracy, requirement of an extremely skilled operator, etc. The identification by experts is reliable, but automated tools would be useful to support experts and also helpful in reducing the cost. The primary goal of this work is to analyze microscopic images by designing a computer-aided diagnosis system (CAD) to support medical activity.

This paper presents a new hybrid technique for acute leukemia classification from the microscopic images based on machine learning approaches. The proposed method mainly consists of three different steps namely, Segmentation, feature extraction, and classification of the disease. Over the years, many automatic segmentation techniques have been proposed for the disease, still they fail to segment the overlapping blood cells. This scheme utilizes discrete cosine features and support vector machine (SVM) for classification of normal and malignant cells.

The rest of the paper is organized as follows: Sect. 2 deals with some of the highly regarded works on the detection of leukemia from the blood smear along with some segmentation schemes. Section 3 presents the proposed work. Section 4 gives a com-

parative performance study of the proposed method with existing schemes. Finally the concluding remarks are provided in Sect. 5.

2 Related Work

A careful study on automatic blood cell recognition reveals that numerous works have been reported since early 2000. All these existing techniques said, giving a near perfect performance under certain constraints. Various segmentation and feature extraction techniques have been examined for the same. The following paper gives an overview of the different segmentation and feature extraction techniques based on their category.

Scotti [6] has proposed a method for automated classification of ALL in gray level peripheral blood smear images. As per the experiments conducted by them on 150 images, it has been concluded that lymphoblast recognition is feasible from blood images using morphological features. Gupta et al. [7] have proposed a suitable support vector machine-based technique for the identification of three types of lymphoblasts. The classification accuracy for the childhood ALL has been promising but needs more study before they are used for the adult. Escalante et al. [8] have suggested an alternative approach to leukemia classification using ensemble particle swarm model selection. Manually isolated leukemia cells are segmented using Markov random fields. This method is useful for ALL versus AML (Acute Myeloblastic Leukemia) classification. Putzu et al. [9] have proposed a scheme for automated detection of leukemia using image processing techniques. The segmentation procedure produces a lower segmentation rate and can be further improved. Mohapatra et al. [10] have suggested an ensemble classifier system to classify a blood cell into normal lymphocyte and an unhealthy one (lymphoblast). The scheme yields good accuracy for detecting the disease, but failed to detect the disease for grouped cell present in an image.

3 Proposed Work

The proposed method comprises of different stages such as the acquisition of images, preprocessing, segmentation of overlapping cells, feature extraction, and classification of the image into a normal (Benign) and abnormal (malignant) one. Figure 1 shows an overall block diagram of the proposed method. Each stage is discussed below in brief. The main contribution in this article is the use of discrete cosine transform (DCT) coefficients in SVM classifier for classification.

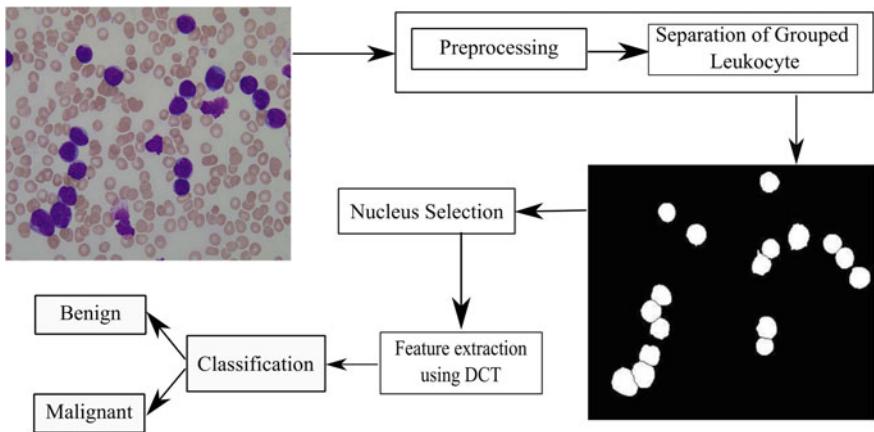


Fig. 1 Block diagram of the method for the classification of ALL

3.1 *Preprocessing*

Due to the presence of noise in the microscopic images under the diverse lighting conditions, the image requires preprocessing prior to segmentation. To generate a better quality image, Weiner filtering followed by contrast enhancement with histogram equalization is used.

3.2 *Separation of Grouped Leukocyte Using Segmentation*

Segmentation is a critical step for correct classification of the objects. Microscopic images are typically in RGB color space. It is very difficult to achieve accurate segmentation in the color image. So the RGB image is converted into Lab color space to reduce the dimension with the same color information. Color-based clustering mainly uses the Lab color space for the segmentation purpose.

Due to the presence of overlapped and grouped objects, marker-based watershed segmentation algorithm [11] is used for separating grouped and overlapped objects. In Lab space, component ‘a’ contains the highest information about the nucleus. So further processing is done on the component ‘a’. After separating the objects from an image, all the lymphocytes are extracted using the bounding box technique for the detection of ALL. Finally, the single lymphocyte sub-image is used in the next process for feature extraction.

3.3 Feature Extraction

The fundamental step of the proposed scheme is to calculate the DCT features from the lymphocyte sub-images of size $M \times N$. The cosine transform generates $M \times N$ DCT coefficients. Since the DCT has the energy compaction property and the energy coefficient are in descending order, the higher order coefficients are significant and the lower coefficient can be neglected. Hence, very few coefficients retain the energy of the whole image and reconstruct the original image with minor loss of information. This particular behavior of DCT has been exploited to use few DCT coefficients as a feature for classification of normal and abnormal cells in the image. Also, the feature extraction capacity of the DCT coupled with fast computation time has made it a worldwide candidate for pattern recognition. The general equation for a 2D ($M \times N$ image) DCT is defined by the following equation [12]:

$$F(u, v) = \frac{1}{\sqrt{MN}} \alpha(u)\alpha(v) \sum_{x=1}^M \sum_{y=1}^N f(x, y) \cos \left[\frac{(2x+1)u\pi}{2M} \right] \cos \left[\frac{(2y+1)v\pi}{2N} \right] \quad (1)$$

Gray scale coordinate of the image of size $M \times N$ is represented by, $f(x, y)$, where $1 \leq x \leq M, 1 \leq y \leq N$. $\alpha(w)$ can be defined as,

$$\alpha(w) = \begin{cases} \frac{1}{\sqrt{2}}, & \text{for } w = 1 \\ 1, & \text{otherwise} \end{cases} \quad (2)$$

The detailed step for feature extraction is articulated in Algorithm 1.

Algorithm 1 Feature Extraction Algorithm

Require: Samples of n lymphoblasts lymphocytes from the segmentation step

Ensure: $X[n : m]$: Feature matrix, $S[1 : n, 1 : m + 1]$: New dataset

- 1: Compute DCT features using function `dct2()` from the microscopic images
 - 2: Store the features in the $X[n \times m]$
 - 3: Append another vector Y to the X and assign a class level for each sample
 - 4: Form a new dataset, $S=(x_i, y_i), x_i \in X, y_i \in Y, 1 \leq i \leq n$
-

3.4 Classification

Classification is the task of assigning an unknown feature vector, to one of the known classes. Each classifier has to be built up in such a way that a set of inputs must produce a desired set of outputs. In this paper support vector machine (SVM), a classifier is used for differentiating normal and malignant cells. Support vector machine is a very powerful supervised learning technique that was first introduced by Vap-

nik [13]. It is a two class supervised classifier. It uses a nonlinear mapping function for transforming the input data into a high-dimensional feature space by creating a hyperplane between the two categories. The entire set of measured data is divided into training and testing data. Here, images from ALL-IDB1 is used for both training and testing purpose. The extracted features from the above step can be classified using three other standard classifiers, i.e., Naive Bayesian [14], KNN [15], BPNN [16], and SVM. The basic steps of classification process are given in Algorithm 2.

Algorithm 2 Classifier System

Require: S : Training dataset with N samples, $S = (x_i, y_i)$ for $i = 1, 2, \dots, N$ and $x_i \in X$ with class labels $y_i \in Y$
 X and Y represents the input and output class respectively.

- 1: Perform K-fold cross validation
- 2: **for** each classifier, ($j=1$ to M) **do**
- 3: Train the classifier
- 4: Calculate the performance measures on test images
- 5: **end for**

4 Experimental Setup and Results

The experiments are carried out on a PC with 3.40 GHz Core-i7 processor and 4 GB of RAM, running under Windows 8 operating system. The proposed algorithm is simulated using MATLAB 2013 toolbox. Specimen blood samples are collected from a public database ALL-IDB [17, 18] having two distinct versions, namely, ALL-IDB1 and ALL-IDB2 containing 108 and 260 images, respectively. The ALL-IDB1 dataset is used for the purpose of training and testing. Among them, 59 are normal images and 49 are affected blood cells that consist of at least one lymphoblast. The ALL-IDB2 dataset is made from ALL-IDB1 by cropping the images having less dimension that is mainly used for testing purpose. The ALL-IDB is a public database employed in the field of medical image processing for the research purpose and detection of the tumor in the blood cells. Each image present in the dataset is represented using three primary colors (red, green, blue), and the image is stored with a size of 1368×1712 array.

To make the classifier more stable and more generalize, a fivefold cross-validation (CV) procedure is utilized. In this work, the abnormal (malignant) and normal (benign) images have been considered to be in the positive and negative class, respectively. Sensitivity is the probability that an investigative step is positive while the patient has the disease, whereas specificity is the probability that a diagnostic report is negative while the patient has not got any disease. For a given sample, a training system leads to four possible categories that are described in Table 1.

Table 1 Confusion matrix

	Positive	Negative	Performance measure
Positive	True Positive (TP)	False Positive (FP)	Positive predictive value = $TP/(TP + FP)$
Negative	False Negative (FN)	True Negative (TN)	Negative predictive value = $TN/(TN + FN)$
Performance measures	True Positive Rate (Sensitivity) = $TP/(TP + FN)$	True Negative Rate (Specificity) = $TN/(TN + FP)$	Accuracy = $(TP + TN)/\text{Total number of samples}$

4.1 Results and Discussion

The first step after the acquisition of the image is to clean the image to differentiate the lymphoblasts from the other component of the cell like RBC, platelets, etc. Weiner filter is being used to reduce noise present in the image along with the contrast enhancement. The next step is the segmentation process. Grouped cells of an image are separated using watershed segmentation. Conventional watershed segmentation is used along with the use of a marker. Marker-controlled watershed transform is a two-way process. It consists of two types. Internal markers represent the blast cell nucleus, and external markers represent the boundary to separate the grouped cells. Figure 2 represents the overall steps associated with the segmentation process.

The next step is to find the DCT coefficients. Here, the DCT coefficients have been taken as the feature for the classification process. Due to the energy compaction properties of the DCT, less computational time is required. The number of extracted

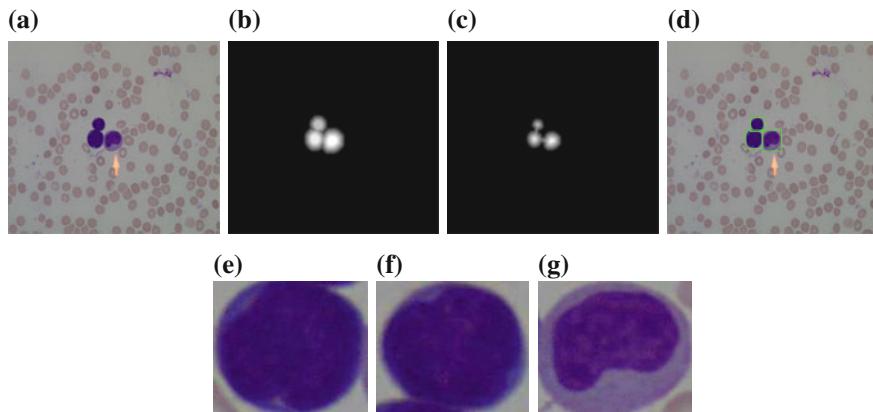


Fig. 2 Different segmentation steps: **a** original image, **b** image after using external marker, **c** image after using internal marker, **d** detected lymphoblasts, **e–g** detected sub-image using bounding box

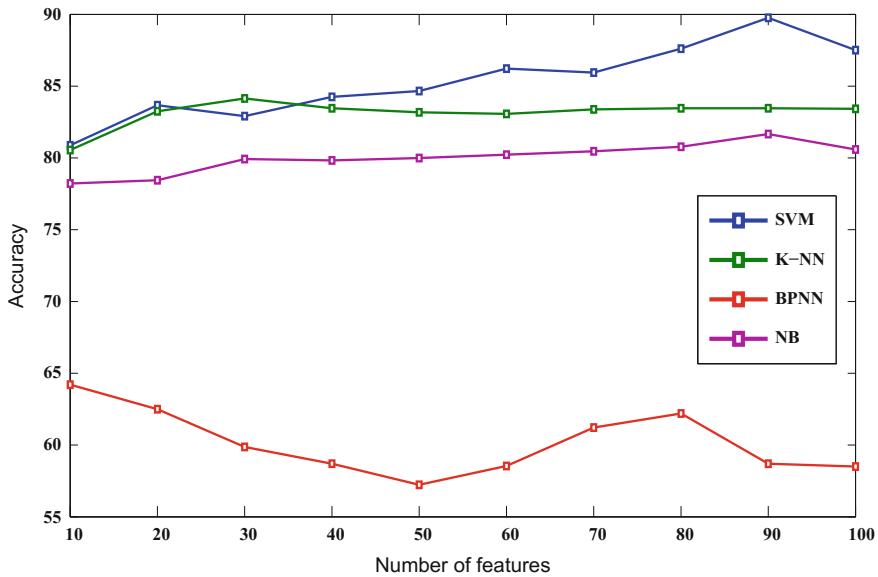


Fig. 3 Accuracy rate with the increase in number of features

Table 2 Comparison of accuracy of various classifiers over fivefold

Classifier	Fold					Average accuracy (%)
	1	2	3	4	5	
NB	78.49	83.78	81.23	83.45	81.35	81.66
KNN	82.85	80.59	85.51	84.23	84.12	83.46
BPNN	58.54	54.26	60.85	56.95	63.20	58.7
SVM	85.21	94.32	85.59	88.62	95.06	89.76

features is found to be 90. In Fig. 3, the classification accuracies of different classifiers with some features are portrayed. It is observed from that, all the classifiers show maximum accuracy at feature number 90. The acquired features are fed to different classifiers to get different performance measures. Tables 2, 3, and 4 present the fold-wise result of the fivefold cross-validation procedure for the determination of accuracy, sensitivity, and specificity for different classifiers. The optimum has been achieved with an accuracy of 89.76 % using the SVM classifier. The obtained values of sensitivity and specificity are found to be 84.67 % and 94.61 % respectively. Note that all the schemes are tested on the same dataset ALL-IDB1.

Table 3 Comparison of average sensitivity of various classifiers over fivefold

Classifier	Fold					Average Sensitivity (%)
	1	2	3	4	5	
NB	88.62	87.32	83.65	91.25	92.66	88.70
KNN	99.80	96.32	97.35	99.21	99.21	98.38
BPNN	77.38	83.25	85.23	79.36	77.98	80.64
SVM	89.36	85.69	88.22	79.24	80.84	84.67

Table 4 Comparison of average specificity of various classifiers over fivefold

Classifier	Fold					Average Specificity (%)
	1	2	3	4	5	
NB	70.32	78.61	76.55	80.29	71.13	75.38
KNN	62.89	68.25	69.41	73.56	72.04	69.23
BPNN	35.23	39.58	36.24	38.22	39.18	37.67
SVM	96.23	95.37	94.35	88.39	98.71	94.61

5 Conclusion

In this work, a hybrid system for the automatic classification of leukemia using microscopic images has been proposed. This system first applies Wiener filter followed by histogram equalization to preprocess the image. The watershed segmentation algorithm has been utilized to correctly separate the lymphocytes sub-image from the preprocessed image. The DCT-based feature is used for deriving a set of features from the sub-images. Subsequently, SVM is used to classify the images as benign and malignant. The simulation results show the efficacy of the proposed scheme while testing the system with SVM classifier. The classification accuracy on dataset ALL-IDB1 is found to be 89.76 % using SVM. However, there is a scope to reduce the computational overhead of the feature extraction step, and also works can be further extended toward the extraction of cytoplasm from the blood cells.

References

1. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2013. CA: a cancer journal for clinicians 63.1, 11–30 (2013)
2. Kulkarni, K.P., Arora, R.S., Marwaha, R.K.: Survival outcome of childhood acute lymphoblastic leukemia in India: a resource-limited perspective of more than 40 years. Journal of pediatric hematology/oncology 33.6, 475–479 (2011)
3. Singh, T.: Atlas and text of hematology. Avichal Publishing Company, New Delhi 136 (2010)
4. Saraswat, M., Arya, K.V.: Automated microscopic image analysis for leukocytes identification: A survey. Micron 65 20–33 (2014)

5. Wright, J. H.: The histogenesis of the blood platelets. *Journal of Morphology*. Vol. 3. No. 1. (1910)
6. Scotti, F.: Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. *IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. (2005)
7. Gupta, L., Jayavant, S., Ramaiah, A.: Identification of different types of lymphoblasts in acute lymphoblastic leukemia using relevance vector machines. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 6675–6678 (2008)
8. Escalante, H.J., et al.: Acute leukemia classification by ensemble particle swarm model selection. *Artificial intelligence in medicine* 55.3 163–175 (2012)
9. Putzu, L., Caocci, G., Ruberto, C.D.: Leucocyte classification for leukaemia detection using image processing techniques." *Artificial intelligence in medicine* 62.3 179–191 (2014)
10. Mohapatra, S., Patra, D., Satpathy, S.: An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications* 24.7-8 1887–1904 (2014)
11. Parvati, K., Rao, P., Das, M.M.: Image segmentation using gray-scale morphology and marker-controlled watershed transformation. *Discrete Dynamics in Nature and Society* (2008)
12. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. *IEEE Transactions on Computers*, 100.1 90–93 (1974)
13. Vapnik, V.N., Vapnik, V.: *Statistical learning theory*. Vol. 1. New York: Wiley, (1998)
14. Duda, R.O., Hart, P.E., Stork, D.: *Pattern classification*. John Wiley & Sons, (2012)
15. Acharya, T., Ray, A.K.: *Image processing: principles and applications*. John Wiley & Sons, (2005)
16. Rumelhart, D.E., Hinton, G. E., Williams, R.J.: Learning representations by back-propagating errors. *Cognitive modeling* 5.3 (1988)
17. Labati, R.D., Piuri, V., Scotti, F.: All-IDB: The acute lymphoblastic leukemia image database for image processing. *18th IEEE international conference on Image processing (ICIP)*, (2011)
18. ALL-IDB Dataset for ALL Classification. <http://crema.di.unimi.it/~fscotti/all/>

Robust Image Hashing Technique for Content Authentication based on DWT

**Lokanadham Naidu Vadlamudi, Rama Prasad V. Vaddella
and Vasumathi Devara**

Abstract This paper presents an image hashing technique for content verification using Discrete Wavelet Transform (*DWT*) approximation features. The proposed technique converts resized *RGB* color images to $L^*a^*b^*$ color images. Further, images are regularized using Gaussian low pass filter. A level 2, 2D *DWT* is applied on L^* component of $L^*a^*b^*$ color image and the LL_2 approximation sub-band image is chosen for feature extraction. The features are extracted by utilizing a sequence of circles on approximation sub-band image. Finally, the robust binary hash is generated from extracted features. The experimental results indicate that the hash of the presented technique is invariant to standard content preserving manipulations and malicious content altering operations. The experiment results of Receiver Operating Characteristics (*ROC*) plots indicate that the presented technique shows strong discriminative and robustness capabilities. Besides, the hash of the proposed technique is shorter in length and key dependent.

Keywords Image authentication • Image hashing • Discrete Wavelet Transform • Approximation features • Robustness

L.N. Vadlamudi (✉)

Information Technology, Sree Vidyanikethan Engineering College,
Tirupati 517102, AP, India
e-mail: vlnaidu1982@gmail.com

R.P.V. Vaddella

Computer Science and Engineering, Sree Vidyanikethan Engineering College,
Tirupati 517102, AP, India
e-mail: vvramaprasad@redffmail.com

V. Devara

Computer Science and Engineering,
JNTU College of Engineering, JNT University, Hyderabad 500085, TS, India
e-mail: rochan44@gmail.com

1 Introduction

With the modern image editing software, digital images can be easily tampered without loss of their perceptual meaning that makes the verification of integrity and authenticity of images critical. Many methods exist for verifying the integrity of images and checking of authenticity. Generally, these are classified as watermarking and digital hashing methods [1, 2]. The watermarking methods insert secret data into images and the inserted secret data will be extracted for verifying images integrity. The second category methods generate a digital hash from images using existing crypto hash algorithms like MD5, SHA-1, etc. The main drawback of these algorithms is that the output of the algorithm will be changed radically, even when small changes occur in input data. Moreover, images undergo some normal image processing operations including geometric transformations, filtering, compression, brightness and contrast adjustment, color conversions, and format conversion. These operations will change image pixel values that result with different distortions in images, but the image content is preserved. Due to the sensitivity property of crypto hash algorithms, they can classify images as unauthentic when images undergo normal image processing operations. The alternative way for verification of integrity of images is Content-Based Image Authentication (*CBIA*) [2, 3]. *CBIA* is a process that uses features of the image like edges, textures, coefficients, etc., in verification procedure to categorize query images as authentic or unauthentic. The image hash is produced by processing the image features using a secret key. The primary goal of this process is that the image features [2, 3] used in hash generation must be sensitive to malicious attacks and insensitive to insignificant modifications [3].

2 Related Work

In recent times, many image hashing techniques have come out for integrity verification. Ahmed et al. [4] proposed a secure hash method using DWT. Initially, image pixels are permuting at block level and then DWT is applied for generating a secure hash. Tang et al. [5] proposed a hash method using DCT coefficients. The block-based dominant DCT coefficients are extracted for generating the image hash. Qin et al. [6] proposed a hash technique in Discrete Fourier transform (DFT) domain. The hash is generated from DFT phase coefficients. Swaminathan et al. [7] proposed a hashing method based on coefficients of discrete transform. Monga and Mhcak [8] proposed image hashing using Non-negative Matrix Factorization (NMF). NMF captures local features of images which are insensitive to geometric attacks and reduces misclassifications. Zhao et al. [9] designed a hash method for identifying forgeries on images like object deletion and addition. This method used Zernike moments and significant texture details of image regions for hash generation. Zhenjun et al. [10] have developed hashing method using pixels entropy. The obtained hash from pixels entropy is unique for malicious manipulations and

shorter in length. Naidu et al. [11] proposed block-based image hash technique using histogram. The image hash is produced based on block histogram bin distribution. The hash is robust to geometric distortions as well as malicious alternations.

3 The Proposed Hashing Technique

To produce invariant hash for normal image processing operations and sensitive hash for malicious image modifications, we present a new image hashing technique. The presented technique includes three stages. They are:

Preprocessing: In this stage, the input image (I_{RGB}) is scaled to $N \times N$ size of pixels using bi-cubic interpolation. The scaled image converts into $L^*a^*b^*$ color image (I_{lab}) [12]. Further, the I_{lab} color image is regularized using Gaussian low pass filter.

Feature Extraction: A level 2, 2D DWT is applied on the L^* component of the I_{lab} color image and features used in hash generation are extracted from the LL_2 approximation sub-band. The approximation coefficients are very robust to normal image processing operations and sensitive for unauthorized attacks. To reduce hash size and for extracting all approximation coefficients from the sub-band image, the proposed technique extracts approximation coefficients by creating circles from the center of the LL_2 sub-band image as shown in Fig. 1a. Let, $P \times P$ is the size of LL_2 sub-band image then its center point is obtained at position $(P/2, P/2)$. The proposed technique creates $d = N/8$ circles. The circles are labeled as $C_1, C_2, \dots, C_i, \dots, C_{d-1}, C_d$, and the radius of these circles are $1, 2, 3, \dots, d-1, d$. The coefficients that appear along the boundary of these circles are considered as features and these are used in hash generation.

The feature coefficient (fc) on circle with radius r and angle θ shown in Fig. 1b is extracted by computing X_{fc} and Y_{fc} coordinates using the determining points on a

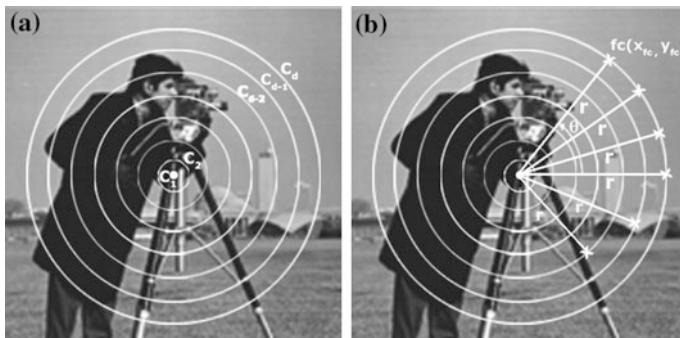


Fig. 1 Feature extraction. **a** A set of *circles* on a cameraman LL_2 approximation sub-band image. **b** Indicating feature coefficients on a boundary of the *circle* using radius r and angle θ

circle procedure [13]. The proposed technique extracts n feature coefficients appearing on the boundary of each circle by varying the angle θ by 1 degree, $0 \leq \theta < 2\pi$. To produce a secure image hash, the LL_2 sub-band coefficients are permuted row and column wise using a key-based random permutation process [14]. The proposed technique extracts $n = 8$ feature coefficients on circle with radius 1, $n = 16$ feature coefficients on circle with radius 2, and so on. The proposed technique creates 64 circles on approximation image. The number of features coefficients extracted on circles 1–64 are $n = \{8, 16, 24, 32, 40, 48, 56, 64, 72, 80, 86, 88, 96, 112, 104, 120, 128, 144, 138, 145, 152, 160, 168, 168, 182, 200, 178, 200, 192, 200, 200, 224, 222, 232, 250, 240, 240, 256, 262, 272, 310, 264, 274, 280, 296, 280, 310, 296, 297, 320, 305, 312, 335, 328, 335, 328, 337, 352, 353, 352, 359, 360, 359, 360\}$. The feature coefficients extracted on circle C_i , $1 \leq i \leq d$ is expressed as $V0_i(j)$, $1 \leq j \leq n$. Later, the final feature vector V_k of length d obtained by computing the mean of coefficients extracted on each circle is described as:

$$V_k = \sum_{j=1}^n V0_i(j)/n, \quad 1 \leq i, k \leq d \quad (1)$$

where k , is the index of the final Feature Vector. The step-by-step process of feature extraction is given in Algorithm 1.

Algorithm 1. Feature Extraction from LL_2 Approximation Sub-band

Algorithm FeatureExtraction(LL_2, d) $// P \times P$ is the size of LL_2 $// (X_c, Y_c)$ is the center point on LL_2 image, $(P/2, P/2)$ $// r$ is the radius and θ is the angle $// fv$ is the vector of features extracted on circle with r $// V_k$ final feature vector used in hash generation $\{$ \quad for $r := 1$ to d do {	\quad for $\theta := 1$ to 2π do { $\quad \quad X_{fc} := X_c + r \times \cos(\theta);$ $\quad \quad Y_{fc} := Y_c + r \times \sin(\theta);$ $\quad \quad fv := LL_2(X_{fc}, Y_{fc});$ $\quad \quad end for$ $\quad \quad V_k = mean(fv);$ $\quad \quad end for$ $\quad }$
---	---

Hash Generation: The binary hash (h) of length $d - 1$ bits is generated from the feature vector V_k using Eq. 2.

$$h_l = \begin{cases} 1 & \text{if } V_{k-1} - V_k \geq 0, 2 \leq k \leq d \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where l , $1 \leq l \leq d - 1$ represents the hash bit index. Finally, the hash bits are permuted using a secret key for producing the secure binary hash.

4 Experimental Analysis

The presented hashing technique is experimented on 1600 original color images selected from various sources [15–17]. The selected images are scaled to 512×512 pixels. A level 2, 2D Daubechies wavelet transform is applied on the L^* component and 64 circles are utilized for features extraction. The 63 bit length hash is produced from extracted features. For verifying the image integrity and to measure the similarity among suspect and original images, the hamming distance is computed using Eq. (3).

$$NHD(\text{hash}_1, \text{hash}_2) = \frac{1}{L} \sum_{m=1}^L |\text{hash}_1(m) - \text{hash}_2(m)| \quad (3)$$

4.1 Experiment Analysis on Standard Image Processing Operations

The robustness experiment on proposed technique is performed with the list of normal image processing operations described in Table 1. Totally, 60 similar copies are produced by modifying images from the first 6 operations listed in Table 1, 8 similar versions are generated by manipulating 10 benchmark original images using the last 2 manipulations listed in Table 1. A total of 96,080 distorted images are used in experiment. Over 1600 images, the average hash distance is estimated on one particular manipulated operation. The achieved results are compared with reported image hashing approach [6]. The results are shown in Figs. 2, 3, 4 and 5.

The proposed technique yields a better performance on scaling and Gaussian distortions. The scaling operation does not change the approximation coefficients that contribute to the proposed technique shows better performance than the method [6]. The results obtained on scaling and Gaussian distortion are shown using Fig. 2a, b, respectively. The performance results under median and wiener filters are presented in Fig. 3a, b, respectively. The proposed technique yields better performance.

Table 1 Standard image processing operations

Tool used	Operation type	Total images
Matlab	Scaling (scale factor: 0.5–0.1–1.5)	10
Matlab	Gaussian noise (variance: 0.01–0.01–0.1)	10
Matlab	Median filter (size of filter: 1–1–10)	10
Matlab	Wiener filter (size of filter: 1–1–10)	10
Matlab	JPEG compression (quality: 10–10–100)	10
Matlab	JPEG 2000 compression (ratio: 1–1–10)	10
Photoshop	Brightness adjustment (adjustment scale: –20–10–20)	04
Photoshop	Contrast adjustment (adjustment scale: –20–10–20)	04

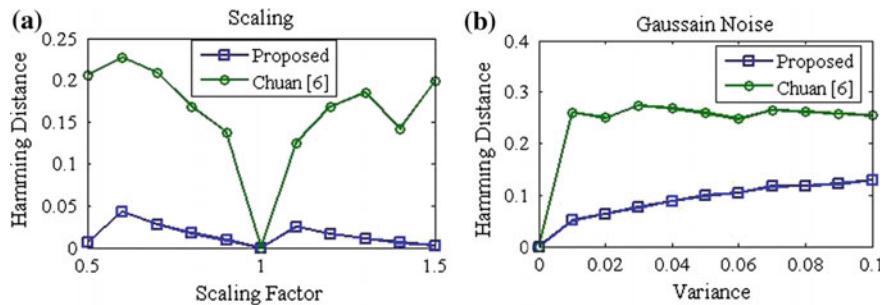


Fig. 2 Performance results on **a** Scaling, **b** Gaussian noise distortion

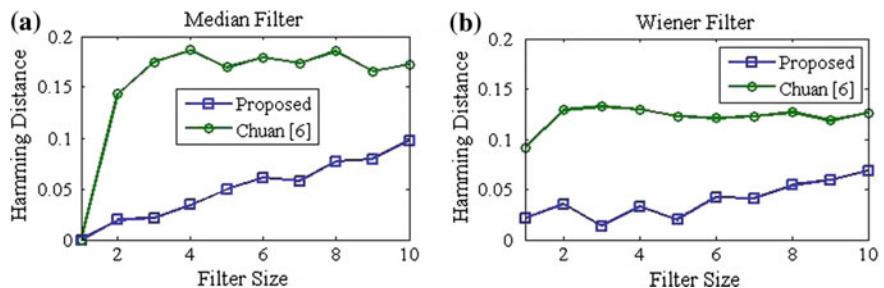


Fig. 3 Performance results on **a** Median filter, **b** Wiener filter

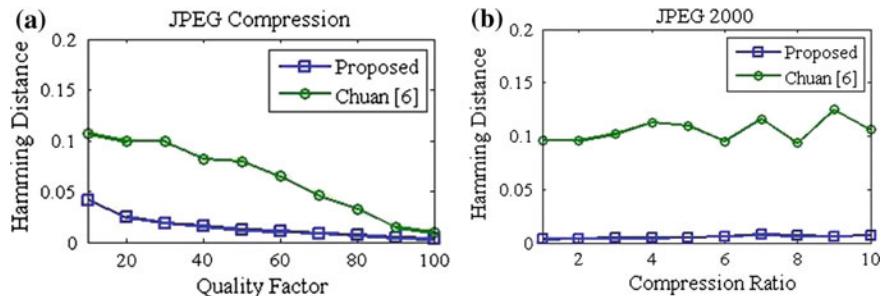


Fig. 4 Performance results on image compression operation using **a** JPEG, **b** JPEG 2000

The results on image compression operation using JPEG are presented in Fig. 4. From Fig. 4a, we noted that the hamming distances are below 0.05 for all quality factors. Under JPEG 2000, the proposed technique has shown hamming distances below 0.02 for compression ratios 1 to 10. The comparison results on brightness and contrast adjustment operations are indicated using Fig. 5. The produced hamming distances are below 0.02 for all brightness and contrast adjustment manipulations.

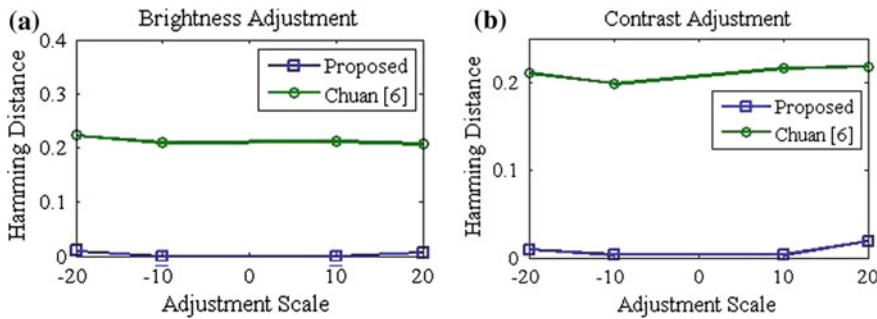


Fig. 5 Performance results of **a** Brightness adjustment. **b** Contrast adjustment

4.2 Experiment Analysis on Malicious Content Alterations

The presented technique is also tested with malicious content alterations including removal of image content and inserting new content on images using cropping operation. The images used in content removal operation are shown in Fig. 6a, b respectively. The image indicated in Fig. 6b is generated by removing the content from the image center shown in Fig. 6a. From each image, four versions of content deletion images are created. Finally, a total of $4 \times 1600 = 6400$ content removal images are used in experiment. The similarity is computed among the actual and content removal images. The performance results on content removal attack are shown in Fig. 7a. For content removal below 80 %, the method [6] yields better performance than the proposed technique. The method [6] extracts more sampling points from the center of the DFT phase image that contributed to produce higher hamming distances.



Fig. 6 Altered images used in content removal and content insertion operations

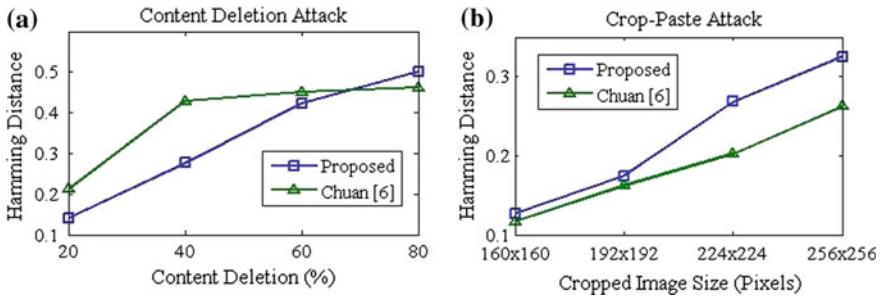


Fig. 7 Performance results on **a** Content removal. **b** Content insertion

The images used in content insertion using cropping operation are indicated in Fig. 6c–h. The two distinct images used under this operation are presented in Fig. 6c, d. Initially, four sub-images of sizes 160×160 , 192×192 , 224×224 , and 256×256 are created using crop operation from Fig. 6d. The created sub-images 160×160 and 224×224 pixels of size are presented in Fig. 6e, f, respectively. Further, the content insertion copies are generated by inserting cropped sub-images on image 6c. The created forgery images are presented in Fig. 6g, h. Totally, $1600 + 4 \times 1600 = 8000$ forgery images are used in this malicious operation. The average hash distance is computed between the actual and forgery images. The obtained results are presented in Fig. 7b. The presented technique on content insertion operation yields better performance than the technique [6].

4.3 Experiment Analysis on Fragility and Robustness

The presented hashing technique robustness is estimated using True Positive Rate (*TPR*). For good robustness, the technique has to produce larger *TPR* values. Similarly, the fragility of the contributed method is computed using False Positive Rate (*FPR*) [11]. For higher fragility, the technique has to yield smaller *FPR* values. To estimate *TPR* and *FPR*, we consider two groups of color images including query and database images. We have selected 1200 query images and divided 1200 query images into 12 groups, each group with 100 images. To categorize query images as unauthentic or authentic, the following list of Threshold (*T*) values is utilized, where $T = [0.1 \ 0.12 \ 0.14 \ 0.16 \ 0.18 \ 0.2]$. The query image is considered as authentic when the hash distance among query image and database images is smaller than *T*, if it is greater than *T*, image is classified as unauthentic. For each group of images, the *TPR* and *FPR* values are computed separately for all *T* values. The plotted *ROC* points with values of *FPR* and *TPR* are indicated using Fig. 8. The plotted *ROC* points appeared in top-left corner on the *ROC* plane. It indicates that the proposed technique is possessing strong robustness and high discriminative capability. The *TPR* and *FPR* values over 1200 images are shown in Table 2. For

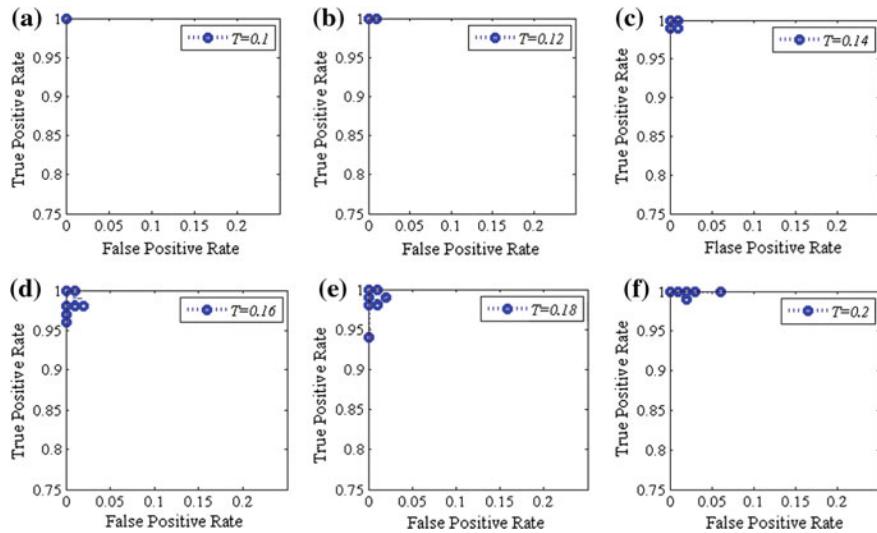


Fig. 8 ROC plots using FPR and TPR values at **a** $T = 0.1$. **b** $T = 0.12$. **c** $T = 0.14$. **d** $T = 0.16$. **e** $T = 0.18$ and **f** $T = 0.2$

Table 2 Performance results on image authentication

Threshold (T)	0.1	0.12	0.14	0.16	0.18	0.2
<i>TPR</i>	1	1	0.9983	0.9858	0.985	0.9991
<i>FPR</i>	0	0.0008	0.0025	0.0041	0.0041	0.015

$T = 0.2$, the proposed technique classified only 12 images as unauthentic out of 1200 authentic images and 18 unauthentic images are misclassified as authentic. The TPR and FPR values shown in Table 2 are almost equal to 1 and 0 respectively, except TPR values at $T = 0.16$ and $T = 0.18$.

4.4 Secure and Key Dependent Hash

In order to show, the image hash of the proposed technique is key dependent, we have generated a hash of cameraman image using a secret key. Later, we have chosen 100 wrong keys and produced 100 hashes of cameraman image. The hamming distance is computed between original hash of the cameraman image and hashes produced using 100 wrong keys. The obtained results are shown in Fig. 9a. We also presented results on 100 different images with the same secret key using Fig. 9b. The hash distance variations present in Fig. 9 show that the produced hash is key dependent. The proposed hashing technique produced a short length hash when compared to reported image hash methods listed in Table 3.

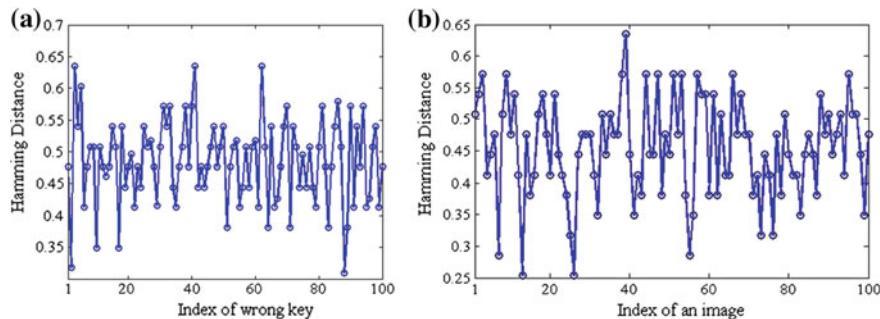


Fig. 9 Key dependent hash on **a** Cameraman image with different keys and **b** Different images with same secret key

Table 3 Hash lengths of various image hashing methods

Method	Proposed	[5]	[7]	[6]	[11]	[9]
Hash length in bits	63	64	420	444	448	560

5 Conclusion

We presented a robust image hashing technique for integrity verification and to authenticate digital images using DWT approximation features. The hash of the proposed technique is invariant to various standard content preserving operations and variant to image rotation operation. The proposed technique is also robust to content removal and insertion manipulations. The evaluation on fragility and robustness experiment shows that, the presented technique has high discriminative and robustness capabilities. The hash of the proposed technique is shorter in length and key dependent.

References

1. S. M Saad, “Design of a Robust and Secure Digital Signature Scheme for Image Authentication over Wireless Channels,” *IET Information Security*, vol. 3(1), pp. 1–8, 2009.
2. Adil Haouzia, Rita Noumeir, “Methods for Image Authentication: A Survey,” *Journal of Multimedia Tools Applications*, vol. 39(1) pp. 1–4, 2008.
3. Shui-Hua Han, Chao-Hsien Chu, “Content-based Image Authentication: Current Status, Issues and Challenges,” *Int. Journal of Information Security*, vol. 9(1), pp. 19–32, 2010.
4. F Ahmed, M. Y. Siyal, V. Uddin Abbas, “A Secure and Robust Hash Based Scheme for Image Authentication,” *Journal of Signal Processing*, vol. 90(5), pp 1456–1470, 2010.
5. Zhenjun Tang et al., “Robust Image Hashing with Dominant DCT Coefficients,” *Int. Journal for Light and Electron Optik*, vol. 125(18), pp. 5102–5107, 2014.

6. Chuan Qin et al., “Robust Image Hashing using Non-uniform Sampling in Discrete Fourier Domain,” *Journal of Digital Signal Processing*, vol. 23(2), pp. 578–585, 2013.
7. Swaminathan A, Yiniyan Mao, Min Wu, “Robust and Secure Image Hashing,” *IEEE Transactions on Information Forensics and Security*, vol. 2(1), pp. 215–230, 2006.
8. V. Monga, M.K. Mhcak, “Robust and Secure Image Hashing via Non-Negative Matrix Factorization,” *IEEE Tr. on Information Forensics and Security*, vol. 2(3), pp. 376–390, 2007.
9. Yan Zhao, Shuzhong Wang, Xinpeng Zhang, and Heng Yao, “ Robust Hashing for Image Authentication using Zernike Moments and Local Features,” *IEEE Tr. on Information Forensics and Security*, vol. 8(1), pp. 55–63, 2013.
10. Zhenjun Tang et al., “Robust Image Hashing using Ring-based Entropies,” *Journal of Signal Processing*, vol. 93(7), pp. 2061–2069, 2013.
11. Lokanadham Naidu et al., “Robust Hash Generation Technique for Content Based Image Authentication using Histogram,” *Multimedia Tools and Applications*, vol. 74(9), 2015.
12. Patel Janak kumar et al., “Color Image Segmentation for Medical Images using $L^*a^*b^*$ Color Space,” *Jr. of Electronics and Communication Engineering*, vol. 1(2), pp. 24–45, 2012.
13. Joey Lott et al., “ActionScript 3.0 Cookbook,” *O'Reilly*, pp. 98–100, 2006.
14. Black, Paul E. (2005), Fisher-Yates Shuffle, “Dictionary of Algorithms and Data Structures,” *National Institute of Standards and Technology*, Retrieved 2007.
15. <http://sipi.usc.edu/database/>.
16. <http://decsai.ugr.es/cvg/dbimagenes/>.
17. <http://tabby.vision.mcgill.ca/>.

Robust Parametric Twin Support Vector Machine and Its Application in Human Activity Recognition

Reshma Khemchandani and Sweta Sharma

Abstract This paper proposes a novel and Robust Parametric Twin Support Vector Machine (RPTWSVM) classifier to deal with the heteroscedastic noise present in the human activity recognition framework. Unlike Par-v-SVM, RPTWSVM proposes two optimization problems where each one of them deals with the structural information of the corresponding class in order to control the effect of heteroscedastic noise on the generalization ability of the classifier. Further, the hyperplanes so obtained adjust themselves in order to maximize the parametric insensitive margin. The efficacy of the proposed framework has been evaluated on standard UCI benchmark datasets. Moreover, we investigate the performance of RPTWSVM on human activity recognition problem. The effectiveness and practicability of the proposed algorithm have been supported with the help of experimental results.

Keywords Human activity recognition · Twin support vector machines · Heteroscedastic noise · Machine learning

1 Introduction

Human Activity Recognition is an interesting field of research in the domain of Computer Vision. From the viewpoint of computer vision, the recognition of activity is to match the observation (e.g., video) with previously defined patterns and then assign it a label, i.e., activity type [1]. The challenges include the endless vocabulary of the activity classes, varying illumination, occlusion, and intraclass differences. In addition to this, the intraclass noise contributes to the complexity of the activity recognition problem [2].

R. Khemchandani (✉) · S. Sharma

Faculty of Mathematics and Computer Science, Department of Computer Science,
South Asian University, New Delhi, India
e-mail: reshma.khemchandani@sau.ac.in

S. Sharma
e-mail: sharma.sweta.2007@gmail.com

A human activity is represented by set of features derived from the corresponding video sequence. An activity recognition problem is divided into two phases: feature extraction from the video sequence followed by classification and labeling. Recently, global space-time features representation and support vector machine (SVM) for human activity recognition have drawn wide attention in the research community [3, 4]. An activity is represented via a motion-context descriptor obtained using histogram of action sequences and optic flow values. Then, classification models are exploited to label a video sequence to the corresponding activity class.

In the recent past, Jayadeva et al. [5] proposed Twin Support Vector Machine (TWSVM) which seeks two nonparallel hyperplanes such that each hyperplane passes through one of the two classes and is atleast one unit distance away from the samples of other class. This model leads to lower computational complexity and better generalization ability when compared with SVMs. TWSVM solves two QPPs of smaller size as compared to SVM where a single large size QPP is solved and hence TWSVM training time is improved by approximately four times when compared with SVM [5]. TWSVM has also been effectively applied to human activity recognition problem [6].

In general, the classical SVM and its extensions anticipate the noise level on training data to be uniform throughout the domain, or at least, the functional dependency is assumed to be known in advance [7]. However, this uniform noise assumption is rarely satisfied in real-life problem including activity recognition framework where the noise in the features is largely dependent on the feature values. Working on the lines of v -support vector machine (v -SVM) [9], Hao proposed parametric margin v -support vector machine (Par- v -SVM) in order to deal with heteroscedastic noise problem in pattern classification problem. Unlike v -SVM, Par- v -SVM seeks to maximize a parametric margin between a pair of nonparallel margin hyperplanes. Wang et al. [7] enhanced the idea of parametric margin to proximal support vector machine which they called Proximal Parametric Support vector classifier (PPSVC). PPSVC maximizes proximal parametric margin between the proximal hyperplanes, unlike Par- v -SVM which maximizes the parametric margin between two nonparallel hyperplanes.

Working on the idea of parametric margin, Peng [10] proposed twin parametric margin SVM (TPSVM) in which author minimized the sum of projection values of negative (positive) samples on the positive (negative) hyperplane with parameter v . Apart from this, TPMSVM constrained the projection values of positive (negative) points on the positive (negative) parametric margin hyperplane to be at least zero. This gives a more flexible parametric classifier. However, TPSVM does not consider the prior structural information inherent in the data to deal with the effect of noise.

This paper introduces a novel parametric TWSVM classifier based on Par- v -SVM and TWSVM termed as Robust parametric Twin Support Vector Machine (RPTWSVM). It seeks a pair of nonparallel parametric hyperplanes, determined using a pair of parametric insensitive proximal functions, by solving two-smaller sized QPPs resulting into a robust heteroscedastic noise handling structure. The goal of our proposed model is to adjust the parametric insensitive zone of arbitrary shape and size so as to capture the noise present in each of the class more accurately.

Computational efficacy of par- ν -SVM, TPSVM, and RPTWSVM in terms of activity class prediction for standard Weizmann dataset [11] along with several UCI [12] benchmark datasets have been reported.

In this paper, let us consider a data set D having total m number of data points in which m_1 data points belong to class +1 and are represented by matrix A and m_2 data points belongs to class -1 and are represented by matrix B . Therefore, the sizes of matrices D , A , and B are $(m \times n)$, $(m_1 \times n)$, and $(m_2 \times n)$, respectively, where n is the dimension of feature space.

Section wise the paper is organized as follows. Section 2 presents the brief discussion of the related work. Section 3 introduces the proposed work in activity recognition framework. Experimental results are presented in Sect. 4. Finally, Sect. 5 summarizes our contributions.

2 Review of Parametric Support Vector Machines

In this section, we briefly dwell upon the par- ν -SVM and TPSVM which forms the base of our proposed work.

2.1 Parametric- ν -Support Vector Machine

Hao [8] replaced the ρ in original ν -SVC [9] algorithm with a parametric margin model $g(x) = (z^T x + d)$. Thus, Par- ν -SVM classifier finds parametric margins of two classes using two nonparallel hyperplane given by $w^T x + b = \pm(z^T x + d)$, and the classifying hyperplane is given by the mean of these parametric hyperplane. The classifying hyperplane $w^T x + b = 0$ is obtained via solving the following problem:

$$\begin{aligned} & \underset{w, b, z, d, \xi}{\text{Min}} \quad \frac{1}{2} w^T w + C(-\nu(\frac{1}{2} z^T z + d) + \frac{1}{m} e^T \xi) \\ & \text{subject to } D(Aw + eb) \geq Az + ed - \xi, \xi \geq 0, \end{aligned} \quad (1)$$

where ξ represents the error variable.

Rather than solving the primal problem, author in [9] solved the corresponding dual problem. Further, with the help of experimental results author in [8] have shown that Par- ν -SVM turns out to be a good parametric insensitive model and is effective in handling the heteroscedastic noise [8].

The class of new test sample x ($i = +1$ or -1) is determined by $\text{sign}\{w^T x + b\}$.

2.2 Twin Parametric Support Vector Machine (TPSVM)

TPSVM [10] finds a pair of hyperplanes given by $f_1(x) = w_1^T x + b_1 = 0$ and $f_2(x) = w_2^T x + b_2 = 0$ that determines the positive and negative parametric margin of the two classes, respectively. The linear TPSVM considers the following pair of optimization problems to obtain the corresponding hyperplanes:

$$\begin{aligned} \text{Min}_{w_1, b_1, \xi} \quad & \frac{1}{2} w_1^T w_1 + \frac{\nu}{l_2} \sum_{i \in B} (w_1^T x_i + b_1) + \frac{c}{m_1} \sum_{i \in A} \xi_i \\ \text{subject to} \quad & w_1^T x_i + b_1 \geq 0 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, m_1 \end{aligned} \quad (2)$$

$$\begin{aligned} \text{Min}_{w_2, b_2, \xi} \quad & \frac{1}{2} w_2^T w_2 + \frac{\nu}{l_1} \sum_{i \in A} (w_2^T x_i + b_2) + \frac{c}{m_2} \sum_{i \in B} \xi_i \\ \text{subject to} \quad & w_2^T x_i + b_2 \geq 0 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, m_2, \end{aligned} \quad (3)$$

where m_1 and m_2 are number of patterns in class +1 and class -1, respectively, and ξ represents the error variable..

After solving quadratic programming problems (2) and (3), the classifier for TPSVM is given as follows:

$$f(x) = \text{sign}[(\hat{w}_1 + \hat{w}_2)^T x + (\hat{b}_1 + \hat{b}_2)] \quad (4)$$

where $\hat{w}_i = \frac{w_i}{||w_i||}$ and $\hat{b}_j = \frac{b_j}{||w_i||}$ for $i = 1, 2$.

On the similar lines, TPSVM has been extended to nonlinear version as well.

3 Robust Parametric Twin Support Vector Machine

In this Section, we present a robust and efficient learning algorithm working on the lines of Par- ν -SVM and TWSVM termed as the robust parametric twin support vector machine (RPTWSVM). Unlike Par- ν -SVM which solves a single large-sized QPP to obtain the classifying hyperplane, RPTWSVM is faster and more efficient as it solves two smaller sized QPPs. Similar to TWSVM, RPTWSVM also derives a pair of nonparallel hyperplanes around the data points through two QPPs. However, RPTWSVM does not make any fixed noise assumption like TWSVM, and thus automatically adjust the hyperplanes to maximize the parametric insensitive region in order to deal with the effect of noise and hence it is more robust.

3.1 Linear Case

RPTWSVM determines two hyperplanes in R^n given as follows:

$$f_1(x) - g_1(x) = 0 \Rightarrow w_1^T x + b_1 = (z_1^T x + d_1)$$

and,

$$f_2(x) + g_2(x) = 0 \Rightarrow w_2^T x + b_2 = -(z_2^T x + d_2)$$

Each determines the (lower/upper) bound for the corresponding class. Here, $f_1(x) - g_1(x) = 0$ determines the positive parametric margin hyperplane, and the negative parametric margin hyperplane is determined by $f_2(x) + g_2(x) = 0$.

The following pair of optimization problems are solved in order to determine the two hyperplanes:

(RPTWSVM 1)

$$\text{Min}_{w_1, b_1, z_1, d_1, \xi_1} \frac{c_1}{2} (||w_1||^2 + b_1^2) + c_2 (-\frac{\nu}{2} (||z_1||^2 + d_1^2) + e^T \xi_1) + \frac{c_3}{2} ||Bw_1 + eb_1 + e||^2$$

subject to

$$\begin{aligned} (Aw_1 + eb_1) &\geq (Az_1 + ed_1) - \xi_1, \\ \xi_1 &\geq 0. \end{aligned} \tag{5}$$

(RPTWSVM 2)

$$\text{Min}_{w_2, b_2, z_2, d_2, \xi_2} \frac{c_1}{2} (||w_2||^2 + b_2^2) + c_2 (-\frac{\nu}{2} (||z_2||^2 + d_2^2) + e^T \xi_2) + \frac{c_3}{2} ||Bw_2 + eb_2 - e||^2$$

subject to

$$\begin{aligned} -(Bw_2 + eb_2) &\geq (Bz_2 + ed_2) - \xi_2, \\ \xi_2 &\geq 0, \end{aligned} \tag{6}$$

where $c_1, c_2, c_3 \geq 0$ are regularization parameters, e represents vectors of ones of appropriate dimension, ν determine the penalty weights and ξ_1 and ξ_2 represent the error vector corresponding to classes A and B , respectively.

Unlike TPSVM, where author seeks for the parametric hyperplanes given by $f_1(x)$ and $f_2(x)$, respectively, the RPTWSVM seeks two parametric margin hyperplanes given by $f_j(x) \pm g_j(x)$, where $j = 1, 2$, one for each class, and classifies points depending upon its proximity to the two classes. The first term in the objective function of (5) or (6) takes care of structural risk minimization of the data and controls the model complexity. For each point x_i , an error up to $(z_j^T x_i + d_j)$ is allowed. Everything beyond this is captured in error variable ξ_i , which is further penalized in the objective function via a regularization parameter c_2 . The size of the parametric insensitive zone is controlled by $\frac{1}{2} (||z_j||^2 + d_j^2)$ which is regulated using constant ν . The constraint ensures that all the samples of the positive class lie beyond the parametric margin hyperplane.

Consider $[w \ b] = u$, $[z \ d] = v$, $[A \ e] = H$, $[B \ e] = G$, then Eq. (5) can be rewritten as

$$\begin{aligned} & \underset{u, v, \xi_1}{\text{Min}} \frac{c_1}{2}(u^T u) + c_2(-v(v^T v) + e^T \xi_1) + \frac{c_3}{2} ||Gu + e||^2 \\ & \text{subject to} \quad Hu \geq Hv - \xi_1, \\ & \quad \xi_1 \geq 0. \end{aligned} \quad (7)$$

Considering the Lagrangian function corresponding to QPP (7) and using the KKT necessary and sufficient conditions, the dual of QPP (5) is obtained as follows:

$$\begin{aligned} & \underset{\alpha}{\text{Max}} -\frac{1}{2}\alpha^T H(c_1 I + c_3 G^T G)^{-1} H^T \alpha + \frac{c_3}{2} e^T G(c_1 I + c_3 G^T G)^{-1} H^T \alpha + \frac{1}{2c_2 v} \alpha^T H H^T \alpha \\ & \text{subject to} \quad 0 \leq \alpha \leq C, \end{aligned} \quad (8)$$

where α is the vector of Lagrange multipliers.

Similarly, the solution of (6) is obtained via solving the following optimization problem:

$$\begin{aligned} & \underset{\beta}{\text{Max}} -\frac{1}{2}\beta^T G(c_1 I + c_3 H^T H)^{-1} G^T \beta + \frac{c_3}{2} e^T H(c_1 I + c_3 H^T H)^{-1} G^T \beta + \frac{1}{2c_2 v} \beta^T G G^T \beta \\ & \text{subject to} \quad 0 \leq \beta \leq C. \end{aligned} \quad (9)$$

A new test sample \hat{x} is assigned a class label depending upon the following decision function:

$$\underset{i=1,2}{\text{argMin}} \left\{ \frac{|(w_1 + u_1)^T \hat{x} + (b_1 + d_1)|}{||w_1 + u_1||_2}, \frac{|(w_2 - u_2)^T \hat{x} + (b_2 - d_2)|}{||w_2 - u_2||_2} \right\}. \quad (10)$$

Similarly, the proposed method can be extended to nonlinear case using kernel methods.

4 Experimental Results

All the experiments have been carried out in MATLAB version 8.0 under Microsoft Windows environment on a machine with 3.40 GHz CPU under 16 GB RAM.

4.1 Benchmark Datasets

To prove the efficacy of proposed work, we performed classification experiments on UCI machine learning datasets [12]. The features in all the datasets are normalized to the range $[0, 1]$ before training the classifier. In our simulations, we performed experiments with Gaussian kernel in order to obtain the classifiers.

Table 1 Classification results with Gaussian Kernel on UCI datasets

Dataset	Par- ν -SVM	TPSVM	RPTWSVM
	Mean accuracy Learning time (in sec)		
Heart-Statlog (270 × 13)	81.85	80.00	83.7
	0.43	0.67	0.49
WPBC (198 × 34)	81.32	80.82	81.32
	0.28	0.44	0.19
Haberman (306 × 3)	72.52	76.15	74.84
	0.52	0.89	0.31
Echocardiogram (131 × 10)	84.78	86.32	87.03
	0.17	0.19	0.10
Spect (267 × 44)	79.00	79.42	79.74
	0.39	0.80	0.28

We have applied grid search method [13] to tune the parameters ν , c_i ($i = 1$ to 4) and kernel parameter σ . For each dataset, a validation set comprising of 10 % randomly selected samples from the dataset is used. For this work, we have selected values of c_1 , c_2 , and c_3 from the range 0 to 1. The parameters ν are tuned in the range {0.1 to 1}. The value of parameter C was tuned in the range of 10^i , where $i = \{1, 2, 3, 4, 5, 6\}$. Mean of classification accuracy is determined using tenfold cross validation [14].

Classification Results We have examined the performance of RPTWSVM against Par- ν -SVM and TPSVM. The experiments are conducted with all the algorithms using tenfold cross validation [14] and the mean classification accuracy is reported.

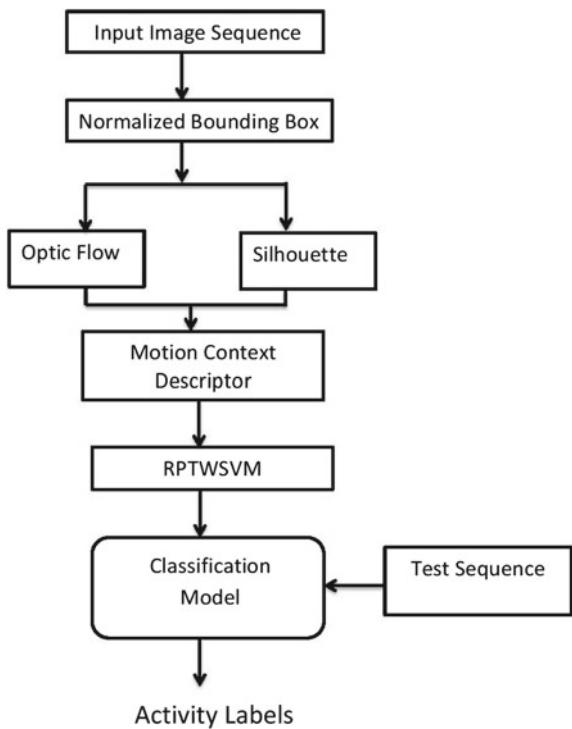
The classification results using Gaussian kernel for Par- ν -TWSVM, TPSVM and RPTWSVM are reported in Table 1. The bold values indicate best result and the mean of accuracy (in %) across tenfolds. The table demonstrates that RPTWSVM mostly outperforms the other methods in terms of generalization ability. In this case, mean accuracy of RPTWSVM, for all the datasets, is 81.32 % as compared to 79.89 and 80.48 % for Par- ν -TWSVM and TPSVM, respectively.

4.2 Application to Human Activity Recognition

Figure 1 depicts the flow chart which briefly explains the feature extraction and activity recognition steps involved in the proposed activity recognition framework.

Motion-Context Descriptor Following Tran et al. [3], we have used global features suggested for representing the activity sequence. The descriptor is a robust and efficient in representing the activity sequence. The first step deals with the extraction of image sequence from the video which is further used for calculating the normalized

Fig. 1 A flowchart depicting the proposed activity recognition framework



bounding box around the actor using the silhouettes obtained with background subtraction. Then pyramidal implementation of Lucas Kanade algorithm [15] is used to obtain optic flow values for each image. The optic flow values are then divided into horizontal and vertical components given by F_x and F_y , respectively. These values are then smoothed using median filter to lessen the effect of noise. The silhouette gives us the third set of values that represents the bounding coordinates of the binary image in four directions, namely, left, right, top, and bottom. The normalized bounding box obtained from each image is divided into 2×2 windows and then each window is divided into 18 pie slices covering 20° each. These pie slices are further histogrammed to obtain the component values. The component values are integrated to obtain resultant 72 ($2 \times 2 \times 18$)-dimensional descriptor. By concatenating the values from all three components, we obtain a 216-dimensional frame descriptor [3].

The feature descriptor obtained gives a very rich representation of the local feature of the image. Adding to this, to make a context of time we concatenate three blocks of five frames representing the past, current, and future to form a 15-frame block. Further the frame descriptor thus obtained corresponding to each block are stacked in order to form a 1080-dimensional vector. This 1080-dimensional vector is then projected onto 70 dimensions using principal component analysis.



Fig. 2 Weizmann Dataset

The resulting motion-context descriptor is joined with the current frame descriptor to form the final 286-dimensional motion-context descriptor [3].

Weizmann Dataset Weizmann dataset [11] consists of 93 (180×144 pixels) videos in low resolution where nine actors are performing 10 activities which include walk (walking), run (running), jump (jumping), side (striding sideways), bend (bending), wave1 (waving with one hand), wave2 (waving with both hands), etc. Some example actions belonging to Weizmann dataset are shown in Fig. 2.

Evaluation Methodology

1. **Leave 1 Actor Out (L1AO):** This methodology removes all the activity sequences performed by an actor from the training set and uses them as the testing set.
2. **Leave One Sequence Out (L1SO):** This methodology removes one activity sequence at a time from the training set.

The activity label of an activity video sequence was assigned based on the majority of the labels assigned to each video frame of that particular activity sequence.

Results For Human Activity Recognition, we choose radial basis function (RBF) kernel for our classifiers because of nonlinear relationship between action classes and histogrammed feature obtained in the descriptor. Optimal values of parameter c_1, c_2, c_3, C and ν , and kernel parameter σ were obtained using grid search with a set comprising 10 % of the frames from each video sequence.

In order to implement activity recognition problem as a binary classification problem, we picked up two activity classes at a time, eg., bend versus wave, etc., and used them to evaluate our results. We performed the activity recognition task using RPTWSVM, Par- ν -SVM and TPSVM for both the evaluation methodologies.

Table 2 Prediction accuracy on Weizmann dataset using L1SO

Dataset	Par- v -SVM	TPSVM	RPTWSVM
	Prediction accuracy learning time (in sec)		
Bend versus Wave2 (1245 × 214)	73.68	72.22	100
	33.32	14.59	16.52
Wave1 versus Wave2 (1259 × 214)	73.68	72.22	100
	38.78	13.71	16.59
Skip versus Jump (925 × 214)	73.68	72.22	89
	36.19	9.31	6.41

Table 3 Prediction accuracy on Weizmann dataset using L1AO

Dataset	Par- v -SVM	TPSVM	RPTWSVM
	Prediction accuracy learning time		
Bend versus Wave2 (1245 × 214)	83.33	100	100
	19.69	13.71	12.86
Wave1 versus Wave2 (1259 × 214)	72.22	100	100
	16.94	13.71	13.34
Skip versus Jump (925 × 214)	77.77	61.11	78.94
	13.24	9.86	8.37

The results have been summarized in Tables 2 and 3. The results in Table 2 show that RPTWSVM performs comparable to other approaches and for the highly confused classes of skip versus jump, RPTWSVM performs better. In case of leave one actor out (L1AO), the test sequence was composed of two activity sequences from each actor. The results are comparable for all the approaches. However, RPTWSVM performs better in terms of training time.

5 Conclusion

In this paper, we have presented a novel robust parametric twin support vector machine (RPTWSVM) for binary classification problems. RPTWSVM successfully extend the parametric margin concept to the twin support vector machine framework leading to better generalization ability and faster training in comparison to Par- v -SVM. The hyperplanes obtained in RPTWSVM are more flexible and automatically adjust itself with the objective to lessen the effect of heteroscedastic noise in generalization ability of the classifier.

Experimental results on the benchmark UCI datasets proves the efficacy of our proposed framework in terms of generalization ability when compared with TPSVM and Par- v -SVM. Further, we investigated the performance of RPTWSVM on human

activity recognition problem, which also validated the effectiveness and practicability of the proposed framework.

In this paper, we have considered dataset, where only one actor is performing a single action. However, it would be interesting to explore the application of the proposed approach for other video sequences, where more than one actor is performing more than one activity. Moreover, the proposed approach can be modified in the least squares sense so as to further reduce the training time complexity.

References

1. Cheng, Guangchun, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, Bill P. Buckles Advances in Human Action Recognition: A Survey. arXiv preprint [arXiv:1501.05964](https://arxiv.org/abs/1501.05964) (2015).
2. Aggarwal, J. K. and L. Xia, Human activity recognition from 3D data: a review, Pattern Recognition Letters, vol. 48, pp. 70–80, (2014).
3. Tran, Du, and Alexander Sorokin. Human activity recognition with metric learning. Computer VisionECCV 2008. Springer Berlin Heidelberg, 548–561, (2008).
4. Manosha Chathuramali, K. G., and Ranga Rodrigo. Faster human activity recognition with SVM. Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on. IEEE, (2012).
5. Jayadeva, Khemchandani, R., and Suresh Chandra. Twin support vector machines for pattern classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29.5: 905–910, (2007).
6. Nasiri, Jalal A., Nasrollah Moghadam Charkari, and Kourosh Mozafari. Energy-based model of least squares twin Support Vector Machines for human action recognition. Signal Processing 104: 248–257 (2014).
7. Wang, Zhen, Yuan-Hai Shao, and Tie-Ru Wu. Proximal parametric-margin support vector classifier and its applications. Neural Computing and Applications 24.3-4: 755–764, (2014).
8. Hao, Pei-Yi. New support vector algorithms with parametric insensitive/margin model. Neural Networks 23.1: 60–73, (2010).
9. Schlkopf, Bernhard, et al. New support vector algorithms. Neural computation 12.5: 1207–1245, (2000).
10. Peng, Xinjun. TPMSVM: a novel twin parametric-margin support vector machine for pattern recognition. Pattern Recognition 44.10: 2678–2692, (2011).
11. Blank, Moshe, et al. Actions as space-time shapes. Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Vol. 2. IEEE, (2005).
12. Blake, Catherine, and Christopher J. Merz. UCI Repository of machine learning databases, (1998).
13. Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification, (2003).
14. Duda, Richard O., Peter E. Hart, and David G. Stork, Pattern classification, John Wiley & Sons, (2012).
15. Lucas, Bruce D., and Takeo Kanade. An iterative image registration technique with an application to stereo vision. IJCAI. Vol. 81 (1981).

Separating Indic Scripts with ‘matra’—A Precursor to Script Identification in Multi-script Documents

**Sk.Md. Obaidullah, Chitrita Goswami, K.C. Santosh,
Chayan Halder, Nibaran Das and Kaushik Roy**

Abstract Here, we present a new technique for separating Indic scripts based on matra (or shirorekha), where an optimized fractal geometry analysis (FGA) is used as the sole pertinent feature. Separating those scripts having matra from those which do not have one, can be used as a precursor to ease the subsequent script identification process. In our work, we consider two matra-based scripts namely Bangla and Devanagari as positive samples, and the counter samples are obtained from two different scripts namely Roman and Urdu. Altogether, we took 1204 document images with a distribution of 525 matra-based (325 Bangla and 200 Devanagari) and 679 without matra-based (370 Roman and 309 Urdu) scripts. For experimentation, we have used three different classifiers: multilayer perceptron (MLP), random forest (RF), and BayesNet (BN), with the target of selecting the best performer. From a series of test, we achieved an average accuracy of 96.44 % from MLP classifier.

Sk.Md. Obaidullah (✉) · C. Goswami
Department of Computer Science & Engineering, Aliah University,
Kolkata, West Bengal, India
e-mail: sk.obaidullah@gmail.com

C. Goswami
e-mail: chtrgswm@gmail.com

K.C. Santosh
Department of Computer Science, The University of South Dakota,
Vermillion, SD, USA
e-mail: santosh.kc@usd.edu

C. Halder · K. Roy
Department of Computer Science, West Bengal State University,
Kolkata, West Bengal, India
e-mail: chayan.halderz@gmail.com

K. Roy
e-mail: kaushik.mrg@gmail.com

N. Das
Department of Computer Science & Engineering, Jadavpur University,
Kolkata, West Bengal, India
e-mail: nibaran@gmail.com

Keywords Handwritten script identification • ‘matra’ based script • Topological feature • Fractal geometry analysis

1 Introduction

To build a ‘paperless’ world, physical documents need to be first converted into digital form, then to textual version. This technique of converting scanned images of handwritten, typewritten, or printed text into machine-encoded form is known as optical character recognition. Increasing efforts from the research community can be found in the literature of this domain. Script identification is an important aspect in India for multilingual document processing as there are 13 official scripts (including Roman) and 23 different languages (including English). In the literature, the need of automatic script identification systems has already been pointed out [1]. Here, we present a new technique to identify two different script types, i.e., with ‘matra’ (like Bangla, Devanagari etc.) and without ‘matra’ (like Roman, Urdu, etc.). Though this paper does not represent a standalone work on automatic Indic script identification, but it guarantees to ease the subsequent script identification process by separating these ‘matra’ based scripts from their counter part.

1.1 Related Works and Our Contribution

State-of-the-art works on script identification based on Indic and non-Indic scripts are reported in the literature since last decade. These works can be classified into different categories namely document-level, line-level, block-level, word-level depending on the type of input document images considered. Character-level script identification is not very common because in general multi-script documents does not occurs at character-level. Among the available works on Indic scripts, Hochberg et al. [2] did identification on a few Indic scripts composed together with non-Indic scripts at document-level, considering features based on connected components. Zhu et al. [3] did work to identify some Indic and non-Indic scripts using different image descriptor with scale, rotation, translation invariant shape codebook-based features. They had used multi-class SVM to classify at document-level. A technique based on texture feature (rotation invariant) and multichannel GLCM to identify Bangla, Roman, Devanagari, and Telugu scripts was proposed by Singhal et al. [4]. Hangarge et al. [5] proposed a scheme to identify Devanagari, Roman, and Urdu scripts using KNN classifier and features like visual discrimination, pixel/stroke density, morphological transformation, etc. Rajput et al. [6] suggested a technique based on DCT and wavelet-based feature with KNN classifier to identify eight Indic scripts. Sarkar et al. [7] proposed a word-level script identification technique using foreground background translation-based approach. Recently, Hangarge et al. [8] suggested a directional DCT-based approach to identify few popular handwritten Indic scripts. Rani

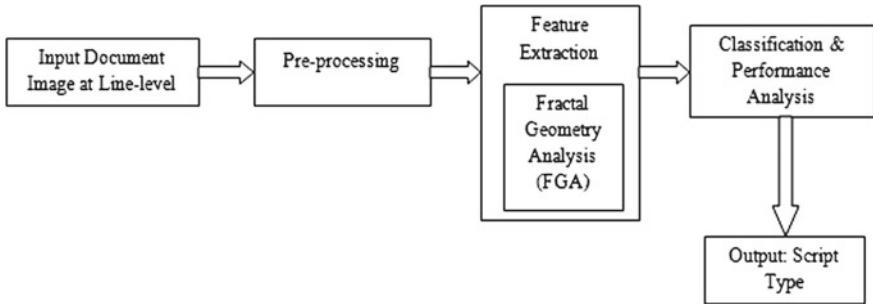


Fig. 1 Block diagram of the proposed work

et al. [9] proposed a character-level script identification technique using gabor filter and gradient-based features for Gurumukhi and Roman scripts.

In this paper, we propose a technique to identify scripts with ‘matra’ from others without ‘matra.’ The concept of fractal analysis (FGA) was first introduced in the field of script identification by one of the authors of the present work [10]. But in that work fractal feature was used in combination with other features making a multidimensional feature set. But novelty of the present work is due to optimization of the feature set (one-dimensional in our case) and faster algorithm as it extracts features directly from the topological distribution of the pixels (presence or absence of ‘matra’). Figure 1 shows block diagram of the present work. Initially line-level document images are feed to the system. Preprocessing is done using our existing technique [11]. Then one-dimensional feature is extracted. Present work considers two types of Indic scripts namely with ‘matra’, e.g., Bangla, Devanagari and without ‘matra’ namely Roman and Urdu. Here we apply the concept of FGA to identify the presence or absence of ‘matra.’ Then best performer classifier is chosen among the three different classifiers. Finally, the script type is produced as an output.

1.2 Our Scope

We have already stated that there are 13 scripts (including Roman) in India and few of the major scripts can be classified by a topological property known as ‘matra.’ A ‘matra’ is a horizontal line present on the upper part of scripts like Bangla, Devanagari, etc. When user starts writing with pen or pencil he/she draws the line at the top and then starts writing the graphemes below this line with some touching component in between. Figure 2 shows an example where a Bangla script word is shown which contains ‘matra’ but a Roman script word contains no ‘matra’. Though there are many scripts with ‘matra’ and without ‘matra’ available in India, for present experiment two majorly used ‘matra’ based scripts namely Bangla and Devanagari and two scripts without ‘matra’ namely Roman and Urdu are considered. These four scripts were chosen observing their wide demographic distribution in India.

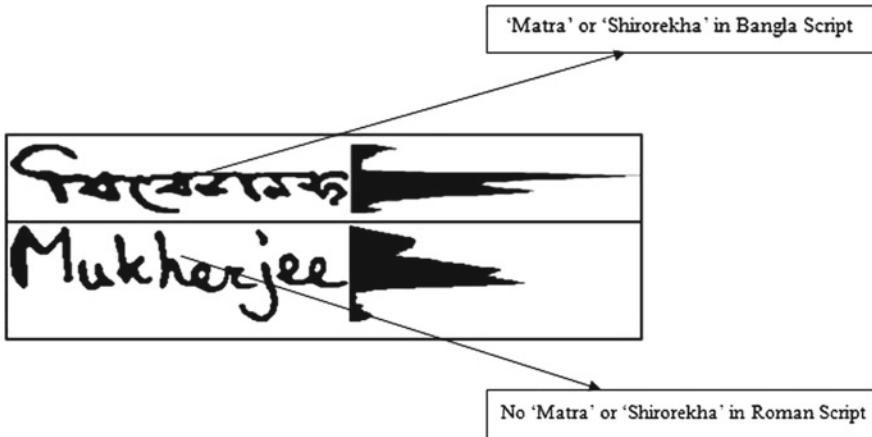


Fig. 2 Presence of ‘matra’ or ‘matra’ in Bangla script, the same is absent in Roman [11]

The paper has been organized in the following manner: In Sect. 2 feature extraction techniques are discussed. In Sect. 3 we discuss experimentation details including data collection, preprocessing and classification. Finally we conclude our paper at Sect. 4.

2 Feature Extraction

The extraction of features is the most imperative task in any work requiring pattern recognition. The features should be robust enough, though, easily computational. Fractal geometry analysis or FGA based intelligent and optimized technique is employed here to compute the feature vector. Our approach is intelligent and optimized because for present work only one-dimensional feature vector is considered. Feature dimensionality reduction is an important issue in the field of machine learning. As the dimension of the feature set will be reduced it will become more computationally effective. The following section discuss about the extracted feature in detail.

2.1 *Fractal Geometry Analysis (FGA)*

Present work is motivated by the concept of fractal geometry analysis or in short FGA of an object [10]. A fractal is formally defined as an irregular geometric object with an infinite nesting of structure at all scales (or self-similarity). The geometric characteristics of the objects or connected components of an image can be under-

stood by its fractal dimension. A fractal is represented as a set for which Hausdorff-Besicovich [12] dimension is larger than the topological dimension. Fractal dimension is an important property for textural analysis. Researchers typically estimate the dimension of connected components in an image by fractal analysis. The fractal dimension of a continuous object is specified by a well-defined mathematical limiting processes.

Mandelbrot and Van Ness derived the fractal theory from the work of Hausdorff and Besikovich. The Hausdorff-Besicovich dimension (D_H) is defined by the following equations:

$$D_H = \lim_{s \rightarrow 0^+} \frac{\ln N_\epsilon}{\ln \frac{1}{\epsilon}} \quad (1)$$

where N_ϵ is the number of elements of ϵ diameter required to cover the object in the embedded space. When working with the data, which is discrete in nature, one is keen to find out the deterministic fractal and the associated fractal dimension (D_f). The fractal dimension (D_f) can be defined as the ratio of the number of self-similar pieces (N) with magnification factor ($1/r$) into which an image may be broken. Our intuitive idea is that, normally the dimensions of surfaces are integers. However we have to understand that, here we are dealing with non-idealized objects (*non-euclidean*), and thus we cannot say that their surfaces have dimensions which is integral in nature. As Mandelbrot said, mountains are not cones and lightnings are not straight lines. Here we encounter a similar situation. Surfaces of natural objects differ from the idealized ones in the aspect that the curves on natural surfaces are continuous, but non-differentiable. These objects, thus have a dimension D_f which is “fractional” in nature. D_f is defined as

$$D_f = \frac{\ln N}{\ln \frac{1}{r}} \quad (2)$$

D_f may be a non-integer value, as opposed to objects strictly Euclidean in nature. However, D_f can only be directly calculated for a deterministic fractal. Mandelbrot pointed out that the dimension of the dimension of the object and the dimension of the embedded space are two different things. Dimension of the embedded space is given by the degrees of freedom in that space. The idea of dimension of an object is how many boxes of the embedded space are necessary to cover it. We reduce the box size to zero progressively and find out where the number D_f converges, the converging point being the dimension of the object. There are varieties of applicable algorithms for estimating D_f , and we have used Box-counting algorithm for the same.

The upper part and the lower part play a significant role in feature extraction from the document image. This observation motivated us to solve the present problem by FGA. Indic scripts can be categorized as ‘matra’ based and without ‘matra’ based with respect to topological structure. So if pixel density of the connected components is calculated, there will be difference in pixel density of upper part and lower part of the components of different scripts.

The algorithm to compute average fractal dimension of an image component is described below

Algorithm_FGA :

- The line-level gray scale images are considered and converted into two tone image using our existing tow-stage based binarization technique.
- The fractal dimension of the top most and bottom most profile of each of the image component is calculated and stored in TFD and BFD variables correspondingly.
- Then the average fractal dimension of the top most and bottom most profile of each of the image component is computed. These average values are stored in ATFD and ABFD variables correspondingly.
- Ratio of the fractal dimension of top most and bottom most profile is calculated and this single valued feature vector is constructed. The ration is stored in variable RFD.

$$RFD = \frac{ATFD}{ABFD} \quad (3)$$

This RFD obtained by Eq. 3 is used as the one and only distinguishing topological feature for the present work.

Sample results are shown in following Fig. 3 for Bangla, Devanagari, Roman, and Urdu scripts. In each of these figures subpart of the original line-level image is shown in **a**. The fractal dimension of the *top* and *bottom* profiles are shown by **b** and **c** correspondingly.

3 Experiments

Any experimental work requires proper experimental setup building to carry out the training, testing, and validation. The first and foremost task is the data collection followed by preprocessing. Next Sect. 3.1 discuss about this step. In Sect. 3.2 brief discussions about experimental set up is provided. Lastly classification and statistical performance analysis of different well-known classifiers are done in Sect. 3.3.

3.1 Data Collection and Preprocessing

Due to unavailability of standard open database in script identification field we have developed our own dataset. Data collection is a tedious and time-consuming job which is solely dependent on other people who may contribute towards building the database. Different people with varying age, sex, educational qualification group were contributed to build our present database. We constructed a line-level database by collecting 1204 document images, comprising of the said scripts. The script dis-

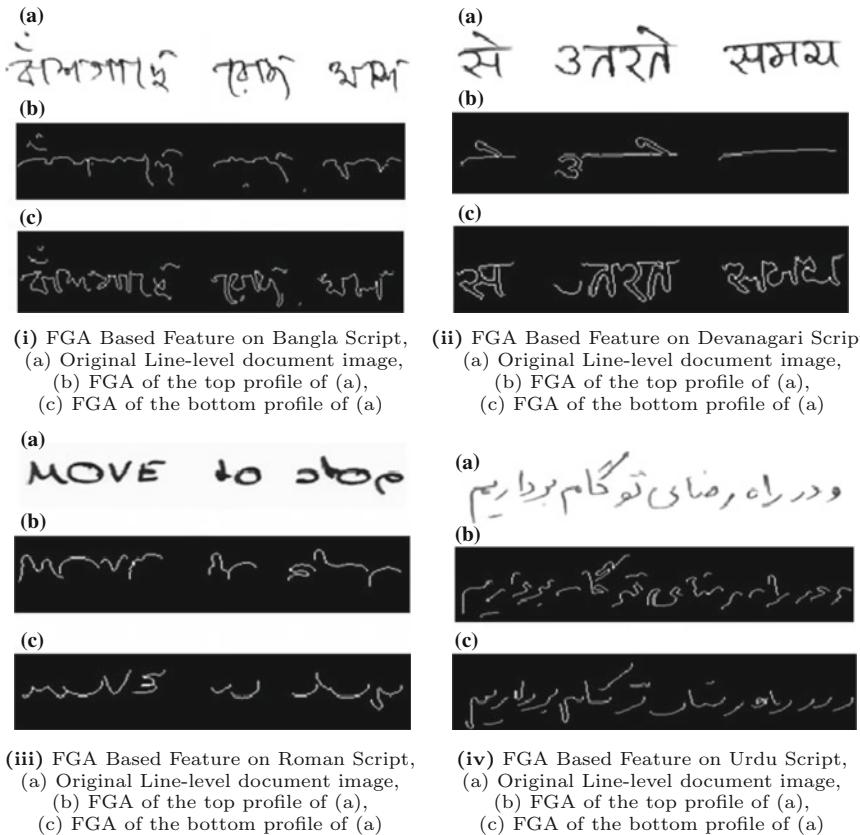


Fig. 3 Illustrating fractal dimension of Bangla, Devanagari, Roman, and Urdu scripts

Table 1 Database distribution (script wise)

Script	Topological property	Database size
Bangla	with ‘matra’	325
Devanagari	with ‘matra’	200
Roman	without ‘matra’	370
Urdu	without ‘matra’	309

tribution is as follows: Bangla 325, Devanagari 200, Roman 370, and Urdu 309. So the distribution of ‘matra’ based script and without ‘matra’ based script becomes 525 and 679 Line-level document images correspondingly. Following Table 1 provides a glimpse of the used dataset.

These images were digitized using HP flatbed scanner and stored in gray scale version with 300 dpi intensity level. We have used our existing two-stage based binarization algorithm [11] for preprocessing of these gray scale images (Fig. 4).

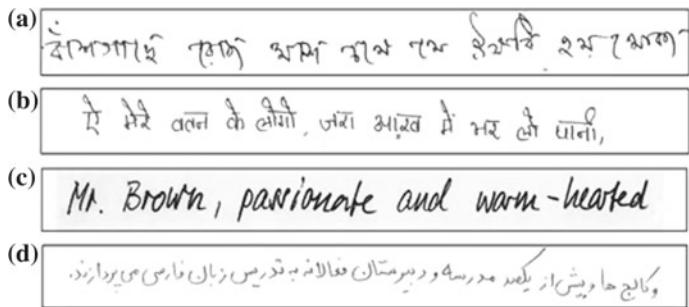


Fig. 4 Scripts with ‘matra’ **a** Bangla **b** Devanagari, and scripts without ‘matra’ **c** Roman **d** Urdu script documents (*top to bottom*)

3.2 Set up

Multilayer perceptron (MLP), random forest (RF), and bayes net (BN) were used for classification and statistical performance analysis which is discussed in the following section. We followed K-fold cross validation approach to train and test the dataset. In this work, the value of K was considered as five experimentally, which means that, the whole dataset will be broken into a ratio of 4:1. Then, it will be repeated five times such that, all the instances would participate for training and testing data distribution.

3.3 Results and Analysis

We have not only used a light weight single-dimensional feature set for classification but different well known classifiers are also tested. MLP, RF, and BN, these three classifiers are considered and their performances are statistically analyzed with respect to standard measuring parameters like average accuracy rate (AAR), model building time (MBT), relative absolute error (RAE), etc. A confusion matrix is shown in Table 2 which was taken using MLP classifier. For global classification, we are considering the complete ‘matra’ dataset(i.e., Bangla and Devanagari) versus the complete without ‘matra’ script dataset (i.e., Roman and Urdu). The ‘matra’ dataset are assigned as class 1, and without ‘matra’ dataset are assigned as class 2. For local classification, we are considering each instance of a ‘matra’ script versus without matra; script. Here, we are showing the resultant data sets that were built using MLP classifier with fivefold cross validation.

(A) For global classification, out of 525 ‘matra’ based scripts from Bangla and Devanagari 487 were classified correctly (on an average). For without ‘matra’ based scripts out of 679 document images, 654 instances were perfectly classified (on an average). Average accuracy rate of 94.43 % for MLP, 95.59 % is obtained by random

Table 2 Confusion matrix on fivefold cross validation using MLP

Classified As	‘matra’	Non-‘matra’
‘matra’	481	44
without ‘matra’	23	656

Table 3 Local classification accuracy

Script combinations	Average accuracy (%)
‘matra’ based	Bangla versus Roman
	Bangla versus Urdu
without ‘matra’ based	Devanagari versus Roman
	Devanagari versus Urdu

Table 4 Statistical performance analysis of different classifiers

Classifier	AAR (%)	MBT (s)	RAE (%)	TP Rate	FP Rate	Precision	Recall	F-Measure
MLP	96.44	0.59	10.67	0.9644	0.0406	0.965	0.9644	0.9642
RF	96.32	0.03	10.45	0.9632	0.0418	0.9632	0.9632	0.9632
BN	94.83	0.02	10.4	0.9484	0.058	0.9484	0.9484	0.9484

forest RF and 94.68 % for bayes net BN classifiers. So, we can observe that at global level RF > BN > MLP, though the difference is very close.

(B) For local classification, we took each instance of ‘matra’ vs without ‘matra’ script, and found the accuracy to be even higher. This is because the deviation is higher for a single instance of ‘matra’ versus without ‘matra’ script (e.g., Bangla vs Roman) than it would be for the whole class. The close proximity of average accuracy rate obtained by different classifiers supports the robustness of our technique. The highest accuracy for each instance of the local classification using MLP classifier are as follows (Tables 3 and 4):

4 Conclusion and Plan

No doubt, script identification from multi-script documents is essential towards making the automation of document processing for a large country like India. This work will definitely accelerate the entire process by offering a new approach for classification of ‘matra’ based scripts from others without ‘matra’. Further, an one-dimensional, lightweight feature ensures computational ease, resulting a faster processing. Not only that, out of three different classifiers namely MLP, RF, and BN, the best performer for different output cases has also been found out.

Future scope includes enriching the volume of the corpus. Scopes can be further extended to review some misclassification issues by combining FGA with few more script dependent features. But a realistic trade-off between feature dimension and accuracy rate must also be kept into mind.

References

1. Ghosh, D., Dube, T., Shivprasad, S. P.: Script Recognition - A Review. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(12), 2142–2161 (2010)
2. Hochberg, J., Bowers, K., Cannon, M., Kelly, P.: Script and Language Identification for Handwritten Document Images. *Int. J. Doc. Anal. Recog.* 2(2/3), 45–52 (1999)
3. Zhu, G., Yu, X., Li, Y., Doermann, D.: Language Identification for Handwritten Document Images Using A Shape Codebook. *Pattern Recog.* 42, 3184–3191 (2009)
4. Singhal, V., Navin, N., Ghosh, D.: Script-based Classification of Hand-written Text Documents in a Multi-lingual Environment. In: 13th RIDE-MLIM. pp. 47–54 (2003)
5. Hangarge, M., Dhanda, B. V.: Offline Handwritten Script Identification in Document Images. *Int. J. Comput. Appl.* 4(6), 6–10 (2010)
6. Rajput, G., H. B., A.: Handwritten Script Recognition using DCT and Wavelet Features at Block Level. *IJCA, Special Issue on RTIPPR.* 3, 158–163 (2010)
7. Sarkar, R., Das, N., Basu, S., Kundu, M., Nasipuri, M., Basu, D. K.: Word level Script Identification from Bangla and Devanagari Handwritten Texts Mixed with Roman Script. *J. Comput.* 2(2), 103–108 (2010)
8. Hangarge, M., Santosh, K. C., Pardeshi, R.: Directional discrete cosine transform for handwritten script identification. In: *ICDAR.* pp. 344–348 (2013)
9. Rani, R., Dhir, R., Lehal, G. S.: Script Identification for Pre-segmented Multi-font Characters and Digits. In: 12th *ICDAR.* pp. 2010–1154 (2013)
10. Roy, K., Pal, U.: Word-wise Hand-written Script Separation for Indian Postal Automation. In 10th *IWFHR.* pp. 521–526 (2006)
11. Roy, K., Banerjee, A., Pal, U.: A System for Word Wise Handwritten Script Identification for Indian Postal Automation. In: *IEEE India Annual Conf.* pp. 266–271 (2004)
12. Mandelbrot, B. B.: *The Fractal Geometry of Nature* (New York: Freeman). (1982)

Efficient Multimodal Biometric Feature Fusion Using Block Sum and Minutiae Techniques

Ujwalla Gawande, Kamal Hajari and Yogesh Golhar

Abstract Biometric is widely used for identifying a person in different area like Security zones, Border crossings, Airports, Automatic teller machines, Passport, Criminal verification, etc. Currently, most of the deployed biometric systems use a single biometric trait for recognition. But there are several limitations of unimodal biometric system, such as Noise in sensed data, Non-universality, higher error rate, and lower recognition rate. These issues can be handled by designing a Multimodal biometric system. This research paper proposes a novel feature level fusion technique based on a distance metric to improve both recognition rate and response time. This algorithm is based on the textural features extracted from iris using Block sum and fingerprint using Minutiae method. The performance of the propose algorithms has been validated and compared with the other algorithms using the CASIA Version 3 iris database and YCCE Fingerprint database.

Keywords Multimodal biometric system • Block sum • Minutiae • Feature level fusion

1 Introduction

Multimodal biometric recognition requires logical fusion of different traits [1–3]. Whenever more than one biometric modality is used for the development of multimodal biometric, the fusion of different modalities can be performed at the sensor, feature, matching score, or decision levels. The majority of the work in this area has been focused on fusion at decision level and matching score-level [4, 5]. The

U. Gawande (✉) · K. Hajari · Y. Golhar
Department of Information Technology, YCCE, Nagpur, India
e-mail: ujwallawande@yahoo.co.in

K. Hajari
e-mail: kamalhajari123@gmail.com

Y. Golhar
e-mail: yj999@ymail.com

amount of information available for fusion at these levels is limited [6]. Sparing references are available for fusion at the feature level [7, 8]. Integration at feature level is expected to achieve high recognition rate. On the contrary feature level fusion is difficult to achieve, due to various issues like: (1) the feature sets of multiple modalities may be incompatible. For example, minutiae points set of fingerprint and Eigen-coefficients of face. (2) The feature spaces of diverse biometric systems may be different and (3) concatenating two feature vectors may lead to a feature vector with very high dimension [6]. Due to these issues False Rejection Rate (FRR) is more. Most of the feature level fusion in the literature is performed using simple concatenation of feature sets, serially or parallel [9, 10].

The remainder of this paper is organized as follows: Sect. 2 describes various unimodal and multimodal system. Section 3 describes the preprocessing and texture feature extraction algorithm for both the modality. Section 4 describes the proposed feature level fusions for single feature set algorithm. Simulation results are presented in Sect. 5 along with the comparison of the proposed approach with existing feature level fusion technique. Finally, Sect. 6 concludes this paper.

2 Unimodal and Multimodal Systems

The fused feature vector usually requires appropriate feature normalization, selection, and reduction techniques [5]. Since the features exploits more useful information about the raw biometric data, fusion at this level is expected to give more accuracy compared to fusion at the score and decision levels [6, 11]. It is observed that feature level fusion is easy to implement for closely related modalities. These feature sets are called homogeneous sets [12]. However, feature level fusion is difficult to achieve for heterogeneous feature sets [5]. Due to these constraints, most of the attempts at feature level fusion have reported limited success. Son and Lee [13] used Daubechies wavelet transform for face and iris images. Kumar et al. [14] proposed palmprint-based verification system by integrating hand geometry features. Feature level fusion of iris and fingerprint has been proposed by Jagadeesan et al. [7]. A similar attempt is made by Conti et al. [15].

3 Preprocessing and Feature Extraction

We have selected Iris and Fingerprint as the two modalities to fuse at feature level.

3.1 Iris Recognition

Iris patterns are unique due to its rich, distinctive, and complex pattern of crypts, furrows, arching, collarettes, and pigment spots. These factors result in discriminating

textural patterns for each eye of an individual and even distinct between twins. Iris recognition consists of three main steps. (1) Preprocessing, (2) Feature Extraction, and (3) Recognition.

3.2 Iris Preprocessing

The preprocessing step consists of iris localization and normalization steps. The iris area between the inner and outer boundaries of the iris is first localized. This step also detects and removes any specular reflection and eyelash or eyelids noise from the image using Hough transform [17]. Daugmans rubber sheet model is used for iris normalization [18]. The next step is to extract distinctive texture features from normalized iris image.

3.3 Iris Feature Extraction

We propose a novel algorithm called block sum [19] for extracting the texture features from normalized image. This method is based on the blocks, which is a rectangular window of pixels in a normalized image, as shown in Fig. 1. The block size is varied from 5×5 to larger values. In this context, the normalized image of size 240×20 , is divided into 60 blocks of size $m \times n$, where $m = 16$ and $n = 5$.

The block size is set to 16×5 and its main aim is to reduce size feature vector, first the entropy for a given block. It is the probability of occurrence of the average intensity value of that block. To obtain the unique feature vector, we subtract the entropy value of a given block from the gray level of each pixel and sum it with the block sum of previous cell. The representative value of each block is computed as follows:

1. Entropy for each block is computed using an entropy function, say it is ' E_i ', where ' i ' varies from 1 to 60.
2. The gray level value for each pixel in a block is represented as C_{ij} , where ' i ' represent the block index for j th pixel in the block, respectively. With a normalized image of size 240×20 , ' j ' varies from 1 to 80.

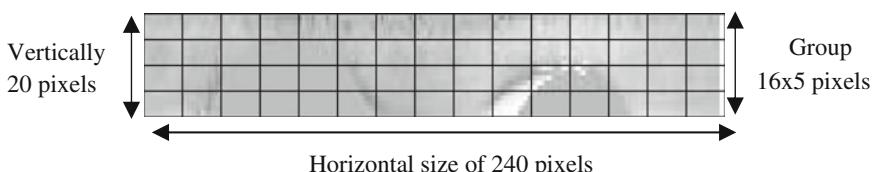


Fig. 1 Division of normalized image into cells

3. For finding out the block sum feature value for the i th block and for every pixel in a block, the value Sj , is calculated using

$$Sj = Sj - 1 + (Cij - Ei) \quad (1)$$

Where, $1 \leq i \leq 60$, $1 \leq j \leq 80$

4. Finally, the value of Sj is scaled by a size of window or block, i.e., final value of block sum feature for the i th block, which is used as a feature value of that block.

$$Si = Sj = \sum_{j=1}^{80} \left(\frac{Sj}{80} \right) \quad (2)$$

In this manner, 60 values for 60 blocks are obtained and stored feature vector of size 1×60 . This block sum feature vector is further used in feature level fusion. Another biometric used for feature level fusion is Fingerprint.

3.4 Fingerprint Recognition

Fingerprint is highly used in society and extensively used by forensic experts in criminal investigations. The fingerprint recognition consists of four steps: (1) Image acquisition, (2) Preprocessing, (3) Feature extraction, and (4) Recognition. These steps are performed on YCCE Fingerprint database [20, 21].

3.5 Fingerprint Preprocessing

Fingerprint images acquired not assured to be of perfect quality. Preprocessing step increases the contrast between ridges and furrows and connects the false broken points of ridges. This step is followed by finding a region of interest, which is performed by segmentation.

3.6 Fingerprint Feature Extraction

We propose a novel feature set algorithm that captures global point, i.e., core and local points (minutiae point's) of a fingerprint. For this, we used the method of [22] due to its more accurate results in detecting the core point. At each pixel the directional component ' θ ' is available based on the gradient. In every block, the difference of direction components is computed using Eqs. (1) and (2).

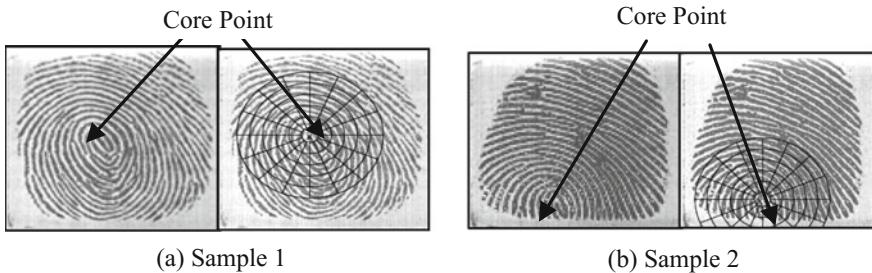


Fig. 2 Samples of core point detection

$$\text{Diff } X = \sum_{k=1}^3 \sin 2\theta(k, 3) - \sum_{k=1}^3 \sin 2\theta(k, 1) \quad (3)$$

$$\text{Diff } Y = \sum_{k=1}^3 \cos 2\theta(3, 1) - \sum_{k=1}^3 \cos 2\theta(1, 1) \quad (4)$$

The core point ‘X’ is located at the pixel (i, j) where Diff X and Diff Y are negative. The result of this core point detection is shown in Fig. 2.

The next step is to mark the ridge termination and bifurcation. We perform a morphological operation that connects the ridges for nonbreakable ridge structure. At each pixel 3×3 windows sliding is performed. The minutiae points are marked based on the number of ones in these 3×3 blocks. These blocks are categorized as terminations, if a central pixel is 1 and count of one valued neighbor is also one. Similarly, blocks are categorized as bifurcations, if a central pixel is 1 and counts of one valued neighbor are three. [23], the fingerprint image consists of genuine marked minutiae points, as shown in Fig. 3. Now, using terminations and bifurcations point, we follow the ridges, starting from core point, in increasing direction of radial distance.

The 60 value is derived based on an observation that maximum fingerprint information will be captured by minutiae points. The distance from the core point to each of initial 60 minutiae points are considered as new features set.

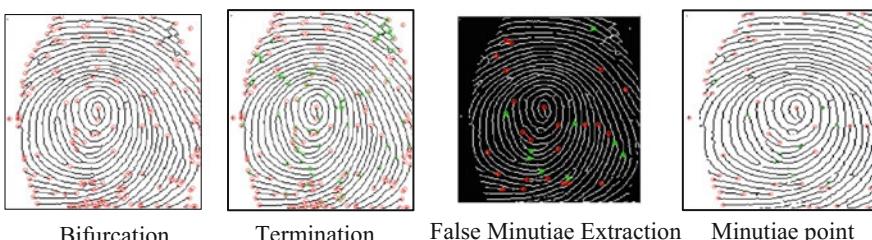


Fig. 3 Sample of marked minutiae points around the core point

4 Feature Level Fusion

A new method of feature fusion is proposed in this section. It consists of two steps. First step is the generation of fused feature vector for reference database. This fused vector is trained using RBFSVM classifier. Second step is the generation of query feature sets, which is used to test the RBFSVM classifier for final recognition.

4.1 Fused Feature Vector for Reference Database

The process of fused feature vector generation for reference database is depicted in Fig. 4. Here, we describe the framework in generic form. Let the total number of subjects is ‘M,’ which is identified by ‘i’ and it varies from 1 to m. Let the number of samples per subject is ‘N,’ which is identified using ‘j’ and ‘j’ varies from 1 to n. The total number of subject, i.e., ‘M,’ is set to 100 and the number of images per subject, i.e., ‘N,’ is set to 4, for our experimentation. The proposed algorithm is described below

1. Each fingerprint feature vector is represented as F_{ij} . Where ‘F,’ stands for fingerprint, ‘i’ represent the subject number with ‘m’ number of subject, ‘i’ varies from 1 to m. The ‘j’ indicates feature vector derived from jth sample image. So ‘j’ varies from 1 to 4, as four images are used per subject. The length of each of these feature vectors is ‘l,’ where ‘l’ is the set of 60 elements.
2. Each iris feature vector is represented as I_{ij} . Where ‘I,’ stands for iris, ‘i’ represent the subject number and ‘j’ indicates feature vector derived from a jth sample image. Each generated feature vector is of length 1×60 .
3. Next, we derive the difference vector for each pair each subject, i.e., DF_{ij} , where ‘ DF_{ij} ’ for difference feature of a fingerprint, for jth sample image of ith subject.

$$DF_{i1} = |F_{i1} - F_{i2}| \quad (5)$$

$$DF_{i2} = |F_{i2} - F_{i3}| \quad (6)$$

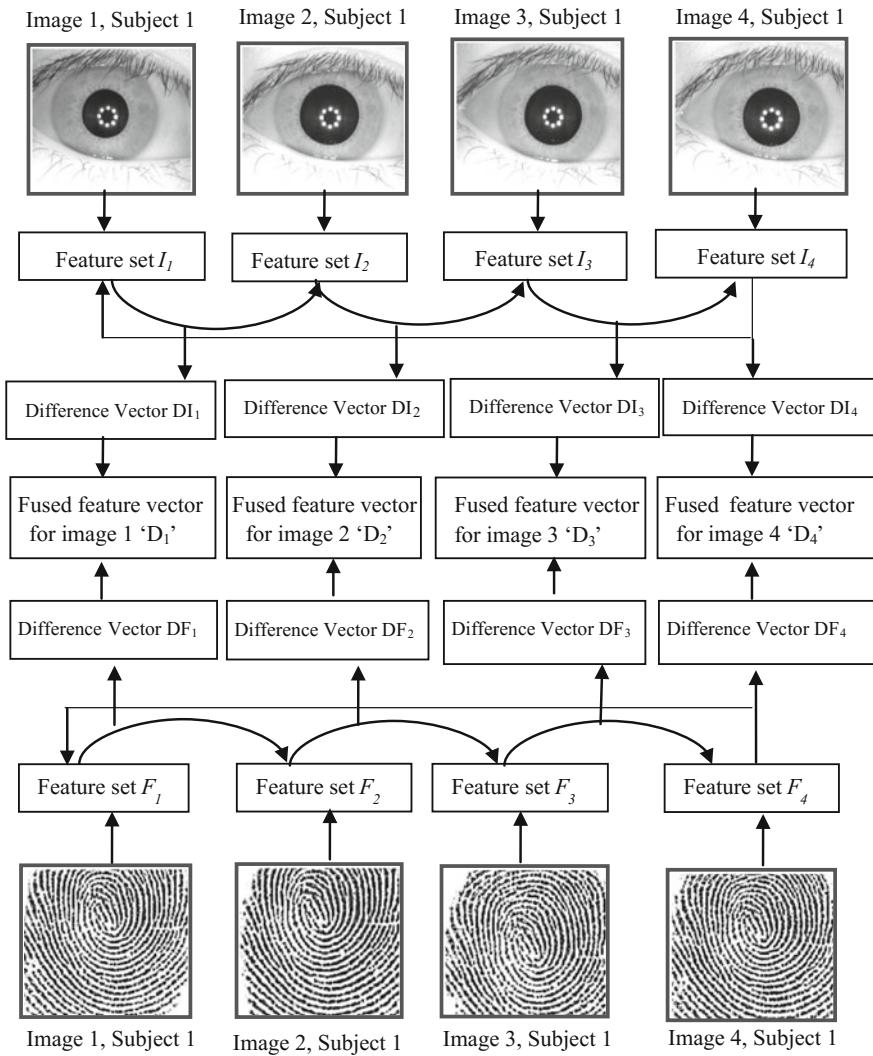
$$DF_{i3} = |F_{i3} - F_{i4}| \quad (7)$$

$$DF_{i4} = |F_{i4} - F_{i1}| \quad (8)$$

4. Similarly, for iris, again we derive the difference vector for each subject, i.e., ‘ DI_{ij} ,’ where ‘ DI_{ij} ’ for difference feature of the iris, for jth sample image of ith subject.

$$DI_{i1} = |I_{i1} - I_{i2}| \quad (9)$$

$$DI_{i2} = |I_{i2} - I_{i3}| \quad (10)$$

**Fig. 4** Reference fusion vector creation

$$DIi3 = |Ii3 - Ii4| \quad (11)$$

$$DIi4 = |Ii4 - Ii1| \quad (12)$$

5. Here each generated difference vector of size is 1×60 . The computational complexity reduces by taking the average of two difference measure derived from modality.

$$D_{ij} = (DF_{ij} + DI_{ij})/2 \quad (13)$$

Were, ' j ' = 1 → n , were n = 4 and

' i ' = 1 → m , were m = 100

The length of the resultant fused feature vector is also 1 × 60, for each subject.

4.2 Feature Level Fusion for Query Image

1. We search for most similar feature vector from reference database to derive difference vector. The proposed feature level fusion for query image is shown in Fig. 5.

$$\Sigma_d = \sqrt{(\bar{x} - \bar{x}') R^{-1} (\bar{x} - \bar{x}')^t} \quad (14)$$

Where \bar{x} and \bar{x}' are the query feature vector and reference feature vector, respectively. 'R' represents the covariance between \bar{x} and \bar{x}' . In our case, the Mahalanobis distance for fingerprint and iris are represented by 'ΣF' and 'ΣI' symbols.

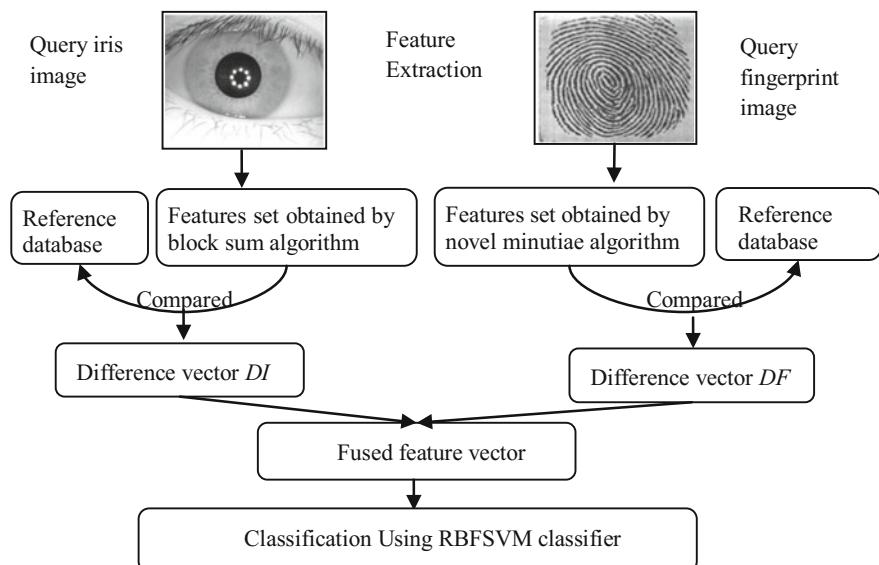


Fig. 5 Proposed feature level fusion

$$\sum_F = \sqrt{(F_t - F_{ij}) R^{-1} (F_t - F_{ij})^t} \quad (15)$$

$$\sum_I = \sqrt{(I_t - I_{ij}) R^{-1} (I_t - I_{ij})^t} \quad (16)$$

2. Next, the most similar feature vector is derived based on the Mahalanobis distance

$$\sum_F = \min_{ij} \left((F_t - F_{ij}) R^{-1} (F_t - F_{ij})^t \right)^{1/2} \quad (17)$$

$$\sum_I = \min_{ij} \left((I_t - I_{ij}) R^{-1} (I_t - I_{ij})^t \right)^{1/2} \quad (18)$$

3. We derive the difference vector between query feature vector and most similar feature vector by

$$DF_t = |F_t - F_{ij}| \quad (19)$$

$$DI_t = |I_t - I_{ij}| \quad (20)$$

Where F_{ij} and I_{ij} are the most similar feature vector in the reference database.

4. For feature level fusion, we merge these two difference vectors derived by taking simple average

$$D_t = (DF_t + DI_t)/2 \quad (21)$$

5. The length of the fused feature vector is also 1×60 . This new fusion vector is used for further classification using RBFSVM classifier for final recognition.

5 Simulation Results

The performance are measures using FAR, FRR, and response time. The database consists of 1000 images from 200 subjects. For database, we use 100 subjects of iris from CASIA and 100 subjects from YCCE fingerprint database. Five pairs of iris and fingerprint are obtained for each subject. Each pair represents one subject. Our approach is compared with the approach of [16]. The results are depicted in Fig. 6. In this experiment best performance of 1 % FAR is archived for proposed algorithm.

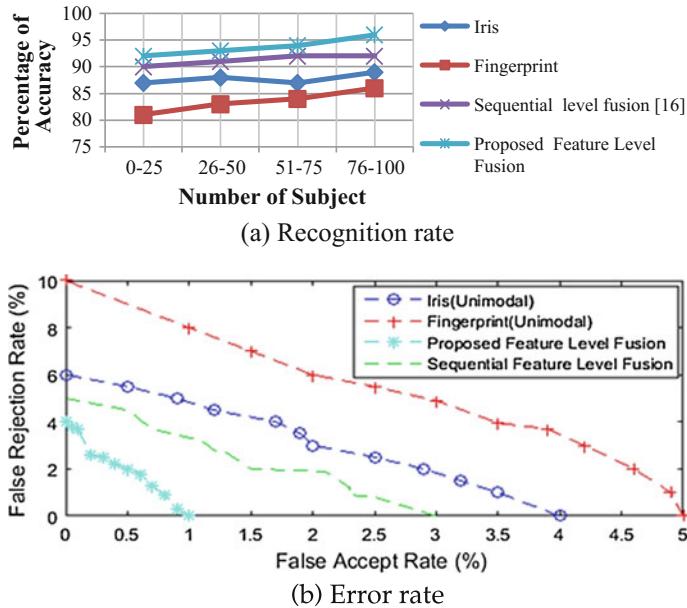


Fig. 6 Comparison of proposed feature level fusion using RBFSVM classifier

6 Conclusion

Some of the issues in feature level fusion in the context of multimodal biometrics have been tackled in the proposed work. The feature level fusion algorithm eliminates the high dimensionality issues of existing feature fusion. The objective is to obtain a higher recognition rate with minimum false acceptance rate and fast response time. With this objective, from the simulation results, it is observed that 96 % of recognition rate and 0 % FAR are obtained using proposed feature level fusion. All these experiments are sufficient to prove the superiority of proposed feature level fusion. Overall, comparing the results, it is observed that the proposed algorithm fuses the features of individual modalities efficiently, which is very much evident from simulation results.

References

1. N. Poh, and S. Bengio, “Database, Protocols and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication”, Journal of Pattern Recognition, ScienceDirect, vol.39, no. 2 pp. 223–233, Feb. 2006.
2. A. Jain, K. Nandakumar and A. Ross, “Score normalization in multimodal biometric systems”, IEEE Journal of Pattern Recognition, vol. 38, no. 12, pp. 2270–2285, Dec., 2005.

3. R. Frischholz and U. Dieckmann, "BioID: A Multimodal Biometric Identification System", IEEE Journal of Computer Science, vol. 33, no. 2, pp. 64–68, Feb. 2000.
4. A. Gongazaga and R. Dacosta, "Extraction and Selection of Dynamic Features of Human Iris", IEEE Journal of Computer Graphics and Image Processing (SIPGRAPI), vol. 22, no. 1, pp. 202–208, Oct., 11–15, 2009.
5. A. Ross, K. Nandakumar and A. K. Jain, "Handbook of Multi-biometric", International series on Biometrics, New York: Springer-Verlag, vol. 6, 2006.
6. M. Faundez-Zanuy, "Data Fusion in Biometrics", IEEE Journal on Aerospace and Electronic Systems Magazine, vol. 20, no. 1, pp. 34–38, Jan. 2005.
7. A. Jagadeesan, Thillaikkarasi. T., K. Duraiswamy, "Protected Bio-Cryptography Key Invention from Multimodal Modalities: Feature Level Fusion of Fingerprint and Iris", European Journal of Scientific Research, vol. 49, no. 4, pp. 484–502, Feb. 2011.
8. Md. MarufMonwar, and Marina L. Gavrilova, "Multimodal Biometric System Using Rank-Level Fusion Approach", IEEE Transaction - Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 39, no. 4, pp. 867–879, Aug. 2009.
9. S. Chikkerur, A. Cartwright, V. Govindaraju, "Fingerprint Enhancement using STFT Analysis", Journal on Pattern Recognition, Elsevier, vol. 40, no.1, pp. 198–211, Jan. 2007.
10. W. Chen and S. Yuan, "A Novel Personal Biometric Authentication Technique using Human Iris based on Fractal Dimension Features", International Conference on Acoustics Speech and Signal Processing, Hong Kong, China, vol. 3, pp. 201–204, April 6–10,2003.
11. A. Nagar, K. Nandakumar, and A. K. Jain, "Multi-biometric Cryptosystems Based on Feature-Level Fusion", IEEE Transactions on Information Forensics and Security, vol. 7, no. 1, pp. 255–268, Feb. 2012.
12. A. Ross, S. Shah and J. Shah, "Image Versus Feature Mosaicing: A Case Study in Fingerprints", SKPIE Conference on Biometric Technology for Human Identification III, Orlando, USA, vol. 6202, pp. 620208-1–620208-12, April 17, 2006.
13. B. Son and Y. Lee, "Biometric Authentication System Using Reduced Joint Feature Vector of Iris and Face", 5th International Conference on Audio and Video Based Biometric Person Authentication (AVBPA), Rye Brook, USA, pp. 513–522, July 20–22, 2005.
14. A. Kumar and D. Zhang, "Personal Authentication using Multiple Palmprint Representation", Pattern Recognition Letters, vol. 38, no. 10, pp. 1695–1704, Oct. 2005.
15. V. Conti, C. Militello, F. Sorbello and S. Vitabile, "A Frequency-based Approach for Features Fusion in Fingerprint and Iris Multimodal Biometric Identification Systems", IEEE Tran. on Systems, Man and Cybernetics, Part C, vol. 40, no. 4, pp. 384–395, July 2010.
16. I. Raglu and Deepthi P.P, "Multimodal Biometric Encryption Using Minutiae and Iris feature map", IEEE Conference on Electrical, Electronics and Computer Science, Madhya Pradesh, India, pp. 926–934, March 1–2, 2012.
17. L. Masek, "Recognition of Human Iris Patterns for Biometrics Identification", B.E. thesis, School of Computer Science and Software Engineering, Uni. Of Western Australia, 2003.
18. J. Daugman, "How Iris Recognition Works", IEEE Transactions on Circuits and Systems for Video Technology, vol.14, no.1, pp. 21–30, Jan. 2004.
19. U. Gawande, M. Zaveri and A. Kapur, "Bimodal biometric system: feature level fusion of iris and fingerprint", ScienceDirect, Elsevier, vol. 2013, no. 2 pp. 7–8, Feb. 2013.
20. U. Gawande, K. Hajari, Y. Golhar, "YCCE Fingerprint Color image database". v, File ID: #52507, Version:1.0 Web Link: <http://www.mathworks.com/matlabcentral/fileexchange/52507-fingerprint-color-image-database-v1>.
21. U. Gawande, K. Hajari, Y. Golhar, "YCCE Fingerprint Grayscale image database. v2, File ID: #5250, Version: 2.0 Web Link: <http://www.mathworks.com/matlabcentral/fileexchange/52508-fingerprint-grayscale-image-database-v2>.
22. A. K. Jain, J. Feng and K. Nandakumar, "On Matching Latent Fingerprint", IEEE Workshop of Computer vision and Pattern Recognition, pp. 36–44, June, 23–28, 2008.

Video Synopsis for IR Imagery Considering Video as a 3D Data Cuboid

Nikhil Kumar, Ashish Kumar and Neeta Kandpal

Abstract Video synopsis is a way to transform a recorded video into a temporal compact representation. Surveillance videos generally contain huge amount of recorded data as there are a lot of inherent spatio-temporal redundancies in the form of segments having no activities; browsing and retrieval of such huge data has always remained an inconvenient job. We present an approach to video synopsis for IR imagery in which considered video is mapped into a temporal compact and chronologically analogous way by removing these inherent spatio-temporal redundancies significantly. A group of frames of video sequence is taken to form a 3D data cuboid with X , Y and T axes, this cuboid is re-represented as stack of contiguous $X - T$ slices. With the help of Canny's edge detection and Hough transform-based line detection, contents of these slices are analysed and segments having spatio-temporal redundancy are eliminated. Hence, recorded video is dynamically summarized on the basis of its content.

Keywords Video synopsis · Video summarization · IR · MWIR · Canny's edge detection · Hough transform-based line detection · Spatio-temporal redundancy

1 Introduction

Popularity of thermal imaging systems in surveillance technology has drawn a lot of attention from vision community in the past few decades. Increasing population of such systems is generating vast amount of data in the form of recorded videos; with help of video summarization a compact but informative representation of video

N. Kumar (✉) · A. Kumar · N. Kandpal
Instruments Research and Development Establishment, Dehradun, India
e-mail: nikhilkumar@irde.drdo.in

A. Kumar
e-mail: ashishkumar@irde.drdo.in

N. Kandpal
e-mail: neeta@irde.drdo.in

sequence may be provided. Since for surveillance purpose timing information of events is important, chronology of events is also maintained in compact representation. Generally, IR (infra-red) signatures of targets are more prominent than background and clutter; this contrast is commonly used as a clue for change detection. We have also decided contrast-based clue for detecting representative segments with motion but in place of processing video sequence in $X - Y$ plane, we have chosen $X - T$ plane. Spatio-temporal regularity [1] is utilized for labelling representative segments with motion.

2 Related Work

The goal of this section is to review and classify the state-of-the-art video synopsis generation methods and identify new trends. Our aim is to extract information from unscripted and unstructured data obtained from recorder of surveillance system. Ding [2] categorized video synopsis techniques in the following three levels:

- Feature-Based Extraction: In such approaches low level features like number of foreground pixels and distance between histograms are used to identify frames with higher information content.
- Object-Based Extraction: In such approaches objects of interest like vehicle, pedestrian are used for labelling frames with higher information content.
- Event-Based Extraction: In such approaches events like entrance of a vehicle, pedestrian in field of view are used for setting pointers with high semantic level. Such approaches are more application specific.

Li et al. [3] presented an optical flow based approach for surveillance video summarization. It is a motion analysis-based video skimming scheme in which play-back speed depends upon motion behaviour.

Ji et al. [4] presented an approach based on motion detection and trajectory extraction. Video is segmented based on the moving objects detection and trajectories are extracted from each moving object. Then, only key frames along with the trajectories are selected to represent the video summarization.

Cullen et al. [5] presented an approach to detect boats, cars and people at coastal area. For this, the region of interest is decided and validated. It is taken as input for video condensation algorithm to remove inactive time space.

Rav-Acha et al. [6] presented a method for dynamic video synopsis in which several activities were compressed into a shorter time, where the density of activities were much higher. For better summarization of video event, chronology is not maintained as several events are merged in few frames.

Petrovic et al. [7] presented an approach for adaptive video fast forward. A likelihood function based upon content of video is formulated and playback speed is modelled accordingly.

Hoferlin et al. [8] presented an information based adaptive fast forward approach in which the playback speed depends on the density of temporal information in the

video. The temporal information between two frames is computed by the divergence between the absolute frame difference and noise distribution.

Porikli [9] presented multiple camera surveillance and tracking system based on object based summarization approach. For this, only the video sequence for each object is stored in place of storing video for each camera. Then, object is tracked by background subtraction and mean shift analysis.

Most of the approaches discussed above rely on motion detection-based techniques in $X - Y$ plane for video summarization but in case of IR sequences with poor SNR and targets limited in very small fraction of $X - Y$ plane it becomes challenging to detect targets, to tackle with such scenarios a novel approach of video summarization is presented in subsequent sections. In place of detecting targets in $X - Y$ plane, trajectory of motion is extracted from $X - T$ slices which covers a relatively larger fraction of $X - T$ slice.

3 Methodology

3.1 Overview of the Approach

Problem of video synopsis can be defined as a mapping generation problem between a video sequence and its temporal compact version. In the present approach for mapping generation, considered video sequence is analysed in $X - T$ plane and spatio-temporal redundant segments are eliminated. Trajectory of moving objects is utilized for this job. First a video sequence is represented as a 3D data cuboid V_{XYT} then this cuboid is chopped in contiguous $X - T$ slices and a set of slices $I(y)_{XT}$ is generated. Set of binary edge maps $E_{XT}(y)$ is obtained from $I(y)_{XT}$ with help of Canny's edge detection. A consolidated edge map ξ_{XT} is generated by registration of all elements of $E_{XT}(y)$. Using Hough transform-based line detection with a number of constraints representative segments with motion are labelled in ξ'_{XT} and ζ_{XT} is generated where ξ'_{XT} is binary edge map obtained from ξ_{XT} . For transformation from V_{XYT} to Ψ_{XYT} a transformation matrix τ_T is needed which is extracted from τ_{XT} where τ_{XT} is formed by subtracting ξ'_{XT} from ζ_{XT} and Ψ_{XYT} is 3D data cuboid representation of temporal compact version of video sequence.

3.2 Data Cuboid Representation of a Video Sequence

As in Fig. 1a, b group of frames of video sequence is taken to form a 3D data cuboid V_{XYT} with X , Y and T as axes. V_{XYT} can be expressed [10, 11] as following:

$$V_{XYT} = \{I(t)_{XY}, \forall t \in \{1, 2, 3, \dots, \dots, p\}\} \quad (1)$$

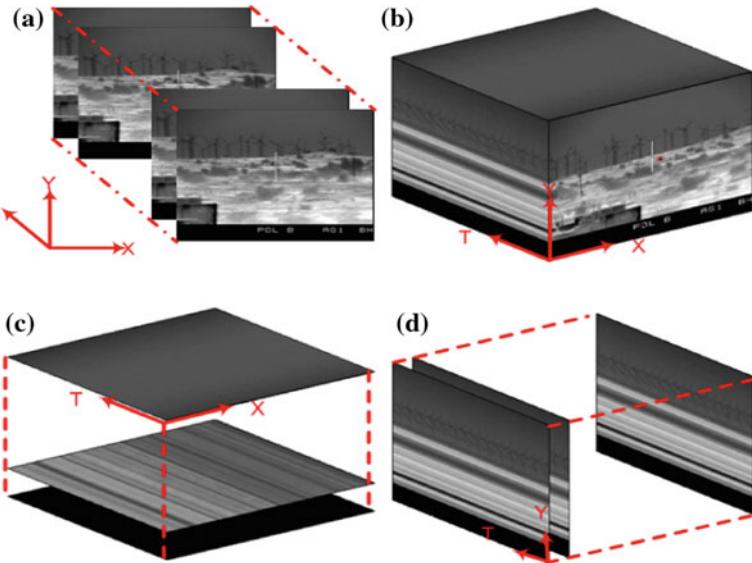


Fig. 1 Data cuboid representation of a video sequence **a** Frames from video sequence *Windmill*, **b** Video sequence *Windmill* represented as a 3D data cuboid, **c** Data cuboid chopped in contiguous $X - T$ slices and **d** Data cuboid chopped in contiguous $Y - T$ slices

Where $I(t)_{XY}$ is a frame of video sequence with X and Y axes at any particular time t and p is number of such frames of size $m \times n$.

As shown in Fig. 1c data cuboid V_{XYT} has an alternative representation [10, 11] as an stack of m number of contiguous $X - T$ slices $I(y)_{XT}$ with size $n \times p$.

$$V_{XYT} = \{I(y)_{XT}, \forall y \in \{1, 2, 3, \dots, m\}\} \quad (2)$$

Yet another way to represent [10, 11] the same data cuboid V_{XYT} is suggested in Fig. 1d by stacking n number of contiguous $Y - T$ slices $I(x)_{YT}$.

$$V_{XYT} = \{I(x)_{YT}, \forall x \in \{1, 2, 3, \dots, n\}\} \quad (3)$$

3.3 Mapping Between Contents of $X - Y$ and $X - T$ Planes of Data Cuboid

If content of $X - Y$ frame is stationary then in $X - T$ slices there will be a number of horizontal features parallel to T axes. Since present approach assumes that video is recorded from a stationary IR system, such horizontal features are most likely content

of $X - T$ slices. Most important conclusion related to present work is that if there are pixels with local motion in $X - Y$ frame then trajectory of motion appears in features of $X - T$ slices having those pixels. Geometry of this trajectory can be approximated by combining a number of inclined line segments. If there is any acceleration in motion, then there will be a number of curves in trajectory but any curve can be approximated by combining a number of small inclined line segments. This fact is utilized for labelling of segments with motion. In Fig. 2, an $X - T$ slice is shown which corresponds to $X - Y$ frame containing stationary as well as moving objects, hence combination of corresponding features is appearing in figure.

Formation of a Set of Binary Edge Maps. From Eq. 2 a set $\{I(y)_{XT}, \forall y \in \{1, 2, 3, \dots, m\}\}$ is obtained from V_{XYT} . In this section we obtain $E_{XT}(y)$ from $\{I(y)_{XT}, \forall y \in \{1, 2, 3, \dots, m\}\}$ using Canny's edge detection [12], which is one of the most widely used edge detection algorithms. Even though it is quite old, it has become one of the standard edge detection methods and it is still used in research [13]. Canny redefined edge detection problem as a signal processing optimization problem and defined an objective function with following

- Detection: Probability of detecting real edge points should be maximized while the probability of falsely detecting non-edge points should be minimized.
- Localization: Amount of error between detected edge and real edge should be minimum.
- Number of responses: For one real edge there should be one detected edge though this point is implicit in first point yet important.

Consolidated Edge Map Generation. Since we are mapping video sequence into a temporal compact representation, information carried along Y axes of V_{XYT} is redundant atleast for labelling of representative segments with motion; therefore for further processing we are using a consolidated edge map formed by utilizing all elements of $E_{XT}(y)$. As $E_{XT}(y)$ is generated from $I(y)_{XT}$ whose elements are contiguous slices, all elements of $E_{XT}(y)$ are already registered in spatial domain; hence consolidated edge map ξ_{XT} of V_{XYT} is generated by using logical OR operation over all elements of $E_{XT}(y)$.

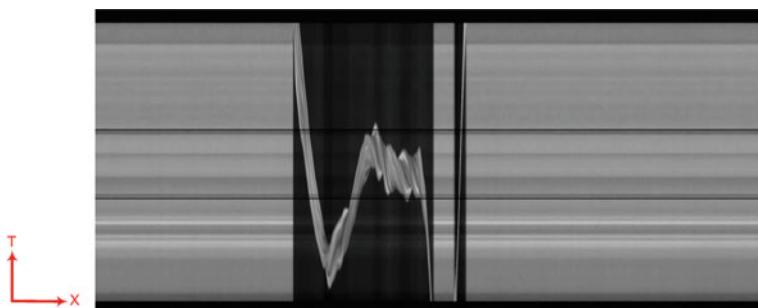


Fig. 2 A typical $X - T$ slice representing features of stationary as well as moving objects in corresponding $X - Y$ plane of video sequence, Moving objects are appearing in curved trajectory

3.4 Extraction of Representative Segments with Motion from Consolidated Binary Edge Map

As discussed earlier, to extract segments having motion we have to extract inclined line segments from $X - T$ slices, hence our goal is to find out set of inclined Y_{XT} . But as $Y_{XT} \subset L_{XT}$ where L_{XT} is set of lines with cardinality r in any $X - T$ slice, elements of Y_{XT} are obtained from L_{XT} with imposed constraints. Hough transform-based line detection is used to find out elements of L_{XT} .

Hough Transform-Based Line Detection. Now we have to explore a set of line segments $L_{XT}(y)$ from binary edge map ξ'_{XT} of consolidated edge map ξ_{XT} which is mathematically a set of points in any V_{XYT} . Hough transform [14] based line detection is a very popular, accurate, easy, and voting-based approach for such kind of operations [15]. Hough transform is based upon line point duality between $X - Y$ and $M - C$ domains, where $y = mx + c$ is equation of line. By quantizing the $M - C$ space appropriately a two-dimensional matrix H is initialized with zeros. A voting-based method is used for finding out elements of H matrix $H(m_i, c_i)$, showing the frequency of edge points corresponding to certain (m, c) values.

Considered Constraints. Following are assumed constraints while implementing Hough transform-based line detection:

- Slope constraint: If $L_{XT} = \{l_{iXT}(y), \forall i \in \{1, 2, 3, \dots, r\}\}$ where $l_{iXT}, \forall i \in \{1, 2, 3, \dots, r\}$ are line segments with slopes $\{\theta_{iXT}, \forall i \in \{1, 2, 3, \dots, r\}\}$ in any $I_{XT}(y), \forall y \in \{1, 2, 3, \dots, m\}$ then $l_{iXT} \in Y_{XT}$ if $\theta_{low} < \theta_{iXT} < \theta_{high}, \forall i \in \{1, 2, 3, \dots, r\}$
Where θ_{high} is dependent upon global motion in $I(t)_{XY} \forall t \in \{1, 2, 3, \dots, p\}$ and θ_{low} is dependent upon velocity of moving object and clutter in scene.
- Maximum Length constraint: In present approach we are using Hough transform based line detection for labelling representative segments having motion, so few constraints have been imposed on this method. It will increase temporal redundancy if an object is with motion with similar pose is part of for more than few frames. Since ξ_{XT} is generating transformation matrix between V_{XYT} and Ψ_{XYT} , inclined line segments of a fixed slope with more than a threshold length are replaced with inclined line segments of a fixed slope with threshold length. By setting an upper threshold on H matrix of Hough transform line segments more than certain length can be avoided.
- Minimum Length constraint: As it is obvious in real time scenarios that there will be a substantial amount of clutter available in captured scenes in form of unwanted motions due to various causes, e.g., motion in leaves due to wind, it becomes necessary to tackle such scenarios for robustness of proposed approach. By analysing such unwanted motions we can conclude that such motions will also generate incline trajectories in $X - T$ slices but shorter in length, hence by selecting a lower length threshold in H matrix of Hough transform these can be eliminated.

3.5 Labelling of Representative Segments with Motion

From Eq. 4 set of representative segments with motion τ_{XT} is difference of ζ_{XT} as in Fig. 3b and ξ'_{XT} as in Fig. 3a.

$$\tau_{XT} = \zeta_{XT} - \xi'_{XT} \quad (4)$$

3.6 Extraction of Representative Segments with Motion

A sparse set τ_T is generated from τ_{XT} with unity entries corresponding to frame numbers with representative motion segments. This set is used as transformation matrix for obtaining Ψ_{XYT} from V_{XYT} .

4 Results

Results of video synopsis along with video sequences are presented based on our approach on two datasets. As there are very limited datasets available for such sequences, we have tried to generate a robust test bed of thermal imaging sequences captured in different environmental conditions using a 640×512 detecting elements-based MWIR imaging system. Number of frames in temporal compact representation are dependent upon motion content of considered video sequence. There are also few false alarms in form of frames containing no motion information, due to outlier line segments during Hough transform-based line detection. As in Fig. 4a there is *Room* dataset containing an IR video sequence of 1393 frames out of which 939 frames contain object(s) with motion, in its temporal summarized representation there are 525 frames out of which 55 frames are false positives; this implies that we are getting almost 2.65 times compressed sequence. There are three randomly moving persons in this sequence, it can be concluded that almost all important events related with motion are captured in its compact representation. Analysis is also done for *Road* dataset containing a MWIR video sequence *Road-2* of 2440 frames out of which 1425 frames contain object(s) with motion, it is transformed in a compact representation containing 1084 frames with almost 2.25 times compression out of which 243 frames are false positives. Similarly as in Fig. 4b for *Road-1* sequence containing a thermal video of 823 frames we are getting almost two times temporal compact representation with 411 frames.

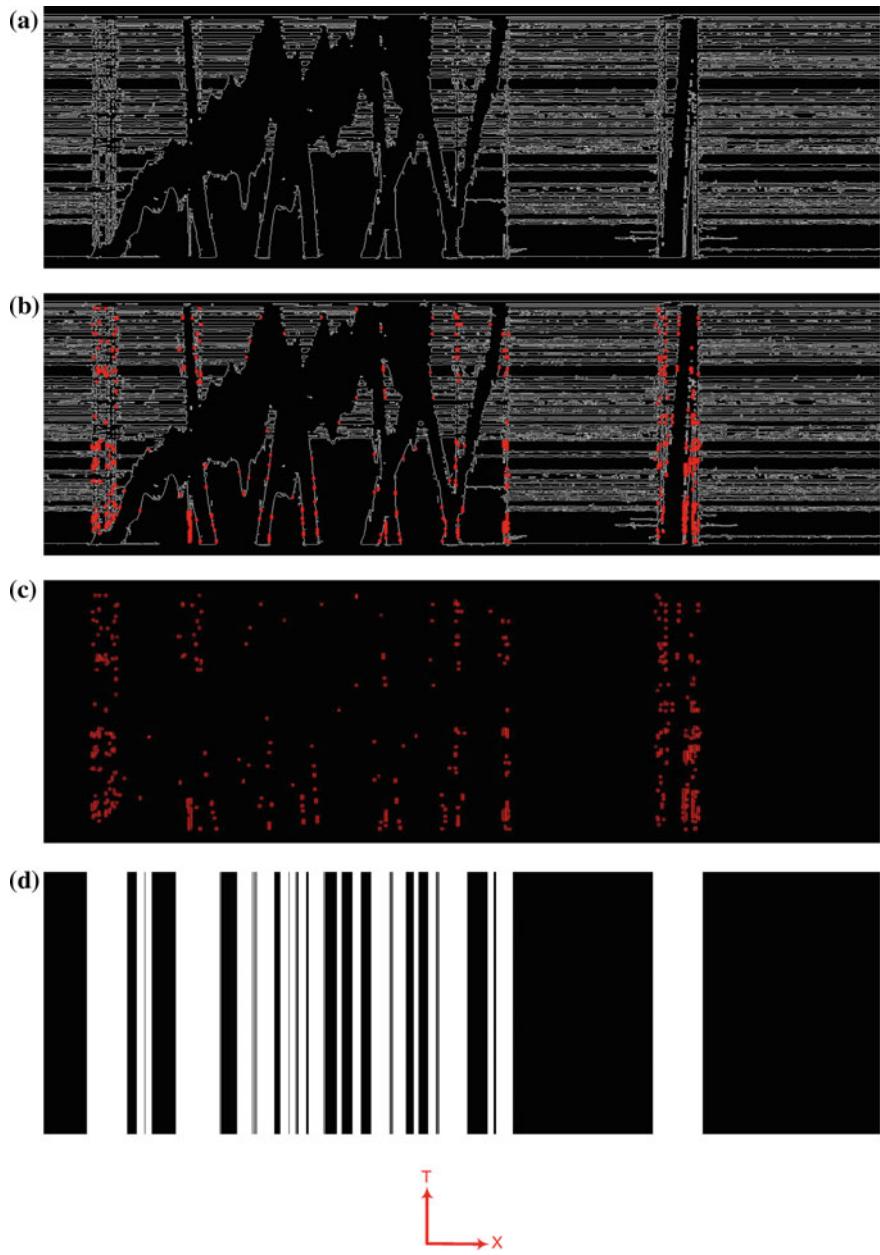


Fig. 3 For *Room* video sequence of 1393 frames X axes representing T (number of frames) and Y axes representing X . **a** ξ'_{XT} : Binary edge map obtained from Canny's edge detection of consolidated binary edge map ξ_{XT} . **b** ζ_{XT} : Result of Hough-based line with imposed constraints (in red). **c** τ_{XT} : Representative segments with motion (in red). **d** Selected frame nos. for compact representation (in white).

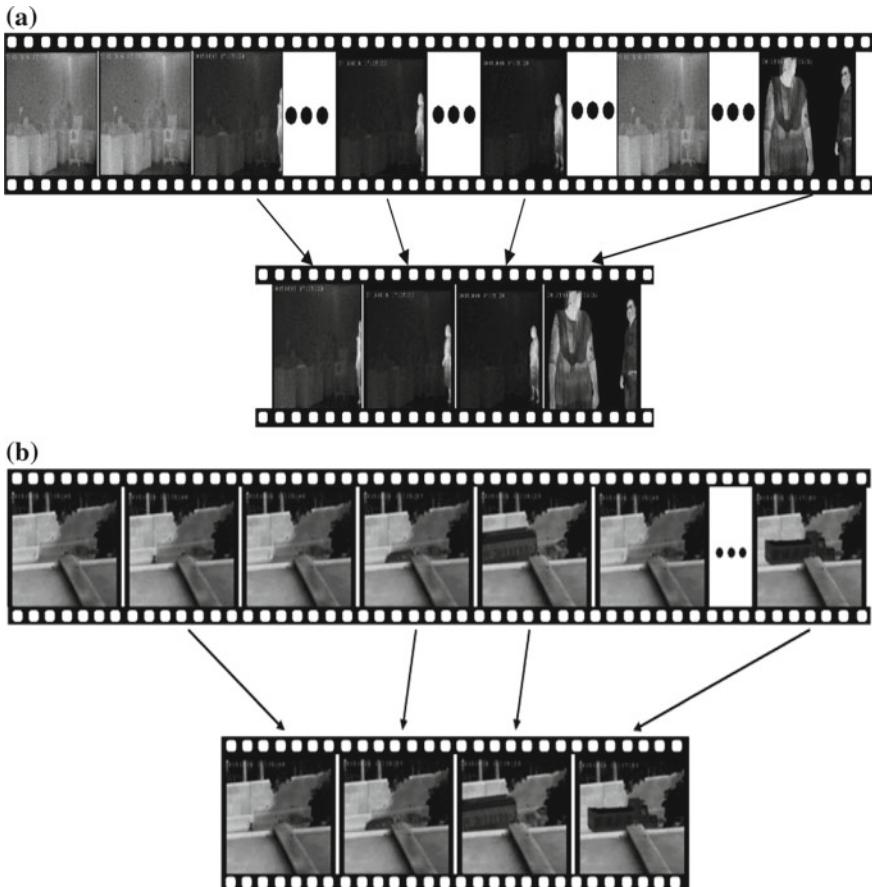


Fig. 4 For both figures **a** and **b** sequence shown above is considered video and sequence shown below is temporal compact representation of considered video **a** Synopsis Generation for *Room* dataset: In its compact representation, there are four non-contiguous frames, first frame corresponds to entrance of person-1, second and third frames correspond to pose change and fourth frame corresponds to entrance of person-2 and **b** Synopsis Generation for *Road-1* sequence: In its compact representation, there are four non-contiguous frames representing entrance of pedestrian, car, bus, and truck, respectively

5 Limitations

Although we have obtained very promising results from present approach there are certain limitations. As Hough transform is a voting-based mechanism for detecting geometries from a set of points and we are using it with some imposed constraints, hence it is obvious that there will be a number of outliers and missing segments. When transformation matrix is generated using these outliers then there are a few frames which unnecessarily become part of temporal compact representation Ψ_{XYT}

and hence we are unable to completely eliminate spatio-temporal redundancy. On the other hand, if the missing segments are part of τ_{XT} then few of important events may be missing from Ψ_{XYT} . Number of such outlier or missing segments can be reduced by adjusting upper and lower thresholds of Canny's edge detection.

6 Conclusion

We considered a novel approach for video synopsis in IR imagery. Although there are a number of approaches suggested in literature yet Hough transform based line detection has barely been used to solve such kind of problems. We are making use of Canny's edge detection and Hough transform based line detection, fortunately both are very old and well established algorithms. This makes implementation aspect of present model very simple. The results are promising barring limitations and model is extremely simple.

Acknowledgements We take this opportunity to express our sincere gratitude to Dr. S.S. Negi, OS and Sc 'H', Director, IRDE, Dehradun for his encouragement. As good things cannot proceed without good company, we would like to thank Mrs Meenakshi Massey, Sc 'C' for not only bearing with us and our problems but also for her support in generating datasets.

References

1. Alatas, Orkun, Pingkun Yan, and Mubarak Shah. "Spatiotemporal regularity flow (SPREF): Its Estimation and applications." Circuits and Systems for Video Technology, IEEE Transactions on 17.5 (2007): 584–589.
2. Ding, Wei, and Gary Marchionini. "A study on video browsing strategies." (1998).
3. Li, Jian, et al. "Adaptive summarisation of surveillance video sequences." Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007.
4. Ji, Zhong, et al. "Surveillance video summarization based on moving object detection and trajectory extraction." Signal Processing Systems (ICSPS), 2010 2nd International Conference on. Vol. 2. IEEE, 2010.
5. Cullen, Daniel, Janusz Konrad, and T. D. C. Little. "Detection and summarization of salient events in coastal environments." Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on. IEEE, 2012.
6. Rav-Acha, Alex, Yael Pritch, and Shmuel Peleg. "Making a long video short: Dynamic video synopsis." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 1. IEEE, 2006.
7. Petrovic, Nemanja, Nebojsa Jojic, and Thomas S. Huang. "Adaptive video fast forward." Multimedia Tools and Applications 26.3 (2005): 327–344.
8. Hoferlin, Benjamin, et al. "Information-based adaptive fast-forward for visual surveillance." Multimedia Tools and Applications 55.1 (2011): 127–150.
9. Porikli, Fatih. "Multi-camera surveillance: object-based summarization approach." Mitsubishi Electric Research Laboratories, Inc., <https://www.merl.com/reports/docs/TR2003-145.pdf> (Mar. 2004) (2004).

10. Paul, Manoranjan, and Weisi Lin. "Efficient video coding considering a video as a 3D data cube." Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on. IEEE, 2011.
11. Liu, Anmin, et al. "Optimal compression plane for efficient video coding." *Image Processing, IEEE Transactions on* 20.10 (2011): 2788–2799.
12. Canny, John. "A computational approach to edge detection." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 6 (1986): 679–698.
13. Azernikov, Sergei. "Sweeping solids on manifolds." *Proceedings of the 2008 ACM symposium on Solid and physical modeling*. ACM, 2008.
14. VC, Hough Paul. "Method and means for recognizing complex patterns." U.S. Patent No. 3,069,654. 18 Dec. 1962.
15. Illingworth, John, and Josef Kittler. "A survey of the Hough transform." *Computer vision, graphics, and image processing* 44.1 (1988): 87–116.

Performance Analysis of Texture Image Retrieval in Curvelet, Contourlet, and Local Ternary Pattern Using DNN and ELM Classifiers for MRI Brain Tumor Images

A. Anbarasa Pandian and R. Balasubramanian

Abstract The problem of searching a digital image in a very huge database is called content-based image retrieval (CBIR). Texture represents spatial or statistical repetition in pixel intensity and orientation. When abnormal cells form within the brain is called brain tumor. In this paper, we have developed a texture feature extraction of MRI brain tumor image retrieval. There are two parts, namely feature extraction process and classification. First, the texture features are extracted using techniques like curvelet transform, contourlet transform, and Local Ternary Pattern (LTP). Second, the supervised learning algorithms like Deep Neural Network (DNN) and Extreme Learning Machine (ELM) are used to classify the brain tumor images. The experiment is performed on a collection of 1000 brain tumor images with different modalities and orientations. Experimental results reveal that contourlet transform technique provides better than curvelet transform and local ternary pattern.

Keywords CBIR • Texture • Brain tumor • DNN and ELM

1 Introduction

In recent years, digital image collection has increased with the rapid growth of the size. The gigabytes of images are generated from military and civilian equipments. We cannot access or make use of the information unless it is organized so as to allow the efficient, browsing, searching, and retrieval. The active research area in

A. Anbarasa Pandian (✉) · R. Balasubramanian
Department of Computer Science and Engineering,
Manonmaniam Sundaranar University, Tirunelveli, India
e-mail: anbuaec@gmail.com

R. Balasubramanian
e-mail: rbalus662002@yahoo.com

image processing is image retrieval. The two major research communities are computer vision and database management. The text-based and visual-based are two research communities in image retrieval. Content-based image retrieval has proposed to overcome those difficulties. The problem of searching for digital image in a very huge database is content-based image retrieval. The text-based keywords, images are manually annotated by their own visual content such as texture, color, shape, etc. [1].

Medical imaging is the technique to create images of the human body for clinical purposes and medical science. The medical image is used in modern techniques like digital radiography (X-ray), ultrasound, microscopic imaging, computed tomography (CT), magnetic resonance imaging (MRI), single photon emission computer tomography (SPECT), and positron emission tomography (SPECT). The MRI uses radio waves originally for brain images like bleeding, aneurysms, tumors, and damages. The MRI is an accurate procedure for hundreds of images in slices per single patient. The T1, T2, and PD are the images produced for specific tissue characteristics of the image. The three types of image orientation in brain images are coronal, sagittal, and axial [2]. The malignant tumor and benign tumor are the two types of brain tumor. The malignant tumor can be classified into primary tumor and secondary tumor. The primary tumor is stored within the brain and secondary tumor that have spread from somewhere else called brain metastasis. The symptoms of brain tumor are headaches, seizures, problem with vision, vomiting, and mental changes [3].

Texture analysis is an important issue used in a variety of image and video applications including image segmentation, image retrieval, object recognition, contour detection, scene classification, and video indexing. Texture features are visual patterns consisting of contrast, uniformity, coarseness, and density. In medical image analysis, the main objective of the texture is used to classify the brain images into gray and white matter of the magnetic resonance (MR) image [4].

Murala et al. [5] developed a new image retrieval algorithm for local mesh pattern using biomedical image retrieval. The significant improvement for retrieval performance LBP with gabor transform and domain methods. Manjunath et al. [6] developed an image retrieval method using gabor texture feature to measure the similarity of the image. The retrieval performance of the texture is useful for region-based retrieval. Pan et al. [7] proposed similarity retrieval uncertain location graph method for brain CT image texture. The similarity image retrieval technique is used to reduce the searching time, high accuracy, and efficiency.

2 Feature Extraction

Feature extraction is a dimensionality reduction that starts with the initial set of measured data for human interpretations. Feature extraction is used to reduce the huge set of data from the resources.

2.1 Preprocessing

The preprocessing step is used to enhance the visual appearance of an image using contrast enhancement techniques. Contrast enhancement will be used to perform adjustments on the darkness or lightness of the image.

2.2 Curvelet Transform

Curvelet Transform (CT) represents a curve-like feature the use of texture and spatial locality information for multiscale directional transform. The digital image processing is important to handle brain visual cortex of images as spatial locality, scale, and orientation. CT is used in the 3-D biological data analysis, seismic data analysis, fluid mechanics analysis, video processing, and partial differential equation analysis. CT is locally implemented ridgelet and is closely related to ridgelets. CT can represent even curve-like features sparsely, whereas ridgelet sparsity is on the straight line like features. Curvelet-based feature extraction is defined as extraction of characteristic and discriminating curvelet coefficient features from the data [8, 9] (Fig. 1).

2.3 Contourlet Transform

Contourlet transform (ConT) is a form of directional multiresolution image representation and made up of a smooth regions partition by smooth boundaries. Contourlet transform is implemented based on two types: Laplace pyramid and directional filter bank. The contourlet transform is based on basis functions with flexible aspect ratios and different directions in multiple scales. It has a small redundancy unlike other transforms. Contourlet has offered a high degree of

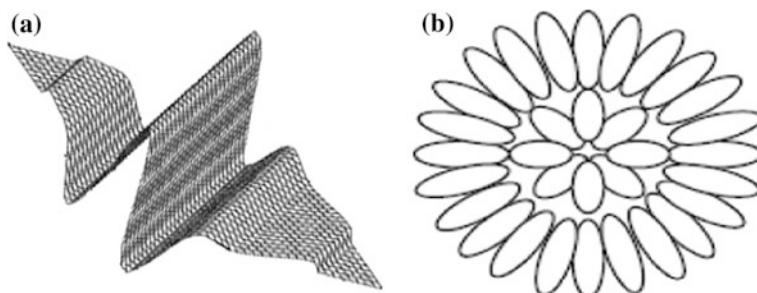


Fig. 1 **a** Ridgelet waveform. **b** Curvelets with different scales

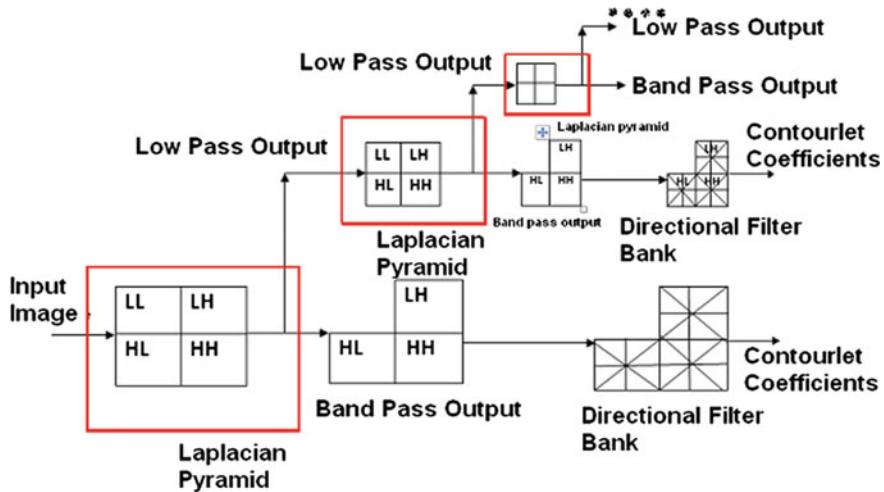


Fig. 2 Decomposition of contourlet transform

directionality and anisotropy and not only access the main features of wavelets like multiscale and time frequency localization. Contourlet is achieving critical sampling, but it takes different and flexible number of directions at each scale [10–12].

Contourlet Transform algorithm.

1. The input image has four frequency components like LL (Low Low), LH (Low High), HL (High Low), and HH (High High).
2. At each stage, Laplacian pyramid produces a low pass output (LL) and a band pass output (LH, HL, HH).
3. The band pass output is passed into a directional filter bank, which produces the results as contourlet coefficient. Then the low pass output is again passed through the Laplacian pyramid to produce more coefficients. The process is repeated until the fine details of the image are retrieved.
4. Finally, the image is reconstructed by applying the inverse contourlet transform (Fig. 2).

2.4 Local Ternary Pattern

Local ternary pattern (LTP) is a texture operator that has higher dimensionality and is more discriminative than LBP. LBP has threshold pixels of 0 and 1 values but LTP uses threshold pixels of 0, 1, and -1 values. Consider p as a neighboring pixel, c as the center pixel, and k as the threshold constant [13, 14] (Fig. 3).

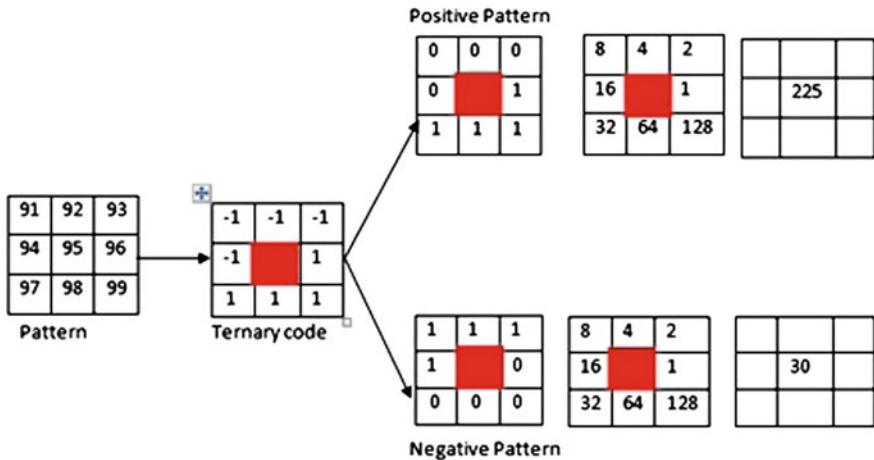


Fig. 3 Calculation of LTP

$$f(x) = \begin{cases} 1, & \text{if } p > c - k \\ 0, & \text{if } p > c - k \text{ and } p > c + k \\ 1, & \text{if } p < c - k \end{cases} \quad (1)$$

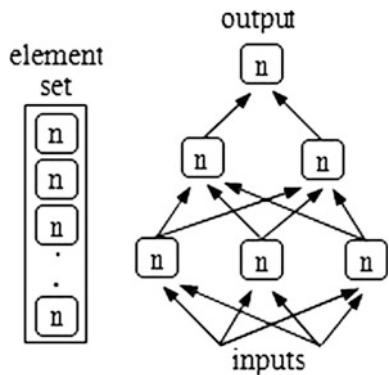
LTP algorithm.

1. Consider a pixel surrounded by eight neighbors.
2. Every pixel in a cell, the left-top, left-middle, left-bottom, right-top, etc., are compared to the center pixel of each of its eight neighbors.
3. The normal pixel values in an image form a values in the pattern matrix.
4. To extract the LTP for the center pixel, C, where the center pixel's value is greater than neighbor's value, assign “1”, center pixel value is less than neighbors value, assign “-1” and center pixel value is equal to neighbors value, assign “0”.
5. Split as positive pattern and negative pattern.
6. Finally, the local ternary pattern value is calculated

3 Classification

Classification is the process in which ideas and objects are recognized, differentiated, and understood. The deep neural network (DNN) and extreme learning machine (ELM) are the classification techniques used for MRI brain tumor images.

Fig. 4 Examples of functions represented by a graph of computations



3.1 Deep Neural Network

Deep neural network (DNN) is a multilayer neural network model that has more than one layer of hidden units between its inputs and its outputs. The two important processes used in the classification are training and testing phases. In the training phase, the features of training data are trained using deep learning classifier. Artificial neural networks provide a powerful tool to analyze, model, and predict. The benefit is that neural networks are data-driven, self-adaptive methods. Commonly used neural network uses back propagation algorithm. But it is not adequate for training neural networks with many hidden layers on large amounts of data. Deep neural networks contain many layers of nonlinear hidden units and a very large output layer. Deep neural networks have deep architectures which have the capacity to learn more complex models than shallow ones. [15, 16] (Fig. 4).

3.2 Extreme Learning Machine

The extreme learning machine is single layer feedforward neural network (SLFNs) which chooses randomly hidden nodes and the output weights of SLFNs. The random choice of input weight and hidden biases of SLFNs can be assigned if the activation functions in the hidden layer are infinitely differentiable. The output weight (linking the hidden layer to the output layer) of SLFNs can be inverse operation of the hidden layer output matrices. The learning speed of the extreme learning machine is thousand times faster than feedforward network algorithm like back propagation (BP) algorithm. The learning algorithm tends to reach the smallest training error, good performance, and obtains the smallest norm of weights. It extremely fast SLFN learning algorithm and easily implemented to extreme learning algorithm [17].

4 Experimental Results

Evaluated with the overall performance of the brain tumor image is the type with sagittal and axial. We used T1 weighted image of MRI brain tumor images. The MRI brain tumor image database contains 1000 images. Brain images contain five classes and each class has 200 images. The resolution of the image is 256×256 . Sensitivity, specificity, accuracy, error rate, and f-measure are the five performance metrics used. For evaluation, both DNN and ELM classification average accuracy are as follows: curvelet transform- 97.5 %, contourlet transform- 97.5 %, and local ternary pattern- 20 %. Let TP be true positive, FN—false negative, FP—false positive, and TN—true negative [18, 19] (Figs. 5, 6, 7 and 8).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (3)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$\text{Errorrate} = 100 - \text{Accuracy} \quad (5)$$

$$F\text{-measure} = \frac{2TP}{2TP + FP + FN} \quad (6)$$

Fig. 5 Retrieval performance of sensitivity, specificity, accuracy, error rate, and F-measure for DNN classifier in CT, ConT, and LTP

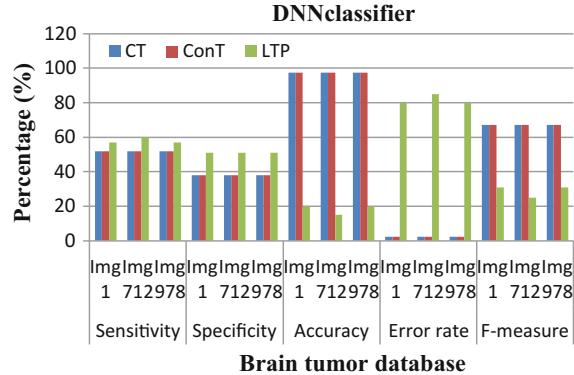


Fig. 6 Retrieval performance of sensitivity, specificity, accuracy, error rate, and F-measure for ELM classifier in CT, ConT, and LTP

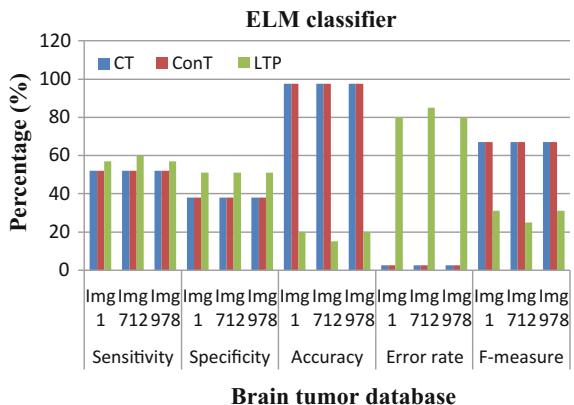


Fig. 7 Retrieval performance of total time taken for DNN classifier in CT, ConT, and LTP

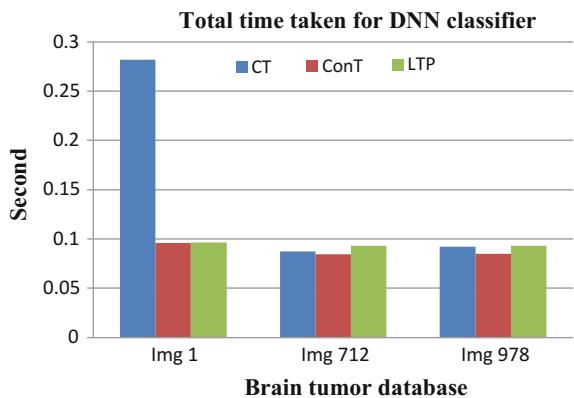
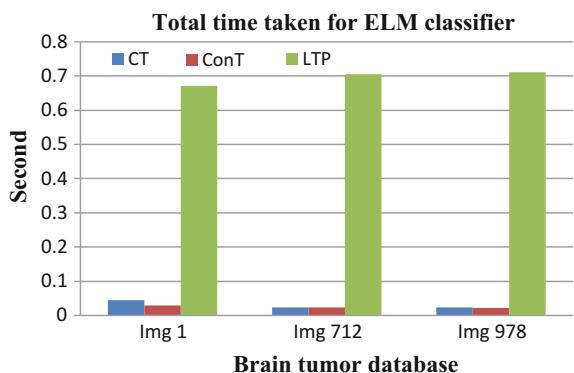


Fig. 8 Retrieval performance of total time taken for ELM classifier in CT, ConT, and LTP



5 Conclusion

In this paper, the performance of texture feature extraction for MRI brain tumor image retrieval is evaluated. The main objective of this study is to investigate and evaluate an effective and robust approach for texture representation and to use it in image retrieval. The curvelet transform, contourlet transform, and local ternary pattern are the techniques used for texture feature extraction. To classify, the brain tumor image for supervised learning algorithms, namely DNN and ELM are used. It is inferred from the results that contourlet transform using DNN and ELM classifier outperform other techniques like curvelet transform and Local ternary pattern. In terms of time, ELM classifier outperforms DNN. Contourlet transform achieves better performance and ELM classifier achieves better performance. It reduces the retrieval time and improves the retrieval accuracy significantly.

References

1. Rui, Yong, Thomas S. Huang, and Shih-Fu Chang. "Image retrieval: Current techniques, promising directions, and open issues." *Journal of visual communication and image representation* 10.1: 39–62, March 1999.
2. Castelli, Vittorio, and Lawrence D. Bergman, Eds. *Image databases: search and retrieval of digital imagery*. John Wiley & Sons, 2004.
3. https://en.wikipedia.org/wiki/Brain_tumor.
4. Castelli, Vittorio, and Lawrence D. Bergman, eds. *Image databases: search and retrieval of digital imagery*. John Wiley & Sons, 2004.
5. Murala, Subrahmanyam, and Q. M. Wu. "Local mesh patterns versus local binary patterns: biomedical image indexing and retrieval." *Biomedical and Health Informatics, IEEE Journal of* 18.3 (2014): 929–938.
6. Manjunath, Bangalore S., and Wei-Ying Ma. "Texture features for browsing and retrieval of image data." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 18.8 (1996): 837–842.
7. Pan, Haiwei, Pengyuan Li, Qing Li, Qilong Han, Xiaoning Feng, and Linlin Gao. "Brain CT Image Similarity Retrieval Method Based on Uncertain Location Graph." *Biomedical and Health Informatics, IEEE Journal of* 18, no. 2 (2014): 574–584.
8. Sumana, Ishrat Jahan, Md Monirul Islam, Dengsheng Zhang, and Guojun Lu. "Content based image retrieval using curvelet transform." In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pp. 11–16. IEEE, 2008.
9. Ma, Jianwei, and Gerlind Plonka. "The curvelet transform." *Signal Processing Magazine, IEEE* 27.2 (2010): 118–133.
10. Manju, K., and Smita Tikar. "Contourlet Transform and PNN Based Brain Tumor Classification." *International Journal of Innovative Research and Development* (2014).
11. Javed, Arshad, Wang Yin Chai, Narayanan Kulathuramaiyer, Muhammad Salim Javed, And Abdulhameed Rakan Alenezi. "Automated Segmentation Of Brain Mr Images By Combining Contourlet Transform And K-Means Clustering Techniques." *Journal of Theoretical & Applied Information Technology* 54, no. 1 (2013).
12. Anbarasa Pandian. A and R.Balasubramainian, "Performance analysis of Texture Image Retrieval in Curvelet Transform, Contourlet Transform and Local Ternary Pattern Using MRI Brain Tumor Images" International Journal in Foundations of Computer Science & Technology (IJFCST) Vol.5, No.6, November 2015.

13. https://en.wikipedia.org/wiki/Local_ternary_patterns.
14. Murala, Subrahmanyam, R. P. Maheshwari, and R. Balasubramanian. "Local tetra patterns: a new feature descriptor for content-based image retrieval." *Image Processing, IEEE Transactions on* 21, no. 5 (2012): 2874–2886.
15. Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *Signal Processing Magazine, IEEE* 29, no. 6 (2012): 82–97.
16. Gladis Pushpa V.P, Rathi and Palani .S, "Brain Tumor Detection and Classification Using Deep Learning Classifier on MRI Images" *Research Journal of Applied Sciences Engineering and Technology*10(2): 177–187, May-2015, ISSN:20407459.
17. Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." *Neurocomputing* 70, no. 1 (2006): 489–501.
18. Somasundaram .K, S.Vijaylakshmi and T.Kaliselvi, "Segmentation of Brain portion from head scans using k means cluster", International Journal of Computational Intelligence and Informatics, Vol. 1 : No. 1, April - June 2011.
19. <https://en.wikipedia.org/wiki/Sensitivity> and specificity.

ROI Segmentation from Brain MR Images with a Fast Multilevel Thresholding

Subhashis Banerjee, Sushmita Mitra and B. Uma Shankar

Abstract A novel region of interest (ROI) segmentation for detection of Glioblastoma multiforme (GBM) tumor in magnetic resonance (MR) images of the brain is proposed using a two-stage thresholding method. We have defined multiple intervals for multilevel thresholding using a novel meta-heuristic optimization technique called Discrete Curve Evolution. In each of these intervals, a threshold is selected by bi-level Otsu's method. Then the ROI is extracted from only a single seed initialization, on the ROI, by the user. The proposed segmentation technique is more accurate as compared to the existing methods. Also the time complexity of our method is very low. The experimental evaluation is provided on contrast-enhanced T1-weighted MRI slices of three patients, having the corresponding ground truth of the tumor regions. The performance measure, based on Jaccard and Dice indices, of the segmented ROI demonstrated higher accuracy than existing methods.

Keywords Segmentation · Discrete curve evolution · Delineation · MRI · GBM · Thresholding

1 Introduction

Image segmentation is ubiquitous in any image analysis system. In the medical image analysis, we need to perform segmentation accurately and efficiently since errors occurring in this step are usually carried over to subsequent processing phases. Accurate segmentation facilitates computerized visualization, feature extraction, and

S. Banerjee (✉) · S. Mitra · B. Uma Shankar
Machine Intelligence Unit, Indian Statistical Institute,
203 B. T Road, Kolkata 700108, India
e-mail: mail.sb88@gmail.com

S. Mitra
e-mail: sushmita@isical.ac.in

B. Uma Shankar
e-mail: uma@isical.ac.in

analysis of the region of interest (ROI). Image segmentation entails pixel wise labeling of the image regions according to similarity in terms of visual characteristics. The segmentation of the ROI is an important step in computer-aided detection and diagnosis (CAD) systems assisting medical practitioners and radiologists in the process of interpreting medical images [14, 20].

Detection of tumors, anomalies, organs of specific interest or any other features in a medical image requires considerable amount of experience and knowledge about visual characteristics of the anatomical features. Image segmentation and detection of such different region of interests (ROIs) is typically performed manually by expert radiologists as part of the whole treatment process. The increasing amount of available data and the complexity of the features of interest made us think differently. It is almost necessary and becoming essential to develop automated delineation system to assist and advance image understanding to allow for reproducible results which are quantifiable for further analysis and inference [3]. Besides this, the existence of inter and intra observer, inter patient and inter scanner variability make computer-guided delineations highly desirable for outlining the tumor or anomalies. Computer-aided quantitative image analysis is thus rising in significance with more and more biomedical researchers exploring the field [5, 15].

1.1 Magnetic Resonance Imaging (MRI) and Glioblastoma Multiforme (GBM)

Glioblastoma multiforme (GBM) is a common and highly lethal primary brain tumor in adults. Since repeated biopsies are challenging in brain tumor patients, brain imaging has become widespread by allowing noninvasive capturing of GBM heterogeneity. One such imaging tool is Magnetic Resonance Imaging (MRI) used routinely in diagnosis, characterization, and clinical management for diseases of the brain. MRI is a powerful and noninvasive diagnostic imaging tool. It works on the principles of magnetic fields and radio frequency waves controlled by a computer to make the body tissues emit distinguishing radio waves (depending on their chemical makeup) of varying intensities in order to map detailed 3D images of soft tissues of the brain. Such images provide insight using the achieved high spatial resolution of normal and diseased tissue molecular signatures (as in solid tumors). Variable image contrast can be achieved by using different pulse sequences while changing the imaging parameters like proton density (PD), longitudinal relaxation time (T1), and transverse relaxation time (T2). Therefore, development of semi-automated medical image segmentation algorithms, which would be more accurate and require negligible user interactions become necessary [1, 12].

1.2 Image Segmentation by Thresholding

Image segmentation can be defined as the process of dividing an image into multiple nonoverlapping partitions with homogeneous properties (for example, similar gray

level intensities, texture, etc.). If I is an image then the segmentation problem is to determine the regions R_1, R_2, \dots, R_m , such that $I = \bigcup_{(i=1)}^m R_i$, where R_i is a connected region; and $R_i \cap R_j = \emptyset$ for all i and j , $i \neq j$; and $C_i \neq \emptyset, \forall i$.

Among the various segmentation strategies, thresholding remains the most efficient in terms of its simplicity in understanding, implementation, as well as processing time. A clearly distinguishable target from the background usually has a bimodal histogram, with the threshold corresponding to its *valley*. In such cases, bi-level non-parametric thresholding algorithms exhaustively search for an optimum threshold. There exists many bi-level thresholding methods in literature [16, 17], including those based on histogram shape, clustering, entropy, and minimum error. However, for real-world images often a single threshold may not suffice to generate a satisfactory segmentation of the ROI and there may not exist any traceable valley. Such scenario led to the development of multilevel thresholding, which computes multiple thresholds to segment images into several classes. Thereby an automatic determination of the appropriate number and values of multiple thresholds that can preserve most relevant details in the original image is still an important issue, therefore an interesting and challenging area of research.

The bi-level methods can easily be extended to a multilevel framework, however, it requires an exhaustive search for determining the optimal set of thresholds while maximizing the optimizing a function. For example, to segment an image with L gray levels in m classes, an exhaustive search needs $L-1 C_m$ combinations of thresholds; thereby involving a huge computational overhead. Several improved algorithms have been proposed in literature [7, 9, 10].

The main objective of this research is to develop a novel semi-automatic, minimally interactive tumor delineation method from MR images. The algorithm produces fast segmentation of the ROI, as compared to that obtained by other interactive segmentation methods. The output is almost as good as those generated manually by radiologists. A Multilevel thresholding is initially employed, based on the meta-heuristic optimization technique called Discrete curve Evolution (DCE) [2], to optimally segment the contrast-enhanced T1-weighted MR image. The tumor is then automatically delineated from only a single seed initialization, on the ROI, by the user; thereby, minimizing user interaction.

2 Meta-Heuristic Optimization for Generating the Intervals from the Histogram

In the multilevel scenario, global thresholding-based segmentation approaches sometimes do over segmentation. Liu et al. [11] proved that the objective function of Otsu's [7, 13] method in multilevel scenario is equivalent to the objective function of the K -means method. The main drawback of K -means clustering is that it cannot discover clusters with non-convex shape or different sizes [6], which is also a limitation of the multilevel Otsu's method. To overcome the drawbacks of traditional multilevel Otsu's thresholding here we have proposed a multilevel thresholding method on localized intervals (see Sect. 3).

Here we apply a meta-heuristic optimization technique [2] over the shape of the histogram, to approximate the contour profile, based on a cost optimization function given in Eq. 1, within a fixed number of iterations, while preserving multiple visually critical points. Subsequently, we detect each of these thresholds with the minimization in terms of the between-class variance. According to the meta-heuristic optimization technique a digital curve P is represented by a set of piecewise linear segments with vertex set $V(P) = \{v_1, v_2, \dots, v_k\}$, and a set of line segments $S(P) = \{s_1, s_2, \dots, s_{k-1}\}$, such that consecutive line segments are: $s_1 = \{v_1, v_2\}, \dots, s_{k-1} = \{v_{k-1}, v_k\}$. For each pair of consecutive line segments s_i and s_{i+1} the cost C is calculated as

$$C = \frac{\theta(s_i, s_{i+1}) \times l(s_i) \times l(s_{i+1})}{l(s_i) + l(s_{i+1})}, \quad (1)$$

where $\theta(s_i, s_{i+1})$ is the angle between segments s_i, s_{i+1} , and $l(\cdot)$ is the length function normalized with respect to the total length of the curve P .

The intensity histogram of a gray level image is first scaled into relative intensity frequencies in the range $[0, 1]$, as $f_i = \frac{f_i - \min}{\max - \min}$ with f_i denoting the frequency of the i^{th} gray level, and max and min corresponding to the maximum and minimum frequency values in the histogram. The normalized histogram profile is represented as a digital curve, with the meta-heuristic optimization technique being applied over $k - (n + 2)$ evolutions, (where $k : k > n + 2$ number of points in the curve), to determine the $n + 2$ most significant points by minimizing the cost F of Eq. (1). Note that the points of the curve, where its shape changes drastically, are denoted as significant points. It is obvious that the significant points typically lie around the peak or valley regions of the histogram. Therefore n intervals are defined around these n significant points. The steps of curve evaluation and interval generation from the histogram of the image using the meta-heuristic optimization technique are given in the Fig. 1. The intensity histogram in Fig. 1a is scaled to generate Fig. 1b. Figure 1c–e show the approximate histogram at the end of 100, 200, 230, and 245 evolutions, followed by the final approximation at $n = 3$, we consider only the three significant points (Point 1, Point 2, Point 3, as evaluated by Eq. (1)) excluding the starting and the ending points, are depicted in Fig. 1f. Each interval is then further processed, using bi-level Otsu's thresholding, to generate the thresholds. Figure 2a depicts the multilevel exhaustive Otsu's thresholding, the points detected by the meta-heuristic optimization technique and thresholds generated by locally applying bi-level Otsu in each intervals. The segmentation by the proposed method is given in Fig. 2b.

The time complexity of the proposed multilevel thresholding algorithm is of the order $O(L^2)$, where L is the number of gray levels. It is thus, independent of the number of thresholds or the size of the image. There is no multilevel thresholding method, reported in literature, which produces such large improvement in computational complexity without compromising on the quality of the thresholding. Our algorithm is less susceptible to noise also, which is one of the inherent characteristics of medical images.

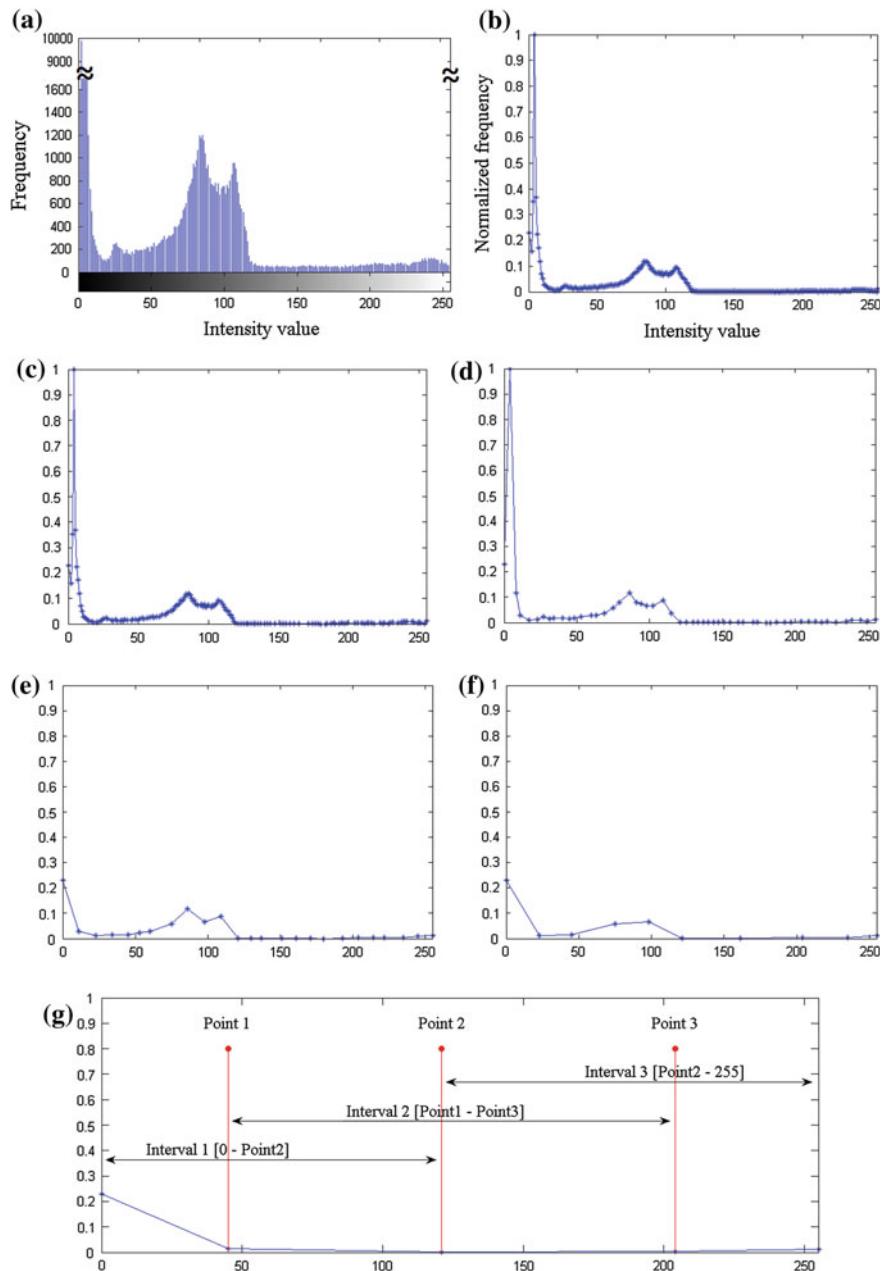


Fig. 1 Steps of the meta-heuristic optimization technique based interval generation from the histogram of the image. **a** Histogram of the image, **b** scaled intensity histogram, meta-heuristic optimization technique approximated histograms after **c** 100, **d** 200, **e** 300, **f** 245 and **g** 250 evaluations

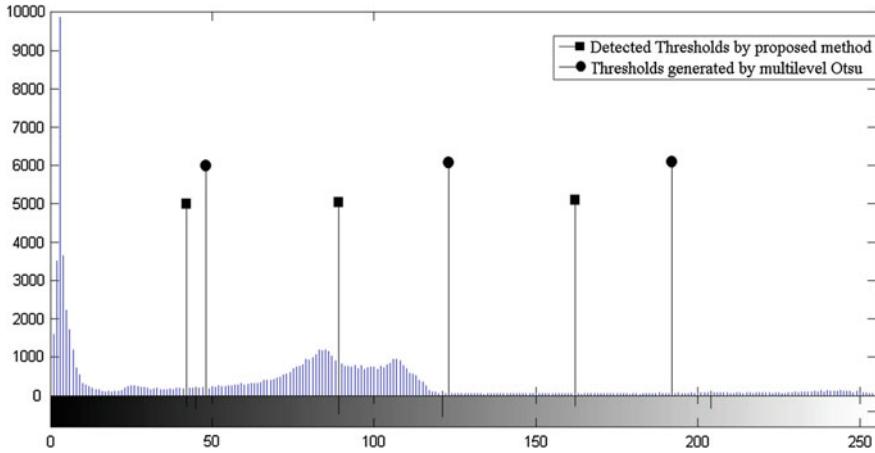


Fig. 2 Thresholds generated by the exhaustive multilevel Otsu's method and by the proposed method

In the next phase the user selects a point (interactively) within the ROI, to extract the ROI from the thresholded image. The seed selection problem is thus minimized in the proposed semi-automatic scheme. Since the segmented image may also contain other objects. A connected-component analysis is applied for labeling the different components based on their similarity, and then extracting the region with the same label as the initial seed pixel. This is followed by the filling of any holes in the ROI, using the flood fill operation [4]. The pseudo code of the proposed method is outlined in Algorithm 1.

Algorithm 1 : Proposed Algorithm

Input: A gray-level image I , and the number of thresholds n .

Output: Segmented image containing only the ROI(s).

Steps

- 1: **procedure** PROPOSED(I, n)
 - 2: Compute the intensity histogram of the input image I .
 - 3: The intensity histogram scaled into relative intensity frequencies in the range $[0, 1]$.
 - 4: Use the meta-heuristic optimization for generating the n intervals from the histogram as outlined in Sect. 2.
 - 5: A threshold is selected, from each of these n intervals, using Otsu's method.
 - 6: Segment the input MR slice into multiple classes using the thresholds generated by the proposed method.
 - 7: User selects the appropriate class (ROI) by placing a single seed point over it.
 - 8: Extract the ROI by using post processing steps like connected component analysis and flood filling.
 - 9: **return** Segmented image containing only the ROI(s).
 - 10: **end procedure**
-

3 Experimental Results

In this section we provide a comparative analysis on the segmentation results of glioma with the proposed method. The results are demonstrated, at first, using a typical example in Fig. 3 and Table 1, which shows a comparison with multilevel Otsu and K-means clustering, as mentioned in Sect. 2. Figure 3a shows a T1-weighted brain MR image and its corresponding manual segmentation (selection of threshold by trial and error) Fig. 3b, with four classes viz. background and CSF (Region 1), gray matter (Region 2), white matter (Region 3), and tumor and skull (Region 4). Figure 3c, d show the segmentation of the same MR image using multilevel Otsu and K-means, respectively, and proposed method in Fig. 3e. The Table 1 illustrates the pixels count in each regions, for these methods. It can be observed from Fig. 3 that multilevel Otsu or K -means fail to produce the proper segmentation of the MR image as the regions are of different sizes (Table 1), which is also justified with the sensitivity scores.

We also considered a more elaborated analysis of the proposed method using MR images of three patients. The comparison is made with ground truth, multilevel Otsu and with the results obtained by a region growing based semi-automatic method called the Grow-Cut [19]. The Grow-Cut has been reported to be a good method for tumor delineation in medical images [18]. The results of Dice and Jaccard index [8] and run time analysis are presented in the Fig. 4 and in Table 2 respectively.

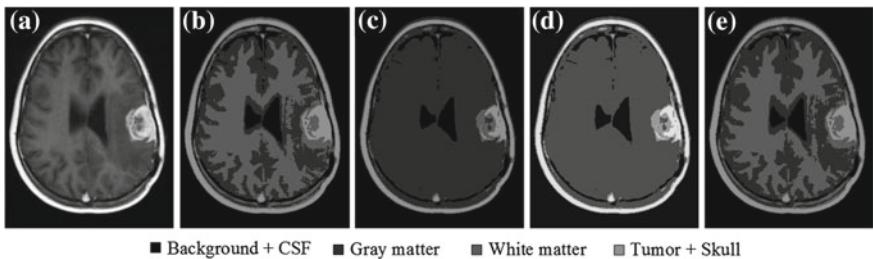


Fig. 3 **a** T1-weighter brain MR image, **b** manually segmentation into four classes, **c** segmentation done by using exhaustive multilevel Otsu, **d** K -means clustering, and **e** segmentation done by the proposed method

Table 1 Number of pixels in different regions in manual segmentation, Otsu's method, K -means clustering, and the proposed method

Regions	Manual	Otsu	K -means	Proposed
Region 1	33126	34316	34531	34707
Region 2	23813	43143	43080	20290
Region 3	22384	3273	3344	24280
Region 4	6437	5028	4805	6483
Sensitivity	–	74.47 %	76.06 %	97.32 %

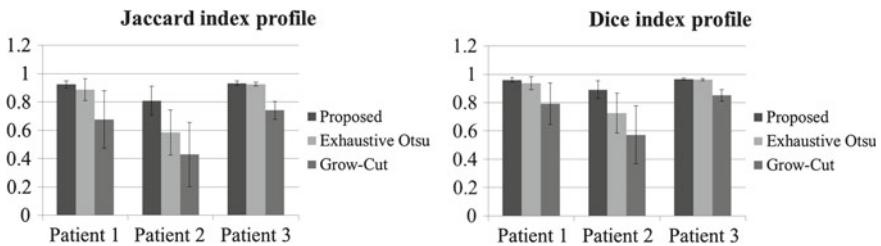


Fig. 4 Jaccard and Dice index profiles for the proposed algorithms, exhaustive Otsu, and Grow-Cut over of 3 patients

Table 2 Average accuracy and runtime (in seconds) for different methods over 30 MR images(slices) for three patients

Mean accuracy over 30 slices (Jaccard and Dice index)

Slice	Proposed		Exhaustive otsu		Grow-Cut	
	Jac	Dice	Jac	Dice	Jac	Dice
Mean	0.89	0.94	0.79	0.88	0.66	0.78
S.D.	0.08	0.05	0.18	0.14	0.22	0.19

Runtime (Sec.)			
Mean	0.99	26.25	3.11
S.D.	0.23	0.52	0.72

The brain tumor images used in this experiment was obtained from a publicly available database.¹ The average accuracy using Jaccard index and Dice index shows that the segmentation accuracy of the proposed method is higher than multilevel exhaustive Otsu and Grow-Cut. The proposed method is also efficient in terms of computational time as observed from Table 2. The computational time was less than 1.0 s on an average for proposed method and it was implemented on a computer with an Intel i5 3.0-GHz processor having 4 GB RAM, in MATLAB R2013. This is a significant improvement in term of quality of segmentation and speed.

Figure 5 illustrates the qualitative evaluation of different segmentation methods over six skull stripped 2D T1-weighted contrast-enhanced brain MR images of the three patients (in the three rows) having GBM. It is visually evident that our algorithm, with only a single seed pixel per ROI, is able to achieve best accuracy (with reference to the ground truth) in extracting the tumor region(s).

¹“Brain tumor image data used in this work were obtained from the MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation (<http://www.imm.dtu.dk/projects/BRATS2012>) organized by B. Menze, A. Jakab, S. Bauer, M. Reyes, M. Prastawa, and K. Van Leemput. The challenge database contains fully anonymized images from the following institutions: ETH Zurich, University of Bern, University of Debrecen, and University of Utah.

Note: the images in this database have been skull stripped”.

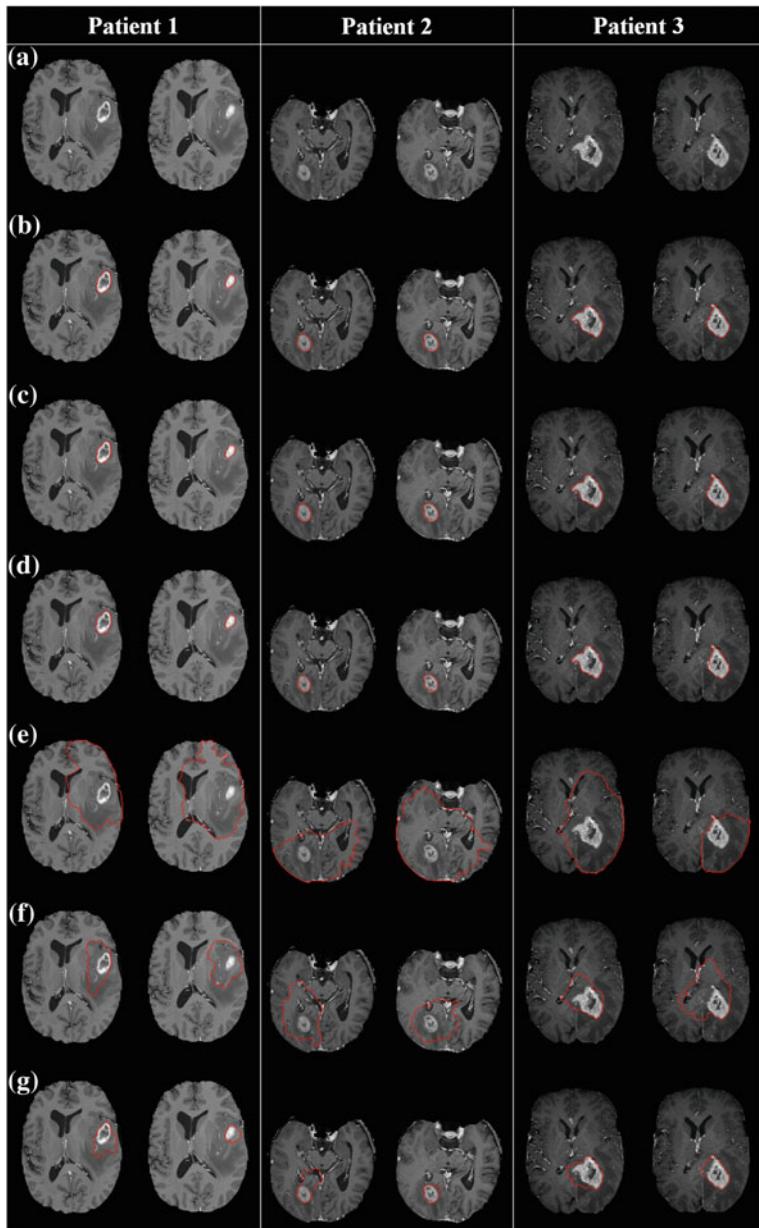


Fig. 5 Six skull-stripped T1-weighted MRI slices of three patients with glioma. **a** Original images (Patient 1: slice no.(100, 103), Patient 2: slice no.(60, 63), Patient 3: slice no.(92, 95)), from *left* to *right*. **b** Expert delineation of the tumor region on the six slices. Segmentation results obtained by **c** proposed method, **d** exhaustive Otsu, and Grow-Cut with one, five, and twelve seed pixels, represented as **e**, **f** and **g**, respectively. (*For interpretation of the references to color in the text, the reader may refer to the web version of this article*)

4 Conclusions

Computer-aided segmentation of the region of interest (ROI) plays the key role in treatment and surgical planning. A wide variety of brain MR image segmentation approaches have been proposed in literature, each having their own advantage and limitations in terms of suitability, applicability, performance, and computational cost. A segmentation method is practically applicable in case of medical images if it is efficient in terms of computational time (preferably producing results in real time), accuracy (i.e., being comparable to the manual segmentation by medical experts) and minimum user interaction. The proposed method is very efficient in terms of computational time, as given in Table 2. With the advancement in hardware the proposed method is expected to produce segmentation within real-time which would be a significant contribution in the field of medical image processing. The segmentation accuracy of the proposed method is also very high with respect to the state-of-the-art methods.

References

1. Bagci, U., Udupa, J.K., Mendhiratta, N., Foster, B., Xu, Z., Yao, J., Chen, X., Mollura, D.J.: Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images. *Med. Image Anal.* 17, 929–945 (2013)
2. Bai, X., Latecki, L.J., Liu, W.Y.: Skeleton pruning by contour partitioning with discrete curve evolution. *IEEE T. Pattern Ana.* 29, 449–462 (2007)
3. Banerjee, S., Mitra, S., Uma Shankar, B., Hayashi, Y.: A novel GBM saliency detection model using multi-channel MRI. *PLoS ONE* 11(1): e0146388 (2016), doi:[10.1371/journal.pone.0146388](https://doi.org/10.1371/journal.pone.0146388)
4. Beucher, S., Meyer, F.: The morphological approach to segmentation: The watershed transformation. *Opt. Eng.* 34, 433–481 (1993)
5. Gatenby, R.A., Grove, O., Gillies, R.J.: Quantitative imaging in cancer evolution and ecology. *Radiology* 269(1), 8–14 (2013)
6. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques. Morgan kaufmann (2006)
7. Huang, D.Y., Wang, C.H.: Optimal multi-level thresholding using a two-stage Otsu optimization approach. *Pattern Recogn. Lett.* 30, 275–284 (2009)
8. Klein, A., et al.: Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage* 46, 786–802 (2009)
9. Liang, Y.C., Cuevas, J.R.: An automatic multilevel image thresholding using relative entropy and meta-heuristic algorithms. *Entropy* 15, 2181–2209 (2013)
10. Liao, P.S., Chen, T.S., C., P.C.: A fast algorithm for multilevel thresholding. *Inf. Sci. Eng.* 17, 713–727 (2001)
11. Liu, D., Yu, J.: Otsu method and k-means. In: Ninth International Conference on Hybrid Intelligent Systems (HIS'09). vol. 1, pp. 344–349. IEEE (2009)
12. Mitra, S., Uma Shankar, B.: Medical image analysis for cancer management in natural computing framework. *Inform. Sciences* 306, 111–131 (2015)
13. Otsu, N.: A thresholding selection method from gray-level histogram. *IEEE T. Syst. Man. Cyb.* 9, 62–66 (1979)
14. Pham, D.L., Xu, C., Prince, J.L.: Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.* 2, 315–337 (2000)

15. Rosenkrantz, A.B., *et al.*: Clinical utility of quantitative imaging. *Acad. Radiol.* 22, 33–49 (2015)
16. Sahoo, P.K., Soltani, S., Wong, A.K.C.: A survey of thresholding techniques. *Comput. Vision Graph.* 41, 233–260 (1988)
17. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. *Electron. Imaging* 13, 146–165 (2004)
18. Velazquez, E.R., Parmar, C., *et al.*: Volumetric CT-based segmentation of NSCLC using 3D-slicer. *Scientific Reports* 3 (2013)
19. Vezhnevets, V., Konouchine, V.: GrowCut: Interactive multi-label N-D image segmentation by cellular automata. In: Proc. of GraphiCon. pp. 150–156 (2005)
20. Withey, D.J., Koles, Z.J.: A review of medical image segmentation: Methods and available software. *Int. J. of Bioelectromagnetism* 10, 125–148 (2008)

Surveillance Scene Segmentation Based on Trajectory Classification Using Supervised Learning

Rajkumar Saini, Arif Ahmed, Debi Prosad Dogra and Partha Pratim Roy

Abstract Scene understanding plays a vital role in the field of visual surveillance and security where we aim to classify surveillance scenes based on two important information, namely scene's layout and activities or motions within the scene. In this paper, we propose a supervised learning-based novel algorithm to segment surveillance scenes with the help of high-level features extracted from object trajectories. High-level features are computed using a recently proposed nonoverlapping block-based representation of surveillance scene. We have trained Hidden Markov Model (HMM) to learn parameters describing the dynamics of a given surveillance scene. Experiments have been carried out using publicly available datasets and the outcomes suggest that, the proposed methodology can deliver encouraging results for correctly segmenting surveillance with the help of motion trajectories. We have compared the method with state-of-the-art techniques. It has been observed that, our proposed method outperforms baseline algorithms in various contexts such as localization of frequently accessed paths, marking abandoned or inaccessible locations, etc.

Keywords Trajectory · Surveillance · HMM · Supervised classification · Scene segmentation · RAG

R. Saini (✉) · P.P. Roy
IIT Roorkee, Roorkee, India
e-mail: rajkr.dcs2014@iitr.ac.in

P.P. Roy
e-mail: proy.fcs@iitr.ac.in

A. Ahmed
Haldia Institute of Technology, Haldia, India
e-mail: arif.1984.in@ieee.org

D.P. Dogra
IIT Bhubaneswar, Bhubaneswar, India
e-mail: dpdogra@iitbbs.ac.in

1 Introduction

Self regulating visual vigilance systems are highly dependent upon motion patterns of moving objects to find out distinct types of activities occurring within the range of sensor. Understanding scene dynamic [1], analyzing traffic patterns [2] or abnormal activity detection [3] can be accomplished through analyzing motion patterns of the objects appear in such scenes. Therefore, computer vision guided approaches are becoming popular to solve some of the aforementioned problems. This is possible largely due to the advancement of object detection and tracking techniques in recent times.

Object detection and tracking has applications in behavior analysis [4], vehicle detection and counting for traffic surveillance [5], anomalous or abnormal activity detection [6–8], foreground segmentation [9] scene segmentation and analysis [1, 10, 11], automatic traffic routing [2], important area segmentation [12], video content analysis [13], etc.

However, context-based scene segmentation still needs close attention of the researchers of computer vision community. As camera-based surveillance has reached almost every corner of our society, it has become necessary to automate such systems for tackling the surveillance task with higher efficiency and lesser depended on human observers. Thus, detection of uncommon movements can be accomplished with computer-based systems. To achieve this goal, installed systems must have better understanding of surveillance scene which cannot be achieved without acceptable scene segmentation. Existing scene segmentation techniques use location of the object center [11] or movement pattern learned through velocity or displacement [10] to partition surveillance scenes.

Unsupervised approach has its own benefits in abnormal activity detection. However, sometimes it is beneficial to use supervised approach for scene understanding since such techniques first learn from the scene and then investigate. In addition to that, supervised learning-based approaches are likely to produce better results because of the availability of ground truths. For learning through supervised approach, one requires training samples and it is desired to have samples within appropriate feature space. It is cumbersome to take decision about the feature space and relevant threshold.

An object's instantaneous position, e.g., (x_i, y_i) at time (t_i) can be used to construct the trajectory of the object. This can be quite helpful to understand the pattern of movement within the scene, however, complex representation of the scene dynamic may not be possible using such simple features. On the contrary, high-level features such as importance of a local block (rectangular) [10] can be useful to understand the scene dynamics. They have proposed a new representation of a scene, referred to as Region Association Graph (RAG). Such a representation can be used to describe the path of a moving object. We have observed that, trajectories of two moving objects may look visually different in-spite of having similar motion paths. Therefore, building a scene model using raw trajectories can be challenging. However, if such trajectories are represented using high-level information, the task can be simplified and

less ambiguous. For example, if trajectories are represented using paths constructed with the help of RAG proposed in [10], classification or grouping of the trajectories will be more accurate. This will definitely lead to better scene segmentation which is the primary objective of this paper.

We have used high-level information such as importance of the nodes and identification numbers as per the RAG to represent raw trajectories. These trajectories are then fed to HMM classifier. Next, a heuristic has been applied to assign the blocks to grow meaningful segments. We compared our technique with the method proposed by Dogra et al. [10] using two datasets, namely, IIT and MIT car datasets. Our proposed method gives better results on scene segmentation as compared to the method proposed in [10].

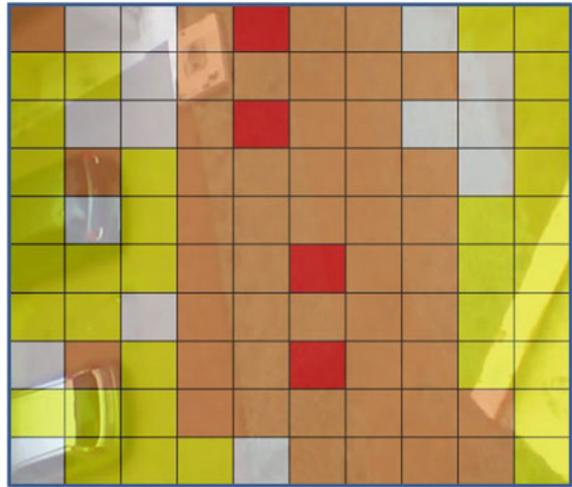
Scene segmentation with the help of trajectory classification is presented in the following section. We have carried out experiments using two datasets and the results of our experiments are presented in Sect. 3. Conclusions of the work are presented in Sect. 4. A few possibilities of future extensions are also mentioned in conclusion section.

2 Scene Segmentation

Raw trajectories are first processed to extract the high-level features using the methodology introduced by the authors of [10, 12]. They have verified the theoretical model of object motion using various publicly available datasets and custom-built dataset. According to their theory, distribution of block importance usually follow a pattern. Thus, if a surveillance scene is divided into rectangular blocks of known dimension, their importance can be estimated using number of objects visiting a block and how much time users or objects spend inside that block. In the next phase, we have constructed a weighted region association graph (RAG) as discussed in [10].

In this proposed framework, high-level features, namely, **node importance** and **node identification**, are used to represent object trajectories. It has been verified that, these two features are important to decide the pattern of movements inside a scene. An example segmentation of a surveillance scene is shown in Fig. 1. The scene is segmented using the aforementioned method with the help of trajectories taken from the IIT human trajectory dataset [10]. We have used different colors to show various types of nodes or regions. The blocks are encoded with unique symbols. Nearby blocks are connected to grow the regions and a final segmentation is achieved. Our scene segmentation approach is solely based on movement patterns and it does not depend upon other information such as texture.

Fig. 1 A typical segmentation map generated using [10] when applied on IIT dataset



2.1 High-Level Feature Extraction

For the purpose of trajectory smoothing and removal of discontinuities or outliers, we have used RANSAC-based approach proposed by [14]. Outliers exist in the signal because of tracking error. Therefore, we need to remove these outliers for better segmentation results. Dogra et al. [12] have extracted some high-level features, such as **label** of a block (b). In their approach, first they have calculated average velocity of a target object from its uniformly sampled trajectory segment. In the next step, they have calculated the total number of times a block is visited by various targets, which is referred to as **global count** or σ_b . Using this count, they filtered out some unimportant blocks. Lastly, they have estimated block importance (τ_b) using Eqs. (1) and (2), respectively.

$$\omega_b = \omega_b + \frac{\bar{v}^{t_i} - v^{t_i}}{\bar{v}^{t_i}} \quad (1)$$

where, ω_b represents the weight of the block b computed from average (\bar{v}^{t_i}) and instantaneous (v^{t_i}) velocities of a target object, t_i .

$$\tau_b = \frac{\omega_b}{\sigma_b} \quad (2)$$

In this paper, we have used the above block importance (τ_b) to construct a high-level feature vector comprising with **label** together with the original (x, y) coordinate values of the trajectory, and the **node-number** of a block as per the region association graph (RAG) proposed in [10]. Hence each point of a trajectory of arbitrary length can be replaced by the four dimensional feature point as given in Eq. (3), where $x(i)$, $y(i)$, $(i)_b$, and $node-number(i)_b$, represent value of x coordinate, y coordinate, label

of block b , and node-number of block b at sequence number i of the time series representation of the original trajectory.

$$F(i) = [x(i), y(i), \text{label}(i)_b, \text{node-number}(i)_b] \quad (3)$$

2.2 Supervised Classification Using HMM

An HMM consists of a number of states, say Q . Each state (j) has an associated observation probability distribution $b_j(O)$ that determines the probability of generating observation O and each pair of states i and j has an associated transition probability a_{ij} . For the purpose of recognition HMM can be learned for each category. HMM can be defined as $\lambda = (A_m, B_m, \pi_m)$, where $m = 1, 2, \dots, C$, $\sum_{m=1}^C \lambda_m = 1$, A_m is a $Q \times Q$ state transition probability distribution matrix, and B_m is the observation probability distribution matrix. For all classifier C , we chose the model that best fits with the observation. This actually tells that with unknown observation and category, we can estimate the posterior $P(\lambda_i|O)$ for each λ_m and select λ_C^* , subject to the condition given in (4),

$$C^* = \arg \max_m P(\lambda_m|O) \quad (4)$$

such that

$$P(\lambda_m|O) = \frac{P(O|\lambda_m)P(\lambda_m)}{P(O)} \quad (5)$$

where $P(O)$ is the evidence that is calculated using (6).

$$P(O) = \sum_1^C P(O|\lambda_m)P(\lambda_m) \quad (6)$$

As HMM is a supervised approach, therefore, we first need to manually classify some of the trajectories to prepare a training set and provide this set as input to the HMM. Classification of trajectories highly depends upon the scene geometry as well as on movement patterns. In general, classification criteria likely to change as the scene and motion patterns change. Based on this, the number of classes may vary. Remaining set of trajectories are fed to the classifier to grouping. using these grouped trajectories, we apply a heuristic to merge blocks of the image. Finally, we get the segmentation of the scene. Segmentation is described in the following section.

2.3 Segmentation of the Surveillance Scene

The scene segmentation heuristic is described in this section. As mentioned earlier, a surveillance scene is first needs to be partitioned into $N \times N$ nonoverlapping blocks. Next, we apply the following methodology to assign or group these blocks based on the number of trajectories passing through these blocks. We may recall, our HMM-based classification has already grouped the trajectories into desired numbers of classes. Therefore, the whole scene is finally segmented into regions depending upon the number of classes in the trajectories plus one. This extra region is introduced because, there may be some blocks that have not visited by any target. The method of segmentation is as follows (Fig. 2).

Let T_i represents the i th trajectory of length n and b_j denotes the j th block of the scene such that B is the set of all blocks. We assume, there are k classes of trajectories, where c_k represents the k th class. Now, for every block, say $b_j \in B$, we determine the set of trajectories λ_{b_j} passing through block b_j . Now, we partition the trajectories from set λ_{b_j} into respective classes. In the next step we find the dominating class within that block having highest number of trajectory footfalls. Suppose c_k be the dominating class for the block b_j , then we assign class c_k to this block. An example of the above block labeling is depicted in Fig. 3. Finally, simple region growing algorithm has been used to merge blocks having similar cluster assignment and connected through 8-connectivity rule.

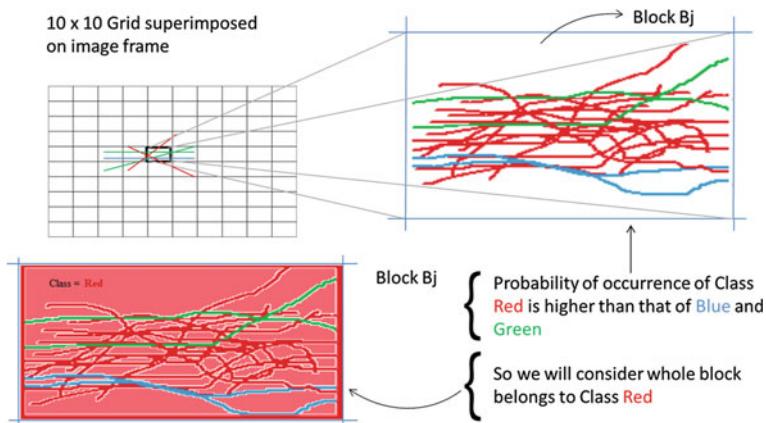


Fig. 2 An example demonstrating the process of association of a block to a particular segment or region

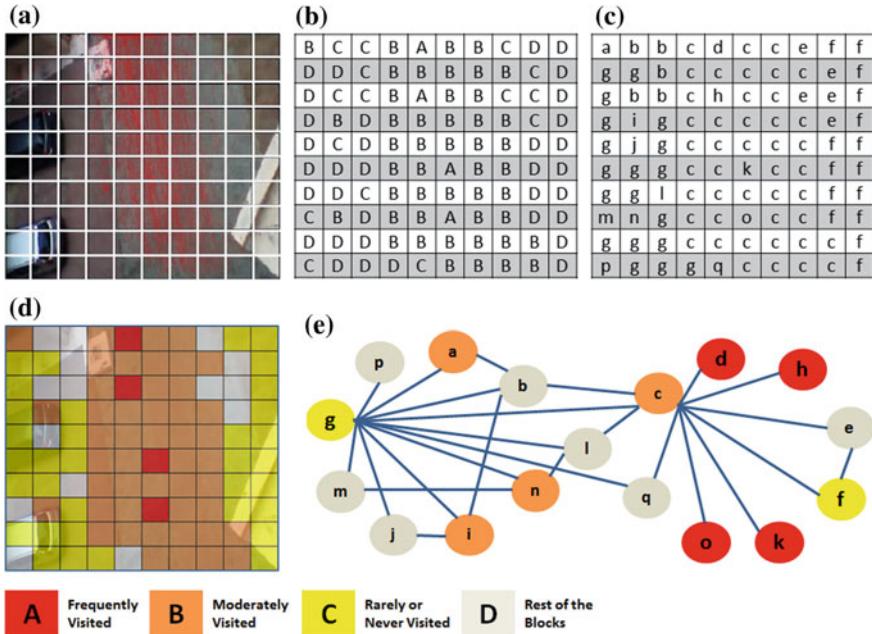


Fig. 3 **a** Image shows surveillance scene of the IIT dataset. The scene image is divided into 100 nonoverlapping blocks of similar dimension and the trajectories are superimposed on the scene. **b** Blocks are labeled using the method proposed in [10]. **c** Creation and representation of the nodes of the graph. **d** A typical color-coded representation of the surveillance scene after segmentation, and **e** Final RAG representation of the scene.

3 Experimental Evaluation

Experimental results are presented and discussed in this section. To support our claim, we have used publicly available datasets in our experiments. A comparative analysis has also been performed to establish the superiority of our approach of scene segmentation. We have compared our results with the method proposed by authors of [10]. Context Tracker proposed by Dinh et al. [15] has been used to extract the raw trajectories representing object movements.

3.1 Description of the Dataset

Our proposed scene segmentation method has been applied on IIT surveillance dataset prepared by the authors of [10] and a car trajectory dataset proposed by Xiaogang et al. [16]. The IIT dataset consists of 191 trajectories representing human

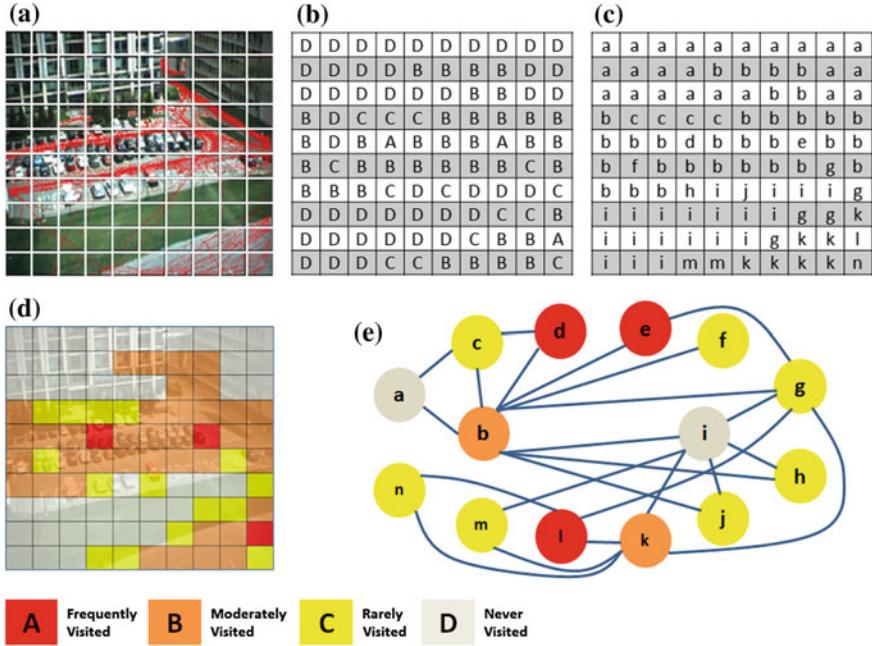


Fig. 4 **a** Surveillance scene of the MIT dataset is partitioned into 100 nonoverlapping blocks of similar dimension and the trajectories are superimposed on the scene **b** Blocks are labeled using the method proposed in [10] **c** Creation and representation of the nodes of the graph **d** A typical color-coded representation of the surveillance scene after segmentation, and **e** Final RAG representation of the scene

movements. As per the authors, primary objective of creating such a dataset was to test scene segmentation and video abnormality detection algorithms. Car dataset is huge in volume (≥ 40 k trajectories), out of which we have randomly selected 400 trajectories to verify our algorithm. Dogra et al. [10] have reported that, their proposed anomaly detection algorithm performs satisfactorily on various other datasets such as VISOR [17] or CAVIAR [18]. However, we have not used these datasets due to less number of trajectories.

We first show the original scenes of IIT and MIT datasets with trajectories plotted over the scenes in Figs. 3a and 4a, respectively. Outputs at various intermediate stages are shown in the subsequent figures.

3.2 Results of HMM Classification

Trajectories were manually labeled and then used for training and testing. HMM was trained and tested using high-level features ‘label’ and ‘node-number’. Based

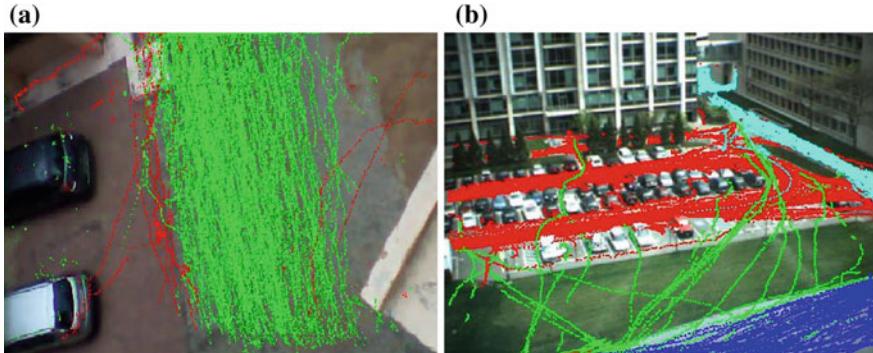


Fig. 5 HMM Classification **a** ($k = 2$) for IIT dataset **b** ($k = 4$) for MIT dataset

on the HMM classification, we segmented the surveillance scene of both datasets. We have trained and tested HMM with varying number of classes. Finally, we have observed that, HMM-based classification produces exciting results for four classes when applied on MIT car dataset. We have classified the IIT dataset trajectories into two classes. It has been observed that, for both datasets, trajectories of suspicious or off-the-track nature have been classified with reasonably high accuracy. Figure 5 presents the HMM classification of both datasets.

Figure 6 shows that, our proposed classification-based segmentation produces better scene segmentation as compared to segmentation obtained by Dogra et al. [10]. Trajectories were preprocessed by a method proposed in [14] to remove the effect of outliers. In Fig. 6b, d, we mark the unvisited blocks with *Tan* color after outlier removal. For IIT dataset, frequently visiting or regular blocks are marked as *Red* and suspicious blocks are marked as *Yellow*. For MIT dataset, frequently visiting nodes are marked as *Red*, moderately visited nodes with *Orange* and *purple* color as they belong to different classes and finally, rarely visited blocks are marked with *Yellow* color.

It can be easily verified from Fig. 6 that, the approach of [10] identifies some blocks as frequently visiting, however, in reality, these blocks have never been visited by any moving target. On the other hand, our proposed segmentation algorithm does a better job. Ground truths and classification results using IIT surveillance dataset and MIT car dataset are presented in Tables 1 and 2, respectively.

4 Conclusion

In this paper, surveillance scene segmentation with the help of trajectory classification using HMM, is introduced. High-level features have been obtained from raw object trajectories and then trajectories were classified into normal and abnormal

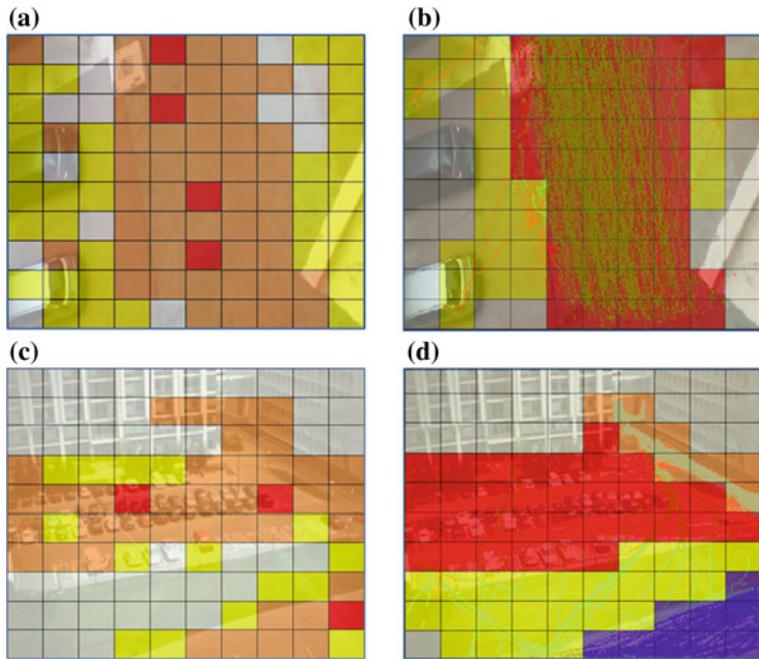


Fig. 6 HMM Classification-based Segmentation **a** ($k = 2$) Segmentation of IIT dataset using [10] **b** Segmentation of IIT dataset using our approach **c** ($k = 4$) Segmentation of MIT car dataset using [10] **d** ($k = 4$) Segmentation of MIT car dataset with our approach

Table 1 Results of IIT dataset (Training: 50 % and Testing: 50 %), GT: Ground Truth

	C1	C2	Result
Dataset: IIT Trajectory = 191; GT: C1 = 190; GT: C2 = 10; K = 2;	91	4	Accuracy = 96.84 % Precision = 75 % Recall = 60 %

Table 2 Results using MIT car dataset (Training: 50 % and Testing: 50 %), GT: Ground Truth

	C1	C2	C3	C4	Result
Dataset: MIT car Trajectory = 400; GT: C1 = 195; GT: C2 = 103; GT: C3 = 77; GT: C4 = 25; K = 4;	92	55	38	13	Accuracy = 97.47 % Precision = 92.31 % Recall = 100 %

movements on MIT parking lot and IIT Bhubaneswar datasets. High-level features are obtained using a recently proposed unsupervised technique to label segments of a given surveillance scene partitioned into nonoverlapping blocks. Our proposed method produces better results (Fig. 6) as compared to the method of [10].

References

1. X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *European Conference on Computer Vision, Proceedings of the*, pages 110–123, 2006.
2. J. Melo, A. Naftel, A. Bernardino, and J. Santos-Victor. Detection and classification of highway lanes using vehicle motion trajectories. *Intelligent Transportation Systems, IEEE Transactions on*, 7(2):188–200, June 2006.
3. C. Piciarelli and G. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15):1835–1842, 2006. Vision for Crime Detection and Prevention.
4. L. Brun, A. Saggese, and M. Vento. Dynamic scene understanding for behavior analysis based on string kernels. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(10):1669–1681, Oct 2014.
5. G. Salvi. An automated nighttime vehicle counting and detection system for traffic surveillance. In *International Conference on Computational Science and Computational Intelligence, Proceedings of the*, 2014.
6. C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1544–1554, Nov 2008.
7. N. Suzuki, K. Hirasawa, K. Tanaka, Y. Kobayashi, Y. Sato, and Y. Fujino. Learning motion patterns and anomaly detection by human trajectory analysis. In *International Conference on Systems, Man and Cybernetics, Proceedings of the*, pages 498–503, 2007.
8. D. Xu, X. Wu, D. Song, N. Li, and Y. Chen. Hierarchical activity discovery within spatio-temporal context for video anomaly detection. In *International Conference on Image Processing, Proceedings of the*, pages 3597–3601, 2013.
9. H. Fradi and J. Dugelay. Robust foreground segmentation using improved gaussian mixture model and optical flow. In *International Conference on Informatics, Electronics and Vision, Proceedings of the*, pages 248–253, 2012.
10. D. Dogra, R. Reddy, K. Subramanyam, A. Ahmed, and H. Bhaskar. Scene representation and anomalous activity detection using weighted region association graph. In *10th International Conference on Computer Vision Theory and Applications, Proceedings of the*, pages 31–38, March 2015.
11. B. Morris and M. Trivedi. Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis. In *International Conference on Advanced Video and Signal Based Surveillance, Proceedings of the*, pages 154–161, 2008.
12. D. Dogra, A. Ahmed, and H. Bhaskar. Interest area localization using trajectory analysis in surveillance scenes. In *10th International Conference on Computer Vision Theory and Applications, Proceedings of the*, pages 31–38, March 2015.
13. D. Dogra, A. Ahmed, and H. Bhaskar. Smart video summarization using mealy machine-based trajectory modelling for surveillance applications. *Multimedia Tools and Applications*, pages 1–29, 2015.
14. Y. Sugaya and K. Kanatani. Outlier removal for motion tracking by subspace separation. *IEICE Trans. Inf. and Syst.*, 86:1095–1102, 2003.
15. T. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1177–1184, 2011.
16. W. Xiaogang, T. Keng, N. Gee-Wah, and W. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1–8, June 2008.
17. Visor dataset. <http://www.openvisor.org>
18. Caviar dataset. <http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1>

Classification of Object Trajectories Represented by High-Level Features Using Unsupervised Learning

Rajkumar Saini, Arif Ahmed, Debi Prosad Dogra and Partha Pratim Roy

Abstract Object motion trajectory classification is an important task, often used to detect abnormal movement patterns for taking appropriate actions to prohibit occurrences of unwanted events. Given a set of trajectories recorded over a period of time, they can be clustered to understand usual flow of movement or detection of unusual flow. Automatic traffic management, visual surveillance, behavioral understanding, and sports or scientific video analysis are some of the typical applications that benefit from clustering object trajectories. In this paper, we have proposed an unsupervised way of clustering object trajectories to filter out movements that deviate large from the usual patterns. A scene is divided into nonoverlapping rectangular blocks and importance of each block is estimated. Two statistical parameters that closely describe the dynamic of the block are estimated. Next, these high-level features are used to cluster the set of trajectories using k-means clustering technique. Experimental results using public datasets reveal that, our proposed method can categorize object trajectories with higher accuracy when compared to clustering obtained using raw trajectory data or grouped using complex method such as spectral clustering.

Keywords Trajectory · Surveillance · Clustering · k-means · Label · Node-no · RAG

R. Saini (✉) · P.P. Roy
IIT Roorkee, Roorkee, India
e-mail: rajkr.dcs2014@iitr.ac.in

P.P. Roy
e-mail: proy.fcs@iitr.ac.in

A. Ahmed
Haldia Institute of Technology, Haldia, India
e-mail: arif.1984.in@ieee.org

D.P. Dogra
IIT Bhubaneswar, Bhubaneswar, India
e-mail: dpdogra@iitbbs.ac.in

1 Introduction

Automatic visual surveillance systems heavily rely on motion patterns of moving targets to understand different types of events. Movement patterns of the objects can be quite informative. They can be used for highway traffic analysis [1], scene semantic analysis [2], anomalous activity detection [3], etc. Research work on computer vision-aided object detection and tracking has advanced significantly during last two decades. Therefore, applications of object tracking have increased leaps and bounds. For example, it is being used for human behavior analysis [4], anomalous activity detection [5–7], semantics analysis [2], automatic traffic management or routing [1], scene segmentation or classification [8, 9], interest area localization [10], video summarization [11], etc.

However, in spite of very good advancement in object detection and tracking; trajectory clustering is yet in its infancy stage. Majority of the existing classification techniques use low-level features such as location of the object center, velocity of the object, pattern of movement, etc. Existing techniques that use high-level features such as region information [12] are limited in scopes and often turn out to be scene specific, domain dependent, and supervised [13]. Unsupervised techniques can be exploited to mitigate above limitation.

However, selection of a good set of features is a challenging task, be it supervised or unsupervised. Simple features like object location (x_i, y_i, t_i) can work in local pattern analysis, however, they cannot be used for analysis of complex patterns. High-level features, such as the importance or label of a block as proposed by Dogra et al. [8] can be used to represent semantic change in a trajectory over its course of execution. For example, the RAG-based segmentation of a scene proposed in their method can be useful to represent a moving object's path. Traces of raw trajectories of two objects, τ_i and τ_j , can be completely different even if they move in similar fashion, therefore, making the clustering process tricky. In this paper, we have used two high-level features proposed in the work of [8], namely **label** and **node-number** of a block to classify object trajectories. Modified representation of a trajectory or path can be fed to a classifier. We have shown through experimental validation that, our proposed high-level feature-based clustering provides better results as compared to low-level feature-based clustering. We have also shown that, our method based on a simple k-means clustering technique done using high-level features outperforms complex methods such as spectral clustering proposed by Ng et al. [14].

The rest of the paper is organized as follows. In the next section, we present the proposed work of trajectory processing and clustering. In Sect. 3, experimental results obtained using two publicly available surveillance datasets, are presented. We conclude in Sect. 4 by highlighting some of the possible extensions of the present work.

2 Classification of Object Trajectories

We have extracted high-level features by processing raw trajectories using the method proposed in [8] that was originally built upon a concept proposed by Dogra et al. [10]. They have assumed a theoretical model of object movement and verified their hypothesis with benchmark datasets. Their model represents the probabilistic importance distribution of localized areas of a surveillance scene. The authors have partitioned a scene into $N \times N$ rectangular blocks and estimated the probabilistic importance or label of a block depending on the number of persons visiting the block and total time spent inside it. Next, a weighted region association graph (RAG) is constructed to represent the scene. This has been used to detect suspicious movements.

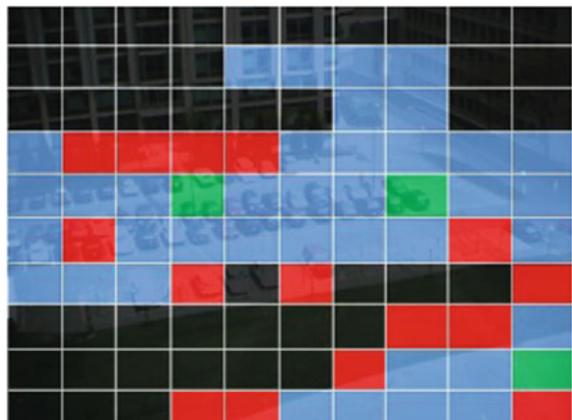
In this paper, we have used **label** and **node-number** as high-level features to represent a raw trajectory. In Fig. 1, we present a sample RAG-based segmentation map of the surveillance scene taken from the MIT car trajectory dataset [15], wherein black colored regions represents never visited, red colored regions are rarely visited, blue colored areas represent moderately visited, and green colored segments are frequently visited blocks. These colored blocks are encoded with decimal values and a region growing technique is adopted to construct the RAG.

2.1 Preprocessing of Labeled Trajectories

Let, τ as mentioned in (1) represents the set of trajectories available for a given scene. Trajectories need not be of identical length.

$$\tau = \{T_1, T_2, \dots, T_N\} \quad (1)$$

Fig. 1 RAG-based segmentation map generated using the method proposed in [8] applied on MIT car dataset [15]



Assume T_i , as given in (2)–(3), is a trajectory of length n . Since we are using k-mean clustering, we need to make sure that the feature vectors become equal length. A heuristic has been applied to accomplish this. We resample each trajectory to a desired number of points. Finally a set of feature vectors, denoted by V , is created where each vector is of equal length.

$$T_i = \{p_1, p_2, p_3, \dots, p_n\} \text{ where } n \text{ is the length of } T_i \quad (2)$$

such that,

$$p_j = \langle x_j, y_j, \text{label}_j, \text{node-number}_j \rangle \quad (3)$$

where (x, y) represents spatial location at particular time instance ‘ j ’ and *label* and *node-number* are calculated using [8, 10].

2.2 Clustering of Labeled Trajectories

Trajectories represented using high-level features (F_v) are then clustered to find groups. We have used k-means clustering for grouping. Length of the feature vector (d) can be chosen empirically. Selecting the number of desired clusters is a tricky process. However, the method proposed by Dogra et al. [8] provides us some clue about the probable optimum value of number of clusters (K). Assume the RAG of a scene contains V nodes. We may recall, the entire scene has been divided into four types of blocks depending upon the importance of each block. Therefore, on an average, $V/4$ number of nodes belong to the same label. Now, we can assume that a person or moving object can visit any one of those $V/4$ number of distinct nodes (with same label). Thus, we have initialized the k-means algorithm with $V/4$. Though the above-mentioned assumption is valid for idealistic situations, however, our experimental evaluation reveals that, it will work on most datasets. The clustering algorithm is discussed briefly in Algorithm 1. The distance used is ‘Euclidean’ distance.

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. Trajectories were extracted using the target detection and tracking algorithm proposed by Dinh et al. [16].

Algorithm 1 Trajectory clustering using high-level features and k-means clustering

```

1: procedure TRAJECTORYCLUSTERING( $F_v, d, K$ )            $\triangleright$  Returns  $C_1, C_2, \dots, C_K$  clusters of
   trajectories using 'label' and 'node-number'
2:  $\triangleright$  Each  $T_i \in F_v$  is a matrix of dimension  $n \times 4$  where 3rd and 4th parameters denote 'label' and
   'node-number'
3:   for  $i = 1 \rightarrow \|F_v\|$  do
4:      $LABEL(i, 1..d) = F_v(1..d, 3)^t$ 
5:      $NODE-NUMBER(i, 1..d) = F_v(1..d, 4)^t$ 
6:   end for
7:   Cluster the arrays 'LABEL' and 'NODENO' using k-means algorithm individually and
   return respective groups.
8: end procedure

```

3.1 Datasets and Ground Truths

Through experiments, we have tested our proposed methodology applied on IIT anomaly detection dataset [8] and MIT car dataset [15]. The first dataset contains a total of 191 distinct human trajectories and it was created for testing scene segmentation and anomaly detection techniques. The car dataset contains 400 car trajectories recorded at a parking lot. Dogra et al. [8] have shown that, the anomaly detection algorithm proposed by them works well on IIT dataset and other public datasets, such as VISOR¹ and CAVIAR.² However, these datasets are not suitable for our application because of insufficient number of trajectories. Thus, we have not tested our method using these datasets. We begin our presentation on results by showing original scenes of IIT and MIT datasets with trajectories superimposed on the scene as depicted in Figs. 2a and 3a. Corresponding level-map, node-number map, scene-segmentation map, and RAG representations are shown in subsequent diagrams of the figures.

3.2 Results of Clustering

Clustering results of trajectories taken from IIT dataset are presented first. Since there are only 17 nodes in the RAG of the scene (refer to Fig. 2), we have applied clustering assuming four clusters as $V/4 = 4$. However, we wanted to separate only those trajectories that visited any one of the red nodes. For comparisons, we have applied our algorithm by varying $K = \{2, 3, 4\}$. Our observations reveal that, $K = 2$ provides better results. The method is able to segregate those trajectories that describe voluntary or unintentional movements of objects visiting inaccessible regions or regions marked as red. Experimental results are presented in Fig. 4.

¹<http://www.openvisor.org>.

²<http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/>.

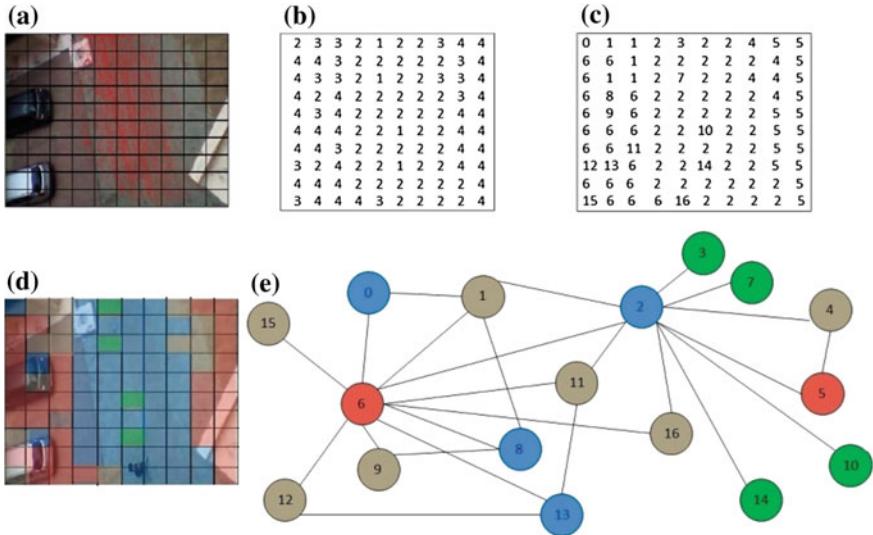


Fig. 2 **a** The background scene of IIT dataset is partitioned into 10×10 number of blocks with overlayed trajectories. **b** Labeling of the blocks using the method described in [8]. **c** Construction of the graph nodes using labels. **d** Color-coded representation of the segmented scene, and **e** RAG corresponding to the segmentation

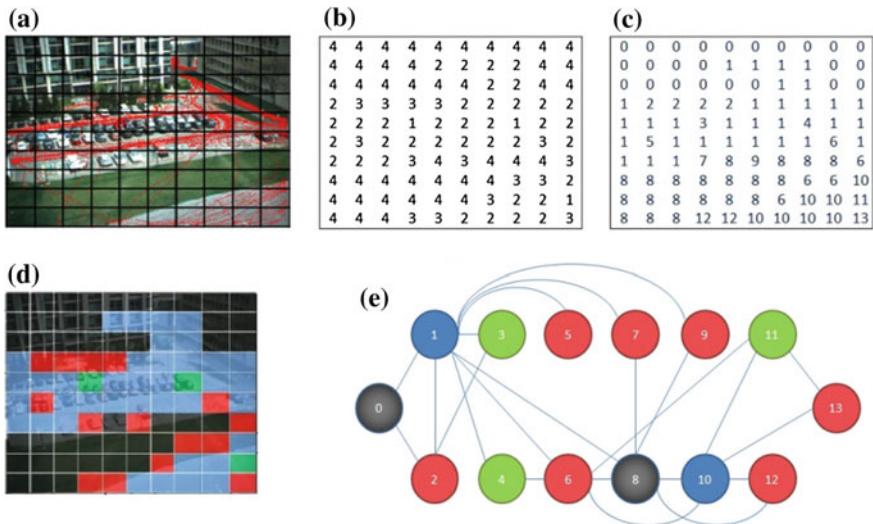


Fig. 3 **a** The background scene of MIT dataset as partitioned into 10×10 number of blocks with overlaid trajectories. **b** Labeling of the blocks using the method described in [8]. **c** Construction of the graph nodes using labels. **d** Color-coded representation of the segmented scene, and **e** RAG corresponding to the segmentation

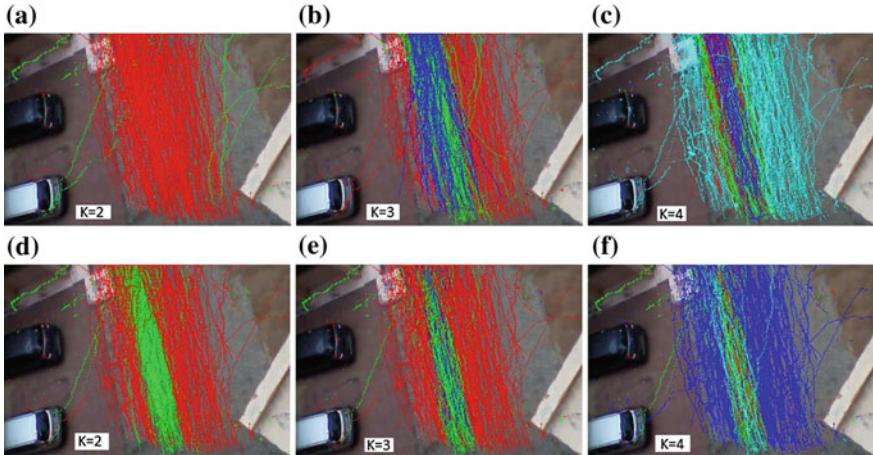


Fig. 4 a–c Clustering of the trajectories of the IIT dataset using **label** by varying $K = \{2, 3, 4\}$. **d–f** Clustering of the trajectories using **node-number** by varying $K = \{2, 3, 4\}$

It can be verified from Fig. 4, best separation is observed when **label** is used as feature in clustering with $K = 2$. On the other hand, when $K > 2$, results degrade. Use of **node-number** to filter out outlier trajectories has also been tried, however, it has failed to predict correct grouping. This has happened because, even clean trajectories that pass through the central region of the scene, usually get separated due to mutually opposite directions if node-number is used as the feature. This can be explained with an example. Let, a person moves from bottom portion of the scene to the top as per the segmentation shown in Fig. 2d. He or she can achieve this by visiting the nodes in the following sequence: $2- > 3- > 2- > 7- > 2$. Ideally, if any other person moves from top portion of the scene to the bottom through $2- > 7- > 2- > 3- > 2$ sequence, it should be considered similar with the previous pattern except a change in direction. But, the classifier usually assumes both to be distinct. However, this is unlikely to happen when we use **label** as the high-level feature since label of similar, but distantly located blocks, can be same.

Clustering results obtained using MIT car dataset trajectories, are presented in Fig. 5. We have recorded best results using $K = 3$. This happens because, the scene has two spatially separated regions of high trajectory density as marked by segments blue blocks; one along the bottom-right corner of the scene as shown in Fig. 3d and the other is toward the central part. Remaining blocks are either rarely visited (red) or never visited (black), except three isolated, but heavily visited green blocks.

A summary of the results is presented in Table 1. We have conducted experiments using various combinations of low-level (x, y) as well as high-level (**label** and **node-number**) features. Result of one such combination (**label + node-number**) is presented in Fig. 6. Our experiments reveal that, other combinations do not perform satisfactorily.

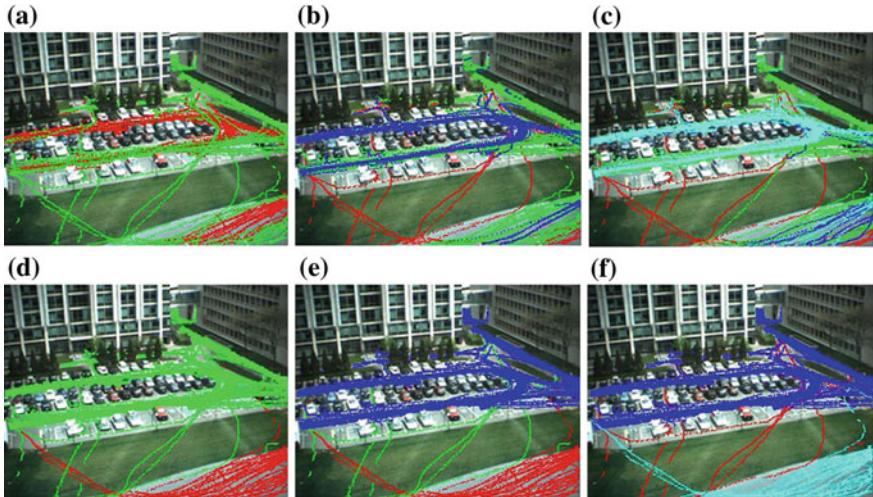


Fig. 5 a–c Clustering of the trajectories of the MIT dataset using **label** by varying $K = \{2, 3, 4\}$. d–f Clustering of the trajectories using **node-number** by varying $K = \{2, 3, 4\}$

3.3 Comparative Analysis

We have compared our proposed trajectory clustering algorithm with spectral clustering method proposed by Ng et al. [14]. In spectral clustering-based classification, we have computed distance separately using Dynamic Time Warping (DTW) [17] and Minimum Variance Matching (MVM) [18].

Results of spectral clustering (using high-level feature) are presented in Fig. 7. It may be observed that, spectral clustering results are not as accurate as we achieved using our proposed high-level feature-based clustering. Computational overhead of our proposed algorithm implemented using MATLAB R2013a on a personal computer with Intel's Core i3 (3rd generation processor) with 4GB RAM and 1GB graphics card running with Windows 8, is presented in Table 2. It can be verified from the table that, the proposed algorithm runs faster as compared to above two baseline methods.

4 Conclusion

In this paper, a novel object trajectory classification methodology has been proposed. We have used high-level features extracted from raw trajectories and applied them to cluster normal and abnormal trajectories of two publicly available datasets. High-level features are extracted using a recently proposed unsupervised way of block-based labeling of surveillance scene. Our proposed method provides better

Table 1 Summary of results using **label** and **node-number** as features applied on both datasets by varying $K = \{2, 3, 4\}$, GT: Ground Truth, C1, C2, C3, and C4 are corresponding clusters

	Feature	K = 2			K = 3			K = 4					
		C1	C2	Results	C1	C2	C3	Results	C1	C2	C3	C4	Results
Dataset: IIT, Trajectory: 191 GT; C1 = 185, GT; C2 = 6	Label	183	8	Precision = 75 % Recall = 100 %	51	46	94	Precision = 6.38 % Recall = 100 %	41	38	24	88	Precision = 6.81 % Recall = 100 %
Node-number	75	116	Precision = 3.44 % Recall = 66.67 %	26	39	126	Precision = 3.17 % Recall = 66.67 %	24	14	27	126	Precision = 3.17 % Recall = 66.67 %	
Dataset: MIT, Trajectory: 400 GT; C1 = 298, GT; C2 = 76, GT; C3 = 26	Label	171	229	Precision = 11.35 % Recall = 100 %	316	34	50	Precision = 45.09 % Recall = 88.46 %	218	99	34	49	Precision = 46.93 % Recall = 88.46 %
Node-number	89	311	Precision = 13.48 % Recall = 46 %	303	84	13	Precision = 80 % Recall = 46.15 %	295	83	9	13	Precision = 78.57 % Recall = 42.30 %	

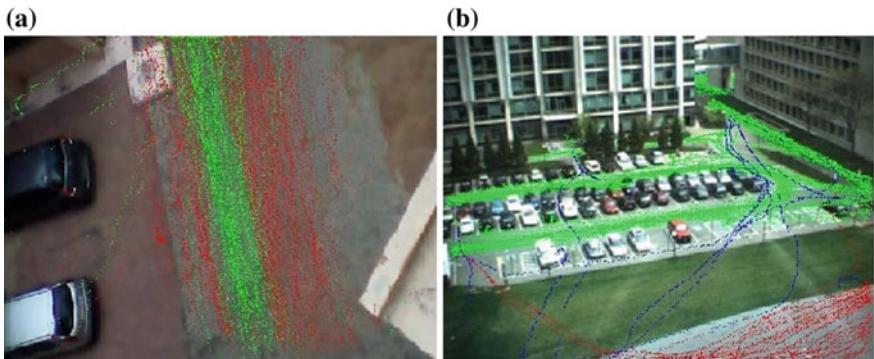


Fig. 6 Result of clustering using label and node-number in combination. **a** Partitioning of the trajectories using $K = 2$ IIT dataset. **b** Partitioning of the trajectories using $K = 3$ MIT car dataset

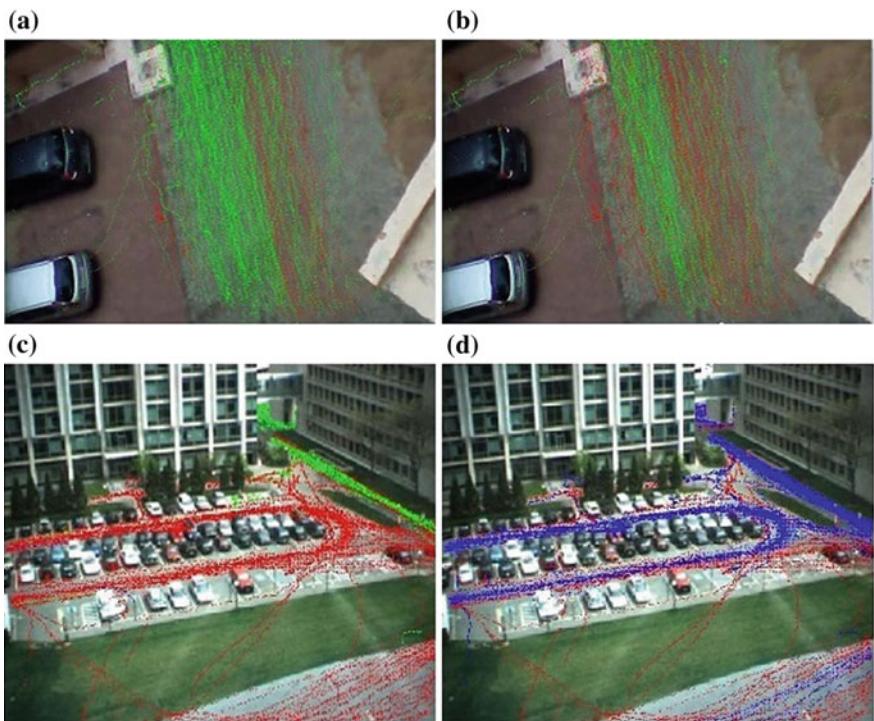


Fig. 7 Result of spectral clustering-based method proposed in [14] using ‘level’ feature applied on both datasets **a** Distance measure using DTW [17] with $K = 2$. **b** Distance measure using MVM [18] with $K = 2$. **c** Distance measure using DTW [17] with $K = 3$. **d** Distance measure using MVM [18] with $K = 3$

Table 2 Comparison of computational overhead against baseline clustering

	Execution time in seconds	
	IIT dataset	MIT dataset
Proposed (high-level feature + k-means)	1.599389	1.564383
DTW + Spectral clustering	80.231930	173.619898
MVM + Spectral clustering	90.163235	179.059711

results as compared to complex and time consuming techniques such as spectral clustering in combinations with DTW and MVM. The proposed trajectory clustering has several applications including traffic management, suspicious activity detection, crowd flow analysis, etc.

References

1. J. Melo, A. Naftel, A. Bernardino, and J. Santos-Victor. Detection and classification of highway lanes using vehicle motion trajectories. *Intelligent Transportation Systems, IEEE Transactions on*, 7(2):188–200, June 2006.
2. X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *European Conference on Computer Vision, Proceedings of the*, pages 110–123, 2006.
3. C. Piciarelli and G. Foresti. On-line trajectory clustering for anomalous events detection. *Pattern Recognition Letters*, 27(15):1835 – 1842, 2006. Vision for Crime Detection and Prevention.
4. L. Brun, A. Saggese, and M. Vento. Dynamic scene understanding for behavior analysis based on string kernels. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(10):1669–1681, Oct 2014.
5. C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1544–1554, Nov 2008.
6. N. Suzuki, K. Hirasawa, K. Tanaka, Y. Kobayashi, Y. Sato, and Y. Fujino. Learning motion patterns and anomaly detection by human trajectory analysis. In *International Conference on Systems, Man and Cybernetics, Proceedings of the*, pages 498–503, 2007.
7. D. Xu, X. Wu, D. Song, N. Li, and Y. Chen. Hierarchical activity discovery within spatio-temporal context for video anomaly detection. In *International Conference on Image Processing, Proceedings of the*, pages 3597–3601, 2013.
8. D. Dogra, R. Reddy, K. Subramanyam, A. Ahmed, and H. Bhaskar. Scene representation and anomalous activity detection using weighted region association graph. In *10th International Conference on Computer Vision Theory and Applications, Proceedings of the*, pages 31–38, March 2015.
9. B. Morris and M. Trivedi. Learning and classification of trajectories in dynamic scenes: A general framework for live video analysis. In *International Conference on Advanced Video and Signal Based Surveillance, Proceedings of the*, pages 154–161, 2008.
10. D. Dogra, A. Ahmed, and H. Bhaskar. Interest area localization using trajectory analysis in surveillance scenes. In *10th International Conference on Computer Vision Theory and Applications, Proceedings of the*, pages 31–38, March 2015.
11. D. Dogra, A. Ahmed, and H. Bhaskar. Smart video summarization using mealy machine-based trajectory modelling for surveillance applications. *Multimedia Tools and Applications*, pages 1–29, 2015.

12. J. Lee, J. Han, X. Li, and H. Gonzalez. Traiclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proceedings of the VLDB Endowment*, 1(1):1081–1094, 2008.
13. J. Lee, J. Han, X. Li, and H. Cheng. Mining discriminative patterns for classifying trajectories on road networks. *Knowledge and Data Engineering, IEEE Transactions on*, 23(5):713–726, 2011.
14. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems, Proceedings of the*, pages 849–856, 2001.
15. W. Xiaogang, T. Keng, N. Gee-Wah, and W. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1–8, June 2008.
16. T. Dinh, N. Vo, and G. Medioni. Context tracker: Exploring supporters and distracters in unconstrained environments. In *Computer Vision and Pattern Recognition, Proceedings of the IEEE Computer Society Conference on*, pages 1177–1184, 2011.
17. I. Nakanishi, H. Sakamoto, N. Nishiguchi, and Y. Fukui. Multi-matcher on-line signature verification system in dwt domain. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 89(1):178–185, 2006.
18. L. Latecki, V. Megalooikonomou, Q. Wang, and D. Yu. An elastic partial shape matching technique. *Pattern Recognition*, 40(11):3069–3080, 2007.

A Hybrid Method for Image Categorization Using Shape Descriptors and Histogram of Oriented Gradients

**Subhash Chand Agrawal, Anand Singh Jalal
and Rajesh Kumar Tripathi**

Abstract Image categorization is the process of classifying all pixels of an image into one of several classes. In this paper, we have proposed a novel vision-based method for image categorization is invariant to affine transformation and robust to cluttered background. The proposed methodology consists of three phases: segmentation, feature extraction, and classification. In segmentation, an object of interest is segmented from the image. Features representing the image are extracted in feature extraction phase. Finally, an image is classified using multi-class support vector machine. The main advantage of this method is that it is simple and computationally efficient. We have tested the performance of proposed system on Caltech 101 object category and reported 76.14 % recognition accuracy.

Keywords Segmentation • K-means clustering • Histogram of oriented gradients • Shape descriptors • Codebook

1 Introduction and Related Work

In computer vision, the problem of image categorization is very easy task for human being but it is very difficult to recognize and categorize an image for the machine. This problem becomes complex due to the several factors such as complex and cluttered backgrounds, position of an object, rotations, shape variations, illumination effects, and occlusion in an image. The segmentation of an image is problematic when image consists of several objects very close to each other or

S.C. Agrawal (✉) · A.S. Jalal · R.K. Tripathi
Department of Computer Engineering and Applications,
GLA University, Mathura, India
e-mail: subhash.agrawal@gla.ac.in

A.S. Jalal
e-mail: asjalal@gla.ac.in

R.K. Tripathi
e-mail: rajesh.tripathi@gla.ac.in

mixed with other objects. This problem is helpful in content-based image retrieval, where the goal is to extract more accurate and faster image search results. The problem of Image categorization mainly consists of three phases: segmentation, feature extraction, and classification. In segmentation, region of interest (ROI) has been extracted from the image. In feature extraction, various features which uniquely represent the image are stored as a feature vector. Finally in the classification, an image is tested against trained classifier and provides the label of category. Some approaches skip the step of segmentation and directly apply the features to the images. However, these methods have low recognition accuracy. The accuracy of the method depends on the segmentation result, i.e., all objects must be removed except required object.

In literature, there are various approaches which exist to categorize the object in the image. Fei-Fei et al. [1] used a generative probabilistic model for the shape and appearance feature of an object. Bayesian classifier is trained with these features. The algorithm was also tested on batch Bayesian and maximum likelihood classifier. The incremental and batch have comparable result but both outperforms maximum likelihood. This method was tested on 101 image category Caltech dataset. Classifier is trained with 1, 3, 6, and 15 examples of each category. The performance of proposed method is worst when training set size is large.

Perronnin et al. [2] utilized bag of visual words for feature extraction and support vector machine for classification on large image sets. SIFT features at multiple scales and codebooks of 4000 visual words are trained. Experiment is performed on three standard benchmarks: Caltech 101, PASCAL VOC07 and ImageNet. With linear classifiers on square-rooted BOV vectors, accuracy is 54.4 % on Caltech 101 and 49.0 % on VOC07. This approach focused on large training set but did not deal with large number of categories.

Csurka et al. [3] presented a simple, computationally efficient method and affine invariant method for visual categorization. This method consists of bag of keypoints approach corresponding to a histogram of the number of occurrences of a particular pattern in an image and then multi-class classifiers are trained using these bags of keypoints. Two descriptors, Harris affine detector and SIFT are computed in feature extraction phase. Two classifiers, Naïve Bayes and SVM are used for categorization. The performance of this method was tested on seven category database.

Grauman and Darrell [4] proposed a new fast kernel-based approach which maps unordered feature set to multiresolution histograms and computes a weighted histograms intersection. In pyramid match kernel input sets are converted to multiresolution histograms. For object recognition, author uses SVM classifiers, which are trained by matrix of kernel values between all pairs of training examples. The method was tested on ETH-80 database with eight object classes, for a total of 400 images. A Harris detector is used to find interest point and several descriptors such as SIFT, JET, patches are used to compute feature vector. Author also tested performance on 101 Caltech database. For each object, 30 samples are used for training and eight runs using randomly selected training sets yield a low accuracy of 43 % on the remainder of other database.

Favorskaya and Proskurin [5] presented unsupervised image categorization method which is robust to scales, shifts, rotations, and lighting conditions. The images are segmented using J-SEG algorithm and only 5–7 large area segments were involved in the categorization. Author used color G-SURF descriptor for the regions. Visual words are constructed using K-means algorithm and matching is performed using SVM algorithm. The experiment was performed on OT8 dataset includes 2688 images with 256 * 256 pixels. From each category, 100 images are selected for training and other set of images from each category was used for testing.

Yao et al. [6] presented fine-grained image categorization approach for image categorization. Author proposed a random forest with discriminative decision tree algorithm to find image patches. SIFT descriptor on each image with spacing of four pixels was applied. Using K-means algorithm, a vocabulary of visual words was constructed. The method was evaluated on two fine grained categorization tasks: human activity recognition and animal species. Human activity recognition dataset PASCAL VOC2010 and subordinate object categorization dataset of 200 birds. This method yields 64.6 % accuracy with PASCAL VOC2010 dataset.

Hoi [7] proposed regularized max pooling. Image is divided into multiple sub-windows at multiple location and scales. The method was tested on PASCAL VOC2012 dataset for human activity recognition on still images and reported 76.4 % accuracy.

Zhang et al. [8] proposed a SVM-KNN-based approach for visual categorization. Shape and texture were extracted and the method was tested on 101 object category and achieved a correct classification rate of 59.05 at 15 training images per class and 66.23 at 30 training images.

In this paper, we have a proposed a novel approach for image categorization which is computationally efficient and tested on large category of objects unlike previous works which were tested on handful of objects category. We have used robust features that are invariant to scale, position, rotation, and lighting conditions. Our method requires less number of training images for classification.

This paper is structured as follows: section II describes the proposed methodology. A description of the experiments undertaken and their results can be found in section III. Finally, the last section presents conclusion and future work.

2 Proposed Methodology

Figure 1 shows the framework of proposed method. Our approach consists of three phases: object segmentation, feature extraction, and classification. In the segmentation phase, an object is segmented from the image using color-based segmentation, features representing the particular object are extracted in feature extraction phase and classifier is trained on these features and object is categorized in classification phase.

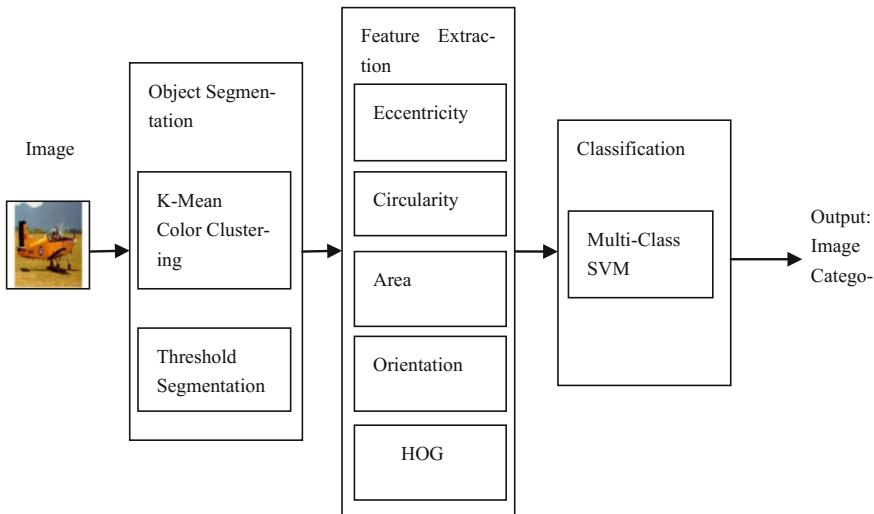


Fig. 1 Proposed framework

2.1 Object Segmentation

Segmentation of an object from the image is very challenging task in image categorization due to complex backgrounds, multiple other small objects closed to each other and different lighting condition. We have applied color-based segmentation using K-means clustering [9] and L*a*b* color space. K-means clustering divides the whole image into three clusters based on different colors. The steps for segmentation are as follows:

Step 1: Convert image to L*a*b* (also known as CIELAB or CIE L*a*b*) color space.

L*a*b* space consists of a lightness layer L*, and color information is contained in two chromaticity layer a* and b* where a* represents the color falls along the green-red axis, and b* represents the color that falls along the yellow-blue axis. Difference between two colors can be measured using the Euclidean distance metric.

Step 2: Classify the Colors in ‘a*b*’ Space Using K-means Clustering

K-means clustering separates each object in an image as having a location in space. It partitions the image into clusters in such a way that object in one cluster is near to each other as possible, and object in other cluster as far as possible. K-means clustering requires two things to specify: number of clusters and a distance measure to quantify how close two objects are to each other in the ‘a*b*’ space. We cluster the objects into three clusters using the Euclidean distance metric.

Step 3: After step 2, we get three images in cluster 1, cluster 2, and cluster 3. The object from image in most cases is contained in cluster 1 and cluster 2. So we add these two clusters to get the object. Sometimes object is in cluster 3.

Step 4: Then we apply a global threshold to segment the object into binary image. We find the area of object in binary image; if this area is less than the threshold (7000 pixels) then object in cluster 3 is converted to binary image using global threshold segmentation. Otherwise, addition of cluster 1 and cluster 2 is taken and threshold is applied to convert image to binary image.

Step 5: Median filter is applied to remove noise in the object.

Step 6: Still this segmented object contains number of small object, a largest blob of the object is taken out for feature extraction.

Figures 2, 3, 4, and 5 show the process of segmentation of object in an image with process description in Fig. 2. In Figs. 2, 3, and 4 segmented image is obtained from cluster1 + cluster2. It is clear from the images that all other objects except ROI are removed. In Fig. 3, other objects are partitioned into cluster 3. Figure 4 shows the importance of color information and portions of the image are extracted in cluster 1 and cluster 2. Figure 5 shows the importance of cluster 3, in this image object is not identified in cluster 1 and cluster 2 so cluster 3 is taken for threshold segmentation.

2.2 Feature Extraction

The goal of feature extraction is to locate the most discriminate information in the images. It plays a vital role in matching and recognition process. We have used global shape descriptors and HOG descriptor for feature extraction. Common simple shape descriptors are area, eccentricity, circularity, and orientation which

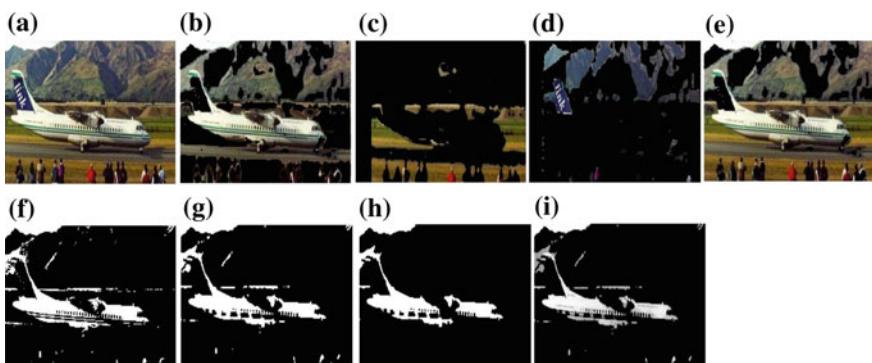


Fig. 2 **a** Original image **b** Image in cluster 1 **c** Image in cluster 2 **d** Image in cluster 3 **e** image in cluster1 + cluster2 **f** Binary segmented image **g** After applying Median Filter **h** Largest blob **i** Masked grey image

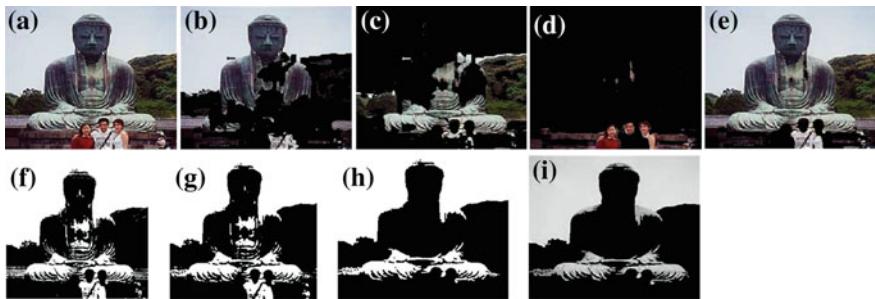


Fig. 3 Other objects in cluster 3

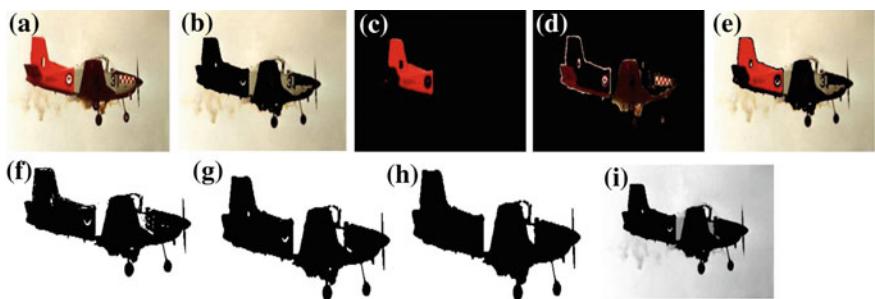


Fig. 4 Color importance (object is of two different colors)

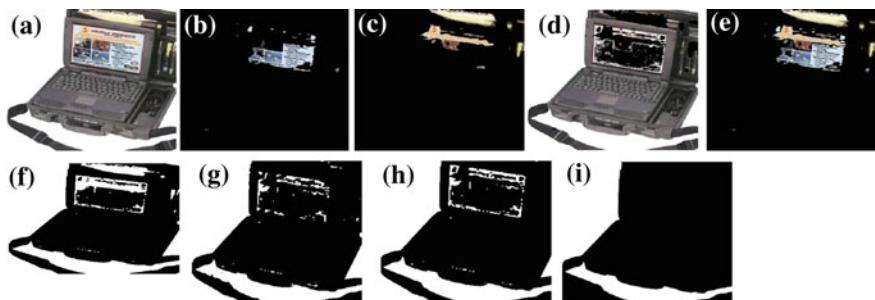


Fig. 5 Importance of cluster 3 as object is identified in cluster 3

decide the overall shape of an object. The limitation of global shape descriptors is that they can discriminate images with only large differences; therefore, they are combined with other shape descriptors [10]. We used these shape descriptors on binary segmented image.

Area.

It is the actual number of pixels in the region.

Circularity [11].

A common shape factor, a function of the perimeter P and Area A which are applicable to all geometric shape and independent of scale and orientation calculated as follows:

$$\text{Circularity} = \frac{4\pi A}{P^2} \quad (1)$$

Eccentricity [11].

It is the ratio of major axis to minor axis.

$$\text{Eccentricity} = \left(\frac{\text{major axis}}{\text{min or axis}} \right) \quad (2)$$

Orientation.

Measures the overall direction of the image.

The second feature that we have extracted is HOG (Histogram of Oriented Gradients) which is generally used for object detection. This descriptor takes gradient orientation into account in localized portion of an image. It converts the image into small regions called cell and then a histogram of each pixel within the cell is computed. Thus, HOG is a concatenation of such histograms [12]. The advantage of HOG is that it is invariant to transformation and can be applied to complex background and cluttered background. The features are built up over a block and are normalized; therefore, it is invariant to illumination effects. In proposed work, we used 9 cells and 9 bin histogram per cell so length of feature vector for HOG is 81. We apply HOG on masked image which is obtained after applying mask operation on original image with segmented image containing only object.

3 Classification

In this phase, image categorization is performed. First, the systems are prepared for training the dataset through support vector machine by extracting the features from the object and store them as feature vector in a codebook. The global shape descriptors help us by calculating the different values of the boundary or shape of an image while HOG helps us by recognizing the exact sign and its meaning when the image have some change in scale, position, or illumination. SVM gained popularity because of its existing feature such as better empirical performance. SVM is a classification and regression technique that uses machine learning theory to maximize the accuracy of prediction [13]. In the proposed work, MSVM is used to

suitably classify the image among multiple classes. SVM is binary classifier which has been converted to multi-class SVM using one-against-one approach which is also known as pair-wise coupling. In this approach, one SVM is used for each pair of classes. For the problem of n classes $n*(n-1)/2$ SVMs are trained to differentiate sample of one class with sample of other classes. The classification of an image is done according to maximum voting for the class by each SVM [14].

4 Experimental Results and Analysis

We have evaluated the performance of our proposed system on Caltech 101 [15]. This data set contains 8677 images of 101 objects. This dataset has wide variation in number of images per object ranging between 40 to 800 images and size of each image is about 300 * 200 pixels. We used 10-fold cross validation to check the recognition rate of proposed algorithm. Then accuracy and confusion matrix is derived. Each time one different fold is used for evaluation. Accuracy is measured in terms of number of images correctly classified. The final result is average of all 101 object category class accuracy. We reported 76.14 % recognition accuracy. Table 1 shows the recognition accuracy of first 10 objects of Caltech 101 and Table 2 shows the confusion matrix for the same. Data in diagonal of the matrix represents the number of images that are correctly classified by the system.

From the experimental result, we observe the categories which have large number of images reported good recognition accuracy. For example, in dataset, airplanes object has 800 images, classification accuracy is 90.25 which is greater compared to other objects. We have also tested our method against other methods that exist in the literature shown in Table 3 and found that our method outperforms these methods.

Table 1 Recognition accuracy

Object	# images	#Correctly classified	Accuracy (%)
Airplanes	800	722	90.25
Anchor	42	33	78.57
Ant	42	30	71.42
Barrel	47	35	74.46
Bass	54	39	72.22
Beaver	46	34	73.91
Brain	98	80	81.63
Buddha	85	65	76.47
Butterfly	91	69	75.82
Camera	50	38	76

Table 2 Confusion matrix

Actual/predicted class	Airplanes	Anchor	Ant	Barrel	Bass	Beaver	Brain	Buddha	Butterfly	Camera
Airplanes	722	2	0	2	4	0	0	0	0	0
Anchor	0	33	0	2	2	0	2	0	7	0
Ant	0	0	30	0	0	3	0	0	9	0
Barrel	32	0	8	35	0	0	0	5	0	7
Bass	35	0	0	0	39	0	7	0	6	1
Beaver	0	0	0	0	3	34	0	1	0	0
Brain	6	5	0	0	6	4	80	8	0	0
Buddha	5	0	0	4	0	5	6	65	0	0
Butterfly	0	2	4	2	0	0	2	6	69	4
Camera	0	0	0	2	0	0	1	0	0	38

Table 3 Comparison with other methods

Method	Recognition accuracy (%)
Zhang et al. [8]	66.2
Yang et al. [16]	73.2
Hu and Guo [17]	65.5
Bilen et al. [18]	75.31
Our method	76.14

5 Conclusions and Future Work

In this paper, we have proposed a method for image categorization using segmentation by color-based K-means clustering and threshold method. We have combined two descriptors shape descriptor and HOG descriptor for feature extraction so that we can extract strong features from the image. Image classification is done using MSVM. Proposed method is invariant to affine transformation and cluttered background and computationally efficient. When two objects are similar and one object is contained in another object, our method gets confused and gives the wrong result. Our future work will focus on these problems and will develop a method which is suitable for large-scale dataset.

References

1. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: Computer Vision and Image Understanding, vol. 106, no. 1, pp. 59–70 (2007).
2. Perronnin, F., Sánchez, J., Liu, Y.: Large-scale image categorization with explicit data embedding. In: proceedings of CVPR (2010).
3. Csurka, G., Dance, C., Bray, C., Fan, L.: Visual categorization with bags of keypoints. In: Proceedings of Workshop on Statistical Learning in Computer Vision (2004).
4. Grauman, K., Darrell, T.: Pyramid match kernels: Discriminative classification with sets of image features. In: Proceedings of ICCV (2005).
5. Favorskaya, M., Proskurin, A.: Image categorization using color G-SURF invariant to light intensity. In: proceedings of 19th International Conference on Knowledge Based and Intelligent Information and Engineering, Elsevier Systems, pp. 681–690 (2015).
6. Yao, B., Khosla, A., Li, F.F.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011).
7. Hoai, M.: Regularized Max Pooling for Image Categorization. In: Proceedings of the British Machine Vision Conference, BMVA Press (2014).
8. Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proceedings of CVPR (2006).
9. Chitade, A., Katiyar, S.K.: Color based image segmentation using K-means clustering. In: International Journal of Engineering Science and Technology, vol. 2, no. 10, pp. 5319–5325 (2010).
10. Zhang, D., Lu, G.: Review of shape representation and description techniques. In: Journal of Pattern Recognition, Elsevier, vol. 37, pp. 1 – 19 (2004).

11. Agrawal, S.C., Jalal A.S., Bhatnagar, C.: Redundancy Removal for Isolated Gesture in Indian Sign Language and Recognition using Multi-Class Support Vector Machine: In: International Journal of Computational Vision and Robotics, InderScience, Vol. 4, No. 1/2, pp. 23 – 38 (2014).
12. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886–893 (2005).
13. Byun, H., Lee, S.W.: Applications of Support Vector Machines for Pattern Recognition: A Survey. In: Proceedings of first International Workshop on Pattern Recognition with Support Vector Machines, Springer, pp. 213–236 (2002).
14. Milgram, J., Cheriet, M., Sabourin, R.: “One Against One” or “One Against All”: Which One is Better for Handwriting Recognition with SVMs?. In: Tenth International Workshop on Frontiers in Handwriting Recognition (2006).
15. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: IEEE. CVPR 2004, Workshop on Generative-Model Based Vision (2004).
16. Yang, K.Y., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of CVPR (2009).
17. Hu, J., Guo, P.: Combined Descriptors in Spatial Pyramid Domain for Image Classification. In: Computer Vision and Pattern Recognition (2012).
18. Bilen, H., Namboodiri, V.P., Gool, L.J.V.: Object and Action Classification with Latent Window Parameters. In: International Journal of Computer Vision, vol. 106, no. 3 (2014).

Local Binary Pattern and Its Variants for Target Recognition in Infrared Imagery

Aparna Akula, Ripul Ghosh, Satish Kumar and H.K. Sardana

Abstract In this research work, local binary pattern (LBP)-based automatic target recognition system is proposed for classification of various categories of moving civilian targets using their infrared image signatures. Target recognition in infrared images is demanding owing to large variations in target signature and limited target to background contrast. This demands robust features/descriptors which can represent possible variations of the target category with minimal intra class variance. LBP, a simple yet efficient texture operator initially proposed for texture recognition of late is gaining popularity in face and object recognition applications. In this work, the suitability of LBP and two of its variants, local ternary pattern (LTP), complete local binary pattern (CLBP) for the task of recognition in infrared images has been evaluated. The performance of the method is validated with target clips obtained from ‘CSIR-CSIO moving object thermal infrared imagery dataset’. The number of classes is four- three different target classes (Ambassador, Auto and Pedestrian) and one class representing the background. Classification accuracies of 89.48 %, 100 % and 100 % were obtained for LBP, LTP and CLBP, respectively. The results indicate the suitability of LBP operator for target recognition in infrared images.

Keywords Local binary pattern • Local ternary pattern • Complete local binary pattern • Infrared • Recognition • Support vector machine

A. Akula (✉) · R. Ghosh · S. Kumar · H.K. Sardana
CSIR-Central Scientific Instruments Organisation (CSIR-CSIO),
Chandigarh 160030, India
e-mail: aparna.akula@csio.res.in

A. Akula · R. Ghosh · S. Kumar · H.K. Sardana
Academy of Scientific and Innovative Research (AcSIR), Chennai, India

1 Introduction

Human beings are capable of detecting and recognizing objects surrounding them through their innate abilities which improve with time by experience and learned competence. Research is being carried out to develop systems which can imitate these human activities for various applications, such as automatic target recognition (ATR) [1, 2], surveillance [3–5], rescue operations [6] and intelligent transportation systems [7, 8]. ATR systems originally meant for defence applications, are gaining relevance and popularity in civilian sector recently for infrastructure security in industries, such as nuclear, energy, oil and gas. In the civilian context, monitoring unauthorized intruding persons and vehicles for security has been one of the most critical issues for society, governments and industries [9, 10].

The objective of ATR algorithm is to detect and classify each target image into one of the number of classes. The first stage detection involves detecting the moving target in the image and removing the background clutter. The next stage of recognition involves computing features which is passed to a classifier for classification [1]. The main focus of this work is the task of recognition.

Infrared imaging owing to its ability to operate even in extreme darkness and harsh weather conditions is one of the preferred sensing technologies for target recognition. However, target recognition in infrared images is difficult because of the large variations in the thermal signatures of targets. As a matter of fact, the thermal heat signature of both target and background clutter vary extensively due to environmental variations [11]. Infrared images fail to provide information regarding edges and boundaries of targets, making it difficult to extract robust features for recognition [12]. This calls for exploration of features that are robust to these variations, which is the core emphasis of this work.

2 Literature Survey

Over a span of 20 years, researchers have proposed several feature extraction approaches. Feature extraction can be looked upon as a two-step process involving feature detection and feature description. Feature detectors can be based on key-points or uniform dense sampling of points across image [13]. Feature descriptors describe the detected point using the local neighbourhood information around it. Some popular feature descriptors used for recognition are SIFT-Scale invariant feature transform [14], HOG-Histogram of oriented gradients [15], SURF-Speeded up robust features [16]. In the past decade, texture operators used for texture analysis and classification were extended to face and object recognition applications and have shown promising results. Ojala et al. for the first time has presented Local binary pattern (LBP) for texture recognition which has become popular as it is very discriminative and computationally efficient [17]. A decade later, the suitability of LBP for face recognition was demonstrated in [18]. Owing to its various

advantages, researchers have proposed variants of LBP, Hamming LBP [19], Extended LBP [20], Completed LBP [21], Local ternary patterns [22], Multi block LBP [23, 24], etc. Also, LBP applicability to various fields, such as image matching, object recognition and scene classification has been explored. In the recent years, researchers have started investigating the use of LBP for military target recognition in infrared images [25, 26]. This work focuses on demonstrating the suitability of LBP and two of its variants—LTP, CLBP for civilian target recognition in infrared images. The reason for choosing these variants is their highly discriminative nature while retaining the computational effectiveness of the basic LBP making them suitable for ATR systems which demand real-time operation capabilities.

This paper is organized as follows. First, we provide a brief introduction of LBP, LTP and CLBP. Then, we describe the infrared ATR database used. Then, we present the in depth recognition methodology, followed by the experimental results and analysis and finally some concluding remarks.

3 Theoretical Background

A brief about the basic LBP and two of its extensions, LTP and CLBP is presented in this section.

3.1 LBP

Local binary pattern (LBP) operator was presented by Ojala et al. for the purpose of texture recognition [27]. It works on the basis of one major assumption that texture consists of two complementary aspects, i.e. textural pattern and its strength. The LBP pattern for each pixel is generated by thresholding the pixel intensity value with its neighbourhood thereby resulting in a binary pattern. The original LBP was proposed for a 3×3 neighbourhood. Later, in 2002 Ojala et al. [17] presented a more generic revised form of LBP which is currently used as LBP. Two main advantages of LBP operator is its invariance against monotonic grey-level variations and ease of computation. LBP label is calculated by weighted multiplication of the obtained LBP binary pattern, and summing up the result as shown in Eq. (1). The thresholded values are computed by comparing its intensity value with its neighbours as shown in Eq. (2).

$$LBP_{P,R} = \sum_{i=1}^P 2^{(i-1)} \times f_t(intensity_i - intensity_c) \quad (1)$$

$$f_t(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{else} \end{cases}, \quad (2)$$

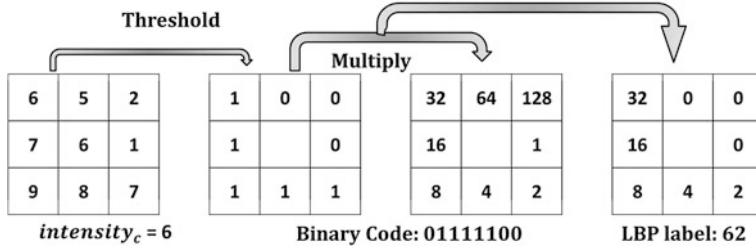


Fig. 1 LBP calculation for a sample window of 3×3

where $intensity_c$ denotes the pixel intensity value of the center pixel, $intensity_i$ is the pixel intensity value of its neighbours, P is the number of neighbours and R represents the radius of the neighbourhood. The LBP procedure is shown in Fig. 1.

The number of different LBP labels possible for a neighbourhood of P pixels is given as 2^P . For a 3×3 neighbourhood, the number of neighbour pixels is 8 with $2^8 = 256$ possible number of different labels. Another popular extension to the initial proposed LBP operator is the *Uniform Patterns* owing to its two main characteristics, reduced size of feature vector and rotation invariant nature of the descriptor [17]. If a LBP contains less than or equal to two bitwise changeovers from 0 to 1 or vice versa when the bit pattern is navigated circularly then it is known as uniform pattern. The binary patterns 11111111, 1000111 with 0 and 2 transitions, respectively, are uniform whereas the patterns 00101100 and 01011011 with 4 and 5 transitions, respectively, are not uniform patterns. It exploits the point that the occurrence of some binary patterns is more frequent than others in texture images. Uniform pattern detects only important local texture, such as spot, edge, flat, corner, etc. While calculating uniform pattern LBP, a distinct label is used for each of the uniform patterns and a single label is used to represent rest all non-uniform patterns, thereby reducing the size of feature vector.

3.2 LTP

The original LBP was extended to a version named local ternary patterns (LTP) by Tan and Triggs [28]. LTP is more robust to noise specifically in near-homogenous image regions. LBP is sensitive to noise because of thresholding exactly at the central pixel intensity. This issue is addressed in LTP by a 3-value encoding scheme, in which intensity values in a zone of width $\pm t$ around $intensity_c$ are quantized to zero, intensity values above this are quantized to +1 and lower than this to -1 as shown in Eqs. (3) and (4).

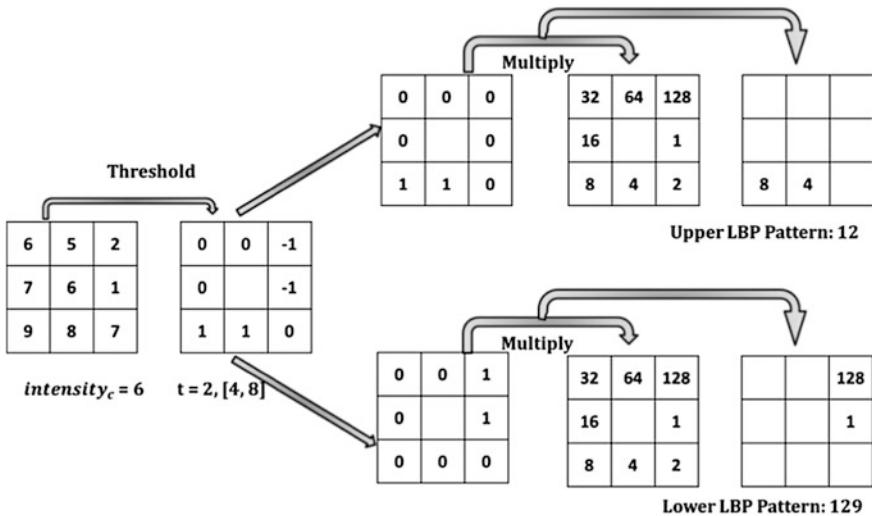


Fig. 2 LTP calculation for a sample window of 3×3 . In the LTP, the ternary pattern is further coded into upper LBP and lower LBP. Upper pattern is coded by holding 1 and substituting 0 and -1 with 0. Lower pattern is coded by substituting 1 and 0 with 0 and -1 with 1

$$LTP_{P,R} = \sum_{i=1}^P 2^{(i-1)} \times f_t(intensity_i - intensity_c) \quad (3)$$

$$f_t(x) = \begin{cases} 1 & x \geq intensity_c + t \\ -1 & x \leq intensity_c - t \\ 0 & |x - intensity_c| < t \end{cases}, \quad (4)$$

where t is the user-specified threshold value. This makes LTP more resistant to noise compromising to an extent its invariance to grey-level transformations. They also proposed a simplified coding scheme whereby each ternary pattern is fragmented into two parts: positive and negative, as shown in Fig. 2.

3.3 CLBP

A generalized and complete LBP, termed as completed LBP (CLBP) was proposed by Guo et al. [29]. Intensity of center pixel and a local difference sign-magnitude transforms (LDSMT) is used to characterize a local region in CLBP. Global thresholding is used to binary code the center pixel which is termed as CLBP_Center (CLBP_C). The image local structure is decomposed into two complementary components: the difference signs termed as CLBP_Sign (CLBP_S) and the difference magnitudes termed as CLBP_Magnitude (CLBP_M). LBP can be treated as a distinctive case of CLBP which uses only CLBP_S. Through

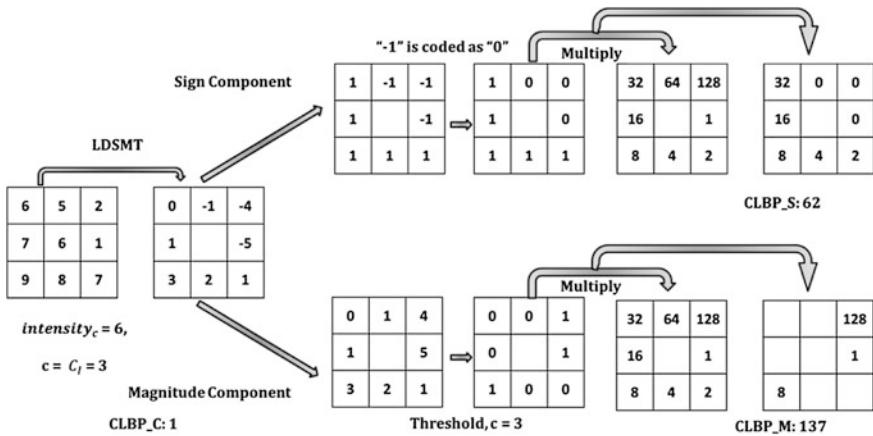


Fig. 3 CLBP calculation for a sample 3×3 window

theoretical and experimental evidence it was observed that local structural information is more preserved in the sign component when compared to the magnitude component. In line with the coding strategy of CLBP_S, CLBP_M is defined as in Eq. (5). The CLBP encoding process is illustrated in Fig. 3.

$$CLBP_M_{P,R} = \sum_{i=1}^P 2^{(i-1)} \times f_t(intensity_i - intensity_c) \quad (5)$$

$$f_t(x) = \begin{cases} 1, & x \geq c \\ 0, & x < c \end{cases}, \quad (6)$$

where c is value of threshold which is computed adaptively. It is generally chosen as global mean intensity value of the image.

$$CLBP_C_{P,R} = f_t(intensity_c, C_I) \quad (7)$$

where f_t is defined as in Eq. (6) and threshold C_I is set as the average intensity of the complete image.

4 Experimental Data

In this work, LBP, LTP and CLBP features are evaluated in the context of infrared target recognition. There are a total of four classes—three different target classes (Ambassador, Auto, Pedestrian) and one class representing background. A total of 376 infrared target and background clips are generated from the ‘CSIR-CSIO Moving Object Thermal Dataset’ consisting of total of 18 thermal video sequences captured using a microbolometer type thermal image [12]. Complexity of IR images



Fig. 4 Illustrative collection of infrared clips of targets (*Top—Bottom* Ambassador, Auto, Pedestrian, Background)

is captured by generating target clips having variations in environmental conditions. We used 80 % of dataset for training, i.e. 300 images (89 Ambassador, 89 Auto, 61 Pedestrians and 61 Background) and 20 % for testing, i.e. 76 images (19 images of each category). Even though the number of classes is only four, the dataset is challenging as each of the classes have large intra-class variations in terms of variations in ambient temperature, target temperature, scale and pose. Representative set of infrared clips of the targets under diverse conditions is shown in Fig. 4.

5 Methodology

In this work, LBP and its derivative models are experimented for the IR target recognition problem. Target recognition consists of two steps, feature extraction/representation and classifier design. LBP is used for feature representation and for classifier SVM (Support Vector Machine) classifier [30] was chosen, as SVM has demonstrated success in various computer vision tasks. In brief, the overall process involves preprocessing the images to a fixed size of 200×200 , extracting the uniform LBP/LTP/CLBP features, generating the feature label

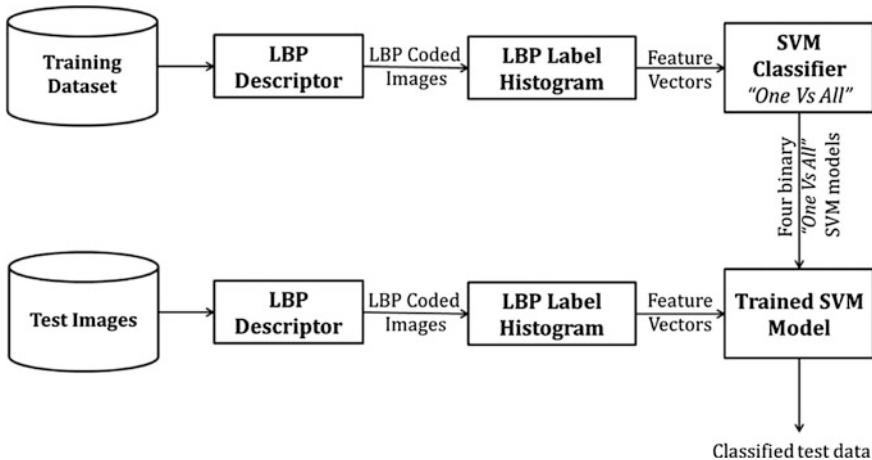


Fig. 5 Graphical representation of LBP-based target recognition system

histogram and training the SVM classifier with these histograms. SVMs are inherently two-class classifiers. We have extended them to handle the multiclass problem by implementing the ‘One-versus-All’ approach [31]. The ‘one against all’ strategy consists of creating one SVM classifier for each class, which is trained to differentiate the samples of one class from the samples of all other classes. This trained SVM model is used to classify the test data. Figure 5 provides a graphical representation of the process.

6 Results and Analysis

The performance of the proposed LBP-based infrared target recognition system is evaluated on IR clips obtained from experimentally generated infrared image sequences. The first stage of ATR, detection is performed using the spatiotemporal information-based method proposed by Akula et al. [12]. The output of this method is used to automatically obtain the IR target and background clips. A collection of randomly chosen 300 clips across three target classes and background is chosen for training and testing of the proposed recognition system. These clips capture variations of target signature in terms of temperature, scale and environmental conditions.

The proposed method has been validated by using confusion matrix and average classification accuracy as performance evaluation metrics. Average classification accuracy is the average of the individual accuracy obtained for each class. To evaluate the efficiency of LBP for target recognition in IR, we have used LBP with varying pixel neighbourhood of 3×3 , 5×5 and 7×7 . The same have been considered for LTP and CLBP. For all the cases uniform pattern representation was

Table 1 Average classification accuracy of LBP and its variants for different radius of pixel neighbourhood

Sr. No	Name of texture operator	Neighbourhood of texture operator		
		3×3	5×5	7×7
1	LBP	89.48 %	89.48 %	88.16 %
2	LTP	100 %	96.05 %	100 %
3	CLBP	98.68 %	96.05 %	100 %

Table 2 Confusion Matrix of LBP operator with pixel neighbourhood of 5×5

Predicted class	Actual class			
	Ambassador	Auto	Pedestrian	Background
Ambassador	18	0	2	0
Auto	0	19	0	5
Pedestrian	0	0	17	0
Background	1	0	0	14

used to retain rotational invariance. Confusion Matrix of LBP operator with pixel neighbourhood of 5×5 is shown in Table 2. The high number of false classification of the Background class is due to low resolution and lack of texture information. Average classification accuracy of LBP and its variants for different radius of pixel neighbourhood is detailed in Table 1. Our preliminary experiments show that LTP and CLBP show a better classification performance in comparison to the basic LBP. Also, LTP shows better performance in comparison to CLBP. The addition of magnitude along with sign component is not very much beneficial due to the temperature variations of targets in IR leading to variable intensity values of the targets. Considering the trade-off of accuracy versus complexity, LTP seems to be more suitable for IR target recognition.

7 Conclusion

Automatic target recognition of civilian vehicles and pedestrians in infrared images is gaining popularity owing to its numerous possible applications, such as perimeter monitoring of critical infrastructures, rescue operations, autonomous driver assistance systems. LBP descriptor has become popular among texture models especially for texture and face recognition applications. The primary aim of this work is to explore the appropriateness of LBP and its variants for target recognition in infrared images under large inter-class variations caused by variations in ambient temperature, target temperature, distance of target from sensor and target pose. It was observed that all three descriptor models, LBP, LTP and CLBP achieved considerably decent recognition rates of more than 87 %. Further, the recognition

rates were also improved by optimally choosing the descriptor radius. Our preliminary experiments show that LTP and CLBP show a better classification performance in comparison to the basic LBP. Considering the accuracy-time trade-off, from the preliminary results LTP appears to be more suitable for civilian target recognition in infrared images. In conclusion, this paper presents LBP-based recognition framework and initial results for civilian target recognition in infrared images. In future, the robustness of the system has to be validated on a larger database containing additional relevant target categories, variations in targets and occlusions.

Acknowledgements The work is supported in part by funds of Council of Scientific and Industrial Research (CSIR), India under the project OMEGA PSC0202-2.3.1. The authors acknowledge the contribution of M.Tech trainee CSIR-CSIO, Ms. T. Pathak and Mr. A. Singh for their contribution towards partial implementation of code.

References

1. V. M. Patel, *et al.*, “Sparsity-motivated automatic target recognition,” *Applied optics*, vol. 50, pp. 1425–1433, 2011.
2. B. Bhanu, “Automatic target recognition: State of the art survey,” *Aerospace and Electronic Systems, IEEE Transactions on*, pp. 364–379, 1986.
3. A. Arora, *et al.*, “A line in the sand: a wireless sensor network for target detection, classification, and tracking,” *Computer Networks*, vol. 46, pp. 605–634, 2004.
4. P. Natarajan, *et al.*, “Multi-Camera Coordination and Control in Surveillance Systems: A Survey,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, p. 57, 2015.
5. R. Ghosh, *et al.*, “Time-frequency analysis based robust vehicle detection using seismic sensor,” *Journal of Sound and Vibration*, vol. 346, pp. 424–434, 2015.
6. R. Xin and R. Lei, “Search aid system based on machine vision and its visual attention model for rescue target detection,” in *Intelligent Systems (GCIS), 2010 Second WRI Global Congress on*, 2010, pp. 149–152.
7. C. N. E. Anagnostopoulos, *et al.*, “A license plate-recognition algorithm for intelligent transportation system applications,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 7, pp. 377–392, 2006.
8. Y. Dong, *et al.*, “Driver inattention monitoring system for intelligent vehicles: A review,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, pp. 596–614, 2011.
9. A. Cavoukian, “Surveillance, then and now: Securing privacy in public spaces,” *Office of the Information and Privacy Commissioner of Ontario*, 2013.
10. H. C. Choe, *et al.*, “Wavelet-based ground vehicle recognition using acoustic signals,” in *Aerospace/Defense Sensing and Controls*, 1996, pp. 434–445.
11. U. Braga-Neto, *et al.*, “Automatic target detection and tracking in forward-looking infrared image sequences using morphological connected operators,” *Journal of Electronic Imaging*, vol. 13, pp. 802–813, 2004.
12. A. Akula, *et al.*, “Moving target detection in thermal infrared imagery using spatiotemporal information,” *JOSA A*, vol. 30, pp. 1492–1501, 2013.
13. T. Tuytelaars and K. Mikolajczyk, “Local invariant feature detectors: a survey,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, pp. 177–280, 2008.

14. D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, 1999, pp. 1150–1157.
15. N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, pp. 886–893.
16. H. Bay, *et al.*, "Surf: Speeded up robust features," in *Computer vision–ECCV 2006*, ed: Springer, 2006, pp. 404–417.
17. T. Ojala, *et al.*, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 971–987, 2002.
18. T. Ahonen, *et al.*, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 2037–2041, 2006.
19. D. Huang, *et al.*, "Local binary patterns and its application to facial image analysis: a survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, pp. 765–781, 2011.
20. M. Pietikäinen, *et al.*, "Local binary patterns for still images," in *Computer Vision Using Local Binary Patterns*, ed: Springer, 2011, pp. 13–47.
21. Z. Guo, *et al.*, "A completed modeling of local binary pattern operator for texture classification," *Image Processing, IEEE Transactions on*, vol. 19, pp. 1657–1663, 2010.
22. X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *Image Processing, IEEE Transactions on*, vol. 19, pp. 1635–1650, 2010.
23. D. Xia, *et al.*, "Real-time infrared pedestrian detection based on multi-block LBP," in *Computer Application and System Modeling (ICCASM), 2010 International Conference on*, 2010, pp. V12–139-V12-142.
24. S. Liao, *et al.*, "Learning Multi-scale Block Local Binary Patterns for Face Recognition," in *Advances in Biometrics*, vol. 4642, S.-W. Lee and S. Li, Eds., ed: Springer Berlin Heidelberg, 2007, pp. 828–837.
25. J. Sun, *et al.*, "Concave-convex local binary features for automatic target recognition in infrared imagery," *EURASIP Journal on Image and Video Processing*, vol. 2014, pp. 1–13, 2014.
26. X. Wu, *et al.*, "Improved Local Ternary Patterns for Automatic Target Recognition in Infrared Imagery," *Sensors*, vol. 15, pp. 6399–6418, 2015.
27. T. Ojala and M. Pietikäinen, "Unsupervised texture segmentation using feature distributions," *Pattern Recognition*, vol. 32, pp. 477–486, 1999.
28. X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," in *Analysis and Modeling of Faces and Gestures*, ed: Springer, 2007, pp. 168–182.
29. Z. Guo and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *Image Processing, IEEE Transactions on*, vol. 19, pp. 1657–1663, 2010.
30. V. N. Vapnik, "An overview of statistical learning theory," *Neural Networks, IEEE Transactions on*, vol. 10, pp. 988–999, 1999.
31. F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, pp. 1778–1790, 2004.

Applicability of Self-Organizing Maps in Content-Based Image Classification

Kumar Rohit, R.K. Sai Subrahmanyam Gorthi and Deepak Mishra

Abstract Image databases are getting larger and diverse with the coming up of new imaging devices and advancements in technology. Content-based image classification (CBIC) is a method to classify images from large databases into different categories, on the basis of image content. An efficient image representation is an important component of a CBIC system. In this paper, we demonstrate that Self-Organizing Maps (SOM)-based clustering can be used to form an efficient representation of an image for a CBIC system. The proposed method first extracts Scale-Invariant Feature Transform (SIFT) features from images. Then it uses SOM for clustering of descriptors and forming a Bag of Features (BOF) or Vector of Locally Aggregated Descriptors (VLAD) representation of image. The performance of proposed method has been compared with systems using k-means clustering for forming VLAD or BOF representations of an image. The classification performance of proposed method is found to be better in terms of F-measure (FM) value and execution time.

Keywords Content-Based Image Classification (CBIC) • Self-Organizing Maps (SOM) • Scale-Invariant Feature Transform (SIFT) • Bag of Features (BOF) • Vector of Locally Aggregated Descriptors (VLAD)

K. Rohit (✉)

National Institute of Technology, Hamirpur 177005, Himachal Pradesh, India
e-mail: rohitkhel@gmail.com

R.K. Sai Subrahmanyam Gorthi · D. Mishra
Indian Institute of Space Science and Technology, Thiruvananthapuram 695547,
Kerala, India

1 Introduction

Images are an important medium of communication these days, second only to text. This has led to increasing size of image databases.

Unlike the past, where text-based approach using metadata was used to index and retrieve the images in a database, content-based image retrieval aims at indexing and retrieval of images based on actual content of images rather than the metadata associated with it. CBIC, a sub-problem of Content-based image retrieval (CBIR) [1] problem, which assigns images to different categories automatically on the basis of content, is addressed in this paper. The system takes a query image as input from the user and returns category of image. The generally used image contents are colour, shape and texture information. Due to its effectiveness in current context, this area of research has attracted lot of attention in recent years.

Efficient storing, indexing and searching are not possible if whole images are stored and compared for search [2, 3]. That will be a waste of time and memory space. An efficient representation is needed so that an image can be represented by some of its unique features which can be later used for various purposes, viz. classify images on the basis of their common features and search similar images from database if a query image is given.

CBIC relies on the processing of set of features extracted from an image which can represent whole image. The key frame features include texture features, shape features and the most common colour features. Only the shape-based features for image classification are used in this work. The proposed CBIC system gives the category to which a query image belongs. This can be used for automatic categorization of image in a large database and searching for similar images. Use of Scale-Invariant Feature Transform (SIFT) features [4] make the system invariant to a substantial affine distortion, change in 3D viewpoint, addition of noise and change in illumination in images.

The proposed method first extracts SIFT features from images. Then it uses SOM for clustering of descriptors and forming a BOF [5–8] or VLAD representation of image. K-means is used in conventional CBIC and we propose and demonstrate the use of SOM in place of k-means for forming visual words. This greatly reduces the time complexity and improves the performance. To classify the images from the database, the similarity between the feature vector of the query image and each sequence of feature vectors in the training database is computed. The vector comparison is performed using simple Nearest Neighbour (NN) classifier using Euclidean distance metric.

Structure of content-based image classification system:

- Feature extraction—Features are extracted from training images using key point detectors and described as a high dimensional vector using descriptors.
- Vocabulary creation—Formation of visual words by clustering descriptors of detected features.

- Representation—All training images are represented as high dimensional vector using the visual words formed in second step.
- Query image—Query image given by user is also represented as a high dimensional vector using the visual words formed in second step.
- Categorization—Vector representation of query image is compared with vector representation of training images using some similarity measure and result is returned.

The proposed method has been tested on Caltech101 dataset [9] available online. The performance of the proposed system has been compared with k-means clustering-based BOF and VLAD systems in terms of precision, recall, F-Measure and execution time values.

Rest of the paper is organized as follows: Sect. 2 gives the basic concepts and related literature for CBIC, Sect. 3 explains the details of approach taken in this work, Sect. 4 discusses and compares the results obtained with previous and current approaches and Sect. 5 gives the conclusion and scope of future work.

2 Basic Methodologies and Related Techniques

- Bag of Features (BOF)

BOF approach, initially proposed in [5], is analogous to Bag of Words approach for indexing and retrieval of text documents. BOF representation groups the local descriptors. Key points are detected using SIFT detector and described using SIFT descriptor. A codebook of k “visual words” is obtained by clustering the key point descriptors obtained, using k-means clustering. Each local descriptor of dimension d of each key point of the image is assigned to the closest visual word on the basis of Euclidean distance. An image is represented as histogram of assignment of each local descriptor to a visual word. This gives a k -dimensional vector representation of image. The BOF representation obtained is normalized using Euclidean normalization. Many improvements to BOF have been proposed with emphasis on efficiency and compactness of the representation [10, 11], weighting scheme for visual words [12, 13] and improving recall [14, 15]. Only basic BOF approach has been used in this paper to limit the complexity.

- Fisher Vector (FV)

FV [16] is another widely used representation which has more compact representation than BOF. It is obtained by calculating gradient of the sample’s likelihood with respect to the parameters of this distribution. It is then scaled by the inverse square root of the Fisher information matrix. It gives the direction in

parameter space into which the learnt distribution should be modified to better fit the observed data. FV gives high dimensional vectors for small vocabulary sizes. In comparison with BOF, fewer visual words are required by this more sophisticated representation. Dimension of vector is $(2D + 1)*(N - 1)$ as gradient with respect to mean of word and variation around the mean are calculated for each dimension of a visual word. We consider three parameters of the distribution, viz. relative frequency of word, mean of word and variation around the mean where ‘D’ is the dimension of feature vector and ‘N’ is the number of visual words. Linear classifiers provide excellent results with fisher kernels.

- Vector of Locally Aggregated Descriptors (VLAD)

VLAD [17, 18] is a simplification of Fisher Vector. Similar to BOF, key points are detected and described using SIFT detector and SIFT descriptor. Key points are clustered using k-means to learn a codebook $C = \{c_1, \dots, c_k\}$ of k visual words. Each local descriptor of key point x , belonging to set D, is assigned to its nearest visual word $c_i = \text{NN}(x)$. VLAD descriptor differs from BOF descriptor by storing the differences $x - c_i$ of the vectors x assigned to each visual word c_i in a vector rather than the number of descriptors assigned to each cluster. This gives the distribution of the vectors with respect to the centre.

$$V_i = \sum_{X \in D | \text{NN}(X) = C_i} X - C_i$$

where

C_i Value of i^{th} visual word

V_i VLAD descriptor of i^{th} visual word

D Set of local descriptors of key points

NN Nearest neighbour function which gives the visual word nearest to the input key point descriptor

The dimension D of VLAD representation is $D = k \times d$, if the local descriptor is d -dimensional. Experimental results show that excellent results can be obtained even with a relatively small vocabulary size. VLAD carries only in-plane rotational invariance from original SIFT descriptor and is partially tolerant to image scaling and clipping. Improvements such as Intra-normalization to address the problem of burstiness and Multi-VLAD to improve the retrieval performance for small objects have been proposed [19]. Basic VLAD proposed in [17] has been used in this paper to limit the complexity.

- Self-organizing maps (SOM)

SOM [20, 21] are a way to automatically cluster the points in a dataset. It is an unsupervised competitive learning algorithm of the artificial neural networks. It is related to classical vector quantization. The development of SOM was

motivated by the way different sensory inputs such as motor, visual and acoustic inputs are mapped onto corresponding areas of the cerebral cortex in a topographically ordered computational map. SOM-based clustering is used in this paper in place of more time-consuming k-means clustering. This is based on the intuition that SOM performs faster and topologically meaning full clusters than K-means through its organized learning process. This is in fact demonstrated based on the simulated experiments carried out in this paper. While the execution time of k-means clustering algorithm, previously used in CBIC algorithms, depends directly on the size of visual vocabulary, SOM's execution time will not rise with data as the convergence can be archived with reasonably sufficient samples.

A SOM is made of a network of neurons that is usually one or two dimensions. At the beginning, the neurons are initiated with random weights. Input samples from the input dataset are randomly selected to tune the weight vectors of neurons in such a way that the most stimulated (winning) neuron and its neighbours are being tuned more so that they become more like the stimulating input sample. At the same time, the tuning effects for those neurons that neighbour the winning neuron gradually decrease inversely to their distance from the winning neuron. Gradually, this network of neurons progresses from an initially unordered map to a stable topologically ordered map of clusters of neurons. Each cluster contains neurons that belong to a particular class of the input dataset. Some of the improvements proposed in the literature for SOM are Distance metric learning method [22], spatial indexing method such as R-Tree in order to speed up the search of the winning neuron to reduce the cost [23] and batch version of the Kohonen algorithm to reduce execution time in some cases [24]. Only SOM with Kohonen learning is used in this paper.

Steps in SOM algorithm are:

- (1) Initialization—Weight vectors of all neurons on the map are initialized randomly or randomly picked from inputs. Weight of the neuron j in the map can be expressed as

$$w_j(0) = \{w_{ji}: j = 1, \dots, n; i = 1, \dots, d\}$$

where n is the total number of neurons and d is the number of dimensions of the input vector.

$w_j(0)$ is the value of weight vector of neurons at iteration 0.

- (2) Sampling—An input sample is selected from dataset randomly and applied to the lattice. It can be represented as

$$x(k) = \{x_i^k: i = 1, \dots, d\}$$

- (3) Similarity Matching—The closest (winning) neuron $c(t)$ is found at iteration t using

$$c(t) = \arg \min_j \|w_j(t) - x(k)\|, j = 1, \dots, n$$

where $\|w_j(t) - x(k)\|$ is the Euclidean distance between randomly selected input sample and the neuron $w_j(t)$ on the map.

- (4) Updating—Weight vectors of winning neuron and its neighbouring neurons are updated according to

$$w_j(t+1) = w_j(t) + \eta(t) h_{j,c(t)}(t) (x(k) - w_j(t))$$

where $h_{j,c(t)}(t)$ is the neighbourhood function between neuron j and the winner neuron $c(t)$ at t^{th} iteration and is calculated as

$$h_{j,c(t)}(n) = \exp\left(-\frac{d_{j,c(t)}^2}{2\sigma(t)^2}\right)$$

where $\sigma(t)$ is the time-varying standard deviation that affects the effective width of the topological neighbourhood at iteration t and is calculated as

$$\sigma(t) = \sigma_0 \exp\left(-\frac{\eta(t)}{\tau_1}\right), t = 0, 1, 2, \dots, N \text{ and } \tau_1 = \frac{N}{\log \sigma_0}$$

where

σ_0 is the initial effective width

τ_1 is the time constant

N is the total number of iterations in the adaptation process

Where $d_{j,c(t)}$ is the Euclidean distance between neuron j and the winner neuron $c(t)$.

Where $\eta(t)$ is the time-varying learning rate at t^{th} iteration. The learning rate serves to moderate the learning step of each iteration. It will decrease with time so as to facilitate the convergence of the map. The learning rate at t^{th} iteration is computed as

$$\eta(t) = \eta_0 \exp\left(-\frac{t}{\tau_2}\right), t = 0, 1, 2, \dots, N$$

where

η_0 is the initial learning rate

τ_2 is a time constant which is set to the total number of iterations in the adaptation process, i.e. N

The steps 2–4 are repeated for N number of iterations in the adaptation process.

3 Proposed Methodology

We propose three methods, viz. VLAD with SOM, BOF with SOM and BOF with SOM and Combined histogram and compare these with the case in which k-means clustering is used in place of SOM-based clustering in all of these proposed methods.

SOM in CBIC.

- Implementation of SOM in CBIC:
 1. Input—Key points are detected using SIFT detector and described using SIFT descriptor. Obtained key points descriptors are passed to SOM function along with number of neurons n, total number of iterations N, initial value of effective width σ_0 and initial value of learning rate η_0 .
 2. Clustering—Key points descriptors are clustered into n number of clusters by SOM function.
 3. Output—A vector with index of cluster centre to which each key point belongs and a vector with value of d-dimensional cluster centres are returned by the SOM function.

1. VLAD with SOM

- Training phase

First, SIFT features are extracted from all the training images. Then all the extracted key point descriptors are stored in single row of a matrix and passed to SOM-based clustering function which clusters the descriptors into clusters and gives the output.

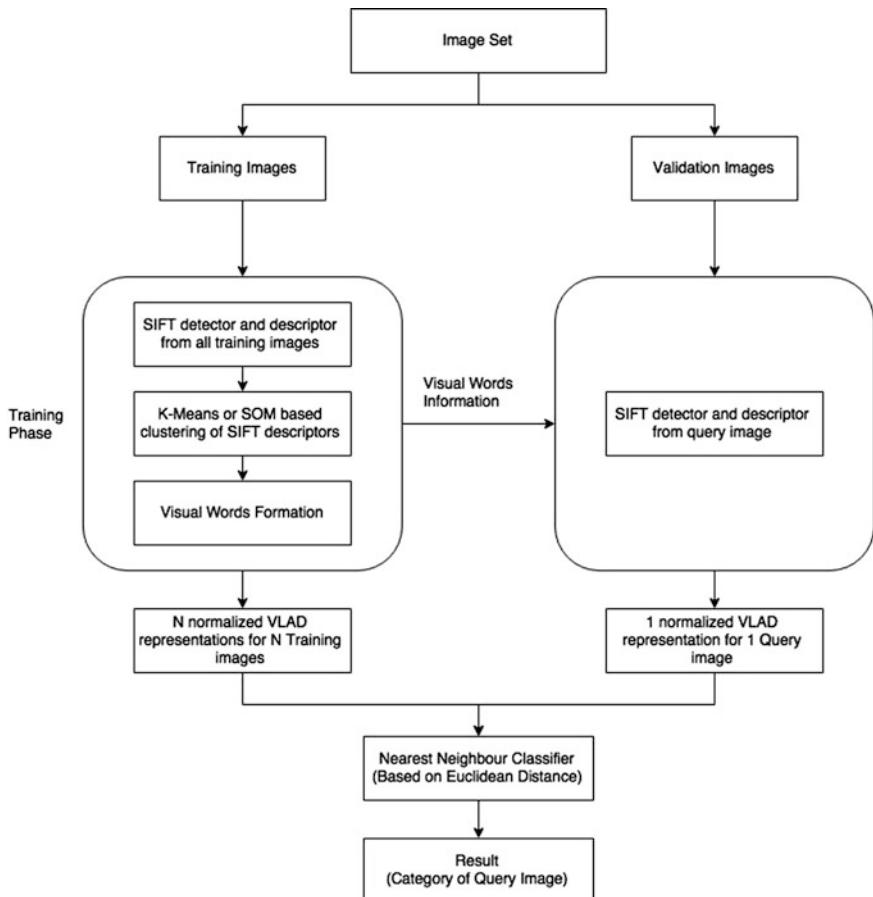
Sum of differences of cluster centre from each key point descriptor assigned to a cluster centre is calculated for each cluster centre and stored in a vector representation. This vector representation is called VLAD representation of image.

- Validation phase

For a given query image, each extracted SIFT descriptor is assigned to a cluster, on the basis of Euclidean distance from the cluster centre, obtained in previous step from SOM function. VLAD representation of a given query image is found out as explained in training phase and its Euclidean distance from VLAD representation of each training image is found. Query image is assigned a category

using NN classification. Flowchart for the process shows the process graphically with left portion showing training phase block and right portion showing testing phase block.

Flowchart for VLAD:



2. BOF with SOM

- Training phase

First, SIFT features are extracted and passed to SOM function to get the required output as in VLAD with SOM.

A histogram for each training image is made in terms of number of key point descriptors assigned to each visual word or cluster centre and stored in a vector. This vector representation is called BOF representation of image.

- Validation phase

For a given query image, each extracted SIFT descriptor is assigned to a cluster as in VLAD with SOM. BOF representation of the given query image is found out as explained in training phase and its Euclidean distance from BOF representation of each training image is found. Query image is assigned a category using NN classification.

3. BOF with SOM and Combined histogram

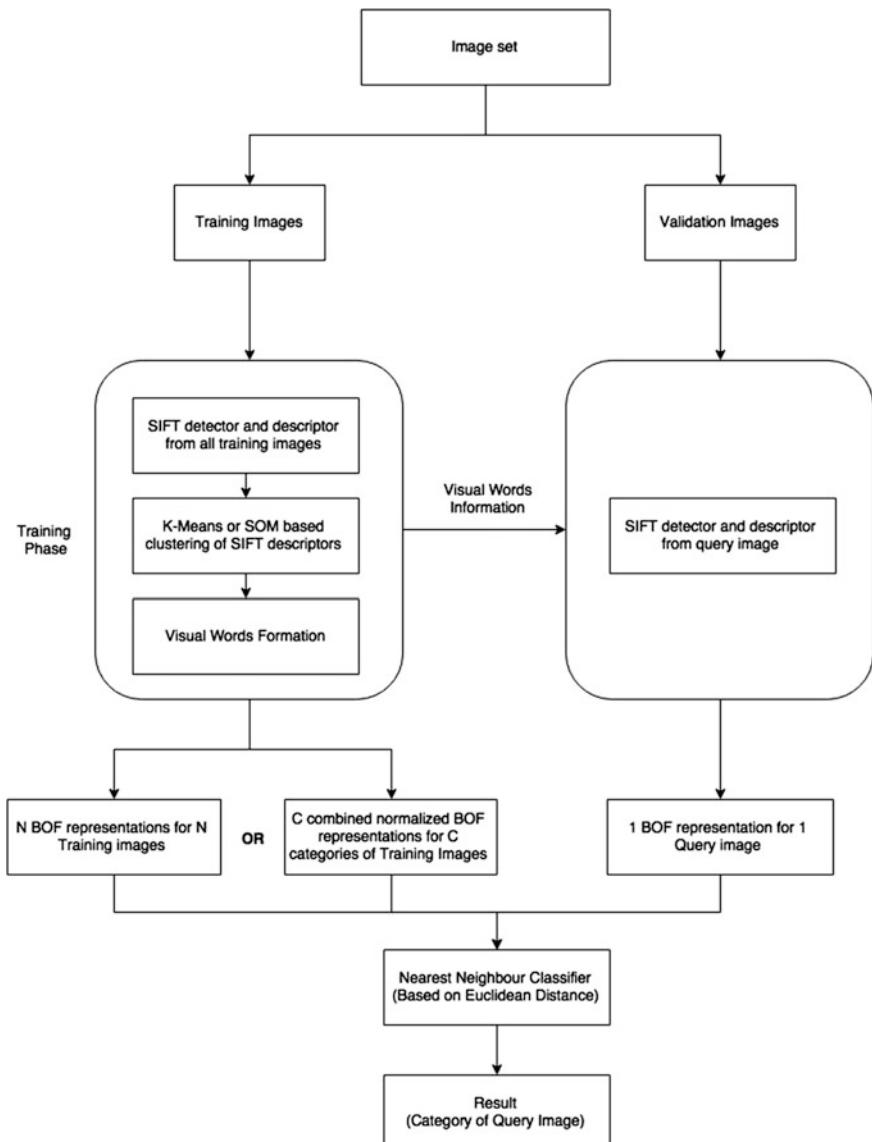
- Training phase

All the steps in this phase are same as that in training phase of BOF with SOM except that instead of creating separate histograms for all training images from each category, only one histogram is created for each category. Sum of all histograms for training images of a category are added and stored in a vector. This vector representation is normalized. This results in only one histogram for each category.

- Validation phase

BOF representation of a given query image is found as explained in BOF with SOM and its Euclidean distance from the histograms obtained in training phase is found. Query image is assigned a category using NN classification. Flowchart for BOF with separate histograms for each training image and with combined histogram is provided. It shows the process graphically with left portion showing training phase block and right portion showing testing phase block

Flowchart for BOF and with combined histogram:



4 Results

The experiments are performed on three categories of images of Caltech101 dataset, viz. ‘airplanes’, ‘motorbikes’ and ‘faces’. Two hundred images were selected randomly from each category and randomly divided into training and validation sets. The results of proposed method are compared with state-of-the-art methods, viz. VLAD with k-means clustering and BOF with k-means clustering. Comparison is done in terms of precision (P), recall (R), F-Measure (FM) and execution time values. FM is calculated as the harmonic mean of P and R (Table 1).

For VLAD, use of SOM gives better results in terms of average F-Measure (FM) and execution time both. Improvement of 6 % is observed in average FM value. Time taken while using k-means is about 3 times more than that of SOM.

For BOF, use of SOM gives better results in terms of execution time for the similar FM values. Average FM values are similar for both cases. Time taken while using k-means is about 1.5 times more than that of SOM.

For BOF with combined histogram, use of k-means gives an improvement of about 7 % in average FM value. But execution time is about 6 times more in case of k-means which is not practical. So, use of SOM gives better results in this case also.

Table 1 Results of different methods

Method	Category	P (%)	R (%)	FM (%)	Avg FM (%)	Execution time (Sec)
VLAD with SOM	Airplanes	98.60	75	85.19	73.52	560
	Motorbikes	73.80	48	58.16		
	Faces	62.89	100	77.21		
VLAD with k-means	Airplanes	78.43	80	79.20	67.01	1800
	Motorbikes	70.76	46	55.75		
	Faces	57.89	77	66.09		
BOF with SOM	Airplanes	87.71	78	82.57	87.79	1500
	Motorbikes	88.99	97	92.82		
	Faces	88.00	88	88		
BOF with k-means	Airplanes	85.71	84	84.84	87.26	2301
	Motorbikes	81.14	99	89.18		
	Faces	98.75	79	87.77		
BOF with SOM and Combined Histogram	Airplanes	92.53	62	74.24	83.14	1500
	Motorbikes	91.26	94	92.60		
	Faces	73.07	95	82.60		
BOF with k-means and Combined Histogram	Airplanes	92.47	86	89.11	90.38	2100
	Motorbikes	100	86	92.47		
	Faces	81.81	99	89.58		

5 Conclusion

Results show the superiority of SOM for clustering purpose as compared to k-means clustering in CBIC techniques. Techniques considered were VLAD, BOF and BOF with combined histogram. Large improvements in execution time with comparable accuracy prove the suitability of SOM for real-time purposes. Accuracy of the techniques considered can be further improved using better classifiers like K-Nearest Neighbour (KNN) or Support Vector Machines (SVM).

References

1. A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. "Content-Based Image Retrieval at the End of the Early Years." *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22(12), pp. 1349–1380, 2000.
2. P.P. Singh and R.D. Garg. "Classification of high resolution satellite image using spatial constraints based fuzzy clustering." *Journal of Applied Remote Sensing*, Vol. 8(1), pp. 083526 (1–16), 2014.
3. P.P. Singh and R.D. Garg. "Land Use And Land Cover Classification Using Satellite Imagery: A Hybrid Classifier And Neural Network Approach." *Proceedings of International Conference on Advances in Modeling, Optimization and Computing*, pp. 753–762, 2011.
4. D. Lowe. "Distinctive image features from scale-invariant key points." *International Journal of Computer Vision*, vol. 60(2), pp. 91–110, 2004.
5. J. Sivic and A. Zisserman. "Video google: A text retrieval approach to object matching in videos." *Proceedings of 9th IEEE International Conference on Computer Vision*, vol. 2, pp. 1470–1477, 2003.
6. G. Csurka, C. Bray, C. Dance, and L. Fan. "Visual categorization with bags of keypoints." *Workshop on Statistical Learning in Computer Vision European Conference on Computer Vision*, pp. 1–22, 2004.
7. H. Jégou, M. Douze, and C. Schmid. "Packing bag-of features." *Proceedings of 12th IEEE International Conference on Computer Vision*, pp. 2357 – 2364, 2009.
8. H. Jégou, M. Douze, and C. Schmid. "Improving bag-of features for large scale image search." *International Journal of Computer Vision*, vol. 87, pp. 316–336, February 2010.
9. L. Fei-Fei, R. Fergus and P. Perona. "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories." *IEEE, CVPR 2004, Workshop on Generative-Model Based Vision*, pp. 178–178, 2004.
10. O. Chum, J. Philbin, M. Isard, and A. Zisserman. "Scalable near identical image and shot detection." *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pp. 549–556, 2007.
11. O. Chum, J. Philbin, and A. Zisserman. "Near duplicate image detection: min-Hash and tf-idf weighting." *Proceedings of 19th British Machine Vision Conference*, pp. 50.1–50.10, 2008.
12. Xin Chen, Xiaohua Hu and Xiaojong Shen. "Spatial Weighting for Bag-of-Visual-Words Representation and Its Application in Content-Based Image Retrieval." *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09)*, pp. 867–874, 2009.
13. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. "Lost in quantization: Improving particular object retrieval in large scale image databases." *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

14. O. Chum, A. Mikulik, M. Perdoch, and J. Matas. "Total recall II: Query expansion revisited." Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 889 – 896, 2011.
15. R. Arandjelović and A. Zisserman. "Three things everyone should know to improve object retrieval." Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911 – 2918, 2012.
16. F. Perronnin and C. R. Dance. "Fisher kernels on visual vocabularies for image categorization." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1–8, 2007.
17. H. Jegou, M. Douze, C. Schmid, and P. Perez. "Aggregating local descriptors into a compact image representation." Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3304 – 3311, June 2010.
18. H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. "Aggregating local image descriptors into compact codes." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, pp. 1704 – 1716, September 2012.
19. R. Arandjelović and A. Zisserman. "All about VLAD." Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1578–1585, 2013.
20. Teuvo Kohonen. "Self-Organizing Maps." Springer Series in Information Sciences, vol. 30, 1995.
21. T.K. Kohonen. "Essentials of the Self-Organizing Map." Neural Networks, vol. 37, pp. 52–65, 2013.
22. Piotr Plonski and Krzysztof Zaremba. "Improving Performance of Self-Organising Maps with Distance Metric Learning Method." Artificial Intelligence and Soft Computing, vol. 7267, pp. 169–177, 2012.
23. E. C. Vargas, R. Francelin Romero and K. Obermayer. "Speeding up algorithms for SOM family for large and high dimensional databases." Proceedings of the Workshop on Self Organizing Maps, pp. 167–172, 2003.
24. T. Kohonen. "Comparison of SOM Point Densities Based on Different Criteria." Neural Computation, vol. 11, pp. 2081–2095, 1999.

Road Surface Classification Using Texture Synthesis Based on Gray-Level Co-occurrence Matrix

Somnath Mukherjee and Saurabh Pandey

Abstract Advance Driving Assistance System (ADAS) has been a growing area of interest in the research community for automotive domain where scene understanding and modeling is one of the principally focus area of activities. Texture synthesis using gray-level co-occurrence matrix (GLCM) of any rigid body is not an exceptional task in image processing area. The additional integration of this method is for texture characterization and use it for the road surface classifications which is the primary focus of this paper. We have also introduced that GLCM based road surface analysis in a line scan manner that can be used as a module for ADAS application.

Keywords Gray-level co-occurrence matrix • Road texture • The Advance Driving Assistance System (ADAS) etc.

1 Introduction

Road quality monitoring and surface type classification of the road and there by detection of pot hole, smoothness or roughness of the road surface is one of the key features of the Advance Driving Assistance System (ADAS). There are some recent work has been addressed in [1] where road quality measurement like potholes, surface cracks has been mentioned for need to detection and to identify and locate the area by continuous monitoring of roads based on Geographical Information systems (GIS) which is totally based on road information system that keeps track of

S. Mukherjee (✉) · S. Pandey

Computer Vision and Image Processing Unit, Kritikal Solutions Pvt. Ltd,
Bangalore, India

e-mail: somnath.7.mukherjee@gmail.com; somnath.mukherjee@kritikalsolutions.com

S. Pandey

e-mail: saurabh@kritikalsolutions.com

the road condition and surface type. It can be considered to inspect the observation of the road by technological solution instead of observing the ground truth from manual inspections. In recent decades most of the vehicle has to be carried out the Advance driving Assisting System (ADAS) which is followed by camera system mounted in the front of the vehicle. The proposed system is only based on the single view monocular camera which is very cost effective now a days and the decision making system is also very low process in terms of algorithmically development as this system is totally depends on simplest methodology for texture analysis which is Gray-Level Co-occurrence matrix (GLCM). There are different methodologies for the analysis of textures like Texton theory, the Wavelet based approach, the Fourier approach. However, the simplest way for texture analysis is using GLCM which is interesting as it is proven to be related to the way the human visual system perceives texture, which is the very first approach to texture analysis defined by Haralik [2] and it is still widely used. Accordingly [2] 14 statistical features has been introduced to estimate similarity between different Gray-Level Co-occurrence matrices. Here we are mainly focusing to develop this system for ADAS technology in which where only highly compute optimal algorithm available for constraint embedded environment can exist, that's why some of relevant features has been considered like dissimilarity, correlation, homogeneity, energy, entropy to reduce the computational complexity so that this can be considered for ADAS along with a very sufficient decision making framework where features that are computed from GLCM are based on the assumption of the texture information in an image can be considered as contained in the overall spatial relationship which Gray levels of neighboring pixels have to one another. GLCM contains information about the frequency of occurrence of two neighboring pixel combination in a particular spatial image domain. An analogous work can be considered in [3] where road surface has been classified in terms of texture analysis on the basis of material of the road using mathematical morphology which is generally used for structuring elements with various shape and sizes. In Ref. [4–6] the various types of Road detection approaches has been described, which can tell us about the review of some research work in related to this. The Ref. [7–13] has been clearly described in various types of texture classification using GLCM in different types of applications like in Image Segmentation, SAR image Analysis etc. In Ref. [14] an application of gray level co-occurrence matrix (GLCM) has been used to estimate texture based similarity of rock images.

The Fig. 1 describe one of the road data which is based on KITTI [Reference] Vision benchmark road data sets for road segmentation and also few patches has been depicted in figure for correspondingly road texture. Accordingly Fig. 2 has been arranged to show the corresponding patch of size 21×21 in terms of gray value for different five road samples. Primarily this paper is focusing for road texture analysis based on some relevant texture characteristics using GLCM and

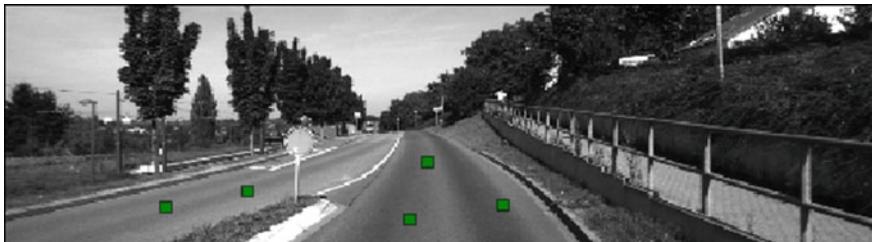


Fig. 1 Original image with patches

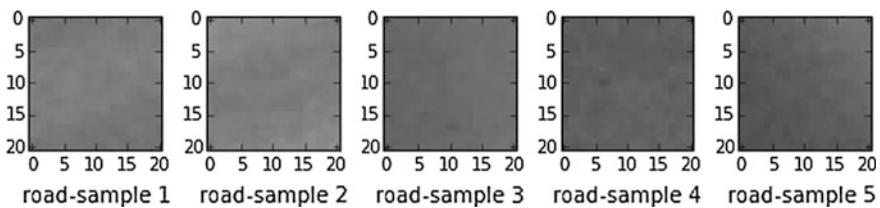


Fig. 2 Five road patches with size 21×21

how to differentiate from various type of surface of road like roughness, smoothness, pot hole presents or not along with decision making system which can be considered as an ADAS features.

2 Overall Methodology

Gray level co-occurrence matrix (GLCM) [2], one of the most well known texture analysis methods considered to be used for estimates image properties. Each entry of the image matrix element in GLCM corresponds to the number of occurrences of the pair of gray levels for each row and column which are a specific distance apart in original image. In the below there is an example of GLCM description, Table-1 consider an image with 4×4 matrix and corresponding value in range in between $\{1, \dots, 10\}$, Table-2 depicts the corresponding GLCM matrix table where it can be easily observed that co-occurrence of gray value in the Table-1 has been mapped into the Table-2. In [15] it has clearly mentioned about the computation of the GLCM is not only the displacement but also the orientation between neighborhood image pixels must be established. The orientations also can be in horizontal (0°) or in vertical (90°) and two diagonal (45° and -45°). As this example has been considered in horizontally (0°) pixel relationship demonstrated in the Table-1 like 5 and 6 gray value has 3 times occurrence and thereby by it has been mapped as 3 in Table-2.

Table 1

	1	2	3	4
1	5	6	10	9
2	4	5	6	10
3	10	9	4	5
4	5	6	10	8

Table 2

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	2	0	0	0	0	0
5	0	0	0	0	0	3	0	0	0	0
6	0	0	0	0	0	0	0	0	0	3
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	1	2	0

According to Mryka Hall-Beyer's web pages [16] for GLCM tutorial as there are some basic criteria are presents for the texture calculations which is required a symmetrical matrix, so the next step is therefore to get the GLCM into as a symmetric form. A symmetrical matrix can be defined in such way that the same values occur in cells on opposite sides of the diagonal. The next step is to make it symmetrical and after that the measures for calculation require that each GLCM cell contain the probability of occurrences rather than to measures the count. So basically the last steps determined to estimate the number of times this gray value relationship occurs, divided by the total number of possible gray value occurrences.

$$P_{i,j} = \frac{V_{i,j}}{\sum_{i,j=0}^{N-1} V_{i,j}} \quad (1)$$

In equation [3] the probability of the GLCM matrix has been described where $V_{i,j}$ denotes the number of occurrences of the specified outcome and the denominator term determines the total number of possible outcomes, I and j are the number of column and rows in GLCM matrix.

2.1 GLCM Feature Estimation

To estimate the road surface using texture analysis we have considered 4 different types of relevant features Homogeneity, Dissimilarity, Correlation and Energy which are basically widely used in the literature.

Homogeneity. Homogeneity can be measured by evaluating the uniformity of the non-zero entries in the GLCM matrix [16]. It can be defined as the weight values by the inverse of contrast weight according to [13].

$$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1 + (i-j)^2} \quad (2)$$

The GLCM homogeneity feature of any texture will be very high when the concentration of Co-occurrence matrix along in the diagonal basis, it indicates that there are a lot of pixels in the image area either with the same or very similar gray value. The larger gray values changes on the image will reflect the lower homogeneity and there by increasing higher GLCM contrast. The range of homogeneity is belongs to in between 0 and 1. If there is any little variation on the image then the value of homogeneity will be very high and so if there is no variation then homogeneity will be equal to 1. Therefore, higher value of homogeneity refers to the textures that contain perfect repetitive structures of arrangement, In other hand lower value of homogeneity refers to isolatable variation in both, texture elements as well as their spatial rearrangements. A texture which has distribution in homogeneously refers to an image that has almost no repetition of texture elements and spatial similarity in it is absent, It can be consider according to Eq. (2).

2.2 Energy

Energy, also called Angular Second Moment [2] and Uniformity is a measure of textural uniformity of an image. Energy reaches its highest value when gray level distribution has either a constant or a periodic form. A homogenous image contains very few dominant gray tone transitions, and therefore the P matrix for this image will have fewer entries of larger magnitude resulting in large value for energy feature. In contrast, if the GLCM probability matrix contains a large number of small entries, the energy feature will have smaller value can be depicted through Eq. (3)

$$Energy = \sqrt{\sum_{i,j=0}^{N-1} P_{i,j}^2} \quad (3)$$

2.3 Dissimilarity

This GLCM feature can be considered as a measure that defines the variation of gray level pairs onto an image. It is the closest to Contrast with a difference of the weight and Contrast unlike the nature of dissimilarity [13]. It is expected that these two different measures behave in the same way for the same texture because they calculate the same parameter with different weights. Contrast will always give slightly higher values than Dissimilarity. Dissimilarity ranges from [0,1] and obtain

maximum when the gray level of the reference and neighbor pixel is at the extremes of the possible gray levels in the texture sample. The dissimilarity can be considered by Eq. (4)

$$\sum_{i,j=0}^{N-1} P_{i,j}|i-j| \quad (4)$$

2.4 Correlation

The Correlation texture measures the linear dependency of gray levels on those of neighboring pixels. GLCM Correlation can be calculated for successively larger window sizes. The window size at which the GLCM Correlation value declines suddenly may be taken as one definition of the size of definable objects within an image. It finds its similarity with autocorrelation as there is a resemblance to the information provided by them and by [16] have demonstrated that GLCM Contrast is identical to semi variance, and GLCM Correlation provides almost identical information as provided by autocorrelation methods using Moran's I or Geary's C. GLCM Correlation is quite a different calculation from the other texture measures described above. As a result, it is independent of them (gives different information) and can often be used suitably in combination with another texture measure. It also has a more intuitive meaning to the actual calculated values: 0 is uncorrelated, 1 is perfectly correlated, it can be estimated by Eq. (5) where μ and σ^2 represents the mean and variance, calculated by GLCM matrix elements.

$$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i - \mu_i)(j - \mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right] \quad (5)$$

3 Testing Road Datasets

We have tested our experiments through the KITTI Vision data sets for road segmentation. The data sets are widely used in the literatures for various purposes like road segmentation and separation, road analysis etc. We have manually marked the area first and apply the patches as described in our introductory part. The patch size we have taken 21×21 described in Fig. 3. If the road segmentation has perfectly done then we can apply for linearly scanning through the road from where we can analyze the road texture in terms of surface classifications like pothole or any kind of road distortion through a single camera.



Fig. 3 Original image with discontinuity of the road

4 Experimental Result

We collected and tested on the basis of the benchmark data sets by using some specific sample where this approach can be evaluated perfectly, Fig. 4 is one of the sample data sets where we can see the discontinuity of the road surface in the middle of the road. Our approach is to calculate Gray-Level Co-occurrence matrix (GLCM) by linearly scanning method on the road area where we used 21×21 patches for every location of the line above the road which can be described as Fig. 5. GLCM features like dissimilarity, correlation, homogeneity and energy has been evaluated accordingly and thereby we plotted the result matrix value in X and Y direction correspondingly. Figure 6 is the plot of GLCM dissimilarity and correlation, same as in Fig. 7 is the energy and homogeneity for the specified line scanned area as in Fig. 5 of the road.

4.1 GLCM Feature Analysis

Figure 6 depicts the plot of GLCM correlation and dissimilarity, which can be analyzed in such way that if any discontinuity has been observed in road texture



Fig. 4 Linearly scanning on the road

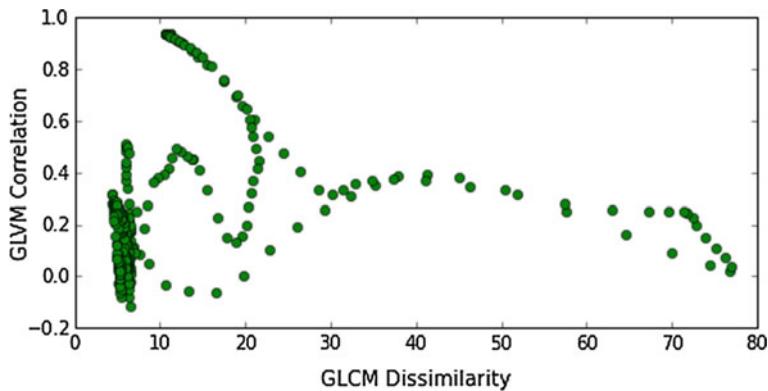


Fig. 5 GLCM—feature analysis using dissimilarity and correlation

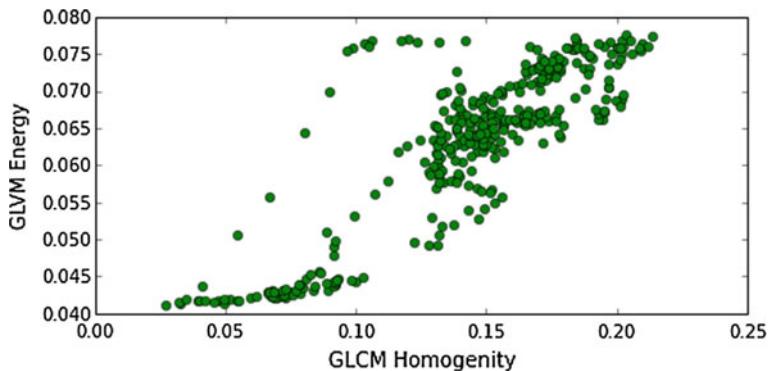


Fig. 6 GLCM—feature analysis using homogeneity and energy

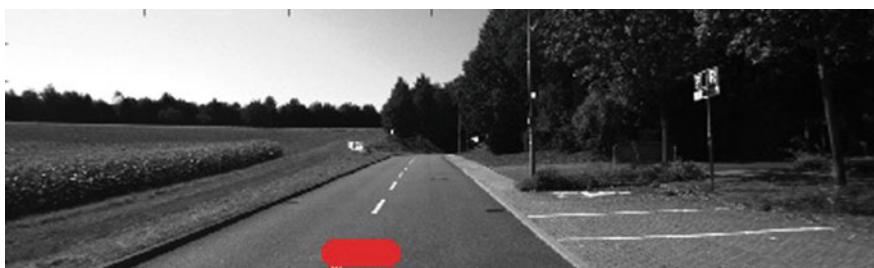


Fig. 7 Detected area which has very low energy and high homogeneity and hence can be easily separable with other area



Fig. 8 Line scanning of the road surface

then it can be separable in simple liner clustering along with if we observe on the Fig. 7 which is homogeneity and energy then it can also be very easily separable by clustering like K-means. Our experimental result has been mapped to the road locations from where it can be easily mapped and it can be decided the discontinuity and here it is described in Fig. 8 where we have marked the region on the road. According to our line by line scanning methodology over the road it can be observed in Fig. 9 which can describe the next scanning process, there by the same GLCM feature has been extracted and plotted in Figs. 10 and 11 which can be distinguishable compare to Figs. 6 and 7. This over all methodology can be easily integrate to any road segmentation methodology for analysis road surface and classified into various categories like pothole, rock cracked, smoothness, roughness of the road which is very essential for any system to identify. The whole system analysis methodology has been clearly mentioned in Fig. 11.

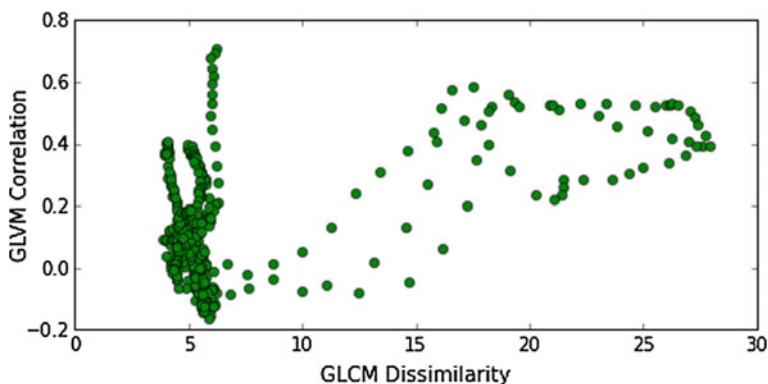


Fig. 9 GLCM—feature analysis using dissimilarity and correlation

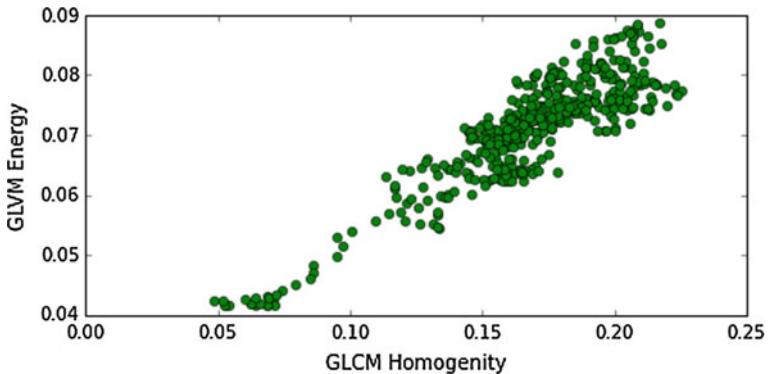


Fig. 10 GLCM—feature analysis using homogeneity and energy

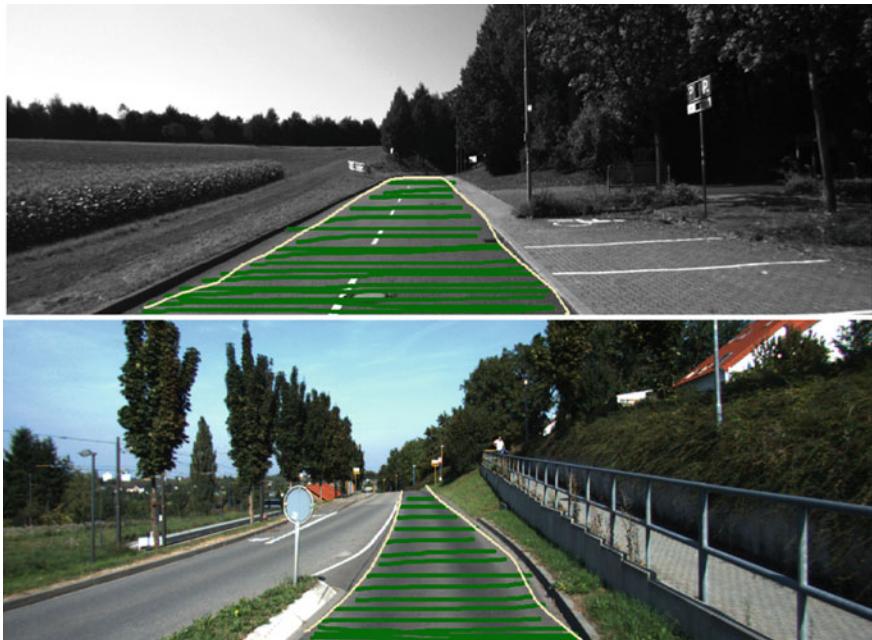


Fig. 11 Line scanning on the road after successfully road segmentation

5 Conclusion

In this paper we have studied the problem of road surface quality analysis using texture method where we have shown that GLCM based features provide valuable inputs can be used as decision making mechanism. In the future work we are will

focus ib fully automatically road quality monitoring where the road segmentation will be executed automatically without any manual input and also the Same system need to apply to make it a purely autonomous system for ADAS purpose where driving assistance system will be more smooth, tangential and safe. There are various kind of opportunity for future work along with this research work which need to be carried out and whether it can be a fully automatic and smart in terms of automatic update with GPS and generate a very smart road map for ADAS integration.

Acknowledgments The work has supported by Computer Vision and Image Processing Lab of Kritikal Solutions Pvt. Ltd., Bangalore, India.

References

1. A Prototype System for Road Condition and Surface Type Estimation by Fusing Multi-Sensor Data, Deepak Rajamohan et all, ISPRS International Journal of Geo-Information ISSN 2220-9964, 2015.
2. R.M. Haralick, K. Shanmungam, and I. Dinstein, "Textural Features of Image Classification," IEEE, vol. 3, pp. 610–621, 1973.
3. Road Surface Textures Classification Using Opening-Based Image Processing St'ephane et all, 2 1Lab. Central des Ponts et Chaussées – Nantes BP 4129.
4. Automatic Road Pavement Assessment with Image Processing: Review and Comparison, Sylvie Chambon et all, HAL Id: hal-00612165.
5. Survey: Vision based Road Detection Techniques, Vipul H. Mistry et all, International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 4741–4747.
6. P. Maillard, "Comparing texture analysis methods through classification," Photogrammetric Engineering and Remote Sensing, vol. 69, pp. 357–367.
7. J.F. Haddon, and J.F. Boyce, "Image segmentation by unifying region and boundary information," IEEE Trans. Pattern Anal. Machine Intell., vol. 12, pp. 929–948, 1990.
8. E. Sali, and H. Wolfson, "Texture classification in aerial photographs and satellite data," International Journal of Remote Sensing, vol. 13, pp. 3395–3408, 1992.
9. G.J. Hay, K.O. Niemann, and G.F. McLean, "An object-specific Image texture analysis of hi-resolution forest imagery", Remote Sensing of Environment, vol. 55, pp. 108–122, 1996.
10. J.H. Xin, and H.L. Shen, "Computational models for fusion of texture and: a comparative study", Journal of Electronic Imaging, vol. 14, pp. 033003, 2005.
11. Zhang, J.; Nagel, H.-H. "Texture based segmentation of road images. In Proceedings of the Symposium of Intelligent Vehicles, Paris, France, 24–26 October 1994.
12. A. Baraldi and F. Parmiggiani, "An investigation of the Textural Characteristics Associated with Gray Level Cooccurrence Matrix Statistical Parameters", IEEE Trans. On Geoscience and Remote Sensing, vol. 33, no. 2, pp. 293–304, 1995.
13. L.K. Soh, and C. Tsatsoulis, "Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices", IEEE Transactions on Geoscience and Remote Sensing, vol 37, 1999.
14. Rock Texture Retrieval Using Gray Level CO-Occurrence Matrix, Mari Partio et all, Tampere University of Technology.
15. Texture Characterization based on Gray-Level Co-occurrence Matrix, A. Gebejes, R. Huertas, ICTIC 2013.
16. Mryka Hall-Beyer's web pages, The GLCM Tutorial Home Page.

Electroencephalography-Based Emotion Recognition Using Gray-Level Co-occurrence Matrix Features

Narendra Jadhav, Ramchandra Manthalkar and Yashwant Joshi

Abstract Emotions are very essential for our day-to-day activities such as communication, decision-making and learning. Electroencephalography (EEG) is a non-invasive method to record electrical activity of the brain. To make Human-Machine Interaction (HMI) more natural, human emotion recognition is important. Over the past decade, various signal processing methods are used for analysing EEG-based emotion recognition (ER). This paper proposes a novel technique for ER using Gray-Level Co-occurrence Matrix (GLCM)-based features. The features are validated on benchmark DEAP database upto four emotions and classified using K-nearest neighbor (K-NN) classifier.

Keywords EEG · HMI · GLCM · K-NN · ER

1 Introduction

Affective Computing is an interdisciplinary branch that relates to emotion or other affective state. It is the bridge between emotions and cognition. Cognition is an important aspect of emotion. Rafael and Sidney [1] discussed theoretical perspectives that view emotions as expressions, embodiments, outcomes of cognitive appraisal, social constructs, product of neural circuitry, and psychological interpretations of basic feelings. It also reviews affect detection modalities such as facial expressions, voice, body language, physiology (EMG, EOG, ECG and EEG), brain

N. Jadhav (✉) · R. Manthalkar · Y. Joshi

Department of Electronics and Telecommunication Engineering, Shri Guru Gobind Singhji

Institute of Engineering and Technology, Nanded, India

e-mail: jadhavnarendra@sggs.ac.in

URL: <http://www.sggs.ac.in>

R. Manthalkar

e-mail: rrmanthalkar@sggs.ac.in

Y. Joshi

e-mail: yvjoshi@sggs.ac.in

imaging (fMRI) and text. EEG signal cannot be intentionally controlled and hence in the past few decades most researchers are trying to recognize emotions through EEG. EEG-based emotion recognition research can be easy than Computed Tomography (CT) and Functional Magnetic Resonance Imaging (fMRI) especially when millisecond temporal resolution is required.

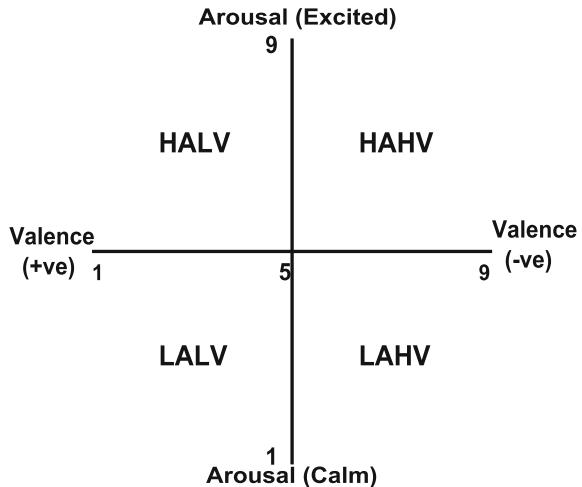
The electrical activity of the brain can be measured by EEG using 32 channel electrodes with 8064 (128 sampling frequency x 63 sec data) samples for each emotion. This 2-D matrix (32 channels x 8064 samples) represent the texture image for each emotion. The texture image can be described by Gray-Level Co-occurrence Matrix (GLCM). It is used as texture descriptor. GLCM is symmetric matrix that contains the count of paired i and j gray levels separated by different distances and directions [2]. GLCM statistical features are useful in texture or image analysis methods [2–4]. GLCM textural features are based on two neighbouring pixel intensity values which are in different directions (0° , 45° , 90° , 135°) and distances. In this work, a novel approach is proposed for EEG-based emotion recognition in terms of texture image using GLCM features. EEG for different emotions such as happy, angry, sad and relax represent texture image for each emotion. Here, texture image is formed due to different emotion effects on EEG 32 channels. As per the literature survey, GLCM-based approach for emotion recognition is not used yet to represent emotion in terms of texture. It is differentiated using spectrogram image also. Here GLCM image intensity is nothing but the amplitude of the EEG samples (8064) per channel used for emotion acquisition. The objective of this paper is to show the relation between EEG 32 channels, i.e., the asymmetry between channels and what is correlation between one channel to another channels using GLCM features. Contrast, correlation, energy and homogeneity are used as GLCM features and classified using K-NN classifier.

The organization of the paper is as follows: Sect. 2 introduced related works of EEG-based emotion recognition using GLCM features. In Sect. 3, materials and methods used in this paper such as DEAP dataset, Valence-Arousal 2-D plane and flow chart of experiment are discussed. We give result in Sect. 4 with discussion and conclude the paper in Sect. 5.

2 Related Work

Haralick et al. [2] were originally developed GLCMs to sort the massive number of satellite images. Here we discussed GLCM features used in physiological signals. The electrocardiogram (ECG) spectrogram image is used in [5, 6] to study the heart abnormalities using GLCM features. Mustafa et al. [7] used EEG spectrogram image from which GLCM texture feature were extracted and in [8] GLCM texture feature used for IQ application. To classify the medical images in various applications using GLCM and K-NN classifier studied in [9].

Fig. 1 2-D Valence-Arousal model



3 Materials and Methods

3.1 DEAP Dataset

In this work, we used preprocessed DEAP EEG dataset [10] and is freely available on [11]. The DEAP dataset consists of two important datasets 1. EEG data 2. Peripheral Physiological data such as EOG, EMG, GSR, Temperature, Heart rate and respiration rate. The EEG and peripheral physiological data of 32 participants were recorded as each watched 40 one-minute long music videos. The EEG data is available for each 32 participants in 3-D format means array shape of 40 (Video/trial) x 40 (32 EEG data and 8 peripheral data) x 8064 (data samples). The videos/trials are reordered from presentation order to video order. The EEG data is recorded for 63 sec using 128 Hz down sampled frequency ($128 * 63 = 8064$ samples). The EOG artefacts were removed and a bandpass filter from 4 to 45 Hz was applied. The pre-processed DEAP EEG data is used in this work.

3.2 Valence-Arousal Plane

In this paper, human emotions are expressed by 2-D Russells valence-arousal plane [12] because of its simplicity. The vertical and horizontal axis of 2-D plane represents arousal and valence respectively as shown in Fig. 1. Arousal scale ranges from excited to calm and valence ranges from positive to negative. The arousal and valence scale from 1 to 9. Low value is scale from 1 to 4 and high value is scale from 6 to 9. The 2-D arousal and valence model is divided into four parts i. e., Low

Arousal Low Valence (LALV), Low Arousal High Valence (LAHV), High Arousal Low Valence (HALV), and High Arousal High Valence (HAHV) as shown in Fig. 1. We used four basic emotions happy (HAHV), angry (HALV), relax (LAHV) and sad (LALV). The 40 videos are divided into 4 parts using valence and arousal values [10, 11].

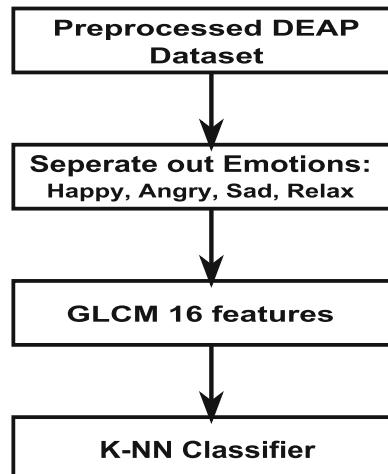
The flow chart of experiment is shown in Fig. 2. The input for our work is pre-processed DEAP data that is 40 (Video/trial) x 32 (EEG data) x 8064 (data samples) for 32 participants. The four emotions: Happy, Angry, Sad and Relax are separated out using valence and arousal value and GLCM four features contrast, correlation, energy and homogeneity for four offset [0, 1], [-1, 1], [-1, 0], [-1, -1] were calculated. Total 16 features calculated for each emotion video. Then we classified emotions using K-NN classifier and calculate the classification accuracy for two, three and four emotions combination.

3.3 GLCM and K-NN Classification Method

A general definition of texture is rough to touch means a difference between high and low values. In image texture, the high and low values means grey levels or brightness values of spatial positions in image. In this work, the spatial positions means the 32 EEG channels arranged accordingly asymmetry and lobe position. A co-occurrence matrix [2] is a matrix over an image to be the distribution of co-occurring values at a given offset is defined using Eq. (1) and GLCM features are defined using Eqs. (2)–(5).

$$CM_{\Delta_r, \Delta_c} = \sum_{x=1}^n \sum_{y=1}^m \begin{cases} 1 & \text{if } I(x, y) = i \text{ and } I(x + \Delta_r, y + \Delta_c) = j, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Fig. 2 Flow chart of experiment



$$contrast = \sum |i - j|^2 p(i,j) \quad (2)$$

$$correlation = \sum_{ij} \frac{(i - \mu_i)(j - \mu_j)p(i,j))}{\sigma_i \sigma_j} \quad (3)$$

$$energy = \sum_{ij} p(i,j)^2 \quad (4)$$

$$Homogeneity = \sum_{ij} \frac{p(i,j)}{1 + |i - j|} \quad (5)$$

where, (r, c) is row and column offset (0 1, -1 1, -1 0, -1 -1) represent the directions (0°, 45°, 90°, 135°) respectively, (i, j) are the image intensity (in this work, amplitude of the EEG samples 8064/channel), x and y are the spatial positions in image I of size n x m (in this work, 32 channels).

The GLCM features (Contrast, correlation, Energy, Homogeneity) calculated for four offset 0°, 45°, 90°, 135° and they are represented as 16 features for one video and one subject. The GLCM matrix for 32 channels and 32 gray-levels used to represent the four emotions in zero direction is shown in Fig. 3.

For each emotion, 40 videos are separated based on the valence and arousal values. Hence for happy 13 videos, angry 11 videos, sad 10 videos and relax 06 videos are separated. The input feature vector of each emotion for classification is the average value of all 32 subjects. The calculated feature vector of each emotion is for happy-13 × 16, angry-11 × 16, sad-10 × 16 and relax-06 × 16. This feature vector matrix is fed to K-NN classifier. In this work, Euclidean distance is used to define the nearest neighbours of a feature vector. The detailed about K-NN classifier is given in [13].

The entire data is used for training and testing the K-NN classifier. The performance of classifier is checked using a fivefold cross validation. The complete data is divided into five equal parts, out of which four parts are used for training and one part is for testing. The same division of data is used for in each fold. The classification



Fig. 3 GLCM matrix for all four emotions in zero direction

accuracy is calculated as the ratio of sum of correctly classified test vectors to total number of test vectors. The two, three and four emotions are classified using K-NN classifier and the value of K varied in odd from 1, 3, 5, 7 and 9.

4 Result and Discussions

In [7], the GLCM features have been calculated for EEG spectrogram image. In this work, 8064 samples of the 32 channels for each emotion are used as texture image for calculation of 16 GLCM features. Each emotion produces a different texture. As per the value of valence-arousal, the 13 happy, 11 angry, 10 sad and 06 relax emotion videos are used.

For two emotions, the combinations are used as follows: happy–angry, happy–sad, happy–relax, angry–sad, angry–relax and sad–relax. For three emotions, the combinations are used as follows: happy–angry–sad, happy–angry–relax, angry–sad–relax, and sad–relax–happy. For four emotions, the combinations are used as follows: happy–angry–sad–relax.

The K-NN classifier accuracy result for two, three and four emotions are shown in Figs. 4, 5 and 6 respectively. If we compared the accuracy then the accuracy of two emotions are larger than the three and four emotions.

As value of K-fold increases, accuracy also increases but after nine nearest neighbours insignificant changes occurred in accuracy hence we consider upto nine. In

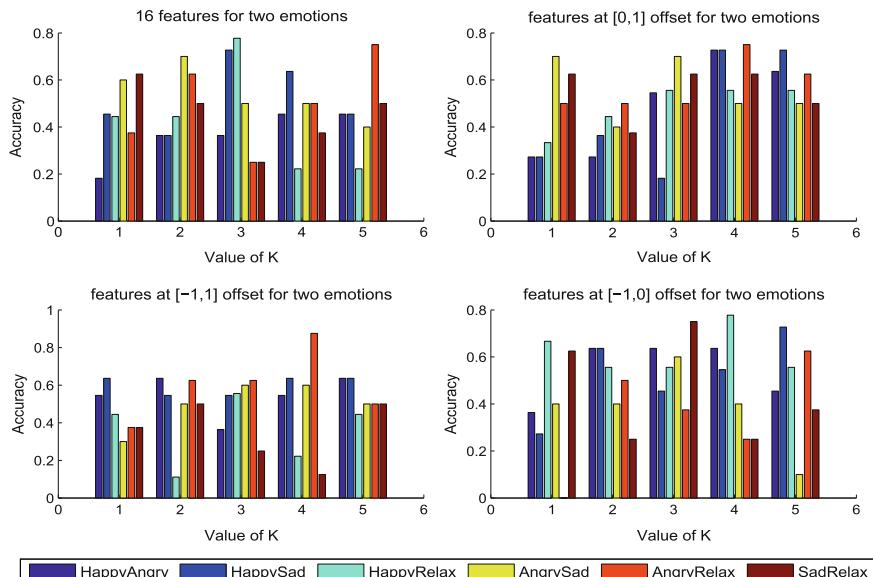
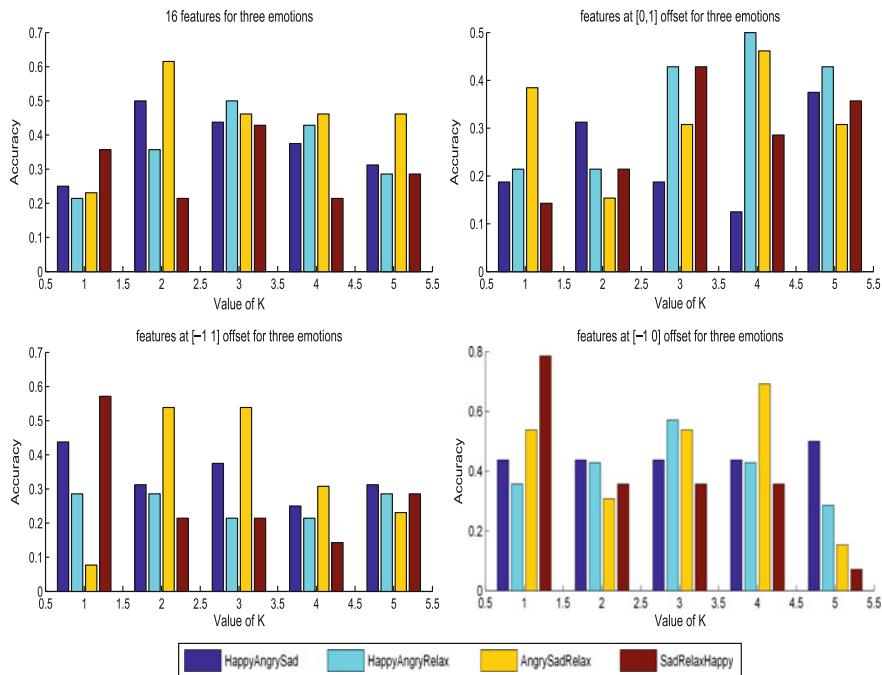
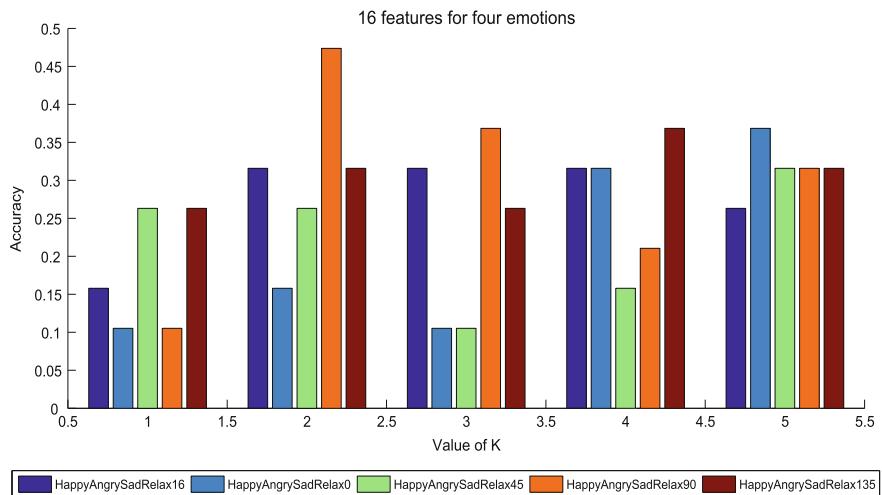


Fig. 4 Accuracy of two emotions

**Fig. 5** Accuracy of three emotions**Fig. 6** Accuracy of four emotions

paper [14], statistical, Power Spectral Density (PSD) and Higher Order Crossing (HOC) features extracted using the DEAP dataset for two emotions calm and stress. The classification accuracy of [14] using K-NN classifier is 66.25 % for statistical, 70.1 % for PSD and 69.59 % for HOC. However the average classification accuracy of our method using GLCM features are 79.58 %. The result obtained by our method for 32 channels and for 32 subjects while in [14], the results obtained for 4 channels for two emotions like calm (arousal below 4 and valence between 4 and 6) and stress (arousal above 5 and valence below 3). If we compared stress and calm emotions with our work then it is angry and relax emotions by valence and arousal values.

5 Conclusion

It could be concluded from the above results that GLCM feature extraction technique is a robust method for recognizing the emotions from EEG with competitive performance. More sample values per class increase the value of K-NN. Not much work in EEG-based emotion recognition using GLCM features has been done. To make Human–Computer Interface more natural, the classification accuracy must be improved by a different classification technique such as SVM as the future work.

References

1. Rafael A. Calvoand., Sidney DMello.: Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing*. vol. 1, no. 1, 18–37 (2010).
2. R. M. Haralick., S. Shanmugam., I. Dinstein.: Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics. SMC.*, Vol. 3, 610–621 (1973).
3. J. F. Haddon., J. F. Boyce.: Co-occurrence matrices for image analysis. *IEEE Electronics & Communication Engineering Journal*. Vol. 5, 71–83 (1993).
4. C. W. D. de Almeida., R.M. C. R. de Souza., A. L. B. Candeias.: Texture classification based on co-occurrence matrix and self organizing map. *IEEE International conference on Systems Man & Cybernetics*. University of Pernambuco Recife. 2487–2491 (2010).
5. A.A. Mohammad., B. Khosrow, J. R. Burk., E. A. Lucas., M. Manry.: A New Method to Detect Obstructive Sleep Apnea Using Fuzzy Classification of Time-Frequency Plots of the Heart Rate Variability. *28th Annual International Conference of the IEEE in Engineering in Medicine and Biology Society*, 2006. EMBS '06. 6493–6496 (2006).
6. M. Saad., M. Nor, F. Bustami., and R. Ngadiran.: Classification of Heart Abnormalities Using Artificial Neural Network. *Journal of Applied Sciences*, vol. 7, 820–825(2007).
7. Mahfuzah Mustafa., Mohd Nasir Taib., Zunairah Hj. Murat., Noor Hayatee Abdul Hamid.: GLCM Texture Classification for EEG Spectrogram Image. *IEEE EMBS Conference on Biomedical Engineering & Sciences (IECBES 2010)*, Kuala Lumpur, Malaysia, 426–429 (2010).
8. Mahfuzah Mustafa., Mohd Nasir Taib., Sahrim Lias., Zunairah Hj. Murat, Norizam S.: EEG Spectrogram Classification Employing ANN for IQ Application. *International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE 2013)*, Konya, Turkey, 199–203 (2013).

9. F. Florea, E. Barbu., A. Rogozan., and A. Bensrhair.: Using texture based symbolic features for medical image representation. 18th International Conference on Pattern Recognition, 2006. ICPR 2006, 946–949 (2006).
10. Sander Koelstra., Christian Muhl., Mohammad Soleymani., Jong-Seok Lee., Ashkan Yazdani., Touradj Ebrahimi., Thierry Pun., Anton Nijholt., Ioannis Patras.: DEAP: A Database for Emotion Analysis using Physiological Signals. IEEE Transactions on Affective Computing. vol. 3, no. 1, (2012).
11. S. Koelstra et al., 2012, DEAP Dataset available at <http://www.eecs.qmul.ac.uk/mmv/datasets/deap/>
12. J.A. Russell.: A Circumplex Model of Affect. *J. Personality and Social Psychology*, vol. 39, no. 6, 1161–1178 (1980).
13. R. O. Duda., P. E. Hart., D. G. Stork.: Pattern Classification. 2nd ed., Wiley-Interscience (2001).
14. T. F. Bastos-Filho., A. Ferreira, A. C. Atencio., S. Arjunan., D. Kumar.: Evaluation of feature extraction techniques in emotional state recognition. 4th Int. Conf. Intell. Hum. Comput. Interact. Adv. Technol. Humanit. IHCI 2012, (2012).

Quick Reaction Target Acquisition and Tracking System

Zahir Ahmed Ansari, M.J. Nigam and Avnish Kumar

Abstract The most relevant application of visual tracking is in the field of surveillance and defense, where precision tracking of target with minimal reaction time is of prime importance. This paper examines an approach for reducing the reaction time for target acquisition. Algorithm for auto detection of potential targets under dynamic background has been proposed. Also, the design considerations for visual tracking and control system configuration to achieve very fast response with proper transient behavior for cued target position have been presented which ultimately leads to an integrated quick response visual tracking system.

Keywords Automatic Target Detection (ATD) • Automatic Video Tracking (AVT) • Dynamic background • Shortest path following • Clutter suppression

1 Introduction

Object tracking is the process of locating and following a moving object (or multiple objects) over time. When motion is estimated from image sequences acquired from camera, it is called visual tracking. Here some features of these objects are selected and matched on to other frames. Closed-loop target tracking involves target acquisition, feature extraction, target representation, target localization, track maintenance, and tracking error minimization using gimbal control system. It also includes occlusion detection and handling. Visual tracking

Z.A. Ansari (✉) · A. Kumar
Instruments Research & Development Establishment, Dehradun, India
e-mail: acadzahir@gmail.com

M.J. Nigam
Indian Institute of Technology Roorkee, Roorkee, India

system precisely consists of two units: Target acquisition and localization unit and servo system to steer line-of-sight. Target motion trajectory and necessary pointing command is calculated by video tracker from the target image. Camera mounted on gimbal platform is steered in closed-loop configuration to point toward the target.

Detection and Visual tracking of moving objects has become one of the extremely important research areas in Electro-Optical surveillance system. It is a multidisciplinary work and involves image processing, computer vision, machine learning, control system, and development of gimbal. Major components of visual tracking have been depicted in Fig. 1.

Target acquisition is the first step for tracking. Here, target is acquired either manually by operator by moving camera/gimbal with joystick or by locking of potential targets with the help of auto detection.

There are various approaches for target localization/target tracking [1, 2] and the approaches depend on various factors such as image quality, size, shape, appearance [3, 4] of objects and the application. Typically, tracking algorithms [1, 2] have evolved and grouped into generative tracking and discriminative tracking [2]. Tracking problem is formulated as a binary classification problem in discriminative tracking methods. Discriminative trackers [5, 6] locate the object region by finding the best way to separate object from background. Generative tracking method is based on the appearance model of target object. Tracking is done via searching target location with best matching score by some metric. Current state of the art motion estimation methods involve target appearance learning [5], target detection along-with basic tracking algorithms. Recently more focus is on discriminative tracking instead of generative tracking.

Tracking of moving object in real scenario is a very challenging task. Most difficult task is to cater for clutter [7], where nearby objects are having similar shape or/and appearance as the target. Also, difficulties arise due to variations of appearances and occlusions. To take care of occlusion [8] of target, continuous prediction of its state is required. Linear prediction is the simplest method of state prediction. Statistical methods are also used for complex scenarios.

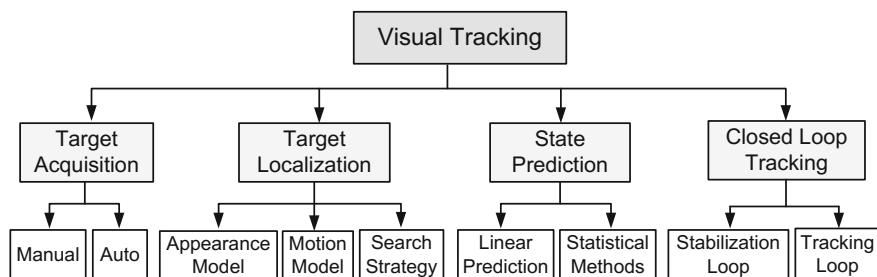


Fig. 1 Visual tracking

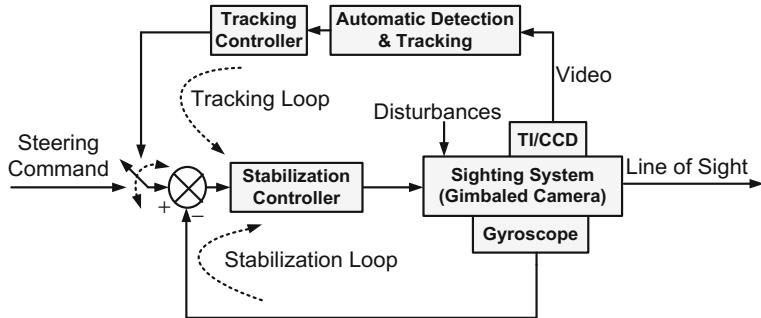


Fig. 2 Closed-loop tracking system

Closed-loop tracking system with its stabilization and tracking loop with essential components has been shown in Fig. 2. Line-Of-Sight (LOS) stabilization of the camera/sighting system facilitates acquisition of stabilized video. It limits the amount of the image motion in the field of view of the sensor during a frame, i.e., sensor integration period. The basic principle of inertial stabilization of LOS rate is to sense the disturbance torque and apply opposite torque to nullify it. This is accomplished by sensing the gimbals rotation in the inertial space with a gyroscope and converting the gyro signal into a torque through an appropriate compensating algorithm and power amplifier using a torquer as an actuator. Track loop controller is used to modify tracking error generated by target localization module and it generates appropriate steering command for gimbal such that tracking error will be minimized and target will be in the center of the FOV of the camera.

The ultimate objective is to develop robust and efficient algorithms for quick command following real-time target detection with dynamic background and tracking.

2 Performance Specifications of Target Tracking System

Starting from operational scenario to camera, tracking algorithm, processing power to control system, there are many important considerations which determine the performance specifications of target tracking system. Table 1 summarizes the design parameters, their effect on tracking performance and corresponding required design focus.

Table 1 Design consideration of tracking system

Parameters	Effect on tracking performance	Design focus
Real-time tracking error estimation	Consistency with current state of target	Development of fast algorithms
	Lag in closed-loop tracking	Real-time implementation
Target size	Different tracking algorithms have capability to handle different minimum/maximum target size	Tracking algorithm and camera to be selected in conjunction
Minimum target contrast	Difficult to track very low contrast target	Suitable tracking algorithm and camera to be selected in conjunction
		Robust feature to be extracted
Tracking rate (Pixels/frame)	Decides maximum trackable speed of target	Bigger search/matching region
		Better closed-loop tracking
Gate size	Helps to eliminate clutter (background)	Manual mode of gate size for robust tracking under heavy clutter for low variation of target size
		Auto gate size is used for unpredictable target size
Tracking algorithms	Decides tracking accuracy and duration	Select algorithms depending upon the operational scenario
	Robustness against adverse conditions	
Camera	Video decides target acquisition and performance of tracking algorithms	Select suitable camera in conjunction with algorithms, considering required size and contrast of target
Control system	Decides tracking accuracy	Nested loop: stabilization loop inside track loop
	Smooth target acquisition (No appreciable transient motion)	High bandwidth and gain, proper damping
		FOV compensation to facilitate tracking with zoom

3 Development of Proposed System

In this paper, focus is to outline the scheme for quick target acquisition of moving target and its fine tracking. For quick acquisition of target, gimbal has to point toward cued target in a short time. After gimbal moves toward target, target will be in the field of view (FOV) of the camera. But it is not necessary that target will be in the center of FOV for its easy manual acquisition. Also, since target will be moving, camera has to be moved accordingly to keep the target in the FOV. Along with the moving target, background will also be dynamic. So, detection algorithm has been developed for detection of moving target with dynamic background. For quick movement of gimbal/camera toward cued target, a scheme has been proposed.

Quick movement of gimbal with high performance control system for stabilization and tracking has been implemented and tested experimentally.

3.1 Scheme for Quick Gimbal Movement

For quick acquisition of target, gimbal motion has been controlled intelligently. Gimbal has a panoramic configuration. Command has been generated such that gimbal will follow shortest path, so that it will take less time to reach the desired position. Since the experimental tracking system can have $n \times 360^\circ$ motion, following the concept of shortest path maximum difference ($\Delta\theta$) in the cued position (Cmd) and current position of the gimbal can be 180° . Hence, if $\Delta\theta$ is greater than 180° , it is modified as follows to make it within $\pm 180^\circ$.

```

if abs ( $\Delta\theta$ ) > 180°
  if Cmd > 0°
    Cmd=Cmd-360°
  else
    Cmd= Cmd+360°
  end if
end if

```

In this way, appropriate direction and command is decided such that gimbal will follow the shortest path to reach cued position. Figure 3 details concept of shortest path. Since, stabilization loop has higher bandwidth than position loop, depending on the error, proper switching of the controller/system mode has been implemented. When error is high, stabilization loop is active with scaled position error as

Fig. 3 Shortest path following

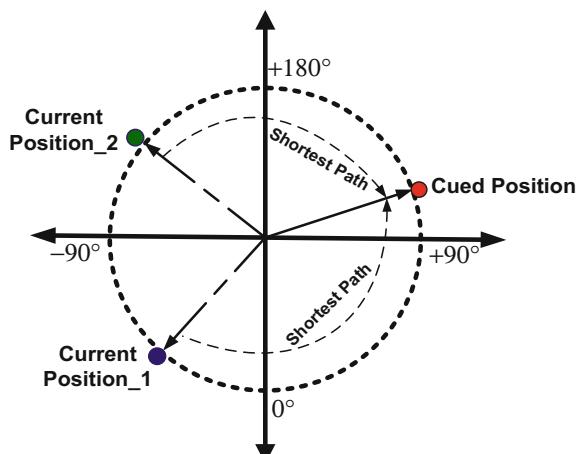


Fig. 4 System mode and command switching

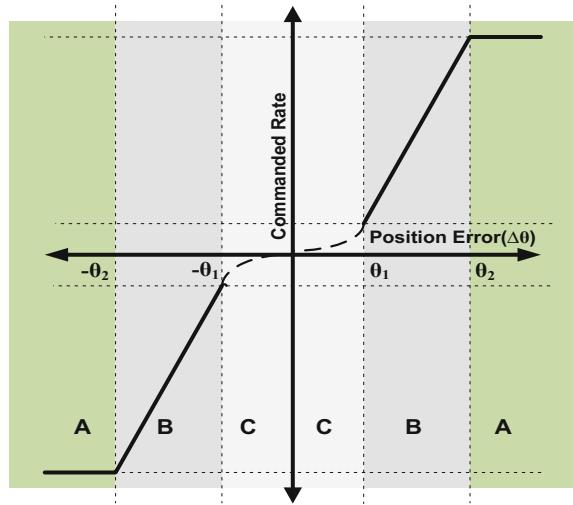


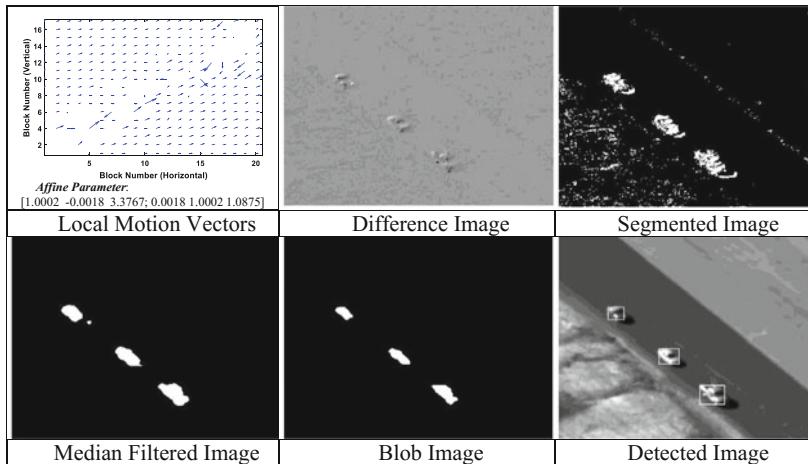
Table 2 System mode and command switching

Operational region	System mode	Command
A: ($\text{abs}(\Delta\theta) \geq \theta_2$)	Stabilization	Maximum possible rate
B: ($\theta_1 < \text{abs}(\Delta\theta) \leq \theta_2$)	Stabilization	Proportional to position error
C: ($-\theta_1 \leq \Delta\theta \leq \theta_1$)	Position	Position error

command. When error is within a very small window, position loop is activated. This switching has helped in very fast position command following with very good transient behavior. Figure 4 depicts scheme for mode switching and Table 2 enlists system mode and command switching for different operational region ($\Delta\theta$).

3.2 Detection of Moving Target with Dynamic Background

Background learning and frame difference algorithms [9–11] are commonly used for automatic target detection (ATD). There are supervised methods [10] for target detection, but they need prior information of the target. Algorithms have been also proposed for small target detection [12]. Gupta and Jain [11] have presented adaptive method for moving target detection in dynamic background. Our proposed algorithm for ATD is based on frame difference method. Here, since training is not needed, prior knowledge about target is not needed. This is a very important requirement for practical scenarios. Following flowchart summarizes the detection procedure.

Table 3 Images at different stages of detection for video *LineOfCars*, Frame # 17

Before taking frame difference, motion between detection windows of consecutive frames has been compensated. Block matching has been used for motion estimation. From the local motion vectors, global motion vector has been estimated using affine motion model by successive outlier rejection [13]. This method is computationally effective and ATD can be achieved with complex motions. Normalization of compensated difference image and logic of windowing based on mean of current difference image and of moving mean of median-filtered previous means help in robust ATD. Dynamic threshold for segmentation has been calculated as

$$\text{Threshold} = \mu \pm 3.5\sigma \quad (1)$$

where μ is the mean and σ is the standard deviation of the normalized difference image. Clutter in the segmented image has been suppressed by median filter [9] and blob has been obtained using morphological filter (Dilation followed by Erosion). Finally all connected components have been declared as detected object. Table 3 shows different stages of proposed ATD algorithm.

If target speed is high, detected object is fragmented. Also, corresponding blobs are larger than the objects. To address these two issues, an approach based on region growing and fragment overlap has been proposed. Blob regions obtained using algorithm of Fig. 5 has been used to find seeds for region growing. Histogram of blob region intensity has been obtained and the intensity having highest peak in the histogram has been selected as the seed. Criteria to grow region has been obtained as intensity limits (lower and higher) almost symmetrical around the seed intensity. Intensities between lower and higher intensity limits cover around 35 % of pixels of the blob. Boundaries of grown regions are boundaries of detected objects using region growing method. Detection boundary obtained through region

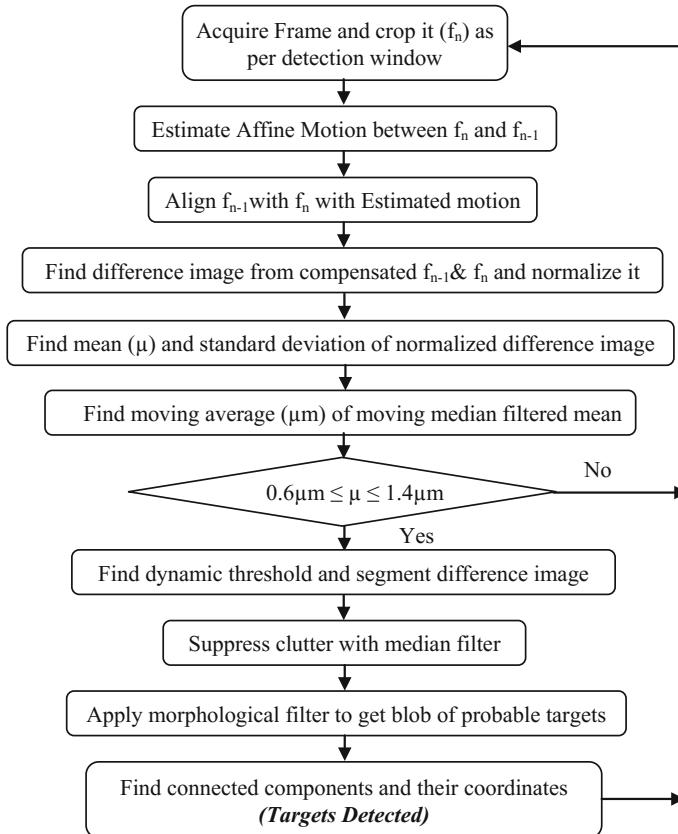


Fig. 5 Flowchart of target detection

growing exactly fits the object size as shown in Table 4 for white car. Now, grown regions, within five pixels neighborhood are associated with the same object. In this way, different fragments of the same object are associated together. Table 4 shows results of region growing and fragment association.

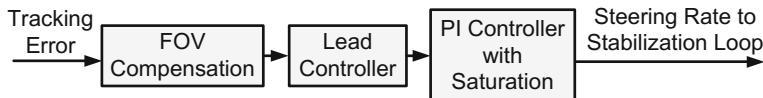
After targets have been detected, a potential target can be acquired. The acquired target can be tracked with either by maintaining and predicting the state of the target from its detection or with a suitable tracking algorithm.

3.3 Closed-Loop Tracking

A high performance servo control system for two axis experimental gimbal has been developed to achieve fine stabilization and tracking. Inertial rate sensing of the stabilization platform is done by a two axis Gyro. A stabilization loop is nested

Table 4 Detection with region growing and association for video *CarChaseWhite*

Frame#52			
Frame#53			
Algo	Flowchart, Fig. 5	Region Growing	Fragment Association

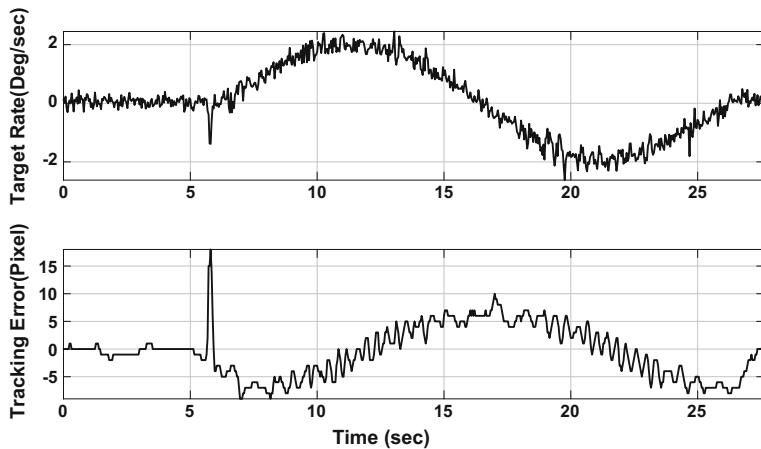
**Fig. 6** Tracking loop controller

within track loop for disturbance isolation and proper damping. Stabilization and tracking loop controllers are basically a combination of Proportional-Integral (PI) controller and a Lead controller with anti-windup, proper scaling, and saturation of intermediate and final signals. Tracker error generated by target localization module has been passed through track loop controller after FOV compensation. Figure 6 shows configuration of track loop controller.

PI controller output has been saturated at 9°/s. This limit covers maximum expected tracking rate of ground targets and also facilitates quick sign reversal of PI output. Different interlocks have been implemented for proper behavior of closed loop (gimbal motion) during target acquisition and target locking and unlocking. Before target locking, joystick or cue command is continuously assigned to the final value of track loop controller to avoid jerk at the time of switching of rate command to the tracking loop output. Also, track loop controller output is immediately replaced by cue signal when tracker stops tracking the target. Suitably modified tracking error (with PI Controller) with respect to aiming point given by target location has been given as rate command to stabilization loop. With this rate command, camera will move and target will always be around aiming point.

Table 5 Performance: stabilization accuracy

Platform	Disturbances		Stabilization accuracy ($\mu\text{rad}(1\sigma)$)
	Amplitude (Degree)	Frequency (Hz)	
Armored fighting vehicle	2.74	1.0	17.99
	0.30	3.0	21.29
Naval ship	25.00	0.1	10.72
Helicopter	2.38	2.0	27.82
	0.318	5.0	42.22
Aerostat	1.00	0.5	8.37

**Fig. 7** Target rate and tracking error (1 Pixel = 22 μ radian)

4 Performance Evaluation

After successful development of control system of tracking system, its performance was evaluated and result has been given in Table 5.

Also, simulated control system model takes 1.2 s for 40° command with position loop. For same command it takes just 0.5 s with proposed algorithm of system mode and command switching. Auto detection module has been implemented in MATLAB and its result has been given in Tables 3 and 4. To check the closed-loop performance of integrated system, a synthetic video was projected and target was locked using joystick. Performance of the control system has been given in Fig. 7.

5 Conclusion and Discussion

A scheme for quick acquisition of target, its tracking, and design considerations of tracking system has been presented. Experimental results show that the proposed tracking system facilitates quick acquisition of target. Automatic target detection is robust to adverse scenarios. It has comprehensively covered development of auto detection. Due to FOV Compensation, this system can track target in variable/different FOV of the camera. Camera can be zoomed in/out to get suitable size of the target located at different distances.

References

1. O. Javed, and M. Shah A. Yilmaz, "Object tracking: A Survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.
2. Dung M. Chu, Simone Calderara, Afshin Dehghan Arnold W. M. Smeulders, "Visual Tracking: an Experimental Survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2013.
3. A. Cavallaro, and L. di Stefano S. Salti, "Adaptive appearance modeling for video tracking: survey and evaluation," *TIP*, 2012.
4. Y. Su, X. Li, and D. Tao X. Gao, "A review of active appearance models," *IEEE Trans Sys Man Cybernetic*, vol. 40, no. 2, pp. 145–158, 2010.
5. Krystian Mikolajczyk, and Jiri Matas Zdenek Kalal, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, July 2012.
6. L Zhang, MH Yang, Q Hu K Zhang, "Robust object tracking via active feature selection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.
7. Andrea Cavallaro Maggio, *Video Tracking: Theory and Practice*.: A John Wiley and Sons, Ltd., 2011.
8. Merwin Amala Roger.J C. Sathish, "Efficient Occlusion Handling Object Tracking System," *International Journal of Innovative Science, Engineering & Technology*, vol. 1, no. 1, March 2014.
9. N. S. Narkhede, Saurabh S. Athalye Mahesh C. Pawaskar, "Detection Of Moving Object Based On Background Subtraction," *International Journal of Emerging Trends & Technology in Computer Science*, vol. 3, no. 3, pp. 215–218, May-June 2014.
10. Ronen Talmon, Israel Cohen Gal Mishne, "Graph-Based Supervised Automatic Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2738–2754, May 2015.
11. Sanket Gupta Yogendra Kumar Jain, "A Highly Adaptive Method for Moving Target Detection in Dynamic Background with a Simplified Manner," *International Journal of Computer Applications*, vol. 102, no. 10, pp. 20–26, September 2014.
12. Yong Ma, Bo Zhou, Fan Fan, Kun Liang, and Yu Fang Jinhui Han, "A Robust Infrared Small Target Detection Algorithm Based on Human Visual System," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 12, pp. 2168–2172, December 2014.
13. Ming-Ting Sun, Vincent Hsu Yeping Su, "Global Motion Estimation from Coarsely Sampled Motion Vector Field and the Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 2, February 2005.

Low-Complexity Nonrigid Image Registration Using Feature-Based Diffeomorphic Log-Demons

Md. Azim Ullah and S.M. Mahbubur Rahman

Abstract Traditional hybrid-type nonrigid registration algorithm uses affine transformation or class-specific distortion parameters for global matching assuming linear-type deformations in images. In order to consider generalized and nonlinear-type deformations, this paper presents an approach of feature-based global matching algorithm prior to certain local matching. In particular, the control points in images are identified globally by the well-known robust features such as the SIFT, SURF, or ASIFT and interpolation is carried out by a low-complexity orthogonal polynomial transformation. The local matching is performed using the diffeomorphic Demons, which is a well-established intensity-based registration method. Experiments are carried out on synthetic distortions such as spherical, barrel, and pin-cushion in commonly referred images as well as on real-life distortions in medical images. Results reveal that proposed introduction of feature-based global matching significantly improves registration performance in terms of residual errors, computational complexity, and visual quality as compared to the existing methods including the log-Demons itself.

Keywords Diffeomorphic Demons · Feature-based registration · Nonrigid image registration

M.A. Ullah (✉) · S.M.M. Rahman

Department of Electrical and Electronic Engineering,

Bangladesh University of Engineering and Technology, 08193 Dhaka, Bangladesh

e-mail: aungkonazim@gmail.com

URL: <http://www.buet.ac.bd/eee/>

S.M.M. Rahman

e-mail: mahbubur@eee.buet.ac.bd

1 Introduction

Image registration refers to establishment of spatial correspondences among the multitemporal, multiview, multisensor, or multimodal source images. Applications of image registration include the atlas construction by geometrically aligning multiple number of source images those are obtained in the remote sensing, medical imaging, navigation, security, and industrial production [1]. In a broad sense, there exists two types of image registration algorithms, viz., rigid and nonrigid. In a rigid-type image registration, the geometric distortions of images are considered to be spatially invariant. The rigid-type registration is usually tackled by conventional affine, projective, or piecewise linear transformations, which require a finite set of parameters. In many cases, however, there exists nonlinear and free-form local deformations (FFLDs) in images, which is addressed by the nonrigid registration techniques [2].

Nonrigid image registration can be broadly classified into two types, viz., feature-based and intensity-based. The feature-based also known as the sparse registration methods involve establishing spatial correspondences in terms of surface landmarks, edges, or ridges. Energy of directional derivatives, local frequency, or phases are used to find the key points of spatial correspondences. These points are used for matching the images using certain cost functions defined in terms of deterministic or probabilistic distances such as the mutual information, cross cumulative residual entropy, and cross-correlation [3]. The feature-based algorithms are computationally efficient, but the robustness of performance of such methods is limited when the imaging systems of source images have significant deviations in the field-of-view [4].

The intensity-based monomodal image registration also known as the dense matching algorithms usually assume brightness constancy restriction such as that in the conventional template matching technique. The multimodal registration considers information theoretic cost functions which are adaptive to intensity corrections or independent of any intensity-based restriction. Optical flow estimation forms a large class of algorithms to solve the problem of nonrigid type image registration using the intensity-based approach. Instead of using the image intensities, the pixel-wise densely sampled scale-invariant feature transform (SIFT) features of images have been used to establish spatial correspondences at the expense of a large complexity [5]. The mutual information-based matching along with the B-Spline-based interpolation has been shown to be effective for tackling nonrigid deformations those occurred in the vicinity of a regular grid [6].

The well-established nonparametric and iterative techniques those are very successful in mechanics such as elastics and fluid dynamics are often adopted for intensity-based nonrigid registration [7]. For example, after the inception of Thirion's Demons algorithm [2], which is primarily based on modeling the image matching in terms of diffusion process, several image registration techniques have been developed for tackling dense deformation field by adopting different regularization approaches. The challenges of classical Demons algorithm include the prediction of image gradient, which often introduces folding of shapes due to local minima in optimization. The introduction of diffeomorphism that preserves the topology of

objects while calculating the gradient iteratively has improved the performance of the Demons algorithms noticeably. Although the diffeomorphic Demons are computationally efficient, the performance of such algorithm for a relatively large deformations in images is not satisfactory. Several approaches have been tried to improve the performance of the Demons algorithm by direct matching of features using the log-spectrum, multiscale decomposition, or samples of intensities when source images have a relatively large deformations [8].

In order to overcome the problems of feature-based and intensity-based image registration algorithms the ‘hybrid’ approach has also been adopted. In this approach, a suitable feature-based approach is applied globally on the source images as a pre-registration technique, and then a finely tuned intensity-based image registration is considered for matching the local deformations. For example, the globally optimized affine-based functional magnetic resonance imaging of brain’s (FMRIB’s) linear image registration tool, which is known as the FLIRT, has been applied along with the Demons registration algorithm for stripping the brain tissues [9]. The software package Elastix, a configurable set of parametric intensity-based image registration, has been used prior to adaptively regularized Demons algorithm for registering synthetic as well as real abdominal computed tomography (CT) images [10]. The non-rigid registration is performed via an integrated squared-error estimator that minimizes a cost function defined in terms of certain regularized SIFT-flow-based sparse feature matching and velocity-based dense matching in [11]. The preregistration of source images in terms of scale, translation, and rotations has been implemented with the use of the convolution neural network-based classification algorithm, which follows the diffeomorphic Demons algorithm for the ultimate registration [12].

In this paper, we argue that instead of using available software packages or certain affine transformations those are specific for a defined class of deformations, a generalized feature-based low-complexity preregistration can follow the well-known diffeomorphic Demons algorithm for developing an efficient nonrigid registration algorithm. In particular, the spatial correspondences of fiducial points in the source images selected by shape-invariant local descriptors such as the speeded up robust features (SURF) [13], SIFT [14], Affine-SIFT (ASIFT) [15] are used for feature matching. The geometric transformation of these fiducial points in the preregistration are performed by the use of orthogonal polynomial transformation. Experiments are conducted on the source images having locally nonlinear synthetic distortions such as the spherical, barrel, pin-cushion as well as images that undergo real-life nonrigid deformations.

The paper is organized as follows. In Sect. 2, the proposed feature-based registration algorithm that uses the diffeomorphic Demons is presented. Section 3 presents the experimental setup, results, and the performance comparisons of the proposed method with the existing registration methods. Finally, the conclusions are drawn in Sect. 4.

2 Proposed Registration Algorithm

Let $I^r(x, y)$ and $I_m(x, y)$ be the reference and moving images to be registered, where (x, y) represents spatial coordinates in two-dimension. The proposed two-layer registration algorithm requires a feature-based global transformation T_g that follows an intensity-based local transformation T_l . If the cost function of similarity of aligned images is denoted as $C(\cdot)$, then the objective of the registration algorithm is to find an optimal set of transformations T^* ($T^* \in T_g^*, T_l^*$) such that

$$T^* = \arg \min_{T_g^*, T_l^* \in S_T} C(I^r, (I^s \circ T_g) \circ T_l) \quad (1)$$

where S_T is the allowable space of transformation and $I^s \circ T$ represents the moving image after it has been transformed by T . Commonly referred cost functions of similarity include the sum of squared distance, correlation coefficient, mutual information, and cross entropy. As practiced in the hybrid-type registration algorithms, the global and local transformations T_g and T_l are considered to be independent to each other. In order to improve the readability of the paper, these two transformations of the proposed algorithm are presented in separate subsections.

2.1 Global Transformation

In order to align the moving image globally, first a set of key or control points is to be identified. The Hessian detector is chosen for selecting the control points in the images, due to the fact that successful robust features such as SIFT, SURF, and ASIFT adopt this detector. Repeated implementation of difference of Gaussian filter and down-sampling the resultant images are usually practiced in localizing the key points. For the purpose of registration, the features of these key points are estimated, especially to establish the spatial correspondences of the reference and moving images. Hence, the robustness and computational complexity of the features play significant role in finding the overall correspondence of two images. For example, the SIFT features, which are calculated from the oriented gradients of the key points, are invariant to scale, rotation, and translation of the patterns in images [14]. The computation complexity SURF features are improved using the Haar wavelet filters for the key points [13]. The ASIFT features, which are invariant to affine transformation of a pattern in images, introduces a new parameter ‘transition tilt’ to accommodate the distortion of appearance of an object in images due to variations of camera angle both in the longitude and latitude directions [15]. The correspondences of the control points in two source images are established by including the inliers and excluding the outliers. The inliers and outliers are chosen by calculating the nearest neighbor of the features. In the proposed method, we do not set any restriction on the choice of the robust features to establish the correspondences on the control points, say N



Fig. 1 An example of set of spatial correspondences obtained from moving image *Lena* that undergoes a pin-cushion-type deformation

number of points, on the images to be aligned. Through extensive experimentation, however, it is found that the correspondences found from the ASIFT features are significantly better than existing SIFT or SURF features. Figure 1 shows an example of set of spatial correspondences found by the ASIFT features, when the moving image *Lena* undergoes pin-cushion-type synthetic deformation. From this figure, it can be seen that there are sufficient number of control points ($N = 1423$ in actual) for which the spatial correspondences of the patterns exist. It is to be noted that the control points for the SURF and SIFT features are 234 and 367, respectively.

Since the geometric distortions are very often nonlinear, instead of the affine transformation, we have chosen the polynomial-type transformation for the alignment of the images globally. Let $\Phi_m(x, y)$ ($m \in 0, 1, 2, \dots, M$) represent a generalized set of polynomial function of order m and I_i^s ($i \in 1, 2, \dots, N$) ($N \gg M$) represent the control points on the moving image that are used for global transformation. The aligned image is obtained from the global transformation as

$$I_g^a(x, y) = \sum_{m=0}^M a_m \Phi_m(x, y) \quad (2)$$

where the parameters a_m can be estimated in the least-square sense, and solving the set of linear equations given by [16]

$$\sum_{m=0}^M a_m \left[\sum_{i=1}^N \Phi_m(x_i, y_i) \Phi_m(x_i, y_i) \right] = \sum_{i=1}^N I_i^s \Phi_m(x_i, y_i) \quad m = 0, 1, 2, \dots, M \quad (3)$$

If the polynomials are orthogonal, the parameters of the transformation are

$$a_m = \frac{\sum_{i=1}^N I_i^s \Phi_m(x_i, y_i)}{\sum_{i=1}^N [\Phi_m(x_i, y_i)]^2} \quad (4)$$

In practice, a generalized set of orthogonal polynomials can be determined by the Gram–Schmidt orthogonalization process and using a set of linearly independent

functions. Examples of a few linearly independent functions include [16]

$$\begin{aligned} h_0(x, y) &= 1 \\ h_1(x, y) &= x, \quad h_2(x, y) = y \\ h_3(x, y) &= x^2, \quad h_4(x, y) = xy, \quad h_5(x, y) = y^2 \end{aligned} \quad (5)$$

that can be used to generate a set of orthogonal polynomials, and hence, to obtain a good approximation of I_g^a .

2.2 Local Transformation

In the proposed method, the globally aligned image I_g^a undergoes the local transformation using the log-Demons algorithm [8]. In log-Demons, the focus is given to estimate the ultimate velocity field v through the increment of velocity field δ^v in each iteration. In order to calculate the updates, a diffeomorphic transformation $\phi = e^v$ of the velocity field is obtained. The forward mapping of the increment field is given by [8]

$$\delta_{f \rightarrow s}^v = -\frac{I^r - I_{g \circ \phi}^a}{\|\nabla I_{g \circ \phi}^a\|^2 + \alpha_f^2 |I^r - I_{g \circ \phi}^a|^2} \nabla I_{g \circ \phi}^a \quad (6)$$

where $I_{g \circ \phi}^a$ represents the matched image via the map ϕ , α_f^2 ($0 < \alpha_f \leq 1$) is a weight parameter, and ∇ is the gradient operator. In a similar fashion, the backward mapping of the increment field $\delta_{s \rightarrow f}^v$ is obtained. Finally, the updated velocity field for the j th step of an iteration is given by [8]

$$v_j = K_1 \otimes \left[v_{j-1} + \frac{1}{2} K_2 \otimes \left(\delta_{f \rightarrow s}^v - \delta_{s \rightarrow f}^v \right) \right] \quad (7)$$

where K_1 and K_2 are smoothing Gaussian kernels having standard deviations σ_1 and σ_2 , respectively, and \otimes is a convolution operator.

3 Experimental Results

Experimentations have been carried out on a number of commonly referred test images in order to evaluate the performance of the proposed registration method as compared to the existing methods. Due to the space limitations, however, the results in this section are given that are obtained from popular grayscale images such as *Cameraman*, *Lena*, *MRI* having pixel sizes 512×512 , 256×256 , and 128×128 , respectively. Three types of nonlinear deformations, namely, spherical, barrel, and

pin-cushion are considered in the experiments with *Cameraman*, *Lena* for which original version of the images are available. In cases of spherical- and barrel-type distortions the origin of deformation lies in the center of the images, whereas that lies in the center of first quadrant of image for the pin-cushion-type distortion. The polar coordinates that undergo the spherical deformation include those within half the maximum radius of image from the center. In the experiments, the cubic deformation parameters of the barrel and pin-cushion-type distortions were chosen as 10^{-3} and -10^{-4} , respectively, so that visual contents of the images remain within an acceptable quality. In addition to the synthetic deformations, prototype distortions on medical images such as the *MRI* images that have been used by other research groups [17] are also included in the experimentations. Four registration algorithms, namely, the B-Spline [6], Demons [8], FLIRT + Demons [9] and the proposed algorithm have been considered in the experiments. The performance of the proposed algorithm has been provided as three independent methods, namely, SIFT + Demons, SURF + Demons, and ASIFT + Demons.

Table 1 shows the root mean-square (RMS) error estimated from the registered images obtained from the methods when the moving images are deformed using three-type synthetic distortions considered in the experiments. It is seen from this table that in general the proposed methods show lower values of RMS error as compared to the existing methods. In particular, the proposed ASIFT + Demons method, which uses ASIFT features as the control points for global transformation, provides the lowest RMS error as compared to the others. This is expected due to the fact that the features provided by the ASIFT are invariant to the affine transformation, and provides the best approximation of the globally aligned images. It is to be also noted from the results of Table 1 that when the image size is relatively small, such as the case of *Lena* as compared to that of the *Cameraman*, in terms of RMS error the SIFT + Demons method performs very close with the ASIFT + Demons method.

Table 1 Results concerning the RMS error provided by the methods considered in the experiments of synthetic deformations

Experimental nonlinear deformations	Comparing methods			Proposed methods		
	B-Spline [6]	Demons [8]	FLIRT + Demons [9]	SURF + Demons	SIFT + Demons	ASIFT + Demons
<i>Lena</i>						
Spherical	50.88	18.15	18.15	11.47	9.28	8.65
Barrel	30.88	55.88	49.41	21.95	14.64	12.41
Pin-Cushion	67.35	66.39	66.39	26.74	26.84	25.11
<i>Cameraman</i>						
Spherical	92.91	40.55	40.55	37.50	32.04	24.28
Barrel	40.50	68.19	60.90	29.48	29.62	23.97
Pin-Cushion	60.45	64.66	64.66	32.36	33.97	27.42

Thus, the SIFT + Demons method can be chosen for registration of a small size image, but the ASIFT + Demons method is recommended for a large size image.

Figure 2 shows the visual outputs of the reference and moving images of *Cameraman* which is synthetically distorted by spherical-type transformation, and the corresponding registered images by the Demons [8] and proposed ASIFT + Demons methods. From this figure, it is seen that the proposed method can align a moving image very close to the reference image better than the Demons algorithm. Figure 3 shows the visual outputs of the reference and deformed images *MRI* used in [17], and the images showing the squared of errors of registered images that are obtained from the Demons [8] and proposed ASIFT + Demons methods. From this figure, it is clearly seen that the error of registration can be significantly reduced by introducing the proposed ASIFT-based global transformation prior to the Demons-based local transformation. Figure 4 shows the variations of RMS error in first few iterations of the Demons and the proposed algorithm that uses the SIFT, SURF, and ASIFT for global transformations. It is seen from this figure that the error of the proposed method converges faster than the Demons algorithm. In addition, the proposed ASIFT + Demons algorithm provides the lowest RMS error among the methods compared. These results are also consistent with the lowest errors shown in Table 1, due to the fact that ASIFT performs better than SIFT or SURF as features to determine the control points for global transformation.

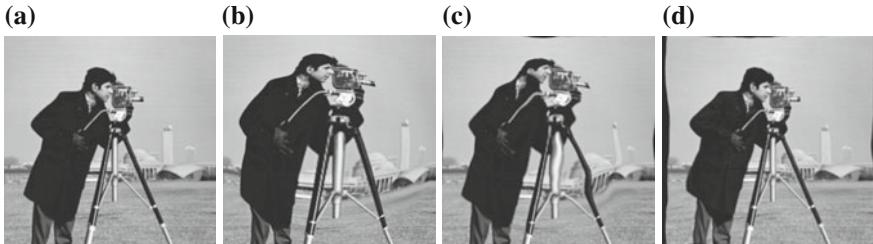


Fig. 2 Comparison of visual outputs of registration on image *Cameraman*. **a** Reference image, **b** moving image distorted by spherical deformation. The registered images are obtained from **c** Demons [9] and **d** proposed ASIFT + Demons, respectively

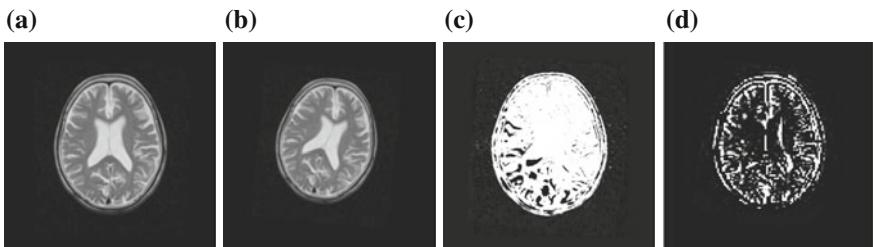
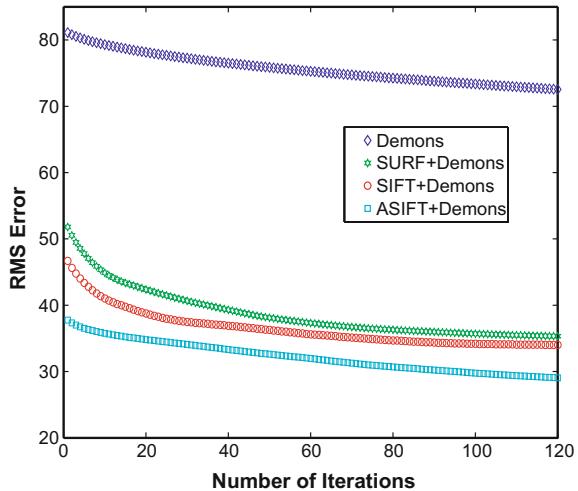


Fig. 3 Comparison of errors of registration on image *MRI*. **a** Reference image, **b** real-life moving image used in [17]. The images depicting squared errors of registered images that are obtained from **c** Demons [9] and **d** proposed ASIFT + Demons, respectively

Fig. 4 Variations of the RMS errors when feature-based global transformations are applied prior to the Demons-based local transformation



4 Conclusion

In this paper, an hybrid-type registration algorithm has been presented to geometrically align a pair of images that are deformed by nonlinear or nonrigid distortions. Instead of using the traditional linear-type affine-registration or supervised and class-specific registration parameters for the global transformation, the proposed method has adopted an approach of robust and feature-based registration, which fits very well for generalized and nonlinear-type distortions in images. In particular, the control points have been identified in the pair of images using the well-known robust features such as the SIFT, SURF, and ASIFT. The global alignment of these key points has been performed using the orthogonal polynomial transformation. The diffeomorphic Demons algorithm, which is well-known for registering images having non-grid-type distortions, has been adopted for local matching of globally aligned images. Experimentations have been carried out on commonly referred test images that undergo both the synthetic-type distortions such as spherical, barrel, and pin-cushion as well as real-life distortions. Results show that the proposed introduction of feature-based global matching and Demons-based local matching significantly improve the registration performance in terms of RMS error and visual quality as compared to the existing methods. It has been also shown that due to the introduction of proposed approach, the computational complexity of Demons algorithm is significantly reduced and at the same time generalized-type of large deformations can be tackled very well by this algorithm.

References

1. Alam, M.M., Howlader, T., Rahman, S.M.M.: Entropy-based image registration method using the curvelet transform. *Signal, Image and Video Processing* 8(3), 491–505 (2014)
2. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwells demons. *Medical Image Analysis* 2(3), 243–260 (1998)
3. Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: LCC-Demons: A robust and accurate symmetric diffeomorphic registration algorithm. *NeuroImage* 81(Nov), 470–483 (2013)
4. Wang, F., Vemuri, B.C.: Non-rigid multi-modal image registration using cross-cumulative residual entropy. *Int. J. Computer Vision* 74(2), 201–215 (2007)
5. Liu, C., Yuen, J., Torralba, A.: SIFT flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Analysis and Machine Intelligence* 33(5), 978–994 (2011)
6. Myronenko, A., Song, X.: Intensity-based image registration by minimizing residual complexity. *IEEE Trans. Medical Imaging* 29(11), 1882–1891 (2010)
7. Huang, X.: Nonrigid image registration problem using fluid dynamics and mutual information. *Biometrics & Biostatistics* S12(004), 1–11 (2015)
8. Lombaert, H., Grady, L., Pennec, X., Ayache, N., Cheriet, F.: Spectral log-Demons: Diffeomorphic image registration with very large deformations. *Int. J. Comput. Vision* 107(3), 254–271 (2014)
9. Wang, Y., Nie, J., Yap, P.T., Shi, F., Guo, L., Shen, D.: Robust deformable-surface-based skull-stripping for large-scale studies. In: *Lecture Notes in Computer Science: Int. Conf. on Medical Image Computing and Computer-Assisted Intervention.* vol. 6893, pp. 635–642. Toronto, Canada (2011)
10. Freiman, M., Voss, S.D., Warfield, S.K.: Demons registration with local affine adaptive regularization: Application to registration of abdominal structures. In: *IEEE Int. Symp. Biomedical Imaging: From Nano to Macro.* pp. 1219–1222. Chicago, IL (2011)
11. Ma, J., Qiu, W., Zhao, J., Ma, Y., Yuille, A.L., Tu, Z.: Robust $L_2 E$ estimation of transformation for non-rigid registration. *IEEE Trans. Signal Processing* 63(5), 1115–1129 (2015)
12. Zhao, L., Jia, K.: Deep adaptive log-Demons: Diffeomorphic image registration with very large deformations. *Computational and Mathematical Methods in Medicine* 2015(836202), 1–16 (2015)
13. Bay, H., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110(3), 346–359 (2008)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision* 60(2), 91–110 (2004)
15. Morel, J.M., Yu, G.: ASIFT: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sciences* 2(2), 438–469 (2009)
16. Goshtasby, A.: Image registration by local approximation methods. *Image and Vision Computing* 6(4), 255–261 (1988)
17. Liu, X., Chen, W.: Elastic registration algorithm of medical images based on Fuzzy set. In: *Lecture Notes in Computer Science: Int. Workshop on Biomedical Image Registration.* vol. 4057, pp. 214–221. Utrecht, The Netherlands (2006)

Spotting of Keyword Directly in Run-Length Compressed Documents

Mohammed Javed, P. Nagabhushan and Bidyut Baran Chaudhuri

Abstract With the rapid growth of digital libraries, e-governance and Internet applications, huge volume of documents are being generated, communicated and archived in the compressed form to provide better storage and transfer efficiencies. In such a large repository of compressed documents, the frequently used operations like keyword searching and document retrieval have to be carried out after decompression and subsequently with the help of an OCR. Therefore developing keyword spotting technique directly in compressed documents is a potential and challenging research issue. In this backdrop, the paper presents a novel approach for searching keywords directly in run-length compressed documents without going through the stages of decompression and OCRing. The proposed method extracts simple and straightforward font size invariant features like number of run transitions and correlation of runs over the selected regions of test words, and matches with that of the user queried word. In the subsequent step, based on the matching score, the keywords are spotted in the compressed document. The idea of decompression-less and OCR-less word spotting directly in compressed documents is the major contribution of this paper. The method is experimented on a data set of compressed documents and the preliminary results obtained validate the proposed idea.

Keywords Compressed document keyword spotting • Run-length compressed domain • Correlation-entropy analysis • Compressed image processing

M. Javed · P. Nagabhushan · B.B. Chaudhuri (✉)

Department of Studies in Computer Science, University of Mysore, Mysore, India
e-mail: bbc@isical.ac.in

M. Javed
e-mail: javedsolutions@gmail.com

P. Nagabhushan
e-mail: pnagabhushan@hotmail.com

B.B. Chaudhuri
CVPR Unit, Indian Statistical Institute, Kolkata, India

1 Introduction

Since past decade, with the advancement in multimedia technology and internet, there has been tremendous growth in the volume of documents handled by digital libraries such as ACM, IEEE, and many online databases [1]. Therefore, keyword searching has become very useful and frequent task while retrieving the related documents from these huge collection of documents. However because of large volume of documents, the documents are generally compressed before archiving in different databases to provide storage and transfer efficiencies. With the existing state-of-the-art techniques, keyword spotting in compressed documents requires an additional step of decompression which warrants additional computing resources. Therefore, it would be novel and interesting to develop algorithms that can search the queried keyword directly in compressed documents without going through the stage of decompression.

The traditional way of tackling the problem of keyword spotting in uncompressed documents is with the use of OCR [5, 14]. Techniques employing OCR carry out text segmentation upto character level and convert the entire document image into machine- readable text. But due to segmentation errors and document degradation, the performance of the OCR decreases gradually. Hence an OCR-less method of spotting keywords for uncompressed documents based on image matching was introduced by Doermann [5, 14]. The OCR-less techniques can tolerate some sort of noise and degradations, and also work irrespective of the script printed in the document. However, the available keyword spotting techniques cannot be directly applied to compressed documents without being decompressed and/or OCRed. Therefore, this research paper aims at developing a word spotting model directly in compressed documents bypassing the stages of decompression and OCRing, which is the major contribution projected in this paper.

There are different document compression formats available in the literature, such as TIFF, PDF, JPEG, and JBIG. But for compressing text documents CCITT Group 3(T.4) [2] and CCITT Group 4(T.6) [3] compression schemes are popularly used and are widely supported by TIFF and PDF formats. T.4 and T.6 compression schemes are also frequently employed for compressing and archiving printed text documents emerging from fax machines and digital libraries. Run-Length Encoding (RLE) forms the backbone of these compression schemes. Therefore, this research study is specifically focused on demonstrating the keyword spotting directly on run-length compressed text documents of CCITT compression schemes.

In the literature, there are some efforts to operate directly on the compressed data of CCITT Group 3 and CCITT Group 4 compression schemes [14]. The following operations such as feature extraction [8, 10], segmentation [11], word spotting [11, 13], document similarity [7], and retrieval [6, 12] have been reported. In the context of keyword spotting, the authors of [13] and [12] propose an OCR-less word searching and document retrieval model for compressed documents. Their proposed method reads the black and white changing elements from the compressed file, builds an outer skeleton of the texts in an uncompressed mode, and then performs

word spotting using the connected components and word object features. Since the extracted skeleton of text and subsequent analysis (connected component, feature extraction) are carried out in uncompressed mode, the method is same as working on uncompressed document. On the other hand, the word spotting model proposed by [11] is directly on the compressed document data. But their word spotting model is based on the usage of an OCR. Here, the bounding box ratio is used to filter the segmented compressed test words and in the subsequent step, the first two characters of the filtered test words are decompressed and sent to OCR. In case the output of the OCR matches with the first two characters of the keyword, the remaining characters of the test words are OCRed and matched with that of the keyword. In this way, their method attempts to spot keywords by minimizing the burden of decompression and OCRing. To our best knowledge in the literature, the idea of decompression-less and OCR-less based word spotting in the compressed document data has not been attempted in document image analysis. Therefore, through this research contribution, the novel idea of word spotting directly in compressed documents using a decompression-less and OCR-less method is initiated in document image analysis.

Overall, the research work aims at developing a novel method of keyword spotting directly in run-length compressed domain with the help of simple and efficient run-length features. The proposed method bypasses the stages of decompression and OCRing as generally found in the conventional approaches. The user-queried keyword is printed as an image and the font size invariant features are extracted from its run-length compressed version. The extracted features of the input keyword are matched with that of the compressed words in the document and the matching words are spotted in the compressed document. The proposed idea is experimentally validated with a data set of compressed English printed documents and the encouraging results are reported. Rest of the paper is organized as follows: Sect. 2 gives an outline of the proposed model and further discusses the different features used for word spotting, Sect. 3 reports the experimental results and related discussions, Sect. 4 concludes the paper with a brief summary.

2 Proposed Method

The different stages involved in the proposed model of word spotting in run-length compressed document are shown in Fig. 1.

In the proposed model, the run-length compressed data is assumed to be extracted from the TIFF-compressed document by employing Modified Huffman decoding algorithm. Further the run-length compressed document is subjected to text line and word segmentation using the recent approach proposed by [11]. From the segmented compressed words also called as test words, the bounding box ratio is extracted and indexed in a database of compressed documents. On the other side, the user query is input through the keyboard, then mapped as an image and run-length compressed to extract the bounding box ratio and other proposed features for word spotting. In this way, based on the bounding box ratio of the keyword, the indexed test words from the

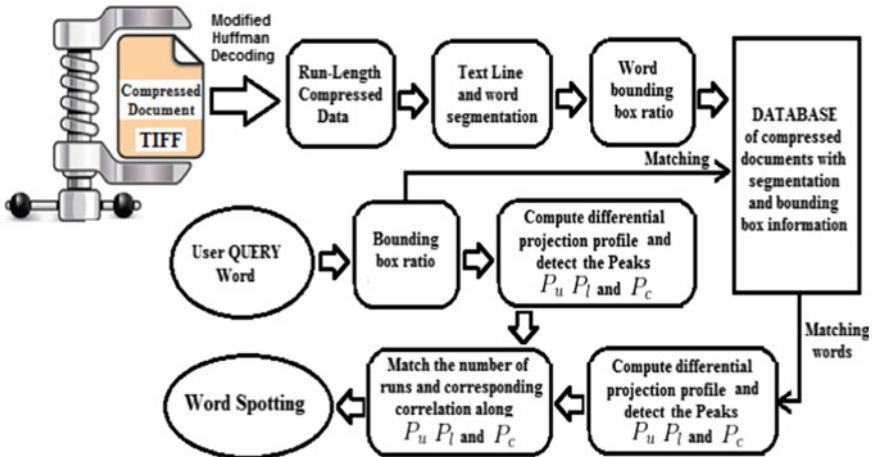


Fig. 1 Proposed compressed document word spotting model

compressed document database are filtered out. The filtered test words are subjected to feature extraction using their compressed run-length data. Finally, the extracted font size invariant features of the test words are matched with that of the keyword, and the matching words are spotted in the compressed documents. The feature extraction and feature matching strategies are illustrated in the upcoming discussions.

2.1 Feature Extraction

Efficient feature extraction is very crucial in word spotting application. The different font size invariant features extracted from the keyword and the segmented test words for matching and keyword spotting in compressed documents are discussed here.

Bounding Box Ratio: Consider the sample word image in Fig. 2 which shows the word ‘THIS’ in both uncompressed and compressed (run-length) versions. In the figure, for the purpose of distinguishing the character runs after compression, each character is represented with a unique color. The bounding box (bb) ratio for a word is the ratio of the word length to word height. For the compressed word in Fig. 2, the number of rows indicate word height (6) and sum of all the runs in any given row (25) indicates the word length. The bounding box ratio computed is 25/6. Experimentally it was observed that the bounding box ratio for any given word represented in different font size (for all the discussions in this paper font size is taken to be in the range 10–18) does not vary much and a threshold of 0.5 on either side of that of the keyword can capture most of the words appearing in such a range of font sizes in the compressed document. This is illustrated in Fig. 3.

Projection Profile Peaks: Generally a word in English language is characterized by three regions [1], where the upper zone is called ascender region, center zone

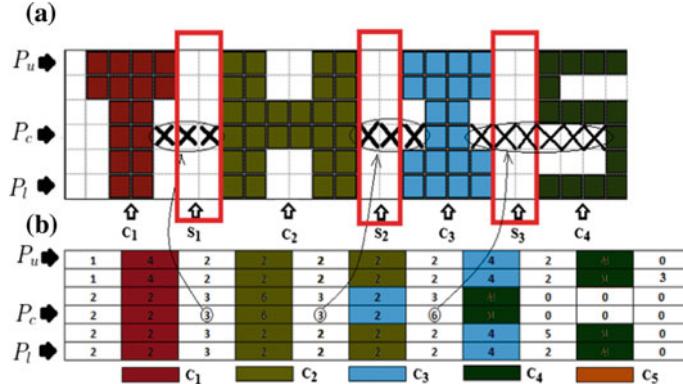


Fig. 2 Sample word in **a** uncompressed mode, **b** run-length compressed mode

Fig. 3 Bounding box ratio of a word in different font size



is base region (or x height region), and the bottom zone is descender region. For extracting the run-length-based word features, three feature locations are defined in the base region namely P_u , P_l , and P_c . The P_u is the topmost row of the base region and P_l is the lower most row of the base region, and P_c is the center of the base region. In a compressed word, P_u , P_l , and P_c can be automatically detected by computing the projection profile and subsequently with a differential projection profile as used in papers [4, 8, 9]. If a represents a word image in uncompressed version with m rows and n columns, then b represents a matrix of compressed runs arranged in m rows and n' columns (where $n' \ll n$) as shown in Fig. 2. Then the projection profile (pp) and the differential projection profile (dpp) computed are as follows,

$$pp(i) = \sum_{j=1}^{n'} b(i,j) \quad (1)$$

$$dpp(i) = pp(i+1) - pp(i), \forall i = 1 \dots m' \quad (2)$$

In the sample word shown in Fig. 2, all characters are of upper case and hence ascenders and descenders do not come into picture. Therefore the top and bottom rows are marked as P_u and P_l ; however, in words with mixed case or lower case characters the locations of P_u , P_l and P_c vary, and hence they have to be detected automatically using $dpp(i)$ as detailed in [9]. The base region peaks detected for a word of different font size (see Fig. 3) are shown in Fig. 4. From the figure, it is evident that the nature of the profile peaks remain invariant to change in font size.

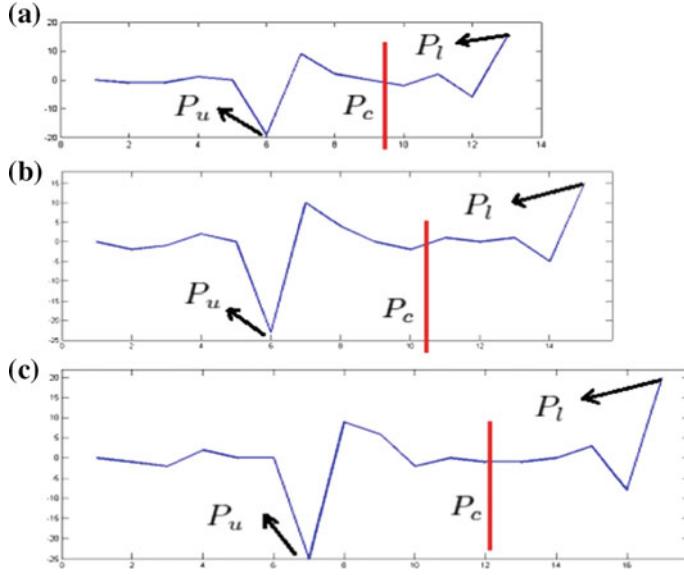


Fig. 4 Detection of P_u and P_l regions in a word with **a** Font size 12, **b** Font size 14, **c** Font size 16

Number of Run Transitions: After computing the locations of P_u , P_l , and P_c for a compressed word, the runs along the corresponding rows are scanned and the feature called number of runs or run transitions are extracted which are identified as N_u , N_l , and N_c respectively. It was observed that, in an ideal case, for any given word the number of run transitions in a particularly positioned row always remain constant irrespective of the font size of the word. However, in practicality due to binarization and presence of noise the run transitions slightly vary which can be handled by fixing a threshold. Therefore, number of run transitions is potentially an important feature for word spotting in run-length compressed documents. It was also observed that, to some extent the proposed feature at the detected base region peaks are also invariant to change in font size.

Correlation: The other important feature is correlation of runs which is computed along the rows of the detected locations P_u , P_l , and P_c of the given compressed word. The computed correlation features are respectively denoted as C_u , C_l , and C_c . The proposed feature captures the similarity of black and white runs existing between the test word and the keyword computed along the detected locations P_u , P_l , and P_c . The correlation between any two rows of compressed runs is computed using the approach proposed by the researchers in [10]. However, in this paper, the correlation feature defined is the sum of the features C_{0-0} and C_{1-1} transitions defined in [10]. Between any two rows of runs in a compressed image, the values of C_{0-0} and C_{1-1} , respectively, indicate the presence of number of 0–0 and 1–1 pixel transitions (which implies similar pixels). Further it is also important to note that, the correlation feature changes with the change in word font size. Therefore, before computing correlation

of runs across the detected peaks of the compressed words having different font size, it is necessary to normalize the runs at the detected peak locations. This is achieved by summing up the runs at the detected peak locations for both keyword and test word to determine the word with longer row value (in pixels). For the detected word, all the runs at the peak locations are proportionally mapped to the runs at the corresponding peak locations of the word having smaller row value. Then, finally after normalizing the runs, the correlation of runs is computed. This makes correlation feature suitable for different font size words.

2.2 Compressed Word Matching

In the proposed algorithm, word matching is done in two stages. First, with the help of bounding box ratio which filters out most of the test words from the compressed word database. In the second stage, word matching is performed on the filtered test words using two of the above discussed features such as number of run transitions and correlation which are computed for both the keyword and the test words from the database along their respective P_u , P_l , and P_c locations. Let the transition features computed for the keyword and test word along P_u , P_l , and P_c locations be represented as N_u^k , N_l^k and N_c^k , and N_u^t , N_l^t and N_c^t . Similarly, the correlation features computed are as follows, C_u^k , C_l^k and C_c^k , and C_u^t , C_l^t and C_c^t , respectively, for the keyword and the test word. Then, word matching is accomplished by matching the features of the keyword and the test word in three steps given below.

```

Step 1: If  $(P_c^k - \tau_1) \leq P_c^t \leq (P_c^k + \tau_1)$  then
    If correlation of  $C_c^k$  and  $C_c^t \geq 90\%$  then
        Goto Step 2
    else Exit
    else Exit
Step 2: If  $(P_u^k - \tau_2) \leq P_u^t \leq (P_u^k + \tau_2)$  then
    If correlation of  $C_u^k$  and  $C_u^t \geq 90\%$  then
        Goto Step 3
    else Exit
    else Exit
Step 3: If  $(P_l^k - \tau_3) \leq P_l^t \leq (P_l^k + \tau_3)$  then
    If correlation of  $C_l^k$  and  $C_l^t \geq 90\%$  then
        Keyword Matching Successful
    else Exit
    else Exit

```

The parameters τ_1 , τ_2 , and τ_3 are the number of run transition thresholds defined for P_c , P_u , and P_l locations. Their values were empirically determined to be 5 (using a training set of 10 compressed documents of Times New Roman font), which gives better performance and can potentially work in presence of some small noise or degradation happening due to image binarization. The order of feature comparison is along P_c , P_u , and P_l locations.

Table 1 Performance of keyword spotting on the dataset of [11]

#Documents	Font style	#Keywords	Precision (%)	Recall (%)	F-Measure
56	TNR	30	90.23	85.37	87.73
26	Arial	30	91.33	90.76	91.04
26	Calibri	30	92.03	88.33	90.14

Table 2 Comparative performance of keyword spotting with [11]

#Documents	Font style	#Keywords	OCR method [11]	Proposed method
			F-Measure	F-Measure
56	TNR	30	75.82	87.73
26	Arial	30	80.67	91.04
26	Calibri	30	81.05	90.14

3 Experimental Results

In order to validate the proposed method, two sets of experiments were performed, first with fixed font size documents and then with variable font size documents. For testing the method with fixed font size documents, the dataset presented in the research work of [11] were utilized. The dataset totally consists of 108 compressed text documents obtained from thesis and student project reports which contains documents in three font styles namely TNR-Times New Roman (56), Arial (26), and Calibri (26), all containing text with fixed font size (12). Based on the topics covered in the documents, 30 keywords were selected for testing purpose. The performance measures used for testing the performance of keyword spotting are precision and recall. The experimental results obtained for the above dataset using the proposed approach are tabulated in Table 1.

Further, the comparative analysis of the proposed method with that of OCR-based word spotting method of [11] is shown in Table 2. A sample result of keyword spotting in fixed font size text is shown in the uncompressed version of the document in Fig. 5. In the figure, the words within the red rectangular box are the one with bounding box ratio near to that of the keyword and the words within the green rectangular box are the spotted keywords. The proposed algorithm is also tested with compressed documents containing variable font size and mixed font size. The text documents contain texts with font size ranging from 10 to 18. The experimental results obtained is tabulated in Tables 3 and 4 respectively.

The preliminary level experimental results reported in this paper are specifically meant to establish the fact that OCR-less and decompression-less word spotting could be possible in compressed documents. Therefore, the efficiency aspects like computational complexity of the proposed algorithm is not reported in this paper. Further, the proposed keyword searching technique has the limitation of spotting only the exact keyword. For example the keyword ‘document’ can spot only the

Fax documents invoices receipts etc are traditionally subjected to compression for the efficiency of data storage and transfer. However in order to process these documents they need to undergo the stage of decompression which incents additional computing resources. This limitation induces the motivation to research on the possibility of directly processing compressed images. In this research work we propose to extract conventionally defined features such as projection profile run-histogram and entropy straight from run-length compressed documents. With the experiments we demonstrate that the features extracted straight from the compressed images are identical to those obtained from uncompressed version eliminating the need for an in-between stage of decompression. Subsequently we use the vertical and horizontal projection profile feature extracted from compressed document to carry out document segmentation. We propose methods to extract the segments of text-line words and characters directly from the compressed data. The proposed methods are experimentally validated.

Fig. 5 Word spotting results shown in uncompressed version of the document for the keyword ‘documents’

Table 3 Performance of keyword spotting in variable font size documents

#Documents	Font style	#Keywords	Precision (%)	Recall (%)	F-Measure
20	TNR	20	84.57	82.30	83.42
20	Arial	20	87.32	80.15	83.58
20	Calibri	20	83.61	78.23	80.83

Table 4 Performance of keyword spotting in mixed font size documents

#Documents	Font style	#Keywords	Precision (%)	Recall (%)	F-Measure
20	TNR	20	87.20	82.39	84.72
20	Arial	20	85.19	84.21	84.69
20	Calibri	20	81.16	80.27	80.71

word ‘document’ from the compressed database and fails to detect the cases such as ‘Document’ (First letter in upper case), ‘DOCUMENT’ (all upper case), and ‘documents/documentation’ (partial keyword). However, idea of decompression-less and OCR-less word spotting extended to the aforesaid limitations along with computational efficiency aspects of the proposed algorithm is anticipated in our next research communication.

4 Conclusion

In this research work, a novel method of word spotting directly in run-length compressed text documents using the simple font size invariant run-length features such as run transitions and correlations have been proposed which bypasses the stages of

decompression and OCRing. The features extracted from the keyword are matched with that of test words from the corpus of compressed documents and the matching words are identified which are the successfully spotted words. The proposed method was experimented and validated with a compressed document dataset and encouraging results have been reported. The word spotting method detects only exact words, however the idea of avoiding decompression and OCRing in word spotting could also be extended to partial keyword matching which could be next future enhancement work based on this research paper.

References

1. Bai, S., Li, L., and Tan, C. L. Keyword spotting in document images through word shape coding. *International Conference on Document Analysis and Recognition (ICDAR)* (2009), 331–335.
2. CCITT-Recommendaion (T.4). Standardization of group 3 facsimile apparatus for document transmission, terminal equipments and protocols for telematic services, vol. vii, fascicle, vii.3, geneva. Tech. rep., 1985.
3. CCITT-Recommendaion (T.6). Standardization of group 4 facsimile apparatus for document transmission, terminal equipments and protocols for telematic services, vol. vii, fascicle, vii.3, geneva. Tech. rep., 1985.
4. Chen, F. R., Bloomberg, D. S., and Wilcox, L. D. Detection and location of multicharacter sequences in lines of imaged text. *Journal of Electronic Imaging* 5, 1 (January 1996), 37–49.
5. Doermann, D. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding* 70, 3 (1998), 287–298.
6. Hull, J. J. Document matching on ccitt group 4 compressed images. *SPIE Conference on Document Recognition IV* (Feb 1997), 8–14.
7. Hull, J. J., and Cullen, J. Document image similarity and equivalence detection. *International Conference on Document Analysis and Recognition (ICDAR) 1* (1997), 308–312.
8. Javed, M., Nagabhushan, P., and Chaudhuri, B. B. Extraction of projection profile, run-histogram and entropy features straight from run-length compressed documents. *2nd IAPR Asian Conference on Pattern Recognition (ACPR)* (November 2013), 813–817.
9. Javed, M., Nagabhushan, P., and Chaudhuri, B. B. Automatic detection of font size straight from run length compressed text documents. *IJCSIT* 5, 1 (February 2014), 818–825.
10. Javed, M., Nagabhushan, P., and Chaudhuri, B. B. Automatic extraction of correlation-entropy features for text document analysis directly in run-length compressed domain. *13th International Conference on Document Analysis and Recognition (ICDAR)* (2015), 1–5.
11. Javed, M., Nagabhushan, P., and Chaudhuri, B. B. A direct approach for word and character segmentation in run-length compressed documents and its application to word spotting. *13th International Conference on Document Analysis and Recognition (ICDAR)* (2015), 216–220.
12. Lu, Y., and Tan, C. L. Document retrieval from compressed images. *Pattern Recognition* 36 (2003), 987–996.
13. Lu, Y., and Tan, C. L. Word searching in ccitt group 4 compressed document images. *International Conference on Document Analysis and Recognition (ICDAR)* (2003), 467–471.
14. Murugappan, A., Ramachandran, B., and Dhavachelvan, P. A survey of keyword spotting techniques for printed document images. *Artificial Intelligence Review* 35, 2 (2011), 119–136.

Design and Implementation of a Real-Time Autofocus Algorithm for Thermal Imagers

Anurag Kumar Srivastava and Neeta Kandpal

Abstract Good image quality is the most important requirement of a thermal imager or any other imaging system in almost all applications. Degree of focus in an image plays a very important role in determining the image quality, thus focusing mechanism is a very important requirement in thermal imagers. A real-time and reliable passive autofocus algorithm has been developed and implemented in FPGA-based hardware. This autofocus module has been integrated with the video processing pipeline of thermal imagers. Prior to the hardware implementation, different algorithms for image sharpness evaluation have been implemented in MATLAB and simulations have been done with test video sequences acquired by a thermal imager with motorized focus control to analyze the algorithms efficiency. Cumulative gradient algorithm has been developed for image sharpness evaluation. The algorithm has been tested on images taken from a thermal imager under varying contrast and background conditions, and it shows high precision and good discriminating power. The images have been prefiltered by a median rank-order filter using a 3×3 matrix to make it more robust in handling noisy images. Complete autofocus algorithm design comprising of a frame acquisition module for acquiring user selectable central region in the incoming thermal imager video, Cumulative Gradient-based image sharpness evaluation module, fixed step size search-based focal plane search module and a motor pulse generation module for generating motor drives have been implemented on Xilinx FPGA device XC4VLX100 using Xilinx ISE EDA tool.

Keywords Autofocus • Gradient • IR • FPGA • VHDL • MATLAB • NUC • DRC

A.K. Srivastava (✉) · N. Kandpal
Instruments Research and Development Establishment, Dehradun, India
e-mail: anurag@irde.drdo.in

N. Kandpal
e-mail: neeta@irde.drdo.in

1 Introduction

Focusing can be done either manually or automatically. Manual focusing is achieved by moving the focus lens assembly manually, assessing the sharpest picture by visual inspection and stop at the position, where the image seen at the system display is sharpest. But visual inspection of best focus position may not be accurate. An autofocus system brings a defocused image into focus by automatically adjusting the focus lens assembly to capture the sharpest image based on some sharpness evaluation metric. Depending on the scene distance from the thermal imager, the focusing lens is required to be adjusted for obtaining sharply focused image [1, 2]. Autofocus can be achieved using an active or passive approach. Active methods are based on ranging techniques to measure the range of target, and then by using the measured range the appropriate lens position is calculated. Passive methods rely solely on the captured image evaluation methods.

Proposed passive autofocus system design scheme has been divided into four modules as shown in Fig. 1.: (a) frame acquisition module which captures image data in selected region of interest (focus window) at different focus lens assembly positions, (b) image sharpness evaluation module that computes sharpness metric from captured frames, (c) focal plane search module that is used to search best focus position of lens assembly, and (d) motor pulse generation module that generates required pulses to drive motor assembly.

Frame acquisition module acquires central region (focus window) of the image at various focus lens assembly positions. Sharpness values are computed for the central focus window of images acquired. These sharpness values are used by the focal plane search module to control the focusing lens assembly to make it reach at

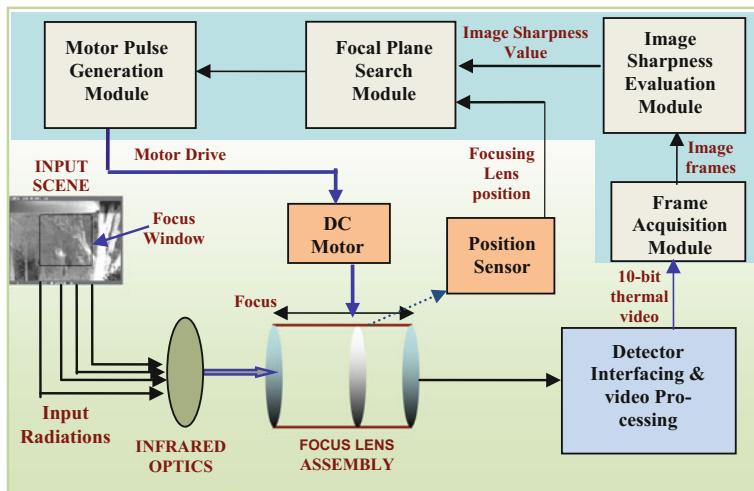


Fig. 1 Autofocus system block diagram

the position where the image sharpness is maximum [3]. The focus lens assembly is powered electrically to reach the position where the computed image sharpness value is maximum. Position sensor input is also used to decide the direction of traversal and limiting positions. Plot of sharpness values corresponding to different focus lens assembly positions gives the sharpness function. A good sharpness function should work for all possible scene conditions and it should be monotonic in nature with respect to the image focusing state [4]. The more defocused the image is, the less its value should be. Moreover, a good sharpness function should have good discrimination power; i.e., it should give a sharper response when the focus point is closer and it should be able to combat noisy and low-contrast imaging conditions [4].

2 Image Sharpness Evaluation Algorithm

Image sharpness is linked to the sharpness of its edge features. Focused images contain more sharp edges compared to defocused images. Image sharpness evaluation approaches involve computing a sharpness value that reflects the sharp edges within a selected region of image. A multitude of image sharpness evaluation algorithms exist today [5–7]. Familiar sharpness evaluation approaches include image histogram, variance, energy of square gradient, entropy function, frequency spectrum function, energy of Laplacian, square Gaussian gradient, Laplacian–Gaussian, Tenengrad, etc. Previous works by researchers have showed that image histogram and variance both have better operating efficiency and globality but poor discrimination power. In the “Tenengrad” method, the horizontal and vertical Sobel operators are used to evaluate the strength of the horizontal and vertical gradients of the image. Then the sum of square of the gradients is defined as the sharpness value. Similarly, Laplacian method employs convolving an image with the Laplacian mask and the sum of square of the pixels in the resulting image is defined as the sharpness value. In contrast to the variance-based methods, these two methods in general have good discrimination power. However, they (especially the Laplacian-based method) are more sensitive to noise than the variance-based methods and have worse operating efficiency. Image sharpness evaluation model based on image absolute central moment (ACM) [8] has better globality, focusing efficiency, reliability, and accuracy but just like the variance, its variation range near the focal plane is narrow. Frequency spectrum function has worse operating efficiency, better locality.

An Image Sharpness Evaluation algorithm based on cumulative gradient measure has been developed for autofocus control of thermal imagers. Image Gradient measure is used to find the transition rate between adjacent pixel values. Sharper edges in focused images lead to a sharper transition in pixel values resulting in larger cumulative gradient value for that image as compared to the defocused images.

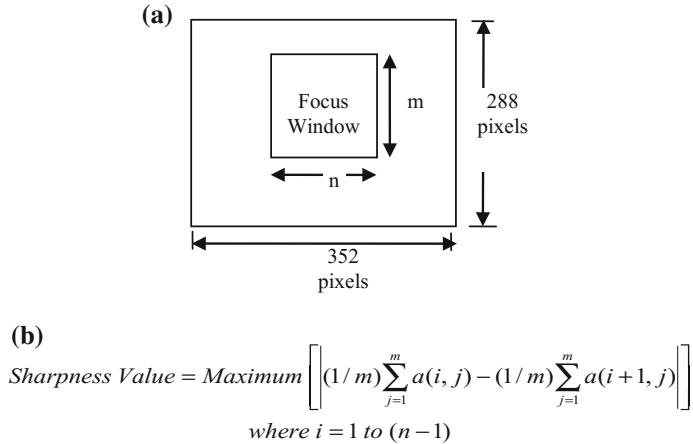


Fig. 2 a Central focus window selection, b Sharpness evaluation formula

For computing cumulative gradient, a central region of image (focus window) is selected as shown in Fig. 2a. Cumulative gradient-based image sharpness value is calculated as per the sharpness evaluation formula (Fig. 2b). Sharpness values are computed for frames acquired at different focus lens assembly positions. This sharpness value increases as the image becomes more focused. Taking a larger region of interest makes the algorithm more robust to image noise but it adds to the computational complexity. This algorithm takes care of vertical edges within the scene as they are more prominent than horizontal edges in typical scenes.

3 Simulation Results for Sharpness Evaluation Algorithm

Video sequences were captured with a thermal Imager having motorized FOV and focus control mounted on a tripod. The focusing lens assembly was moved from one extreme position to the other extreme position so that the video shows a transition from defocused state to focused state and then again to defocused state. Frames extracted were 8-bit gray scale images of 288×352 pixels resolution. Cumulative gradient values were computed for central 160×160 windows of the image frames.

Sharpness values were computed for 160×160 central region of eight sequential frames of video sequence 1 (Fig. 3a). In actual autofocus scenario, the extracted image frames refer to images at different focus lens assembly positions. Sharpness values corresponding to different frames were plotted to give the sharpness function. Sharpness function plot (Fig. 3c) shows frame no. 5 as the best focused frame that is clearly validated visually (Fig. 3b). Also, sharpness function plot shows large variation range of 0.3–1.0 which is very important for the focal plane search module to work efficiently.

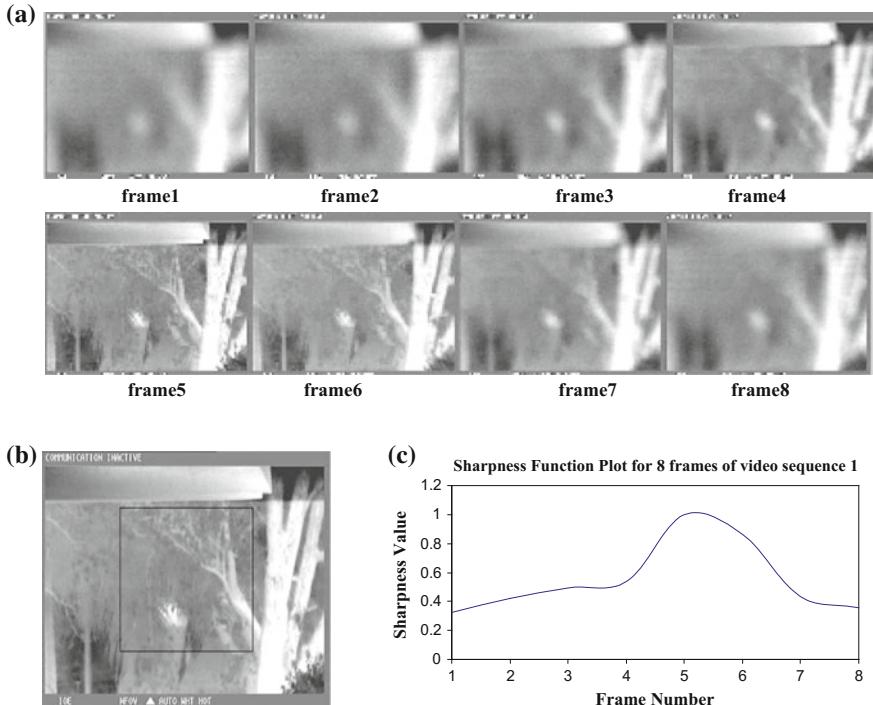


Fig. 3 **a** 8 frames of video sequence 1 in different focusing states **b** Best focused frame (frame 5) with focus window highlighted **c** Sharpness function plot for eight sequential frames

Same algorithm was applied to eight frames of another video sequence having very low-contrast 4-bar target. The simulation results are shown below (Fig. 4):

The simulation results show that the sharpness functions, for each of the two video sequences, have one peak at the best focus position and it strictly decreases away from this peak.

To analyze the effect of different focus window sizes on the sharpness evaluation algorithm, the sharpness values were calculated for three different center focus windows of sizes 80×80 , 160×160 and 240×240 of video sequence 1 (Fig. 5a). Simulation results showed that larger focus window size resulted in smoother sharpness function with more variation range compared to smaller window (Fig. 5b). The best focused frame was found the same (frame 5) for all three focus windows selected.

To see the effect of noise robustness of the algorithm, speckle noise was added to video sequence 1 frames using standard MATLAB [9] function. Sharpness values were computed for different focus window sizes of noise added frames and median filtered frames. Larger focus window size (240×240) was found more effective as compared to 160×160 window sizes under noisy conditions. As illustrated in

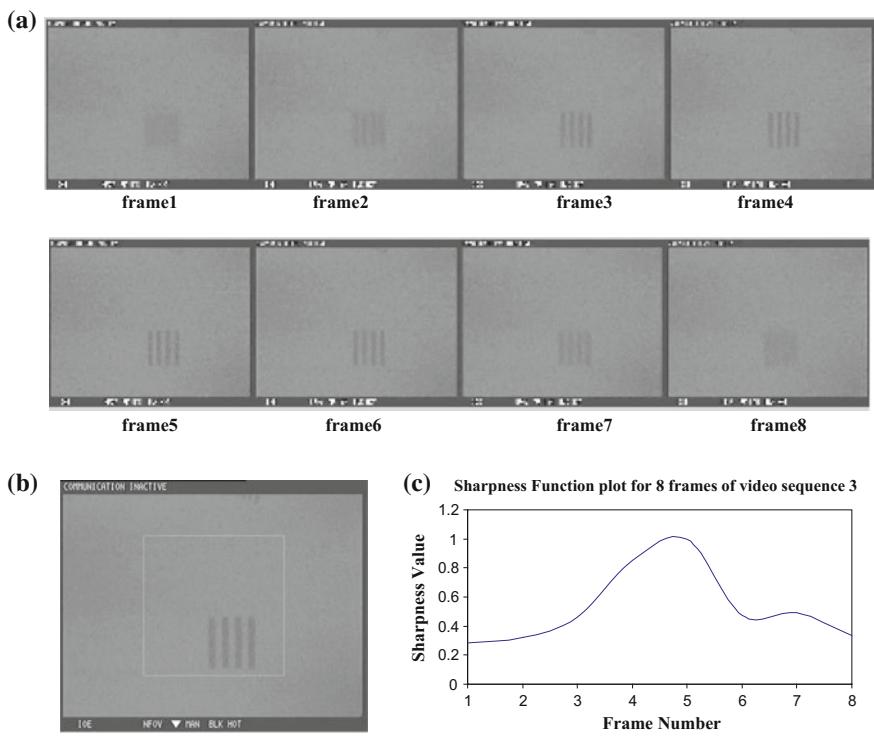


Fig. 4 **a** 8 frames of video sequence two in different focusing states. **b** Best focused image: frame no. 5; **c** Sharpness function plot for 8 sequential frames of video sequence 3

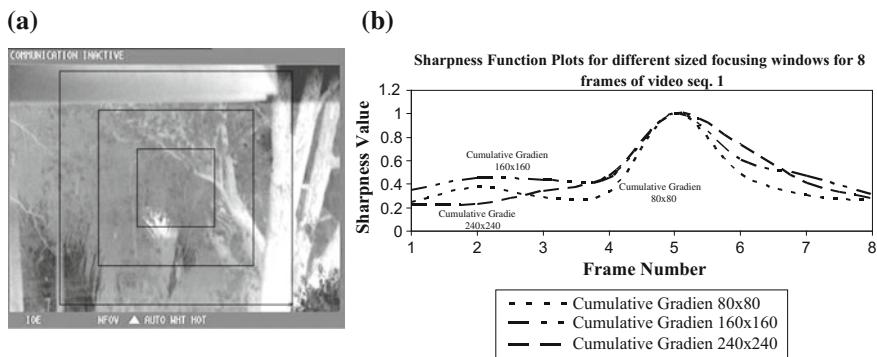


Fig. 5 **a** Focused Frame of video seq. 1 with three different-sized focus windows. **b** Sharpness function plots for different focus window sizes for eight frames of video seq. 1

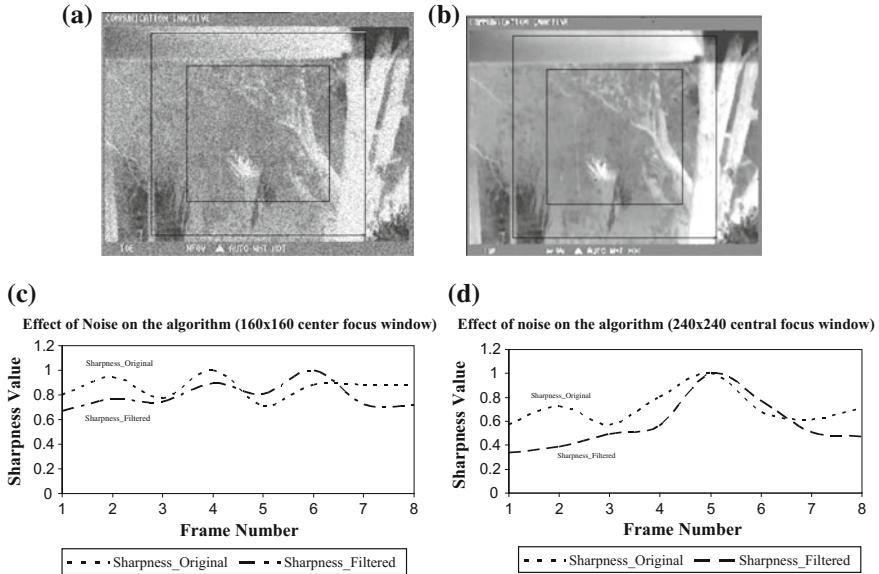


Fig. 6 **a** Noise added best focused frame with two focus windows highlighted, **b** Median filtered frame with two focus windows highlighted. **c** Sharpness function plots for 160×160 size focus window, **d** Sharpness function plots for 240×240 center focus window

Fig. 6c, under noisy condition using 160×160 window size, resulted in false detection of best focus frame (frame 4 was detected in noisy image and frame 6 was detected in filtered image).

The sharpness function for 240×240 center focus window was found to perform better in noisy condition. This window size was found effective in combating this much amount of noise and identifying the best focused frame correctly. It also showed larger variation range of sharpness values (Fig. 6d) as compared to the 160×160 focus window size (Fig. 6c). Filtering resulted in a smoother sharpness curve. Considering the effect of focus window size on the algorithm

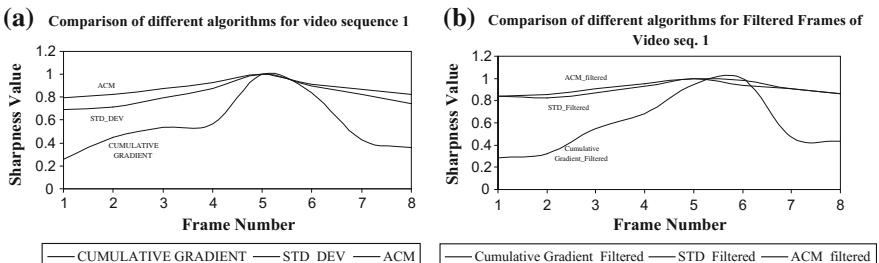


Fig. 7 **a** Comparison of three algorithms for eight frames of video sequence 1. **b** Comparison of three algorithms for eight filtered frames of video sequence 1

results, selection of window size was given as user controlled through some external interface. The algorithm is computationally less complex, so sharpness can be evaluated for full frame resolution also in real time.

The cumulative gradient algorithm was compared with two other widely used sharpness evaluation algorithms: standard deviation (STD_DEV) and absolute central moment (ACM). Sharpness values corresponding to all three algorithms were calculated for central 160×160 portions of eight sequential frames and eight filtered frames of video sequence 1. All three sharpness functions were scaled to $(0, 1)$ range so that the results can be compared. Cumulative gradient algorithm was found to be having greater variation range of sharpness values as compared to the variation range for other two algorithms (Fig. 7).

4 Focal Plane Search Module

Focal plane searching is a process of driving the focusing motor to determine the focal plane location. The focal plane search algorithms that have been used in various autofocus systems include: global scan, Fibonacci searching, semi-searching, variation proportion searching, etc. Among these searching algorithms, global scan has better stability and reliability; Fibonacci searching reduces a great deal of image frames by introducing the Fibonacci array, but it requires higher quantification accuracy of the evaluation function and it is more sensitive to the delay of the driving motor.

Fixed step size search has been employed for focal plane search of thermal imager best focus position. Image sharpness value corresponding to the starting focus lens position is computed and the result is stored in a register. Then the lens assembly is moved by one step in predetermined direction and image sharpness is computed at that position. These two sharpness values are then compared which gives the direction of movement of lens assembly in next step. In this way different positions of focus lens assembly position are scanned and the image sharpness corresponding to those positions are computed. Once peak is found and the sharpness value starts to decrease, the focus lens assembly will not stop at that position, instead it will go three steps further in the same direction to check that the peak found is not the one of various sub peaks (Fig. 8). If the sharpness values corresponding to these three positions are less than the peak found, the focal plane search module will generate a signal to take the lens assembly back to the peak value position and stop. If another peak greater than the earlier is found then the motor will continue to search in the same manner. In our implementation, searching for three subsequent positions is based on the step size we selected. Number of search positions can be optimized as per the focus step size. This module has been implemented in FPGA using VHDL.

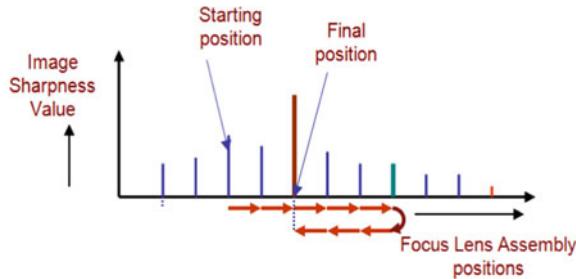


Fig. 8 Focal plane search method

5 Hardware Implementation

The hardware module for autofocus control (Fig. 10) has been implemented in Xilinx FPGA device XC4VLX60 using Xilinx ISE EDA tool and VHDL [10]. Autofocus module has been integrated with the video processing pipeline (Fig. 9) of thermal imager. 15-bit Infrared (IR) sensor data is processed and various functions like nonuniformity correction (NUC), bad pixel replacement (BPR), automatic gain control (AGC) and dynamic range compression (DRC) are implemented in FPGA. Autofocus module interfaces with this video pipeline output (10-bit data) and other external inputs required to initiate this function. Sufficient resources were available in XC4VLX60 FPGA device to integrate both video processing pipeline and autofocus module.

Autofocus (AF) module has five inputs: clock (32 MHz board clock), asynchronous reset, vertical and horizontal sync signals (vbl, hbl), lens position and a switch interface (AF, i.e., autofocus switch to initiate the autofocus functionality). Although focus window size has been fixed to 160×160 in the current implementation, it can be given as an input to this module.

AF module generates two outputs: motor_pulse_positive and motor_pulse_negative to drive the motors in two directions. The thermal imager will be in manual focus mode till the AF switch is pressed. This switch enables this module and initiates the autofocus operation. 40 ms is taken as the debouncing time for this

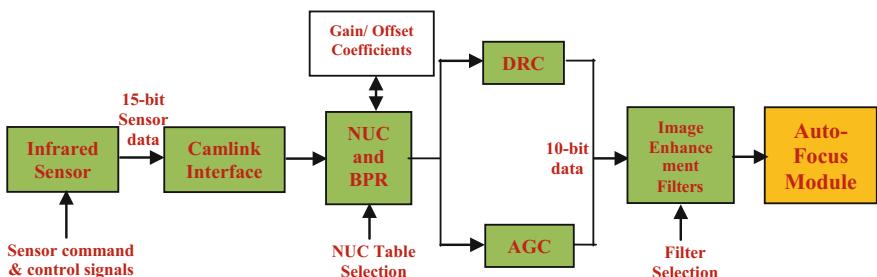


Fig. 9 Video processing pipeline of thermal imager

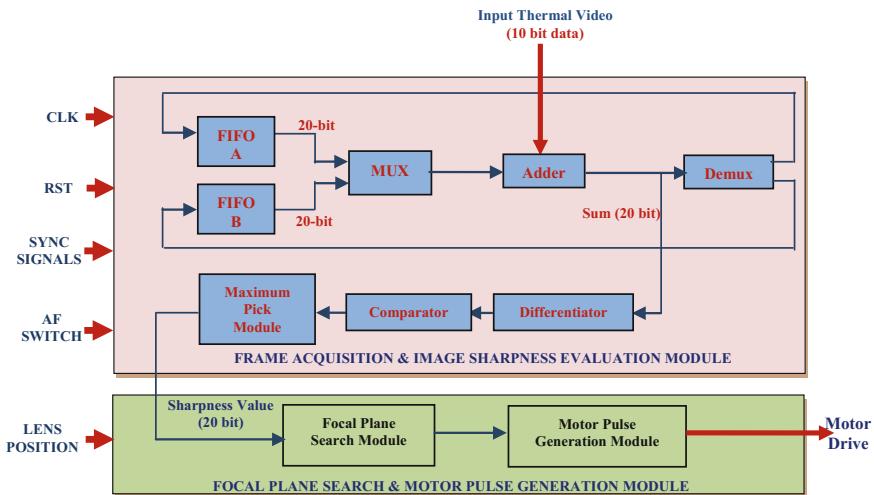


Fig. 10 Data flow diagram of complete autofocus module

switch. 32 MHz onboard clock is internally divided into 8 MHz clock which is used as a reference to trigger all processes. FIFO A and B are parameterizable hardware specific synchronous modules instantiated in the VHDL design. The two FIFOs work in ping pong fashion to store the intermediate results. Adder, differentiator, and comparator modules are implemented in VHDL and are used to compute the image sharpness value. Focal plane search module compares two sharpness values and provides directional decision input to the motor pulse generation module which generates the drive according to decision taken. Motor pulse generation module has two output lines. Pulse at the two output lines drives the focus lens assembly in two directions. The termination criteria for stopping search are also decided by focal plane search module. The search process is terminated when

(a) The peak is found and there is no greater peak in next three steps in same direction of motor movement, or (b) The extreme position is reached and the focus lens assembly has traversed two times to-and-fro. In this case the focus lens assembly will stop at the extreme position.

This autofocus module accepts 640×512 resolution, 10-bit image data, computes sharpness value for center 160×160 focus window, and drives focus motor accordingly. This implementation can work up to 220 MHz clock rate that

Table 1 Performance data of Autofocus module

Logic utilization	Used	Available	Utilization (%)
# Slices	525	6244	8
# Slice Registers	587	12288	4
# Slice LUT	829	12288	6
# Block RAM/FIFO	2	48	4

meets the real-time requirements of 25 Hz frame rate. From resource utilization point of view (Table 1), no. of slices used is only 8 %, so this module was easily integrated with the video processing electronics of thermal imager.

6 Conclusion and Future Work

A real-time autofocusing algorithm is designed and implemented for thermal imagers. An accurate and reliable image sharpness evaluation method based on the cumulative gradient technique is presented in this paper. This algorithm gives good results with thermal scenes with different-sized targets having varying contrast and noise levels. The reported sharpness evaluation algorithm has been found very effective when the scene contains prominent features. The algorithm will be further optimized to provide autofocus capability under all types of imaging scenario.

Acknowledgments We would like to thank Dr. S S Negi, Director, I.R.D.E for allowing us to work in this area.

References

1. R.D. Hudson, Infrared Systems Engineering, John Wiley, New York, 1969.
2. Gregory A. Baxes, “Digital Image Processing- Principles and Applications”, 1994.
3. J. Baina and J. Dublet, “Automatic focus and iris control for video cameras”, Fifth International Conf. on Image Processing and its Applications, pp. 232–235, July 1995.
4. Mark Antunes and Michael Trachtenberg, “All-In-Focus Imaging From A Series Of Images On Different Focal Planes”, B.Sc thesis report, Faculty of Engineering, University of Manitoba, March 2005.
5. Feng Li, Hong Jin, “A Fast Auto Focusing Method For Digital Still Camera”, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18–21 August 2005.
6. Mukul V. Shirvaikar, “An Optimal Measure for Camera Focus and Exposure” Proceedings of the IEEE SSST 2004.
7. Ng Kuang Chern, Nathaniel Poo Aun Neow and Marcelo H. Ang Jr., “ Practical issues in pixel-based autofocusing for machine vision”, Proceedings of the 2001 IEEE International Conference on Robotics & Automation, Seoul, Korea, May 21–26, 2001 M.
8. Chun-Hung Shen and Homer H. Chen, “Robust Focus Measure for Low-Contrast Images”, Proceedings of the IEEE SSST 2006.
9. MATLAB version 2012.
10. J. Bhasker, A VHDL Primer, AT&T, 1999.

Parameter Free Clustering Approach for Event Summarization in Videos

Deepak Kumar Mishra and Navjot Singh

Abstract Digital videos, nowadays, are becoming more common in various fields like education, entertainment, etc. due to increased computational power and electronic storage capacity. With an increasing size of video collection, a technology is needed to effectively and efficiently browse through the video without losing contents of the video. The user may not always have sufficient time to watch the entire video or the entire content of the video may not be of interest of user. In such cases, user may just want to go through the summary of the video instead of watching the entire video. In this paper we propose an approach for event summarization in videos based on clustering method. Our proposed method provides a set of key frames as a summary for a video. The key frames which are closer to the cluster heads of the optimal clustering are combined to form the summarized video. The evaluation of the proposed model is done on a publicly available dataset and compared with ten state-of-the-art models in terms of precision, recall and F-measure. The experimental results demonstrate that the proposed model outperforms the rest in terms of F-measure.

Keywords Event summarization • Key frames • Clustering • SD index • Video skimming

1 Introduction

In recent past, due to rapid advancement in imaging techniques, internet, and other communication media, a huge amount of videos are produced for different purposes on daily basis, for instance, surveillance videos in various security systems, videos

D.K. Mishra · N. Singh (✉)

Department of Computer Science and Engineering, National Institute of Technology,
Srinagar, Uttarakhand, India
e-mail: navjot.singh.09@gmail.com

D.K. Mishra

e-mail: deepakkumarmishra94@gmail.com

for entertainment and educational purposes, etc. However, an effective and efficient method of browsing the contents of a video is one of the most crucial challenges in accessing such a large amount of videos.

Video summarization [1] is a tool that is used to create a short summary of a video without losing the events occurring in video. It involves three steps namely segmentation, feature detection and representation. There are basically two approaches used for video summarization: video skimming and keyframe extraction. Video skimming is a method which focuses on summarization of a video without losing the important events along with its semantics. It can be considered as a concise representation of a video clip. Video skimming is basically based on the selection of local representative frames from the entire set of video frames [2]. Video skimming is a difficult task because semantic level analysis of contents of a video by an automated machine is still not possible by the computational technologies we currently have. On the other hand, keyframe extraction [3] refers to the selection of important frames (or salient frames) from a video. Here frames are ranked based on importance, so as to break the video into unique events.

In this paper, our objective is to select key frames from a global view point by using clustering algorithm and optimizing the number of clusters which are required to effectively and efficiently divide and represent the entire content of the video. In this paper, we are not going to focus on comprehensive coverage of the entire video but to extract most representative key frame set for the video. The proposed algorithm is divided into two phases. In the first phase achromatic signals representing individual frames undergo a clustering procedure. A cluster validation technique is used to make the clustering parameter free by identifying the optimal number of clusters for a given video. Then in the second phase, the frames closest to the respective cluster heads are chosen as the key frames for the video content.

In Sect. 2 related works pertaining to video summarization is discussed. Then in Sect. 3 proposed model is presented, followed by the experimentation and results discussion in Sect. 4. Finally the Sect. 5 concludes the paper.

2 Related Works

There are basically two main approaches used in previous works for automatic video summarization. The first approach is frame based approach, such as color histogram, image difference [4–6]. This approach extensively segments the video into individual frames and then the summarization algorithm is applied over these frames. The second approach is based on the scene detection or motion based approach. Scenes are detected using motion information which are computed based on some global models, local models or global models fitted into local context [7, 8]. These image based and scene based models are capable to provide reasonable segmentation of the video. However, the number of key frames may become large in order to provide an overview of a long video if the key frames are considered at every cut point. To overcome this, the clustering algorithms [9–11] can be applied

for grouping the shots. The clustering is mainly based on color information and shape and spatial correlation [10, 12]. Furthermore, some authors [13] have proposed the scene transition graph for video content representation.

Some authors have proposed the video summarization systems based on high-level semantic properties of the video [14]. In this approach, video's semantic contents are analyzed and along with the information of user content preference, the decision whether a frame will be in the summary or not is taken. This approach is basically used for video skimming. Li et al. [15] considered temporal rate distortion MINMAX optimization problem for selection of key frames. Ma et al. [16] used human attention to prioritize and filter the information more effectively and efficiently and proposed a model for video understanding. One of the latest approaches formulated the problem of video summarization as sparse dictionary selection [17]. Local feature based key frames selection approaches have also been proposed. Guan, et al. [18] proposed a method which exploits the distinctive nature of local features for identification of key frames.

Some of the other models include open video project storyboard (OVP) [19], Delaunay Clustering (DT) [20], still and moving video storyboard (STIMO) [21], video summarization (VSUMM) [22] as suggested by deAvila et al., sparse dictionary (SD) [23], keypoint based keyframe selection (KBKS) [24], and three variants of offline MSR based methods (OffMSRm, OffMSRa, OnMSR) [25].

3 Proposed Approach

In this paper, we have proposed the video summarization method which is based on global view point and key frame extraction. The resultant summary will be based on the selection of most representative frames rather than a comprehensive coverage of the entire video content. The selection of key frames is based on the result of the clustering algorithm where the frames closest to the respective cluster heads represent the key frames. The proposed summarization technique can be explained in following steps:

3.1 Preparation of Video for Clustering

The very first step is to segment the given video into frames and convert resulting RGB frames into grayscale frames for easier computations. The grayscale frames so obtained is converted into a linear data set so as to apply the clustering algorithm on the frames of the video. Let the total number of frames in the given video be N of size $M = W \times H$, where W and H represent width and height of a frame respectively.

The resultant data for clustering is as follows:

$$\mathbf{V} = [\mathbf{f}_1 \mathbf{f}_2 \dots \mathbf{f}_N]^T \quad (1)$$

where \mathbf{f}_i represents the i -th frame of the video.

3.2 Optimal Clustering to Select Key Frames

Now the data set consisting of N frames are clustered into k^* clusters using k-means clustering algorithm. The optimal value k^* is obtained as a result of learning process based on SD Validity Index [26]. The SD Validity Index is one of the most recent clustering validity approaches, which is based on the *average scattering* for the clusters and the *total separation* between the clusters. The index is good at measuring the homogeneity and compactness of the clusters. Let \mathbf{C}_i be a cluster of vectors. Let \mathbf{X} be the entire dataset to be clustered and n be the number of desirable clusters.

Average scattering for the clusters is defined as

$$Scat(n) = \frac{1}{n} \sum_{i=1}^n \frac{\|\sigma(\mathbf{C}_i)\|}{\|\sigma(\mathbf{X})\|} \quad (2)$$

where \mathbf{C}_i is the i -th cluster, \mathbf{X} is the entire dataset and n is the number of clusters. $\sigma(\mathbf{C}_i)$ is the standard deviation of the i th cluster and $\sigma(\mathbf{X})$ is the standard deviation of the entire dataset.

Total separation between clusters is defined as

$$Dis(n) = \frac{D_{max}}{D_{min}} \sum_{k=1}^n \left(\sum_{z=1}^n \|v_k - v_z\| \right)^{-1} \quad (3)$$

where $D_{max} = \max(\|v_i - v_j\|)$ and $D_{min} = \min(\|v_i - v_j\|)$; $\forall i, j \in \{1, 2, 3, \dots, n\}$ represents the maximum and minimum separation between the cluster centers respectively.

Now, the SD validity index can be defined as follows

$$SD(n) = K \cdot Scat(n) + Dis(n) \quad (4)$$

where K is a weighting factor equal to $Dis(c_{max})$ where c_{max} is the maximum number of input clusters.

Table 1 Algorithm of the proposed model

Algorithm: The clustering based video summarization algorithm.

Input: Video V

Output: Event summary E consisting of key frames.

1. Select the entire set of frames of the video and convert the RGB frames into grayscale frames.
 2. Convert the 2-D array of pixel values of grayscale frames into 1-D array of pixel values.
 3. for $k = k_{min}$ to k_{max}
 - apply k-means clustering algorithm for k clusters.
 - Find $SD(k)$
 - end
 4. $k^* = \operatorname{argmin} SD(k)$
 5. Find k^* frames closest to the cluster heads and combine them into E .
 6. return E .
-

The number of clusters that minimizes the above SD validity index can be considered as an optimal value, k^* , for the number of clusters present in the dataset, which can be computed as

$$k^* = \operatorname{argmin}_{k_{min} \leq k \leq k_{max}} SD(k) \quad (5)$$

where k_{min} and k_{max} are the minimum and maximum value of the number of clusters respectively. Finally k^* frames from the video are picked that are closest to the cluster heads and constitutes the event summarization video E .

$$E = \left[\mathbf{f}_{e_1} \mathbf{f}_{e_2} \cdots \mathbf{f}_{e_{k^*}} \right]^T \quad (6)$$

where \mathbf{f}_{e_i} represents the closest frame to the i -th cluster.

The algorithm pertaining to the proposed model can be seen in Table 1.

4 Experimental Results and Discussions

In this section, first we will describe the datasets that we have used for evaluation and then introduce the evaluation metric. Similar evaluation metric is employed among different summarization approaches for fair comparison.

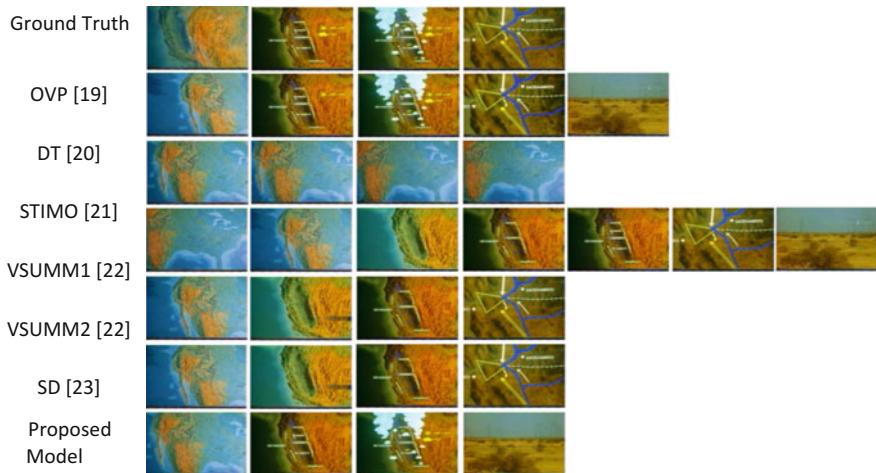


Fig. 1 Qualitative comparison of the proposed model with other state-of-the-art models

4.1 Experimental Setting

4.1.1 Benchmark Datasets

To evaluate the performance of our algorithm, we have conducted our experiments on a publicly available dataset, VSUMM dataset, which consists of videos of different genres.

*VSUMM Dataset*¹: It consists of 50 videos in MPEG-1 format (30fps, 352×240 pixels). These videos are of different genres (documentary, educational, historical, lecture) and varies in duration between 1 to 4 min. It provides 250 user summaries, 5 summaries for each video, created manually by 50 different users, each one dealing with 5 different videos. These user summaries are considered as the ground truth summaries for our experiment which is provided by the group which released the dataset [20]. Figure 1 shows the qualitative comparison of the proposed model with other state-of-the-art models. It can be clearly seen that the results of the proposed model are the most closer to the ground truth.

¹<https://sites.google.com/site/vsummsite/download>.

4.1.2 Evaluation Metric

To measure the quantitative performance and effectiveness of the proposed approach for event summarization, we employed the F-score measure which provides the collective information about the precision and recall. In general, F-score is defined as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} & Recall &= \frac{TP}{TP+FN} \\ F_{\beta} &= \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \end{aligned} \quad (7)$$

where β represents the effectiveness for which recall has β times as much importance as that of precision. In the evaluation of our experimental results, we have used $\beta=1$ to give equal importance to both precision and recall.

4.2 Performance Evaluation on Datasets

We compared the summarization results of the proposed method with ten state-of-the-art methods to evaluate the performance of the proposed method. It can be seen from the Table 2 and Fig. 2 that the summarization performance of the proposed approach outperforms the other summarization methods with the highest average value of the F-score **0.609** for VSUMM dataset.

Table 2 Performance of various algorithms on the dataset

Algorithms	Precision	Recall	F-measure
OVP [19]	0.43	0.64	0.514
DT [20]	0.47	0.50	0.485
STIMO [21]	0.39	0.65	0.488
VSUMM1 [22]	0.42	0.77	0.544
VSUMM2 [22]	0.48	0.63	0.545
SD [23]	0.40	0.61	0.483
KBKS [24]	0.31	0.89	0.460
OffMSRm [25]	0.58	0.58	0.580
OffMSRa [25]	0.60	0.57	0.585
OnMSR [25]	0.50	0.66	0.569
Proposed model	0.62	0.60	0.609

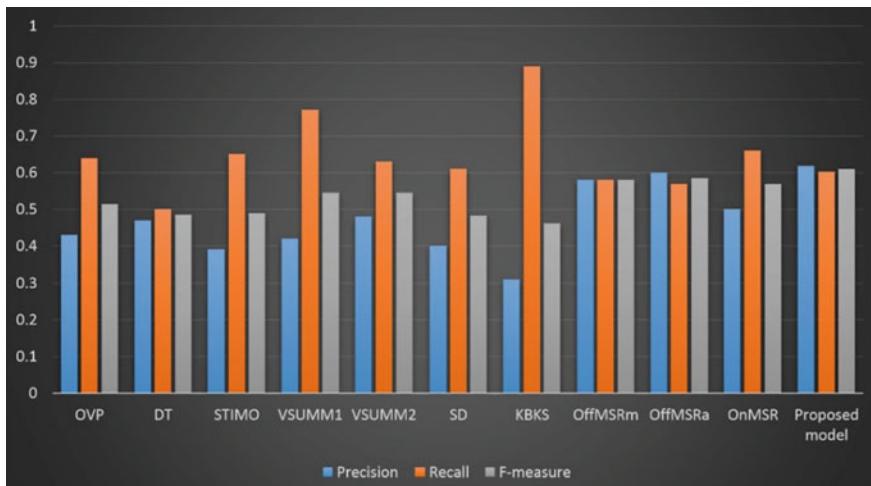


Fig. 2 Quantitative Evaluation of different algorithms in terms of Precision, Recall and F-measure for the VSUMM dataset

5 Conclusion and Future Work

In this paper, we have proposed an event summarization technique that is based on extraction of key frames such that the number of frames in the summary should be as few as possible. The technique is mainly based on clustering the entire set of video frames into an optimal number of clusters. The frames closest to the respective cluster heads of each cluster are considered as the key frames in the summary of that video. Experiments have been conducted on the VSUMM dataset (which contains videos from different genres) for the performance evaluation and comparison of the proposed approach with ten state-of-art techniques. The selection of optimal number of clusters can be extended to the expectation optimization where clustering is done based on certain probability function and the clusters can be merged if the similarity between clusters is above a certain threshold.

References

1. M. Shaohui, G. Genliang, W. Zhiyong, W. Shuai, M. Mingyi, D. F. David, “Video summarization via minimum sparse reconstruction”, in Pattern Recognition, vol. 48, ELSEVIER, 2014, pp. 522–533.
2. A. Hanjalic, H. Zhang, “An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis”, IEEE Trans. Circuits Syst. Video Technol. 9 (8) (1999) 1280–1289.

3. Y. Zhuang, Y. Rui, T. S. Huang, S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering", in: Proceedings of the International Conference on Image Processing, vol. 1, IEEE, 1998, Chicago, Illinois, USA, pp. 866–870.
4. M. M. Yeung, B. Liu, "Efficient matching and clustering of video shots", in: International Conference on Image Processing, vol. 1, IEEE, Washington, D.C., USA, 1995, pp. 338–341.
5. K. Otsuji and Y. Tonomura, "Projection-detecting filter for video cut detection," Multimedia Syst., vol. 1, pp. 205–210, 1994.
6. H. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partition of full-motion video," Multimedia Syst., vol. 1, pp. 10–28, 1993.
7. B. Yeo and B. Liu, "Rapid scene analysis on compressed video," IEEE Trans. Circuits Syst. Video Technol., vol. 5, pp. 533–544, Dec. 1997.
8. H. Zhang, C. Y. Low, Y. Gong, and S. W. Smoliar, "Video parsing using compressed data," in Proc. IS&T/SPIE Conf. Image and Video Processing II, 1994, pp. 142–149.
9. H. Aoki, S. Shimotsuji, and O. Hori, "A shot classification method of selecting effective key-frames for video browsing," in ACM Multimedia 96, 1996, pp. 1–10.
10. M. M. Yeung, B. Yeo, W. Wolf, and B. Liu, "Video browsing using clustering and scene transitions on compressed sequences," in Multimedia Computing and Networking, vol. SPIE-2417, 1995, pp. 399–413.
11. D. Zhong, H. Zhong, and S. Chang, "Clustering methods for video browsing and annotation," in Storage and Retrieval for Still Image and Video Databases IV, vol. SPIE-2670, 1996, pp. 239–246.
12. M. M. Yeung, B. Yeo, and B. Liu, "Extracting story units from long programs for video browsing and investigation," in Proc. IEEE Multimedia Computing & Syst., 1996, pp. 296–305.
13. M. M. Yeung and B. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," IEEE Trans. Circuits Syst. Video Technol., vol. 7, pp. 771–785, Oct. 1997.
14. S. S. Intille and A. F. Bobick, "Closed-world tracking," in Proc. IEEE Int. Conf. Comput. Vision, June 1995, pp. 672–678.
15. Z. Li, G. Schuster, and A. Kataggelos, "Minmax optimal video summarization," IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 10, pp. 1245–1256, Oct. 2005.
16. Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. Multimedia, vol. 7, no. 5, pp. 907–919, Oct. 2005.
17. C. Yang, J. Shen, J. Peng and J. Fan, "Image collection summarization via dictionary learning for sparse representation," Patter Recog., vol. 46, no. 3, pp. 948–961, 2013.
18. G. Guan, Z. Wang, S. Lu, J. Da Deng, and D. Feng, "Keypoint based keyframe selection," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 729–734, Apr. 2013.
19. OpenVideoProject, "<http://www.open-video.org/>", 2011.
20. P. Mundur, Y. Rao, Y. Yesha, "Keyframe-based video summarization using Delaunay clustering," Int. J. Digit. Libr., vol. 6, no. 2, pp. 219–232, 2006.
21. M. Furini, F. Geraci, M. Montangero, M. Pellegrini, "Stimo: still and moving video storyboard for the web scenario," Multimed. Tools Appl., vol. 46, no. 1, pp. 47–69, 2010.
22. S. E. F. deAvila, A. P. B. Lopes, et al., "Vsumm: a mechanism designed to produce static video summaries and a novel evaluation method," Pattern Recognit. Lett., vol. 32, no. 1, pp. 56–68, 2011.
23. Y. Cong, J. Yuan, J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," IEEE Trans. Multimed., vol. 14, no. 1, pp. 66–75, 2012.
24. G. Guan, Z. Wang, S. Lu, J. Da Deng, D. Feng, "Keypoint based keyframe selection," IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 4, pp. 729–734, 2013.
25. S. Mei, G. Guan, Z. Wang, S. Wan, M. He, D. D. Feng, "Video summarization via minimum sparse reconstruction," Pattern Recognition, vol. 48, pp. 522–533, 2015.
26. M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques," Journal of Intelligent Information Systems, vol. 17, no. 2, pp. 107–145, 2001.

Connected Operators for Non-text Object Segmentation in Grayscale Document Images

Sheshera Mysore, Manish Kumar Gupta and Swapnil Belhe

Abstract This paper presents an unconventional method of segmenting nontext objects directly from a grayscale document image by making use of connected operators, combined with the Otsu thresholding method, where the connected operators are realized as a maxtree. The maxtree structure is used for a simplification of the image and at a later stage, it is used as a structure from which to extract the desired objects. The proposed solution is aimed at segmenting halftone images, tables, line drawings, and graphs. The solution has been evaluated and some of its shortcomings are highlighted. To the best of our knowledge, this is the first attempt at using connected operators for page segmentation.

Keywords Page segmentation · Connected operators · Otsu thresholding

1 Introduction

The segmentation of a page into its text and nontext objects forms one of the preliminary tasks in a document image analysis task. Page segmentation generally precedes a region classification step to group nontext objects into one of many different types of nontext regions so that an analysis specific to the object type may then be applied to it. In addition, the segmentation also serves to ensure that any OCR tasks applied to any text does not evaluate a nontext region and lead to an incorrect result.

Here we propose a method of segmentation of the nontext objects in an image that makes use of connected operators realized as maxtrees [5]. Connected operators form

S. Mysore (✉)
Independent Student Researcher, Pune, India
e-mail: mssheshera@yahoo.com

M.K. Gupta · S. Belhe
Center for Development of Advanced Computing, Pune, India
e-mail: mgupta@cdac.in

S. Belhe
e-mail: swapnilb@cdac.in

a set of morphological operators which operate on grayscale images, as opposed to structuring element-based morphological operators which act on binary images or their extensions which act on grayscale images [1]. Our use of connected operators therefore is a departure from the way in which conventional page segmentation algorithms work, in that segmentation is conventionally performed on the binarized image. The solution we propose aims to achieve segmentation of nontext regions while withstanding any degradation that may be present in the image. The proposed approach aims to segment nontext components such as halftone images, tables, graphic elements, line drawings, and graphs.

The use of connected operators for document image analysis tasks has been limited despite their suitability to tasks often necessary in document image analysis; tasks such as low level filtering while preserving contours in the image and high level object extraction [6]. The suitability of connected operators for document image analysis was suggested by Lazzara et al. [1] in the very recent past. Prior to this however, connected operators have been applied to document image analysis tasks. Naegel and Wendling applied these operators to the task of document image binarization [3] while Wilkinson and Oosterbroek experimented with connected operators for the task of character segmentation [8]. Apart from these solutions, to the best of our knowledge, connected operators have not been investigated for use in document image analysis tasks.

The remainder of this paper is organized as follows; Sect. 2 briefly describes connected operators and the way it has been realized. Section 3 describes the solution we propose. Section 4 discusses the dataset we use to evaluate the proposed solution and discusses some drawbacks of the solution. Section 5 presents our conclusions.

2 Connected Operators

Connected operators are region-based filtering tools in that they act on the connected components where the image is constant (flat zones) solely by removing boundaries between these flat zones. This gives these operators very good contour preserving properties. One of the ways in which connected operators are implemented in practice is by use of a tree representation of the image [6]. The representation we make use of is the maxtree [5]. Each node \mathcal{N}_i of a maxtree represents a connected component extracted by a repeated thresholding process where for each threshold k where $k \in [0, 255]$, considering sets $X_k = \{n, \text{ such that } I[n] \geq k\}$ and $Y_k = \{n, \text{ such that } I[n] = k\}$ the node \mathcal{N}_i represents the connected components C of X such that $C \cap Y = \emptyset$. The links between nodes denote inclusion relationships. In the case of a maxtree therefore, a link from $\mathcal{N}_i^{k_a}$ to $\mathcal{N}_j^{k_b}$ exists when the component C^{k_a} from $\mathcal{N}_i^{k_a}$ contains C^{k_b} from $\mathcal{N}_j^{k_b}$ where $k_b > k_a$. The leaf nodes therefore contain the regional maxima. The processing of the image is done by means of pruning the tree and reconstructing the image from the pruned tree. Filtering with the maxtree may be thought intuitively as the grayscale image being thresholded at each value

of k , a binary filter being applied to the result and the results of all of these filters stacking up to result in the final grayscale result. We refer readers to [6] for a more elaborate description of maxtrees and algorithms for their implementation.

3 Proposed Solution

In this section we first present an outline of the proposed solution. Following this outline we present the details of the solution we propose.

The proposed solution makes use of the previously described connected operator approach combined with the more conventional Otsu thresholding. Given a grayscale image, we first construct a maxtree representation of the image, perform a bounding box attribute opening operation by pruning the tree, and reconstruct an image from the simplified tree so as to obtain a simplified image. The Otsu threshold and a metric of effectiveness of the threshold for the simplified image are then estimated. The image is then partially thresholded, in that pixels of the simplified image which fall below the estimated threshold are set to zero, subject to the condition that the effectiveness metric and the threshold are greater than previously set threshold values for each one of these two values. This partially thresholded image is then used to construct another maxtree representation of the image. Nodes of this maxtree that satisfy certain size criteria are then extracted. The image components that these nodes form are potential object candidates in the segmented image. Of these possible candidates, the object candidate that is the largest and the bounding box of which includes the other candidates is chosen as the object, which will represent a given nontext object in the segmented image. A sample result after completion of each stage of processing is presented in Fig. 1.

Next we elaborate on and discuss the proposed solution and provide some intuition on the various processing steps we employ as part of the proposed solution.

Attribute Opening

Attribute opening is achieved by the construction of a maxtree representation of the input image followed by a pruning of this tree. The pruning is done on the basis of the bounding box size attribute of the image component that a given node would form. Therefore any node with bounding box smaller than a preset minimum height, h_{min} and a preset minimum width, w_{min} gets pruned. The pixels of the pruned nodes now become owned by their parent nodes. The resulting image is often much easier to process, in the case of document images, given that it has far fewer components. In the case of document images, with well chosen parameter values, the attribute opening often has the effect of removing all the text in the image. In our implementation we set $w_{min} = N_w/12$ and $h_{min} = N_h/12$ where N_h and N_w are the image height and width, respectively. The attribute opening operation is followed by the reconstruction of the image from the pruned tree to give us the simplified image. At this point, however, we must mention that pruning of the leaves and their pixels getting owned by their parent may often result in the generation of flat-zones which

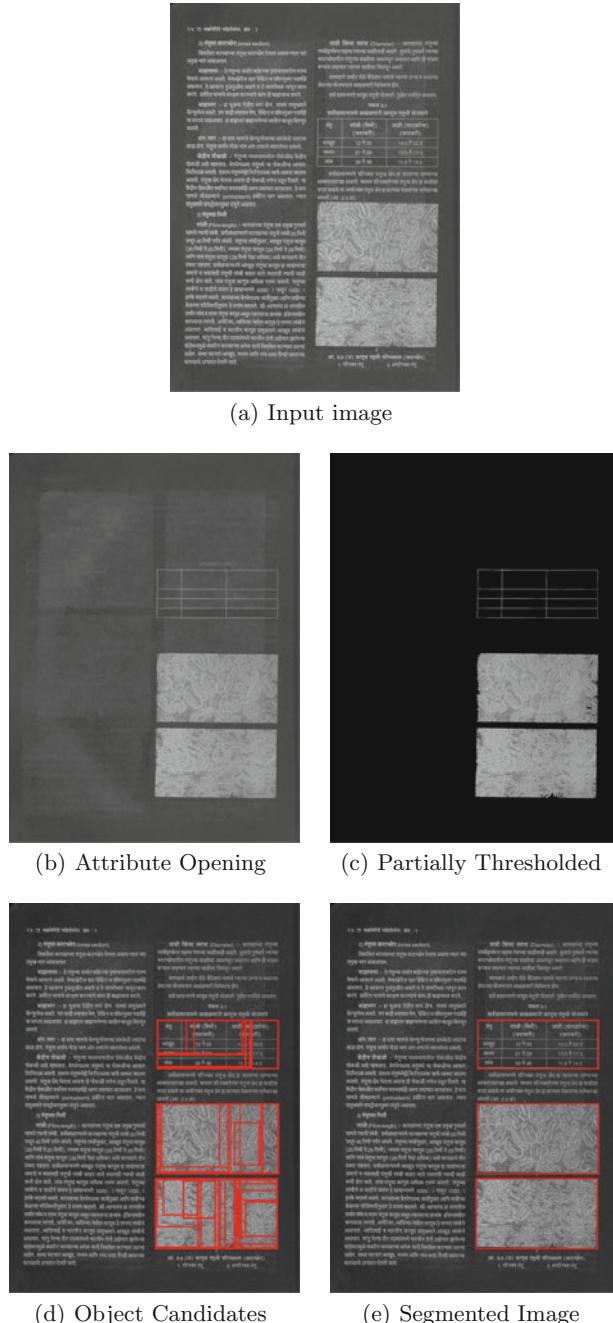


Fig. 1 A sample results at every stage of the proposed solution

were previously not present in the image, due to a merging of the boundaries of flat-zones, this is specifically so in the case of degraded document images (Fig. 1b); this may therefore interfere with the segmentation process if not dealt with.

Partial Thresholding

The image histograms of the simplified images are often bimodal where the objects which have been retained represent maxima and form a separate cluster from the cluster formed by the large background areas that are formed as a result of the attribute opening. A separation of the two clusters of the histogram therefore allows for a further simplification of the image. The Otsu thresholding method [4], which is ideally suited to this task, is therefore employed here so as to estimate a threshold value k^* , at which the clusters are optimally separated. In addition, an estimate of the effectiveness of the optimal threshold is obtained by calculation of the effectiveness metric $\eta(k^*)$ (also denoted by η^*); given by:

$$\eta(k^*) = \frac{\sigma_B^2(k^*)}{\sigma_T^2} \quad (1)$$

where σ_B^2 denotes the between class variance and σ_T^2 denotes the total variance of graylevels. And here $1 \leq \eta^* \leq 0$, with $\eta^* = 1$ denoting an image with two separate graylevels and $\eta^* = 0$ denoting an image with a single graylevel. We make use of these values to decide the ability of the subsequent steps to extract objects accurately from the image. If the estimated values of k^* and η^* are below certain threshold values η_T^* and k_T^* we mark the image as either having no objects or as being too noisy to be segmented. The justification for this comes from the observations that in an image with no nontext objects, following the attribute opening operation, the histogram will primarily be unimodal and have a low value of η^* and k^* ; this is evidenced by the (Table 1) mean values of η^* before and after the attribute opening for the different image types in our dataset. This process is one that is necessary specifically in the case of degraded document images where following the attribute opening, the flatzones formed by the degradations might qualify as object candidates and subsequently be segmented as objects. An illustration of these flatzones formed by the degradations can be observed in Fig. 1b. In our implementation $\eta_T^* = 0.6$ and $k_T^* = 0.3$. The values of these thresholds were determined experimentally for our dataset. The images with η^* and k^* values above the threshold values (η_T^* and k_T^*) are partially thresholded as:

Table 1 Mean values of η^* for different input image types

		Degraded	Nondegraded
Objects absent	Before	0.84	0.93
	After	0.46	0.59
Objects present	Before	0.83	0.90
	After	0.71	0.86

$$I_{PT}(x, y) = \begin{cases} I_{AO}(x, y) & I_{AO}(x, y) \geq k^* \\ 0 & I_{AO}(x, y) < k^* \end{cases} \quad (2)$$

where \mathbf{I}_{AO} is the image obtained after attribute opening and \mathbf{I}_{PT} is the partially thresholded image. The partial thresholding allows the proposed solution to deal with degraded document images while not influencing the performance on nondegraded document images.

Object Extraction

The partially thresholded image is once again used to build the maxtree representation. This representation allows us to extract objects which satisfy certain criteria and filter these further so as to obtain the objects which will form the final segmentation. In our case, we make use of the bounding box sizes for this extraction and filtering. Therefore, from the tree representation we extract the nodes which are such that $h_{min} \leq bb_h^i \leq h_{max} \wedge w_{min} \leq bb_w^i \leq w_{max}$; where bb_h^i and bb_w^i are the bounding box height and width of a given sub-image component formed by a particular node. This step will typically yield a number of possible candidate objects (Fig. 1d). At this stage we obtain the objects, the bounding boxes of which are the largest with all other candidate objects enclosed within the larger bounding box. This largest bounding box is marked as being the object to obtain the segmented image. In the case of multiple such largest bounding boxes being present we choose the first bounding box which satisfies the above criteria. In our implementation $h_{max} = 0.9 \times N_h$ and $w_{max} = 0.9 \times N_w$.

At this stage we must mention however, that attributes other than the bounding box could have been explored; attributes such as the number of children nodes of a given node, the topological height of the node or the area of a given node [7]. The reason we choose to employ the bounding box attribute over others is its inherent simplicity and the fact that it is a less abstract attribute than attributes derived from the tree representation of the image and therefore better suited to a task such as segmentation. However, we also concede that the use of more abstract attributes to segmentation will require further investigation.

4 Results and Discussion

In this section we describe the nature of the dataset on which the proposed solution was tested, next we describe briefly the performance of the proposed solution, this is followed by a discussion of some of the shortcomings of the proposed solution.

The proposed solution was tested on a dataset consisting of a total of 545 images. One subset of the dataset consisted of 291 grayscale scanned images of technical documents. This subset of the dataset was representative of degraded document images. The images in this dataset were such that they suffered majorly from bleed through distortions. The other subset of the dataset consisted of 254 high

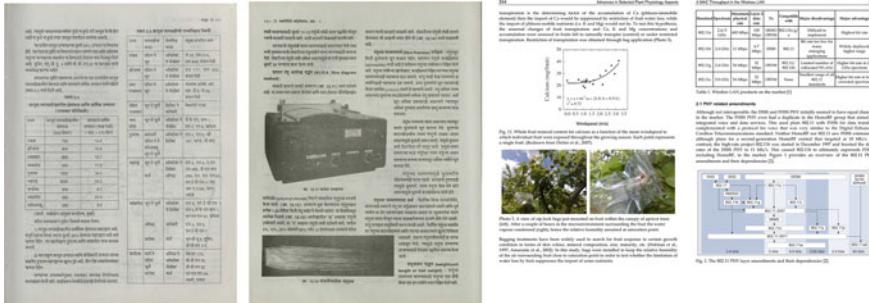


Fig. 2 A set sample images from our dataset

Table 2 Percentage accuracy values of the proposed solution

	Degraded	Nondegraded
Accurately segmented nontext	93.59	90.30
Unsegmented or under-segmented nontext	6.41	9.70

quality images of technical documents. This dataset however suffered from no distortions. The nontext objects present in our dataset consisted majorly of tables, halftone images, line drawings and graphs; these are the nontext objects which we aimed to segment with the proposed solution. The layouts of the images were however largely simple in that the objects consisted majorly of rectangular layouts. Sample images from the datasets are presented in Fig. 2. In our evaluation of the performance of the proposed solution we maintain the separation of our dataset into the two classes of degraded and nondegraded input document images. In addition, we also make a representative set of result images available for examination.¹

The proposed solution was evaluated by means of determining the number of nontext objects, which the method was able to segment and mark as nontext and the number of nontext objects, which the method was either not able to segment at all or one which was under segmented. Due to the way the method is formulated, in our dataset, the method never marks a text region as being a nontext object or oversegments a given nontext object. The results of this evaluation are presented in Table 2. Note however that here the evaluation presents overall accuracy rather than that of individual types of nontext objects. In practice, however, the proposed solution is able to segment halftone images, tables, line-drawings, and graphs. Sample results which depict these images are shown in Fig. 3. Aside from the nontext types mentioned above, since the proposed solution is essentially able to pick out boxes which satisfy certain criteria, any object enclosed in a box will be segmented.

¹Results: <https://github.com/MSheshera/TNTClassify>.

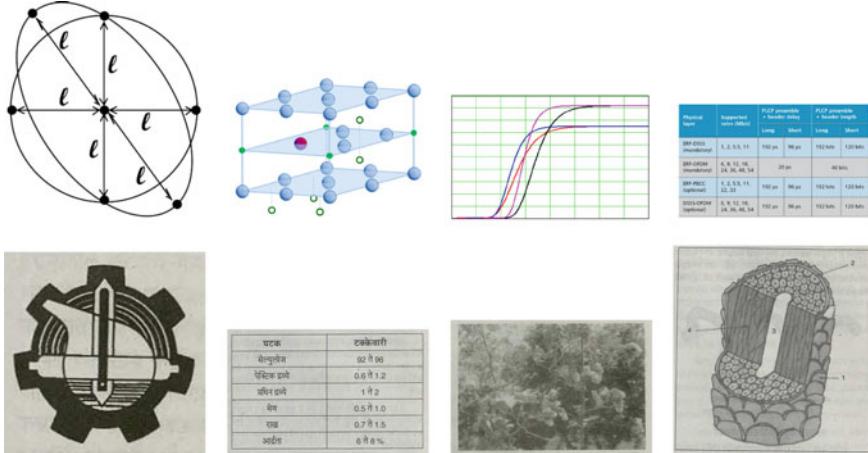


Fig. 3 Object types which get segmented accurately

परमाणु	दर्शकाती
सिंगलेट	92 ± 96
ट्रिप्लेट इन्वे	0.6 ± 1.2
सिंगल इन्वे	1 ± 2
इन्वे	0.5 ± 1.0
एन्वे	0.7 ± 1.5
डिप्लो	9 ± 9%

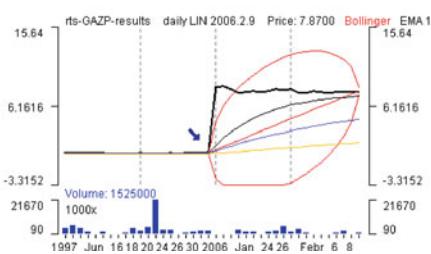


Fig. 4 Failure cases; Objects in the semantic sense alone

The proposed solution does have certain shortcomings, we highlight those next. Due to the nature of connected operators, where objects which are connected can be segmented, objects which are considered objects in a semantic sense while not being connected physically do not get segmented by the proposed solution. Effective segmentation of such objects will likely need to involve some form of geometric reasoning in addition to the connected operator approach or involve an extension to the way in which connectivity is defined for connected operators [8]. Examples of such objects may include tables, which are formed purely by horizontal lines or graphs such as the ones shown in Fig. 4.

In addition to the images where objects aren't connected, the proposed solution also fails to segment images where the nontext objects are poorly contrasted with the background. This shortcoming can be attributed to the partial thresholding step, and our rejection of images as being either too noisy or not containing objects by use of the η^* and k^* values as described in Sect. 3. This step however is one that is necessary in the case of degraded document images.

5 Conclusion

In this work we propose a novel method of segmenting non-text objects from grayscale document images. The proposed method employs connected operators implemented as maxtrees and combines this approach with Otsu thresholding. We stress at this point that, since the maxtree is a region-based representation of the image it allows for tasks such as object segmentation. Although this work has assumed simple layouts of input images, the proposed solution can be extended to deal with more complex layouts. This forms one of the future areas of work. In addition, as has been mentioned, the bounding box-based criteria used in this solution can be extended to include other criteria such as topological heights of the node, the number of children of the sub-tree rooted at node or other attributes [7], which may aid in effective segmentation of different kinds of objects. In addition, other forms of trees which may potentially aid in segmentation, such as an inclusion tree [2] could be investigated for suitability to segmentation.

References

1. Lazzara, G., Géraud, T., Levillain, R.: Planting, growing and pruning trees: Connected filters applied to document image analysis. In: Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS). pp. 36–40. IAPR, Tours, France (April 2014)
2. Monasse, P., Guichard, F.: Fast computation of a contrast-invariant image representation. *Trans. Img. Proc.* 9(5), 860–872 (May 2000)
3. Naegel, B., Wendling, L.: Document binarization based on connected operators. vol. 0, pp. 316–320. IEEE Computer Society, Los Alamitos, CA, USA (2009)
4. Otsu, N.: A threshold selection method from gray-level histograms. *Systems, Man and Cybernetics, IEEE Transactions on* 9(1), 62–66 (January 1979)
5. Salembier, P., Oliveras, A., Garrido, L.: Antiextensive connected operators for image and sequence processing. *Image Processing, IEEE Transactions on* 7(4), 555–570 (April 1998)
6. Salembier, P., Wilkinson, M.: Connected operators. *Signal Processing Magazine, IEEE* 26(6), 136–157 (November 2009)
7. Silva, A., de Alencar Lotufo, R.: New extinction values from efficient construction and analysis of extended attribute component tree. In: *Computer Graphics and Image Processing, 2008. SIBGRAPI '08. XXI Brazilian Symposium on.* pp. 204–211 (October 2008)
8. Wilkinson, M., Oosterbroek, J.: Mask-edge connectivity: Theory, computation, and application to historical document analysis. In: *Pattern Recognition (ICPR), 2012 21st International Conference on.* pp. 1334–1337 (November 2012)

Non-regularized State Preserving Extreme Learning Machine for Natural Scene Classification

Paheding Sidike, Md. Zahangir Alom, Vijayan K. Asari
and Tarek M. Taha

Abstract Scene classification remains a challenging task in computer vision applications due to a wide range of intraclass and interclass variations. A robust feature extraction technique and an effective classifier are required to achieve satisfactory recognition performance. Herein, we propose a nonregularized state preserving extreme learning machine (NSPELM) to perform scene classification tasks. We employ a Bag-of-Words (BoW) model for feature extraction prior to performing the classification task. The BoW feature is obtained based on a regular grid method for point selection and Speeded Up Robust Features (SURF) technique for feature extraction on the selected points. The performance of NSPELM is tested and evaluated on three standard scene category classification datasets. The recognition accuracy is compared with the standard extreme learning machine classifier and it shows that the proposed NSPELM algorithm yields better accuracy.

Keywords Scene classification · Non-regularization · Extreme learning machine · Bag-of-words

P. Sidike (✉) · M.Z. Alom · V.K. Asari · T.M. Taha

Department of Electrical & Computer Engineering, University of Dayton,
Dayton 45469, USA

e-mail: pahedings1@udayton.edu

M.Z. Alom

e-mail: alomm1@udayton.edu

V.K. Asari

e-mail: vasari1@udayton.edu

T.M. Taha

e-mail: ttaha1@udayton.edu

1 Introduction

As a human, we are capable of distinguishing, categorizing, and associating objects without too much efforts. This recognition and interpretation ability of a human is the key foundation for any higher level of processing. This concept can be synonymous as clustering, classification, and prediction in the field of machine learning. Most of these algorithms intend to mimic human brain to perform object recognition or other diverse tasks. Researchers have conducted extensive research of the functionality of brain associated with vision, speech, sensing, etc.

In the near future, it is envisaged that the machines would match or exceed the capabilities of human visual system performance. It is interesting to note that we receive tremendous amount of different types of data (e.g., audio, visual, etc.) in every moment, and our brain can rapidly process this data for extracting and interpreting information. It is a basic process for us to recognize an object in a scene as a table, a ball, a book, etc. Here, we attempt to model the human visual system behavior for interpreting a scene and classify it to various categories such as a coast, street, etc. Figure 1 shows some sample images of scene categories.

To perform a recognition task, training data set is required and this data contains sample images from all the scene categories that need to be classified. Then an advanced learning algorithms such as Support Vector Machine (SVM) [4] or Extreme Learning Machine (ELM) [11] are used to train the system. This is a basic strategy to enable us to categorize a scene. The main tasks involved in this work is listed as follows:

- Development of a nonregularized state preserving ELM (NSPELM) with Bag-of-Words (BoW) [6] features for scene classification.
- Evaluation of the individual performance of ELM and NSPELM using BoW model on different benchmark scene classification datasets.

The rest of the paper is organized as follows. In Sect. 2, related work on scene classification is provided. The mathematical details of ELM and NSPELM are given in Sect. 3. Discussion on the datasets and test results are provided in Sect. 4. Finally, the conclusion is drawn in Sect. 5.



Fig. 1 Examples of scene categories. (Images are from the dataset provided in [15])

2 Related Work

Differentiating scene category is a challenging task in the area of computer vision. Although human vision is still much powerful than computers in many object classification tasks, recent works have showed the great progress in enhancing capability of computers to perform scene categorization. Scene representation and feature learning are two main tasks for scene category classification. Low level features of images, such as color histogram and object orientation features [8, 17], can be used for scene representation, although they may not provide preferred results. Studies in [19] presented a local semantic concept for natural scene classification, whereas Anna et al. [3] improved their method by using a hybrid generative scene representation approach. Intermediate representations [6, 13, 14] have shown better performance than low level feature representation. These methods often make use of low level local features and construct them in holistic manner. For instance, study in [6] introduced a new type of scene representation method which uses a collection of local features and build them as visual words through clustering. This method is referred to the BoW. Lazebnik et al. [14] improved the BoW method by encoding spatial information through partitioning an image into small grid cells.

In this paper, we present a new scene classification scheme that is developed using the BoW features and a biologically inspired ELM model. The classical ELM can be viewed as a Single-Hidden Layer Feedforward Neural Network (SLFN). Recent studies showed many successful applications of ELM in both effectiveness and efficiency way. However, ELM produces a fluctuated accuracy when there are several repeated experimental trials/runs. In contrast, SPELM presented in [1] shows significant advantages with better recognition rate and faster processing speed over the other type of ELMs, such as standard ELM and regularized ELM (RELM) [5, 10]. In this work, we introduce a NSPELM classifier with BoW features for scene recognition.

3 Methodology

In this section, we first provide a brief introduction of ELM formulation, and then introduce our NSPELM concept.

3.1 *Extreme Learning Machine (ELM)*

ELM is a new type of SLFN which utilizes randomly generated parameters in hidden layer instead of iteratively tuning network parameters. This makes it different from traditional neural networks and enable ELM to produce smaller training error with fast performance [7, 9–11].

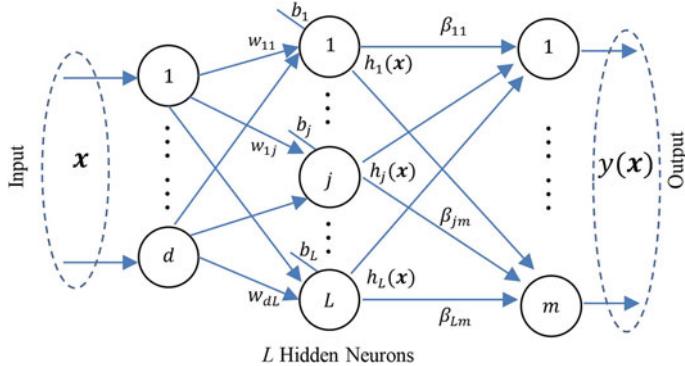


Fig. 2 ELM structure

Figure 2 shows a sample structure of ELM. The output function of an ELM can be obtained by [11]

$$y(\mathbf{x}) = \sum_{j=1}^L \beta_j h_j(\mathbf{x}) = h(\mathbf{x})\boldsymbol{\beta}. \quad (1)$$

where \$y(\mathbf{x})\$ is the output of the neural network, \$\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_L]\$ is the \$L \times m\$ weight matrix between hidden layer of \$L\$ nodes and output layers of \$m\$ nodes. \$h(\mathbf{x})\$ is the hidden layer activation function with respect to the input \$\mathbf{x}\$, and \$h_j(\mathbf{x})\$ corresponds to the \$j^{th}\$ hidden node. For \$n\$ data samples with \$d\$ input neurons, \$h_j(\mathbf{x}_i)\$ can be expressed as

$$h_j(\mathbf{x}_i) = G(\mathbf{w}_j, b_j, \mathbf{x}_i), \mathbf{w}_j \in \mathbb{R}^d, \mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \dots, n \quad (2)$$

where \$\mathbf{w}_j\$ represents the \$j^{th}\$ weight vector from the input layer to the \$j^{th}\$ hidden neuron and \$b_j\$ is the bias of \$j^{th}\$ hidden neuron. \$G(\mathbf{w}_j, b_j, \mathbf{x}_i)\$ is a nonlinear piecewise continuous function (e.g., sigmoid function). ELM aims to obtain the smallest training error by minimizing \$\|h(\mathbf{x}_i)\boldsymbol{\beta} - \mathbf{T}_i\|\$. For \$n\$ input samples, a more compact form can be written as

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{T} \quad (3)$$

where \$\mathbf{T} \in \mathbb{R}^{n \times m}\$ is the desired output matrix in training samples and \$\boldsymbol{\beta} \in \mathbb{R}^{L \times m}\$. \$\mathbf{H}\$ refers to the hidden layer output matrix which maps the data from \$d-\$dimensional input space to \$L-\$dimensional feature space, expressed as

$$\mathbf{H} = \begin{bmatrix} h(\mathbf{x}_1) \\ \vdots \\ h(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} G(\mathbf{w}_1, b_1, \mathbf{x}_1) & \dots & G(\mathbf{w}_L, b_L, \mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ G(\mathbf{w}_1, b_1, \mathbf{x}_n) & \dots & G(\mathbf{w}_L, b_L, \mathbf{x}_n) \end{bmatrix} \quad (4)$$

β can be estimated by a least squares solution as [11]

$$\hat{\beta} = \mathbf{H}^\dagger \mathbf{T} \quad (5)$$

where \mathbf{H}^\dagger is the Moore–Penrose generalized inverse of matrix \mathbf{H} [12, 18] and $\mathbf{H}^\dagger = \mathbf{H}^T (\mathbf{H}\mathbf{H}^T)^{-1}$ [10]. Hence, the prediction value matrix, denoted as \mathbf{Y} , for all input samples is expressed by

$$\mathbf{Y} = \mathbf{H}\hat{\beta} = \mathbf{H}\mathbf{H}^\dagger \mathbf{T} \quad (6)$$

The error matrix can be described as

$$\varepsilon = \|\mathbf{Y} - \mathbf{T}\|^2 = \|\mathbf{H}\mathbf{H}^\dagger \mathbf{T} - \mathbf{T}\|^2 \quad (7)$$

The above-mentioned SLFN-based learning method is called ELM.

3.2 Non-Regularized State Preserving ELM (NSPELM)

ELM is considered to be an efficient and effective SLFN, however, many experimental results, such as in [16, 20], have shown that ELM yields inconsistent accuracy during different trials of experiments although using the same parameters such as the number of hidden nodes. Furthermore, it is observed that accuracy inconsistency occurs more than once during subsequent trials [1]. Accordingly, we observed that if weights and biases that contribute to better accuracy during numbers of trials are preserved, the overall performance of ELM would be improved. In other words, we refer this strategy as a monotonically increasing learning through a number of trials. In [1], we proposed a regularized version of SPELM for face recognition and the results showed superior performance when compared to ELM and RELM. In this work, we propose a nonregularized form of SPELM (refer as NSPELM) for scene classification. The basic concept of SPELM or NSPELM is that a higher accuracy with respect to relevant parameters (i.e., weights and biases) are preserved during successive trials such that the resultant outputs will be improved or retained. The same process will be repeated until completing all required trials. The mathematical details of NSPELM are described below.

Assume that the experiments are repeated N times, or there are N number of trials, that introduces different states of the ELM network. Let S_t be the t -th state, then the accuracy for S_t is denoted as A_{S_t} . The hidden layer output matrix \mathbf{H} in S_t state is denoted as \mathbf{H}^{S_t} , and the activation function \mathbf{G} and the weight vector β are represented by \mathbf{G}^{S_t} and β^{S_t} , respectively. Accordingly, the output function can be written as

$$y^{S_t}(\mathbf{x}_i) = \sum_{j=1}^L \beta_j^{S_t} G^{S_t}(\mathbf{w}_j^{S_t}, b_j^{S_t}, \mathbf{x}_i^{S_t}) \quad (8)$$

Then the weights $w_j^{S_t}$ and the bias $b_j^{S_t}$ are updated according to current state accuracy A_{S_t} and the immediate previous state accuracy $A_{S_{t-1}}$. The updating rule is defined by Eqs. (9) and (10).

$$w_j^{S_t} = \begin{cases} w_j^{S_t} & , \text{ if } A_{S_t} > A_{S_{t-1}} \\ w_j^{S_{t-1}} & , \text{ otherwise} \end{cases} \quad (9)$$

$$b_j^{S_t} = \begin{cases} b_j^{S_t} & , \text{ if } A_{S_t} > A_{S_{t-1}} \\ b_j^{S_{t-1}} & , \text{ otherwise} \end{cases} \quad (10)$$

If the present state variables (i.e., weights and biases) provide better accuracy, NSPELM attempts to find the smallest training error by minimizing errors between the predicted value and the labeled value: $\|\mathbf{G}^{S_t} \boldsymbol{\beta}^{S_t} - \mathbf{T}\|$. Similar to ELM, this can be solved by a least squares solution as in Eq. (5) and the output function for input data is computed by

$$\mathbf{Y}^{S_t} = \mathbf{H}^{S_t} \hat{\boldsymbol{\beta}}^{S_t} \quad (11)$$

The NSPELM can be generalized to a regularized version as in [1]. In this work, we explore the performance of NSPELM for scene category recognition.

4 Scene Classification Using NSPELM

4.1 Dataset Description

The scene classification method using NSPELM is evaluated on the following three publicly available datasets. Figure 3 shows example images from these datasets.

8-scene categories dataset [15]: This dataset includes 8 outdoor scenes such as are coast, mountain, forest, open country, street, inside city, tall buildings, and highways. There are 2688 color images in total, with the number of images in each category ranging from 260 to 410, and each one is with a size of 256×256 pixels.

13-scene categories dataset [6]: This dataset contains 3859 gray-scale images classified into 13 natural scene categories, in which eight categories are from [15] and the other five categories are offices, kitchen, bedroom, suburb, and living rooms. The average size of each image is approximately 300×250 pixels.

15-scene categories dataset [14]: This dataset consists of 4485 gray-scale images divided into 15 categories where 13 categories are from [6] and the two additional categories are industrial and store.



Fig. 3 Sample images from the three datasets. The first two rows are from 8-scene category dataset, row three is from 13-scene category dataset, and row four is from 15-scene category dataset

4.2 Results and Discussion

For each category in the datasets, we randomly chose 100 images for training and the rest for testing. All experiments are repeated 20 times and the average accuracy is reported. For the parameters in BoW model, we use a regular grid method with an 8×8 step size for interest point selection, and a speeded up robust features (SURF) [2] technique is employed for extracting features. Then a codebook with 200 visual words is constructed by k -means clustering. The vocabulary size is set equal for all the datasets. We use a ‘*sine*’ function as the activation function with 8000 hidden neurons for both ELM and NSPELM.

In the experiments, we compare our proposed BoW+NSPELM (i.e., BoW is used for feature extraction and NSPELM is used for classification) scheme with BoW+ELM. Classification results for 8-scene, 13-scene, and 15-scene datasets are shown in Figs. 4, 5, and 6, respectively. Figure 4 shows the classification results on the 8-scene dataset. It is obvious that the proposed BoW+NSPELM outperforms BoW+ELM in most of the trials and yields higher average recognition accuracy as shown in Table 1. A noticeable finding is that the recognition accuracy of the BoW+NSPELM gradually increases as the number of trials increases. This is because of the inherent character of NSPELM in which it always preserves better weights and biases for achieving a better or equal recognition accuracy in each trial.

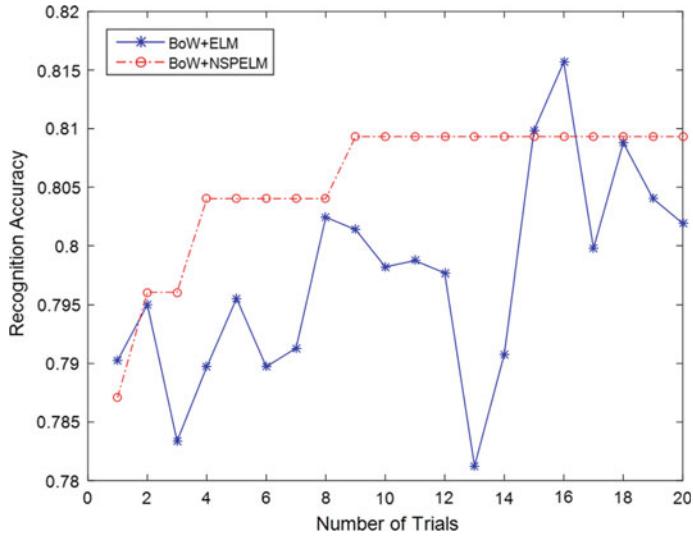


Fig. 4 Recognition accuracy for 8-scene dataset

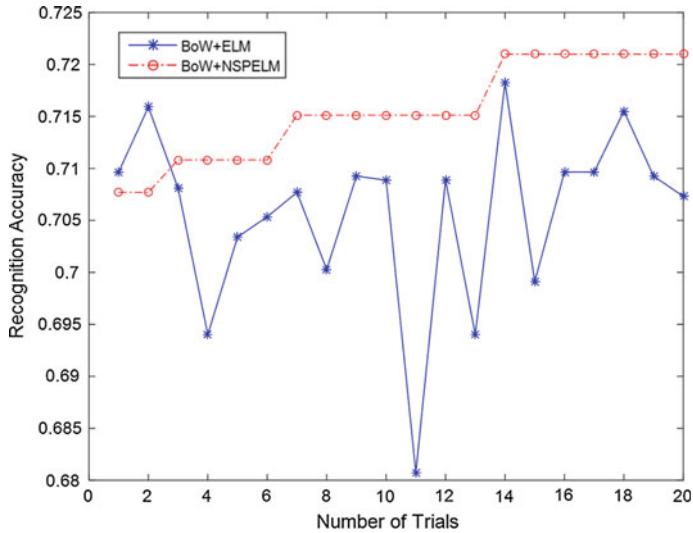


Fig. 5 Recognition accuracy for 13-scene dataset

Considering the recognition accuracy versus number of trials in Figs. 4, 5, 6, it is evident that our NSPELM illustrates consistently improving behavior throughout scene recognition process. In addition, comparing the classification performance on 13-scene and 15-scene categories, accuracy of both ELM and NSPELM drops down

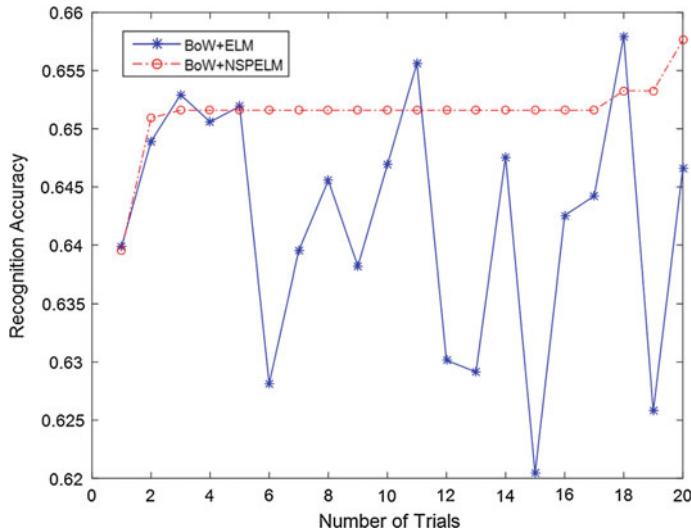


Fig. 6 Recognition accuracy for 15-scene dataset

Table 1 Performance comparison (The best accuracy is shown in bold)

Methods	8-category	13-category	15-category
BoW + ELM	79.73	70.57	64.21
BoW + NSPELM	80.56	71.56	65.14

due to more complex scene categories compared to the 8-scene dataset. However, NSPELM based method still provides the better classification accuracy as shown in Figs. 5 and 6.

5 Conclusion

In this paper, we presented an effective scene classification method by incorporating BoW model with NSPELM. The proposed NSPELM's recognition characteristics illustrate the consistency and stability in learning the input features for classification as opposed to inconsistent behavior of the conventional ELM. Experimental results on three widely used data sets showed that our approach yields better classification rate compared to BoW + ELM. Our future research will employ more effective feature extraction techniques with a kernel based NSPELM to further improve scene classification accuracy.

References

1. Alom, M., Sidike, P., Asari, V.K., Taha, T.M.: State preserving extreme learning machine for face recognition. In: 2015 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2015)
2. Baya, H., Essa, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding 110(3), 346–359 (2008)
3. Bosch, A., Zisserman, A., Muoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4), 712–727 (2008)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
5. Deng, W., Zheng, Q., Chen, L.: Regularized extreme learning machine. In: IEEE Symposium on Computational Intelligence and Data Mining. CIDM'09. pp. 389–395. IEEE (2009)
6. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. vol. 2, pp. 524–531. IEEE (2005)
7. Feng, G., Huang, G.B., Lin, Q., Gay, R.: Error minimized extreme learning machine with growth of hidden nodes and incremental learning. IEEE Transactions on Neural Networks 20(8), 1352–1357 (2009)
8. Gorkani, M.M., Picard, R.W.: Texture orientation for sorting photos “at a glance”. In: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. vol. 1, pp. 459–464. IEEE (1994)
9. Huang, G.B., Chen, L., Siew, C.K.: Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Transactions on Neural Networks 17(4), 879–892 (2006)
10. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42(2), 513–529 (2012)
11. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing 70(1), 489–501 (2006)
12. Johnson, C.R.: Matrix theory and applications. American Mathematical Soc. (1990)
13. Kwitt, R., Vasconcelos, N., Raswasia, N.: Scene recognition on the semantic manifold. In: Computer Vision–ECCV 2012, pp. 359–372. Springer (2012)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 2169–2178. IEEE (2006)
15. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision 42(3), 145–175 (2001)
16. Peng, Y., Wang, S., Long, X., Lu, B.L.: Discriminative graph regularized extreme learning machine and its application to face recognition. Neurocomputing 149, 340–353 (2015)
17. Pietikäinen, M., Mäenpää, T., Viertola, J.: Color texture classification with color histograms and local binary patterns. In: Workshop on Texture Analysis in Machine Vision. pp. 109–112. Citeseer (2002)
18. Rao, C.R., Mitra, S.K.: Generalized inverse of matrices and its applications, vol. 7. Wiley New York (1971)
19. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision 72(2), 133–157 (2007)
20. Wang, Y., Cao, F., Yuan, Y.: A study on effectiveness of extreme learning machine. Neurocomputing 74(16), 2483–2490 (2011)

A Local Correlation and Directive Contrast Based Image Fusion

Sonam and Manoj Kumar

Abstract In this paper, a local correlation and directive contrast-based multi-focus image fusion technique is proposed. The proposed fusion method is conducted into two parts. In first part, Discrete Wavelet Packet Transform (DWPT) is performed over the source images, by which low and high frequency coefficients are obtained. In second part, these transformed coefficients are fused using local correlation and directive contrast-based approach. The performance of the proposed method is tested on several pairs of multi-focus images and compared with few existing methods. The experimental results demonstrate that the proposed method provides better results than other existing methods.

Keywords Image fusion · Discrete wavelet packet transform · Directive contrast · Peak-signal-to-noise-ratio · Correlation coefficient · Standard deviation

1 Introduction

In recent years, image fusion has become an important and useful tool to enhance the visual perception of images [1, 2]. It has been used in various fields, such as medical diagnosis, surveillance, robotics, remote sensing, biometrics, and military, etc. Image fusion provides an effective way to integrate the information of multiple source images and produce a single fused image containing enhanced description of scene without introducing any artifacts. Fusion techniques have been applied on several pairs of images, such as multi-focus images, multi-spectral images, visible images, infrared images, medical images, and remote sensing images [3], etc. The images with different focuses contain less information than focused images.

Sonam (✉) · M. Kumar
Department of Computer Science, Babasaheb Bhimrao Ambedkar University,
Lucknow, India
e-mail: sonam870115@gmail.com

M. Kumar
e-mail: mkjnuitr@gmail.com

Therefore, to get an image with entire scene in focus from multi-focus images is a difficult task. This problem can be solved by multi-focus image fusion. The objective of multi-focus image fusion is to combine all relevant information from multiple source images having different focuses and to produce a single image with entire scene in focus.

Image fusion can be broadly classified into three categories. This categorization includes pixel-level [4], feature-level [5], and decision-level [6] image fusion. Pixel-level image fusion defines the process of combining pixel by pixel information from all source images into a single fused image. It preserves more source image information and represents the lowest level of image fusion. Feature-level image fusion, fuses the features such as, edges, color, and texture that have already been extracted from individual source images. Finally, decision-level is the highest level of image fusion, which fuses the results obtained from multiple algorithms to achieve a final fused image. Among these, pixel-level image fusion is a popular method because it contains the advantages of preserving source information, computationally easy to implement, and time efficient. Image fusion techniques can be performed in spatial and transform domains. The spatial domain-based methods directly deals with the image pixel values to obtain a desired result. Generally, several undesired effects are produced by spatial domain methods, such as contrast reduction and distortion. To address these problems, various methods based on transform domain have been explored. Transform domain methods produce good results in less computation and in less time, therefore it is more suitable for fusion purposes. At different resolutions, transform domain methods contain unique information and also provide directional information in all decompositions. The basic aim of transform domain-based methods is to perform multiresolution decomposition on each source image and generate a composite image by combining all these decompositions using certain fusion rules. Finally, inverse multiresolution transform is performed to reconstruct the fused image [3]. Usually, transform domain-based methods use transform methods like, Discrete Wavelet Transform (DWT) [7, 8], Dual Tree Complex Wavelet Transform (DT-CWT) [9], contourlet transform [10], and Nonsubsampled Contourlet Transform (NSCT) [11], etc. In [12], a DWT-based fusion is performed using maximum selection rule. This scheme selects the largest absolute wavelet coefficients at each location from the source images and integrates into a single fused image. In PCA-based fusion technique [13], source images are transformed into uncorrelated images and principal components are evaluated from eigenvalues. These principal components represent variance of the pixels contribute to the weights used for fusion process. Wavelet Packet Transform (WPT) [14] domain based fusion technique is used to increase the contrast of the fused image. The low frequency coefficients are combined using median rule, and high frequency coefficients are combined through directive contrast method in wavelet packet transform domain.

In this paper, an effective multi-focus image fusion scheme is proposed. Due to the advantages of easy implementation and time efficiency, this paper concentrates on pixel-level image fusion. For image fusion, it is assumed that the source images are already registered. The objective of this paper is to fuse two multi-focus (defocused) images of same scene into a single image with entire scene in focus and

better visual information. The key challenge of multi-focus image fusion problem is to evaluate the blurring of each image and then select more relevant information from the processed image to increase the visual quality of image. In this paper, a local correlation and directive contrast-based image fusion method in discrete wavelet packet transform domain is proposed to achieve the final fused image.

The rest of this paper is organized as follows; Sect. 2 describes the basic concepts of wavelet packet transform and directive contrast. In Sect. 3, the proposed fusion method is explained. Experimental results and discussions are given in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Basic Concepts of Wavelet Packet Transform and Directive Contrast

In this paper, wavelet packet transform and directive contrast are used for multifocus image fusion. In this section, the basic theories of WPT and directive contrast are discussed.

2.1 Wavelet Packet Transform

The Wavelet Packet Transform (WPT) is a generalization of the DWT and provides a more flexible tool for time-scale analysis of the data [15, 16]. It retains all the properties of wavelet transform because the wavelet basis is in the repertoire of bases available with the wavelet packet transform [14]. For first level decomposition, when DWT is applied over the image, it decomposes into low and high frequency subbands. Where, the low frequency subband is usually referred as the approximation part and the high frequency subbands as the detail parts containing horizontal, vertical, and diagonal information. For the second level decomposition, low frequencies are further decomposed into another set of low and high frequency subbands and the process can be repeated upto the desired level [17], as given in Fig. 1a. During the WPT decomposition, process is iterated on both low and high frequency subbands shown in Fig. 1b. Hence, wavelet packet transform decomposes the frequency space into more number of frequency subbands and provide better frequency localization of the signals. Finally, a fused image is obtained by performing inverse WPT.

2.2 Directive Contrast

The concept of contrast of an image is developed by Toet et al. [18]. Further in [19], it was defined as:

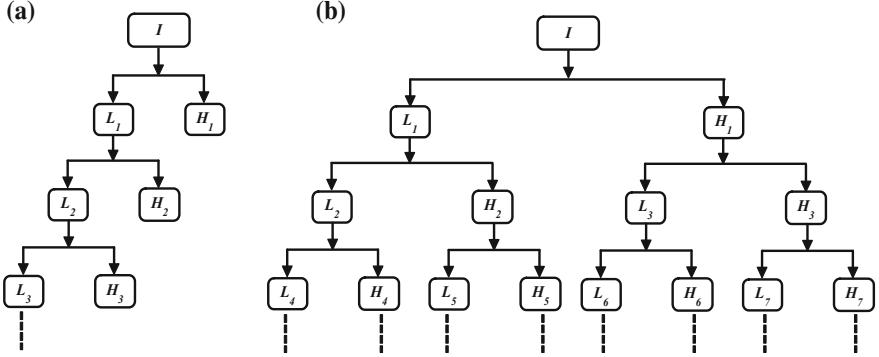


Fig. 1 **a** Wavelet decomposition; **b** wavelet packet decomposition

$$R = (L - L_B)/L_B = L_H/L_B \quad (1)$$

where, L represents the local grey level, L_B local brightness of the background which corresponds to local low frequency component, and $L_H = L - L_B$ corresponds to local high frequency components.

In our work, directive contrast is applied on high frequency subbands of Discrete Wavelet Packet Transform (DWPT). Where, DWPT decomposes the image into four frequency subbands, one low frequency component, and three high frequency components. The level wise different high frequencies are selected for finding directive contrast. Therefore, three directional contrasts: horizontal, vertical, and diagonal are obtained. Thus, a directional contrast for l -level DWPT can be defined as [14]:

$$\text{Horizontal contrast} : R_{l,p}^H = \frac{H_{l,p}}{A_{l,p}} \quad (2)$$

$$\text{Vertical contrast} : R_{l,p}^V = \frac{V_{l,p}}{A_{l,p}} \quad (3)$$

$$\text{Diagonal contrast} : R_{l,p}^D = \frac{D_{l,p}}{A_{l,p}} \quad (4)$$

where, $1 \leq p \leq 2^{l+1}$.

3 Proposed Method

In this section, the proposed fusion method is described in detail. For multi-focus image fusion, two multi-focus images are required which are considered as source images in our method. Let us consider, each source images is of size $m \times n$. In the

proposed work, DWPT is applied upto l -level decomposition over the source images X and Y . It decomposes X and Y images into one low ($A_{l,p}^X, A_{l,p}^Y$) and three high frequency coefficients ($H_{l,p}^X, V_{l,p}^X, D_{l,p}^X$, $(H_{l,p}^Y, V_{l,p}^Y, D_{l,p}^Y)$), respectively, as discussed in Sect. 2.1. The low and high frequencies obtained by DWPT are fused as given in following subsections:

3.1 Fusion of Low Frequency Coefficients

The obtained low frequency coefficients ($A_{l,p}^X, A_{l,p}^Y$) are fused using local correlation-based approach. In this method, a block wise correlation coefficients are computed from these low frequencies using 8×8 block size. These correlation coefficients [20] are computed using the following formula:

$$C = \frac{\sum_{i=1}^m \sum_{j=1}^n (A_{XX} - \mu(A_{XX})) \cdot (A_{YY} - \mu(A_{YY}))}{\sqrt{\sum_{i=1}^m \sum_{j=1}^n (A_{XX} - \mu(A_{XX}))^2 \cdot \sum_{i=1}^m \sum_{j=1}^n (A_{YY} - \mu(A_{YY}))^2}} \quad (5)$$

where $A_{XX} = A_{l,p}^X(i,j)$ and $A_{YY} = A_{l,p}^Y(i,j)$.

$\mu(A_{l,p}^X(i,j))$ and $\mu(A_{l,p}^Y(i,j))$ represent mean values of their respective low frequency components. In correlation-based fusion strategy, obtained correlation value (C) is compared with the threshold value (T). Here, in this work the threshold value T is considered as 0.6. If the value of correlation coefficients is less than or equal to the threshold value, then maximum method is performed. In maximum method, fusion is performed by selecting the largest values from both of the transformed images. Otherwise, averaging method is employed, which computes the average value using both of the transformed images to perform fusion. Correlation-based fusion strategy is given as follows:

$$A_{l,p}^{new} = \begin{cases} \max(A_{l,p}^X, A_{l,p}^Y), & \text{if } (C \leq 0.6) \\ \text{avg}(A_{l,p}^X, A_{l,p}^Y), & \text{otherwise} \end{cases} \quad (6)$$

where, max and avg stand for maximum and average values, respectively, and $A_{l,p}^{new}$ for the fused coefficients.

3.2 Fusion of High Frequency Coefficients

The high frequency coefficients ($H_{l,p}^X, V_{l,p}^X, D_{l,p}^X$) and ($H_{l,p}^Y, V_{l,p}^Y, D_{l,p}^Y$) from both of the transformed images are fused using directive contrast method discussed in Sect. 2.2 and horizontal, vertical, and diagonal contrasts are obtained [14].

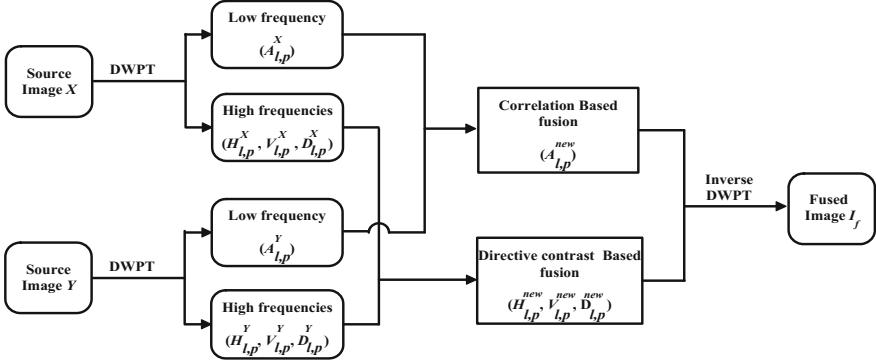


Fig. 2 Block diagram of proposed method

$$H_{l,p}^{new} = \begin{cases} H_{l,p}^X, & \text{if } |R_{l,p}^{H,X}| \geq |R_{l,p}^{H,Y}| \\ H_{l,p}^Y, & \text{otherwise} \end{cases} \quad (7)$$

$$V_{l,p}^{new} = \begin{cases} V_{l,p}^X, & \text{if } |R_{l,p}^{V,X}| \geq |R_{l,p}^{V,Y}| \\ V_{l,p}^Y, & \text{otherwise} \end{cases} \quad (8)$$

$$D_{l,p}^{new} = \begin{cases} D_{l,p}^X, & \text{if } |R_{l,p}^{D,X}| \geq |R_{l,p}^{D,Y}| \\ D_{l,p}^Y, & \text{otherwise} \end{cases} \quad (9)$$

Obtained fused wavelet coefficients from low ($A_{l,p}^X, A_{l,p}^Y$) and high frequency coefficients ($(H_{l,p}^X, V_{l,p}^X, D_{l,p}^X), (H_{l,p}^Y, V_{l,p}^Y, D_{l,p}^Y)$) are represented as $A_{l,p}^{new}$ and $H_{l,p}^{new}, V_{l,p}^{new}, D_{l,p}^{new}$. These four new fused coefficients are used for inverse DWPT. After performing inverse DWPT, a fused image I_f is obtained. The block diagram of proposed method is shown in Fig. 2.

4 Experimental Results and Discussion

The proposed method is applied on pepper, mandrill, pepsi images of size 512×512 , and office image of size 256×256 . Among these, pepper and mandrill images are considered as reference images I_r . But before applying the proposed method, these two reference images are blurred by convolving a Gaussian filter using 13×13 window and standard deviation $\sigma = 5$, and obtained images are referred as source images. The source images are shown in Figs. 3a, b, 4a, b, 5a, b, and 6a, b. The pepper images, which are highly concentrated on right and left part are given in Fig. 3a, b. Mandrill images are shown in Fig. 4a, b are blurred on upper and lower

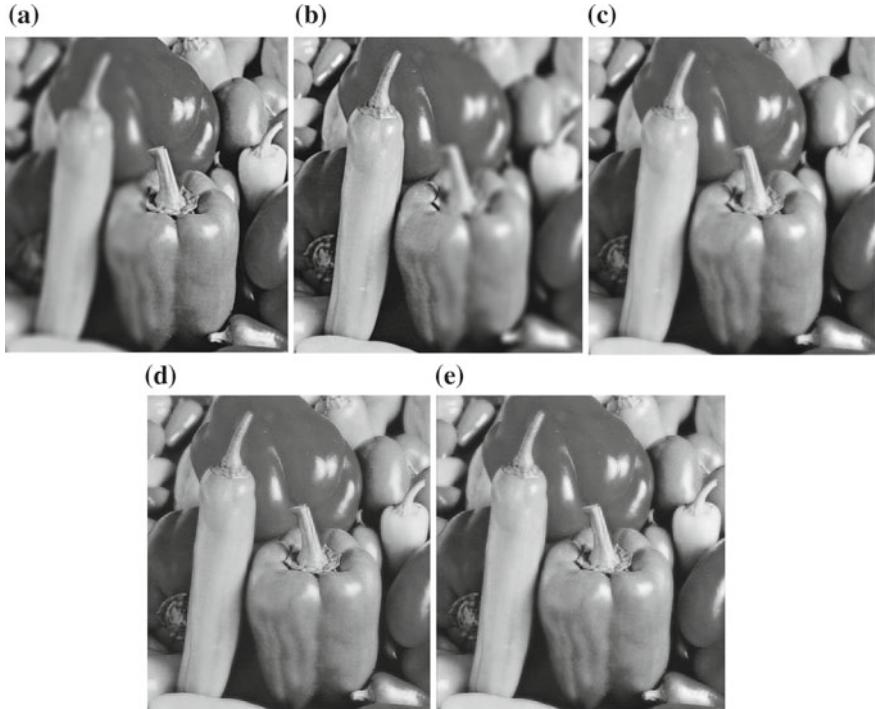


Fig. 3 Fusion results for pepper image. **a** Image blurred on the *left*; **b** image blurred on the *right*; **c** fused image by PCA; **d** fused image by directive contrast; **e** fused image by proposed method

parts. In pepsi images, barcode and container are focused as given in Fig. 5a, b. The office images are blurred on middle and corner as shown in Fig. 6a, b. Over the source images, proposed method is performed. The experimental results obtained by proposed method are compared with, PCA [13] and directive contrast [14] based image fusion techniques. The obtained results by existing methods are shown in Figs. 3c, d, 4c, d, 5c, d, and 6c, d. The results obtained by proposed method are shown in Figs. 3e, 4e, 5e, and 6e. It can be visually analyze from the obtained fused images that the proposed method gives better results than existing methods. But only visual inspection is not sufficient to measure the quality of fused images. Therefore, to evaluate the performance of existing methods and proposed method quantitatively, several parameters mean, standard deviation, correlation coefficients, and Peak-Signal-to-Noise-Ratio (PSNR) are also computed. The obtained results are shown in Table 1 and it can be observed that the results of proposed method are better than other existing methods. All these parameters are defined as follows:

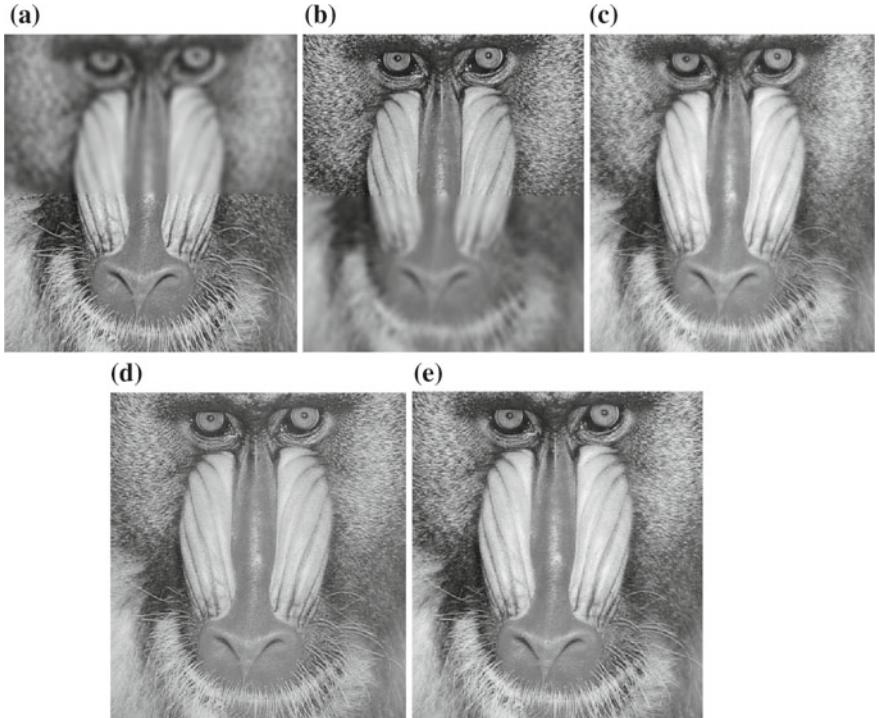


Fig. 4 Fusion results for mandrill image. **a** Image blurred on the *upper*; **b** image blurred on the *lower*; **c** fused image by PCA; **d** fused image by directive contrast; **e** fused image by proposed method

4.1 Mean and Standard Deviation

The mean and standard deviation (S.D.) is given as:

$$\hat{\mu} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I_f(i,j) \quad (10)$$

$$\sigma = \sqrt{\frac{1}{mn-1} \sum_{i=1}^m \sum_{j=1}^n (I_f(i,j) - \hat{\mu})^2} \quad (11)$$

where I_f represents fused image. The higher value of standard deviation represents the high contrast image.

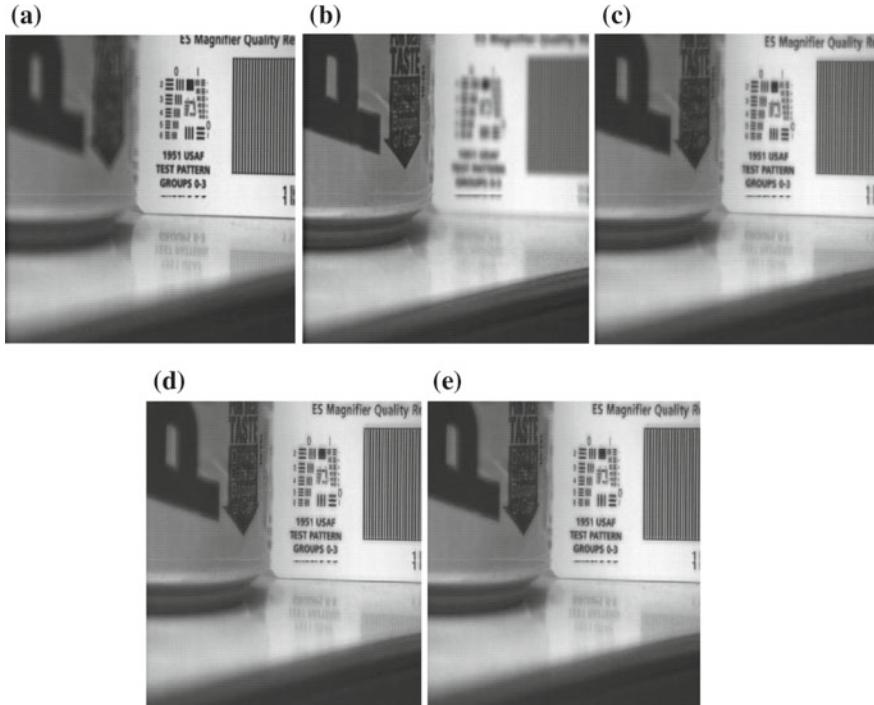


Fig. 5 Fusion results for pepsi image. **a** Image focus on barcode; **b** image focus on container; **c** fused image by PCA; **d** fused image by directive contrast; **e** fused image by proposed method

4.2 Correlation Coefficients

The correlation coefficient measures (C) the degree of two linearly related variables. It computes the relation between reference and fused image. The correlation coefficient value lies in [0, 1]. Correlation coefficient is computed from given in Eq. 5. The correlation coefficient approaches to 1 represents more similar information.

4.3 Peak-Signal-to-Noise-Ratio (PSNR)

PSNR evaluates the error between one of the reference and fused image.

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \quad (12)$$

where MSE is defined as

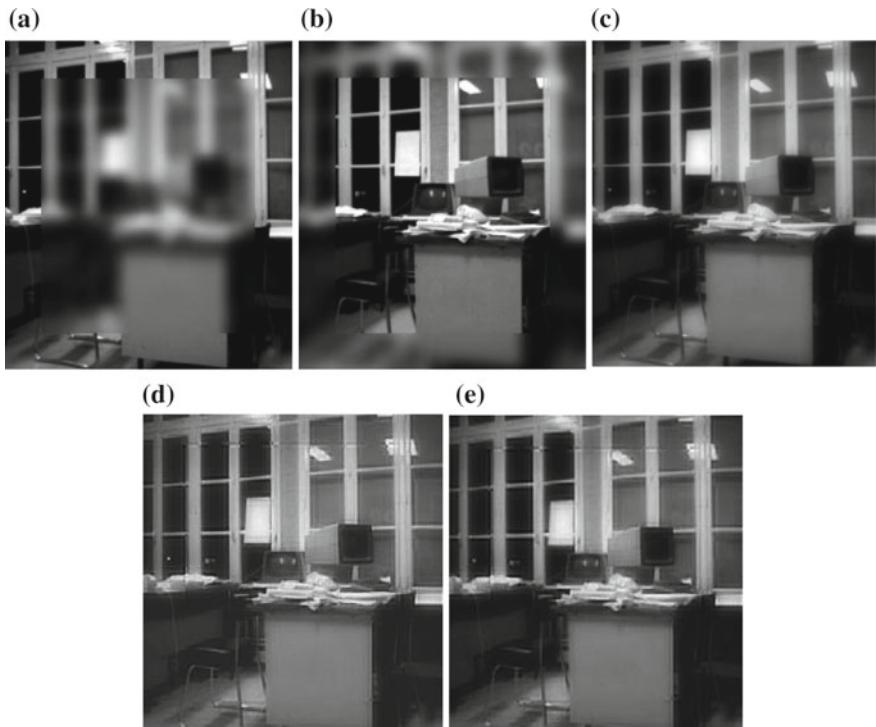


Fig. 6 Fusion results for office image. **a** Middle side blurred; **b** corner side blurred image; **c** fused image by PCA; **d** fused image by directive contrast; **e** fused image by proposed method

Table 1 Evaluation metrics using fused images

Source images	Evaluation indices	PCA	Directive contrast	Proposed method
Peppers	Mean	119.1268	119.1555	119.5538
	S.D.	50.8325	51.1246	51.2989
	C	0.9643	0.9847	0.9889
	PSNR	34.2197	34.5824	34.9147
Mandrill	Mean	128.4620	128.9738	129.5436
	S.D.	35.6187	38.0478	38.1671
	C	0.9103	0.9189	0.9399
	PSNR	30.2266	30.2827	30.7864
Pepsi	Mean	97.5710	97.5725	97.8974
	S.D.	43.9852	44.3332	44.4092
Office	Mean	80.3662	80.3571	80.5871
	S.D.	59.4153	60.7419	60.9577

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I_r(i,j) - I_f(i,j)]^2$$

where, I_r and I_f represent the reference image and fused image. Less MSE value denotes less error and high PSNR value represents better visual quality.

5 Conclusions

In this paper, a multi-focus image fusion technique based on local correlation and directive contrast is proposed using discrete wavelet packet transform domain. In the proposed work, DWPT is used which provides better localization of low as well as high frequency subbands. After performing the DWPT decomposition, obtained low frequency coefficients are combined using local correlation and high frequency coefficients are fused using directive contrast method. Using these fusion techniques, we obtained an improved fused image in which the contrast and details from each original image are enhanced. Several pairs of multi-focus images are used to test the performance of the image fusion method. The experimental results demonstrate that the proposed method is better than existing methods qualitatively as well as quantitatively.

References

1. Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. Proc. IEEE, 85 (1), 6–23 (1997)
2. Mitchell, H.B.: Image Fusion: Theories, techniques and applications. Springer-Verlag Berlin Heidelberg, (2010)
3. Zhou, Z., Li, S., Wang, B.: Multi-scale weighted gradient-based fusion for multi-focus images. Information Fusion, 20, 60–72 (2014)
4. Mitianoudis, N., Stathaki, T.: Pixel-based and region-based image fusion schemes using ICA bases. Inf. Fusion, 8 (2), 131–142 (2007)
5. Sasikala, M., Kumaravel, N.: A comparative analysis of feature-based image fusion methods. Inf. Tech. J., 6 (8), 1224–1230 (2007)
6. Tao, Q., Veldhuis, R.: Threshold-optimized decision-level fusion and its application to biometrics. Pattern Recogn., 42 (5), 823–836 (2009)
7. Pajaez G., Cruz J.: A wavelet-based image fusion tutorial. Pattern Recognition, 37(9), 1855–1872 (2004)
8. Chipman, L.J., Orr, T.M., Graham, L.N.: Wavelets and image fusion. International Conference on Image Processing, 3, 248–251 (1995)
9. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.C.: The dual-tree complex wavelet transform. IEEE Signal Process, Mag, 22(6), 123–151 (2005)
10. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans, Image Process, 14(12), 2091–2106 (2005)
11. da Cunha, A.L., Jianping, Z., Do, M.N.: The nonsubsampled contourlet transform: theory, design and applications. IEEE Trans, Image Process, 15(10), 3089–3101 (2006)

12. Hui, L., Manjunath, B.S., Mitra, S.K.: Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing*, 57 (3), 235–245 (1995)
13. Naidu, V.P.S., Raol, J.R.: Pixel-level image fusion using wavelets and principle component analysis. *Defence Science Journal*, 58(3), 338–352 (2008)
14. Bhatnagar, G., Raman B.: A new image fusion technique based on directive contrast. *Electronic letters on computer vision and image analysis*, 8(2), 18–38 (2009)
15. Walczak, B., Bogaert, B.V.D., Massart, D.L.: Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data. *Analytical Chemistry*, 68(10), 1742–1747 (1996)
16. Wang, H.H., Peng, J.X., Wu, W.: A fusion algorithm of remote sensing based on discrete wavelet packet. *IEEE, Proc. Machine learning and cybernetics*, 4, 2557–2562 (2003)
17. Amiri, G.G., Asadi, A.: Comparison of different methods of wavelet and wavelet packet transform in processing ground motion records. *Internatinal journal of civil engineering*, 7(4), 248–257 (2009)
18. Toet, A., Ruyven, L.J.V., Valeton, J.M.: Merging thermal and visual images by a contrast pyramid. *Opt Eng.* 28(7), 789–792 (1989)
19. Guixi, L., Wenjin, C., Wenjie L.: An image fusion method based on directional contrast and area-based standard deviation. *Electronic Imaging and Multimedia Technology IV, Proc. of SPIE*, 5637, 50–56 (2005)
20. Wenzhong, S., ChangQing, Z., Yan, T., Janet, N.: Wavelet-based image fusion and quality assessment. *International Journal of Applied Earth Observation and Geoinformation*, 6, 241–251 (2005)

Multi-exposure Image Fusion Using Propagated Image Filtering

Diptiben Patel, Bhoomika Sonane and Shanmuganathan Raman

Abstract Image fusion is the process of combining multiple images of a same scene to single high-quality image which has more information than any of the input images. In this paper, we propose a new fusion approach in a spatial domain using propagated image filter. The proposed approach calculates the weight map of every input image using the propagated image filter and gradient domain postprocessing. Propagated image filter exploits cumulative weight construction approach for filtering operation. We show that the proposed approach is able to achieve state-of-the-art results for the problem of multi-exposure fusion for various types of indoor and outdoor natural static scenes with varying amounts of dynamic range.

Keywords Computational photography · Multi-exposure fusion · HDR imaging · Propagated image filter

1 Introduction

Optical system and sensor of a digital camera play a crucial role during image acquisition. The lens of the camera can capture sharp details only for some finite depth of field around focal plane leading to blur of objects outside that depth region. The amount of blur increases as the object is more distant from the focal plane. This makes impossible to focus all the objects in a scene. Also, natural scenes contain very high dynamic range which is not possible to capture with a single exposure due to the limitations of the digital sensor dynamic range. These two aspects raise the need of combining a number of images to generate a single high-quality image.

D. Patel (✉) · B. Sonane · S. Raman

Electrical Engineering, Indian Institute of Technology, Gandhinagar, India

e-mail: diptiben.patel@iitgn.ac.in

B. Sonane

e-mail: bhoomika.sonane@iitgn.ac.in

S. Raman

e-mail: shanmuga@iitgn.ac.in

Image fusion combines a set of input images and creates more informative (in the sense of human perception or machine perception) output image. Image fusion finds application in numerous areas, such as computer vision, medical imaging, remote sensing, security and surveillance, and microscopic imaging. For satellite imaging, image fusion is useful for combining high-spatial resolution panchromatic image and low-spatial resolution multi-spectral image to produce image with high-spatial and spectral resolution [1]. With the advanced use of medical imaging techniques like X-ray, computed tomography (CT), magnetic resonance imaging (MRI), they can capture better information for bones, tissues, or blood vessels, respectively. Multimodal image fusion has been found as a solution to combine all information in a single image to diagnose a disease [2]. We focus on the multi-exposure fusion application in this work which enables one to reconstruct a single high contrast low dynamic range (LDR) image from multiple differently exposed LDR images.

Computational photography application which captures static scene using multiple exposures can be fused using the proposed approach. The primary contributions of the proposed approach are listed below.

1. We propose a novel approach to estimate the weight for each input image using propagated image filter.
2. Propagated image filter can overcome the problem of inter region mixing without the need of explicit spatial filter kernel for larger filter size.
3. The proposed image fusion approach does not need tone mapping after fusion as the final image is of LDR.

The rest of the paper is organized as follows. Section 2 describes related work in image fusion techniques. Section 3 explains the proposed image fusion approach using propagated image filtering. Results are discussed in Sect. 4. The paper is concluded in Sect. 5 with suggestions for future improvement.

2 Related Work

We survey the broad spectrum of fusion techniques in this section apart from multi-exposure fusion. The primary objective of image fusion techniques is to select the best information available in a set of input images. Spatial domain techniques perform weighted sum of the pixel values of set of input images to compose the output image ([3–9]). Transform domain techniques perform weighted sum of the transformed coefficients after transforming set of input images into other domain ([10–13]). The output image is obtained by inverse transform operation. Weight for every pixel is calculated using different sharpness criteria. Li et al. calculated pixel weight by morphological filtering followed by image matting technique [3]. Morphological filtering does rough segmentation and image matting provides accurate focus region to it. This method combines the focus information and correlation between the nearby pixels to provide accurate results. Weights of pixels for set of images are obtained by a bilateral sharpness criterion which exploits strength and phase coher-

ence of gradient information in [4]. Luo et al. computed image driven weight map based on the similarity characteristics between source images [5]. Similarity characteristics of source images are described by luminance, contrast, and structure. Rather than going for pixel-based fusion, Ma et al. proposed a patch-based fusion approach improving three image features: signal strength, signal structure, and mean intensity [6]. Decomposing a color image into these three features, and reconstructing a color patch allows one to process all color channels together providing better color appearance.

Multi-resolution based image fusion techniques have been adopted more as they combine input images at multiple resolution level via different fusing rules. They find efficient as they fuse features rather than intensities. Various multi-resolution transforms exploited for the purpose are Laplacian pyramid [10], discrete wavelet transform (DWT) [11], and nonsubsampled contourlet transform (NSCT) [12]. Mertens et al. fused bracketed exposure sequences by simple sharpness measures like saturation, contrast, and well-exposedness at each level of Laplacian pyramid [10]. Having examined the difference between marginal distribution of the wavelet coefficients for different focus depths, Tian et al. proposed a new statistical sharpness measure based on the marginal wavelet coefficient distribution and locally adaptive Laplacian mixture model to perform adaptive image fusion in wavelet domain [11]. After decomposing images using NSCT, Zhang et al. calculated weight for different sub-band coefficients in a different manner based on imaging principle of the defocused optical system, the characteristics of the human vision system and the property of NSCT [12]. A survey of fusion strategies can be found in [14].

The primary objective of this paper is to device an algorithm for multi-exposure image fusion in spatial domain. Typical approaches combined multi-exposure images into a single HDR image ([15–17]). These methods require the estimation of camera response function to fuse the images in the irradiance space. A detailed survey is given in [18, 19]. Whenever an image is subjected to exposure change, texture, and small-gradient edges are blurred more than sharp edges. Edge preserving filters have the property of preserving sharp edges and filtering texture and small-gradient edges more. This property proves to be advantageous for weight map construction for image fusion techniques. Edge preserving filters like the bilateral filter ([20, 21]) and the guided image filter [22] are used for image fusion. Raman et al. proposed a matte function for each input images using the bilateral filter [20]. Matte function is computed from the difference image created by subtracting the bilateral filtered image from original image. This difference image highlights texture region and small-gradient edge region which is given more weight while compositing. Li et al. proposed an image fusion technique using the guided filter [22]. The guided filter is used to construct a weight map along with salient detail of each image.

3 Proposed Image Fusion Approach

The bilateral filter and the guided filter are edge preserving filters which exploit spatial and range filter kernels. The bilateral filter makes use of spatial and photometric distance for filtering purpose [23]. The guided filter explores local linear transformation to filter an image using guidance image [24]. The guided filter is advantageous in removing gradient reversal artifacts over the bilateral filter. Despite this, both the filters fail when the filter size is too large. They exploit spatial filter kernel which defines the neighboring pixels while deciding the filter size. Hence, the pixel which does not belong to the same region (but included in filter kernel size) is also assigned a comparable weight in modifying a center pixel. This elevates the problem of inter region mixing. To overcome this problem, Chang et al. defined a probabilistic model of a filter named propagated image filter which considers context information between two pixels [25].

Proposed image fusion approach using propagated image filter is shown in Fig. 1. Propagated image filter is used to construct weight map for each input image I_n . Then original images are fused through the propagated image filter-based weight map.

3.1 Propagated Image Filter

Propagated Image filter uses cumulative weight construction approach between two pixels to be considered [25]. Let P_0 is a pixel to be filtered and P_t is a pixel within the window size w . As shown in Fig. 2a, the pixels along a path from P_t to P_0 are

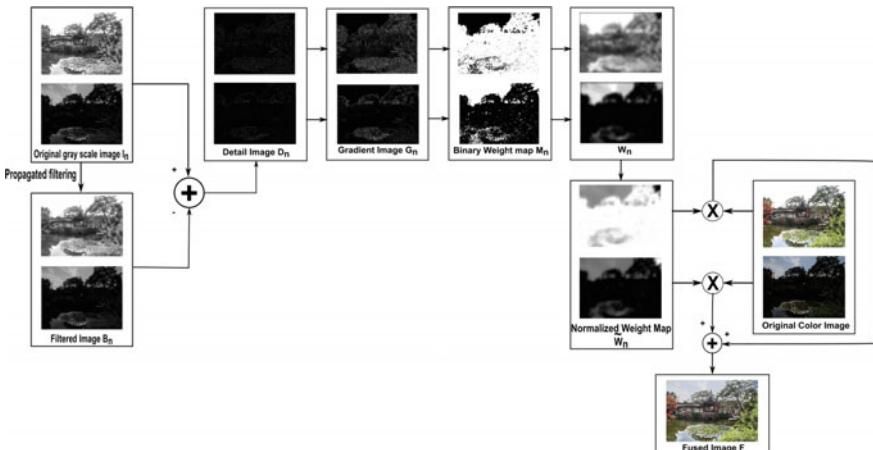


Fig. 1 Proposed image fusion approach using two images

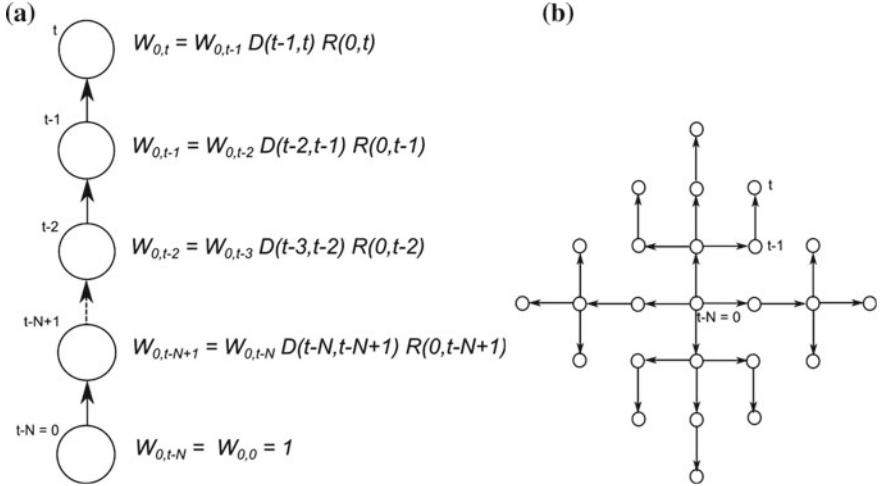


Fig. 2 Propagation filter: **a** Calculation of weight $w_{0,t}$ **b** 2D path pattern used for image filtering using Manhattan distance [25]

defined as $P_{t-1}, P_{t-2}, \dots, P_{t-N}$. Where, $P_{t-N} = P_0$. The weight contribution of pixel P_t for P_0 is defined as Eq. (1):

$$w_{0,t-k+1} = w_{0,t-k} \times D(t-k, t-k+1) \times R(0, t-k+1) \quad (1)$$

where, $k = N, N-1, \dots, 1$. Initially, $w_{0,t-N} = w_{0,0} = 1$. As per Eq. (2), D calculates adjacent photometric relation between adjacent pixels along a path, which is proportional to the value of Gaussian function of their pixel value difference.

$$\begin{aligned} D(x, x-1) &= g(\|P_x - P_{x-1}\|; \sigma_a) \\ &= \exp\left(\frac{\|P_x - P_{x-1}\|^2}{2 \times \sigma_a^2}\right) \end{aligned} \quad (2)$$

where, P_x is intensity value of pixel at location x . $\|P_x - P_{x-1}\|$ is the Euclidean distance between intensity values of two consecutive pixels along a path. It supplies higher value for the pixel values being similar (Less intensity difference).

As defined in Eq. (3), R calculates photometric relation between pixels x and y which is defined as:

$$\begin{aligned} R(x, y) &= g(\|P_x - P_y\|; \sigma_r) \\ &= \exp\left(\frac{\|P_x - P_y\|^2}{2 \times \sigma_r^2}\right) \end{aligned} \quad (3)$$

where, P_x is intensity value of pixel at location x . $\|P_x - P_y\|$ is the Euclidean distance between intensity values of two pixels (not necessary adjacent to each other)

along a path. σ_a and σ_r control the width of Gaussian kernel. For simplicity we consider $\sigma_a = \sigma_r$.

For image filtering, we need to find a path connecting every pixel in a window w to center pixel and then find a weight for each pixel traversing along that path. Path is found using Manhattan distance as proposed in [25]. The 2D pattern used in this paper is shown in Fig. 2b.

3.2 Weight Map Construction with Propagated Image Filtering

Let us explain the weight map construction for the set of S multi-exposure images indexed by $n = 1, 2, \dots, S$. After filtering gray scale input image I_n , the filtered image obtained is L_n . Subtracting filtered image from input image gives the detail image D_n as given by Eq. (4).

$$D_n(x, y) = I_n(x, y) - L_n(x, y) \quad (4)$$

Each detail image D_n is converted into gradient domain by taking absolute value of first derivative in each direction and summing it as defined in Eq. (5):

$$G_n(x, y) = \left| \frac{\partial D_n(x, y)}{\partial x} \right| + \left| \frac{\partial D_n(x, y)}{\partial y} \right| \quad (5)$$

Binary weight map for each image is constructed by considering the pixel which belongs to have maximum value of G_n . It is defined as Eq. (6):

$$M_n(x, y) = \begin{cases} 1, & \text{if } G_n(x, y) = \max(G_1(x, y), G_2(x, y), \dots, G_S(x, y)) \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Binary weight map is multiplied with input image to weigh each pixel according to input intensity value after morphological operation such as *hole filling*. Average filter is applied to overcome local effects due to sensor, scene, etc. as shown in Eq. (7),

$$\tilde{M}_n(x, y) = (M_n(x, y) \cdot I_n(x, y)) * H \quad (7)$$

where, \tilde{M}_n is modified weight map. H is average filter of size $l \times l$. The practical value of l used here is 20. Each weight map \tilde{M}_n is normalized to range (0, 1) to make all weight maps into same range. Normalization operation is defined as Eq. (8).

$$W_n(x, y) = \frac{\tilde{M}_n(x, y) - \tilde{M}_{n,min}}{\tilde{M}_{n,max} - \tilde{M}_{n,min}} \quad (8)$$

Here, $\tilde{M}_{n,\min}$ and $\tilde{M}_{n,\max}$ are the minimum and maximum values of \tilde{M}_n , respectively. As per Eq. (9), W_n is normalized at pixel level so that it sums to 1 for every pixel location along the set of input images S .

$$\begin{aligned}\tilde{W}_n(x, y) &= \frac{1}{Z(x, y)} W_n(x, y) \\ Z(x, y) &= \sum_{n=1}^S W_n(x, y)\end{aligned}\tag{9}$$

3.3 Image Fusion

Input images are fused using the weight map \tilde{W} to obtain fused high contrast LDR image F as defined in Eq. (10).

$$F = \sum_{n=1}^S \tilde{W}_n \times I_n\tag{10}$$

For color image fusion, same weight map \tilde{W}_n is used for each of RGB color channel.

4 Results and Discussion

We tested our proposed algorithm on a variety of indoor and outdoor natural images with more than two input images captured with different exposure settings. Figure 3 shows sample input images used for a scene in each row. Images used to evaluate the algorithm are of size $340 \times 512 \times 3$. Number of images varies from 3 to 6 for different exposure settings of a same static scene. Window size w , σ_a and σ_r for propagated image filtering are chosen to be 6, 5 and 5, respectively.

Figure 4 shows the fused images of the proposed method compared with the techniques of [6, 10, 20]. Considering the first row of Fig. 4, we can observe that the sky and the cloud color are well preserved than other fusion techniques. In the second row, information captured with least exposure (far from camera like from window) is more accurately fused from corresponding first input image of second row in Fig. 3. Landscape image preserves contrast and haze well from input images as per the third row of Fig. 4. We observe from the fourth row of Fig. 4 that building structure at left part of the image is fused with more contrast and good illumination as compared to [6, 10, 20]. It can be observed that the results are little under saturated with loss of building structure in the fused images of [6, 10, 20]. Fifth row shows the fused images of a cave with fall outside it. The proposed approach is observed to fuse far

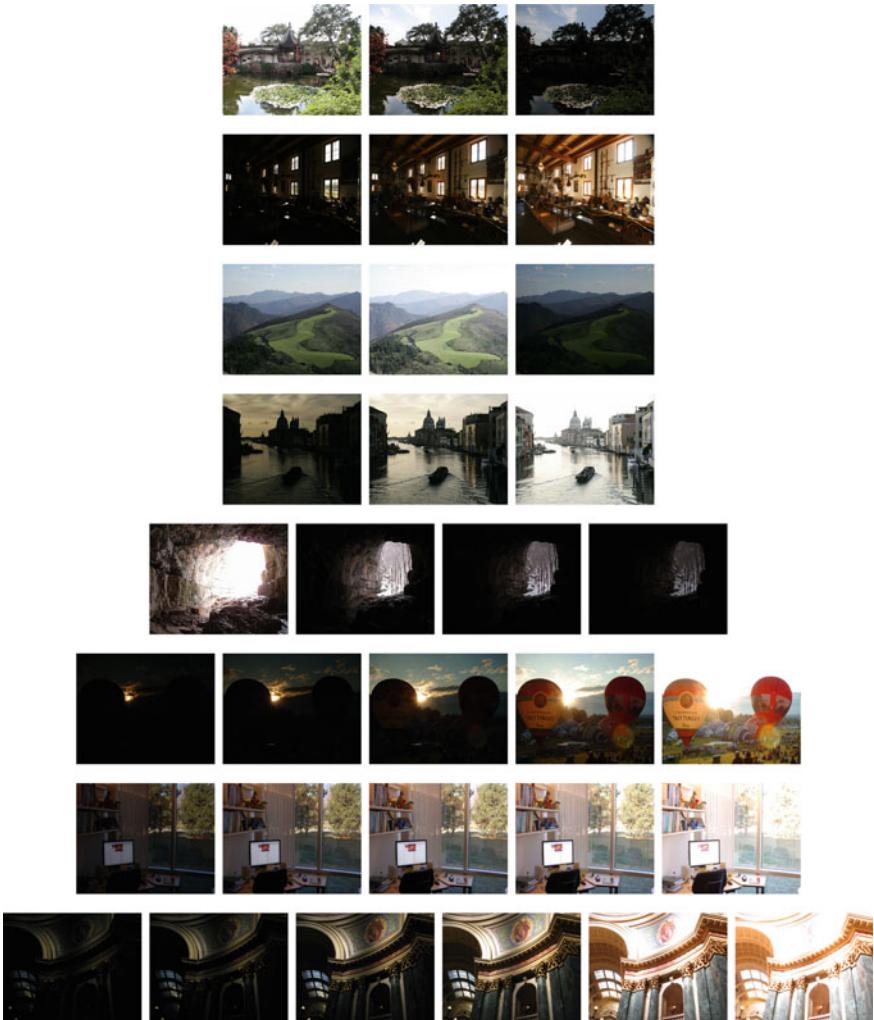


Fig. 3 Each row corresponds to a set of multi-exposure images of a scene

distant illumination better than other fusion techniques. In the sixth row, the proposed algorithm produces visible artifacts in the sky region because dataset consists of more number of less exposure images which is a limitation to be addressed in future. The proposed fusion algorithm preserves the reflection of indoor objects in a mirror as shown in the seventh row. The eighth row shows fused images for one part of capitol building at Madison. We observe that the proposed approach maintains proper illumination at hemisphere of a building and at wall border below portrait of a man as compared to that of [6, 10, 20].

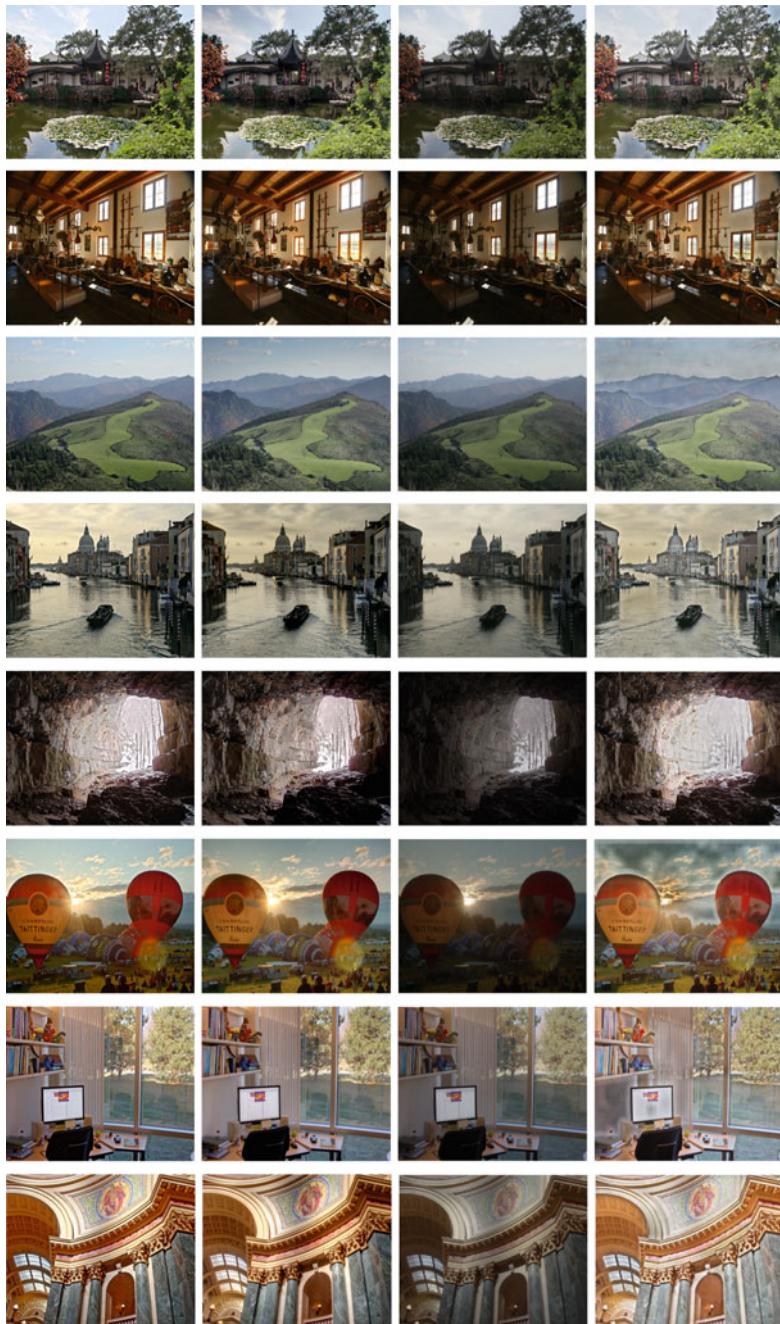


Fig. 4 Fused images obtained by different techniques: First column shows results of Ma and Wang [6], second column shows results of Mertens et al. [10], third column shows results of Raman and Chaudhuri [20] and fourth column shows results of proposed approach

5 Conclusion

We have proposed a novel image fusion approach using propagated image filtering. The proposed algorithm exploits gradient domain weight map construction along with propagated image filter. It is more effective than the existing state-of-the-art image fusion techniques even for images captured with less exposure time. It also preserves contrast and brightness well in the fused images. Moreover, the proposed approach does not require tone mapping after fusion process as it generates the LDR image directly. We would like to extend the proposed approach to work more effectively independent of how the exposure values are distributed in a given dynamic range of the scene. Extension to scenes with dynamic objects could also be explored in future.

References

1. M. R. Metwalli, A. H. Nasr, O. S. F. Allah, and S. El-Rabaie, "Image fusion based on principal component analysis and high-pass filter," in *Computer Engineering & Systems, 2009. ICCES 2009. International Conference on*. IEEE, 2009, pp. 63–70.
2. G. Bhatnagar, Q. Wu, and Z. Liu, "Directive contrast based multimodal medical image fusion in nsct domain," *Multimedia, IEEE Transactions on*, vol. 15, no. 5, pp. 1014–1024, 2013.
3. S. Li, X. Kang, J. Hu, and B. Yang, "Image matting for fusion of multi-focus images in dynamic scenes," *Information Fusion*, vol. 14, no. 2, pp. 147–162, 2013.
4. J. Tian, L. Chen, L. Ma, and W. Yu, "Multi-focus image fusion using a bilateral gradient-based sharpness criterion," *Optics communications*, vol. 284, no. 1, pp. 80–87, 2011.
5. X. Luo, J. Zhang, and Q. Dai, "A regional image fusion based on similarity characteristics," *Signal processing*, vol. 92, no. 5, pp. 1268–1280, 2012.
6. K. Ma and Z. Wang, "Multi-exposure image fusion: A patch-wise approach," in *ICIP*. IEEE, 2015.
7. S. Li and X. Kang, "Fast multi-exposure image fusion with median filter and recursive filter," *Consumer Electronics, IEEE Transactions on*, vol. 58, no. 2, pp. 626–632, 2012.
8. W. Zhang and W.-K. Cham, "Gradient-directed multiexposure composition," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 2318–2323, 2012.
9. Z. G. Li, J. H. Zheng, and S. Rahardja, "Detail-enhanced exposure fusion," *Image Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 4672–4676, 2012.
10. T. Mertens, J. Kautz, and F. Van Reeth, "Exposure fusion: A simple and practical alternative to high dynamic range photography," in *Computer Graphics Forum*, vol. 28, no. 1. Wiley Online Library, 2009, pp. 161–171.
11. J. Tian and L. Chen, "Adaptive multi-focus image fusion using a wavelet-based statistical sharpness measure," *Signal Processing*, vol. 92, no. 9, pp. 2137–2146, 2012.
12. Q. Zhang and B.-l. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Processing*, vol. 89, no. 7, pp. 1334–1346, 2009.
13. P.-w. Wang and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*. IEEE, 2008, pp. 965–968.
14. R. Szeliski, "Image alignment and stitching: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 2, no. 1, pp. 1–104, 2006.
15. S. Mann and R. Picard, "Being undigital with digital cameras." MIT Media Lab Perceptual, 1994.
16. P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *ACM SIGGRAPH*. ACM, 1997.

17. T. Mitsunaga and S. K. Nayar, "Radiometric self calibration," in *CVPR*, vol. 1. IEEE, 1999.
18. R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.
19. E. Reinhard, W. Heidrich, P.Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.
20. S. Raman and S. Chaudhuri, "Bilateral filter based compositing for variable exposure photography," in *Proceedings of Eurographics*, 2009.
21. F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM transactions on graphics (TOG)*, vol. 21, no. 3, pp. 257–266, 2002.
22. S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *Image Processing, IEEE Transactions on*, vol. 22, no. 7, pp. 2864–2875, 2013.
23. S. Paris, P. Kornprobst, J. Tumblin, and F. Durand, "Bilateral filtering: Theory and applications in computer graphics and vision." 2008.
24. K. He, J. Sun, and X. Tang, "Guided image filtering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1397–1409, 2013.
25. J.-H. Rick Chang and Y.-C. Frank Wang, "Propagated image filtering," in *IEEE CVPR*, 2015, pp. 10–18.

Tone Mapping HDR Images Using Local Texture and Brightness Measures

Akshay Gadi Patil and Shanmuganathan Raman

Abstract The process of adapting the dynamic range of a real-world scene or a photograph in a controlled manner to suit the lower dynamic range of display devices is called tone mapping. In this paper, we present a novel local tone mapping technique for high-dynamic range (HDR) images taking texture and brightness as cues. We make use of bilateral filtering to obtain *base* and *detail* layer of the luminance component. In our proposed approach, we weight the base layer using local to global brightness ratio and texture estimator, and then combine it with the detail layer to get the tone mapped image. To see the difference in contrasts between the original HDR Image and the tone mapped image using our model, we make use of an online dynamic range (in)dependent metric. We present our results and compare it with other tone mapping algorithms and demonstrate that our model is better suited to compress the dynamic range of HDR images preserving visibility and information and with minimal artifacts.

Keywords Computational photography · HDR imaging · Tone mapping

1 Introduction

The perception of real-world scenes by human beings is natural and consists of a wide range of luminance values [1]. HDR Imaging aims to capture all of these luminance values present in the natural scene and can simultaneously incorporate detailed information present in the deepest of shadows and brightest of light sources ([2, 3]). But the range of luminance values that a given display device can reproduce is limited. For instance, only 8 bits (256 levels) of brightness information per channel is assigned for every pixel in video cameras or digital still cameras [4]. This is

A. Gadi Patil (✉) · S. Raman

Electrical Engineering, Indian Institute of Technology, Gandhinagar, India
e-mail: akshay.patil@iitgn.ac.in

S. Raman
e-mail: shanmuga@iitgn.ac.in

because the display devices can reproduce luminance range of about 100:1 Candela per square meter (Cd/m^2) as opposed to human vision which ranges from 100 000 000:1 [5]. In today's world most applications in gaming industry, augmented reality, and photography require that the display devices mimic the images with real-world scene luminance to provide a more realistic and natural experience ([6, 7]). Tone mapping operators (TMOs) aim at maintaining the overall contrast and brightness levels imitating images with high-dynamic range (HDR). Myszkowski in [8] confronts incompatibility between images and display devices while performing dynamic range compression. One of the important aspects of tone mapping is that the same HDR image can be tone mapped into different styles depending upon the situational or contextual considerations [9].

Most often texture, contrast, and brightness information is lost in the process of tone mapping. The relative amount of information of the above-mentioned features is one of the parameters used to measure the quality of the tone mapped image. We propose a novel approach for tone mapping using a computationally efficient texture and brightness estimator which makes use of the local variations in the given HDR image. Variation in the texture information is captured by this texture estimator whereas the relative local brightness information is captured by the brightness estimator. Using this we work on the luminance channel of the HDR image and try to obtain a tone mapped image that leads to same response in the observer as that of a real world scene.

The main contributions of this paper are listed below.

1. Our proposed model uses texture and brightness estimator that is computationally simple compared to most of the existing tone mapping algorithms.
2. A novel framework for preserving contrast and brightness information from HDR to tone mapped images.
3. This approach for tone mapping is an local adaptation that does not affect other similar regions present in the image.

We discuss the existing tone mapping algorithms in Sect. 2. In Sect. 3, we describe the algorithm of the proposed approach. We then present our results in Sect. 4 and compare them with the results of other tone mapping algorithms using the online dynamic range (in)dependent metric assessment [10]. We end the paper with conclusions, discussions and scope for future work in Sect. 5.

2 Related Work

The earliest works on tone mapping were by Tumblin and Rushmeier [11] where they developed nonlinear global mapping functions that characterized the human visual system's brightness and contrast perception. In [12], Ward made use of a simple linear function relating the world luminance and the display luminance using a proportionality constant known as scale factor. Larson et al in [1] used histogram equalization technique which disregarded empty portions of the histogram and was

able to achieve efficient contrast reduction. But contrast reduction was not achieved when input image exhibited a uniform histogram. The past decade has seen considerable work deriving motivation from the above ideas and much attention has been given to the tone mapping algorithms since then. Almost all tone mapping algorithms work on the luminance channel of the HDR Image and then integrate this modified luminance channel to the CIE_Lab color space from which the tone mapped RGB image is obtained. Fattal et al. described that large changes in gradients are due to abrupt change in luminance values in HDR images [13]. So large gradients were attenuated keeping the smaller ones unaltered.

The tone mapping methods used can be broadly classified into two groups. The first one is the *spatially uniform* group where in a global mapping function is used that does not account for local variations in the image. It means that a single function is used for all the pixels in the image and usually the global mapping functions are sigmoid or close to sigmoid in nature that preserve some details in highlights and shadows [7]. Because of this, the mapping functions in this group do not directly address contrast reduction. Intuitively this must be computationally inexpensive and the experiments have proven so. The second one is the *spatially varying* group which is local in the sense that mapping functions are applied over a given neighborhood of a pixel just as human vision is sensitive to local contrasts. Some of these operators take motivation from what McCann suggests in [14], lightness perception. The operators presented in ([13, 15, 16]), on the other hand, compute the arithmetic mean of luminance values in a pixel neighborhood. Color is treated as a separate entity by these tone mapping operators [17]. Mantiuk et al. in [18] proposed post processing techniques to reduce color saturation. Durand and Dorsey [16] employed edge preserving smoothing operators to justify locally adaptive processes which help in minimizing haloing artifacts. We base our approach on the one suggested by Ashikmin in [19] by making use of ratio of Gaussian filtered images to get the texture information. We select same kernel size for Gaussian filters in our approach. Due to the localization of the approach, these methods tend to be computationally expensive than spatially uniform or global methods.

Other distinctions among tone mapping operators exist, such as *static* and *dynamic operators*. Our approach falls in the *static operator* category where we work on static images unlike *dynamic operators* which process a stream of images. Information on making use of texture for capturing the details in images is given in Tao et al. [20]. So we derive our motivation from the above and would like to use local features, such as contrast, texture and brightness for tone mapping but at the same time, we want our algorithm to be of same computational complexity as that of spatially uniform methods. Making use of local features will allow us to capture more information that can be used appropriately on display devices than the global compression techniques that store information which produce not so realistic images. It is necessary to preserve every line or edge and every minute detail of the HDR image. Such textures on images can be visibly affected by geometric, radiometric and other distortion as mentioned in [21]. We make use of what Durand and Dorsey suggested in [16], the base layer and the detail layer obtained by splitting a HDR image into two layers using a bilateral filter. In our approach, the detail layer is not processed and only

the base layer is processed to compress the dynamic range. We also process only the luminance component of the HDR image while leaving the color components unaltered.

3 Proposed Approach

The block diagram in Fig. 1 explains our methodology of tone mapping. Our proposed approach comprises of four steps for tone mapping of HDR Images. All the four steps are discussed below sequentially.

3.1 Luminance Channel of HDR Image

As mentioned earlier, we work on the luminance channel of the HDR image because the information on the dynamic range of the intensity values is contained in this channel. We extract the luminance channel of the HDR image I using the equation in [22] which is given below.

$$L = 0.2126I_R + 0.7152I_G + 0.0722I_B \quad (1)$$

I_R , I_G , and I_B are the red, green, and blue channels of the image I , respectively. The above equation is used because it closely relates to the human vision perception. Similarly ‘a’ and ‘b’ channels are computed and stored unaltered. Once the

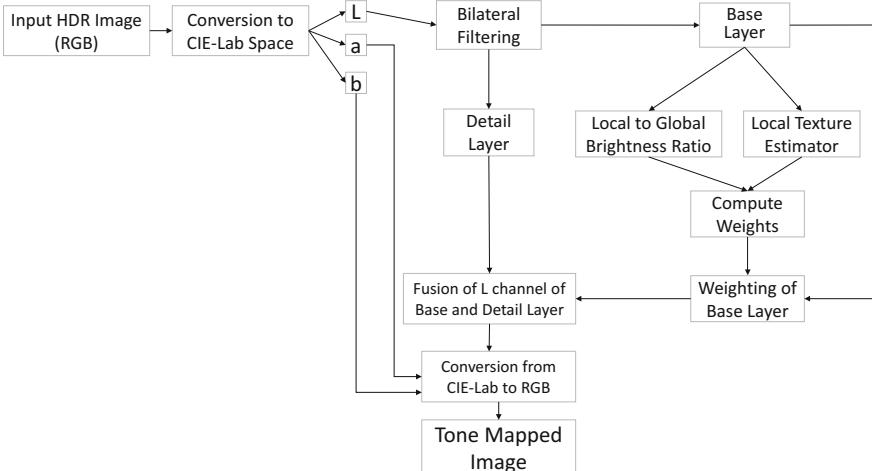


Fig. 1 Block diagram of the proposed approach

luminance channel is obtained, we would like to split it into two layers using bilateral filter. The two layers are: *base layer* and *detail layer*. We work only on the base layer and keep the detail layer as it is.

3.2 Bilateral Filtering

Once the luminance channel of the HDR image is available, we process it using an edge preserving bilateral filter as suggested by Durand and Dorsey in [16]. The output of the bilateral filter gives two layers, the base layer *BL* and the detail layer *DL*. We preprocess the base layer *BL* to extract the texture and brightness information.

3.3 Pre-processing of Base Layer

Once the base layer *BL* is obtained, it is preprocessed to obtain information about texture and brightness as explained below. We compute the weights from local texture estimator and local to global brightness ratio. The base layer is modified by these weights as explained by Eq. (3).

Local Texture Estimation The amount of local texture present is estimated by taking the ratio of local smoothing around a pixel (x, y) of the image I using gaussians of different standard deviations. In the numerator, we use a Gaussian filter $G_{\sigma_1}(x, y)$ of size 9×9 with a standard deviation 1 around that neighborhood of the pixel and we use another Gaussian $G_{\sigma_2}(x, y)$ with a standard deviation of 4. This quantity we call, the texture estimator $T(x, y)$. The equation for the texture estimator is given below.

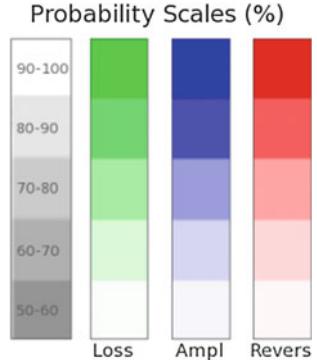
$$T(x, y) = \frac{G_{\sigma_1}(x, y) \otimes \|g(x, y)\|}{G_{\sigma_2}(x, y) \otimes \|g(x, y)\|} \quad (2)$$

Here $g(x, y)$ is the gradient around the 9×9 neighborhood of a pixel (x, y) of the base layer *BL* obtained after applying the bilateral filter on the HDR image [20]. We next focus on preserving the brightness in the tone mapped image.

Local Brightness Estimation In the process of tone mapping, the brightness may often get reduced perhaps making the tone mapped image dark. In order to preserve the brightness in the process, we first estimate the average brightness in the neighborhood of a pixel (x, y) and then, we find the average brightness of the entire image. We call Brightness estimator $B(x, y)$, which is the ratio of average local brightness in the neighborhood of a pixel to that of the global average brightness value.

Weighting the Base Layer Now once we have the local texture $T(x, y)$ and brightness $B(x, y)$ estimators, we take a linear combination of these two in accordance with the equation (3) below to compute the weights W and then we weight the base layer and modify it using Eq. (4) given below.

Fig. 2 Dynamic range
(in)dependent metric legends
[10]



$$W = (\alpha \times T(x, y)) + ((1 - \alpha) \times (B(x, y))^{\beta}) \quad (3)$$

$$\widehat{BL} = BL \times ((\alpha \times T(x, y)) + ((1 - \alpha) \times (B(x, y))^{\beta})) \quad (4)$$

In the above Eq. (4), the first term inside the brackets works to preserve the contrast and texture information and the second term helps in maintaining the brightness information. In Eq. (4) above, we have used two parameters, α and β . Both of them can be varied by the user. Here α takes care of the overall contrast in the image and β takes care of the brightness in the tone mapped image. For the results presented in this paper we have used $\alpha = 0.3$ and $\beta = 0.4$.

3.4 Fusion and Tone Mapped Image

Once the preprocessing is done, we then combine the detail layer DL on the processed base layer \widehat{BL} , thus bringing back every information present in detail layer DL back into the image. We are still in the CIE-Lab space. So to get our tone mapped RGB image, we need to convert the image from CIE-Lab to RGB space. Thus, we obtain the final tone mapped image.

4 Results and Discussion

We present the results for a set of nine HDR images on six different state-of-the-art tone mapping methods including the proposed approach which are presented from top row to the bottom row of Fig. 3. The rows are in the order, namely: AdobeLobby, Backyard, diffuse_map, elephant_fg, face_fg, puma_fg, ostrich_fg, smallOffice,

and tahoe1 [6]. For every image, we compare our tone mapped image with reference to other standard tone mapping methods. For the results of other tone mapping algorithms, we have used the best possible values of user input parameters. As could be seen from Fig. 3, ChiuTMO gives a blurred image that is not at all appealing and has spilling across edges [6]. TumblinRushmeierTMO increases the brightness in the resulting image and gives good results [11]. FattalTMO gives a dull image with reduced brightness and more dark regions in shadows [13]. WardHistTMO works well but amplifies much of the invisible contrast in the image irrespective of the environment [12]. ReinhardTMO is found to work better only in good lighting ambiance [7]. Our proposed approach better preserves the texture information in all types of environment and closely matches the best among all the other TMOs which is TumblinRushmeierTMO as could be seen from Fig. 4. Since our method uses gradients around a neighborhood of a pixel to estimate texture information and averaging to get brightness, sudden changes in lighting conditions in the image may result in very minute loss of aesthetic appeal but our method is simple and effective. As mentioned earlier, we make use of an online metric to check the quality of our tone mapped image. In order to compare our approach with the other approaches, we present the results of the dynamic range (in)dependent quality metric in Fig. 4 for all the set of images shown in Fig. 3 [10]. This shows that the proposed approach leads to very less distortion in the final tone mapped images. The color codes for various errors can be seen in Fig. 2.

We performed the experiments in MATLAB environment on a laptop that runs 64 bit Windows 7 and has Intel core i5 (1.7 GHz) processor with 6 GB RAM. From Figs. 3 and 4, we see that our proposed TMO is better than ChiuTMO in terms of overall appearance, FattalTMO, ReinhardTMO and WardHist in terms of visibility, WardHistTMO in terms of non-amplification of invisible contrast and gives real-world HDR like image in comparison with all TMOs while closely matching TumblinRushemeirTMO in terms of maintaining the actual contrast.

5 Conclusion and Future Work

The proposed approach tone maps HDR images based on the local texture and brightness cues. The motivation behind using this technique is its simplicity to produce high-quality tone mapped images. Our approach takes only two parameters as input from the user unlike the other TMO's that require three to four parameters on an average. The proposed approach keeps the details intact of the bright and dark regions in the HDR image in the process. Future scope involves subjective studies of tone mapped images instead of using an online dynamic metric to compare the visual perception and aesthetic appeal of our tone mapped images. Based on the descriptions of the images given by the subjects we would like to work on improving our approach in comparison to other TMOs if the subjects feel a big difference in the quality of the tone mapped images.

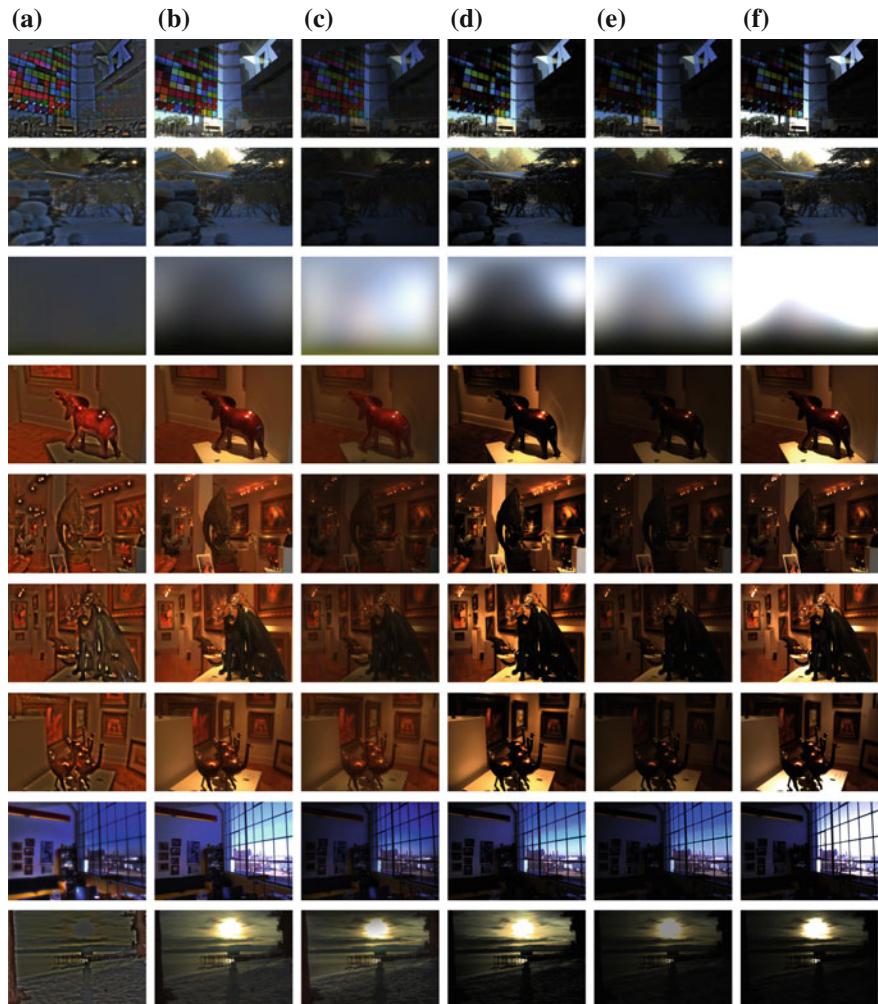


Fig. 3 Tone mapped images for different methods: **a** ChiuTMO [6], **b** TumblinRushmeierTMO [11], **c** FattalTMO [13], **d** WardHistTMO ([1, 12]) **e** ReinhardTMO [23], **f** Proposed TMO



Fig. 4 Dynamic range (in)dependent metric comparison for **a** ChiuTMO [6], **b** TumblinRushmeierTMO [11], **c** FattalTMO [13], **d** WardHistTMO ([1, 12]) **e** ReinhardTMO [23], **f** Proposed TMO. The corresponding color code is shown in Fig. 2

References

1. G. W. Larson, H. Rushmeier, and C. Piatko, “A visibility matching tone reproduction operator for high dynamic range scenes,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 3, no. 4, pp. 291–306, 1997.
2. P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in *ACM SIGGRAPH 2008 classes*, p. 31, ACM, 2008.
3. S. Mann and R. Picard, *Beingundigitalwith digital cameras*. MIT Media Lab Perceptual, 1994.
4. S. K. Nayar and T. Mitsunaga, “High dynamic range imaging: Spatially varying pixel exposures,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, pp. 472–479, IEEE, 2000.
5. K. Devlin, “A review of tone reproduction techniques,” *Computer Science, University of Bristol, Tech. Rep. CSTR-02-005*, 2002.
6. F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *advanced high dynamic range imaging: theory and practice*. CRC Press, 2011.
7. E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” in *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 267–276, ACM, 2002.
8. R. Mantiuk, G. Krawczyk, D. Zdrojewska, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, *High Dynamic Range Imaging*. na, 2015.
9. A. O. Akyüz, K. Hadimli, M. Aydinlilar, and C. Bloch, “Style-based tone mapping for hdr images,” in *SIGGRAPH Asia 2013 Technical Briefs*, p. 23, ACM, 2013.
10. T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, “Dynamic range independent image quality assessment,” in *ACM Transactions on Graphics (TOG)*, vol. 27, p. 69, ACM, 2008.
11. J. Tumblin and H. Rushmeier, “Tone reproduction for realistic images,” *Computer Graphics and Applications, IEEE*, vol. 13, no. 6, pp. 42–48, 1993.
12. G. Ward, “A contrast-based scalefactor for luminance display.”
13. R. Fattal, D. Lischinski, and M. Werman, “Gradient domain high dynamic range compression,” in *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 249–256, ACM, 2002.
14. J. J. McCann and A. Rizzi, *The art and science of HDR imaging*, vol. 26. John Wiley & Sons, 2011.
15. P. Choudhury and J. Tumblin, “The trilateral filter for high contrast images and meshes,” in *ACM SIGGRAPH 2005 Courses*, p. 5, ACM, 2005.
16. F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” *ACM transactions on graphics (TOG)*, vol. 21, no. 3, pp. 257–266, 2002.
17. E. Reinhard, T. Pouli, T. Kunkel, B. Long, A. Ballestad, and G. Damberg, “Calibrated image appearance reproduction,” *ACM Transactions on Graphics (TOG)*, vol. 31, no. 6, p. 201, 2012.
18. R. Mantiuk, A. Tomaszecka, and W. Heidrich, “Color correction for tone mapping,” in *Computer Graphics Forum*, vol. 28, pp. 193–202, Wiley Online Library, 2009.
19. M. Ashikhmin, “A tone mapping algorithm for high contrast images,” in *Proceedings of the 13th Eurographics workshop on Rendering*, pp. 145–156, Eurographics Association, 2002.
20. M. W. Tao, M. K. Johnson, and S. Paris, “Error-tolerant image compositing,” *International journal of computer vision*, vol. 103, no. 2, pp. 178–189, 2013.
21. S. F. El-Hakim, L. Gonzo, M. Picard, S. Girardi, A. Simoni, E. Paquet, H. L. Viktor, and C. Brenner, “Visualisation of highly textured surfaces.” in *VAST*, pp. 203–212, 2003.
22. M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, “A standard default color space for the internetsrgb, 1996,” URL <http://www.w3.org/Graphics/Color/sRGB>, 2012.
23. E. Reinhard, W. Heidrich, P. Debevec, S. Pattanaik, G. Ward, and K. Myszkowski, *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010.

Pre- and Post-fingerprint Skeleton Enhancement for Minutiae Extraction

Geevar C. Zacharias, Madhu S. Nair and P. Sojan Lal

Abstract Automatic personal identification system by extracting minutiae points from the thinned fingerprint image is one of the popular methods in a biometric system based on fingerprint. Due to various structural deformations, extracted minutiae points from a skeletonized fingerprint image may contain a large number of false minutiae points. This largely affects the overall matching performance of the system. The solution is to validate the minutiae points extracted and to select only true minutiae points for the subsequent matching process. This paper proposes several pre- and post-processing techniques which are used to enhance the fingerprint skeleton image by detecting and canceling the false minutiae points in the fingerprint image. The proposed method is tested on FVC2002 standard dataset and the experimental results show that the proposed techniques can remove false minutiae points.

Keywords Fingerprint • False minutiae structures • Minutiae validation

1 Introduction

Personal identification system using fingerprint is one of the widely used biometric systems because of its availability, uniqueness, and inexpensiveness. A fingerprint has a rich structure as it is composed of several ridges and valleys. Personal identi-

G.C. Zacharias (✉)

Department of Computer Applications, MES College of Engineering,
Kuttippuram, Kerala, India
e-mail: geevarcz@gmail.com

M.S. Nair

Department of Computer Science, University of Kerala, Kariavattom,
Thiruvananthapuram 695581, Kerala, India
e-mail: madhu_s_nair2001@yahoo.com

P. Sojan Lal

School of Computer Sciences, Mahatma Gandhi University,
Kottayam, Kerala, India
e-mail: sojanlal@gmail.com

fication using fingerprint is carried out either by extracting global features like core and delta points or local features like minutiae from fingerprint image. The most important minutiae types are ridge ending and ridge bifurcation.

Various approaches proposed for automatic minutiae extraction in the literature are carried out either directly from the gray-level image [1, 2], or from the binarized image [3, 4] or from a thinned skeleton image [5, 6]. Extracting minutiae from the thinned image has advantages like computational efficiency and simplicity over the other methods. Usually, the minutiae extraction is carried out from a binarized thinned image. Consistent extraction of these features is crucial for automatic fingerprint recognition as these extracted feature points are matched against fingerprint database for personal identification.

There are several factors that determine the effective extraction of minutiae points from a fingerprint image. The quality of the fingerprint image is largely affected by the way the fingerprint is acquired. Different conditions like quality of the fingerprint scanners, excessive dryness, wetness, scars, and cuts affect the quality of the fingerprint image. The pre-processing steps like binarization and thinning can cause connectivity issues in the fingerprint ridges. These problems can lead to the extraction of a large number of false minutiae points, reducing the overall system accuracy. Hence, it is necessary to apply a post-processing technique to validate and eliminate false minutiae points.

Most of the post-processing algorithms in the literature make use of the statistical and structural characteristics of the fingerprint ridges to eliminate the false minutiae points. Zaho [6] exploited the duality property of the ridge and valley structure of the fingerprint image to eliminate short spurs, holes, and bridges. Ratha et al. [3] proposed heuristic rules to eliminate false minutiae points. Tico [5] made use of the structural properties of ridges to validate minutiae points and Kim et al. [7] proposed combination of ridge orientation and structural properties to eliminate the false minutiae points.

The paper is organized as follows: Sect. 2 describes the different pre-processing steps required to extract minutiae from an input fingerprint image. Section 3 elaborates different post-processing techniques to remove the false minutiae points from the fingerprint image. Section 4 gives experimental results and Sect. 5 draws the conclusion.

2 Pre-processing

A skeleton-based feature extraction method generally consists of binarization and thinning as its pre-processing steps. However, a low-quality image leading to poor binarization and thinning may extract large number of spurious minutiae points. This necessitated the inclusion of fingerprint enhancement in the pre-processing step, where it corrects broken ridge connectivity. After enhancement, fingerprint features like minutiae points and angles (ridge angle and bifurcation angle) are extracted. The following are the different steps in the pre-processing stage:

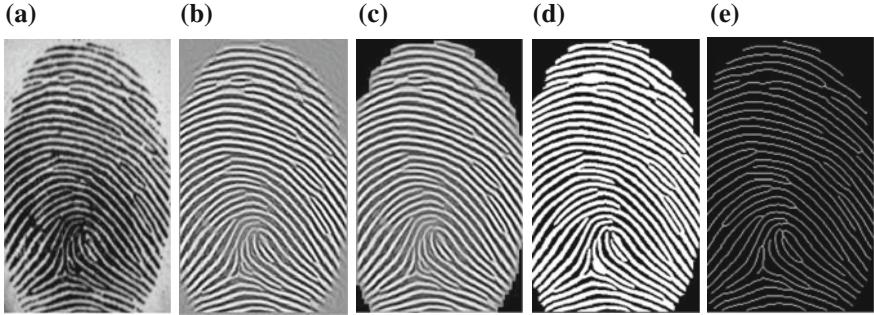


Fig. 1 Pre-processing results: **a** Original fingerprint image. **b** Enhanced image. **c** Image after segmentation. **d** Binarized image. **e** Thinned image

2.1 Fingerprint Enhancement and Thinning

Short-time Fourier transform (STFT) method introduced by Chikkerur et al. [8] is used to enhance the fingerprint image. This method collects many intrinsic properties like ridge orientation and frequency of the fingerprint to enhance the image. Fingerprint ridges can be segmented from its background using the derived region mask that distinguishes between ‘recoverable’ and ‘unrecoverable’ regions of the fingerprint image [8]. A local adaptive thresholding algorithm [9] is used to compute the binary image from the enhanced gray-scale fingerprint image. The skeleton of the fingerprint image is obtained from the binary image by the process of thinning where pixels are eliminated from the boundaries of fingerprint ridges without destroying the connectivity [10, 11]. A rule-based thinning algorithm for character recognition proposed by Ahamed and Ward [10] is used for fingerprint thinning. Later Patil et al. [11] extended the rule set specifically for fingerprint images. Figure 1 shows the images in the different stages of pre-processing steps.

2.2 Minutiae Extraction

Minutiae points are extracted using the concept of crossing number (CN) [3, 6]. The fingerprint skeleton image is scanned and all the pixels are labeled by the properties of CN which is defined in Eq. 1:

$$CN = \frac{1}{2} \sum_{i=1}^8 |P_i - P_{i+1}| \quad (1)$$

where each P'_i 's represent the neighboring binary pixel values of P and $P_1 = P_9$.

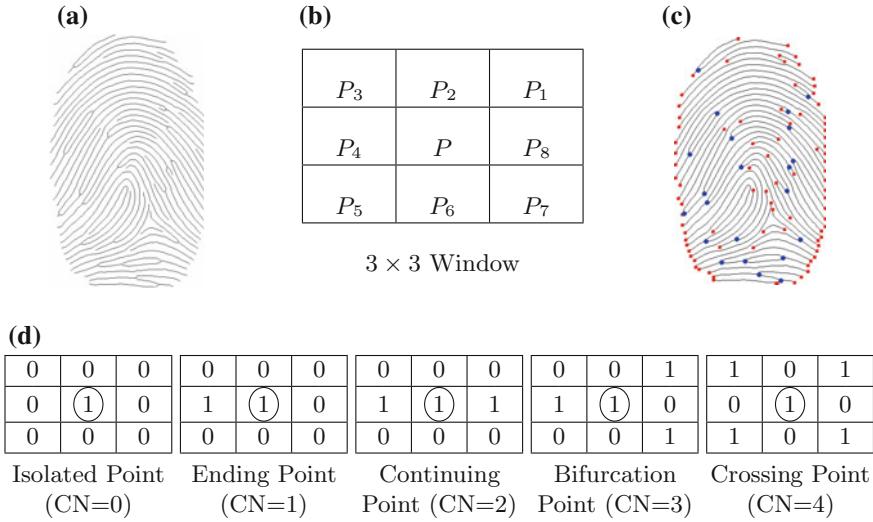


Fig. 2 Minutiae point detection process: **a** Thinned fingerprint image. **b** 3×3 window used for counting CN **(c)** Image with minutiae points marked: end points marked with '■', bifurcation and crossing points with '●'. **d** An example neighborhood values for detecting various CN properties for the pixel P .

Using the CN values, each pixel is classified as isolated point, ending point, continuing point, bifurcation point, or crossing point. Since the isolation points ($CN = 0$) correspond to the noise pixels in the fingerprint valley, this is treated as fingerprint background. A *minutiae map image* (see Fig. 4a) is created by recording the different values of CN. A minutiae map image is defined as an image with each pixel valued 0 through 4, where 0 represents the background and values 1 through 4 represent different CN values as shown in Fig. 2d. A fingerprint image with all the minutiae points marked is shown in Fig. 2c.

2.3 Angles

Two different angles are measured, along with the minutiae extraction step, which can be used to identify the structural properties of the fingerprint ridge segment: a *ridge angle* for each individual ridge segment and a *bifurcation angle* for each bifurcation point. The ridge angle is computed (see Fig. 3b) as a distance-averaged angle over a distance D from a bifurcation point. The bifurcation angle is computed (see Fig. 3b) as the average of the ridge angles associated to a bifurcation point.

For each ridge segment r , the ridge angle (ψ_r) is defined as

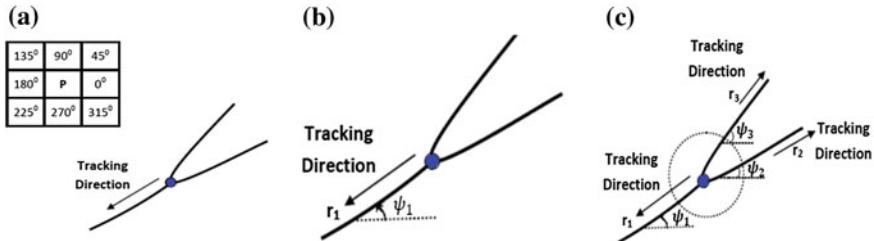


Fig. 3 Angles: **a** A sample ridge segment and 3×3 direction window for finding pixel angle (γ_i). **b** Ridge angle (ψ_r) of a ridge segment. **c** Bifurcation angle (ϕ_b) measured over a distance of radius from the bifurcation point shown in ‘●’

$$\psi_r = \frac{1}{D} \sum_{i=1}^N \gamma_i \quad (2)$$

where γ_i is the associated angle of the ridge skeleton pixel i (45° step in a 3×3 neighborhood), D is taken to be half of the average inter-ridge segments' distance, and N is the number of pixels in the ridge skeleton within the distance D .

For each bifurcation point b , the bifurcation angle (ϕ_b) is defined with respect to the ridge angles ψ_r (for $r = 1, 2, 3$ near a bifurcation point) as

$$\phi_b = \frac{1}{3} \sum_{r=1}^3 \psi_r \quad (3)$$

3 Post-processing Algorithms

The number of minutiae points detected in the pre-processing step is usually much larger than the genuine minutiae points in the fingerprint image. Various structural deformations may cause the pre-processing step to extract spurious minutiae points such as spur, bridge, lake, and island as shown in Fig. 5. Therefore, a post-processing algorithm is required to reliably identify the genuine minutiae from the extracted minutiae points.

Most of the current post-processing algorithms that work on the thinned image require an additional scan over the skeleton to validate the genuine minutiae points. As processing time is critical, our proposed algorithm avoids this additional processing using the minutiae map image. Post-processing algorithms can be categorized into two: first is the minutiae validation algorithm which removes both ridge ending points and ridge bifurcations points that are ambiguous; second is the false minutiae elimination where different minutiae points detected with respect to the false minutiae structures are identified and removed.

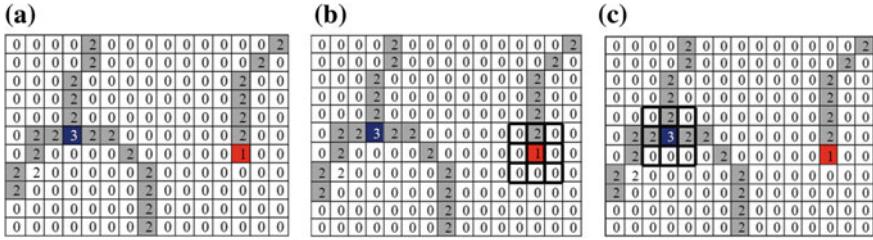


Fig. 4 Minutiae validation: **a** Part of the minutiae map image where each value represents the CN property of each pixel location. **b** Ending point validation using a 3×3 window. **c** Bifurcation point validation using a 3×3 window

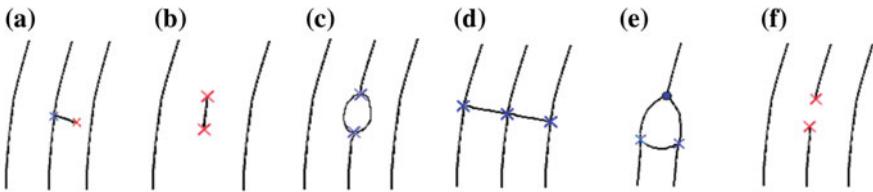


Fig. 5 False minutiae structures: **a** Spike. **b** Spur. **c** Hole. **d** Bridge. **e** Triangle. **f** Broken ridge. False bifurcation points are marked as ‘ \circlearrowleft ’, false endpoints are marked as ‘ \times ’ and true bifurcation points are marked as ‘ $*$ ’

3.1 Minutiae Validation Algorithm

Some candidate minutiae points are false even though they satisfy the CN property. This is because of the ambiguous pixels generated by the thinning algorithm near the candidate minutiae points. Ideally, the thinning algorithm should give only a one-pixel-wide fingerprint ridge image.

The detected candidate minutiae points (Ending and Bifurcation Points) are validated by making a full clockwise scan along eight neighborhood points using a 3×3 window centering the candidate minutiae point in the minutiae map image and counting the number of transitions (T_{02}) from 0 to 2. In the minutiae map image, the ridge ending points and the ridge bifurcation points are represented by CN = 1 and CN = 3, respectively.

For each candidate minutiae point P in the minutiae map image, perform the following steps:

1. Candidate minutiae point is a ridge ending point, if and only if there are only transitions involving 0 and 2 in the eight neighborhoods of the point and the number of 0 to 2 transitions is 1 (i.e., $T_{02} = 1$). If not, mark it as false ridge ending point (see Fig. 4b).
2. Candidate minutiae point is a ridge bifurcation point, if and only if there are only transitions involving 0 and 2 in the eight neighborhoods of the point and the number of 0 to 2 transitions is 3 (i.e. $T_{02} = 3$). If not, mark it as false ridge bifurcation point (see Fig. 4c).

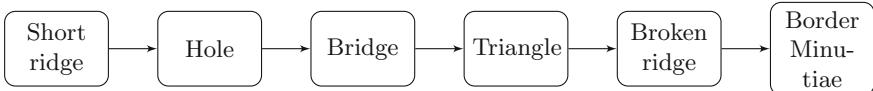


Fig. 6 Removal order of false minutiae structures

3.2 False Minutiae Removal

The false minutiae structures (see Fig. 5) can be categorized either under the ones which appear in the ridge (for e.g., Bridge, Ladder, Broken ridge, Hole, Spike, Spur, Triangle) or at the border of the image where every fingerprint ridge ends. However, ladder is a special case of bridge and both spike and spur are treated as short ridge structures. Some of these structures are easy to identify while others are very difficult due to their structural properties. We propose different algorithms to eliminate these structures. In each stage, the algorithm either removes endpoints or bifurcation points or both that are involved in the false minutiae structure. Since the removal of one false minutiae structure may affect the performance of detecting the others, it is important to specify the correct order of processing. In this paper, the order of detecting and removing false minutiae structure is given in Fig. 6.

3.2.1 Short Ridge Structure Detection

Short ridge structures occur in a thinned fingerprint image because of the noises present in the fingerprint image or they may be introduced by false binarization or thinning [5]. Both spur and spike can be treated as short ridge structure and can be eliminated in a single ridge tracing. Start tracing the ridges in the minutiae map from a ridge endpoint ($CN = 1$) until it meets another endpoint ($CN = 1$) or a bifurcation point ($CN = 3$). If the distance traversed is within a threshold (T_1), these two minutiae points are considered to be false minutiae. The former case is identified as spur and latter is treated as spike.

3.2.2 Hole Structure Detection

Hole structure may occur due to a wide ridge structure [5]. Ridge pores can also cause the hole structure. Hole structure can be detected by starting from a bifurcation point ($CN = 3$) and tracing the three individual ridges. If any two of these three ridges meet at another bifurcation point ($CN = 3$) and if the distance between these two bifurcation points are within a certain threshold (T_2), it is treated as a hole structure and both the bifurcation points are removed.

3.2.3 Bridge Structure Detection

The bridge and ladder structures usually occur if the fingerprint ridges come very close. Considering the fact that the bridge structure is a connection between two fingerprint ridges that are running almost in parallel, it can be eliminated by comparing its bifurcation angle.

- Step 1: Start tracing each ridge segment starting from an end point ($CN = 1$) until it encounters a bifurcation point ($CN = 3$) A .
- Step 2: If one of the bifurcated ridge segments meets another bifurcation point B , calculate bifurcation angles ϕ_A and ϕ_B as defined in Eq. 3.
- Step 3: If the distance between the points A and B is less than a threshold (T_3) and the difference between ϕ_A and ϕ_B is less than a specified angle ($\frac{\pi}{4}$ used in this paper), then the two bifurcation points are identified as a bridge structure.

3.2.4 Triangle Structure Detection

In a fingerprint image, triangular structure is formed when a bridge structure comes near a bifurcation point. The triangular structure is identified when the ridge tracing algorithm detects three bifurcation points within a threshold distance. To remove the triangular structure, the algorithm needs to identify only two false bifurcation points from the detected three bifurcation points. Now, by applying the steps for detecting the bridge structure, the algorithm will detect the two false bifurcation points formed by the bridge structure by its property.

3.2.5 Broken Ridge Structure Detection

A broken ridge may be caused by cuts and scars in the fingerprint. Due to this, two false ridge endpoints will be detected. Two endpoints $A(x_1, y_1)$ and $B(x_2, y_2)$ are considered to be the part of the broken ridge if they satisfy all the following conditions:

- Step 1: distance between two endpoints A and B is less than a threshold (T_4).
- Step 2: the tangent orientation of the points A and B , $\tan^{-1}(\frac{y_2-y_1}{x_2-x_1}) \approx \frac{\pi}{2}$. This ensures that these points are connectable.
- Step 3: $\psi_{r_1} - \psi_{r_2} > \frac{\pi}{4}$, where r_1 represents the ridge segment with the endpoint A , r_2 represents the ridge segment with the endpoint B , and ψ_{r_1}, ψ_{r_2} are computed using Eq. 2. This ensures that the two ridge segments flow in the opposite direction.

Table 1 Typical threshold values proposed by [7]: *freq* indicates fingerprint ridge frequency computed at fingerprint enhancement stage

Threshold label	Phase	Threshold value
T_1	Short ridge	$1.7/freq$
T_2	Hole	$2/freq$
T_3	Bridge	$1.5/freq$
T_4	Broken ridge	$2/freq$
T_5	Border minutiae	$2.5/freq$

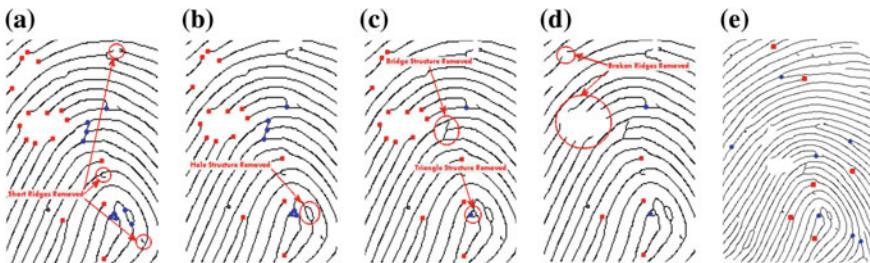


Fig. 7 Result of different post-processing techniques after removing: **a** short ridges. **b** Holes. **c** Bridges and Triangles. **d** Broken ridges. **e** True minutiae points after removing border minutiae

3.2.6 Border Minutiae Points Detection

This is the final stage in the false minutiae detection. Border minutiae points are the points in the fingerprint image where ridges end. Image border is determined by the derived region mask in the fingerprint enhancement stage which is used to segment the fingerprint image. In this stage all the minutiae points within a certain distance threshold (T_5) from the image border are removed.

In this paper, different threshold values are used at different stages to determine the false minutiae structures. Since the fingerprint ridge structure may change using several deformations, we have used an adaptive threshold value proposed by Kim et al. [7] rather than static value. Table 1 gives the different threshold values used.

Figure 7a–e shows the intermediate result of different false minutiae detection stages and Fig. 8a–d shows the fingerprint image with all true minutiae points marked after removing the false minutiae structures (bottom row).

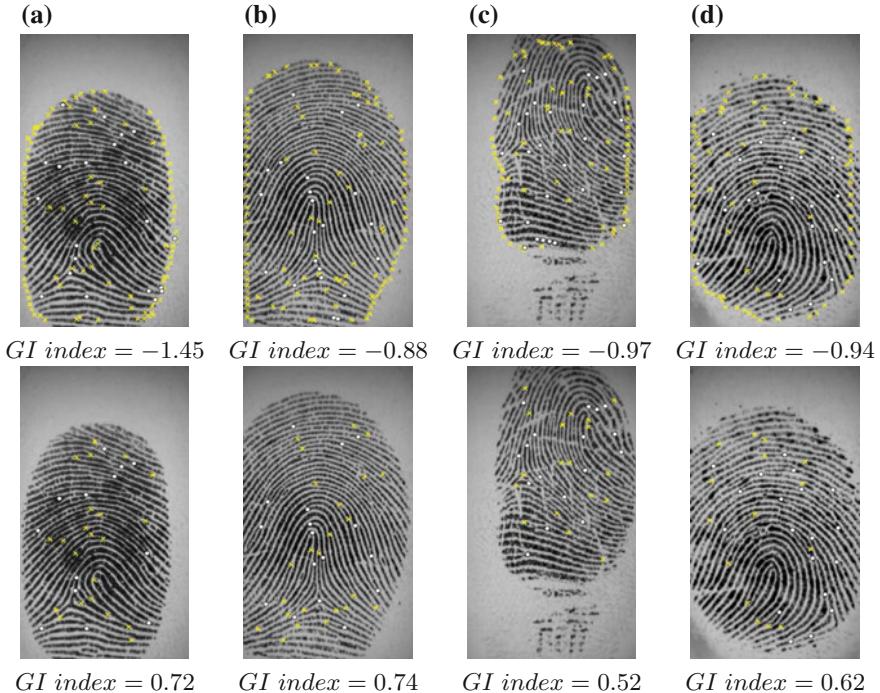


Fig. 8 Results of final minutiae extracted from some fingerprint images in the FVC2002 Db2_a dataset: Endpoints are marked as ‘ \times ’ (shown in yellow color) and bifurcation points are marked as ‘ \circ .’ Top row shows images with all the initial minutiae points marked. Bottom row shows images with true minutiae points marked by removing all the false minutiae points. The *GI index* value before and after processing shows that a large number of false minutiae points are eliminated after processing

4 Results and Discussions

To quantitatively measure the performance of our post-fingerprint enhancement algorithm we have used the *Goodness Index* (GI) proposed by [3] by taking the true minutiae points obtained from the fingerprint image using a domain expert and the extracted minutiae points after post-processing. We have used the enhanced fingerprint image from the pre-processing step to extract the final minutiae points. The goodness index is defined as

$$GI = \frac{P - D - I}{T} \quad (4)$$

where P is the paired minutiae points in the whole fingerprint image, D is the missed minutiae points (includes both dropped (d) and type exchanged minutiae points (e)), I is the spurious minutiae points, and T is the total number of true minutiae points.

Table 2 GI index value for a dataset of 10 fingerprint images

Fingerprint	T	P	D		I	GI
			d	e		
DB2_B_101_1.tif	41	36	5	9	7	0.37
DB2_B_102_7.tif	45	41	4	2	1	0.76
DB2_B_103_5.tif	22	19	3	2	3	0.50
DB2_B_104_7.tif	27	22	5	3	6	0.30
DB2_B_105_2.tif	33	28	5	7	1	0.45
DB2_B_106_1.tif	34	30	4	7	2	0.50
DB2_B_107_3.tif	34	29	5	6	4	0.41
DB2_B_108_5.tif	31	23	8	2	2	0.35
DB2_B_109_7.tif	23	18	5	3	4	0.26
DB2_B_110_8.tif	41	36	5	5	9	0.41

Since the minutiae points are extracted from the thinned fingerprint image, there may be some deviation in the distance from the minutiae points marked by the human experts. Therefore, an extracted minutiae point $m_1(x_1, y_1)$ is said to be paired with the true minutiae point $m_2(x_2, y_2)$ marked by the human expert if the euclidean distance error between these points is not larger than 10 pixels (average fingerprint ridge width). A high value of GI indicates a high performance. The maximum value of GI is 1, and will be achieved when $P = T, D = 0, I = 0$, i.e., all extracted minutiae points are true, no missed points, and no spurious points. Table 2 gives the GI values of a set of 10 randomly selected fingerprint images from the FVC2002 Db2_b standard dataset [12]. The GI index for this set ranges from 0.26 to 0.76 with an average of 0.43, which is better than the result reported in [4, 7].

We have compared the performance of our proposed fingerprint enhancement algorithm for minutiae extraction with a popular public-domain MINDTCT minutiae detector from NBIS (NIST Biometric Image Software) developed by Institute of Standards and Technology (NIST) [13] and a method proposed by Shi et al. [4] by calculating the average error rates of dropped ($\frac{d}{T}$), type exchanged ($\frac{e}{T}$), spurious ($\frac{I}{T}$) minutiae, and total error rate. Table 3 gives the result of this comparison. Even though the dropped minutiae rate is higher when compared with other methods, our algorithm is able to eliminate large number of spurious minutiae points. The overall performance of our proposed algorithm is clearly evident when we compare the total error rate. It is important to remove spurious minutiae as these are the false feature points that may adversely affect the overall fingerprint matching accuracy.

To further evaluate the performance of our proposed algorithm, we have conducted another experiment on a set of 100 randomly taken fingerprint images from the standard FVC2002 Db2_a dataset [12]. It contains 800 fingerprint images from 100 different fingers (eight images from each finger). Table 4 shows the result by comparing our algorithm with MINDTCT [13] and Shi et al. [4]. The performance

Table 3 Comparison of average error rates

Method	Dropped	Type exchanged	Spurious	Total error
Proposed	0.1480	0.1389	0.1178	0.4047
Shi and Govindaraju [4]	0.1178	0.1359	0.2507	0.5044
MINDTCT [13]	0.0392	0.1087	0.4561	0.6040

Table 4 Comparison of average error rates on FVC2002 Db2_a dataset

Method	Dropped	Type exchanged	Spurious	Total error
Proposed	0.0967	0.1225	0.1322	0.3514
Shi and Govindaraju [4]	0.0838	0.1096	0.2774	0.4708
MINDTCT [13]	0.0367	0.1161	0.4838	0.6366

comparisons from Tables 3 and 4 picturize a consistent and reliable performance of the proposed algorithm, as compared to the other two, irrespective of the number of samples selected from varied dataset.

5 Conclusion

In this paper, we have presented few pre- and post-processing algorithms to enhance the fingerprint skeleton image for fingerprint minutiae extraction. Our proposed algorithm can validate true end, bifurcation points and can detect many false minutiae structures like bridge, broken ridge, hole, spike, spur, and triangle. The experimental results show that the proposed algorithm eliminates a large number of false minutiae points.

References

1. Jiang, X., Yau, W.Y., Ser, W.: Detecting the fingerprint minutiae by adaptive tracing the gray-level ridge. *Pattern Recognition*. 34(5), 999–1013 (2001)
2. Gao, X., Chen, X., Cao, J., Deng, Z., Liu, C., feng, J.: A Novel Method Of Fingerprint Minutiae Extraction Based On Gabor Phase. In: Proc. IEEE International Conference on Image Processing, pp. 3077–3080 (2010)
3. Ratha, N.K., Chen, S., Jain, A.K.: Adaptive flow orientation-based feature extraction in fingerprint images. *Pattern Recognition*. 28(11), 1657–1672 (1995)
4. Zhixin Shi, Venu Govindaraju: A chaincode based scheme for fingerprint feature extraction. *Pattern Recognition Letters*. 27, 462–468 (2006)
5. Tico, M., Kuosmanen, P.: An algorithm for fingerprint image postprocessing. In: Proc. of the Thirty-Fourth Asilomar Conference on Signals Systems and Computers, pp. 1735–1739 (2000)

6. Zahoor, F., Tang, X.: Preprocessing and postprocessing for skeleton-based fingerprint minutiae extraction. *Pattern Recognition*. 40(4), 1270–1281 (2007)
7. Kim, S., Lee, D., Kim, J.: Algorithm for detection and elimination of false minutiae in fingerprint image. In: Proc. of the Third International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'01), Halmstad, Sweden, pp. 235–240 (2001)
8. Chikkerur, S., Cartwright, A.N., Govindaraju, V.: Fingerprint enhancement using stft analysis. *Pattern Recognition*. 40(1), 198–211 (2007)
9. Parker, J.R.: Gray level thresholding in badly illuminated images, *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(8), 813–819 (1991)
10. Ahmed, M., Ward, R.: A rotation invariant rule-based thinning algorithm for character recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(12), 1672–1678 (2002)
11. Patil, P., Suralkar, S., Sheikh, F.: Rotation invariant thinning algorithm to detect ridge bifurcations for fingerprint identification. In: 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) 2005
12. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: *Handbook of fingerprint recognition*. Springer, New York (2003)
13. Institute of Standards and Technology, <http://www.nist.gov/itl/iad/ig/fpmv.cfm> (accessed on: 12/05/2015)

Content Aware Image Size Reduction Using Low Energy Maps for Reduced Distortion

Pooja Solanki, Charul Bhatnagar, Anand Singh Jalal
and Manoj Kumar

Abstract On different devices images are often viewed with different resolutions which require image resizing. Resizing images often affects the quality of the images. To better resize images to different resolutions content aware image resizing should be done so that important features are preserved. Seam carving is one such content aware image resizing technique. In this work seam carving is used to downsize an image. This is achieved by carving out an optimal seam (either vertical or horizontal), which contains less information. Each seam removes a pixel from every row (or column) to reduce the height (or width) of the image. To prevent distortion resulting from uniform seam carving, we propose an algorithm that uses a new energy gradient function. In this method minimum of three neighboring pixels is calculated in both energy map and cumulative map and these values are added to find the value of pixel for the new cost matrix.

Keywords Image resizing • Image resolution • Seam carving

1 Introduction

The screen size is the diagonal measurement of the physical screen in inches, while the resolution is the number of pixels on the screen displayed as width by height. There is a lot of variation in the resolution of devices available in the market.

P. Solanki (✉) · C. Bhatnagar · A.S. Jalal · M. Kumar
GLA University, Mathura 281406, India
e-mail: poojasweet.solanki2@gmail.com

C. Bhatnagar
e-mail: Charul@gla.ac.in

A.S. Jalal
e-mail: anandsinghjalal@gmail.com

M. Kumar
e-mail: manoj.kumar@gla.ac.in

It could vary from 240×320 in some smart phones to 2048×1536 in some tablets [1]. Also, while desktop and laptop displays are in landscape, many mobile devices can be rotated to show images in both landscape and portrait. Image resizing is needed to display the same image on different devices. Apart from displaying an image on smaller screens, image size reduction is also needed to create thumbnails [2, 3].

Cropping and scaling are two of the simplest techniques of reducing the size of an image [4, 5]. In cropping, the outer parts of an image are removed which could lead to the removal of some important portions of the image [6]. In scaling, although the content of the image does not change, it affects the important details as much as the unimportant details. Figure 1 shows the effect of cropping and scaling. Thus change in resolution using these techniques reduces the quality of the image [7, 8]. Seam carving, on the other hand, is content aware image resizing [9, 10]. Every object in the image should be scaled down. Seam carving removes more pixels from uninteresting portions of an image as compared to interesting portions of the image [11].

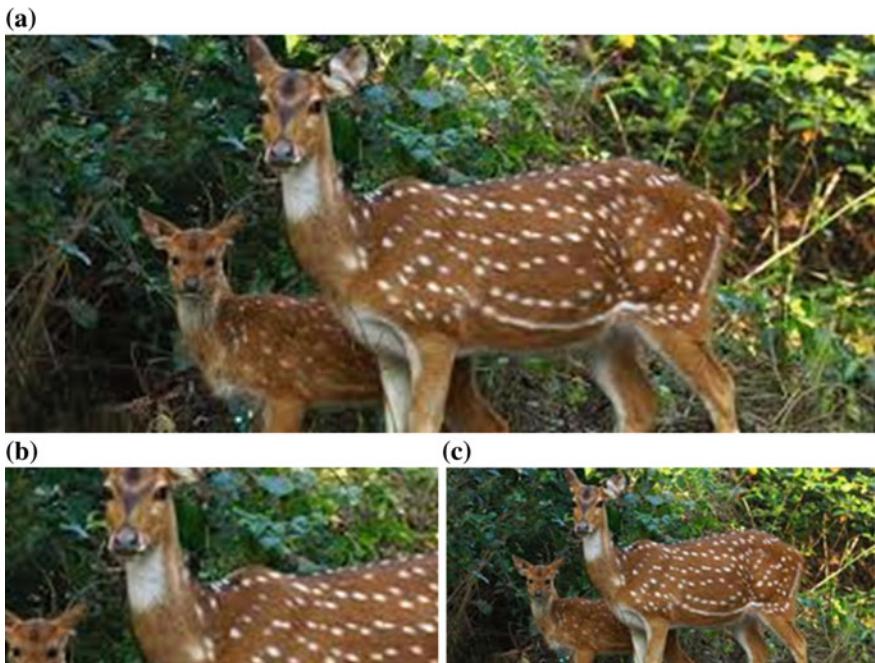


Fig. 1 **a** Original image of size 523×242 , **b** image cropped to size 250×116 , **c** image resized to 250×116

2 Related Work

Avidan and Shamir [12] have used a seam that is an optimal 8-connected path, where optimality is computed using an image energy function. By repeatedly deleting or inserting seams the aspect ratio of an image can be changed. This method however sometimes introduces artifacts because the algorithm selects the seam that has a minimum amount of energy in the image and it ignores the pixels energy. Also, once a seam is removed from the image the entire energy matrix has to be updated.

Rubinstein et al. [9] have proposed method seam carving for video. Seam carving is improved using a 2D seam manifolds from 3D space-time volumes. Instead of the traditionally used dynamic programming, they have used graph cut method. However, the algorithm fails to keep straight lines as straight when resizing.

Saliency maps that assign higher importance to visually prominent whole regions and not just to edges have been used by Achanta and Susstrunk [13]. The algorithm calculates global saliency of pixels using intensity as well as color features. Independent of the number of seams added, the saliency maps are calculated only once. The method fails to give desired results when there are complex objects in the image or when the image contains more than one salient objects.

The algorithm proposed by Domingues et al. [14] uses appearance-based temporal coherence model. It allows improved visual quality, affords greater flexibility, and is scalable for large videos. It highly structures scenes or videos having past paces scenes, and the method sometimes gives unsatisfactory results.

He et al. [15] combine gradient and saliency map which best describe the image content. In this paper optimization method is also provided. This method improves efficiency of seam carving.

3 Image Size Reduction Using Minimum Energy Seams

In this section we discuss the proposed algorithm for reducing the size of an image, resizing the image by sequentially removing the seams, either vertical or horizontal, as required. A vertical seam is an 8-connected path from a pixel in the top row of the image to a pixel in the last row of the image. Analogously we have a horizontal seam, which is an 8-connected path from a pixel in the leftmost column of the image to a pixel in the rightmost column of the image. One vertical (horizontal) seam removal reduces the size of the image by exactly one column (row). Figure 2 shows the result after each step of the proposed algorithm.

Step 1: Calculate the Energy Map of the image

Less prominent pixels are those that blend well with the background. The energy basically represents the importance of the pixel and shows how the pixels are

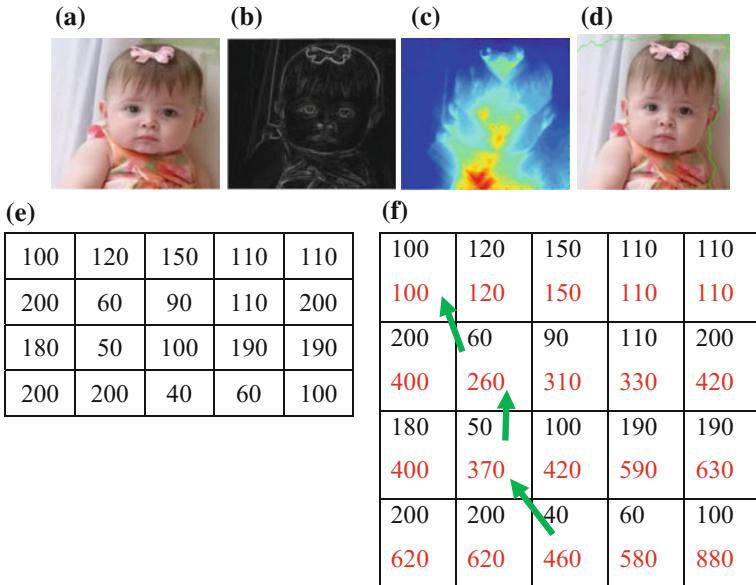


Fig. 2 Steps to remove the optimal seam. **a** Input image, **b** gradient energy image, **c** cumulative energy map, **d** energy map of sub part of the image, **e** *numbers in red* indicate the cumulative energy map of a image subpart. **f** The *green arrows* show the backtracking in dynamic programming

related to their neighbors. For this, we use the Sobel derivative operators to find the energy of the pixels. The smaller the value, less important is the pixel.

Algorithm: Calculate the Energy Map of the Input Image

1. Convert image from RGB into gray scale.
2. Apply Sobel filter for gradient calculation:
 $G_x = [-1 \ 0 \ 1; -2 \ 0 \ 2; -1 \ 0 \ 1]$ and $G_y = [-1 \ -2 \ -1; 0 \ 0 \ 0; 1 \ 2 \ 1]$
3. Energy of the pixel is then calculated as follows:
 $E(x,y) = \sqrt{(G_x^2 + G_y^2)}$

Step 2: Calculate the Cumulative Minimum Energy Map

Dynamic programming is used to compute the cumulative minimum energy map. To compute the lowest energy vertical seam, a 8-connected path of pixels, running from first row to the last row, is found in such a way that one and only one pixel from each row is taken. The cost of a seam is the sum of the energy function along the seam.

Algorithm: Calculate the Cumulative Minimum Energy Map

1. For the pixels in the first row, $C(1, j) = E(1, j)$
2. For $r = 2$: rows
3. {
4. $C(r, 1) = E(r, 1) + [\min(E(r-1, 1), E(r-1, 2)) + \min(C(r-1, 1), C(r-1, 2))]$
 //For pixels in the first column
5. For $c = 2$: col-1
6. {
7. $C(r, c) = E(r, c) + [\min(E(r-1, c-1), E(r-1, c), E(r-1, c+1)) + \min(C(r-1, c-1), C(r-1, c), C(r-1, c+1))]$
8. }
9. $C(r, col) = E(r, col) + [\min(E(r-1, col-1), E(r-1, col)) + \min(C(r-1, col-1), C(r-1, col))]$
 //For pixels in the last column
10. }

Step 3: Remove the Seam with Minimum Energy

From the cumulative energy map, find the vertical seam with the minimum energy and delete the pixels lying on the seam from each row. Analogous technique can be applied to remove the horizontal seam.

Algorithm: Removal of Minimum Energy Seam

1. For vertical seam start from last row in image.
2. Choose the pixel with the minimum cumulative energy in the last raw.
3. Choose the minimum of three neighboring pixels in the row above.
4. Repeat this process until the first row is reached.
5. Delete the connected path of the pixels which is the seam with minimum energy.

4 Results and Discussion

We have tested the proposed method on 100 images in the following categories: Natural scenes, animals, human faces, buildings, and market scenes. Original images were of different sizes. We resized them to 50 % of the original size and compared the proposed results with those of Frankovich and Wong [16]. The results of one image from each category are shown in Fig. 3.

We showed the results of applying [16] and the proposed method on 30 images to ten different viewers. The views were asked whether or not they were satisfied with the results of resizing. The result of the survey is shown in Fig. 4.

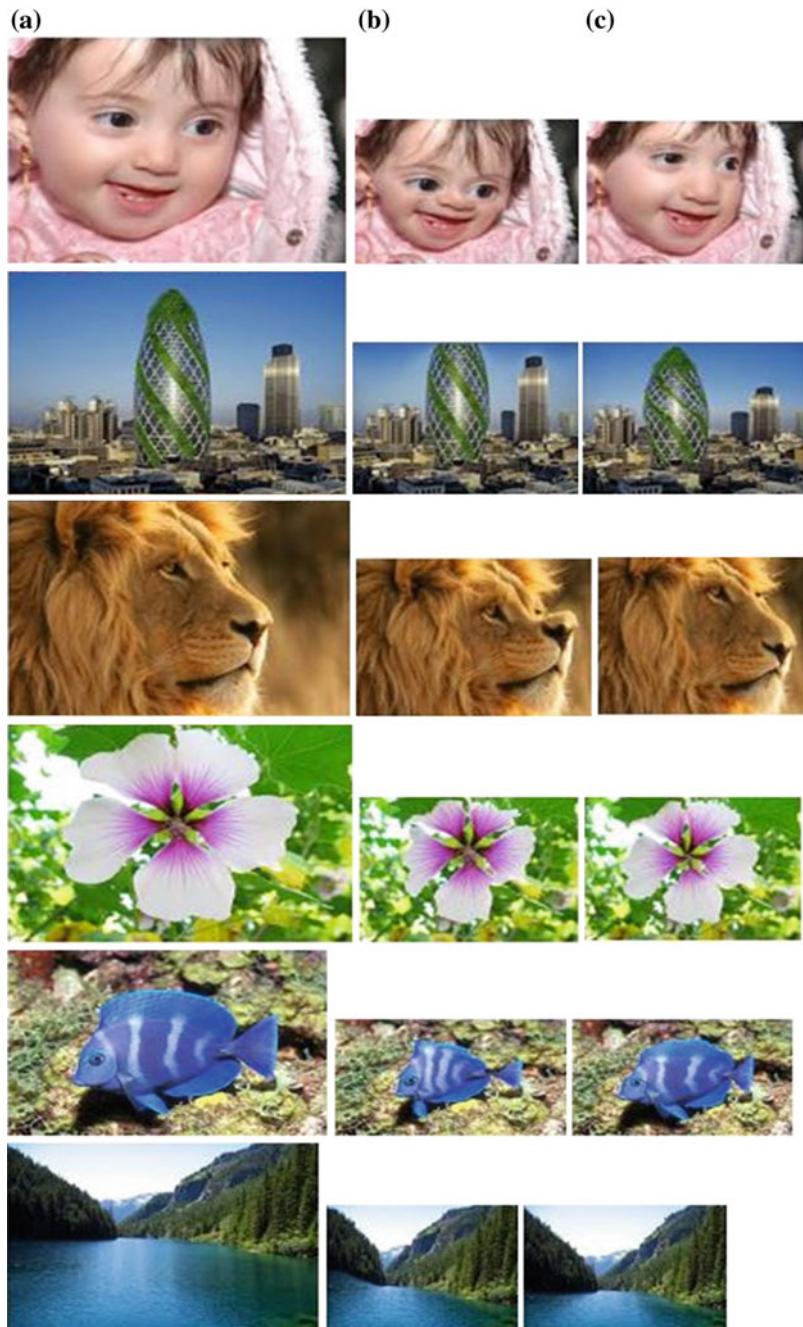
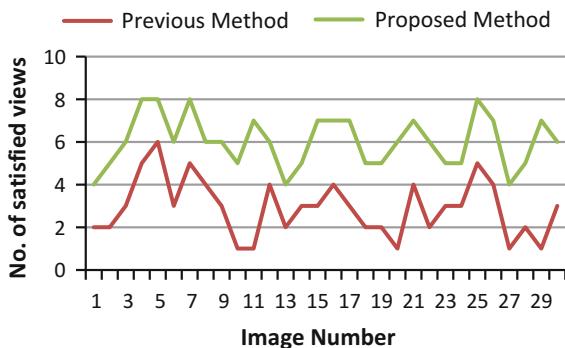


Fig. 3 **a** Original images, **b** results of [16], **c** results of proposed method

Fig. 4 Result of survey of 30 images by 10 viewers



5 Conclusions

With the plethora of display devices available in the market and new ones coming everyday, the need for image resizing techniques has become the call of the day. Seam carving removes minimum energy pixels from the unimportant areas of the image. Various approaches have been proposed by researchers to avoid the distortion in the image using seam carving. Still some distortions in the form of broken straight lines, deformation of object parts, etc. still occur. To minimize such distortions in the image while using seam carving, we use a modified energy function to better handle the image content. This energy function protects the important parts of the image from distortion and reduces the image size. In future, we plan to work on finding a modified energy function that can be used to increase the size of the image with minimum distortion.

References

1. Fan, X., Xie, X., Ma, W. Y., Zhang, H. J., & Zhou, H. Q.: Visual attention based image browsing on mobile devices. In Proceedings of IEEE International Conference on Multimedia and Expo, ICME'03. Vol. 1, pp. I-53, (2003).
2. Chen, L. Q., Xie, X., Fan, X., Ma, W. Y., Zhang, H. J., & Zhou, H. Q.: A visual attention model for adapting images on small displays. Multimedia systems. vol. 9, no. 4, pp. 353–364, (2003).
3. R Samadani, R., Lim, S. H., & Tretter, D.: Representative image thumbnails for good browsing. In IEEE International Conference on Image Processing, ICIP 2007. Vol. 2, pp. II-193, (2007).
4. Liu, F., & Gleicher, M.: Automatic image retargeting with fisheye-view warping. In Proceedings of the 18th annual ACM symposium on User interface software and technology. pp. 153–162, (2005).
5. Setlur, V., Takagi, S., Raskar, R., Gleicher, M., & Gooch, B.: Automatic image retargeting. In Proceedings of the 4th ACM international conference on Mobile and ubiquitous multimedia. pp. 59–68, (2005).

6. Ren, T., Guo, Y., Wu, G., & Zhang, F.: Constrained sampling for image retargeting. In IEEE International Conference on Multimedia and Expo. pp. 1397–1400, (2008).
7. Wang, Y. S., Tai, C. L., Sorkine, O., & Lee, T. Y.: Optimized scale-and-stretch for image resizing. ACM Transactions on Graphics. Vol. 27, no.5, pp. 118, (2008).
8. Mansfield, A., Gehler, P., Van Gool, L., & Rother, C.: Visibility maps for improving seam carving. In Trends and Topics in Computer Vision, Springer Berlin Heidelberg, pp. 131–144, (2012).
9. Rubinstein, M., Shamir, A., & Avidan, S.: Improved seam carving for video retargetting. In ACM transactions on graphics. Vol. 27, No. 3, pp. 16, (2008).
10. Hwang, D. S., & Chien, S. Y.: Content-aware image resizing using perceptual seam carving with human attention model. In IEEE International Conference on Multimedia and Expo. pp. 1029–1032, (2008).
11. Criminisi, A., Perez, P., & Toyama, K.: Object removal by exemplar-based inpainting. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2, pp. II-721, (2003).
12. Avidan, S., & Shamir, A.: Seam carving for content-aware image resizing, In ACM Transactions on graphics, Vol. 26, No. 3, p. 10, (2007).
13. Achanta, R., & Süstrunk, S.: Saliency detection for content-aware image resizing. In 16th IEEE International Conference on Image Processing (ICIP). pp. 1005–1008, (2009).
14. Domingues, D., Alahi, A., & Vandergheynst, P.: Stream carving: an adaptive seam carving algorithm. In 17th IEEE International Conference on Image Processing (ICIP). pp. 901–904, (2010).
15. He, Z., Gao, M., Yu, H., Ye, X., Zhang, L., & Ma, G.: A new improved seam carving content aware image resizing method. In 8th IEEE Conference on Industrial Electronics and Applications (ICIEA). pp. 738–741, (2013).
16. Frankovich, M., & Wong, A.: Enhanced seam carving via integration of energy gradient functionals. IEEE Signal Processing Letters. vol. 18, no. 6, pp. 375–378, (2011).

Artificial Immune Hybrid Photo Album Classifier

Vandna Bhalla and Santanu Chaudhury

Abstract The personal photo collections are becoming significant in our day today existence. The challenge is to precisely intuit user's complex and transient interests and to accordingly develop an adaptive and automated personalized photo management system which efficiently manages and organizes personal photos. This is increasingly gaining importance as it will be required to browse, search and retrieve efficiently the relevant information from personal collections which may extend from many years. Significance and relevance for the user also may undergo temporal and crucial shifts which need to be continually logged to generate patterns. The cloud paradigm makes available the basic platform but a system needs to be built wherein a personalized service with ability to capture diversity is guaranteed even when the training data size is small. An Artificial Immune Hybrid Photo Album Classifier (AIHPAC) is proposed using the nonlinear biological properties of Human Immune Systems. The system does event based clustering for an individual with embedded feature selection. The model is self learning and self evolving. The efficacy of the proposed method is efficiently demonstrated by the experimental results.

Keywords Clonal selection • Antibody • Antigen • Avidity • Affinity maturation

1 Introduction

There is a need for a personalized photo management system which efficiently organizes and manages personal digital photos. Following points need to be noted for personal collections. (1) Personal photos are very different from normal images or videos as they are related to the specific user and are integrated by a context. The user

V. Bhalla (✉) · S. Chaudhury

Indian Institute of Technology, Hauz Khas, New Delhi, India

e-mail: vballa.du@gmail.com

URL: <http://www.iitd.ac.in>

S. Chaudhury

e-mail: schaudhury@gmail.com

will have personal associated memories like for instance the place, the environment, the time etc. specifically relating to the photo clicked. The low level features fail to capture the rich information of personal stories in entirety. A personal collection will typically contain photos of a trip with family or friends or wedding or series of vacation taken with family. (2) The techniques for image retrieval based only on content have been proved unsuccessful as they are unable to connect the semantic import with the content. (3) The photos taken during a particular event, such as a birthday party, show coherence but not like a video sequence. The changes in frames in a conventional video are very strongly related and are highly similar unlike two similar photos where the camera is randomly and frequently tilting, panning or zooming. (4) They are very unlike the MRI/CT scans or other medical images where images are largely similar. (5) Last but not the least the training data sets in these environs are not very large and are inherently diverse. Its growingly becoming more and more laborious to locate a specific photo from a collection of millions using conventional photo organization systems. In this paper we put forth a self learning and self evolving hybrid model for managing personal photos which is adaptive too and can assist users to organize their vast and diverse personal photo collections. In contrast with classical machine learning algorithms the proposed model performs very well even when the size of the training data is not very large. The proposed Artificial Immune Hybrid Photo Album Classifier (AIHPAC) model is inspired by the intuitive and clever information managing technique of the biological immune systems to generate a robust feature set using Clonal Selection principles taking the input images as seeds. This is particularly suitable in the context of personal collections, where for each training sample different points of view are gathered in parallel using clonal selection. The rest of this paper is organized as follows. Section 2 is about the related work. Section 3 describes the AIHPAC model, Sect. 4 tabulates experimental results. Section 5 gives a brief discussion about the clonal effects and the paper concludes with Sect. 6.

2 Related Work

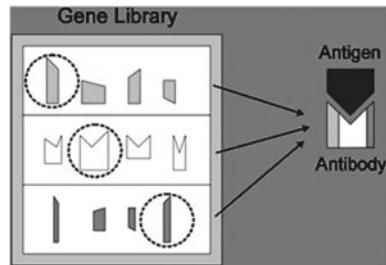
Most personalized systems built so far require explicit user intervention [7]. These current photo management systems are based on text/keywords based annotations [3] which are intuitive but the user needs to acquaint themselves with concepts like class, property relations and instances etc. Early systems, like FotoFile [9] and Photo finder [8], annotate content with keywords and names and use archives to generate tags and annotations. Though the search performance is good but the tedious process of manually annotating each photo by the user is a requirement. Then the issue of capability of knowledge inference where the preceding commented data is inadequate to yield inference for the prospective photos of future. PhotoMesa [1] maximizes the screen-space usage, but here also the photos have to be arranged personally by the user.

The lexically motivated keyword based approach does not resolve multiplicity and semantic relevance remains a issue. Naaman et al. [11] automatically generate additional metadata from each photo and based on a user study and survey identified the useful and relevant categories of contextual metadata for retrieval. The fast visual scanning mechanism like pan and zoom do not scale to manage the large personal photo collections. Some of these techniques perform well for other data scope as per the literature review but their performance is ordinary in personal photo collections. Though tedious, wearisome and time-consuming, yet a lot of researches were dedicated to enhance and ease the annotation tools [5]. Another popular approach is grouping photos using visual features. Using time stamp is also a common but popular approach. Orii et al. [12] shows that congregating on the basis of timestamp makes very little difference for unfamiliar photo collections and in their subsequent work say that it helps browsing experience. This is useful when you want to cluster contiguous photos and not personal photo collections. A classification completely driven by data and based on entropy and evolution was given by Tao Chunmei et al. [14]. The design uses a bottom up mechanism for searching and despite using lesser number of dictum achieves high precision. With Support Vector Machines, the samples near the SVM hyperplanes tend to be misclassified. The classification based on SVM and reverse KNN presented by Chen Li et al. [4] shows a better average forecast accuracy. A decision tree rule based classification by Tan et al. [13] aims at improving precision of classification and Liu et al. [10] gave a model based on KNN for small sample sets. Cao Gen et al. [2] present a locally weighted KNN based weighted classification algorithm. Overall, these algorithms convey that though the supervised techniques for classification have evolved but these are steady specifically under KNN, SVN, Bayesian or similar evolutionary calculation. Our work explores more efficient classification platforms and presents a supervised classification algorithm which simulates the intelligent data processing techniques of the human immune system imitating their distinctive features of organization, adaptability and learning. Shaojin has proposed classification based on AIS but not for personal photo collections [6]. Personal photo collections do not have much training data available class wise and are very different from hyper spectral images or medical images which have been the datasets for most of the past work with artificial immune systems. Kai-En Tsay et al. [15] did develop an organizer but their results are case studies and not statistically significant analysis. Also very broad categories like for instance people/nonpeople and indoor/outdoor were chosen.

3 AIHPAC

Our proposed system uses Clonal Selection principles from Artificial Immune System which we will briefly introduce and then present our model.

Fig. 1 Generation of Diversity



3.1 Artificial Immune System (AIS)

The primary function of our Immune System is to protect the human body from the invading disease causing microorganism called antigens. When a microorganism first attacks, the Adaptive Immunity stores this into memory for future reference and when the same pathogen strikes later it is countered with a more intensified response. The cells that match the best multiply to generate clones. High repetitive mutation continuously evolves these clones and sooner or later a better and more appropriate solution manifests itself. This entire process is termed Clonal Selection. We are inspired by two main ideas which are very pertinent for our framework. Diversity: In the human system, antibodies are generated by highly specialized cell called the B-cells explicitly for a particular antigen. The Gene Library consists of gene fragments which are the building blocks for these antibodies. As the library repository is limited, it does not have genes that can create antibodies for each conceivable antigen. A big heterogeneous collection of antibodies is created by the random combination of the various gene fragments in the library and invariably a specific antibody gets created for every possible antigen, even the new unfamiliar ones. This process is illustrated in Fig. 1.

Avidity: It is a accrued total measure of collective affinities of an antigen with all possible antibodies which are similar to the specific antigen. The biology of the Human Immune System has inspired many Artificial Immune Systems [6]. The Clonal Selection Algorithm, CLONALG, [5] by Castro substantiates that clonal selection theory can help in pattern recognition. The Algorithm is outlined as follows: (1) A community of strength, N , is initialized randomly. (2) Determine the similarity metric of each input pattern I (analogous to an antigen) with every member of N . (3) Choose b best matched members of N and spawn clones of these corresponding to the extent of their similarity with I . Higher number of clones are generated if the matching is higher (4). These clones are now mutated and evolved with a rate proportionate to the degree of match similarity. (5) The evolved and matured clones are finally added to the initial community N . The best n members from this set are stored in the memory. (6) Repeat the algorithm till convergence happens or the stop criteria is achieved.

3.2 The Architecture

Our immune system is able to capture the antigen (foreign elements) through an antibody if one exists. If not, then a process of mutation and proliferation produces the desired antibody. This self-learning and self-adapting immune system inspires our model. Each photo is characterized by its $m \times m$ color correlogram matrix which involves quantization of the image into the color bins and then representing the image by the square matrix where m is the number of bins. We have used this discerning feature in our work because of the following favourable characteristics of color correlogram: (i) It captures the color spatial relationship. (ii) it portrays the global spread of local spatial connection of colors (iii) uncomplicated calculation (iv) the feature dimension is compact and (v) this image feature is found to be resilient to large variations spawned by variations in camera zooming positions, occlusions, background variations, viewing positions etc. The statistical descriptor of color images, the color correlogram has been widely used and is an accepted feature for content based image retrieval (CBIR) systems. The parameters are tuned and adjusted for best results. Inner product metric is used to measure affinities as the relative distance does not provide good performance. The model has training and testing phase. Training phase includes feature extraction.

3.3 Training

The feature set extracted from the Correlogram Matrix of all training images are used as the starting population of each class, N. Training Algorithm is as follows:

- Randomly initialize a population of antibody class (N) and memory cell class (M).
- Select a memory cell pattern from the population M and establish its affinity, A, with every member of set N using Inner Product and select n highest affinity memory cell of M.

The Inner Product between two image I_1 and I_2 is given by:

$$\langle I_1, I_2 \rangle_\gamma = \sum \gamma_{c_i, c_j}^{(k)}(I_1), \gamma_{c_i, c_j}^{(k)}(I_2) / \|\gamma(I_1)\|_2 \|\gamma(I_2)\|_2 \quad (1)$$

where: $i, j \in \{1 \dots m\}$ and, $k \in \{1 \dots n\}$

$$\text{and } \|\gamma(I)\| = \sqrt{\sum_{\substack{i, j \in \{1 \dots m\} \\ k \in \{1 \dots n\}}} [\gamma_{c_i, c_j}^{(k)}(I)]^2}$$

- As per the rules of Clonal Selection, we first clone, then mutate and finally do crossover for each cell separately for each class. Figure 2 illustrates this mechanism. This process produces supplementary data to generate additional

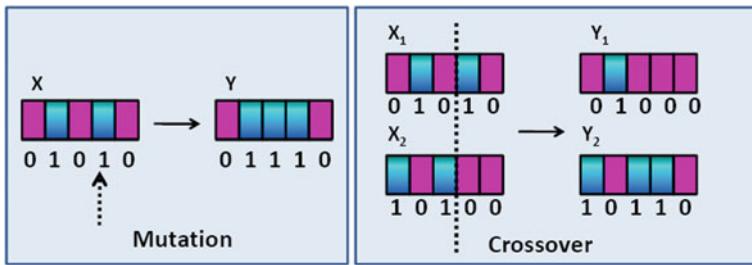


Fig. 2 Mutation and Crossover

features that satisfy the minimum threshold criteria for that class. The total quantity of clones generated is calculated by

$$C_{num} = \eta * A(I_1, I_2) \quad (2)$$

Here the affinity between two photos is A and the cloning constant is η . Higher the affinity of match the greater the clone stimulus gets, the more the cloning number is. If the similarity index which shows the degree of matching is higher then greater number of clones are generated as the stimulus produced for cloning gets higher. On the other hand if the similarity is less then the number of clones produced is less which is in sync with the human immune system process mechanism. Accordingly MF, the mutation frequency, is given by

$$MF = \alpha * 1/A(I_1, I_2) \quad (3)$$

where α is mutation constant and higher the affinity of match, the smaller the clone stimulus gets, the lower the mutation frequency is. On the contrary, the mutation frequency is higher.

- The generated new features, m , are added to the initial population N . To keep a check on the size and the algorithm convergence speed, the elimination principle of biological system is followed. The elimination principle calculates the similarity of an antigen with every antibody present in the community N . The weak affinity members will automatically get removed and this way the community size will be maintained.
- Repeat until all patterns in initial memory set are processed.

After the completion of training phase we have a large pool of feature sets (antibodies) for each class. N images from each class result in around $3N = (N + m)$ valid features (antibodies) for that class.

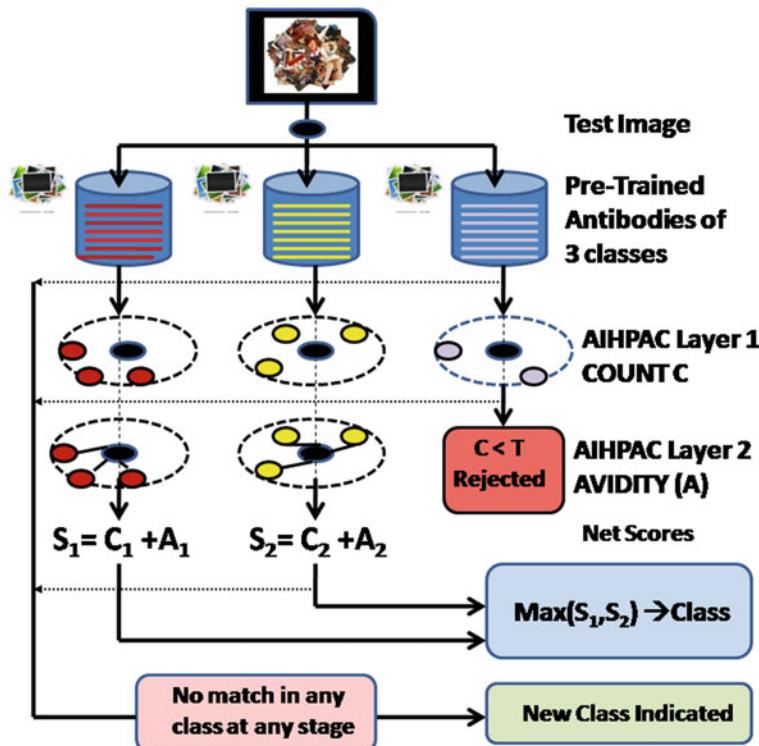


Fig. 3 AIHPAC Model

3.4 Testing

Take the testing image Correlogram Matrix. The model performs classification using a 2-layered approach, Fig. 3. The first layer of innate classification matches the testing image (antigen) with the pre trained antibodies of each class. We predefine a threshold value for the number of matches and a count C for each class is ascertained. If the count C is above the threshold then it qualifies for Layer-2 testing and are subsequently fortified by the acquired knowledge of combined strengths of various matches within a class. This second phase for acquired knowledge computes the average affinity of many strengths of all authorized matches which have cleared C . We calculate the average inner product score between the test image and every member of the given class which has passed the first layer criterion. This is called Avidity, A , and it is established for every class. The class is finally determined on the basis of this net aggregate score $S = (C + A)$. The test images used are different from training images and belong to different individuals. The model has the capability to generate a new class. If none of the existing classes come up with an acceptable score at

layer 1 or 2 then the system suggests a new class and the particular test image may initialize a new class.

4 Experimental Results

The model was tested for three different classes i.e. Picnic, Wedding and Conference. There are no standard datasets available for our domain. We have taken images from the INTERNET and also from some personal albums. Some samples images used are shown in Fig. 4. The results are based on a system correctly recognizing the picture belonging to the specific class which have been aggregated in results under True Acceptance (TA) in Table 1 and not belonging to the class and hence rejected correctly have been aggregated in results under True Rejection (TR) in Table 2. The number of images used for training the model and number of images used for testing are listed in the third column, Training/TA for Table 1 and Training/TR for Table 2 respectively. The number of antibodies created by the model for each class are given in the fourth column. The following columns show classification performance with



Fig. 4 Sample Images

Table 1 Result for true acceptance

S.N	Class	Training/TA	Antibodies (num)	NN (%)	1-layer AIHPAC (%)	2-Layer AIHPAC (%)
1	Picnic	40/37	110	60	62	74
2	Wedding	30/20	85	67	72	75
3	Conference	40/25	115	74	77	79

Table 2 Result for true rejection

S.N	Class	Training/TR	Antibodies (num)	NN (%)	1-layer AIHPAC (%)	2-Layer AIHPAC (%)
1	Picnic	40/95	110	78	82	85
2	Wedding	30/204	85	77	85	92
3	Conference	40/107	115	75	80	85

Table 3 Result for new class indication

	Picnic	Wedding (%)	Conference (%)	NEW (Birthday) (%)
Classification	79	83	81	87

standard Neural Network, 1-layer and 2-layer AIHPAC model. The 2-layer AIHPAC model shows superiority in performance. We added pictures from a fourth class, Birthday, to the testing set and the model indicated a new class, Table 3.

5 The Effect of Clonal Selection

The Clonal Selection technique is very simple with limited number of specifications which are also completely determined by the user. Literature has proved that clonal theory can be used for solving problems pertaining to pattern recognition. Unlike most of the evolutionary methodologies which are probabilistic, this concept inspired by immune theory is highly deterministic in terms of all its features like cloning, replacement, selection etc. This is a highly desirable concept because it not only helps to distinguish between processes but in addition it also gives discernibility of all that is happening within the model along with a great degree of standardization. The flexible and self modifying capability of the model makes it possible for information to be reused. This saves the system restarting from the beginning every time a new feature set comes in. In-fact the model learns and evolves to become more and more robust with every new exposure. We now present the impact of the configurable components on the model.

Antibody pool size is the absolute quantity of antibodies to be sustained by the system. We used the simple approach of selecting an $m \leq N$. Here N is the memory portion and the remaining size is determined by $N-m$ which is represented by r .

Selection pool size represents those antibodies which have the highest similarity and have been drawn from the initial population.

Clonal factor gives a scaling factor and refers to the total count of clones that are generated. Common values for Clonal factor are $\in (0, 1]$. The algorithm executes search depending on the value of the clonal factor. If the clonal value is low then the algorithm probes search in local regions.

6 Conclusion

The AIS based classification used in our model is particularly suited for supervised learning problems. Clonal selection augments data at the feature level and therefore restricts the parameters to particular regions that assist to capture the input distribution. There are scenarios where the camera position may cause variations in the apparent size proportions of the geometry of the scene and this introduces distortions in detected images. Our model develops a set of memory cells through the utilization of artificial evolutionary mechanism which have the intrinsic ability to handle intra and inter album diversity despite small size of the initial datasets. The final result of the algorithm is the memory pool and it captures the dominant and the essential characteristics of class intuitively at the feature level yielding high classification performance. The AIS inspired classifier has the robust and efficient ability to capture diversity and this is because of the intuitive rules of nature which produce antibodies in a human immune system. Past work on classifications based on biological immune mechanisms primarily use medical images, monuments or hyper spectral images as datasets with very inspiring results but the same algorithms deteriorate in performance when tested on personal photo datasets. This new proposed model is self learning and gives novel results on Personal Photo collections. The classifier does not show a majority class tilt because the number of antibodies produced are irrespective of the initial population of the class. The majority and minority class eventually end up with an equable number of antibodies. This method alleviates the tediousness of manual annotations and their associated complexities. We worked with a few variations in threshold criteria and also experimented with different mutations. The results show that the model proposed is effective and feasible with high classification precision. A new class can be added to the existing set of classes dynamically replicating the behavioural aspects of self-learning and self evolving of human system.

References

1. Bederson, B.B.: Photomesa: a zoomable image browser using quantum treemaps and bubblemaps. In: Proceedings of the 14th annual ACM symposium on User interface software and technology. pp. 71–80. ACM (2001)
2. Cao, G., Ge, X., Yang, L.: Locally weighted naive bayes classification algorithm based on k-nearest neighbour. Jisuanji Yingyong yu Ruanjian 28(9) (2011)
3. Chai, Y., Xia, T., Zhu, J., Li, H.: Intelligent digital photo management system using ontology and swrl. In: Computational Intelligence and Security (CIS), 2010 International Conference on. pp. 18–22 (Dec 2010)
4. Chen, L., Chen, J., Gao, X.T., Wang, L.S.: Classification algorithm research based on support vector machine and reverse k-nearest neighbor. Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications) 46(24) (2010)
5. Elliott, B., Özsoyoglu, Z.M.: A comparison of methods for semantic photo annotation suggestion. In: Computer and information sciences, 2007. iscis 2007. 22nd international symposium on. pp. 1–6. IEEE (2007)

6. Feng, S.: Supervised classification algorithms based on artificial immune. In: Natural Computation (ICNC), 2012 Eighth International Conference on. pp. 879–882. IEEE (2012)
7. Guldogan, E., Olsson, T., Lagerstam, E., Gabbouj, M.: Instance based personalized multi-form image browsing and retrieval. *Multimedia tools and applications* 71(3), 1087–1104 (2014)
8. Kang, H., Shneiderman, B.: Visualization methods for personal photo collections: Browsing and searching in the photofinder (2000) (2005)
9. Kuchinsky, A., Pering, C., Creech, M.L., Freeze, D., Serra, B., Gwizdka, J.: Fotofile: a consumer multimedia organization and retrieval system. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. pp. 496–503. ACM (1999)
10. Liu, Y., Niu, H.: Knn classification algorithm based on k-nearest neighbor graph for small sample. *Computer Engineering* 37(9), 198–200 (2011)
11. Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., Paepcke, A.: Context data in geo-referenced digital photo collections. In: Proceedings of the 12th annual ACM international conference on Multimedia. pp. 196–203. ACM (2004)
12. Orii, Y., Nozawa, T., Kondo, T.: Web-based intelligent photo browser for flood of personal digital photographs. In: Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on. vol. 3, pp. 127–130. IEEE (2008)
13. Tan, J.L., Wu, J.H.: Classification algorithm of rule based on decision-tree. *Computer Engineering and Design* 31(5), 1017–1019 (2010)
14. TAO, C.m., WANG, H.l.: Data-driven classification algorithm based on organizational evolution and entropy. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)* 4, 017 (2009)
15. Tsay, K.E., Wu, Y.L., Hor, M.K., Tang, C.Y.: Personal photo organizer based on automated annotation framework. In: Intelligent Information Hiding and Multimedia Signal Processing, 2009. IIH-MSP'09. Fifth International Conference on. pp. 507–510. IEEE (2009)

Crowd Disaster Avoidance System (CDAS) by Deep Learning Using eXtended Center Symmetric Local Binary Pattern (XCS-LBP) Texture Features

C. Nagananthini and B. Yogameena

Abstract In order to avoid crowd disaster in public gatherings, this paper aims to develop an efficient algorithm that works well in both indoor and outdoor scenes to give early warning message automatically. It also deals with high dense crowd and sudden illumination changing environment. To address this problem, first an XCS-LBP (eXtended Center Symmetric Local Binary Pattern) features are extracted which works well under sudden illumination changes. Subsequently, these features are trained using deep Convolutional Neural Network (CNN) for crowd count. Finally, a warning message is displayed to the authority, if the people count exceeds a certain limit in order to avoid the crowd disaster in advance. Benchmark datasets such as PETS2009, UCSD and UFC_CC_50 have been used for experimentation. The performance measures such as MSE (Mean Square Error), MESA (Maximum Excess over Sub Arrays) and MAE (Mean Absolute Error) have been calculated and the proposed approach provides high accuracy.

Keywords Crowd disaster • Texture feature • Convolutional neural network • People counting

1 Introduction

In reality, public safety needed places such as malls, stadiums, festivals and in public gatherings, crowd control and crowd management becomes paramount. One of the basic descriptions of the crowd status is crowd density. Counting its flow is an important process in crowd behavior analysis. It can also be used to measure the comfort level of the crowd for detecting potential risk in order to prevent overcrowd

C. Nagananthini (✉) · B. Yogameena

Department of ECE, Thiagarajar College of Engineering, Madurai, India
e-mail: nagananthiniece2010@gmail.com

B. Yogameena

e-mail: b.yogameena@gmail.com

disasters. Crowd size is the important descriptor to detect threats like riots, fights, mass panic and stampedes. In case of traditional CCTV (Closed Circuit Television) cameras, the task of monitoring the crowd level is tedious. It is because of the requirement of large number of human resources to monitor surveillance cameras constantly over a long period of time. Recently, on 24 September 2015, about 2,070 pilgrims were died due to the crowd overflow during Hajj pilgrimage in Mina, Mecca and on 14 July 2015, at least 27 pilgrims died due to the stampede caused during Maha Pushkaralu festival on the banks of Godavari River, Andhra Pradesh, India. The above mentioned stampedes are due to lack of crowd control in advance. Hence, automated techniques should be involved for observing crowd to avoid crowd crush by estimating crowd count and crowd density. The proposed system focuses on warning the authority in advance to avoid such deadly accidents due to crowd crush. The major challenges faced by crowd detection algorithms are presence of too many people in the scene which makes the scene cluttered and occluded, complex backgrounds, shadows of people who are static and moving, sudden illumination changes and poor resolution of image.

In most of the public places, due to changing environment conditions, the surveillance camera fails in crowd density estimation and accurate counting. Most previous works in the field of crowd counting only count passing people robustly with slightly varying illumination. Hence, this paper aims to provide Crowd Disaster Avoidance System (CDAS) based on crowd density estimation. It involves extracting XCS-LBP (eXtended Center Symmetric Local Binary Pattern) texture features which works well in sudden illumination changing environment [1]. Subsequently, the Deep Convolutional Neural Network has been used for crowd density estimation and counting is done using dot annotations.

2 Related Work

The literature survey for crowd density estimation and people counting given by Abdulla et al. [2] shows that the indirect approach performs very well in detecting persons in highly dense crowd and also in presence of occlusions. The methods of indirect approach are pixel based, texture based and corner point based methods. Among these, texture based methods outperforms other in high dense crowd environments. Rahmalan et al. [3] proposed various texture based feature extraction methods such as GLCM (Gray Level Co-occurrence Matrix), MFD (Minkowski Fractal dimension) and TIOCM (Translation Invariant Orthonormal Chebyshev Moments) for crowd density estimation as well as people counting. In [4, 5], LBP (Local Binary Pattern, [6]) and LBPCM (LBP co-occurrence matrix) are also proposed for the process of crowd density estimation. Of these textures based methods, for estimating crowd density in indoor scenes, Marana et al. [7] proposed that MFD is computationally efficient compared to GLCM. MFD requires only one feature to be extracted which is a fractal dimension obtained after ‘n’ dilation

whereas GLCM requires 16 features such as Energy, Homogeneity, etc., The disadvantage of MFD is that the classifiers are not able to distinguish between high and very high crowd densities.

To estimate the crowd density in outdoor scenes, both GLCM and TIOCM perform well when compared to other methods [3]. But in case of small variations in illumination, TIOCM is better than GLCM but fails in case of clutter, shadow and noise. The extension of LBP (Advanced Local Binary Pattern (ALBP)) proposed by Ma et al. [8] has high distinctive power in handling noise and provides high accurate crowd degree in unconstrained environment but fails with increase in crowd size. In [9], it is given that CS-LBP (Center Symmetric LBP) descriptor is tolerant to small illumination changes and it is computationally simpler than the SIFT descriptor. The XCS-LBP descriptor is efficient in case of sudden illumination changes in background modeling and subtraction in videos [1]. As this paper aims to develop an efficient algorithm that works well in both indoor and outdoor scenes which is robust to instantaneously changed illumination, XCS-LBP texture features have been utilized for crowd density estimation. The main contributions towards this work are

- From the literature, XCS-LBP descriptor has been used only for background modeling and subtraction in video surveillance applications [1]. Therefore, the proposed CDAS make use of it for crowd density estimation and people count.
- The existing Deep learning based crowd count model [10] fails under sudden illumination. Hence, the proposed system adopts XCS-LBP texture feature based deep learning concept for crowd density estimation and people count which is robust under sudden illumination condition.
- The proposed CDAS displays a warning message to the authority when crowd count exceeds a limit. Thus, preventing deadly accidents in advance.

3 Proposed Method

At present, the accuracy of crowd density estimation and people count is reduced due to the sudden illumination changes in the real time environment. The proposed method involves extracting XCS-LBP texture feature which provides good result in case of sudden illumination changing environment [1]. The next step of the proposed methodology is to train the extracted feature samples with the Convolutional Neural Network which provides the learned feature vector. Subsequently, the test samples are tested to provide accuracy of correctly classified samples and error percentage of misclassified samples. Following that, the feature vector is mapped into crowd density map which provides the estimation of people in high dense crowd scenario.

Finally, the people count is done based on dot annotations which gives the count of people in the scene and a warning message box has been displayed to the

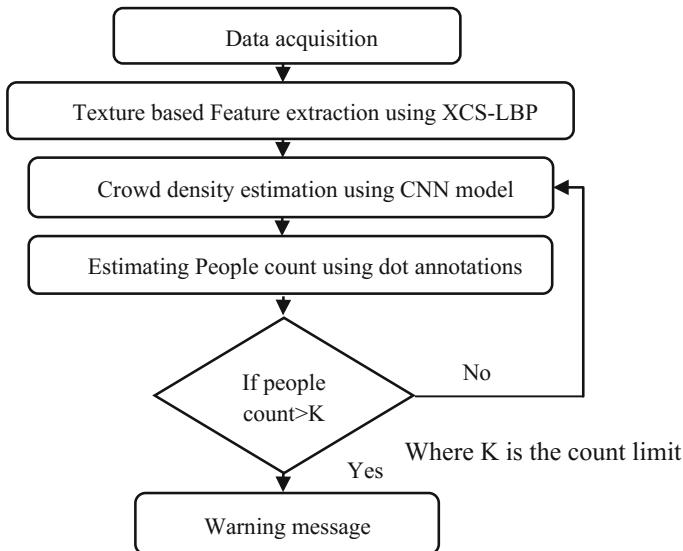


Fig. 1 Proposed methodology

authority in case of exceeding the count limit. The count limit threshold is set by the authority as per the statistics. The methodology of the proposed model is shown in Fig. 1.

3.1 Texture Feature Extraction

Texture is one of the important properties which require the analysis of image patches. The various texture feature extraction methods such as Local Binary Pattern (LBP), Center Symmetric LBP (CS-LBP) and Extended CS-LBP (XCS-LBP) are computed and the results of these methods are compared.

Extended CS-LBP (XCS-LBP) texture feature extraction. The first step of the process is to extract the XCS-LBP texture features. The extension of CS-LBP texture descriptor is given as XCS-LBP operator. The LBP texture descriptor compares the center pixel with its neighborhood pixels and the CS-LBP texture descriptor produces shorter histogram than LBP by comparing the neighborhood pixel values. In the XCS-LBP texture descriptor, both the center pixel and neighborhood pixel are also considered. This combination makes the resulting XCS-LBP descriptor less sensitive to sudden illumination changes [1]. The XCS-LBP operator is shown in Fig. 2.

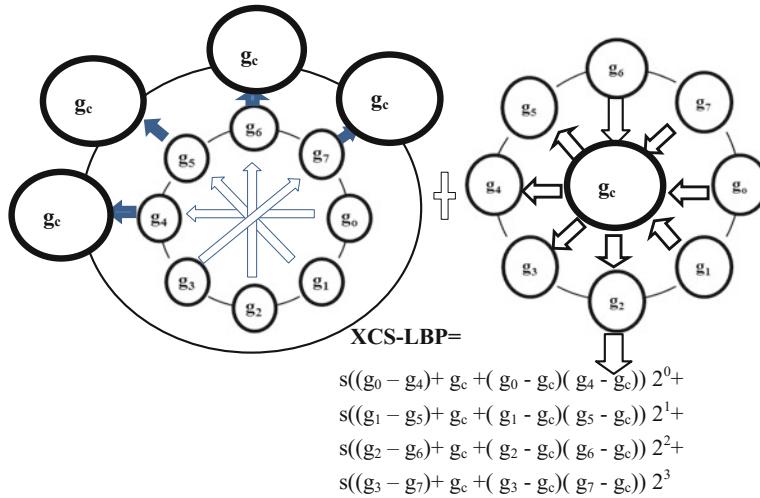


Fig. 2 XCS-LBP texture feature operator

The XCS-LBP (eXtended CS-LBP) is expressed as:

$$\text{XCS-LBP}_{P,R}(C) = \sum_{i=0}^{(P/2)-1} s(g_1(i,c) + g_2(i,c)) 2^i \quad (1)$$

where g_c is the center pixel and g_0, g_1, \dots, g_7 are the neighbourhood pixels. As per the result, the XCS-LBP provides histogram shorter than LBP, as short as CS-LBP. Also more image details are extracted using XCS-LBP than CS-LBP. It is less sensitive to noise and sudden illumination changing environment than both LBP and CS-LBP [1]. Hence, in the proposed methodology XCS-LBP texture features are used. After the extraction of XCS-LBP texture features, these feature samples are given as an input to the deep Convolutional Neural Network.

3.2 Deep Learning

Deep learning is about learning multiple levels of representation and abstraction in data. It replaces handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction. It performs hierarchy of representations with increasing level of abstraction. The traditional machine learning and deep learning concept is shown in Fig. 3.

Crowd density estimation using Convolutional Neural Network (CNN) model. After the extraction of XCS-LBP texture features, these feature samples are fed to the classifier. The concept of Deep learning provides accurate classification rather than the traditional classifiers because Deep learning uses trained features instead of hand-crafted features to obtain the learned feature vector. In a crowded

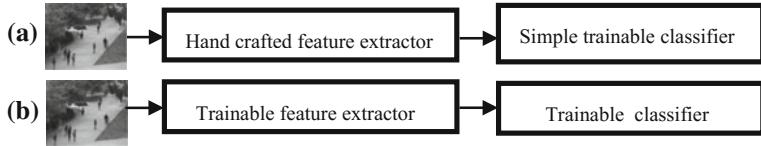


Fig. 3 **a** Traditional concept, **b** deep learning concept

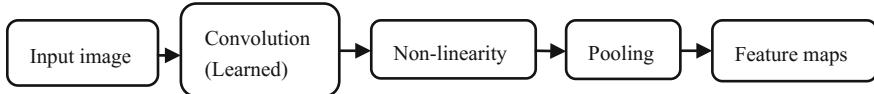


Fig. 4 Different layers in CNN

scenario, the prior knowledge and descriptions regarding the crowd with the human effort lacks the classifier's performance. CNN overcomes this problem and hence the proposed work makes use of it by creating a feed forward neural network with 20 neurons. In CNN, the features are learned with hierarchy of data abstraction with convolutional and pooling layer. After which the fully connected layer with the learned feature vector is obtained. The CNN consists of layers such as Convolution, non-linearity and pooling layers. The different layers employed in CNN are shown in Fig. 4.

The proposed CNN model learns crowd specific features which are more effective and robust than handcrafted features. No matter whether the crowd is moving or not, the crowd texture would be captured by the CNN model and can obtain a reasonable counting result [10]. The feature vector obtained from the CNN model is mapped into a crowd density map. The main objective of this crowd CNN model is to learn a mapping $F: X \rightarrow D$, where X is the set of XCS-LBP features extracted from training images and D is the crowd density map of the image. For example, consider frame I_k and a set of D_k texture features are extracted at their respective locations $\{(x_i, y_i), 1 \leq i \leq D_k\}$, the corresponding crowd density map of the image C_k is defined as follows:

$$C_k(x, y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{D_k} \exp - \left(\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2} \right) \quad (2)$$

where σ is the bandwidth of the 2D Gaussian kernel. The resulting crowd density map characterizes the spatial distributions of pedestrians in the scene. The architecture of proposed model is shown in Fig. 5.

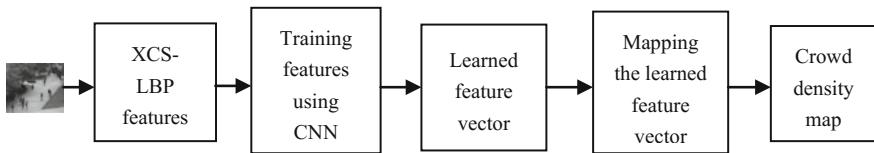


Fig. 5 Architecture of proposed system

3.3 Estimating People Count Using Dot Annotation

After training the XCS-LBP features in CNN model and extracting the learned feature vector, the next step is to count the number of people in the frame by dot annotating the learned feature vector. If the count exceeds a particular limit, the system has been developed in order to display a warning message to the authority. The threshold limit is set by the authority from the statistics of normal crowd in such places. Thus the proposed Crowd Disaster Avoidance System (CDAS) will efficiently prevent the deadly accidents in advance.

4 Results and Discussion

The proposed algorithm has been evaluated using Matlab 2013a. Experimentations are carried on PETS2009, UCSD and UCF_CC_50 bench mark datasets. Such bench mark datasets and their specifications are shown in the Table 1.

The sample frames of benchmark datasets are shown in Fig. 6. As the proposed approach concentrates on sudden illumination changes and high dense crowd, datasets which have frames with illumination changing environment and high dense crowd have also been collected and shown in Fig. 7.

The outputs for texture based feature extraction using LBP and CS-LBP with their histograms are shown in Figs. 8, 9.

The first step of the process is to extract the XCS-LBP texture features. The output of XCS-LBP feature extraction and its histogram are shown in Fig. 10.

As per the result, the XCS-LBP provides shorter histogram than LBP and CS-LBP. It also extracts more texture details of the image than CS-LBP. It is less

Table 1 Benchmark datasets with specifications

Datasets	Pets2009	UCSD	UCF_CC_50
Year	2009	2008	2012
Number of frames	S1 (4X1229)	2000	50
Resolution	768 × 576	238 × 158	158 × 238
Place	Outdoor	Outdoor	Outdoor
Density	0–42	11–46	94–4543



Fig. 6 The sample frames of benchmark datasets. **a** UCSD dataset, **b** PETS 2009 dataset, **c** UCF_CC_50 dataset

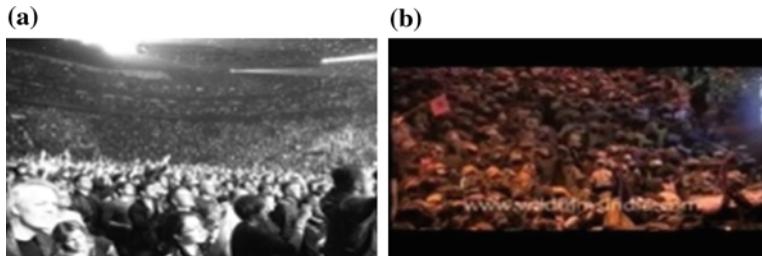


Fig. 7 Frames with illumination changing environment in high dense crowd. **a** UCF_CC_50 dataset, **b** temple dataset

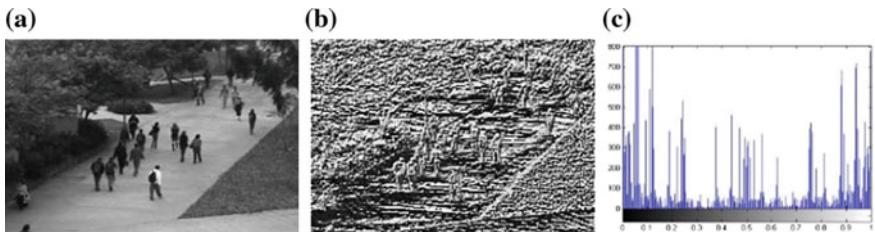


Fig. 8 The output of LBP texture feature extraction and its histogram. **a** UCSD dataset-original frame (frame number: 6), **b** UCSD dataset-LBP feature extracted frame (frame number: 6), **c** histogram

sensitive to noise and sudden illumination changing environment than both LBP and CS-LBP and provides better details for crowd density estimation. After the extraction of XCS-LBP texture features, they are trained in CNN by creating a feed forward neural network with 20 neurons and a learned feature vector is obtained. The vector of testing samples is given and is classified into low, medium and high dense crowd. Finally, the density estimation is labeled to any one of the category as mentioned above. The classification accuracy of 98.24 % and error rate 1.76 % is obtained. The number of persons in the scene is counted using dot annotations [11]

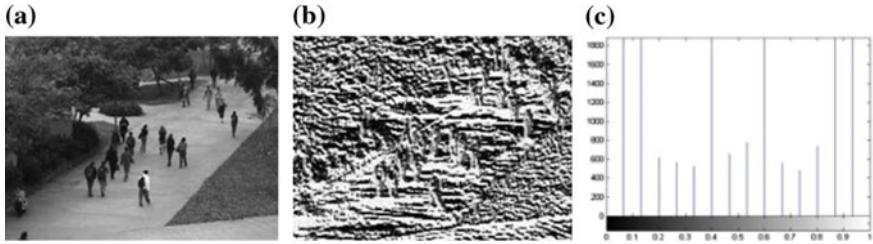


Fig. 9 Output of CS-LBP texture feature extraction and its histogram. **a** UCSD dataset original frame (frame number: 6), **b** UCSD dataset-CS-LBP feature extracted frame (frame number: 6), **c** histogram



Fig. 10 Output of XCS-LBP texture feature extraction and its histogram. **a** UCSD dataset-original frame (frame number: 24), **b** UCSD dataset-XCS-LBP feature extracted frame (frame number: 24), **c** histogram

and a warning message is displayed in case of exceeding the count limit of 25. The training samples and testing samples used for the classification are given in Fig. 11.

Following that, the features are trained using CNN and the learned feature vector is obtained. Subsequently, it is mapped into crowd density level according to the Eq. (2). The number of people in the scene is counted by dot annotations in which the feature vectors are labeled with red dots. The example of people count estimation and warning message display is shown in Fig. 12.

The features are labeled using red dots as shown in Fig. 14 and the number of people in the ROI region is 16 (UCSD), 11 (PETS2009) and 26 (Crowd). As the proposed approach aims at avoiding the crowd disaster in advance, the warning message is displayed in case of exceeding a certain limit. Here, the limit is set to 16. The accuracy of the proposed approach is quantified by the performance measures such as MESA [10], MAE and MSE.

Fig. 11 Training samples used in the proposed method



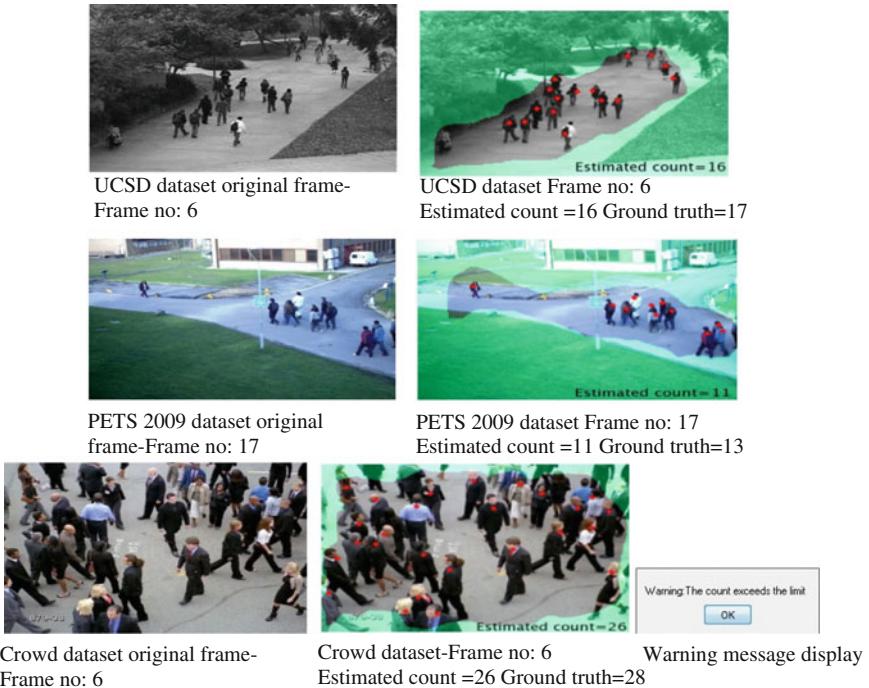


Fig. 12 People count estimation and warning message to alert crowd flaw

DMEA (Maximum Excess over Sub Arrays). Given an image I, the MESA distance D_{MESA} between two functions $F_1(p)$ and $F_2(p)$ on the pixel grid is given as

$$D_{\text{MESA}}(F_1, F_2) = \max_{B \in B} \left| \sum_{p \in B} F_1(p) - \sum_{p \in B} F_2(p) \right| \quad (3)$$

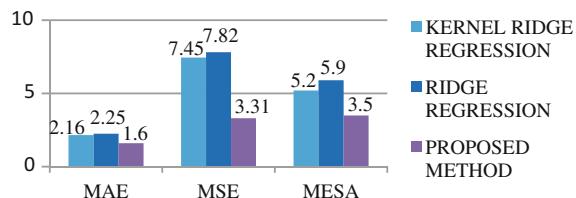
MAE (Mean Absolute Error). Given N as the total number of test frames, y_n is the actual count, and \hat{y}_n is the estimated count of nth frame, the MAE is computed as

$$\epsilon_{\text{abs}} = \frac{1}{N} \sum_{n=1}^N |y_n - \hat{y}_n| \quad (4)$$

MSE (Mean Square Error). Given N as the total number of test frames, y_n is the actual count, and \hat{y}_n is the estimated count of nth frame, the MSE is computed as

$$\epsilon_{\text{sqr}} = \frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2 \quad (5)$$

Fig. 13 Comparisons with various methods used for people counting



The comparison of performance measures of various methods used for people counting are represented as bar chart which is shown in Fig. 13. From which, it is inferred that the MSE, MAE and MESA for the proposed model is less compared to other regression methods are used for people count such as Kernel ridge regression by Liu et al. [12] and Ridge regression by Chen et al. [13]. It shows that the proposed model provides high accurate crowd count under sudden illumination changing environment and also in high dense crowd.

5 Conclusion

Deep CNN based Crowd Disaster Avoidance System (CDAS) using XCS-LBP texture features which have better capability for crowd density estimation and counting even under sudden illumination is proposed. Also, CDAS alerts the authority by displaying a warning message whenever the crowd count exceeds a certain limit, thus reducing the deadly accidents due to crowd crush. Though the proposed system alerts the crowd disaster in advance, the specific location where the crowd crush occurs is not found immediately. The future work aims at formulating the new problem of crowd overflow localization which is useful in crowded scenario to locate the direction where there is a crowd crush is possible. This helps the authority to take necessary action without any delay.

Acknowledgments This work has been supported under DST Fast Track Young Scientist Scheme for the project entitled, Intelligent Video Surveillance System for Crowd Density Estimation and Human Abnormal Analysis, with reference no. SR/FTP/ETA-49/2012.

References

1. Silva, C., Bouwmans, T., Frelicot, C.: An eXtended Center-Symmetric Local Binary Pattern for Background Modeling and Subtraction in Video. In: 10th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pp. 1–8. (2015)
2. Sami Abdulla, Mohsen Saleh, Shahrel Azmin Suandi, Haidi Ibrahim.: Recent survey on crowd density estimation and counting for visual surveillance. In: Engineering Applications of Artificial Intelligence, vol. 41, pp. 103–114. Pergamon Press, Inc. Tarrytown, NY, USA (2015)

3. Rahmalan, H., Nixon, M.S., Carter, J.N.: On crowd density estimation for surveillance. In: The Institution of Engineering and Technology Conference on Crime and Security, pp. 540–545. IET, London (2006)
4. Pratik P. Parate, Mandar Sohani.: Crowd Density Estimation Using Local Binary Pattern Based on an Image. In International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, Issue 7, pp. (2015)
5. Zhe Wang, Hong Liu, Yueliang Qian, Tao Xu.: Crowd Density Estimation Based On Local Binary Pattern Co-Occurrence Matrix. In: IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 372–377. IEEE, Melbourne, VIC (2012)
6. Ojala, T., Pietikainen, M., Maenpaa, T.: Multi resolution gray-scale and rotation invariant texture classification with local binary patterns. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, Issue 7, pp. 971–987. IEEE (2002)
7. Marana, A.N., da Fontoura Costa, L., Lotufo, R.A., Velastin, S.A.: Estimating crowd density with minkowski fractal dimension. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, pp. 3521–3524. IEEE, Phoenix, AZ (1999)
8. Wenhua Ma, Lei Huang, Changping Liu: Advanced Local Binary Pattern Descriptors for Crowd Estimation. In: Pacific-Asia Workshop on Computational Intelligence and Industrial Application, PACIIA'08, pp. 958–962. IEEE, Wuhan (2008)
9. Heikkila Marko, Pietikainen Matti, Schmid Cordelia: Description of Interest Regions with Center-Symmetric Local Binary Patterns. In: 5th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP, vol. 4338, pp. 58–69. (2006)
10. Cong Zhang, Hongsheng Li, Xiaogang Wang, Xiaokang Yang: Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 833–841. IEEE, Boston, MA (2015)
11. Carlos Arteta, Victor Lempitsky, Alison Noble, J., Andrew Zisserman: Interactive Object Counting. In: 13th European Conference Computer Vision ECCV, vol. 8691, pp. 504–518. Springer (2014)
12. An, S., Liu, W., Venkatesh, S.: Face Recognition Using Kernel Ridge Regression. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR'07, pp. 1–7. IEEE, Minneapolis, MN (2013)
13. Chen, K., Loy, C.C., Gong, S., Xiang, T.: Feature mining for localized crowd counting. In: Proceedings of British Machine Vision Conference, BMVC (2012)

A Novel Visualization and Tracking Framework for Analyzing the Inter/Intra Cloud Pattern Formation to Study Their Impact on Climate

Bibin Johnson, J. Sheeba Rani and Gorthi R.K.S.S. Manyam

Abstract Cloud Analysis plays an important role in understanding the climate changes which will be helpful in taking necessary mitigation policies. This work mainly aims to provide a novel framework for tracking as well as extracting characteristics of multiple cloud clusters by combining dense and sparse motion estimation techniques. The dense optical flow (Classic-Nonlocal) method estimates intra-cloud motion accurately from low contrast images in the presence of large motion. The sparse or feature (Region Overlap)-based estimation technique utilize the computed dense motion field to robustly estimate inter-cloud motion from consecutive images in the presence of cloud crossing, splitting, and merging scenario's. The proposed framework is also robust in handling illumination effects as well as poor signal quality due to atmospheric noises. A quantitative evaluation of the proposed framework on a synthetic fluid image sequence shows a better performance over other existing methods which reveals the applicability of classic-NL technique on cloud images. Experiments on half hourly infrared image sequence from Kalpana-1 Geostationary satellite have been performed and results show closest match to the actual track data.

Keywords Cloud motion vector • Clustering • Optical flow • Tracking

1 Introduction

Atmospheric clouds serve as one of the major factor in determining the weather condition of a given area. Accumulation of multiple cloud clusters over a region accounts for a severe weather condition, whereas their splitting and drifting results in

B. Johnson (✉) · J.S. Rani · G.R.K.S.S. Manyam
Indian Institute of Space Science and Technology, Trivandrum, India
e-mail: bibinjohnson.13@iist.ac.in

J.S. Rani
e-mail: sheeba@iist.ac.in

G.R.K.S.S. Manyam
e-mail: gorthisubrahmanyam@iist.ac.in

weakening of the weather condition. Geostationary satellite images allow meteorologists to analyze clouds propagation characteristics, evolution process, and life cycle which helps to improve the forecasting of extreme weather conditions. Cloud motion can be considered as special case of fluid motion with complex dynamic motions. Cloud systems contain smaller as well as larger cloud clusters which moves at different speed and direction. The task of cloud motion estimation is challenging due to the orthographic projection, complex dynamics of the imager's, nonlinear physical process underlying the cloud formation.

Until 1980s, the cloud systems were tracked manually using an Expert eyeball tracking technique [1] which was a time-consuming process. The results were highly accurate, but were sensitive to the user's expertise. From the pioneer work by Woodley in [2], many authors have worked on finding automatized method for tracking clouds. There are mainly two sets of such work in the literature; (a) Feature (sparse)-based motion estimation (b) Dense pixel-based motion estimation. Dense estimation methods computes the individual pixel motion between consecutive frames, but lacks in segmenting clouds from noncloudy regions. The feature-based methods segment the clouds initially based on thresholding or active contours or clustering techniques and then estimate the motion across the consecutive frames using cross correlation or pattern matching techniques.

The feature-based methods discussed in [3], and [4], extract geometric properties of the objects for tracking. A combination of search box and area overlap method for tracking is proposed in [5]. The paper [6] combines multilevel thresholding, block matching algorithm and vector median regularization for tracking. The major drawback of the feature-based methods is that they do not capture movement within the object. The problem is resolved using a local dense approach from computer vision literature to capture intra-cloud motions. In [7] author use affine motion models to estimate small local cloud motions. Based on a comparative study among the various existing dense methods in the literature, Classic-NL [8] method is found to be suitable for tracking clouds. The Classic-NL method is based on classic Horn and Schunck [9] formulation combined with most modern optimization and implementation strategies which helps in handling large motions, considering nonlinearity in the cloud motion, handling occlusions, etc.

In this work, we propose a novel framework that combines the dense and the sparse methods to estimate inter/intra-cloud motions robustly. In [10] authors computes dense flow estimate, but does not use this intra-cloud motion for computing inter-cloud motion. The proposed method is able to estimate cloud characteristics like the size of cloud shield, brightness temperature, evolution direction, trajectory, speed accurately. The main features of the proposed framework are: (i) Robustness against illumination variation using structure-texture decomposition, (ii) Handling large intra-cloud motions accurately using dense multiresolution approach, (iii) Segmentation based on dense flow in conjunction with adaptive intensity and size parameters, (iv) Global motion compensation using histogram-based approach, (v) Robust in handling large inter-cloud motion based on region overlap method in conjunction with Hungarian technique and dominant flow values.

The rest of the paper is organized as follows: Sect. 2 introduces the general framework and then discusses the proposed framework for robust cloud tracking. In Sect. 3, tracking results of the proposed methodology on real-world satellite images are discussed. Section 4 concludes the work by giving possible future directions.

2 A Novel Framework for Cloud Tracking

A general framework for cloud tracking involves a series of operations involving registration, preprocessing, segmentation, and tracking. Of these, the segmentation and tracking are the most complex and time-consuming ones. Most common techniques used for tracking are based on feature-based approaches. But, these fail in computing the intra-cloud motion. Recent literature on segmentation mostly use complex and intensive methods for segmentation which are time-consuming and inaccurate when the underlying dynamics are unknown [11].

In this work, we propose a method to integrate the dense flow estimation technique in to the sparse framework to improve the robustness of the estimated track. The dense flow values are added as extra parameters in segmentation based on thresholding which improves the accuracy. The main steps in the proposed methodology are described below.

2.1 Preprocessing

The low-resolution infrared satellite images used for tracking suffer from noise, illumination effects (due to the variation in sun's reflectance from day to night) and are of low contrast. The main task in preprocessing stage is the impulse noise removal and contrast enhancement. The proposed methodology uses ROF structure-texture decomposition [12] to decompose the image into structure and texture part, which is unaffected by illumination changes. The structural part $I_S(x)$ of an intensity image is given as,

$$\min_{I_s} \int \{ |\nabla I_s| + \frac{1}{2\theta} (I_s - I)^2 \} dx \quad (1)$$

Solution is given by,

$$I_s = I + \theta \mathbf{div}(p) \quad (2)$$

The textural part $I_T(x)$ is computed as the difference between original image and its structural part, $I_T(x) = I(x) - I_s(x)$. The artifacts due to shadow and shading reflections show up in the original image and structural part but not in the textural part which helps in removing the illumination effects.

2.2 Motion Estimation: Dense Flow Approach

The proposed methodology employs a dense method to estimate the intra-cloud motion accurately which helps in studying the life cycle characteristics of the cloud clusters. Optical flow methods are accurate computer vision techniques which gives dense motion accurately. Using first-order Taylor series approximation the intensity of moved pixel is given by,

$$I(x, y, t) = I_o(x, y, t) + u \frac{dI}{dx} + v \frac{dI}{dy} + \frac{dI}{dt} \quad (3)$$

According to the brightness constancy assumption, $I(x, y, t) = I(x + u, y + v, t + 1)$. This gives the optical flow constraint equation,

$$u \frac{dI}{dx} + v \frac{dI}{dy} + \frac{dI}{dt} = 0 \quad (4)$$

Classic methods like Lucas Kanade [13], Horn and Schunck (HS) techniques fail to compute large intra-cloud motions. To overcome this issue, various multiscale approaches based on image pyramids with iterative warping schemes are proposed in the literature. A performance comparison of different flow estimation techniques on synthetic fluid images is described in Table 1. This clearly demonstrates the lower error performance of Classic-NL method [8], as compared to other existing techniques.

It is based on classical flow formulations combined with modern optimization and implementation techniques. The different stages involved are: (a) Graduated non-convexity scheme, (b) Pyramid-based multiresolution approach, (c) Warping, (d) Median filtering, (e) Occlusion detection. The proposed methodology employs this method to compute dense intra-cloud motion accurately. The classical HS cost function containing data term (ρ_D) and spatial smoothness term (ρ_S) is modified to add weighted nonlocal median term which integrate information from large neighborhood considering the discontinuities at motion boundaries.

$$\begin{aligned} E(u, v) = & \{\rho_D(I_1(i, j) - I_2(i + u_{i,j}, j + v_{i,j})) + \\ & \lambda[\rho_S(u_{i,j}u_{i+1,j}) + \rho_S(u_{i,j}u_{i,j+1}) + \\ & \rho_S((v_{i,j}v_{i+1,j}) + \rho_S(v_{i,j}v_{i,j+1}))\} + \\ & \lambda_N \sum_{i,j} \sum_{(i',j')} (|u_{i,j}u_{i',j'}| + |v_{i,j}v_{i',j'}|) \end{aligned} \quad (5)$$

The new median computation is given by,

$$\begin{aligned} \hat{u}_{i,j}^{(k+1)} &= median(Neighbors^{(k)} \cup Data) \\ Neighbors^{(k)} &= \hat{u}_{i',j'}^{(k)} \end{aligned} \quad (6)$$

Table 1 Compare the error performance of different standard OF methods on synthetic fluid image sequence

Dataset/Metric	AAE	STDAAE	EPE	Entropy	RMSE	Density (%)
1. Poiseuille						
Classic NL	0.7135	0.9858	0.0232	1.6024	3.5999	100
Hierarchical LK[1]	1.8842	5.4203	0.0889	0.2493	6.2358	95.3674
LK	4.5851	6.1547	0.2244	0.2632	25.5794	96.8994
HS-SUN[3]	1.2355	1.1547	0.0398	2.8877	3.5447	100
BA-SUN[3]	0.8691	0.7866	0.0283	2.5251	3.2969	100
Lucas-Barron[2]	16.7767	12.1275	0.6382	1.0625	118.9861	79.3213
2. Sink flow						
Classic NL	1.0458	1.7336	0.0232	2.8144	2.9191	100.0000
Hierarchical LK[1]	2.2491	4.0377	0.0473	2.7186	2.0508	95.3674
LK	1.6584	2.6551	0.0339	2.7515	2.8756	96.8994
HS-SUN[3]	2.2547	2.0951	0.0438	2.7709	2.2677	100.0000
BA-SUN[3]	1.9641	1.4097	0.0388	2.7953	2.6624	100.0000
Lucas-Barron[2]	1.7444	4.9271	0.0395	2.4606	5.1402	79.3213
3. Vortex flow						
Classic NL	15.8194	11.3702	0.2871	2.7494	2.8040	100.0000
Hierarchical LK[1]	14.5690	12.0090	0.2685	2.6962	2.1145	95.3674
LK	14.4555	11.4884	0.2639	2.7177	2.6284	96.8994
HS-SUN[3]	17.1875	11.8910	0.3130	2.7103	2.2151	100.0000
BA-SUN[3]	16.9393	11.7379	0.3081	2.7141	2.3551	100.0000
Lucas-Barron[2]	1.7130	4.6170	0.0390	2.2614	31.7256	79.3213

$$Data = \{u_{i,j}, u_{i,j} \pm \frac{\lambda 3}{\lambda 2}, u_{i,j} \pm \frac{2\lambda 3}{\lambda 2} \dots u_{i,j} \pm \frac{|N_{i,j}| \lambda 3}{2\lambda 2}\} \quad (7)$$

The objective function is modified by introducing a weight into the nonlocal median term given by,

$$\sum_{i,j} \sum_{(i',j') \in N_{i,j}} w_{i,j,i',j'} (|\hat{u}_{i,j} \hat{u}_{i',j'}| + |\hat{v}_{i,j} \hat{v}_{i',j'}|) \quad (8)$$

2.3 Histogram-Based Global Motion Compensation

The image registration of geostationary satellite images play an important role in computation of accurate motion field. Image registration of Kalpana-1 satellite is done using three consecutive images by matching valid tracers in forward and backward direction. The inaccuracy of registration in the images will generate the errors in wind speed and direction. This method helps in compensating the registration effects in the satellite images caused by the global motion, due to the deviation in orbital parameters, spacecraft attitude, thermal distortions, and sensor biases. We propose a histogram-based method to compensate the global motion. The proposed method is based on the assumption that the satellite images contains fewer cloud pixels as compared to the background objects. Separate histograms are computed for optical flow vectors: u & v . Since the pixels belonging to static areas does not have any motion, the u & v histograms are supposed to have a zero mean. This is not true for the images with global motion, which means that there is a nonzero value (offset) of velocity present in the static background pixels. This offset can be evaluated from the u & v values and is used to compensate the entire flow vectors. This compensation results in much clearer flow vector computation.

2.4 3D Spatio Temporal Segmentation

Segmentation is one of the most challenging tasks in cloud tracking. Generally used methods in segmenting clouds are level-sets, active contours which are complex and computationally intensive. Even though these methods are accurate in most situations, still it lack in accuracy when the global dynamic description of the underlying object is not clearly known [11]. Motion is a very useful clue for image segmentation. The main idea of this work is to develop a spatiotemporal image segmentation technique for image sequences. In this approach segmentation utilizes information from two image frames. Different features like velocity, pixel separation, multispectral intensity (bright regions are more important than the background which is mostly dark) are extracted and clustered using an ISO Data clustering algorithm. The clustering can be performed on the optical flow vectors obtained by L-K method.

$$A = \begin{pmatrix} 1 & 1 & mag_{11} & \phi_{11} & IR_{11} & WV_{11} & VIS_{11} \\ 1 & 2 & mag_{12} & \phi_{12} & IR_{12} & WV_{12} & VIS_{12} \\ & & . & & . & & \\ M & N & mag_{MN} & \phi_{MN} & IR_{MN} & WV_{MN} & VIS_{MN} \end{pmatrix} \quad (9)$$

A Seven-column matrix A is formed with the first two columns contains the pixels coordinates, the third and fourth; the magnitude of pixel velocities and their direction, the remaining five, six, seven columns contain multispectral image intensities

(IR/WV/VIS). This input matrix A is appropriately weighted and normalized before feeding into the ISODATA clustering algorithm. The criterion for new clusters formation was the degree of similarity between objects computed by the Euclidean distances between their centroids. The weighting factor of the parameters and the ISODATA initialization values (splitting and merging thresholds) are chosen experimentally.

2.5 *Morphological Operation and Dominant Flow Computation*

The Morphological Operation [14] helps to strengthen the irregular and thin shapes in the segmented cloud image. This gives better visual representation of cloud patterns. This is performed in two steps, first step uses disk-shaped structuring element with 12 pixels in the diameter, to remove the noisy regions present in the segmented image. The next step fills any holes existing in the segmented image using disk-shaped structuring element with 10 pixels in the diameter. The Dominant flow, as the name indicates gives the prominent direction and velocity of each cloud cluster. Instead of directly taking the median of the all flow vectors in each cluster, we propose a method to choose only flow vectors which shows strong motion. The median is only computed among a selected group of vectors and the value is assigned to the cloud centroids. The computed dominant flow used in conjunction with the flow estimated by feature based methods provide a robust track.

2.6 *Tracking: Feature Based Approach*

The final stage in the proposed framework is the sparse feature tracking stage for computing the inter-cloud motion. Instead of using a single algorithm, a combination of different methods is proposed in this work. Initial tracking of clouds is performed using a forward and backward region-based overlap method [15]. This technique is based on a simple assumption that there are a set of common pixels in consecutive images with a particular size and temperature. The matching of the consecutive images is performed in forward and backward direction to identify the existence of splitting and merging conditions. Each of these overlapping clusters are given a unique number which is useful in generating statistical information about the life cycle characteristics (generation, dissipation, continuity, splitting, merging) of the cloud system. This method holds valid since the temporal resolution of the satellite images are high, there exists no drastic changes in cloud shape. But, this is not always the case since there exists clouds which moves faster than temporal sampling rate of half hourly interval. Overlap region method fails to classify such type of clouds. So in the second step, the unclassified clouds are matched by the James

Munkres variant of Hungarian algorithm [16] to detect the missing tracks as well as generation of new tracks. In last step, the corrected track vectors are then compared with dominant flow velocities to finalize the flow direction.

$$d(c_k, c_{k+1}) = w_0 * O(c_k, c_{k+1}) + w_1 * \text{delta1}(c_k, c_{k+1}) + w_2 * \text{delta2}(c_k, c_{k+1}) \quad (10)$$

The Overlap Metric

$$O(c_k, c_{k+1}) = 1 - \left\{ \frac{n_o(c_i^k, c_j^{k+1})}{2} * \left[\frac{1}{s_i^k} + \frac{1}{s_j^{k+1}} \right] \right\} \quad (11)$$

The Centroid Metric

$$\delta_c(c_i^k, c_j^{k+1}) = \frac{\sqrt{(X_i^k - X_j^{k+1})^2 + (Y_i^k - Y_j^{k+1})^2}}{\sqrt{I_{height}^2 + I_{width}^2}} \quad (12)$$

The Size Metric

$$\delta_s(c_i^k, c_j^{k+1}) = \frac{|s_i^k - s_j^{k+1}|}{\max(s_i^k, s_j^{k+1})} \quad (13)$$

2.7 Confidence Measure

The tracking system generates a confidence index for each tracked cloud segments which indicate the how good is the estimated track during its entire life cycle. There are mainly three components that can affect the computation of the confidence index. They are minimum lifetime of the cloud segments (depends on the type of clouds, some are persisting strong clouds which have greater chances in getting accurate tracking results), number of surrounding cloud segments, and distance of the segments from centroid to the image border (border cloud clusters have chances to get out of the frame). Lower weight is given for clusters which are very close to other clusters or grouped together into colonies.

3 Results and Discussion

The present experiment was conducted on 8 km resolution IR (10.5–12.5 m) images obtained from Kalpana-1 of the Indian sector. Indian meteorological satellite Kalpana-1 was launched in the year 2002 and features a Very High Resolution

scanning Radiometer (VHRR) which records images in three different bands visible (VIS), thermal infrared (TIR) and water vapor (WV). The input images of Kalpana-1 have dimensions of 808×809 with 10-bit pixel intensity values. The temporal frequency of all these geostationary satellite image sequence is 30 min. The input images are preprocessed for illumination correction and fed to the dense flow estimation unit, which estimate the intra-cloud motion accurately as denoted by the dominant flow vectors shown at the centroid of each cloud segments (Fig. 1).

The cloud clusters are displayed with closed boundary lines and represented using a unique cloud number computed during the tracking stage. The Fig. 2 represents the track of the cloud system on top of Indian region during the month June 2014. Due to the absence of ground truth data, the validation of computed motion field is done visually by a field expert. The estimated track shows close match to the true path in visual inspection. Similarly other cloud parameters like size, speed, life cycle, etc., are computed for making several scientific conclusions about the cloud propagation characteristics.

4 Conclusion

The proposed tracking framework produce competitive results on Kalpana-1 satellite images. This methodology is novel in its kind in combining dense and feature-based methods to generate robust track of each cloud segments. The tracking framework is robust in terms of: (a) illumination independence, (b) dense intra-cloud motion, (c) robust segmentation, (d) dominant flow computation, (e) robust track assignment based on region overlap and Hungarian technique. The proposed framework gener-

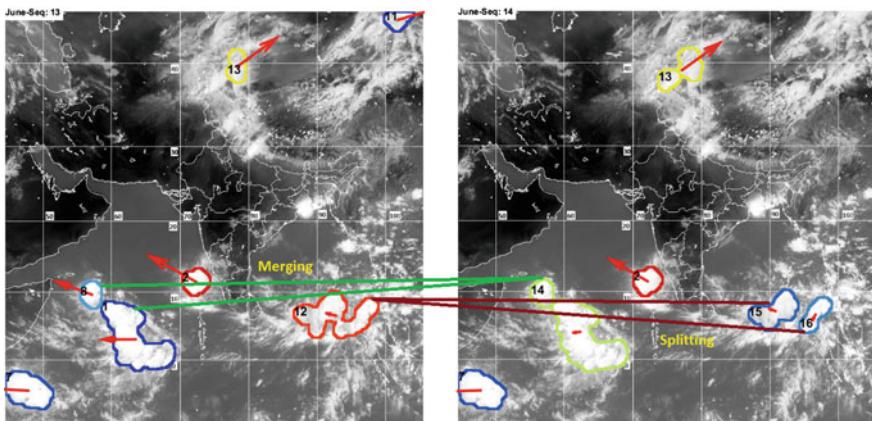


Fig. 1 Two consecutive segmented Images (kalpana-1 June-2014) from showing splitting and merging conditions. The cloud segment 12 is being divided into 15 and 16, and a new cloud segment 14 is formed by the merging of cloud segment 1 and 8

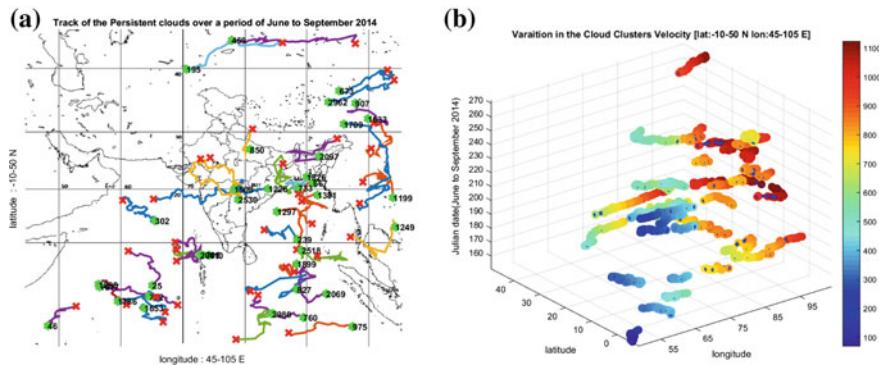


Fig. 2 Trajectories of cloud segments over a period of one month. The cloud clusters which persists for minimum of 50 frames are only displayed. *Green blob* indicates the birth, the *red cross* represents the death and the *small black dot* shows each frame instance

ates information about the number of cloud clusters, the variation in the cloud size, speed of each cloud cluster, their life cycle that is necessary to study the climatic changes over a region over a period of time. The result obtained in this study encourages to further explore the different possibilities in each of the proposed stages. Further extensions to the work include (a) use of clustering for segmentation, (b) use of multispectral images (visible, water vapor) to improve the accuracy, and (c) use of multiple image sequences. Future work also focuses on the design of hardware accelerator, which can improve the tracking performance for near real-time analysis.

References

1. Yves Arnaud, Michel Desbois, and Joel Maizi. Automatic tracking and characterization of african convective systems on meteosat pictures. *Journal of Applied Meteorology*, 31(5):443–453, 1992.
2. William L Woodley, Cecilia G Griffith, Joseph S Griffin, and Scott C Stromatt. The inference of gate convective rainfall from sms-1 imagery. *Journal of Applied Meteorology*, 19(4):388–408, 1980.
3. Daniel Laney, P-T Bremer, Ajith Mascarenhas, Paul Miller, and Valerio Pascucci. Understanding the structure of the turbulent mixing layer in hydrodynamic instabilities. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):1053–1060, 2006.
4. P-T Bremer, Gunther Weber, Julien Tierny, Valerio Pascucci, Marc Day, and John Bell. Interactive exploration and analysis of large-scale simulations using topology-based data segmentation. *Visualization and Computer Graphics, IEEE Transactions on*, 17(9):1307–1324, 2011.
5. Arvind V Gambheer and GS Bhat. Life cycle characteristics of deep cloud systems over the indian region using insat-1b pixel data. *Monthly weather review*, 128(12):4071–4083, 2000.
6. Remus Brad and Loan Alfred LETIA. Extracting cloud motion from satellite image sequences. In *Control, Automation, Robotics and Vision, 2002. ICARCV 2002. 7th International Conference on*, volume 3, pages 1303–1307. IEEE, 2002.
7. Chandra Kambhamettu, Kannappan Palaniappan, and A Frederick Hasler. Automated cloud-drift winds from goes images. In *SPIE's 1996 International Symposium on Optical Science*,

- Engineering, and Instrumentation*, pages 122–133. International Society for Optics and Photonics, 1996.
- 8. Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.
 - 9. Berthold K Horn and Brian G Schunck. Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics, 1981.
 - 10. Harish Doraiswamy, Vivek Natarajan, and Ravi S Nanjundiah. An exploration framework to identify and track movement of cloud systems. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2896–2905, 2013.
 - 11. Claire Thomas, Thomas Corpetti, and Etienne Memin. Data assimilation for convective-cell tracking on meteorological image sequences. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(8):3162–3177, 2010.
 - 12. Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.
 - 13. Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004.
 - 14. Lei Liu, Xuejin Sun, Feng Chen, Shijun Zhao, and Taichang Gao. Cloud classification based on structure features of infrared images. *Journal of Atmospheric and Oceanic Technology*, 28(3):410–417, 2011.
 - 15. Daniel Alejandro Vila, Luiz Augusto Toledo Machado, Henri Laurent, and Ines Velasco. Forecast and tracking the evolution of cloud clusters (fortracc) using satellite infrared imagery: Methodology and validation. *Weather and Forecasting*, 23(2):233–245, 2008.
 - 16. James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Cancelable Biometric Template Security Using Segment-Based Visual Cryptography

P. Punithavathi and S. Geetha

Abstract The cancelable biometric system is susceptible to a variety of attacks aimed at deteriorating the integrity of the authentication procedure. These attacks are intended to either ruin the security afforded by the system or deter the normal functioning of the authentication system. This paper describes various threats that can be encountered by a cancelable biometric system. It specifically focuses on preventing the attacks designed to extract information about the transformed biometric data of an individual, from the template database. Furthermore, we provide experimental results pertaining to a system combining the cancelable biometrics with segment-based visual cryptography, which converts traditional biocode into novel structures called shares.

Keywords Bioencoding • Biocode • Template security • Segment-based visual cryptography

1 Introduction

Biometrics is a powerful tool against repudiation. So it has been widely deployed in various security systems. At the same time, the biometric characteristics are largely immutable, resulting in a permanent biometric compromise when a template is stolen. The concept of cancelable biometrics was introduced to make a biometric template to be canceled and be revoked like a password (in case of biometric compromise), as well as being unique to every application. Cancelable biometrics are repeatable distortions of biometric signals based on transforms at signal level or feature level.

P. Punithavathi (✉) · S. Geetha

School of Computing Science & Engineering,
VIT University Chennai Campus, Chennai, Tamil Nadu, India
e-mail: p.punithavathi2015@vit.ac.in

S. Geetha
e-mail: geetha.s@vit.ac.in

The transformed templates generated using cancelable biometric algorithm must meet the four major requirements.

- Irreversibility: The transformation should be only in one way such that it should be computationally hard to reconstruct the original template from the protected template, despite the fact that it should be easy to generate the transformed biometric template.
- Unlinkability: Multiple versions of transformed biometric templates can be generated corresponding to the same biometric data (renewability), at the same time these templates should not allow cross-matching (diversity). Hence, the inversion of such transformed biometric templates is not feasible for potential imposters.
- Reusability/Revocability: Straightforward revocation of a template like a password and reissue a new template in the incident of biometric template compromise.
- Performance: The recognition performance should not be deteriorated by the transformation function.

1.1 Cancelable Biometrics System

A typical cancelable biometric authentication system is similar to the traditional biometric authentication system except for the transformation module. The modules included in the cancelable biometric authentication system are as follows:

- A sensor module.
- A feature transformation module.
- A matching module.
- A decision module.

Thus, a cancelable biometric system can be viewed as a pattern identification method which is able to classify a biometric signal into one of the several identities (i.e., through identification) or into one of two classes—authentic and impostor users (i.e., through verification).

1.2 Attacks on Cancelable Biometrics System

Different levels of attacks were identified by Patel et al. in [1], which can be launched against biometric system. The cancelable biometric system is also susceptible to such attacks which are shown in Fig. 1.

- a fake biometric trait such as an artificial finger may possibly be presented at the sensor,
- illegally captured data may be resubmitted to the system,
- the feature transforming module may be replaced by illegal Trojan horse program to produce pre-determined feature sets,

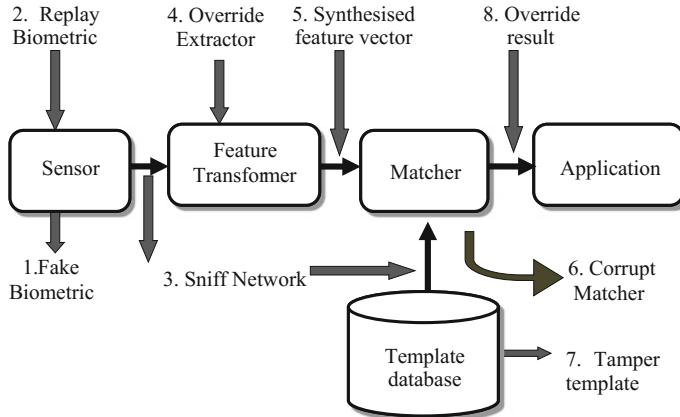


Fig. 1 Possible attacks against cancelable biometrics

- genuine feature sets may be replaced with fake feature sets,
- the matcher may be replaced by illegal programs like Trojan horse which will always output high scores thereby challenging system security,
- the stored templates may be modified or removed, or new templates may be introduced into the database,
- the data within the communication channel between various modules of the system may be modified, and
- the final decision of the matcher may be overwritten.

It is clear from Fig. 1 that the attacks of Type 1 are aimed at the sensor, and can be carried out using artifact. Attacks of Types 4, 6, and 7 may be performed as Trojan horse attacks, bypassing the feature extractor, the matcher, and the template database, respectively. Attacks of Types 2, 3, 5, and 8 may focus on communication channels either to intercept information or insert new information into the channel.

Attacks of Type 1 can be prevented by liveness detection at the biometric sensor. The attacks of Types 2, 5, and 8 can be prevented using encrypted communication channels. But what happens if a transformed template is stolen from the database (attack of type 7) or during transmission to matcher from the database (Type 3 attack)?

1.3 Related Work

There are several methods to generate non-invertible cancelable biometric template. The transformations are applied to the biometric input either at signal level or at function level. The first attempt towards the direction of transformed biometrics was recorded by Soutar et al. in 1998 [2], but the actual idea of cancelable biometrics was furnished by Bolle et al. in 2002 [3]. The fingerprint data has been transformed

by a sequence of three functions based on cartesian transformations, polar transformations, and surface folding of the minutiae positions in [4]. A cancelable biometric template for face has been generated in [5] using an appearance-based approach. In another technique [6], adaptive Bloom filter-based transforms have been applied to combine binary iris biometric templates at feature level, where iris codes are obtained from both eyes of a single subject.

These transformed templates can be canceled and revoked easily like password, in the case of biometric compromise. Also a single user can have several transformed templates corresponding to several applications. The transformed templates are stored into the database and not the biometric template itself. The matching is done in transformed domain. This makes using cancelable biometrics more trustworthy than using traditional biometric authentication system.

Despite of several advantages of cancelable biometrics, the protection of these transformed templates should be taken into account to make the cancelable biometric authentication system more successful. The transformed template itself is enough for an imposter to deceive the entire authentication system.

Several attempts were made to secure the stored biometric templates. A finger-print image watermarking method was proposed in [7] that can embed facial information into host fingerprint images. A similar method was proposed in [8], which combined discrete wavelet transform and least significant bit-based biometric watermarking algorithm that securely embeds the face template of the user into the fingerprint image of the same user. But the disadvantages of both these methods are that an external secret key had to be used during encryption and decryption.

Another method for protecting fingerprint biometrics was proposed in [9], which combines two fingerprints from two different fingers to generate a new template. For authentication, two query fingerprints are required and a two-stage matching process is proposed for matching two query fingerprints against a combined template. But this method cannot be adapted in case of cancelable biometrics where the transformed templates are mostly vectors.

A method was proposed in which face image and fingerprints were secured using visual cryptography [10]. The method splits the biometric input into two shares and saves them into two different databases, hence resulting in the setup of an additional database. Moreover, it is applied to biometric templates in image form. In case of cancelable biometrics the transformed templates are mostly vectors. There is a requirement for a system which can take the transformed template in the vector form and secure it without any external key.

2 Proposed System for Transformed Template Security

A new system has been proposed which uses segment-based visual cryptography to secure biocode—a vector obtained as result of transforming the biometric input using bioencoding method. Two different shares are generated from the input biocode, out of which, one is handed over to the user and the other one is stored into

the database. During authentication the biocode is generated just by applying logical ‘OR’ operation on the shares, and compared with the currently generated biocode from the biometric input. The proposed system has the following advantages:

- No requirement of external key, during share generation.
- No complex recovery of the hidden biocode. The hidden biocode can be generated just by applying logical ‘OR’ operation on the shares.
- No problem if the user has lost the share or if the database has been cracked, because an imposter can get no information from single share.
- No additional memory space is required to store the share which is going to have more or less the same size as the biocode.

2.1 Enrolment Module

Step 1: The iris code is generated from the biometric input which is provided by user.

Step 2: The consistent bits are extracted from the iris code.

Step 3: The biocode is generated from the consistent bits.

Step 4: The biocode is converted into two shares (share 1 and share 2) using segment-based visual cryptography. One of the shares (share 2) is stored into database, while the other share (share 1) is given to the user in form of smart card, tag, etc. The process is illustrated in Fig. 2.

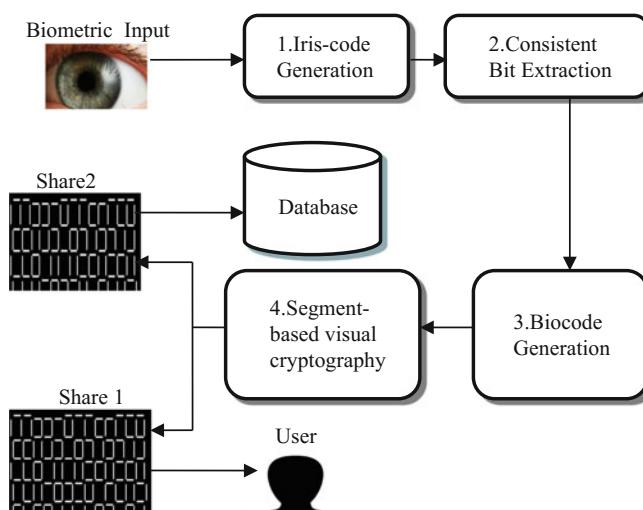


Fig. 2 Enrolment module of the proposed system

2.2 Authentication Module

- Step 1: The iris code is generated from the biometric input which is provided by user.
- Step 2: The consistent bits are extracted from the iris code.
- Step 3: The biocode is generated from the consistent bits.
- Step 4: The segment-based visual cryptography is used to recover hidden biocode by superimposing share 1 (provided by user) on share 2 (retrieved from data base).
- Step 5: The recovered biocode is compared with currently generated biocode the matching unit and the decision is passed to the authenticating unit.

The authentication procedure is illustrated in Fig. 3.

2.3 Bioencoding

Bioencoding is a trustworthy, tokenless cancelable biometrics scheme for protecting biometrics. A unique non-invertible compact bit string, referred to as biocode, is randomly derived from iris code. Rather than the iris code, the derived biocode can be used to verify the user identity, without degrading the recognition accuracy. Additionally, bioencoding satisfies all the requirements of the cancelable biometrics construct. Iris has been selected due to high accuracy and easy availability of

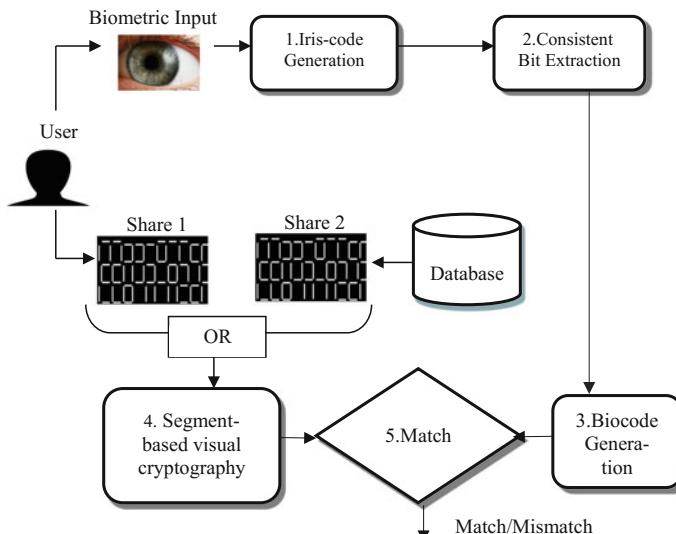


Fig. 3 Authentication module of the proposed system

dataset. The base procedure of bioencoding for iris codes is composed of three major stages:

- Iris code generation

This includes segmentation, normalization, and feature extraction. Segmentation is performed to extract the iris from the eye image. By employing circular Hough transform, the iris boundary is searched. The eyelids are detected using linear Hough transform, and the eyelash is separated using a threshold technique. Daugman's rubber sheet model [11] is used to remap each pixel within the iris region to a pair of polar coordinates, and normalize iris region. Feature extraction is done by convolving the normalized iris pattern into one-dimensional Log Gabor wavelets. The resulting phase information for both the real and the imaginary response is quantized, generating a bitwise template of iris code which is a matrix of size 20×480 .

- Consistent bit extraction

The Hollingsworth method [12] is used to extract the consistent bit out of iris code. The most consistent bits are extracted by identifying fragile bits and masking them out. Fragile bits are bits that have a substantial probability of being '0' in some images of an iris and '1' in other images of the same iris. Their positions are collected in a position vector.

- Biocode generation

The biocode is generated using the technique proposed in [13] with the consistent bits and their position vectors.

It has been selected due to its gaining popularity and the fact that the output is in the form of vector. This vector can be easily secured using segment-based visual cryptography.

2.4 Segment-Based Visual Cryptography

This type of visual cryptography is segment-based. The advantage of the segment-based visual cryptography [14] is that it may be uncomplicated to adjust the secret images and that the symbols are clearly recognized by the human visual system. It is used to hide messages comprising numbers which can be represented by seven-segment display. The seven-segment display is composed of seven bars (three horizontal and four vertical) which are arranged like a digit '8'.

The principle of pixel-based visual cryptography [15] is applied to the segment-based visual cryptography. Assume that every segment 'Sn' ('—' or 'l') in the seven-segment display model comprised two parallel segments 'Sn1' and 'Sn2' ('—' or 'l'), which are very close to each other but do not intersect each other. The digit '8' has been illustrated using this logic, in Fig. 4a.

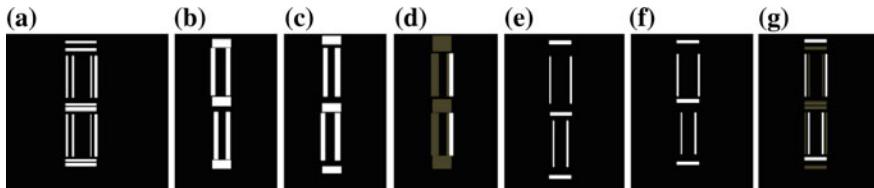


Fig. 4 Principle applied to the seven-segment display

Generation of First Share

Let each segment in a seven-segment display model be represented as ‘ S_n ’ (where $n = 0\text{--}6$). Assume that a certain digit has to be revealed in white color on a black background. Consider a subset (C) of segments, among the segments $S_0\text{--}S_6$ that have to be highlighted in a seven-segment display such that the required digit can be shown.

Similar to the pixel-based visual cryptography, the first share is generated randomly. This means that one of the parallel segments, ‘ S_{n1} ’ or ‘ S_{n2} ,’ is selected randomly for every segment S_n irrespective of whether the segment belongs to ‘ C ’ or not. The selected segment (say ‘ S_{n1} ’) is kept white (or transparent), while the parallel segment (say ‘ S_{n2} ’) is left black (equal to the color of the background of the share). Such a random selection is illustrated with digits ‘1’ and ‘0’ in Fig. 4b and e, respectively.

Generation of Second Share

- If a segment ‘ S_{n1} ’ or ‘ S_{n2} ’ of $S_n \in C$ is kept white in first share, then in the second share the same selection of ‘ S_{n1} ’ or ‘ S_{n2} ’ is made and it is kept white. At the same time, the other parallel segment is turned black. (It is due to this selection that the white segment is shown off while overlaying the shares.)
- On the other hand, if S_n does not belong to the subset C , then in the second share, the segment (‘ S_{n1} ’ or ‘ S_{n2} ’) chosen on the random share (share 1) is turned black. This has the effect that while overlaying the shares, this segment is not highlighted.

In total, exactly the segment belonging to the subset C is showed off, when the shares are overlaid. Therefore, after overlaying, the required digit is displayed to the user which is illustrated in Fig. 4d and g.

3 Experimental Results

The system has been implemented using MATLAB version 13a. The iris images of 10 subjects were randomly selected from IITD iris image database [16]. The database is collection of the iris images (in bitmap format) collected from the students and staff at IIT Delhi, India. The datasets in the database have been

acquired in the Biometrics Research Laboratory using JIRIS, JPC1000, digital CMOS camera. The database of 2240 images has been acquired from 224 different users. All the subjects in the database are in the age group 14–55 years comprising 176 males and 48 females. The resolution of the images in the database is 320×240 pixels and all these images were acquired in the indoor environment. The biocode is generated using bioencoding technique which is better than biohashing technique as shown in Fig. 5. The shares (share 1 and share 2 shown in Fig. 6) are generated from the biocode, using the segment-based visual cryptography.

One of the shares is stored into template database during enrolment phase, instead of storing the biocode directly. The other share is given to the user. Every character in each share resembles digit ‘8’ in seven-segment display. Any imposter who gains access of the template database cannot guess the character from the individual share. The shares are just superimposed using simple logical ‘OR’ operation to recover the hidden biocode during authentication process. The recovered biocode is in seven-segment display pattern. During authentication

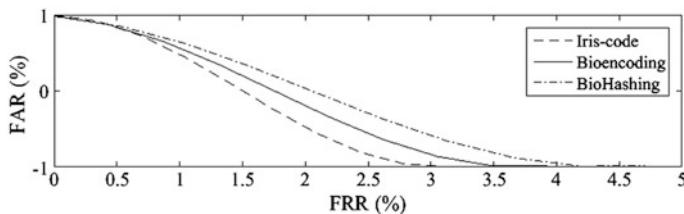


Fig. 5 Performance comparison of bioencoding and biohashing with respect to iris code

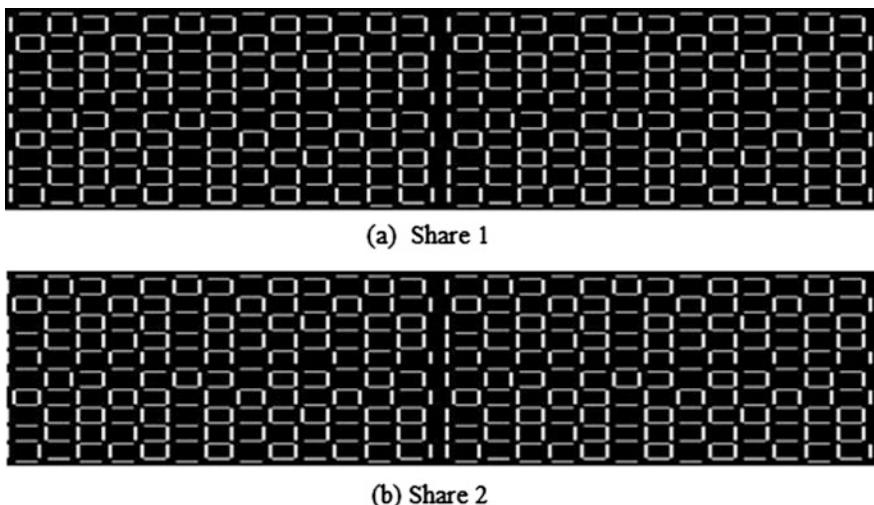


Fig. 6 Shares generated using segment-based visual cryptography

Table 1 EER (%) when using different datasets from IITD iris image database

IITD iris image database	EER (%)
001	4.65
002	4.67
003	4.63
004	4.58
005	4.63
006	4.76
007	4.72
008	4.78
009	4.76
010	4.79

process, the user has to present the share with him/her along with the biometric input. The biocode is generated again, using same method used during the enrolment process. A matcher is designed efficiently to match the recovered biocode (recovered by superimposing shares) with the biocode generated using the input provided by the user at present.

The equal error rate (EER), calculated with respect to false rejection rate (FRR) and false acceptance rate (FAR), has been used to examine the matching performance of the original as well as the reconstructed biocodes during the authentication phase. This resulted in ~4.7% as illustrated in Table 1.

4 Conclusion

This paper has explored the possibility of using segment-based visual cryptography for imparting privacy to the cancelable biometric templates. Since the spatial arrangement of the segments in each share varies, it is hard to recover the original template without accessing both the shares. The logical ‘OR’ operator is used to superimpose the two shares and fully recover the original biocode. It is observed that the reconstructed biocode is similar to the original biocode. Finally, the experimental results demonstrate the difficulty of exposing the identity of the secret image using only one of the shares; further individual shares cannot be used to perform cross-matching between different applications.

References

1. Patel, V.M., Ratha, N. K., Chellappa, R.: Cancelable Biometrics: A Review. *Signal Processing Magazine, IEEE* 32.5, 54–65 (2015)
2. Soutar, C., Roberge, A., Vijaya Kumar, B.V.K.: Biometric Encryption using Image Processing. *SPIE*: 178188 (1998)

3. Bolle, R.M., Connell, J.H., Ratha, N.K.: Biometrics Perils and Patches. Elsevier Pattern Recognition 35: 27272738 (2002)
4. Ratha, N., Chikkerur, S., Connell, J.H., Bolle, R.: Generating cancelable fingerprint Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 561–572 (2007)
5. Jeong, M.Y., Teoh, A.B.: Cancellable Face Biometrics System by Combining Independent Component Analysis Coefficient. Springer Berlin Heidelberg (2011)
6. Rathgeb, C., Busch, C.: Cancelable multi-biometrics: Mixing iris-codes based on adaptive bloom filters. Computers & Security, vol. 42, pp. 1–12 (2014)
7. Jain, A.K., Uludag, U., Hsu R.L.: Hiding a Face in a Fingerprint Image. Pattern Recognition, IEEE (2002)
8. Vatsa, M., Singh R., Noore A., Houck M.M.: Robust Biometric Image Watermarking for Fingerprint and Face Template Protection. IEICE Electronics Express 3.2 (2006)
9. Li, S., Kot, A.: Fingerprint Combination for Privacy Protection. IEEE Transactions on Information Forensics and Security, vol. 8, no. 2, pp. 350–360 (2013)
10. Ross, A., Othman, A.: Visual Cryptography for Biometric Privacy, IEEE Transactions on Information Forensics and Security, vol.6, no.1, pp. 70–81 (2011)
11. Daugman, J.: High Confidence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, no. 11, pp. 1148–1161 (1993)
12. Hollingsworth, K., Bowyer, K., Flynn, P.: The Best Bits in an Iriscode. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pp. 964–973 (2009)
13. Osama, O., Tsumura, N., Nakaguchi, T.: Bioencoding: A Reliable Tokenless Cancelable Biometrics Scheme for Protecting iris-codes. IEICE Transactions on Information and Systems (2010)
14. Borchert, B.: Segment-based visual cryptography. <http://www.bibliographie.uni-tuebingen.de> (2007)
15. Naor, M., Shamir, A.: Visual cryptography. Advances in Cryptology, In: Eurocrypt 1994. Springer, Heidelberg (1995)
16. IIT Delhi Iris Database version 1.0, http://web.iitd.ac.in/~biometrics/Database_Iris.htm

PCB Defect Classification Using Logical Combination of Segmented Copper and Non-copper Part

Shashi Kumar, Yuji Iwahori and M.K. Bhuyan

Abstract In this paper, a new model for defect classification of PCB is proposed which is inspired from bottom-up processing model of perception. The proposed model follows a non-referential based approach because aligning test and reference image may be difficult. In order to minimize learning complexity at each level, defect image is segmented into copper and non-copper parts. Copper and non-copper parts are analyzed separately. Final defect class is predicted by combining copper and non-copper defect classes. Edges provide unique information about distortion in copper disc. In this model, circularity measures are computed from edges of copper disc of a Copper part. For non-copper part, color information is unique for every defect type. A 3D color histogram can capture the global color distribution. The proposed model tries to compute the histogram using nonuniform bins. Variations in intensity ranges along each dimension of bins reduce irrelevant computations effectively. The bins dimensions are decided based on the amount of correlation among defect types. Discoloration type defect is analyzed independently from copper part, because it is a color defect. Final defect class is predicted by logical combination of defect classes of Copper and Non-copper part. The effectiveness of this model is evaluated on real data from PCB manufacturing industry and accuracy is compared with previously proposed non-referential approaches.

Keywords PCB · AVI · Copper part · Non-copper part · SVM

S. Kumar (✉) · M.K. Bhuyan

Indian Institute of Technology Guwahati, Guwahati 781039, India

e-mail: shashi.kumar@iitg.ernet.in

URL: <http://www.cvl.cs.chubu.ac.jp/>

M.K. Bhuyan

e-mail: mkb@iitg.ernet.in

Y. Iwahori

Chubu University, Kasugai 487-8501, Japan

e-mail: iwahori@cs.chubu.ac.jp

1 Introduction

Inception of Electronic Industry marked an era in digital revolution. Nowadays, electronic goods are everywhere in our lives. The new technology Internet of Things (IoT) plans to bring electronics even closer to our lives. The main basic component of Electronic Industry is Printed Circuit Board (PCB), on which ICs are mounted. Any alterations in PCB correctness may result in different circuit behaviors which are undesirable. Conventionally, industries employed humans to identify and classify defects in PCB. This is very slow, painful, and unreliable process. With the developments in Computer Vision, Automatic Visual Systems (AVI) were proposed for PCB inspection which are very fast and reliable.

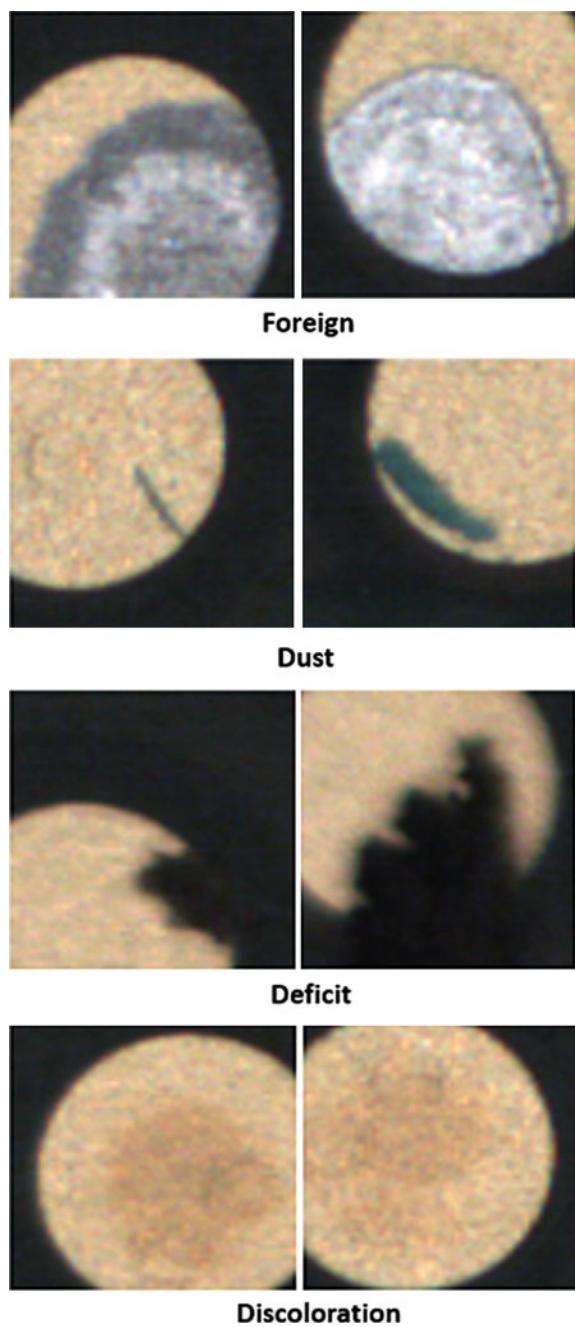
AVI models can broadly be divided into three classes: Referential Approach, Non-Referential Approach, Hybrid Approach. In referential approach, defect area is detected by subtracting test image with corresponding aligned reference image. But, perfect alignment between test and reference image is difficult. Non-referential approach involves extracting features from whole image based on truth establishment or properties. But, it is difficult to characterize every defect type. Hybrid approach combines models from both referential and non-referential approach.

Defect types include Open, Short, Deficit, Foreign, Discoloration, Dust, etc., as shown in Fig. 1. Some of these are true defects like Short, Deficit and some are pseudo defects like Dust. Pseudo Defects can be easily removed from PCB. In this model, types of defects covered are Deficit, No Defect, Foreign, Dust, Discoloration. Open, Short could not be covered due to unavailability of data.

Owing to the difficulties in referential approach, in this paper a non-referential model is proposed. As human vision model dictates, low-level abstract extraction first, then combining these to go upward pyramidal. That is what we have tried to model. To make learning complexity less, segmentation into copper and non-copper part is done. Identifying deviations independently in copper and non-copper parts and then combining these deviations to predict final defect is the goal of this paper.

2 Related Work

Many previous approaches has been proposed in all the three classes of solution model, Referential Approach, Non-Referential Approach and Hybrid Approach. Inoue et al. [1] proposed a non-referential method using Bag of Keypoints (BoK) and SVM as classifier. They formed Visual Word dictionary (VWD) of RootSIFT features from whole image using BoK. BoK Histogram features are then used for SVM learning and classification. Ibrahim et al. [2] proposed a new method to enhance performance of image difference operation in terms of computation time using wavelet transform. Defect Region in Referential approach is detected by taking difference of test image and reference image. Second level Haar wavelet is used for wavelet transform. They compute wavelet transform of both test and reference image and

Fig. 1 Defect classes

then perform the image difference operation. Heriansyah et al. [3] proposed a technique that adopts referential-based approach and classifies the defects using neural network. The algorithm segments the image into basic primitive patterns. These patterns are assigned and normalized and then used for classification. Classification are done using binary morphological image processing and Learning Vector Quantization (LVQ) Neural Network. For performance comparison, a pixel-based approach developed by Wu et al. was used. West et al. [4, 5] implemented a boundary analysis technique to detect small faults using Freeman Chain Coding [6]. Small faults are easily distinguishable features from normal electronic board. Freeman Chain Coding translates boundary into polygonal approximation which eliminates noise. In this method, first Euclidean distance and boundary distance of two boundary points are compared which are at constant number of chain segments apart.

In Hybrid Approach, Tsai et al. [7] proposed a Case-Based Reasoning (CBR) approach to classify defects. Indexing of past cases are done using k-means clustering. Similar cases are extracted from database using indexes for any new case.

3 Defect Classification by Logical Combination

3.1 *Outline of Proposed Approach*

In the proposed non-referential approach, test image is segmented into copper and non-copper part. This is done in order to reduce the learning complexity from the whole image. Segmentation is based on true color of observed copper.

1. For the Copper Part, center of outlined copper circle is detected using Hough Circle Transform. Range of radius used in Hough Circle algorithm is chosen around observed radius of copper disc. Observed radius r depends upon magnification or distance of scanner from the PCB. r can be considered constant for particular system, thus constant radius is assumed here. Using the deduced center coordinates, value of circle equation: $(x - h)^2 + (y - k)^2 - r^2$ is computed which should be around 0 for an undistorted circle. This value is computed for each pixel and pushed back in a vector. 1vR type SVM is used to learn these features.
2. For the non-copper part, 3D nonuniform color histogram is extracted. Dimensions of nonuniform intensity blocks are decided based on the measure of correlation or uncorrelation of non-copper images of different defect types. The pixels of non-copper part are distributed among the bins. Number of pixels in each bin are used as features. SVM with polynomial kernel is used to learn these features.
3. Discoloration type defect is detected separately because of requirement of different features from copper part. Histograms of RGB image are extracted as features and learned using SVM classifier with polynomial kernel.
4. Logical combination of identified defect classes of copper and non-copper part is done as shown in Table 1. Statistical error minimization is avoided so as not to include extra error.

Table 1 Logical combination table

Copper defect class	Non-copper defect class	Final defect class
No Defect	No Defect	No Defect
Deficit	No Defect or Deficit	Deficit
No Defect or Deficit	Foreign	Foreign
No Defect or Deficit	Dust	Dust
Discoloration	Any class	Discoloration

Outline of proposed approach is shown in Fig. 2.

3.2 Segmentation into Copper and Non-copper Part

PCB images taken from rear acute angle tend to reflect particular color with a little variation. Most of the systems use this type of lightning technique in capturing PCB images. Since, this color is retained in all of the copper part, true color-based segmentation is used to avoid computational complexity. Copper patches in Fig. 2 are extracted from test images. These patches display a range of RGB intensity levels. Discolored copper are also extracted as a part of copper due to saturation similarity. Pixels in test images are classified as copper pixels if they are within thresholding boundaries of RGB intensity levels of copper patches. Thresholds were decided empirically. The fact that human eye has different sensitivity for different colors is used for deciding thresholds for example green is more perceptible than red to human eye. Figure 3 shows copper and non-copper part of a test image.

3.3 Copper Part Feature and Classification

Copper part can be either good or chunked up from inside, losing connectivity, or losing circularity from boundary. Thus, two defect classes for copper part were identified: Deficit or No Defect. For copper part, edges can totally provide difference between circular outline and deviations from it. Keeping this in mind, Canny Edges are extracted. Minimum and maximum thresholds are decided empirically for Hysteresis thresholding. Assuming circular joint points in PCB, if AVI system is taking images with M magnification then,

$$r' = \sqrt{M}r \quad (1)$$

where r' is the new radius and r is original radius. Hough Circle transform [8] is used with restricted range of radius to detect center of copper edges (h, k) . The range of radius is bounded by the magnification M . Criteria for identification of circle by

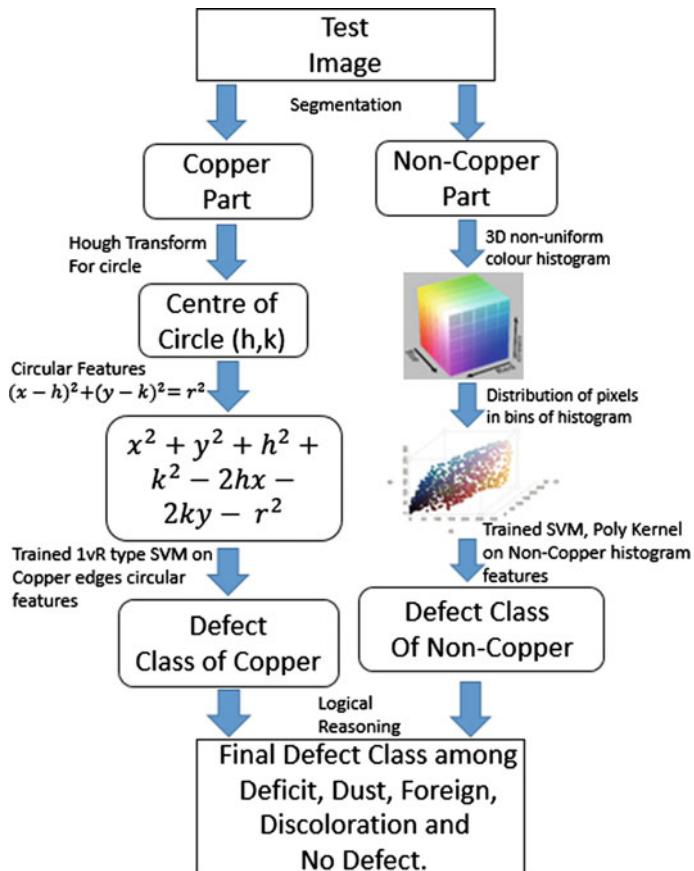


Fig. 2 Outline of proposed approach



Fig. 3 Defect image, Copper part, Non-copper part C

Hough Circle algorithm is the number of intersections in parameter space. Requirement of number of intersections are kept low considering the fact that most of the edge pixels may not lie on original circle owing to defects. Required center is determined by choosing center of circle whose radius is closest to required radius. It diminishes the possibility of detecting another circle because of low number of intersections criteria in parameter space and closeness to radius measure. Now, mathematical measure of circularity is computed for each of the edge pixel. Equation of circle with radius r , center (h, k) is

$$(x - h)^2 + (y - k)^2 - r^2 = 0 \quad (2)$$

Considering randomness caused by sampling, quantization, and Canny edge detector in (x, y) and (h, k) to be small, the value of Eq. 2 should be close to 0 for No Defect copper edges in every dimension. Each dimension is essentially each pixel. Owing to the method of sampling and quantization, expectation of the randomness caused can be assumed to be 0. Thus, there exist a cylinder with some radius r_1 such that in 2D

$$|(x - h)^2 + (y - k)^2 - r^2| < r_1^2 \quad (3)$$

where $|.|$ is modulus. Since, No Defect copper edge values are shown to be bounded, a tight wrapper can be found around the feature vector of No Defect in feature space. Any distortion in circle should cause the expected value to go outside wrapper. So, feature vectors of Deficit and No Defect in copper part can be well separated in feature space. 1vR type SVM is used for training which does exactly what is needed here. Enough iterations are provided to obtain good support vectors. The trained SVM is used for classification.

3.4 Non-copper Part Feature and Classification

Uniqueness of information in each types of defects lie in distribution and blending of colors. These features can be attributed by 3D color histogram in RGB space [9]. Non-copper part of defect classes like Foreign, Dust, etc., differ only in some intensity ranges of RGB color space. So, to reduce computation time as well as to increase performance, RGB color space can be divided into nonuniform bins. Bins intensity ranges in each color dimension can be decided by similarity or dissimilarity measure in histogram of defect types. RGB channel histogram of non-copper defect images are compared based on their correlation coefficients. To compare two histograms H_1 and H_2 , a metric $d(H_1, H_2)$ is computed, where

$$d(H_1, H_2) = \frac{\sum_l (H_1(l) - \bar{H}_1)(H_2(l) - \bar{H}_2)}{\sqrt{\sum_l (H_1(l) - \bar{H}_1)^2 (H_2(l) - \bar{H}_2)^2}} \quad (4)$$

and

$$\bar{H}_k = \frac{1}{N} \sum_J H_k(J) \quad (5)$$

where N is the total number of histogram bins. Total number of bins used are 256. Intensity ranges where $d(H_1, H_2)$ is fairly large in all of defect classes can have large bin size in that dimension of color. This divides the RGB space into nonuniform bins adaptive to color differences in non-copper part of defect types. Now, pixels are distributed in these bins. Number of pixels in bins are used as feature. SVM with polynomial kernel of degree 4 is used for training and classification. Kernel can be written as

$$K(x, y) = (\mathbf{x}^T \mathbf{y} + c)^4 \quad (6)$$

where \mathbf{x} and \mathbf{y} are feature vectors in input space.

3.5 Discoloration

Discoloration can be considered as true color defect in copper part. For copper classification, edge features are extracted as feature vector. But edge features cannot satisfy the purpose in discoloration. So for color features, histogram of RGB channel separately is extracted with 256 number of bins and intensity range 0–255. Histograms are the simplest color representation of an image and provide a very good global color representation of an image. Final feature dimension is RGB channel histogram of 256 bins pushed back, thus $256 * 3 = 768$. For this very high-dimensional feature vector SVM with polynomial kernel of degree 4 is used for training and classification. Termination criteria of SVM is selected to be iteration count but enough to find optimal Support Vectors. Accuracy obtained is 100 %.

3.6 Logical Combination of Copper and Non-copper Defect Classes

In order to reduce learning complexity, defect image is segmented to learn independent characteristics from copper and non-copper part. The same defect class can be obtained by focusing on different unique bunch of information posed by the individual part. Stochastic or probabilistic approach is not adopted because it will introduce unavoidable errors. These errors can be the result of overfitting on training data. Also, it is computationally expensive which might not be a good choice for a real-time system.

The final defect class can be easily perceived in terms of defects classes of the copper and non-copper part. The same analogy can be given for the logical combination reasons. Suppose p be the number of defect classes of copper part and q be

Table 2 Accuracy

Model	True defect		Pseudo defect		Accuracy (%)
	Correct	Incor.	Correct	Incor.	
Paper [10]	496	104	452	148	79.0
Paper [1]	532	68	572	28	92.0
Proposed	711	48	309	7	94.88

the number of defect classes of non-copper part, where $p, q \in \mathbb{N}$, then total number n of combinations of final defect types.

$$n = pq \quad (7)$$

Total number of final possible defect classes are finite and can be assumed to be fairly small, like order of 10. But final defect classes are Foreign, Dust, Discoloration, Deficit, No Defect. It shows that there are many repetitions. After combining all the classes of copper and non-copper part and ruling out repetitions, Table 1 is prepared.

So, final defect class is predicted from the classes obtained in copper and non-copper part using Table 1. It may be noticed that all the variations, local or global, in copper and non-copper parts are captured by this table.

4 Results

Accuracy obtained from proposed method and paper [10] and paper [8] is compared. Defect classes are No Defect, Deficit, Foreign, Discoloration, and Dust.

Result of accuracy is shown in Table 2.

The accuracy of deficit in copper part is slightly less because of large magnitude of random error in coordinates than expected. Also, large chunk off near original boundary of copper might have caused unpredictable results.

5 Conclusion

This paper proposes a non-referential model for PCB Defect classification. The accuracy can be seen to be better than other non-referential methods proposed in the literature. This paper classifies only one defect per image which may not be true in all cases. Maximum Likelihood Estimation (MLE) can be used to predict confidences for defects in a test image. Deep Learning Architectures like CNN can be also explored in extension of this research.

Acknowledgements This research was done while Shashi Kumar was visiting Iwahori Lab. as his research internship. Iwahori's research is supported by Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (C) (#26330210) and Chubu University Grant. The authors would like to thank the related lab member for the useful discussions and feedback.

References

1. Inoue, H., Iwahori, Y., Kjksirikul, B., Bhuyan, M. K.: SVM Based Defect Classification of Electronic Board Using Bag of Keypoints. ITC-CSCC 2015, 31–34 (2015)
2. Ibrahim, Z., Al-Attas, S. A. R., Aspar, Z.: Model-based PCB Inspection Technique Using Wavelet Transform. 4th Asian Control Conference, September 25–27 (2002)
3. Heriansyah, R., Al-Attas, S. A. R., Zabidi, M. M.: Neural Network Paradigm for Classification of Defects on PCB. Master Thesis, University Teknologi Malaysia (2004)
4. West, G. A. W., Norton-Wayne, L., Hill, W. J.: The Automatic Visual Inspection of Printed Circuit Boards. Circuit World, Vol. 8, No. 2, 50–56 (1982)
5. West, G. A. W.: A System for the Automatic Visual Inspection of BarePrinted Circuit Boards. IEEE Transactions of Systems, Man and Cybernetics, Vol. SMC-14, No. 5, 767–773, September/October (1984)
6. Freeman, H.: Computer Processing of Line-Drawing Images. ACM Computing Surveys, Vol. 6, No. 1, 57–97, (1974)
7. Tsai, C.-Y., Chiu, C.-C., Chen, J. S.: A case-based reasoning system for PCB defect prediction. Elsevier Expert Systems with Applications (2005)
8. Ioannou, D., Huda, W., Laine, A. F.: Circle recognition through a 2D Hough Transform and radius histogramming. Elsevier Image and Vision Computing, Vol. 17, 15–26 (1999)
9. Jones, M. J., Rehg, J. M.: Statistical Color Models with Application to Skin Detection. International Journal of Computer Vision, Vol. 46, No. 1, 81–96 (2002)
10. Inoue, H., Hagi, H., Iwahori, Y.: Defect Classification of Electronic Board Using RealAdaboost SVR in AVI. Tokai Section Joint Conference on Electrical, Electronics, Information and Related Engineering, H1-2 (2014)

Gait Recognition-Based Human Identification and Gender Classification

S. Arivazhagan and P. Induja

Abstract The main objective of this work is to identify the persons and to classify the gender of those persons with the help of their walking styles from the gait sequences with arbitrary walking directions. The human silhouettes are extracted from the given gait sequences using background subtraction technique. Median value approach is used for the background subtraction. After the extraction of the silhouettes, the affinity propagation clustering is performed to group the silhouettes with similar views and poses to one cluster. The cluster-based averaged gait image is taken as a feature for each cluster. To learn the distance metric, sparse reconstruction-based metric learning has been used. It minimizes the intraclass sparse reconstruction errors and maximizes the interclass reconstruction errors simultaneously. The above-mentioned steps have come under the training phase. With the help of the metric learned in the training and the feature extracted from the testing video sequence, sparse reconstruction-based classification has been performed for identifying the person and gender classification of that person. The accuracy achieved for the human identification and gender classification is promising.

Keywords Affinity propagation (AP) • Cluster-based averaged gait image (C-AGI) • Sparse reconstruction-based metric learning (SRML) • And sparse reconstruction-based classification (SRC)

S. Arivazhagan (✉) · P. Induja
Department of ECE, Mepco Schlenk Engineering College,
Sivakasi, India
e-mail: sarivu@mepcoeng.ac.in

P. Induja
e-mail: induja.r.s@gmail.com

1 Introduction

Gait Recognition is a task to identify or verify the individuals by their manner of walking. Gait offers several unique characteristics. Unobtrusiveness is the most attractive characteristic which does not require subject's attention and cooperation that is observed. Without the requirement of the physical information from subjects, human gait can be captured at a far distance. Moreover, gait recognition offers great potential for the recognition of low-resolution videos, where other biometrics technologies may be invalid due to the insufficient pixels to identify the person. There are three basic approaches for the gait recognition [1]. They are moving video-based, floor sensor-based, and wearable sensor-based gait recognitions. Moving video-based gait recognition uses video camera for capturing the gait at a distance. In Floor sensor-based gait recognition, a set of sensors are installed on the floor. When a person walks on this sensor floor, those sensors are enabled to measure the features which are related to the gait. For example, maximum time and amplitude values of the heel strike, etc. The most general method is based on the moving video-based gait recognition. The general gait recognition process contains three steps. They are background subtraction and silhouette extraction, Feature extraction and Recognition. In Background subtraction and silhouette extraction step, the moving object that is the moving persons are identified first in the frame. Then some of the background subtraction techniques are applied on those frames. It subtracts each frame from the background frame so that it identifies the moving objects that are differed from the background model. This is the preprocessing steps. This background subtraction method also generates binary images that contain black and white pixels which are also known as the binary silhouettes. This post processing step is used for the silhouette extraction with less noise. For that, some morphological operations like erosion, dilation can be used. This background subtraction technique is useful in many applications mainly in surveillance. There are some challenges in developing a good algorithm for the background subtraction. Some of them are robustness against for the changes in illumination, avoidance of detecting non stationary background objects like moving leaves, shadows of moving objects, etc. The next step in the general gait recognition would be feature extraction. It is a form of dimensionality reduction. If the input data is very large to process, then it will be transformed into a set of features which are in reduced representation. This type of transformation of input data into a set of features is called as feature extraction. The final step in the general gait recognition process is recognition of the person's gait. The input test video features are compared with the features of the training set sequence videos to identify the person's using their gait sequence. There are many different types of classifiers are available. Some of them are multilinear discriminant analysis (MDA), linear discriminant analysis (Lda), etc.

The paper is organized as follows Sect. 2 reviews the related work done in this area. The system model is described in Sect. 3. Results and Discussions are presented in Sect. 4 and Conclusions are given in Sect. 5.

2 Related Work

There are many research works have been done on gait recognition. The gait recognition can be of four types based on its view of walking directions and shape of the person's. They are shape invariant, shape variant, view invariant, and view variant. View-invariant gait recognition is the useful biometric gait recognition because people usually walk freely along the arbitrary directions that is not to walk along the predefined path or the fixed direction. Shape-invariant gait recognition is also very useful because people can carry any objects. The person's should be identified even though they carry some objects like bag, wearing coats, etc. So the efficient gait recognition should satisfy shape invariant property, view-invariant property or both of them. Some of the related works for the above-mentioned types are given below.

Han and Bhanu proposed a new spatial-temporal gait representation, called gait energy image (GEI), to characterize human walking properties for the gait recognition [2]. In this paper, a novel approach had been proposed for solving the problem of the lack of templates for the training by combining the statistical gait features from the templates of the real and synthetic. Real templates had been directly computed from the silhouette sequences which are taken for the training. Synthetic templates had been generated by the simulation of silhouette distortion from the training sequences. It is the type of view variant gait recognition.

Frey and Dueck proposed affinity propagation clustering for the view-invariant gait recognition [3]. For processing sensory signals and for detecting patterns in data, clustering of data is important that is done with the identification of a subset of representative examples. With the help of choosing randomly an initial subset of data points and refining it iteratively, those exemplars can be found out. It works well if and only if that choosing the initial subset will close to the good solution. For that “Affinity Propagation” method had been proposed. The input to this method is the measures of similarity between two data points. Real-valued messages have been exchanged between these data points. This exchange will be continued until the corresponding clusters gradually emerge and a high-quality set of exemplars have been obtained. This method of affinity propagation has been applied for clustering images of faces, detecting genes in microarray data, etc. It found the clusters which have much lower error than other clustering methods and it is a less time consuming process of clustering than other techniques of clustering.

Lu and Zhang proposed a gait recognition method using multiple gait features representations for the human identification at a distance based on the independent component analysis (ICA) [4]. The generic fuzzy support vector machine (GFSVM) can also be used for the representation of the gait features. This comes under the category of view-invariant gait recognition. The binary silhouettes have been obtained by simply performing the background modeling in which the moving objects that is human were subtracted from the background frame. Three kinds of gait representations have been used to characterize these binary silhouettes. They are Fourier descriptor, Wavelet descriptor, and Pseudo-Zernike moment. For the recognition step ICA and GFSVM classifiers have been chosen. To overcome the

limitation for the method of the recognition on the single view and make this method to robust against the view variant problem, one approach which is the multiple views of fusion recognition was introduced. That approach was based on the product of sum (POS) rule. When compared to the other traditional rank-based fusion rules approach, this method gave better performance.

Goff redo et al. presented a new method for viewpoint independent gait biometrics [5]. In this method, there is no need for the camera calibration and it relied on a single camera itself. It worked with the multiple ranges of camera views. These requirements have been achieved when the gait was self calibrating. The proposed method on this paper is suitable for the identification of human by gait with the help of the above-mentioned properties. This method is very useful to encourage using it in the application of the surveillance scenarios. This also comes under the category of the view invariant gait recognition type.

Kusakunniran et al. presented a novel solution using support vector regression (SVR) to create a view transformation model (VTM) from the different point of view [6]. A new method had been proposed to facilitate the process of regression to make an attempt to find the local region of interest under one angle of view to predict the motion information corresponding to another angle of view. This method achieved view independent gait recognition performance. Before the similarity measurements had been done, gait features had been normalized which are taken under various view angle into a single common view angle.

Muramatsu et al. proposed the arbitrary view transformation model (AVTM) for making the recognition method as a robust for the view invariant [7]. The 3D volume of gait sequences of the training subjects was constructed. By projecting the 3D volume gait sequences on the same view as target view, the 2D gait silhouette sequences had been generated for the training subjects. Part dependent view selection scheme (PdVS) had been included with the AVTM which divides the gait features into several parts. This method leads to an improved accuracy for the recognition.

Jia et al. proposed a method based on silhouette contours analysis and view estimation for the view-independent gait recognition [8]. Gait flow image and the head and shoulder mean shape of a human silhouette can be extracted by using Lucas–Kanade’s method. The static and dynamic features of a gait sequence can be preserved by using this LKGFI-HSMS method. To overcome the view variations, the view between a camera and a person can be identified for the selection of the gait feature of the target one.

Jeevan et al. proposed a representation of Gait for each cycle of the silhouettes using Pal and Pal Entropy (GPPE) [9]. Background subtraction was used for the silhouette extraction. Morphological operations had been carried out to remove the noises in the silhouettes. The Gait using Pal and Pal Entropy (GPPE) method had been proposed for the feature extraction. This proposed method was robust against the shape variance. Principal component analysis had been used for making the feature matrix from the features extracted. SVM classifier had been used for the classification.

From these related works on the gait recognition, a view invariant recognition using gait has been proposed [10]. For the identification and gender classification of a human who is walking in arbitrary directions, sparse representation-based classification is used.

3 System Model

The block diagram for the proposed method which is robust for the view invariant is shown in Fig. 1. This method includes two phases. They are training phase and the testing phase. Training phase includes video to frame conversion, binary silhouette extraction, clustering based on affinity propagation, feature extraction on each clusters, and computation of the sparse reconstruction-based metric for the training sequences. The testing phase includes binary silhouette extraction of the testing video sequence, clustering on that silhouette using affinity propagation method and obtaining the features for each clusters as cluster-based averaged gait image. Then with the help of the learned metric which was obtained in the training phase and the feature extracted from the testing video sequences, the human is identified and based on the residuals value calculated in the sparse reconstruction-based classification, the gender of that person is classified. This method uses sparse

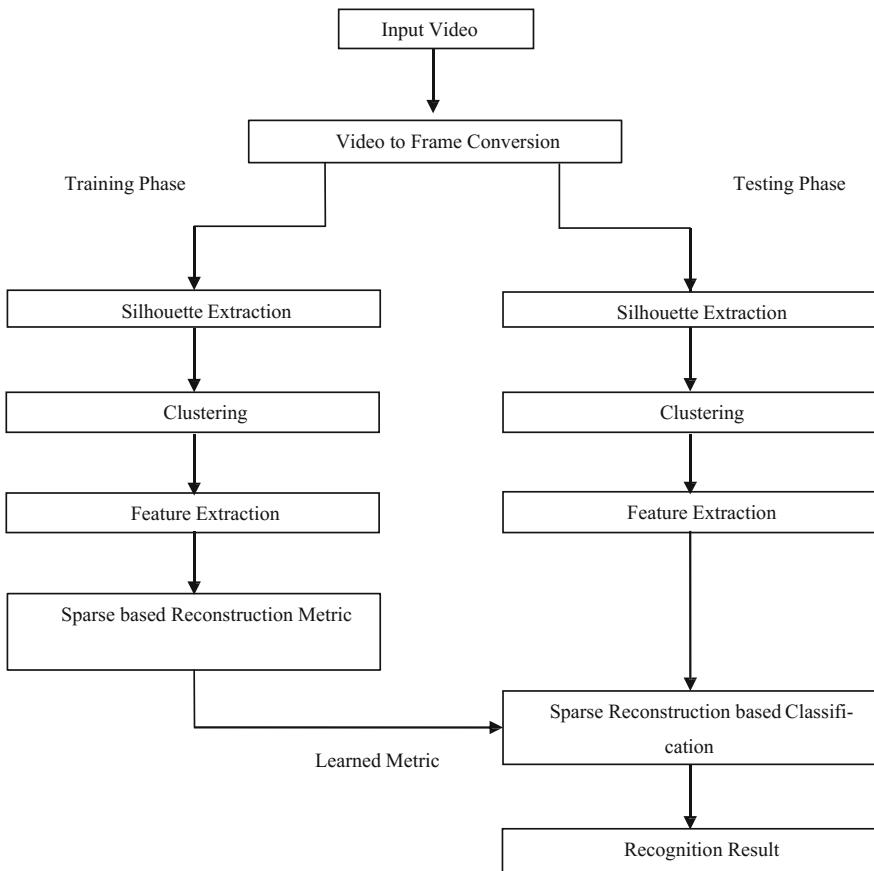


Fig. 1 Block diagram of the proposed method for the human identification and gender classification based on the gait recognition

reconstruction-based metric learning which gives more accuracy than other metric learning methods even though it is a time-consuming process [10]. This method is robust against the view variance but sensitive to the shape variance.

3.1 Training Phase

3.1.1 Video to 2.2.2 Frame Conversion

The input videos for training phase contain the gait sequence of the individual person. The videos are converted into frames. These frames are used for the further steps. The video to frame conversion is the initial step of the proposed method of the identification of human and the classification of the gender based on the gait recognition.

3.1.2 Silhouette Extraction

The frames which got from the input video of gait sequences are given for the silhouette extraction technique. Silhouette Extraction is nothing but the extraction of the human body from the background of the video. It is a type of the moving object detection. For the background subtraction technique, median value approach is used. In this approach, the background frame is created by taking the median value of each pixel among all the frames of the video. Median Value approach is the better technique than other background subtraction technique because the information cannot be lost due to the usage of the median value among the pixels. The moving object will be extracted from that frame, i.e., for our case, the moving object would be the person. Even though the moving objects are extracted from the foreground, there will be some noise in each frame. So, morphological operations are performed to obtain the perfect silhouette. Then thresholding takes place. Here Binary thresholding is used. The frames which have been obtained after this thresholding step contain the silhouettes of the human in the video. The silhouette frames are aligned to the same size because the silhouette size will vary due to the distance between the moving person and the fixed camera. Otherwise, the silhouettes of the same person will be resulted as a different person due to the size variation of the silhouettes in the frames. For that alignment of the silhouettes in each frame, bounding box method is used.

3.1.3 Clustering

The clustering is the next step in the training phase after performing the preprocessing steps which includes background subtraction, silhouette extraction. In this proposed method, affinity propagation clustering method is used for the clustering. The gait energy image (GEI) feature is powerful and common feature in

representing human gaits. Since the proposed method is view invariant, it is very hard to estimate the gait period in the gait sequence. In addition to that difficulty, the GEI feature is very sensitive to the view change. Hence, the computation of the GEI feature for the whole gait sequence is not possible. To address this difficulty, clustering is performed to cluster each gait sequence into many clusters in which each cluster is obtained by gathering human silhouettes which has similar views or poses. K -means clustering which is the popular clustering method to obtain the clusters for each gait sequence has the disadvantage of the requirement of knowing the number of clusters in prior. But the number of clusters in testing video will not know in prior. To address this, the affinity propagation (AP) clustering method is used to obtain the clusters that have the main advantage that it does not require any information about the number of clusters in prior.

Affinity propagation takes input as a collection of similarities of real values between two data points. The similarity $\text{similar}(i, k)$ indicates that how exactly the data point having the index k is behaving as an exemplar for the data point having index i . For example, consider two data points' x_i and x_k , the similarity between these two points would be given in Eq. (1).

$$\text{similar}(i, k) = \|x_i - x_k\|^2 \quad (1)$$

The data points with the larger values of $\text{similar}(k, k)$ is chosen as preferences. The responsibility $\text{res}(i, k)$, is sent from data point i to candidate exemplar point k . It reflects the evidence that how well the data point indexing k is an exemplar for the data point indexing i which also taking into account other potential exemplars for data point indexing i . The availability $\text{avail}(i, k)$ is sent from candidate exemplar point k to point i . It reflects evidence that the how data point i confidently chosen the data point k as its exemplar, which also taking into account the support from other points that point k should be an exemplar to that data points. To start with, the availabilities are initialized to zero, i.e., $\text{avail}(i, k) = 0$. Then, the responsibilities are computed using the rule which is given in Eq. (2)

$$\text{res}(i, k) < -\text{similar}(i, k) - \max_{k' \neq k} \left\{ \text{avail}(i, k') + \text{similar}(i, k') \right\} \quad (2)$$

In later iterations, the availabilities of some data points which are effectively assigned to other exemplar data points will drop below zero. The corresponding candidate exemplars will be removed by these negative availabilities from competition. The availability formula is given below in Eq. (3)

$$\text{avail}(i, k) < -\min \left\{ 0, \text{res}(k, k) + \sum_{i' \neq i, i' \neq k} \max \left\{ 0, \text{res}(i', k) \right\} \right\} \quad (3)$$

The positive parts of incoming responsibilities are added in each data points because there is a necessity for a good exemplar to explain certain data points and

explain other data points poorly. The self-availability $\text{avail}(k, k)$ is updated differently which is given by the Eq. (4).

$$\text{avail}(k, k) < - \sum_{i' \text{ s.t. } i' \neq k} \max\left\{0, \text{res}(i', k)\right\} \quad (4)$$

Availabilities and Responsibilities are combined for identifying exemplars in this affinity propagation clustering. For data point indexing i , the value of the data point indexing k that maximizes $\text{avail}(i, k) + \text{res}(i, k)$ either identifies data point indexing i as an exemplar if $k = i$, or identifies the data point that is the exemplar for point data point indexing i . This message passing may be terminated after a fixed number of iterations after which there is no much change in the exemplars of each data point.

3.1.4 Feature Extraction

The cluster-based averaged gait image is extracted as the gait feature. Consider for a given a gait sequence, the k th cluster contains N_k frames. The formula for obtaining the average gait image is given in Eq. (5)

$$G_k(x, y) = \frac{1}{N_k} \sum_{p=1}^{N_k} I_{pk}(x, y) \quad (5)$$

where— $I_{pk}(x, y)$ is the p th human silhouette in the k th cluster and x and y are the 2D Image Coordinates.

Since Gait cycle estimation is very difficult in the arbitrary walking directions, cluster-based averaged gait image is extracted as a feature. In gait feature, a high intensity-valued pixel represents more variations of poses which are static and a low intensity indicates that more dynamic information will present at that position since human walking occurs frequently in that direction or position.

3.1.5 Sparse Reconstruction-Based Metric Learning

The sparse reconstruction-based metric learning (SRML) is used to learn the distance metric [10]. It minimizes the intraclass reconstruction errors and at the same time this metric learning maximizes the interclass reconstruction errors. The main advantage of the usage of this metric learning is that it does not require the information about the number of frames in each cluster. The algorithm for SRML is given below:

Algorithm:

Input: Training set for the different subjects $Q = [Q^1, Q^2, Q^3, \dots, Q^c]$, iteration number J , convergence error ϵ ,

Output: Distance metric H .

Step 1 (Initialization): Initialize $H: H = I^{d \times d}$

Step 2 (Local optimization): For $e = 1, 2, \dots, R$, repeat
2.1. Compute A and B by using equation (6) and (7)

$$\min_{a_{ij}^e} (a_{ij}^e - SS_{ij}^e a_{ij}^e)^T W (q_{ij}^e - SS_{ij}^e a_{ij}^e) + \lambda \|a_{ij}^e\|_1 \quad (6)$$

$$\min_{b_{ij}^e} (q_{ij}^e - DD_{ij}^e b_{ij}^e)^T W (q_{ij}^e - DD_{ij}^e b_{ij}^e) + \lambda \|b_{ij}^e\|_1 \quad (7)$$

2.2. Compute $V1$ and $V2$ by using equation (8) and (9)

$$v1 \triangleq \sum_{c=1}^C \sum_{l=1}^L \sum_{j=1}^{K_{cl}} (q_{ij}^e - DD_{ij}^e b_{ij}^e) (q_{ij}^e - DD_{ij}^e b_{ij}^e)^T \quad (8)$$

$$v2 \triangleq \sum_{c=1}^C \sum_{l=1}^L \sum_{j=1}^{K_{cl}} (q_{ij}^e - SS_{ij}^e a_{ij}^e) (q_{ij}^e - SS_{ij}^e a_{ij}^e)^T \quad (9)$$

2.3. Solve the Eigen value problem as

$$(v1 - v2)h = \lambda h \quad (10)$$

2.4. Obtain $H_r = [h1, h2, \dots, hL]$.

2.5. Update

$$q_{ij}^e; q_{ij}^e = H^T q_{ij}^e$$

2.6. If $e > 2$, and $|H^r - H^{r-1}| < \epsilon$, go to step 3.

Step 3 (Output Distance Metric): Output Distance Metric: $H = H_r$

3.2 Testing Phase

In the testing phase, given a testing gait sequence T , the human silhouettes are extracted and they are clustered into K groups of clusters. For each group, the C-AGI is calculated as the gait feature which are denoted as $T1, T2, T3, \dots, Tk$. The residuals $\text{residual}_c(G_{RNB(k)}^{RC})$ using sparse reconstruction-based classification (SRC) for the k th C-AGI Tk , under the learned metric H are calculated by using the Eq. (11).

$$\text{residual}_c(T_k) = \left\| H^T G_{RNB(k)}^{RC} - H^T Q \delta_c(G_{RNB(k)}^{RC}) \right\|_2 \quad (11)$$

$$\text{Correct person} = \min(\text{residual}(G_{RNB(k)}))$$

where $\delta_c(G_k^{RC})$ is a correlation coefficient between the reconstructed image and the original silhouette image

After calculating the residuals, the subject for which holds the minimum of the residual value will be the correct person. The gender classification also is performed based on these residual values. The minimum residual value subjects will be labeled as female and maximum value residuals will be labeled as male. The label assignment is given in Eq. (12).

$$Z(c) = \sum_{k=1}^K \text{residual}_c(G_{RNB(k)}) * \delta_c(G_{RNB(k)}^{RC}) \quad (12)$$

$$\min(z) - \text{Label 0(Female)} \max(z) - \text{Label 1(Male)}$$

4 Results and Discussions

Videos from the Advanced Digital Sciences Center-Arbitrary Walking Directions (ADSC-AWD) dataset were used for the analysis. This dataset is for the gait recognition in arbitrary walking directions.

The ADSC-AWD gait dataset was collected by the Microsoft Kinect depth sensor. The Kinect depth camera fixed on a tripod is used to capture gait sequences on two different days in two rooms. Participation in the collection process was voluntary. The size of each room is about $8 \times 6 \times 4$ m. The distance from camera to the person is between 1.5 to 5 m [10].

The video is converted into frames and the silhouettes are extracted using background subtraction technique. All the silhouette frames are arranged to the same size using the bounding box method. Some of the aligned silhouette frames are shown in Fig. 2 and the affinity propagation clustering is performed. The CAGI features for some of the clusters are shown in Fig. 3.

With the help of these features, the metric is calculated using sparse reconstruction-based metric learning which is explained in the algorithm mentioned in Sect. 3. After the calculation of the metric, the testing has been performed.

The testing video is given and the steps up to feature extraction have been performed and the sparse representation-based classification has been performed in which the residuals are calculated for each cluster with respect to each subject for the given testing video sequence. Based on these residuals the gender of that person is also classified. The Experimental results for Human Identification and Gender Classification using ADSC-AWD dataset are shown in Table 1.

- Total Number of Subjects—6
- Number of Sequences taken for training per subject—2
- Number of Sequences taken for testing per Subject—2



Fig. 2 Silhouette frames

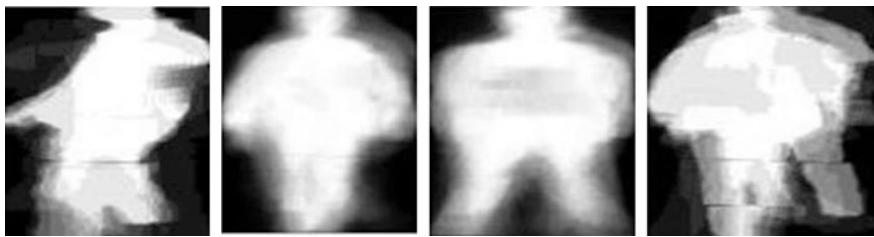


Fig. 3 CAGI features for some clusters

Table 1 Experimental results for human identification and gender classification using ADSC-AWD dataset

S. No	Testing video name	Whether person is correctly identified or not	Whether gender classification is correct or not
1	David_3	Yes	Yes
2	David_4	Yes	Yes
3	Hlk_3	Yes	Yes
4	Hlk_4	No	Yes
5	Cq_3	Yes	No
6	Cq_4	Yes	Yes
7	Hw_3	Yes	Yes
8	Hw_4	Yes	Yes
9	Hzy_3	Yes	Yes
10	Hzy_4	Yes	Yes
11	ljw_3	Yes	No
12	ljw_4	Yes	Yes

From the Table 1, it is observed that the human identification gives the better result and the accuracy of 91.6 % has been achieved. The gender classification gives some misclassification. The accuracy of 83.3 % has been achieved.

5 Conclusion

In this proposed method, human identification and gender classification have done by using gait recognition. The background subtraction technique is used for extracting silhouettes which includes the median value approach. Then clustering is performed in which the human silhouettes with similar views and poses are grouped together to form a cluster. For that Affinity Propagation is used which has the special advantage that there is no need to specify the number of clusters prior to the clustering. After that Feature is extracted as a cluster-based averaged gait image.

Sparse reconstruction-based Metric Learning is used to learn the distance metric. Sparse reconstruction-based classification is used for the recognition as well as the gender classification. The experiments are conducted on the ADSC-AWD dataset. The future scope will be made this method for the large number of databases and to make it as not sensitive to the shape of the subjects.

Acknowledgment The authors wish to express humble gratitude to the Management and Principal of Mepco Schlenk Engineering College, for the support in carrying out this research work.

References

1. Prena Arora and Rajni, "Survey on Human Gait Recognition," *International Journal of Engineering Research and General Science.*, Vol. 3, Issue 3, May-June, 2015.
2. J. Han and B. Bhanu, "Individual Recognition Using Gait Energy Image," *IEEE Trans. Pattern Anal. Mach. Intelligence.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.
3. B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.
4. J. Lu and E. Zhang, "Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fu- sion," *Pattern Recognition Letter*, vol. 28, no. 16, pp. 2401–2411, 2007.
5. M. Goffredo, I. Bouchrika, J. Carter, and M. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 997–1008, Aug. 2010.
6. W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Support vector regression for multi-view gait recognition based on local motion feature selection," in Proc. IEEE Int. Conf. Computer. Vis. Pattern Recognition., June. 2010, pp. 974–981.
7. Daigo Muramatsu, Akira Shiraiishi, Yasushi Makihara, Md. Zasim Uddin, and Yasushi Yagi, "Gait-Based Person Recogni- tion Using Arbitrary View Transformation Model", *IEEE Transactions On Image Processing*, Vol. 24, no. 1, January 2015.
8. Songmin Jia, Lijia Wang, and Xiuzhi Li, "View-invariant Gait Authentication Based on Silhouette Contours Analysis and View Estimation", on *IEEE/CAA Journal Of Automatica Sinica*, Vol. 2, No. 2, April 2015.
9. M.Jeevan, Neha Jain, M.Hanmandlu, Girija Chetty, "Gait Recognition based on Gait PAL and PAL Entropy Image", in *Proc. 8th IEEE Int. Conf. Image Processing*, 2013, pp. 4195–4199.
10. Jiwen Lu, Gang Wang, and Pierre Moulin, "Human Identity and Gender Recognition from Gait Sequences with Arbitrary Walking Directions", *IEEE Transactions on Information Forensics and Security*, Vol. 9, No. 1, January 2014.
11. J. Lu and Y.-P. Tan, "Uncorrelated Discriminant Simplex Analysis For View-Invariant Gait Signal Computing," *Pattern Recognition. Letter*, vol. 31, no. 5, pp. 382–393, 2010.
12. W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technology.*, vol. 22, no. 6, pp. 966–980, June. 2012.
13. Daigo Muramatsu, Akira Shiraiishi, Yasushi Makihara, Md. Zasim Uddin, and Yasushi Yagi, "Gait-Based Person Recogni- tion Using Arbitrary View Transformation Model", *IEEE Transactions On Image Processing*, Vol. 24, no. 1, January 2015.
14. Songmin Jia, Lijia Wang, and Xiuzhi Li, "View-invariant Gait Authentication Based on Silhouette Contours Analysis and View Estimation", on *IEEE/CAA Journal Of Automatica Sinica*, Vol. 2, No. 2, April 2015.

Corner Detection Using Random Forests

Shubham Pachori, Kshitij Singh and Shanmuganathan Raman

Abstract We present a fast algorithm for corner detection, exploiting the local features (i.e. intensities of neighbourhood pixels) around a pixel. The proposed method is simple to implement but is efficient enough to give results comparable to that of the state-of-the-art corner detectors. The algorithm is shown to detect corners in a given image using a learning-based framework. The algorithm simply takes the differences of the intensities of candidate pixel and pixels around its neighbourhood and processes them further to make the similar pixels look even more similar and distinct pixels even more distinct. This task is achieved by effectively training a random forest in order to classify whether the candidate pixel is a corner or not. We compare the results with several state-of-the-art techniques for corner detection and show the effectiveness of the proposed method.

Keywords Feature extraction • Corner detection • Random forests

1 Introduction

Many computer vision and image processing tasks require highly accurate localization of corners along with fast processing. With these needs, many recent works have been proposed to address the corner detection problem. However, despite the increasing computational speed, many state-of-the-art corner detectors still leave a little time for processing live video frames at the capture rate. This motivates the need for faster algorithms for corner detection. Our method tackles the two key issues

S. Pachori (✉) · K. Singh · S. Raman
Electrical Engineering, Indian Institute of Technology Gandhinagar,
Gandhinagar, Ahmedabad, India
e-mail: shubham_pachori@iitgn.ac.in

K. Singh
e-mail: kshitij.singh@iitgn.ac.in

S. Raman
e-mail: shanmuga@iitgn.ac.in

of corner detection—high accuracy and computational time. Apart from these, the algorithm is generic which can be used to detect corners in a given image of any natural scene.

We propose a method to detect corners trained on random forests on a set of corners obtained from training images. We use the trained random forests in order to detect corners in a new image. We show that the proposed approach is effective in the detection of corners in natural images. The algorithm could be used in a variety of computer vision tasks which rely on corner detection such as object recognition, image registration, segmentation, to name a few.

The primary contributions of this work are listed below.

1. A random forest-based learning framework for fast and accurate corner detection.
2. A simple and efficient way to obtain the feature descriptors for training images and the test image.

The rest of the paper is organized as below. In Sect. 2, we briefly review different major approaches for corner detection attempted in the past. In Sect. 3, we discuss the proposed method for corner detection. In Sect. 3.2, we present and compare our results with the state-of-the-art methods proposed earlier. In Sect. 4, we conclude our paper with directions for extending this work.

2 Related Work

The corner detection algorithms proposed in the past could be broadly divided into three main categories: (a) contour-based detectors, (b) model-based detectors and (c) intensity-based detectors as described in [1]. A more detailed survey is presented in [2]. Intensity-based corner detectors rely on the grey-level information of the neighbourhoods of pixels in order to detect the corners. Moravec gave the idea of defining points where a high intensity variation occurs in every direction as points of interest [3]. Harris and Stephens introduced a rotation invariant corner detector [4]. But, Harris corner detector gave poor results while detecting higher order corners as shown in [5]. Mikolajczyk and Schmid presented a novel technique to detect corners invariant to scale and affine transformation [6]. The calculation of first-order and second-order derivatives is computationally expensive. Moreover, the second-order derivatives are very sensitive to noise. SUSAN, proposed by Smith and Brady, uses a mask ‘USAN’ and compares the brightness corresponding to the nucleus of the mask, with the intensity of each pixel within the mask [7]. Corners in the image are represented by the local minima of the USAN map. Rosten and Drummond put forward the FAST algorithm which uses machine learning technique for corner detection [8]. It outperforms the previously proposed methods in computation time and high repeatability but leads to bad accuracy, giving some responses even along the edges. To improve this detection technique AGAST was proposed by Mair et al. [9]. This method apart from being less computationally expensive gives high performance for arbitrary environments. Later, Rosten et al. came with a new version of

their early proposed FAST, called FASTER, which increased the accuracy but with the slightly increased computational time [10]. Jinhui et al. presented a double circle mask and presented their intuition behind the working of the double mask in the noisy environments [11]. We follow their procedure and use a double mask in our corner detection framework.

3 Proposed Approach

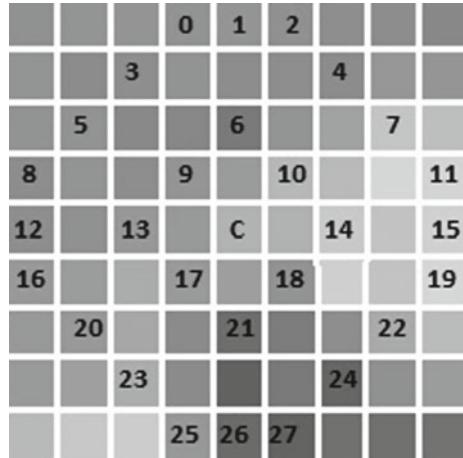
3.1 Motivation for Random Forest

Classification forest [12] is built from an ensemble of classification trees which form a class of non-parametric classifiers [13]. Each tree aims to achieve the classification task by splitting the node into multiple nodes based on some decision criterion. The splitting is performed at each internal (non-leaf) node based on problem-specific metrics for achieving the best splits. Generally, we do not make any assumption about the relationships between the predictor variables and dependent variables in the classification decision trees. Terminal nodes contain responses which classify the candidate into different classes, given its attributes. In this paper, we use information gain as the metric to make a split in each tree. Constructing a single decision tree has been found not giving desired results in most practical cases. This is because even small perturbations and noise in the data can cause changes in the results [14]. The key motivation for using random forest is to reduce the degree of correlation between different trees in forest, which then ultimately leads to a better generalization. We would like to train a random forest using corners detected in a set of natural images in order to detect corners in a given new natural image.

3.2 Random Forest Training and Corner Detection

We propose the usage of a double ring circle structure to be the feature descriptor as proposed in [15]. The intuition behind using this arrangement is that if there is a corner then there is a continuous change in its intensity values rather than abrupt changes. Hence, we could ignore the nearest 8-pixel neighbourhood circle around the candidate pixel. We could have also used the same three-circle mask as proposed in FASTER [10]. But we could remove the middle circle in those three consecutive circles, without much loss in performance due to the same reason. The two alternate circles are sufficient enough to classify a candidate pixel as corner or not. A double circle is also robust to noise as is shown in [15]. Therefore, we propose to use the circular mask as shown in Fig. 1 for extracting the feature descriptors. Let the intensity of the pixel at the nucleus of the mask be denoted by I_c and intensity of pixels within the mask be denoted by $I(x, y)$. The corner locations in the images are extracted using the Harris corner detector [4].

Fig. 1 The mask used by the proposed approach for feature description. The candidate pixel is marked by C



We then take the difference between the intensity value of the candidate pixel and the pixels of the double circular mask, $I_c - I(x, y)$ and form a vector \vec{d} , consisting of elements, $d_i, i = 0, 1, \dots, 27$. Rather than generically grouping them into darker, similar and lighter pixels as done in some of the previous works, we rather feed these values into a function as shown in Fig. 2a or b. The idea is to make the similar pixels look more similar and different pixels look more different. Any function which looks similar to these functions could do the job for us. One such function could be built using the hyperbolic tangent function. The equation of the function used by us could be written as:

$$\hat{d}_i = a(\tanh(b \times d_i + c)) + a(\tanh(b \times d_i - c)) \quad (1)$$

where a , b , and c are real constants.

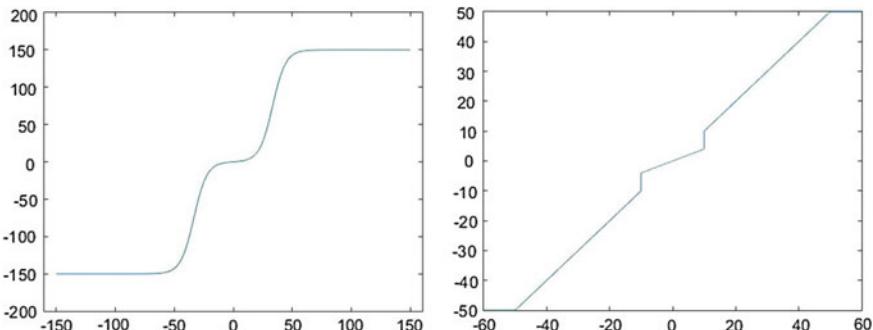


Fig. 2 Here d_i and \hat{d}_i are in x- and y- axes, respectively. **a** Hyperbolic tangent function shown in Eq. 1, and **b** function built using ReLU [16] shown in Eq. 2

But the function built using \tanh is computationally expensive. Therefore we resort to a modified version of the function ReLU (Rectified Linear Unit) as proposed in [16] which is defined as $\max(0, d_i)$. The function used by us built on ReLU is:

$$\hat{d}_i = \begin{cases} b \times d_i, & d_i \in [-a, +a] \\ \min(-a + \max(0, d_i + a), a), & \text{otherwise} \end{cases} \quad (2)$$

where a and b are constants.

The first motivation behind this kind of learning is that small changes in intensities which are sufficient enough to classify a pixel as corner could be detected. Therefore, we have not defined some global threshold to group the pixels into the classes of darker, similar and lighter. Second, by not grouping the pixel values in these classes, we have not tried to make the trees learn by the class categories (similar, darker and lighter) as this method was observed to give poor results. The reason may be that dividing the neighbour pixels into three or five classes would have probably overfitted the data as the training examples are huge and would have got the trees to memorize the pattern. The difference between these two methods will be discussed in the next section through results. The values obtain after processing the difference of intensities using the function in Eq. 2 would then serve as the final feature vector for the training set used to train the forest. Given a new image, we perform the same operation to extract the feature descriptors at all the pixel locations and mark the corners depending on the responses from the trained random forest. The decision is made for a given pixel to be a corner based on the average corner probability of the multiple tree outputs in the random forest. The complete procedure is described in Algorithm 1.

Algorithm 1 Corner Detection using Random Forest

- 1: **Step 1:** Detect corners in the N training images using Harris corner detection [4].
 - 2: **Step 2 :** Train a random forest (with k trees) using the feature descriptors (shown in Fig. 1) around the detected corners and non-corner pixels.
 - 3: **Step 3:** Extract feature descriptors from the test image using the pattern shown in Fig. 1.
 - 4: **Step 4:** Use the trained random forest to detect corners in the test image.
-

We trained our random forests on an Intel i7-3770 processor at 3.40 GHz, with 8GB installed memory (RAM) on 64 bit Windows operating system in Python. We used OpenCV-Python library for image processing tasks and for comparisons between different methods of corner detection. We fixed the value of a and b in Eq. 2 to be 50 and 0.3, respectively, for the work. We compare the results obtained by our method using two cases. Case (i): we do not divide the pixels based on their relative intensities and Case (ii): we classify them as darker (0), similar (1) and lighter (2) as shown in Eq. 3.

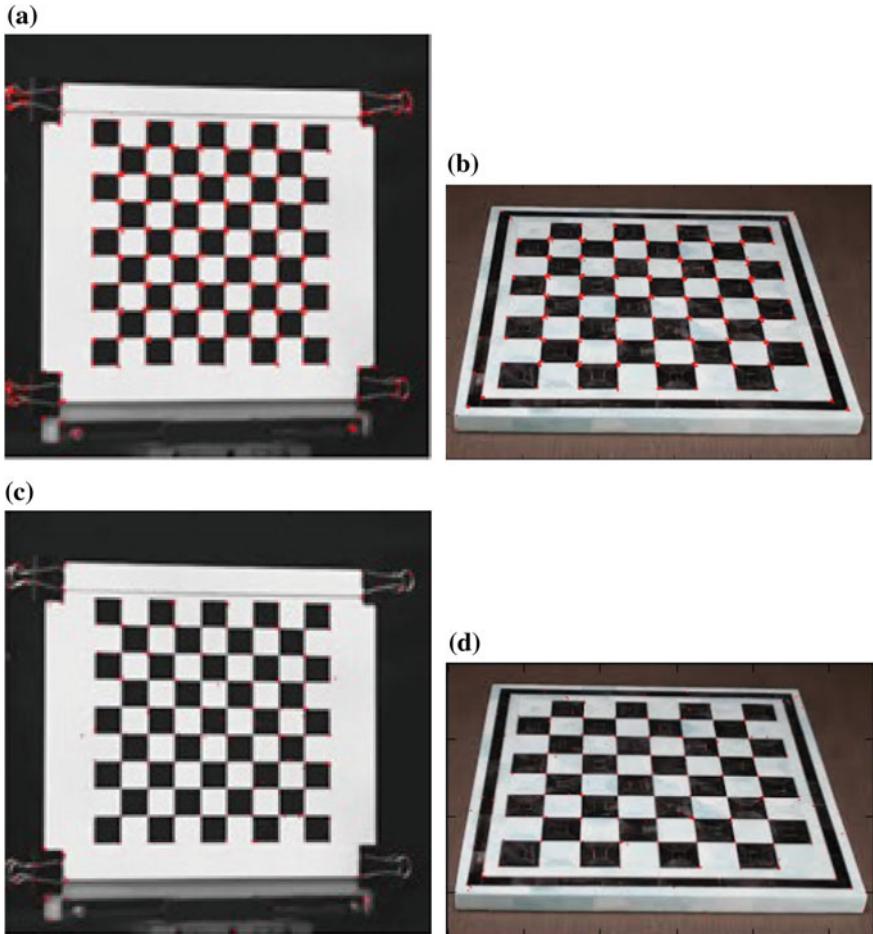


Fig. 3 **a** and **b** are the images obtained by case (i), **c** and **d** are the images obtained by case (ii)

$$d_i = \begin{cases} 0, & I_c - I(x, y) < -10 \\ 1, & -10 < I_c - I(x, y) < 10 \\ 2, & I_c - I(x, y) > 10 \end{cases} \quad (3)$$

For this comparison, we obtained training data from 35 images and the random forest comprised 15 trees. The results obtained are shown in Fig. 3. It can be observed that case (i) leads to better corner detection compared to that of case (ii). Therefore, we use case (i) to compare with other state-of-the-art methods.

For comparison, the number of trees trained was chosen to be 15 without any pruning of leaves. We created the training set from each and every pixel of 52 natural images. The training of the random forest took around 5 hours. We compare the

results of our algorithm with that of Harris detector [4], SUSAN Corner Detector [7] and FAST-9 detector [8]. Average time taken by the methods for a typical 256×256 image were 0.07 s, 7.98 s and 6.86 s respectively. Harris corner detector is faster due to the inbuilt OpenCV-Python function. The best time taken by our method to compute the feature vector on the same image while using the proposed function built using Eq. 2 was 6.14 s and time taken to predict the corners in image, via 15 trees, was 0.3 s. It is much faster than the proposed function built on tanh function in Eq. 1 which took 31.3 s to compute the feature vector without any decrease in the accuracy. The time taken to compute the feature vector without using the proposed functions reduced to 1.12 s.

The images (none of which were trained) shown in Fig. 4 depict the accuracy of our algorithm. Harris corner detector still outperforms the other approaches in terms



Fig. 4 First column—original images, corners detected using: second column—Harris corner detector [4], third column—FAST corner detector [8] fourth column—SUSAN [7], fifth column—proposed method

of accuracy. FAST corner detector showed poor results in terms of accuracy giving some points along the edges as could be seen in images in fifth and third rows of third column. SUSAN detected some extra potential non-corner points as corners. Our approach, though giving less number of corner points, gives nearly the same results as obtained from the Harris corner detector.

4 Conclusion

We have presented a fast learning-based method to detect the corners, whose performance is comparable to that of the state-of-the-art corner detection techniques. The choice of learning multiple decision trees using pre-detected corners along with the feature descriptor enables us to achieve this task. We are able to achieve high degree of accuracy in terms of detected corners using the proposed random forest framework. With very less number of trees and training images, we are able to detect corners in a given new image. In future, we would like to exploit random forests to detect the scale-invariant interest points in a given image and estimate the feature descriptors at these interest points. We would like to accelerate the framework proposed using GPU for various computer vision tasks. We believe that this framework will lead to increased interest in the research of learning-based corner detectors suitable for various computer vision applications.

References

1. A. Dutta, A. Kar, and B. Chatterji, Corner detection algorithms for digital images in last three decades, *IETE Technical Review*, vol. 25, no. 3, pp. 123133, 2008.
2. T. Tuytelaars and K. Mikolajczyk, Local invariant feature detectors: a survey, *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177280, 2008.
3. H. P. Moravec, Obstacle avoidance and navigation in the real world by a seeing robot rover. DTIC Document, Tech. Rep., 1980.
4. C. Harris and M. Stephens, A combined corner and edge detector. in *Alvey vision conference*, vol. 15. Citeseer, 1988, p. 50.
5. P.-L. Shui and W.-C. Zhang, Corner detection and classification using anisotropic directional derivative representations, *Image Processing, IEEE Transactions on*, vol. 22, no. 8, pp. 3204 3218, 2013.
6. K. Mikolajczyk and C. Schmid, Scale and affine invariant interest point detectors, *International journal of computer vision*, vol. 60, no. 1, pp. 6386, 2004.
7. S. M. Smith and J. M. Brady, Susan a new approach to low level image processing, *International journal of computer vision*, vol. 23, no. 1, pp. 4578, 1997.
8. E. Rosten and T. Drummond, Machine learning for high-speed corner detection, in *Computer VisionECCV 2006*. Springer, 2006, pp. 430 443.
9. E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, Adaptive and generic corner detection based on the accelerated segment test, in *Computer Vision ECCV 2010*. Springer, 2010, pp. 183196.

10. E. Rosten, R. Porter, and T. Drummond, Faster and better: A machine learning approach to corner detection, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 105119, 2010.
11. J. Lan and M. Zhang, Fast and robust corner detector based on doublecircle mask, *Optical Engineering*, vol. 49, no. 12, pp. 127204127 204, 2010.
12. L. Breiman, Random forests, *Machine learning*, vol. 45, no. 1, pp. 532, 2001.
13. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
14. L. Breiman, Bagging predictors, *Machine learning*, vol. 24, no. 2, pp. 123140, 1996.
15. M. Trajkovic and M. Hedley, Fast corner detection, *Image and vision computing*, vol. 16, no. 2, pp. 7587, 1998.
16. V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807814.

Symbolic Representation and Classification of Logos

D.S. Guru and N. Vinay Kumar

Abstract In this paper, a model for classification of logos based on symbolic representation of features is presented. The proposed model makes use of global features of logo images such as color, texture, and shape features for classification. The logo images are broadly classified into three different classes, viz., logo image containing only text, an image with only symbol, and an image with both text and a symbol. In each class, the similar looking logo images are clustered using K-means clustering algorithm. The intra-cluster variations present in each cluster corresponding to each class are then preserved using symbolic interval data. Thus referenced logo images are represented in the form of interval data. A sample logo image is then classified using suitable symbolic classifier. For experimentation purpose, relatively large amount of color logo images is created consisting of 5044 logo images. The classification results are validated with the help of accuracy, precision, recall, F-measure, and time. To check the efficacy of the proposed model, the comparative analyses are given against the other models. The results show that the proposed model outperforms the other models with respect to time and F-measure.

Keywords Appearance-based features • Clustering • Symbolic representation • Symbolic classification • Logo image classification

1 Introduction

With the rapid development of multimedia information technology, the amount of image data available on the internet is very huge and it is increasing exponentially. Handling of such a huge quantity of image data has become a more challenging and

D.S. Guru · N. Vinay Kumar (✉)
Department of Studies in Computer Science, University of Mysore,
Manasagangotri, Mysuru 570 006, India
e-mail: vinaykumar.natraj@gmail.com

D.S. Guru
e-mail: dsg@compsci.uni-mysore.ac.in

at the same time it is an interesting research problem. Nowadays, to handle such image data, there are lot many tools available on the internet such as ArcGIS, Google, Yahoo, Bing, etc. Currently, those tools perform classification, detection, and retrieval of images based on their characteristics. In this work, we consider logos which come under image category for the purpose of classification. A logo is a symbol which symbolizes the functionalities of an organization.

Once a logo is designed for any organization, it needs to be tested for its originality and uniqueness. If not, many intruders can design logos which look very similar to the existing logos and may change the goodness of the respective organization. To avoid such trade infringement or duplication, a system to test a newly designed logo for its originality is required. To test for the originality, the system has to verify the newly designed logo by comparing with the existing logos. Since the number of logos available for comparison is very large, either a quick approach for comparison or any other alternative need to be investigated. One such alternative is to identify the class of logos to which the newly designed logo belongs and then verifying it by comparing against only those logos of the corresponding class. Thus, the process of classification reduces the search space of a logo verification system to a greater extent. With this motivation, we address a problem related to classification of logos based on their appearance.

1.1 Related Works

In literature, we can find couple of works carried out on logo classification. Especially, these works are mainly relied on black and white logo images.

In [1], an attempt toward classifying the logos of the University of Maryland (UMD) logo database is made. Here, the logo images are classified as either degraded logo images or non-degraded logo images. In [2], a logo classification system is proposed for classifying the logo images captured through mobile phone cameras with a limited set of images. In [3], a comparative analysis of invariant schemes for logo classification is presented. From the literature, it can be observed that, in most of the works, the classification of logos has been done only on logos present in the document images. Also, there is no work available for classification of color logo images. Keeping this in mind, we thought of classifying the color logos.

In this paper, an approach based on symbolic representation of features for classification of logo images is addressed. The logo images which represent the functionalities of organizations are categorized into three classes, viz., images fully text, images with fully symbols, and images containing both texts and symbols. Some of these color logo images are illustrated in Fig. 1. Among three classes, there exist samples with large intra-class variations as illustrated in Fig. 1. In Fig. 1, the both category logo images have several intra-class variations like the shape of text, the color of text and the texture of symbols. In the second category of logo images, the internal variations are, the color of the text and the shape of text. In the last

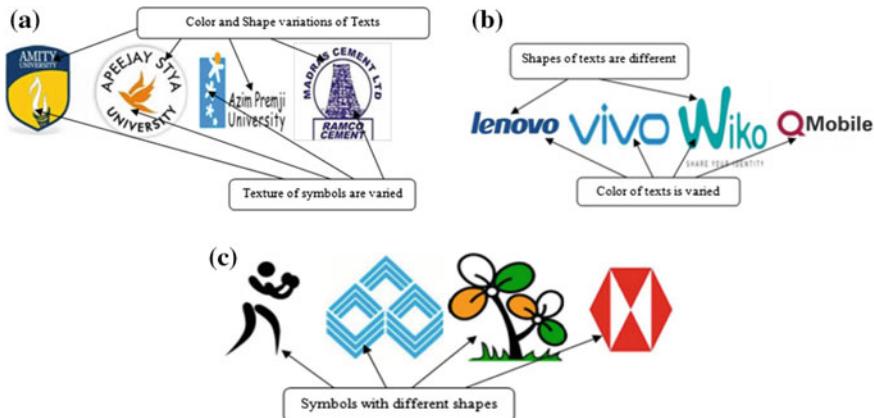


Fig. 1 Illustration of color logo images with several intra-class variations exists with respect to **a** both class, **b** only text class and **c** only symbol class

category, i.e., symbols, the variations are with respect to the shapes of symbols measured at different orientations. The intra-class variations may lead to the misclassification of a logo image as a member of given three classes. To avoid such misclassification, methods which can take care of preserving the intra-class variations present in the logo images are needed. One such method is symbolic data analysis. In symbolic data analysis, the interval representation of data can take care of intra-class variations of features which help in classification [4], clustering [5], and regression [6] of data.

Due to the presence of very large logo image samples in every class of the dataset used, it is better to group the similar looking logo images in every class. This results with the class containing logos with lesser variations within the class. It further helps in representing them in symbolic interval form.

The paper mainly concentrates on only classification. In the literature, there exist some symbolic classifiers [7, 8], but the limitation of these classifiers is present at classification stage. As these classifiers make use of interval representation of both the reference samples and query samples while classification. But in our case, the reference logo images are represented as interval data and the query images are represented as a conventional crisp type data [9]. So to handle such data, a suitable symbolic classifier [9] which can take care of the above-said scenario is used for classification.

Finally, the proposed system is compared with the other models, viz., a model which make use of only clustering and a model which neither uses clustering nor uses symbolic representation. Our system has outperformed with the former and latter models in terms of validity measures and time, respectively.

The rest of the paper is organized as follows. In Sect. 2, the details of the proposed logo classification system are explained. The experimentation setup and the detailed results with reasoning are presented in Sect. 3. Further, comparisons to

the proposed model against the conventional models are given. Finally, the conclusion remarks are given in Sect. 4.

2 Proposed Model

Different steps involved in the proposed logo classification model are shown in Fig. 2. Our model is based on classifying a color logo as either a logo with only text or a logo with only symbols or a logo with both text and symbol. The different stages of the proposed model are explained in the following subsections.

2.1 Preprocessing

In this stage, we recommend two different preprocessing tasks, namely, image resizing and gray-scale conversion. Initially, we resize all the logo images of dimension $M \times N$ into $m \times n$ to maintain the uniformity in the dimensions of the logo images, where the former dimension relates with original logo images and latter dimension relates with resized logo images. Then, we convert the RGB logo images into gray-scale images. The conversion helps in extracting the texture and shape features from gray-scale images [10].

2.2 Feature Extraction and Fusion

In this work, we recommend three different appearance-based (global) features, namely, color, texture, and shape features for extraction from an input color logo image. These features are recommended, as these features are invariant to

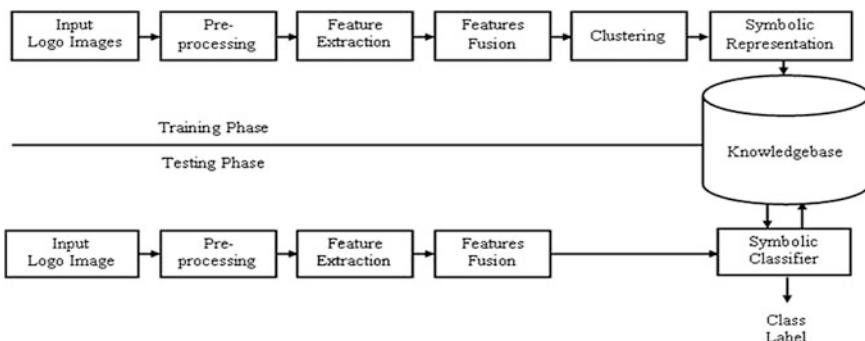


Fig. 2 Architecture of the proposed model

geometrical transformations [10]. Color, texture, and shape features are extracted for all color logo images as explained in [11].

For color feature extraction, the resized RGB logo image is divided into eight coarse partitions or blocks and determines the mean and percentage of individual block with respect to each color component (red/blue/green). For texture and shape feature extraction, the gray-scale logo images are used because these two features do not depend on the information of color properties of an image [10]. Initially for texture feature extraction, an image is processed using steerable Gaussian filter decomposition with four different orientations (0° , -45° , $+45^\circ$, 90°), and then the mean and standard deviation at each decomposition is computed. For shape features, Zernike moments shape descriptor is used in two different orientations (0° and 90°) [11]. Further, these features are fused together to discriminate the logo images.

2.3 Clustering

In this step, the logo images which look similar are grouped together with respect to each class. For clustering the similar logo images, the partitional clustering approach is adopted. The partitional clustering approach is very simple as it makes use of feature matrix for clustering instead of proximity matrix as in Hierarchical clustering [12]. Hence, in concern with the preprocessing efficiency of the proposed model, partitional clustering is chosen over hierarchical clustering.

The K-means clustering algorithm [12] is used to cluster the similar logo images. The value of K is varied to group the similar looking color logos within each class. The clustering results play a significant role in our proposed model in deciding the goodness of the proposed model.

2.4 Symbolic Logo Representation

In this step, the clustered logo images within each class are further represented in the form interval valued data, a representation which preserves the intra-class (cluster) variations [9]. The details of interval representation are given below.

Consider a sample $X_i = \{x^1, x^2, \dots, x^d\}$ ($X_i \in R^d$) belongs to i th class containing d features. Let there be totally N number of samples from m number of classes. If clustering is applied on the samples belong to i th class, the number of clusters obtained from each class is k . Then, the total number of samples present in j th cluster belongs to class i be n_j^i ($j = 1, 2, \dots, k$ and $i = 1, 2, \dots, m$). To preserve the intra-class variations present in each cluster, the mean-standard deviation interval representation is recommended. As it preserves the internal variations present within the samples of each cluster [9], the mean and standard deviation computed for the clustered samples are given by (1) and (2):

$$\mu_{j_i}^l = \frac{1}{n_j^i} \sum_{h=1}^{n_j^i} x_h^l \quad (1)$$

$$\sigma_{j_i}^l = \sqrt{\frac{1}{(n_j^i - 1)} \sum_{h=1}^{n_j^i} (x_h^l - \mu_{j_i}^l)^2} \quad (2)$$

where $\mu_{j_i}^l$ and $\sigma_{j_i}^l$ are the mean and standard deviation value of l th feature which belongs to j th cluster corresponding to class i , respectively.

Further, the mean and standard deviation is computed for all features belonging to j th cluster corresponding to i th class.

After computing the mean and standard deviation for each cluster belonging to a respective class, these two moments are joined together to form an interval cluster representative which belongs to each class. The difference between mean and standard deviation represents lower limit of an interval and the sum of mean and standard deviation represents the upper limit of an interval. Finally, k number of such cluster interval representatives is obtained from each class. Cluster representative is given by

$$\begin{aligned} CR_j^i &= \left\{ \left[(\mu_{j_i}^1 - \sigma_{j_i}^1), (\mu_{j_i}^1 + \sigma_{j_i}^1) \right], \left[(\mu_{j_i}^2 - \sigma_{j_i}^2), (\mu_{j_i}^2 + \sigma_{j_i}^2) \right], \dots, \left[(\mu_{j_i}^d - \sigma_{j_i}^d), (\mu_{j_i}^d + \sigma_{j_i}^d) \right] \right\} \\ CR_j^i &= \{ [f_1^-, f_1^+], [f_2^-, f_2^+], \dots, [f_d^-, f_d^+] \} \end{aligned}$$

where, $f_l^- = \left\{ (\mu_{j_i}^l - \sigma_{j_i}^l) \right\}$ and $f_l^+ = \left\{ (\mu_{j_i}^l + \sigma_{j_i}^l) \right\}$

Finally, we arrived at an interval feature matrix of dimension $(k * m)xd$, considered as a reference matrix while classification.

2.5 Logo Classification

To test the effectiveness of the proposed classification system, suitable symbolic classifier is needed. Here, we make use of a symbolic classifier proposed in [9] for classification of logo images. Here, the reference logo images are represented in the form of interval data as explained in the earlier sections. Let us consider a test sample $S_q = \{s^1, s^2, \dots, s^d\}$ contains d number of features. The test sample S_q needs to be classified as a member of any one of the three classes. Hence, the similarity is computed between a test sample and all reference samples. Here, for every test sample the similarity is computed at feature level. So, the similarity between a test crisp (single valued) feature and a reference interval feature can be computed as follows: The similarity value is 1, if the crisp value lies between the upper limit and lower limit of an interval feature, else 0. Similarly, the similarity between S_q and all remaining samples is computed. If S_q is said to be a member of any one of the three

classes, then the value of acceptance count $AC_q^{j_i}$ is very high with respect to the reference sample (cluster representative) that belongs to a particular class.

The acceptance count $AC_q^{j_i}$ for a test sample corresponding to j th cluster of i th class is given by

$$AC_q^{j_i} = \sum_{l=1}^d \text{Sim}(S_q, CR_j^l) \quad (3)$$

where $\text{Sim}(S_q, CR_j^l) = \begin{cases} 1 & \text{if } s^l \geq f_l^- \text{ and } s^l \leq f_l^+ \text{ and } i = 1, 2, \dots, m; j = 1, 2, \dots, k; \text{ and } l = 1, 2, \dots, d \\ 0 & \text{otherwise} \end{cases}$

3 Experimentation

3.1 Dataset

For experimentation, we have created our own dataset named “UoMLogo database” consisting of 5044 color logo images. The “UoMLogo Database” mainly consists of color logo images of different universities, brands, sports, banks, insurance, cars, industries, etc. which are collected from the internet. This dataset mainly categorized into three classes, BOTH logo image (a combination of TEXT and SYMBOL), TEXT logo image, SYMBOL image. Within class, there exist ten different subclasses. Figure 3 shows the sample images of the UoMLogo dataset. The complete details of the dataset are found in [11].

3.2 Experimental Setup

In preprocessing step, for the sake of simplicity and uniformity in extracting the features from a color logo image, we have resized every image into 200×200 dimensions. To extract texture and shape features, the color logo images are converted into gray-scale images, as these two features are independent of color

Fig. 3 Sample logo images of UoMLogo dataset



features. In feature extraction stage, three different features, mainly color, texture, and shape features, are extracted. These features are extracted from logo images as discussed in Sect. 2.2. From feature extraction, we arrived at 48, 8, and 4 of color, texture, and shape features, respectively. Further, these features are normalized and fused to get 60 features, which represent a logo image sample.

After feature extraction and fusion, these features of logo images are represented in the form of feature matrix, where the rows and the columns of a matrix represent the samples and features, respectively. Further, the samples of a feature matrix are divided into training samples and testing samples. Training samples are used for clustering and symbolic representation. Testing samples are used for testing the classification system.

During symbolic representation, the training samples are clustered using K-means algorithm. The values of K varied from 2 to 10, depending on the formation of cluster samples. The upper bound of the clusters is limited to 10, because the clustering algorithm fails to cluster the similar logo images beyond 10. Further, the clustered samples are represented in the form of interval data as explained in Sect. 2.5. The total number of logo images present in our database is 5044 with three different classes. So, the total number of samples present in a reference matrix after symbolic representation is 6 ($2 \times 3 = 6$; 2: number of cluster interval representative; 3: No. of Classes), 9, 12, 15, 18, 21, 24, 27, and 30 samples for clusters varied from 2 to 10, respectively. For classification, a symbolic classifier is adopted.

In our proposed classification system, the dataset is divided randomly into training and testing. Seven sets of experiments have been conducted under varying number of training set images as 20, 30, 40, 50, 60, 70, and 80 %. At each training stage, the logo images are represented in the form of interval data with respect to the varied number of clusters from 2 to 10. While at testing stage, the system uses remaining 80 %, 70 %, 60 %, 50 %, 40 %, 30 %, and 20 % of logo images, respectively, for classifying them as any one of the three classes. The experimentation in testing is repeated for 20 different trials. During testing, the classification results are presented by the confusion matrix. The performance of the classification system is evaluated using classification accuracy, precision, recall, and F-measure computed from the confusion matrix [13].

3.3 Experimental Results

The performance of the proposed classification system is evaluated not only based on classification accuracy, precision, recall, and F-measure computed from the confusion matrix but also it is done with respect to time.

Let us consider a confusion matrix CM_{ij} , generated during classification of color logo images at some testing stage. From this confusion matrix, the accuracy, the precision, the recall, and the F-measure are all computed to measure the efficacy of the proposed logo image classification system. The overall accuracy of a system is given by

$$\text{Accuracy} = \frac{\text{No. of Correctly classified Samples}}{\text{Total number of Samples}} * 100 \quad (4)$$

The precision and recall can be computed in two ways. Initially, they are computed with respect to each class and later with respect to overall classification system. The class-wise precision and class-wise recall computed from the confusion matrix are given in Eqs. (5) and (6), respectively:

$$P_i = \frac{\text{No. of Correctly classified Samples}}{\text{No of Samples classified as a member of a class}} * 100 \quad (5)$$

$$R_i = \frac{\text{No. of Correctly classified Samples}}{\text{Expected number of Samples to be classified as a member of a class}} * 100 \quad (6)$$

where $i = 1, 2, \dots, m$; $m = \text{No. of classes}$

The system precision and system recall computed from the class-wise precision and class-wise recall are given by

$$\text{Precision} = \frac{\sum_{i=1}^m P_i}{m} \quad (7)$$

$$\text{Recall} = \frac{\sum_{i=1}^m R_i}{m} \quad (8)$$

The F-measure computed from the precision and recall is given by

$$F - \text{Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} * 100 \quad (9)$$

The average time is computed at a particular testing stage while classifying a test sample as a member of any one of the three classes. It is done using a MATLAB built in command *tic—toc*.

The classification results are thus obtained for different training and testing percentages of samples under varied clusters from 2 to 10. These results are measured in terms of accuracy (minimum, maximum, and average), precision (minimum, maximum, and average), recall (minimum, maximum, and average), and F-measure (minimum, maximum, and average). The minimum, maximum, and average of respective results are obtained due to the 20 trials of experiments performed on training samples. Here, precision and recall are computed from the results obtained from the class-wise precision and class-wise recall, respectively.

The results obtained from cluster 2 to 10 under varied training and testing samples are consolidated and tabulated in Table 1. These results are judged based on the best average F-measure obtained under respective training and testing

Table 1 Best results obtained from all clusters under varied training and testing percentage of samples

Train test %	Accuracy			Precision			Recall			F-measure			Cluster#
	Min	Max	Avg										
20–80	60.20	74.21	68.56	52.77	81.95	65.35	53.85	58.48	56.07	54.78	65.39	60.14	10
30–70	60.90	74.07	68.28	55.55	77.19	64.06	55.07	58.91	56.92	56.33	64.93	60.13	9
40–60	61.92	74.38	70.47	56.78	80.08	65.74	54.67	59.16	56.65	56.21	65.77	60.71	5
50–50	60.97	73.70	69.81	59.05	71.22	65.06	55.79	60.28	57.45	57.88	63.14	60.95	10
60–40	61.71	73.61	70.75	58.08	69.76	65.11	56.16	59.77	57.52	57.10	63.45	61.03	5
70–30	66.27	73.94	71.73	58.70	72.94	66.89	55.38	59.29	57.27	58.09	64.20	61.63	4
80–20	65.77	73.31	70.23	59.27	72.21	64.55	56.41	61.22	58.34	58.70	63.91	61.24	8
Best	66.27	73.94	71.73	58.70	72.94	66.89	55.38	59.29	57.27	58.09	64.20	61.63	4

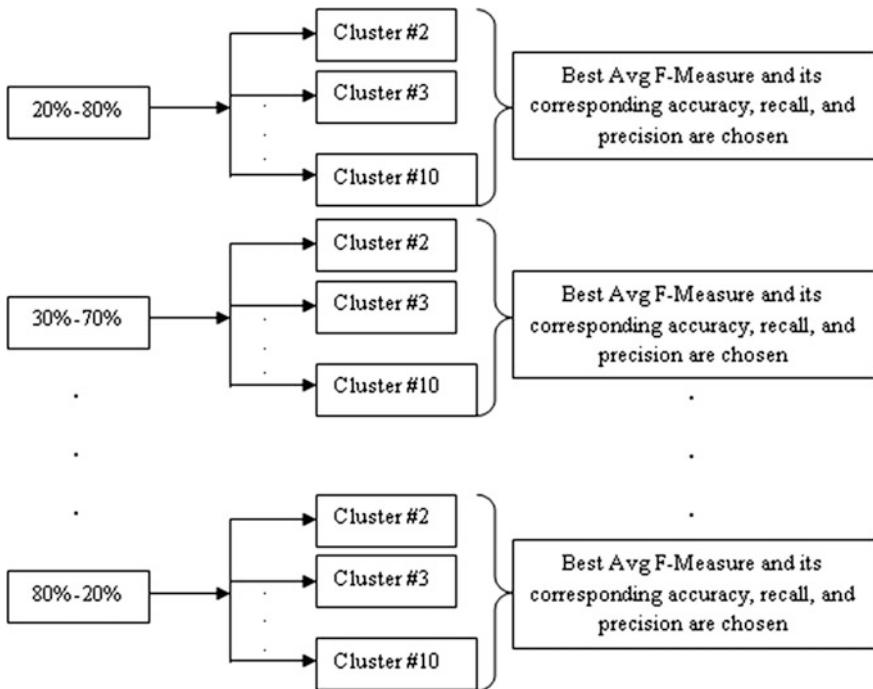


Fig. 4 Selection procedure followed in choosing the best results obtained from all clusters under varied training and testing percentage of samples

percentage among all clusters (varied from 2 to 10). This has been clearly shown in Fig. 4.

From the above table, it is very clear that the best classification results are obtained only if the samples within each class are clustered into four groups. This shows that our model is very robust in choosing the number of clusters for clustering samples within each class. The last row in the above table reveals the best results obtained for respective cluster and for respective training and testing percentage of samples.

3.4 Comparative Analyses

The proposed symbolic logo image classification model is compared against the two different models in classifying the same color logo image database. As we know our model makes use of clustering before representing the samples in symbolic representation, we thought of comparing our model against the other models: (a) a model which never preserves intra-class variations and never groups the similar logo images, and (b) a model which makes use of clustering for grouping

similar logo images within each class but does not preserve intra-class variations (i.e., this model never makes use of symbolic representation; it only uses with conventional cluster mean representation). This comparison helps us to test robustness of the proposed model in terms of F-measure and time.

The experimental setup for the other two models is followed as given below:

For Conventional Model (Model-1):

Totally, 60 features are considered for representation. K-NN classifier ($K = 1$) is used for classification. The training and testing percentage of samples are divided and varied from 20 to 80 % (in steps of 10 % at a time). The experiments are repeated for 20 trials and results are noted based on the minimum, maximum, and average values obtained from 20 trials.

For Conventional+Clustering (Co+Cl) Model (Model-2):

Totally, 60 features are considered for representation. Then K-means partitional clustering algorithm is applied to group similar logo images within each class (K : varied from 2 to 10). K-NN classifier ($K = 1$) is used for classification. The training and testing percentage of samples are divided and varied from 20 to 80 % (in steps of 10 % at a time). The experiments are repeated for 20 trials and results are noted based on the minimum, maximum, and average values obtained from 20 trials.

The classification results for the above-said models are validated based on the confusion matrix obtained during the classification. The same validity measures used in our proposed model (accuracy, precision, recall, F-measure, and time) are used for validating these two models. With respect to model-1, the best results are obtained for 80–20 % of training and testing samples and are shown in Table 2. With respect to model-2, the best results are obtained when the similar logo samples are grouped into two clusters. The results are tabulated in Table 3. Similarly, the results of the proposed model are tabulated in Table 4.

The results shown in the Tables 2, 3, and 4 are for the respective models in classifying the color logo images. From Tables 2 and 3, it is very clear that the model-1 is superior model compared to model-2 in terms of average F-measure and remaining other measures. But in terms of efficiency in classification, model-2

Table 2 Best results obtained from different training and testing percentages of samples based on Avg F-measure

	Min	Max	Avg	Train-test %
Accuracy	69.31	73.51	71.48	80–20
Precision	59.40	66.72	62.36	
Recall	58.46	65.12	60.94	
F-Measure	59.09	65.43	61.64	

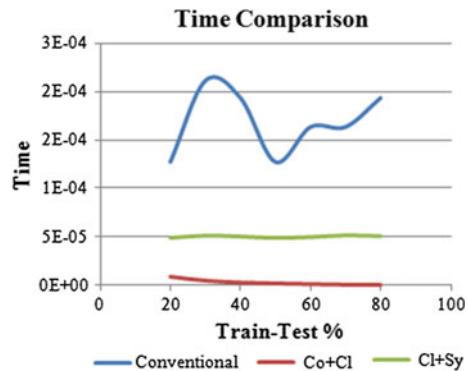
Table 3 Best results obtained for conventional +clustering classification model

	Min	Max	Avg	Cluster #	Train test %
Accuracy	53.08	60.02	56.97	2	80–20
Precision	42.98	50.75	47.57		
Recall	44.08	53.63	49.58		
F-Measure	43.53	51.88	48.55		

Table 4 Best results obtained for symbolic +clustering classification model

	Min	Max	Avg	Cluster #	Train test %
Accuracy	66.27	73.94	71.73	4	70–30
Precision	58.70	72.94	66.89		
Recall	55.38	59.29	57.27		
F-Measure	58.09	64.20	61.63		

Fig. 5 Comparison of time utilization of the proposed model vs conventional model vs conventional + clustering model in classifying the logo images



outperforms model-1 as shown in Fig. 5. With respect to model-2 and proposed, it is clearly observed from Tables 3 and 4 that the proposed model is far superior than the model-2 in terms of average F-measure (and other remaining measures), and also in terms of efficiency, our model is very stable model irrespective of training and testing percentage of samples. Similarly, if we compare model-1 with our own model, our model is somehow equivalent to model-1 with an epsilon difference in average F-measure. But, if we consider with respect to efficiency in classification, our model outperforms model-1 in terms of stability and efficiency in classification. Hence, the proposed model suits better for classifying the huge color logo image database.

For better visualization on classification of color images, the confusion matrices obtained for the best results shown in Tables 3 and 4 are given in Tables 5 and 6, respectively; and also, the misclassifications occurred while classifying the logo images for model-2 and proposed model are given in Fig. 6a, b, respectively.

Table 5 Confusion matrix obtained for model-2 (Conventional+clustering method for cluster #2 (80–20 %))

	Both	Text	Symbol
Both (634)	399	137	98
Text (249)	87	118	44
Symbol (125)	63	23	39



Fig. 6 Negative classification. **a** Results obtained for conventional+clustering classification model; **b** Results obtained for symbolic+clustering classification model

Table 6 Confusion matrix obtained for the proposed model (Symbolic+clustering method for cluster #4 (70–30 %))

	Both	Text	Symbol
Both (951)	818	86	47
Text (419)	154	194	25
Symbol (188)	93	24	71

From the above discussion, it is very clear that the proposed logo image classification system is very stable and efficient compared to other models in classifying the color logo images.

4 Conclusion

In this paper an approach based on symbolic interval representation in classifying the color logo images into the pre-defined three classes is proposed. In classifying a logo image, the global characteristics of logo images are extracted. Then, the partitional clustering algorithm is adopted for clustering the similar logo images within each class. Later, the symbolic interval representation is given for clusters belonging to the corresponding classes. Further, a symbolic classifier is used for logo image classification. The effectiveness of the proposed classification system is validated through well-known measures like accuracy, precision, recall, F-measure and also with respect to time. Finally, the paper concludes with an understanding that the better classification results are obtained only for the symbolic interval representation model compared to other models.

Acknowledgments The author N Vinay Kumar would like to thank Department of Science and Technology, India, for the financial support through INSPIRE Fellowship.

References

1. Jan Neumann, Hanan Samet, Aya Soffer: Integration of local and global shape analysis for logo classification. Pattern Recognition letters 23, pp. 1449–1457, (2002)
2. Shu-kuo sun and Zen Chen: Logo recognition by mobile phone cameras. Journal of Info. Sci. and Engg. 27, pp. 545–559, (2011)
3. Yasser Arafat, S, Muhammad Saleem, Afaq Hussain, S.: Comparative Analysis of Invariant Schemes for Logo Classification. Int. Conf. on Emerging Techn., pp. 256–261, (2009)
4. Lynne Billard and Edwin Diday, Symbolic Data Analysis- Conceptual Statistics and Data Mining, Wiley Publications, (2006)
5. Antonio Giusti and Laura Grassini, Cluster analysis of census data using the symbolic data approach, Adv. Data Anal Classification 2, pp. 163–176, (2008)
6. Roberta A.A. Fagundes, Renata M.C.R. de Souza, Francisco José A. Cysneiros, Robust regression with application to symbolic interval data, Engineering Applications of Artificial Intelligence 26, pp. 564–573, (2013)
7. Alberto Pereira de Barros, Francisco de Assis Teñorio de Carvalho and Euffasio de Andrade Lima Neto, A Pattern Classifier for Interval-valued Data Based on Multinomial Logistic Regression Model, IEEE Int. Conf. SMC, pp. 541–546, (2012)
8. Antdnio Pedro Duarte Silva and Paula Brito, Linear Discriminant Analysis for Interval Data, Computational Statistics 21, pp. 289–308, (2006)
9. D.S. Guru and H.N. Prakash, Online Signature Verification and Recognition: An Approach Based on Symbolic Representation, IEEE Trans. on PAMI, vol. 31, No. 6, pp. 1059–1073, (2009)
10. Xiang-Yang Wang, Yong-Jian Yu, Hong-Ying Yang: An effective image retrieval scheme using color, texture and shape features. Comp. Stds. and Interfaces 33, pp. 59–68, (2011)
11. N Vinay Kumar, Pratheek, V Vijaya Kantha, K. N. Govindaraju, and D S Guru, Features fusion for classification of logos, Elsevier Procedia Computer Science, vol. 85, pp. 370–379, (2016).
12. Anil K. Jain and Richard C. Dubes. Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, (1988)
13. Powers, David M W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, J. of Mach. Learning Tech. 2 (1), pp. 37–63, (2011)

A Hybrid Method Based CT Image Denoising Using Nonsubsampled Contourlet and Curvelet Transforms

Manoj Diwakar and Manoj Kumar

Abstract Computed tomography (CT) is one of the most widespread radio-logical tools for diagnosis purpose. To achieve good quality of CT images with low radiation dose has drawn a lot of attention to researchers. Hence, post-processing of CT images has become a major concern in medical image processing. This paper presents a novel edge-preserving image denoising scheme where noisy CT images are denoised using nonsubsampled contourlet transform (NSCT) and curvelet transform separately. By estimating variance difference on both denoised images, final denoised CT image has been achieved using a variation-based weighted aggregation. The proposed scheme is compared with existing methods and it is observed that the performance of proposed method is superior to existing methods in terms of visual quality, image quality index (IQI), and peak signal-to-noise ratio (PSNR).

Keywords Image denoising · Wavelet transform · Curvelet transform · Nonsubsampled contourlet transform · Thresholding

1 Introduction

In medical science, computed tomography (CT) is one of the important tools, which helps to provide the view of the human body's internal structure in the form of digital images for diagnosis purpose. In computed tomography, X-rays are projected over the human body where soft and hard tissues are observed and other side, a detector is used to collect the observed data (raw data). Using Radon transform, these raw data are further mathematically computed to reconstruct the CT images. X-ray radiation dose beyond a certain level could increase the risk of cancer [1, 2]. Thus, it

M. Diwakar (✉) · M. Kumar

Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India
e-mail: manoj.diwakar@gmail.com

M. Kumar
e-mail: mkjnuitr@gmail.com

is an important to give lower amount of radiation during CT image reconstruction. However, reducing the radiation dose may increase the noise level of CT reconstructed images which may not be usable for diagnosis. Because of acquisition, transmission, and mathematical computation, the reconstructed CT images may be degraded in terms of noise. To surmount noisy problem, various methods have been investigated for noise suppression in CT images where wavelet transform-based denoising has achieved good results over the last few decades. In wavelet transform domain, wavelet coefficients are modified using various thresholding methods. For the modification of wavelet coefficients, numerous strategies have been proposed to improve denoising performance. These strategies can be broadly categorized into two categories: (i) intra-scale dependency-based denoising and (ii) inter-scale dependency-based denoising. In intra-scale dependency-based denoising [3–9], the wavelet coefficients are modified with in same scale. SureShrink [3], BayesShrink [4], and VisuShrink [5] are the popular methods of intra-scale dependency-based denoising, where SureShrink and BayesShrink provide better performance than the VisuShrink. The inter-scale dependency defines that if parent coefficients are large, then its child coefficients are also large. With this consideration, the wavelet coefficients are modified across the scale using denoising methods such as bivariate shrinkage function [10, 11]. The dependency between parent and child coefficients is helpful for better denoising. Wavelet transform-based thresholding shows remarkable results and outperforms those derived from the independent assumption. To improve the problem of shift invariance, aliasing, and poor directionality in traditional discrete wavelet transform, many other directional transforms have also been used such as tetrolet, dual-tree complex wavelet, curvelet, contourlet, and directionlet [12–15]. From various directional transforms, curvelet and nonsubsampled contourlet transforms have received a great deal of edge preservation and noise reduction because of it offers directionality and shift invariance with low computational complexity [16, 17].

With different shrinkage rules and different transforms, it cannot be surely predicted that which one is better in terms of preserving edge, local features, and noise reduction specially in case of medical images. With this consideration, we propose a new hybrid method for reduction of pixel noise with structure preserving using nonsubsampled contourlet and curvelet transforms which combines the advantages of hard and soft thresholdings. The paper is structured as follows. In Sect. 2, we describe a brief overview of nonsubsampled contourlet transform for image denoising. In Sect. 3, we give a brief introduction of curvelet transform-based image denoising. The proposed method for CT image denoising is presented in detail in Sect. 4. In Sect. 5, experimental results and comparison with other standard denoising methods are discussed. Concluding remarks are summarized in Sect. 6.

2 Nonsubsampled Contourlet Transform

The contourlet transform helps to provide multi-scale decomposition and directional decomposition using Laplacian pyramid and direction filter bank. Enhance version of contourlet transform is nonsubsampled contourlet transform (NSCT) which helps to improve the frequency aliasing problem of contourlet transform. NSCT is based on the contourlet conception which helps to avoid the sampling steps during decomposition and reconstruction stages. The structure of NSCT contains two filter banks: non-subsampled pyramid and nonsubsampled directional filter banks. The combination of both filter banks helps to provide the features of shift invariance, multi-resolution, and multi-dimensionality for image presentation. The design and reconstruction of NSCT filters (analysis and synthesis) are easy to be realized and help to provide better collection of frequency and more regularity [15].

2.1 NSCT-Based Thresholding

Due to the non-orthogonal property of NSCT, the noise variance contains different values for each direction on respective sub-bands [17]. Therefore, the noise is estimated independently for their respective sub-bands. In this article, the noise from each NSCT coefficient is estimated for each level and direction. Further, denoising process is performed using a threshold value. For respective level L and direction D , a threshold value can be estimated as follows:

$$\lambda_{L,D} = \frac{\sigma_{\eta_{L,D}}^2}{\sigma_{W_{L,D}}} \quad (1)$$

The variance σ_W^2 of clean (noiseless) image can be obtained as

$$\sigma_{W_{L,D}} = \max(\sigma_Y^2 - \sigma_\eta^2, 0) \quad (2)$$

where $\sigma_Y^2 = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N Y_{i,j}^2$, $Y_{i,j}$ is the NSCT low frequency component and MN is the size of respective sub-band.

The noise variance (σ_η^2) using robust median estimation approach [11] can be obtained as

$$\sigma_{\eta_{L,D}}^2 = \left[\frac{\text{median}(|Y(i,j)|)}{0.6745} \right]^2 \quad (3)$$

To apply thresholding in NSCT domain, soft thresholding function is used which is defined as

$$R^{NSCT} := \begin{cases} \text{sign}(Y)(|Y| - \lambda_{L,D}), & |Y| > \lambda \\ 0, & \text{Otherwise} \end{cases} \quad (4)$$

3 Curvelet Transform

Curvelet transform is a multi-scale transform followed by ridgelet transform which provides several features such as multi-directional and multi-resolution [16]. The point discontinuity through Fourier transform is not well preserved. This problem was recovered by wavelet transform. But wavelet transform was failed to handle the curve discontinuity. To handle this, curvelet transform performs well using small number of curvelet coefficients. The curvelet transform is based on the combination of ridgelet and wavelet transform. Here wavelet transform is helpful for spatial partitioning where each scale of wavelet transform is divided into number of blocks. Further, these blocks are processed using ridgelet transform to obtain the curvelet coefficients [12]. Therefore, curvelet transform is considered as a localized ridgelet transform which gives more sharp edges such as curves. To gain the fine scalability, it also follows the properties of scalable rule.

Discrete curvelet transform has four major steps which can be defined over the function $f(x_1, x_2)$ as below:

(a) Sub-band decomposition: Here, an image f is decomposed into sub-bands using à trous method [12] as given below:

$$f \mapsto (P_0 f, \Delta_1 f, \Delta_2 f, \dots) \quad (5)$$

where $P_0 f$ is a low-pass filter bank and $\Delta_1, \Delta_2, \dots$ are the high-pass (bandpass) filters.

(b) Smooth Partitioning: To gain smooth partitioning, each sub-band is smoothly windowed into a set of dyadic squares:

$$h_Q = (w_Q \cdot \Delta_S f) \quad (6)$$

where w_Q is the collection of smooth windowing function to localized near dyadic squares Q_S , $Q \in Q_S$, and $S \geq 0$.

(c) Re-normalization: Here, re-normalization is performed over the each dyadic square to the unit scale of $[0,1] \times [0,1]$ using the following function:

$$g_Q = (T_Q^{-1} \cdot h_Q) \quad (7)$$

where $(T_Q f)(x_1, x_2) = 2^S f(2^S x_1 - k_1, 2^S x_2 - k_2)$ is the renormalizing operator.

(d) Ridgelet analysis: For each $a > 0, b \in \mathfrak{R}$, and $\theta \in [0, 2\pi]$) in a given function $f(x_1, x_2)$, discrete ridgelet coefficients can be obtained as

$$\mathfrak{R}_f(a, b, \theta) = \int f(x_1, x_2) \Psi_{a,b,\theta}(x_1, x_2) dx_1 dx_2 \quad (8)$$

where a is a scale parameter, b is a location parameter, θ is the orientation parameter, and Ψ is the wavelet function as $\Psi_{a,b,\theta}(x) = a^{1/2}\Psi(\frac{x_1 \cos \theta + x_2 \sin \theta - b}{a})$. Discrete ridgelet transform is obtained through radon transform, and can be analyzed as

$$\mathfrak{R}_f(a, b, \theta) = \int \mathcal{G}_f(t, \theta) a^{1/2} \Psi\left(\frac{t-b}{a}\right) dt \quad (9)$$

where \mathcal{G}_f represents the Radon transform over the variable t , which can be expressed as below:

$$\mathcal{G}_f(t, \theta) = \int f(x_1, x_2) \delta(x_1 \cos \theta + x_2 \sin \theta - t) dx_1 dx_2 \quad (10)$$

where δ is the Dirac distribution.

2D fast discrete curvelet transform (2D FDCT) is the version of curvelet transform which can be developed via USFFT or wrapping methods. Both methods are helpful for curvelet transformation to provide multi-scale and multi-directions. 2D FDCT with wrapping method is implemented on the basis of parabolic scaling function, anisotropic concept, tight framing, and wrapping.

3.1 Curvelet-Based Thresholding

To denoise the noisy CT images (Y), a shrinkage rule is performed using curvelet transform. Before performing shrinkage rule over the curvelet coefficients, the noise variance (σ_γ^2) of respective coefficients must be estimated. The noise estimation [12] can be described as below:

$$\sigma_\gamma^2 = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N Y_\gamma^{Curvelet} \cdot Y_\gamma^{Curvelet*}}{N^2}} \quad (11)$$

Using noise estimation, a hard thresholding function is performed over the curvelet coefficients, as given below:

$$R_\gamma^{Curvelet} := \begin{cases} Y_\gamma^{Curvelet}, & |Y_\gamma^{Curvelet}| \geq K\sigma\sigma_\gamma \\ 0, & \text{Otherwise} \end{cases} \quad (12)$$

where, $Y_\gamma^{Curvelet*}$ is a complex conjugate of $Y_\gamma^{Curvelet}$, σ is estimated noise variance of (Y), and K is the noise control parameter.

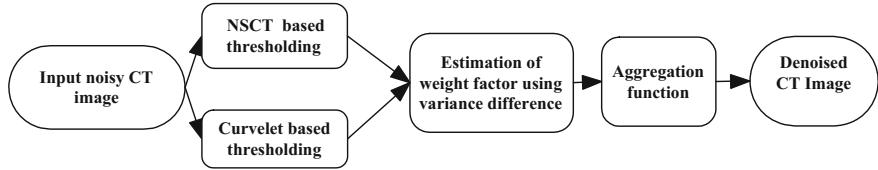


Fig. 1 Proposed scheme

4 Proposed Methodology

A scheme is proposed to denoise the noisy CT images by combining the advantages of nonsubsampled contourlet transform (NSCT)-based thresholding and curvelet transform-based thresholding. Here, noisy CT images are considered as Gaussian noise with zero mean and different variances.

Let the noisy image $Y(i,j)$ can be expressed as

$$Y(i,j) = W(i,j) + \eta(i,j) \quad (13)$$

where $W(i,j)$ is a noiseless image and $\eta(i,j)$ is an additive noise. The block diagram of the proposed scheme is shown in Fig. 1 and can be outlined with the following steps:

- Step 1 : Denoise the input noisy CT image by NSCT-based thresholding (R_1) as well as curvelet transform-based thresholding (R_2) separately.
- Step 2 : Estimate patch-wise variance (Var_{R_1} and Var_{R_2}) at each pixel of denoised images R_1 and R_2 .
- Step 3 : Estimate the weight value using variance difference of both R_1 and R_2 images:

$$\alpha = \sqrt{(Var_{R_1})^2 - (Var_{R_2})^2} \quad (14)$$

Normalize α in the range [0,1].

- Step 4 : Apply aggregation function using an adaptive weight factor α , as shown in the following equation:

$$R_3 = \alpha.R_1 + (1 - \alpha).R_2 \quad (15)$$

where R_3 is the denoised CT image.

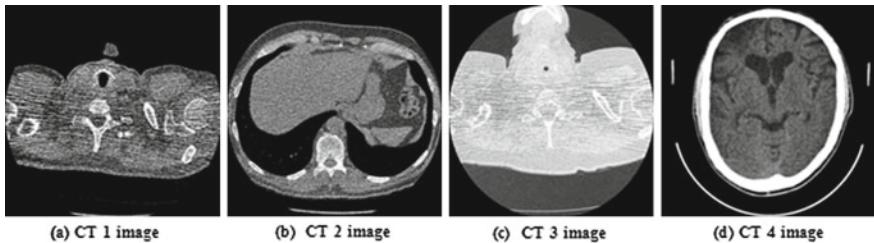


Fig. 2 Original CT image data set

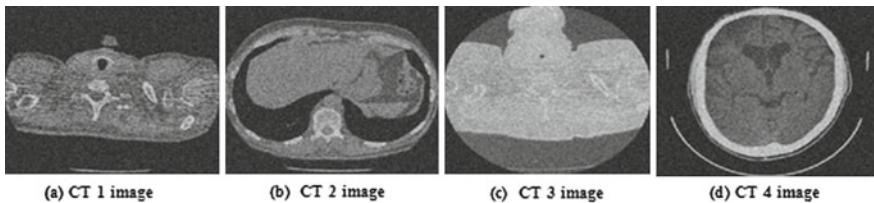


Fig. 3 Noisy CT image data set ($\sigma = 20$)

5 Results and Discussion

The results of proposed scheme are evaluated on the noisy CT images with the size 512×512 . The CT images as shown in Fig. 2a–c are acquired from the public access database (<http://www.via.cornell.edu/databases>), while Fig. 2d CT image is taken from a diagnosis center. The results are tested by applying additive Gaussian white noise at four different noise levels (σ): 10, 20, 30, and 40. Figure 2a–d shows the image named as CT1, CT2, CT3, and CT4, respectively. Figure 3a–d shows the noisy CT image data set where (σ) = 20.

In our experimental results, input noisy CT image is denoised using nonsubsampled contourlet transform (NSCT)-based thresholding and curvelet transform-based thresholding separately. To achieve maximum edge preserving and noise reduction, a weight value is estimated using variance difference of both denoised images R_1 and R_2 where the patch size is used as 3×3 . Using this weight factor, both images are fused using aggregated function and final denoised image has been achieved. The proposed scheme is also compared with some existing methods. The existing methods for comparison are adaptive dual-tree complex wavelet transform-based bivariate thresholding (DTCWTBT) [11], NSCT-based thresholding (NSCTBT) [17], and curvelet-based denoising (CBT) [12]. Figures 4a–d, 5a–d, 6a–d, and 7a–d show the results of DTCWTBT [11], NSCTBT [17], CBT [12], and proposed scheme, respectively. For CT images (1–4), image quality index (IQI) and peak signal-to-noise ratio (PSNR) are also measured for proposed method and existing methods. The IQI is used to observe the performance of the denoised images where the performance is measured in terms of correlation, luminance, and contrast distortions. PSNR also

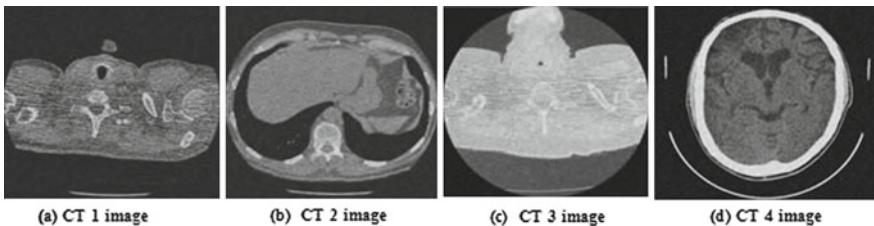


Fig. 4 Results of dual-tree complex wavelet transform-based bivariate thresholding (DTCWTBT)

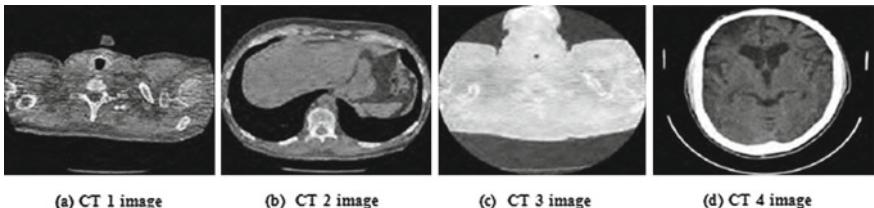


Fig. 5 Results of NSCT-based thresholding (NSCTBT)

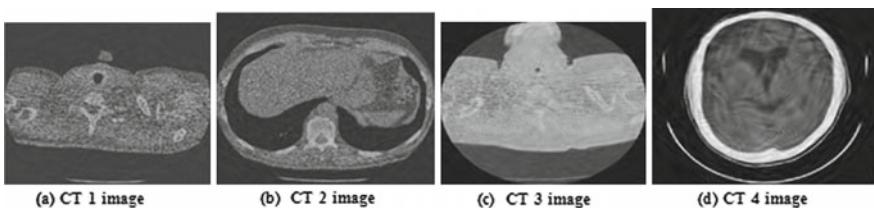


Fig. 6 Results of curvelet-based thresholding (CBT)

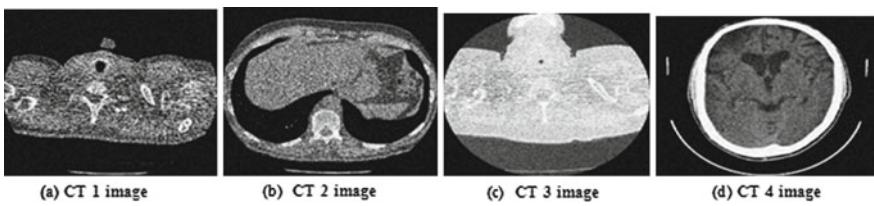


Fig. 7 Results of proposed method

helps to analyze the performance of denoised images in terms of signal-to-noise ratio. Table 1 indicates the performance of proposed and some existing methods in terms of IQI and PSNR. The IQI and PSNR values represent that proposed scheme gives better outcomes in most of the cases. The best values in Table 1 are indicated in bold.

Table 1 PSNR and IQI of CT denoised images

Image	σ	PSNR				IQI			
		10	20	30	40	10	20	30	40
CT1	DTCWTBT [11]	32.31	27.51	25.02	23.91	0.9972	0.9031	0.8871	0.8682
	NSCTBT [17]	31.11	28.14	26.38	24.91	0.9964	0.9181	0.8977	0.8715
	CBT [12]	32.12	28.34	26.56	24.11	0.9981	0.9161	0.8854	0.8556
	Proposed	32.98	28.62	27.33	24.93	0.9990	0.9197	0.8988	0.8801
CT2	DTCWTBT [11]	32.19	28.52	26.22	24.01	0.9985	0.9035	0.8969	0.8678
	NSCTBT [17]	31.73	29.12	26.34	23.61	0.9981	0.9187	0.8712	0.8634
	CBT [12]	32.12	28.52	26.55	24.67	0.9989	0.9067	0.8802	0.8512
	Proposed	32.88	29.11	27.02	24.93	0.9991	0.9087	0.8998	0.8871
CT3	DTCWTBT [11]	32.09	28.15	26.43	24.13	0.9942	0.8911	0.8791	0.8652
	NSCTBT [17]	32.23	29.11	26.33	23.91	0.9990	0.9081	0.8934	0.8745
	CBT [12]	32.86	28.05	26.31	24.02	0.9992	0.9062	0.8833	0.8575
	Proposed	32.01	29.43	27.12	24.96	0.9981	0.9189	0.8999	0.8879
CT4	DTCWTBT [11]	33.39	28.05	26.12	24.91	0.9982	0.9013	0.8919	0.8695
	NSCTBT [17]	33.82	29.21	26.43	23.01	0.9991	0.9068	0.8923	0.8781
	CBT [12]	32.86	28.05	26.63	24.80	0.9990	0.9086	0.8830	0.8571
	Proposed	33.01	29.42	27.81	24.98	0.9989	0.9189	0.8992	0.8883

6 Conclusions

In this paper, a post-processing approach is performed to reduce noise from the CT images using hybrid method using nonsubsampled contourlet and curvelet transforms. An aggregation concept is used in proposed methodology which helps to improve the signal-to-noise ratio. The outcomes of proposed scheme indicate that the visually results are good, especially in terms of edge preservation. The performance metrics (IQI and PSNR) also indicate that results are good in most of the cases. These experiments indicate that noise is suppressed effectively and clinically relevant details are well preserved.

References

1. Boone, J. M., Geraghty, E. M., Seibert, J. A., Wootton-Gorges, S. L.: Dose reduction in pediatric CT: a rational approach. *Journal of Radiology*, vol. 228(2), pp 352–360 (2003).
2. Kim, D., Ramani, S., Fessler, J. A.: Accelerating X-ray CT ordered subsets image reconstruction with Nesterov first-order methods. In *Proc. Fully Three-Dimensional Image Reconst. in Radiology and Nuclear Medicine*, pp. 22–25 (2013).
3. Chang, S. G., Yu, B., Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. on Image Process.*, vol. 9(9): 1532–1546 (2000).

4. Donoho, D. L., Johnstone, I. M.: Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, vol. 81(3): 425–455 (1994).
5. Donoho, D. L.: Denoising by soft thresholding. *IEEE Transactions on Information Theory*, vol. 41(3): 613–627 (1995).
6. Chang, S. G., Yu, B., Vetterli, M.: Spatially adaptive thresholding with context modeling for image denoising. *IEEE Transactions on Image Process.*, vol. 9(9): 1522–1531 (2000).
7. Donoho, D. L., Johnstone, I. M.: Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.*, vol. 90(432): 1200–1224 (1995).
8. Fathi, A., Naghsh-Nilchi, A. R.: Efficient image denoising method based on a new adaptive wavelet packet thresholding function. *IEEE Transactions on Image Processing*, vol. 21(9): 3981–3990 (2012).
9. Borsdorf, A., Raupach, R., Flohr, T., Hornegger, J.: Wavelet Based Noise Reduction in CT-Images Using Correlation Analysis. *IEEE Transactions on Medical Imaging*, vol. 27(12): 1685–1703 (2008).
10. Rabbani, H., Nezafat, R., Gazor, S.: Wavelet-Domain Medical Image Denoising Using Bivariate Laplacian Mixture Model. *IEEE Transaction on Biomedical Engineering*, vol. 56(12): 2826–2837 (2009).
11. Sendur, L., Selesnick, W. I.: Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Transaction on signal processing*, vol. 50(11): 2744–2756 (2002).
12. Bhaduria, H. S., Dewal, M. L.: Performance evaluation of curvelet and wavelet based denoising methods on brain computed tomography images. In *IEEE Int Conf Emerg Trends Electr Comput Technol (ICETECT)* pp. 666–670, (2011).
13. Kingsbury N. C.: The dualtree complex wavelet transform: a new efficient tool for image restoration and enhancement. In *Proc. 9th European Signal Processing Conference (EUSIPCO 98)*, pp. 319–322 (1998).
14. Diwakar, M., Sonam, Kumar M.: CT image denoising based on complex wavelet transform using local adaptive thresholding and Bilateral filtering. In *ACM Proceeding of the Third International Symposium on Women in Computing and Informatics* pp. 297–302, (2015).
15. Cunha A. L. da, Zhou J. P., Do M. N.: The Nonsubsampled Contourlet Transform: Theory, Design, and Applications. *IEEE Transactions on Image Processing*, vol. 15(10):3089–3011 (2006).
16. Donoho DL, Duncan MR: Digital curvelet transform: strategy, implementation and experiments. Standford University, (1999).
17. Ouyang H. B., Quan H. M., Tang Y. , ZengY. Z.: Image Denoising Algorithm using Adaptive Bayes Threshold Subband Based on NSCT. In *Electronic Design Engineering*, vol. 19(23):185–188 (2011).

Using Musical Beats to Segment Videos of *Bharatanatyam Adavus*

**Tanwi Mallick, Akash Anuj, Partha Pratim Das
and Arun Kumar Majumdar**

Abstract We present an algorithm for audio-guided segmentation of the Kinect videos of *Adavus* in *Bharatanatyam* dance. *Adavus* are basic choreographic units of a dance sequence in *Bharatanatyam*. An *Adavu* is accompanied by percussion instruments (Tatta Palahai (wooden stick)—Tatta Kozhi (wooden block), Mridangam, Nagaswaram, Flute, Violin, or Veena) and vocal music. It is a combination of *events* that are either postures or small movements synchronized with rhythmic pattern of beats or *Taals*. We segment the videos of *Adavus* according to the percussion beats to determine the events for recognition of *Adavus* later. We use *Blind Source Separation* to isolate the instrumental sound from the vocal. Beats are tracked by onset detection to determine the instants in the video where the dancer assumes key postures. We also build a visualizer for test. From over 13000 input frames of 15 *Adavus*, 74 of the 131 key frames actually present get detected. Every detected key frame is correct. Hence, the system has 100 % precision, but only about 56 % recall.

Keywords Music-driven dance video segmentation • Multimodal Indian classical dance data captured by kinect • Onset detection on Indian music • Music-to-dance video synchronization

T. Mallick (✉) · A. Anuj · P.P. Das · A.K. Majumdar

Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur
721302, India

e-mail: tanwimallick@gmail.com

A. Anuj

e-mail: akashanuj.iitkgp@gmail.com

P.P. Das

e-mail: ppd@cse.iitkgp.ernet.in

A.K. Majumdar

e-mail: akmj@cse.iitkgp.ernet.in

1 Introduction

India has a rich tradition of classical dance. *Bharatanatyam* is one of the eight Indian classical dance forms. *Adavus* are basic choreographic units that are combined to form a dance sequence in *Bharatanatyam*. These *Adavus* are performed in synchronization with rhythmic pattern of beats known as *Taal*. The *Adavus* are classified according to the style of footwork employed and the *Taal* on which they are based (synchronized).

Every *Adavu* of *Bharatanatyam* dance is a combination of events, which are either *Key Postures* or *Short yet Discrete Movements*. These events are synchronized with the *Taal*. Our objective here is to find the beat pattern or *Taal* from the audio of the musical instrument and locate the corresponding events of the *Adavus*. The beat detection and *Taal* identification from an *Adavu* leads to meaningful segmentation of *Bharatanatyam* dance. We propose to adapt an algorithm for onset detection to achieve effective segmentation of videos of *Adavus* into events. We also build a visualizer to validate segmentation results.

After an overview of related work in Sect. 2, we explain the concept of *Taal* in Sect. 3. The methodology of our work is outlined in Sect. 4 followed by the elucidation of data capture and data sets in Sect. 5. *Blind Source Separation (BSS)* to segregate the instrumental (typically, percussion) sound from the vocal music is discussed in Sect. 6. The beat tracking/onset detection for the audio to locate the events in the corresponding video is elaborated in Sect. 7 followed by video segmentation and visualization in Sect. 8. We talk about the results in Sect. 9 and conclude in Sect. 10.

2 Related Work

Indian classical music, as used in Indian classical dance like *Bharatanatyam*, is based on a sophisticated rhythmic framework, where the rhythmic pattern or *Taal* describes the time scale. Beat detection and *Taal* recognition are challenging problems as Indian music is a combination of instrumental audio and vocal speech. Several attempts [1, 2] have been made to separate the audio streams into independent audio sources without any prior information of the audio signal. Further, several researchers have worked [3–7] to extract the rhythmic description in music through various *Beat Tracking algorithms* to extract the long duration as well as the short duration rhythmic structures. *Onset Detection* is a dominant and effective approach for *Beat Tracking*. Bello et al. [8] present a nice tutorial on *Onset Detection in Music Signals*.

There is, however, no work that uses the rhythms of music to identify key body postures in videos of Indian classical dance.

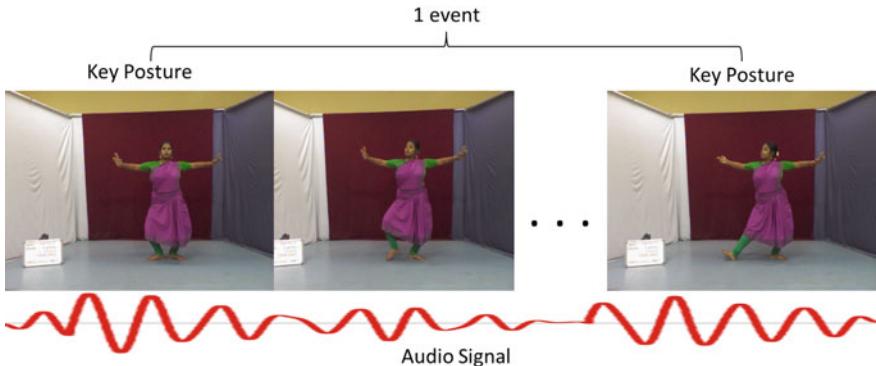


Fig. 1 An Event in a *Bharatanatyam Adavu*

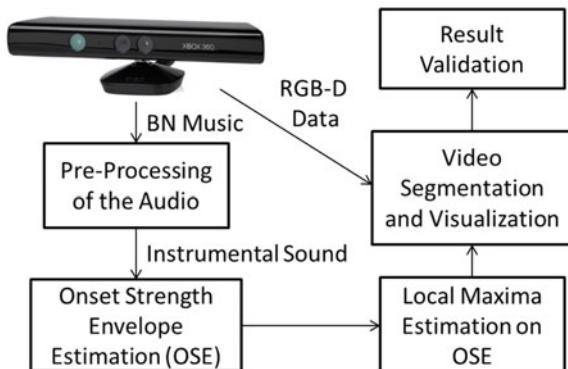
3 Concept of *Taal* in *Bharatanatyam*

Adavus are performed along with the rhythmic syllables played in a particular *Taal* or rhythmic pattern of beats that continue to repeat in cycles. Rhythm performs the role of a timer. Between the interval of beats, the dancer changes the posture. We define these as *Key Postures*. The sequence of frames between two key postures corresponding to two consecutive beats is defined as an *Event* that depicts a primitive audio-visual correspondence in an *Adavu* (Fig. 1).

4 Methodology

We intend to track beats in the audio stream to determine the time instant of the beat and then to extract the RGB frame corresponding to the same instant to determine the events of an *Adavu* video. The steps are as follows (Fig. 2):

Fig. 2 Flowchart of *Bharatanatyam* Video Segmentation



1. Use *Non-diagonal Audio Denoising* [9] through adaptive time–frequency block thresholding to denoise the audio stream.
2. Extract different sources from the audio stream by *Blind Source Separation (BSS)* [1, 10]. Select the instrumental sound for further analysis.
3. Estimate the *Onset Strength Envelope (OSE)* [3].
4. *Onset Detection* is done on the OSE to estimate the time instant of a beat. Before using *Onset Detection*, dynamic programming from [3] was tried to compute the time instant of a beat. This often detected more beats than were actually there. Hence, we improved the algorithm from [3] by finding local maxima in OSE to estimate onset.
5. Extract the video frame at *Onset* or the estimated time instant of a beat. This gives the key postures and segments the video. A tool is built to visualize segments.
- 6 Match the results with the segmentation by experts.

5 Capturing Data Sets

We recorded *Adavus* using *nuiCapture* [11] on Windows records and analyze Kinect data at 30 fps. RGB, skeleton, audio, and depth streams were captured for 15 *Adavus* using Kinect for Windows. These 15 *Adavus*—*Tatta*, *Natta*, *Utsanga*, *Tirmana*, *Tei Tei Dhatta*, *Sarika*, *Pakka*, *Paikkal*, *Joining*, *Katti/Kartari*, *Kuditta Nattal*, *Mandi*, *Kuditta Mettu*, *Kuditta Tattal*, and *Sarrikkal*—together cover all constituent postures and movements of *Bharatanatyam*. All 15 *Adavus* are used in our experiments.

Each *Adavu* was recorded separately by three dancers to study individual variability.

6 Blind Source Separation

The recorded audio streams are often noisy. So we first need to denoise the stream. Audio denoising aims at attenuating environment and equipment noise while retaining the underlying signals. We use *Non-diagonal Audio Denoising* through adaptive time–frequency block thresholding by Cai and Silverman [9]. We find that this is effective in reduction of noise in musical streams. Next we perform source separation.

The musical (beating) instrument used for an *Adavu* is a *Tatta Palahai* (wooden block) and a *Tatta Kozhi* (wooden stick). This is played alongside the vocal sound and is mixed. We separate the sound of the instrument (has *beats*) from the vocal music using *Flexible Audio Source Separation Toolbox (FAAST)* [1, 10]. It was able to segment the audio stream into four parts—*Melody*, *Bass*, *Drums*, and *Other sources*. We selected the *Drums* as we need the beating instrument. Experiments with our *Adavu* videos show good separation for the beating sound even in the presence of multiple instruments.

7 Beat Tracking

We attempt to track the beats from the denoised audio stream using two methods as discussed below.

7.1 Method 1. Beat Tracking by Dynamic Programming

We first explore the beat tracking algorithm by Ellis [3]. It starts with an estimation of a global tempo to construct a transition cost function, and then uses dynamic programming to find the best-scoring set of instants for beats that reflect the tempo as well as correspond to moments of high *onset strength* derived from the audio. This goes as follows:

Onset Strength Envelope (OSE) is calculated as follows:

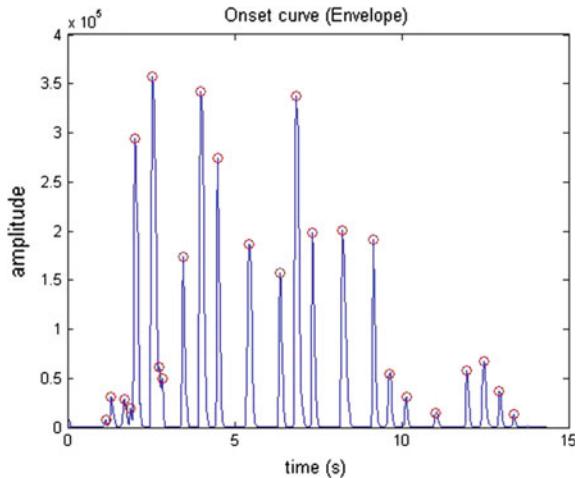
- Audio is re-sampled at 8KHz, and then STFT¹ (spectrogram) is calculated using 32 ms windows and 4 ms advance between frames.
- This is then converted into an approximate auditory representation by mapping to *40 Mel bands* via a weighted sum of the spectrogram values.
- The Mel spectrogram is converted into dB, and the first-order difference along time is calculated in each band. Negative values are set to zero (half wave rectification), and then the remaining positive differences are summed up across all frequency bands.
- This signal is passed through a high-pass filter with a cutoff around 0.4 Hz to make it locally zero mean, and is smoothed by convolving with a Gaussian envelope of about 20 ms width. This gives a *1D OSE* as a function of time that responds to proportional increase in energy summed across approximately auditory frequency bands.

Tempo Period (TP) is the inter-beat interval, τ_p . Autocorrelation of the OSE $O(t)$ is computed to reveal the regular, periodic structure of the *Tempo Period Strength (TPS)* by $TPS(\tau) = W(\tau) \sum_t O(t)(t - \tau)$, where $W(t)$ is a *Gaussian Weighting Function* on a log-time axis. The τ for which $TPS(\tau)$ is largest is then the estimate for τ_p .

Dynamic Programming (DP) Given OSE and TP, we can find the sequence of time instants for beats that correspond to both the perceived onsets in the audio signal and also constitute a regular, rhythmic pattern in them. The objective function $C(t_i) = \sum_{i=1}^N O(t_i) + \alpha \sum_{i=2}^N F(t_i - t_{i-1}, \tau_p)$ combines both these goals, where t_i is the sequence of N beat instants found out by the beat tracker, $O(t)$ is the OSE, τ_p is the TP, α is a weight to balance the relative importance, and $F(., .)$ is a function that measures the consistency between the inter-beat interval and the ideal spacing τ_p defined by the target tempo. We use a simple squared-error function $F(\Delta t, \tau) = -(\log \frac{\Delta t}{\tau})^2$ applied to the log ratio of actual and ideal time spacing.

¹Short-Time Fourier Transform.

Fig. 3 Unequal separation of the onsets



For the objective function above the best-scoring time sequence can be assembled recursively to calculate the best possible score $C^*(t) = O(t) + \max_{\tau=0 \dots t} \{\alpha F(t - \tau, \tau_p) + C^*(\tau)\}$, of all sequences that end at time t .

This follows from the fact that the best score for time t is the local onset strength, plus the best score to the preceding beat time τ that maximizes the sum of that best score and the transition cost from that time. In the process, the actual preceding beat that gave the best score is also recorded as $P^*(t) = O(t) + \arg \max_{\tau=0 \dots t} \alpha F(t - \tau, \tau_p) + C^*(\tau)$.

To find the set of optimal beat times for an OSE, C^* and P^* are computed for every time starting from zero. The largest C^* forms the largest beat instant. Next we backtrack via P^* , find the beat time $t_{N-1} = P^*(t_N)$, and continue backwards till the beginning to get the entire beat sequence $\{t_i\}^*$.

The DP performs well only for a limited set of *Taals* as used in *Bharatanatyam*. This is because it assumes that the beats reflect a locally constant inter-beat interval. This is not true for all *Bharatanatyam Taals*, and any two consecutive onsets might have variable time gaps between them. Figure 3 shows a *Taal*, where the beats/onsets are not equally separated.

The DP solutions lead to the overdetection of beats. This is not acceptable, since we only want good onsets corresponding to salient body postures in the dance. Hence, we propose the method of local maxima detection.

7.2 Method 2. Detection of Local Maxima in OSE

Our proposed method uses the OSE found earlier. We detect the local maxima in the envelope. The local maxima would correspond to the key postures. Figure 4 shows the detection on onsets for the *Utsanga* and *Tirmana Adavus*.

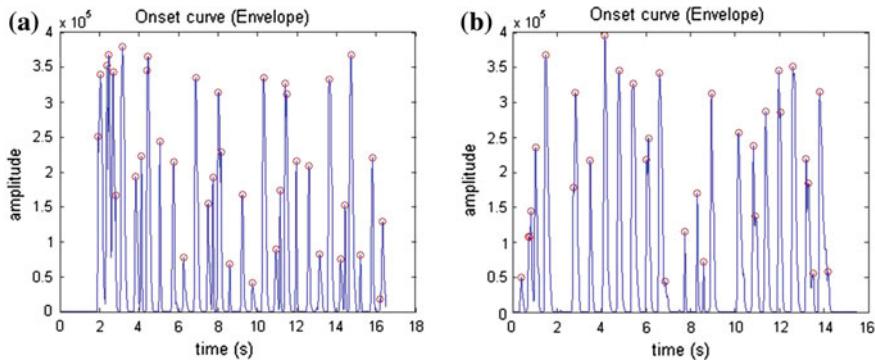


Fig. 4 Onset detection in *Adavus* **a** *Utsanga* **b** *Tirmana*

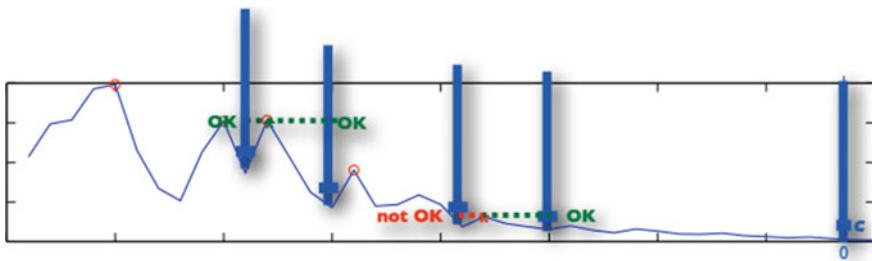


Fig. 5 Avoiding overdetection of local maxima

Avoiding Overdetection of Local Maxima Naive detection of local maxima usually leads to overdetection. To avoid this, a given local maximum is considered as a peak if the difference of amplitude with respect to both the previous and successive local minima (when they exist) is higher than a threshold $cthr$ (0.1, by default). This distance is expressed with respect to the total amplitude of the input signal. A distance of 1, for instance, is equivalent to the distance between the maximum and the minimum of the input signal.

This is implemented from MIRtoolbox [12] and illustrated in Fig. 5.

Retaining Good Onsets It is important that we represent an *Adavu* by a minimal set of body key postures. If two local maxima are very close to each other in time (the difference being less than a threshold $tthr = 0.15$ s), then there would be almost no change in the posture at the corresponding onsets. In such cases, we retain the maxima with the higher peak. A maxima with a higher peak corresponds to an onset with higher confidence. Figure 6b, d shows the removal of unwanted local maxima for the *Utsanaga* and *Tirmana* *Adavu*.

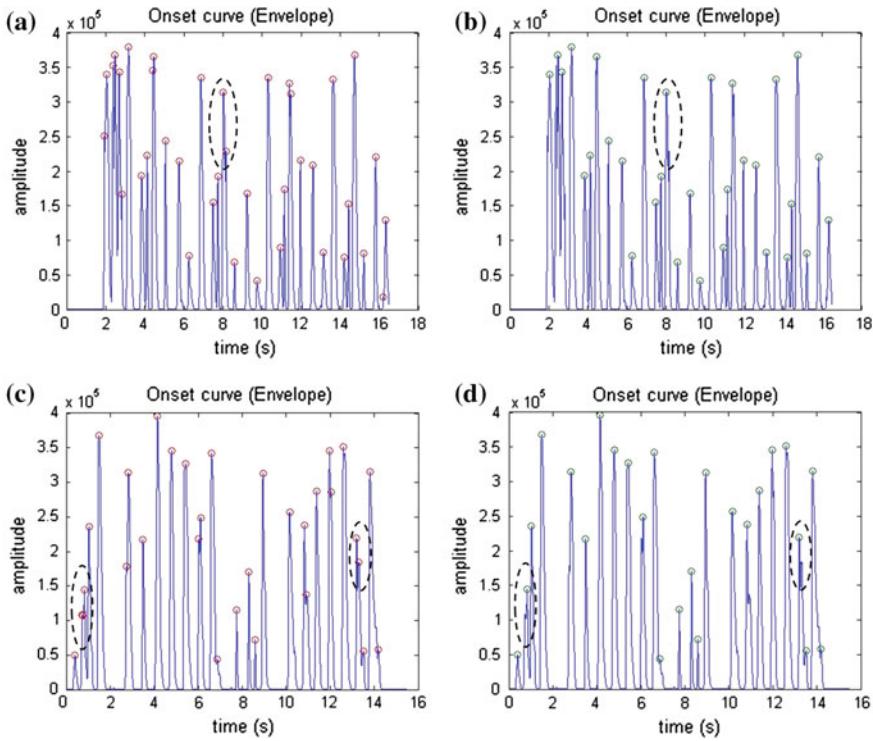


Fig. 6 Retaining good onsets **a** Detection in *Utsanga* **b** Retention **c** Detection in *Tirmana* **d** Retention

8 Video Segmentation and Visualization

Next we use the detected beat instants to segment the videos into events and visualize the key postures in them.

8.1 Segmentation into Events

Since the recording has been done at 30 fps, we know the time stamp for each frame in the RGB, skeletal, or depth stream by the frame number. Hence, given the onset times of beats in the audio stream we can find the corresponding frames (frame numbers) at the onset times by simple temporal reasoning. The frame number corresponding to an onset time t would be $(30 \times t)$, where t is in seconds. Since $(30 \times t)$ might be a floating point value, we round it off to the nearest integer and obtain the corresponding frames for RGB, depth, and skeleton.



Fig. 7 Visualizing correspondence between onset (audio) and RGB frame—*selecting an onset by mouse automatically takes one to the corresponding frame*

8.2 Visualization of Key Postures

A visualization tool has been built to view the correspondence between the onsets and the key posture frames. This helps us to validate if the postures selected are actually those at the onsets of the audio signal. Using this tool we select any of the onset points as given by local maxima detection. It then displays the corresponding RGB frame. Figure 7 shows a snapshot of the tool.

9 Results and Discussion

Using the visualizer we have tested the method for videos of 15 *Adavus*. In total, 74 key posture frames were detected by the system based on the onsets from a total of over 13000 frames in 15 videos. *Bharatanatyam* experts reviewed and verified that every detected key posture was indeed correct.

Independently, the experts were asked to identify key postures in the 15 videos. They manually inspected the frames and extracted 131 key posture frames from the 15 videos including the 74 key postures as detected above. So our system has 100 % precision, but only about 56 % recall.

10 Conclusions

Here we have attempted segmentation of videos of *Adavus* in *Bharatanatyam* dance using beat tracking. We engaged a dynamic programming approach [3] using the global tempo period (uniform inter-beat interval) estimate and the onset strength envelope. It performed well only on some *Adavus*, while on the others, it over-detected beat instants due to the non-uniformity of inter-beat intervals for a number of *Taals*.

We have adapted an algorithm for OSE with detection of local maxima to estimate beats. This does not need the assumption of global tempo period (uniform inter-beat interval) as in [3]. Further, we propose heuristics to avoid overdetection of onsets and retain only the good peaks to get a minimal sequence of key postures to represent an *Adavu*. From a set of onset times, we find the corresponding RGB (skeleton/depth) frames. We have also developed a visualization tool for validation.

We have tested the method for 15 *Adavus*. We find that our system has 100 % precision, but only about 56 % recall. So we need to strike a balance between the overdetection of the DP approach and the over-precision of the local maxima method. We also need to use the domain knowledge of the structure of *Bharatanatyam* to aid the segmentation. In addition, we are focusing on the classification of the *Adavus* based on the detected event sequences.

Acknowledgements The work of the first author is supported by TCS Research Scholar Program of Tata Consultancy Services of India.

References

1. Alexey Ozerov, Emmanuel Vincent, and Frdric Bimbot. A general flexible framework for the handling of prior information in audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20, 2012.
2. Yun Li, K. C. Ho, and Mihail Popescu. A general flexible framework for the handling of prior information in audio source separation. *Biomedical Engineering, IEEE Transactions on*, 61, 2014.
3. Daniel P.W. Ellis. Beat tracking by dynamic programming. *Journal of New Music Research*, 36, 2007.
4. Jonathan T. Foote and Matthew L. Cooper. Visualizing musical structure and rhythm via self-similarity. In *MULTIMEDIA '99 Proceedings of the seventh ACM international conference on Multimedia*, pages 77–80, 1999.
5. Jonathan T. Foote and Matthew L. Cooper. Media segmentation using self-similarity decomposition. In *Proc. SPIE Storage and Retrieval for Media Databases*, 2003.
6. Zafar Rafii and Bryan Pardo. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21, 2012.
7. Ajay Srinivasamurthy, Gregoire Tronel, Sidharth Subramanian, and Parag Chordia. A beat tracking approach to complete description of rhythm in indian classical music. In *2nd Comp-Music Workshop*, 2012.

8. Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, 13, 2005.
9. Guoshen Yu, Stephane Mallat, and Emmanuel Bacry. Audio denoising by time-frequency block thresholding. *Signal Processing, IEEE Transactions on*, 56, 2008.
10. Yann Salaun. Flexible audio source separation toolbox(faast). <http://bass-db.gforge.inria.fr/faast/> Last accessed on 10-Jan-2016, 2016.
11. Cadavid Concepts. nuiCapture Analyze. <http://nuicapture.com/> Last accessed on 10-Jan-2016, 2016.
12. Mathworks. Mirtoolbox: An innovative environment for music and audio analysis. <http://www.mathworks.in/matlabcentral/fileexchange/24583-mirtoolbox> Last accessed on 10-Jan-2016, 2016.

Parallel Implementation of RSA 2D-DCT Steganography and Chaotic 2D-DCT Steganography

G. Savithri, Vinupriya, Sayali Mane and J. Saira Banu

Abstract Information security has been one of the major concerns in the field of communication today. Steganography is one of the ways used for secure communication, where people cannot feel the existence of the secret information. The need for parallelizing an algorithm increases, as any good algorithm becomes a failure if the computation time taken by it is large. In this paper two parallel algorithms—parallel RSA (Rivest Shamir Adleman) cryptosystem with 2D-DCT (Discrete Cosine Transformation) steganography and parallel chaotic 2D-DCT steganography—have been proposed. The performance of both algorithms for larger images is determined and chaotic steganography is proved to be an efficient algorithm for larger messages. The parallelized version also proves to have reduced processing time than serial version with the speed-up ratios of 1.6 and 3.18.

Keywords Chaos · Communication · Encryption · Frequency domain · Information Security · RSA · Steganography

1 Introduction

With widespread distribution of digital data and internet, protecting the privacy of data to be communicated is a challenge. There are several techniques available for protection of data, among which is steganography. The aim of steganography is to embed a piece of information like text or image into a larger piece of information,

G. Savithri (✉) · Vinupriya · S. Mane · J. Saira Banu
School of Computing Science and Engineering (SCSE), VIT University,
Vellore, India
e-mail: savithri.g2015@vit.ac.in

J. Saira Banu
e-mail: jsairabunu@vit.ac.in

so that the secret information is totally hidden inside the large and hence not easily detected by the hacker. Steganography can be classified into two domains: spatial and frequency. In spatial domain, the modifications are made directly on the pixels of the original image. However, since the pixels of the original image itself are altered, the distortion tends to be higher and is easily attacked by an external. In frequency domain, the carrier image is transformed from spatial domain to frequency domain using domain transformation techniques. The secret message is then embedded into the transformed coefficients of the cover to form the stego image [1, 2]. Frequency domain is more tolerable to cropping, shrinking, and image manipulation compared to spatial domain. There are many transforms used to map a signal into frequency domain like Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT). Many types of images can be used as cover media such as Bitmap File Format (BMP), Joint Photographic Experts Group (JPEG), and Graphics Interchange Format (GIF) images [3]. This research mainly focuses on .png and .jpeg formats.

Steganography has attracted the attention of many researchers. Many techniques have been developed to hide secret messages. Least significant bit (LSB) technique is used for image steganography with a new security conception that uses secret key to encrypt hidden information [4, 5]. DCT algorithm is more suitable for the steganography application compared to the LSB and the DWT-based algorithms [6, 7]. To utilize computing power of multicore processors Compute Unified Device Architecture (CUDA) is used to implement steganography [8]. An area efficient row-parallel architecture for 8×8 2-D DCT computation based on real-time implementation of algebraic integer (AI) [9] that reduces the number of DCT cores. RSA is computationally heavy because of modular arithmetic, therefore to overcome this disadvantage and to implement public key algorithms in a faster way, CUDA and Pthread are used for decryption in RSA when large numbers are created by homomorphic encryption [10]. Various parallel implementations of RSA algorithm involving variety of hardware and software implementations have also been done [11]. Image steganography based on LSB insertion and RSA encryption technique for the lossless jpeg images has been proposed [12] for attaining maximum embedding capacity in an image and provides more security to data. A secure DCT steganography method that hides a secret image in another image randomly using Chaos is proposed in [13]. It is robust image embedding process than LSB insertion [14, 15]. This paper uses serial method of implementing chaos which requires more time for execution. A new technique for image steganography (Hash-LSB) with RSA algorithm to provide more security is proposed in [16].

In this paper, two different encryption techniques—parallel RSA with steganography and Chaos with steganography—are proposed (Figs. 1 and 2).

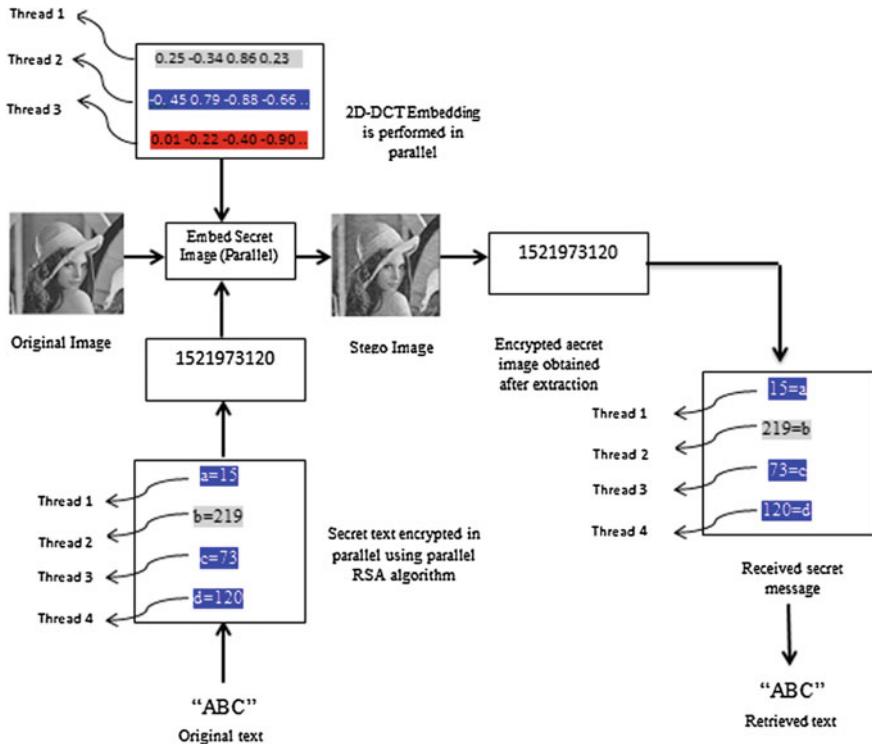


Fig. 1 Parallel RSA with steganography

Both the algorithms are analyzed against various factors and the later encryption technique is found to be better. This paper also focuses on the research of parallelization of the above two algorithms. Since each algorithm contains multiple sections, each section can be executed in parallel. OpenMP platform provides a good set of parallel directives, which is utilized for this research as the platform. Compared to the original sequential code the methods proposed here have better speed-up ratios of 1.6 and 3.18.

Organization of the report is as follows: Sect. 2 explains the parallel RSA–steganography algorithm. Section 3 introduces the parallel chaotic–steganographic algorithm. The hardware platform and software used are discussed in Sect. 4. Section 5 presents different simulation and experimental results with a brief discussion. Finally, Sect. 6 concludes the research work.

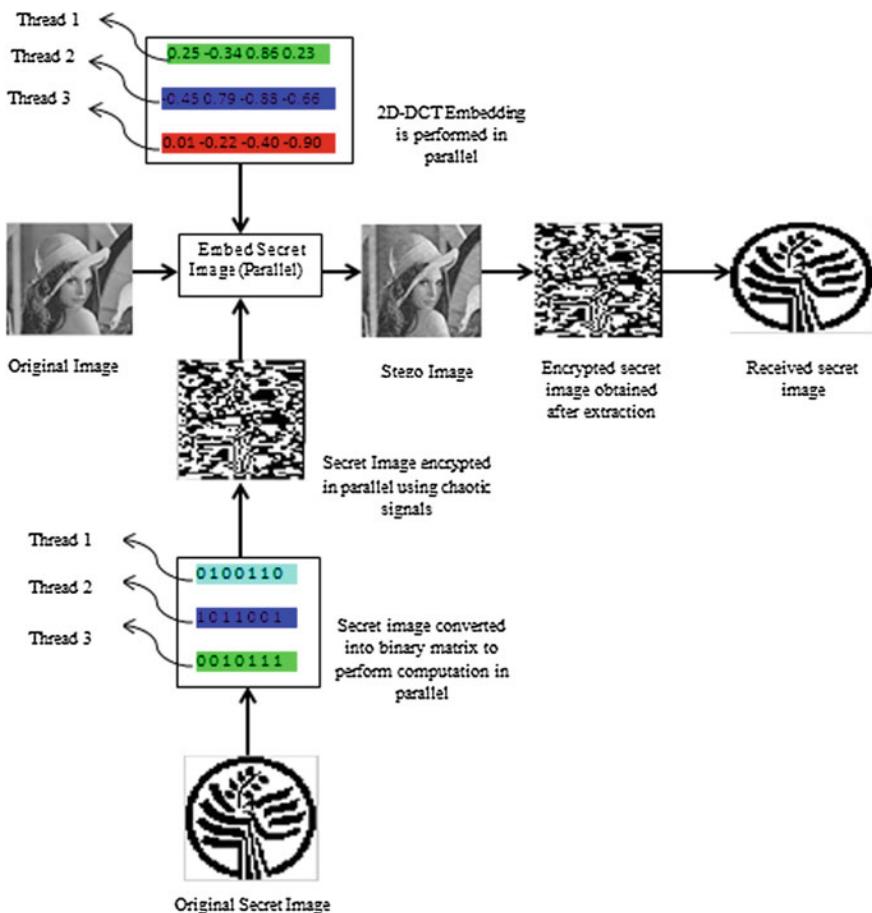


Fig. 2 Parallel Chaos and steganography

2 RSA–Steganography Algorithm

RSA algorithm is used to encrypt and decrypt data while transmitting on the internet. It has two different keys and thus is an asymmetric cryptographic algorithm. RSA key is derived from factoring large integers that are product of two large prime integers. An RSA algorithm can be divided into three parts as key generation, encryption, and decryption.

2.1 Parallel RSA

The RSA algorithm is based on modular arithmetic. It consumes more time in the encryption and decryption which includes exponential and reduction processes. Parallelizing these individual parts in the algorithm will result in better efficiency. The method used to implement parallel RSA is repeated square and multiply method [11]. The value of e in order to parallelize should be even. By this it can be divided into four or more number of cores depending on the system. Then each core performs computation independently and parallelly and at last all computation results are combined to get final result. One can then reframe the equation as given in Eq. 1:

$$g^e \bmod m = (g^{e/2} * g^{e/2}) \bmod m \quad (1)$$

The recursive definition of repeated square and multiply method is described in Eq. 2:

$$\text{ModExp}(g, e, m) = \begin{cases} 1 & \text{if } e = 1 \\ g * \text{ModExp}(g, (e - 1), m) & \text{if } e = \text{odd} \\ \text{ModExp}(g, e/2, m)^2 \bmod m & \text{if } e = \text{even} \end{cases} \quad (2)$$

2.2 RSA and Steganography

The parallel RSA and steganography algorithms when executed individually on a parallel platform proved out to be more efficient. Therefore, in this section an attempt to combine these two individual algorithms—parallel RSA and Steganography—is made (Fig. 3). The main advantage of this combination is that a double encryption technique is obtained and the encryption can be easily split into multiple processors as well.

ALGORITHM: PARALLEL RSA STEGANOGRAPHY	
Input : Secret message, cover image	
Output : Encrypted image	
<ol style="list-style-type: none"> 1. Secret text is converted to cipher text using parallel RSA algorithm. 2. Cover image is converted from spatial domain to frequency domain using 2D-DCT algorithm. 3. The converted cipher text is embedded to cover image and transmitted to receiver side. 4. Inverse DCT is applied to obtain cover image back to spatial domain. 5. Cipher text is obtained from cover image. 6. At the last step original message is decrypted from cipher text. 	

Fig. 3 Parallel RSA steganographic algorithm

3 Parallel Chaos–Steganography Algorithm

3.1 Chaos

Technically chaos can be defined as the extreme “sensitivity to initial conditions mathematically present in the systems.” Chaos recently has taken an important place in the security domain in today’s world. A clear definition of chaos was first given by Poincare [17] using the example of a sphere. Here, Duffing map is considered, as it is easy to implement and has good randomness properties.

3.2 2D-DCT Steganography

An image can be converted into its frequency domain using one of the transform equations like Fourier, cosine, or wavelet. In this paper, cosine transform equation is considered which is given by Eq. 3:

$$DCT_{ij} = \alpha_i \alpha_j \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} C_{mn} \cos \frac{\pi(2m+1)i}{2M} \cos \frac{\pi(n+1)j}{2N} \quad (3)$$

where

$$\alpha_i = \begin{cases} \frac{1}{\sqrt{M}} \dots i=0 \\ \sqrt{\frac{2}{M}} \dots 1 \leq i \leq M-1 \end{cases} \quad 0 \leq i \leq M-1$$

$$\alpha_j = \begin{cases} \frac{1}{\sqrt{N}} \dots j=0 \\ \sqrt{\frac{2}{N}} \dots 1 \leq j \leq N-1 \end{cases} \quad 0 \leq j \leq N-1$$

3.3 Algorithm

Figure 4 specifies the algorithm for chaos. Each image can be represented as a set of pixel matrix. Considering this as the advantage, each pixel matrix is further divided into smaller sub-matrices and allotted to a processor. All the operations are carried out on these smaller matrices which are later combined to form the main matrix. Each of the individual steps is explained in detail.

Encryption of secret image

- a. $M \times N$ (total number of pixels) elements are generated using chaotic map equation with appropriate initial conditions. The $M \times N$ elements are uniformly distributed in n-cores so that the calculation becomes more efficient.

ALGORITHM: CHAOTIC STEGANOGRAPHY
Input: Secret Image, Cover Image
Output: Encrypted message
<ol style="list-style-type: none"> 1. Encryption process <ol style="list-style-type: none"> a. Chaotic Sequence generator b. Smaller image encryption 2. Embedding process <ol style="list-style-type: none"> a. Generation of 2D-DCT coefficients of encrypted message. b. Generation of 2D-DCT coefficients of host image. c. Embedding of encrypted message to larger image. 3. Extraction process <ol style="list-style-type: none"> a. Extraction of encrypted message from large image. b. Re-generation process of 2D-DCT coefficients of encrypted message. 4. Decryption process <ol style="list-style-type: none"> a. Smaller image decryption. b. Image recovering

Fig. 4 Parallel Chaotic steganographic algorithm

- b. The average value of the obtained elements from the chaotic map is taken and made it as threshold value. Any value above it will be assigned as 1 and below it is assigned as 0. Thus the obtained chaotic elements are converted into binary sequence.
- c. This binary matrix is spilt into smaller matrix and each sub-matrix is given to an individual core for further encryption.
- d. For security purpose the binary image is encrypted using the binary-converted chaotic elements. A ‘xor’ operation is performed between the binary secret image and binary-converted chaotic matrix. This will result the secret image to become noise-like pattern.

Embedding Process.

- a. 2D-DCT of the encrypted image is done. The sub-divided matrix is used so that it can be easily distributed to n-cores. This will transform the encrypted image from spatial to frequency domain and convert the matrix values form binary into double.
- b. The host image is also converted into frequency domain. Before taking the 2-D DCT of the host image, the 8-bit gray-level image is converted into 64-bit double image. If 8-bit image is used in the DCT operation, it will produce a 64-bit output. Now for retrieving the original image Inverse DCT (IDCT) of the sequence is taken. The IDCT operation also produces 64-bit output. But originally, 8-bit image was used for the input. If the 64-bit double output is converted back into 8-bit unsigned integer image, an error will be produced as the pixel magnitude after the decimal point is rounded off while conversion. So the

best solution for this problem is to convert the host image into double image from the beginning itself. There is no need of converting back, since the whole watermarking procedure uses 64-bit images as its input and output.

- c. The encrypted image is embedded in a portion of host image, which has the size equal to the size of secret image. In case the smaller image is 4×4 image and host a 4×8 image, a 4×4 portion is selected, anywhere in the host image, and then the smaller image is embedded in that portion. The selection of the position is completely user defined, and the user should provide the information about the position of the secret image in the host image, to the receiver, since it is very much required to know the position for the process of extraction.
- d. In the embedding process a linear combination equation of 2-D DCT coefficients of both encrypted secret image and the host image using a linear equation is used, to obtain a new sequence of 2-D DCT coefficients as shown in Eq. 4. The 2-D DCT coefficients of the selected 4×4 portion of host image are then replaced, with the new sequence of 2-D DCT coefficients.

$$c1 = c + \alpha.w \quad (4)$$

where

- c1— the new sequence of coefficients
- c— the sequence of 2-D DCT coefficients as in the selected 4×4 portion
- w— the sequence of 2-D DCT coefficients of encrypted watermark image, calculated as in the table
- α — the strength of the encrypted image

α is a very important parameter in the embedding process. The value of α is responsible for two characteristics of the watermarked image:

- Visibility of the secret image in the stego image.
- Robustness of the stego image to external attacks and modifications.

Generally, the value of α varies within the range 0.1–0.5. Lower the value of α , lower the visibility of the secret image, which is desired. But in the meantime the robustness of secret image reduces, making the image a little fragile. Conversely, if α has a little higher magnitude, like 0.4–0.5, the robustness of the encrypted image increases significantly, making the image more robust. But the visibility of the secret also increases, which is not desired. So a trade-off between visibility and robustness should be considered and based on it the value of α is selected. In this paper $\alpha = 0.15$ is chosen. After choosing the value of α , we perform embedding process by evaluating Eq. 4.

Extraction and Decryption Process: The reverse algorithm is applied at the receiver's side to retrieve back the secret image.

4 Platform Used

The simulation is carried out using Visual Studio version 2012. Visual Studio supports OpenMP parallel programming language. OpenMP is an API that supports multi-platform shared memory multiprocessing programming in languages like C, C++, and Fortran. A system configuration of 4 GB RAM with Windows 8 as operating system is used. Intel i5 processor is used for the experiments.

5 Results and Discussion

In this section the simulations and obtained results are explained. In the tests, the two proposed algorithms have been applied on several standard host images like “Lena,” “Baboon,” and “Balloon” of various sizes. To measure the image quality, signal-to-noise ratio (SNR), peak signal-to-noise ratio (PSNR), and correlation factor is used. SNR and PSNR measure the difference between host and stego image. Correlation factor is used to measure the difference between original and retrieved secret image. The execution time of both algorithms in parallel as well as in sequential is compared. It is observed that the performance is increased significantly in case of parallel computing.

Considering the secret text of four characters and the cover image of size 50*50 for various standard steganographic images (Lena, Baboon, and Balloon), the results are obtained as shown in (Graph 1, Graph 2, and Graph 3) Fig. 5. The algorithm provides a good visual quality as the distortion is very less. However, as the size of the image and secret message increases, retrieval of exact text message is not possible. This is because 2D-DCT method is a lossy compression technique and RSA being a lossless algorithm the combination proves to be less efficient.

Now considering the secret message as 50*50 binary image (tree) and a gray-scale host image of 200*200 (lena), the following set of images (Fig. 6) is obtained, after applying parallel chaotic steganography (Fig. 7).

The above algorithm is repeated for various host images and the results are analyzed (Graph 4). The algorithm is also tested for various image sizes (Graph 5 and Graph 6) and it is observed that the serial execution time is way more compared to parallel execution time. Thus proving the algorithm chosen is more suitable for parallel computing (Graph 7). For various image sizes the PSNR is constantly above 20. Therefore, the encryption technique is efficient. The distortion to naked eye is minimal making it robust to external attacks. Even with increase in image sizes the original secret image is retrieved accurately.

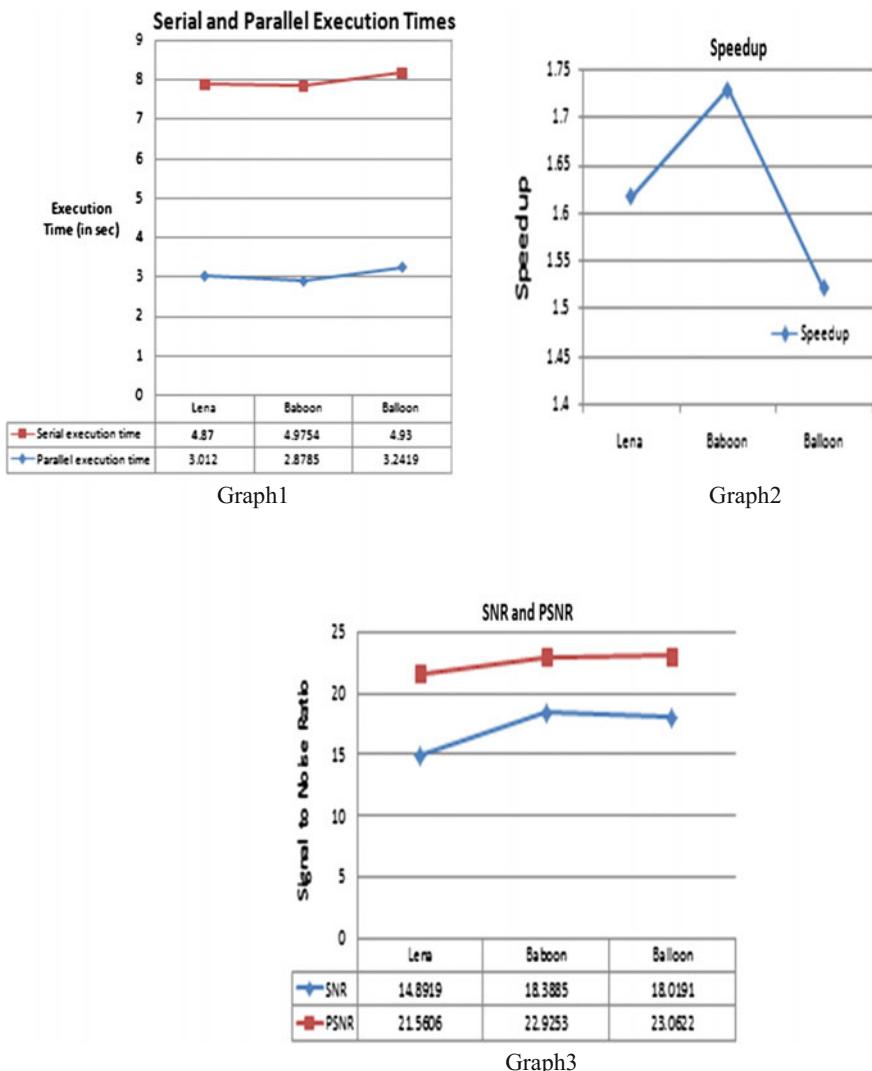


Fig. 5 Performance analysis graphs for parallel RSA and steganography

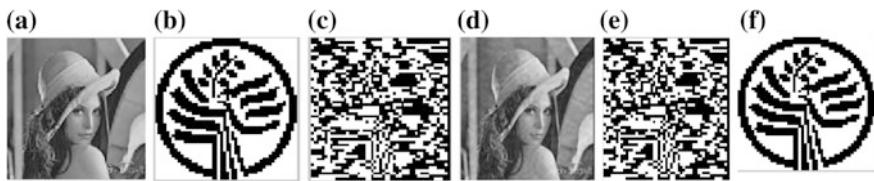
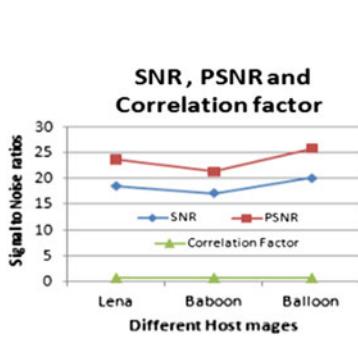
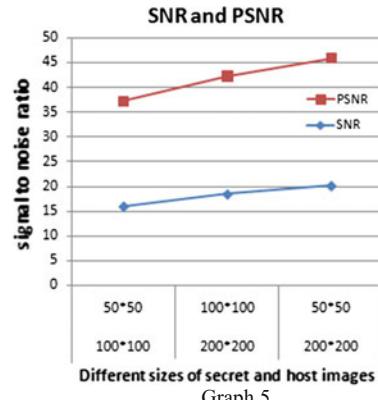


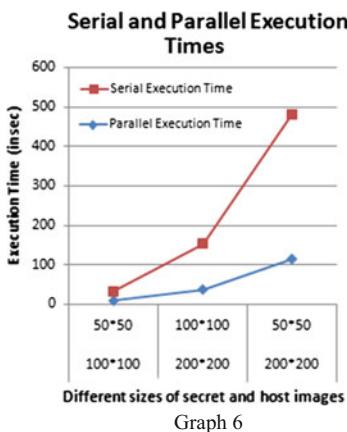
Fig. 6 a Original host image. b Original secret image. c Encrypted chaotic secret image. d Stego image. e Encrypted chaotic secret message obtained after extraction. f Retrieved secret image



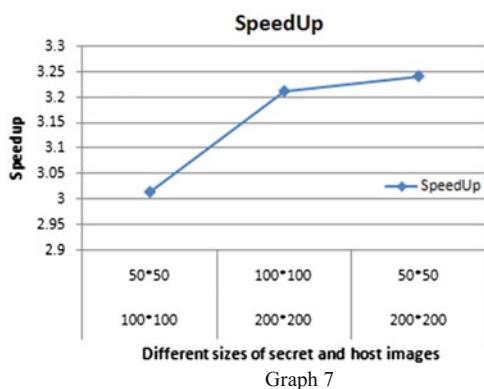
Graph 4



Graph 5



Graph 6



Graph 7

Fig. 7 Performance analysis graphs for parallel chaotic steganography

6 Conclusion

In this paper two parallel encryption techniques are proposed. In the first algorithm text message is encrypted using RSA to obtain the stego image. In the second algorithm a new parallelized chaotic 2D-DCT method is presented. Using Duffing map the secret image is encrypted into a uniform set of pixels and embedded into the frequency domain of the larger host image. Therefore, the locations of secret data in the host image are very random. Both the algorithms are compared. According to the simulation result the chaotic method of encryption proves to be more efficient. The above algorithms can be extended to larger color images.

References

1. S. Bhattacharyya, "A survey of steganography and steganalysis technique in image, text, audio and video as cover carrier." *Journal of global research in computer science* 2, no. 4 (2011).
2. S. Saejung, A. Boondee, J. Preechasuk, and C. Chantrapornchai, "On the comparison of digital image steganography algorithm based on DCT and wavelet," in *Computer Science and Engineering Conference (ICSEC)*, 2013 International, 2013, pp. 328–333.
3. N. Sathisha, K. Suresh Babu, K. B. Raja, K. R. Venugopal and L. Patnaik, "Embedding Information In DCT Coefficients Based On Average Covariance" *International Journal of Engineering Science and Technology (IJECT)*, 3 (4), 3184–3194. 2011.
4. Shamim Ahmed Laskar1 and Kattamanchi Hemachandran "High Capacity data hiding using LSB Steganography and Encryption", *International Journal of Database Management Systems (IJDMS)* Vol.4, No.6, December 2012
5. S. M. Masud Karim, Md. Saifur Rahman and Md. Ismail Hossain, "A New Approach for LSB Based Image Steganography using Secret Key" in *Proceedings of 14th International Conference on Computer and Information Technology (ICCIT 201 I)* 22–24 December, 201 I, Dhaka, Bangladesh
6. Dr. Ekta Walia, Payal Jain, Navdeep An Analysis of LSB & DCT based Steganography, *Global Journal of Computer Science and Technology* Vol. 10 Issue 1 (Ver 1.0), April 2010
7. Saravanan Chandran and Koushik Bhattacharyya, "Performance Analysis of LSB, DCT, and DWT for Digital Watermarking Application using Steganography" in *International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO) – 2015*
8. Samir B. Patel, Shrikant N. Pradhan and Saumitra U. Ambegaokar, "A Novel approach for implementing steganography with Computing Power Obtained by Combining CUDA and MATLAB " in *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 6, No.2, 2009
9. Amila Edirisuriya, Arjuna Madanayake, Renato J. Cintra, Vassil S. Dimitrov, and Nilanka Rajapaksha, "A Single-Channel Architecture for Algebraic Integer-Based 8×8 2-D DCT Computation" In *Ieee Transactions On Circuits And Systems For Video Technology*, Vol. 23, No. 12, December 2013
10. Abu Asaduzzaman, Deepthi Gummadi, and Puskar Waichal, "A Promising Parallel Algorithm to Manage the RSA Decryption Complexity" in *Proceedings of the IEEE Southeast Conf 2015*, April 9–12, 2015 - Fort Lauderdale, Florida
11. Sapna Saxena and Bhanu Kapoor, "State of the art parallel approaches for RSA public key based cryptosystem" in *International Journal on Computational Sciences & Applications (IJCSA)* Vol. 5, No.1, February 2015

12. Jatinder Kaur and Ira Gabba, “Steganography Using RSA Algorithm” in International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-3, August 2013
13. Milia Habib, Bassem Bakache, Dalia Battikh and Safwan El Assad, “Enhancement using chaos of a Steganography method in DCT domain” in ISBN: 978-1-479-4129-2/15/ \$31.00©2015 IEEE
14. Bhavana S and Dr K L Sudha “Text Steganography Using Lsb Insertion Method Along With Chaos Theory”, site: <http://arxiv.org/ftp/arxiv/papers/1205/1205.1859.pdf> as on october 30th 2015.
15. Dr. K. L. Sudha, and Manjunath Prasad, (Aug. 2011), “Chaos image encryption using pixel shuffling with henon map,” Elixir Elec. Engg. 38, pp 4492–4495.
16. Anil Kumar and Rohini Sharma “A Secure Image Steganography Based on RSA Algorithm and Hash-LSB Technique” in International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 7, July 2013, ISSN: 2277 128X
17. MazharTayel, Hamed Shawky, Alaa El-Din Sayed Hafez, “A New Chaos Steganography Algorithm for Hiding Multimedia Data”, Advanced Communication Technology (ICACT), 2012, 14th International Conference on 19–22 Feb, 2012 ISSN: 1738-9445

Thermal Face Recognition Using Face Localized Scale-Invariant Feature Transform

Shruti R. Uke and Abhijeet V. Nandedkar

Abstract Biometric face reorganization is an established means for the prevention of frauds in financial transactions and security issues. In particular, face verification has been extensively used to endorse financial transactions. Thermal face recognition is an upcoming approach in this field. This work proposes a robust thermal face recognition system based on face localized scale-invariant feature transform (FLSIFT). FLSIFT tackles the problem of thermal face recognition with complex backgrounds. Experimental results of proposed FLSIFT thermal face recognition system are compared with the existing Blood Vessel Pattern method. To test the performance of proposed and existing method, a new thermal face database consisting of Indian people and variations in the background is developed. The thermal facial images of 113 subjects are captured for this purpose. The test results show that the recognition accuracy of Blood Vessel Pattern technique and FLSIFT on face images with simple background is 79.28 % and 100 %, respectively. Moreover, the test performance on the complex background for the two methods is found to be 5.55 % and 98.14 %, respectively. It may be noted that FLSIFT is capable to handle background changes more efficiently and the performance is found to be robust.

Keywords Thermal face • Scale-invariant feature transform • Blood vessel pattern • ITFDB dataset

S.R. Uke (✉) · A.V. Nandedkar
Department of Electronics and Telecommunication Engineering,
S.G.G.S.I.E & T, Nanded, India
e-mail: ukeshruti@sggs.ac.in

A.V. Nandedkar
e-mail: avnandedkar@sggs.ac.in

1 Introduction

Biometric identification techniques can be done by two ways, either using physical characteristics or using behavior characteristics. Identification techniques that are based on physical characteristics are more difficult to counterfeit than behavior methods. Physical characteristic identification includes face recognition, voice recognition, vein recognition, fingerprint recognition, and iris recognition. Face recognition has a benefit that it does not require direct physical interaction with machine. Face recognition can be done in visual and thermal domain; each one has its advantages. Currently, most researchers are tending to use the thermal domain as thermal face characteristics are unique for a person [1] and represent a heat pattern emitted by an object.

Thermal mid-wave infrared (MWIR) is one of the partitions of electromagnetic (EM) spectrum that can fix problems like light illumination that may appear in visual domain. Also, any fake object present over face could be detected easily as they have different emissivity than human face. Socolinsky et al. [2, 3] suggested that thermal images of human faces can be a valid biometric and is superior as compared to visual images. Moreover, a thermal domain technique is not affected by scattering and absorption of smoke or dust and gives satisfactory result in complete darkness as well.

There are appearance-based techniques mostly used in thermal face recognition like principal component analysis (PCA) [4], linear discriminant analysis (LDA) [4], Kernel component analysis [5], local binary pattern [6], etc. Utilizing anatomical information of human face one can extract vascular information and can be used in recognition [7]. However, these techniques face problems in thermal domain due to amorphous nature of images and lack of sharp boundaries which makes object boundary extraction challenging [8]. It is observed that if thermal face image background consists of different objects, the recognition becomes difficult. The thermal pattern of background objects affects the performance. In this work a face localized scale-invariant feature transform (FLSIFT) based on scale-invariant feature transform (SIFT) [9] is proposed to address this issue. SIFT is a very robust feature descriptor for object recognition and matching [9].

The main contribution of this work includes development a new Indian thermal face database (ITFDB), evaluation of existing blood vessel pattern (BVP) technique [7] on ITFDB and development of FLSIFT face recognition system. The Sect. 2 briefly describes about thermal face recognition using BVP. Section 3 elaborates detailed setup about ITFDB creation. The proposed thermal face recognition using FLSIFT is detailed in Sect. 4. The experimental results are discussed in Sect. 5.

2 A Brief on Thermal Face Recognition Using BVP

This section briefly discusses the thermal face recognition system using BVP [7]. In this system, a thermal pattern is analyzed for the recognition obtained by superficial blood vessels pattern present over the bones and muscles. A typical thermal pattern is caused due to temperature gradient of warm blood flowing through the superficial vessels against the surrounding tissues. By knowing the thermal characteristics; pattern of blood vessels can be extracted. The implementation is done in four steps as shown in Fig. 1. In the first step, a reference image of a subject is used to do registration of three other images of the same subject. In second step, the face region is extracted by region growing segmentation algorithm with predefined seed points. For enhancing edge information, a standard Persona-Malik anisotropic diffusion filter [10] is applied over all images.

Extraction of enhanced edges is required for thermal pattern generation. White top-hat segmentation is used for this purpose. This gives an individual signature pattern for a subject as shown in Fig. 2. For each subject, four signatures are created using four different images. This signature is used to match with signatures of other

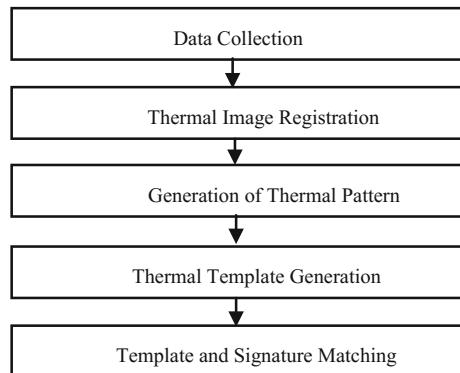


Fig. 1 Thermal face recognition using BVP

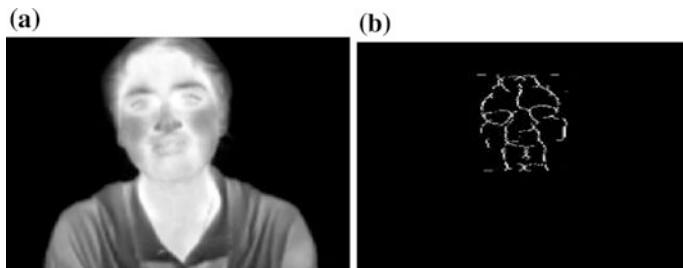


Fig. 2 **a** Thermal image **b** Extracted signature

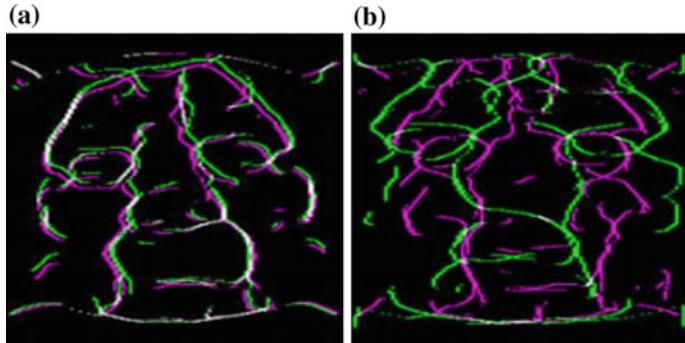


Fig. 3 Overlay of signature over **a** same subject **b** Different subject

images using distance-based similarity metric, such as Euclidean distance and Manhattan distance [7, 11]. Figure 3 shows the overlay of signature of a subject signature with signature of itself and with other subject.

3 Development of a New Indian Thermal Face Database

To evaluate performance of thermal face recognition systems in the Indian conditions, a need of an Indian thermal face dataset with complex background was felt. This work proposes ITFDB consisting of thermal face images of 113 subjects. The database is created using TESTO-875-1i thermal camera equipped with an uncooled detector with a spectral sensitivity range from 8 to 14 μm and provided with a standard optical lens. The thermographic camera provides a resolution of 160 \times 120 pixels for thermal.

The dataset was generated at a room temperature. For each subject, four frontal views were taken. Specific arrangement was maintained for the creation of dataset; camera was mounted on tripod stand, a chair was fixed at a distance of 1 m from tripod stand. Each subject was asked to seat straight in front of thermal camera, looking straight into the lens and snapshots were captured, as detailed in [7]. Database is available online at [12]. Other database details are as follows:

Figure 4 shows sample images from ITFDB. The frontal face images were captured with simple background, i.e. wall and with complex background; which



Fig. 4 Dataset Samples **a** simple background, **b** and **c** complex background

Table 1 Dataset details

Particular	No. of images
Number of subject	113
Images with simple background	452
Images with complex background	108
Total images in dataset	560

contains glass, wood, iron material, etc. As the different material has different emissivity and reflection it creates different temperature gradients in thermal image which can be clearly observed (Table 1).

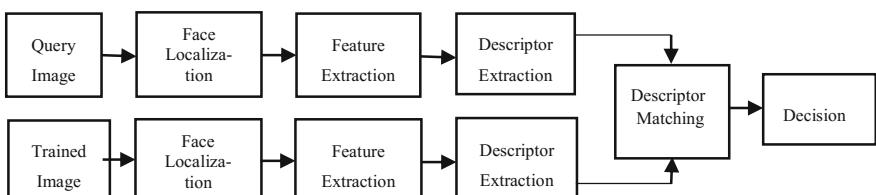
It may be noted that I2BVSD dataset with occluded thermal faces without complex background is proposed in [13, 14].

4 Thermal Face Recognition Using FLSIFT

In this work, a thermal face recognition system using face localized SIFT is proposed. The detail flow of the system is given in Fig. 5. Proposed system consists of five steps for the recognition. These include (1) face localization, (2) key point extraction, (3) descriptor Extraction, (4) descriptor matching, and (5) decision-making based on the maximum matching score of descriptors. The details of the proposed system are depicted in Fig. 5.

4.1 Face Localization

The first step, localizes face in the given thermal image is done. For a robust system, this is a crucial stage so that the disturbances due to background are minimized. The key points of these localized face regions are extracted using SIFT [9]. These robust features are invariant to affine transform, rotation, scale, and having distinctive features which are highly required in recognition was invented by Lowe [9]. To extract the face region, the thermal image is first binaries and using connected component labeling the larger connected area having same labeling is cropped and extracted as a face [15] the result is as shown in Fig. 6b.

**Fig. 5** Thermal face recognition using FLSIFT

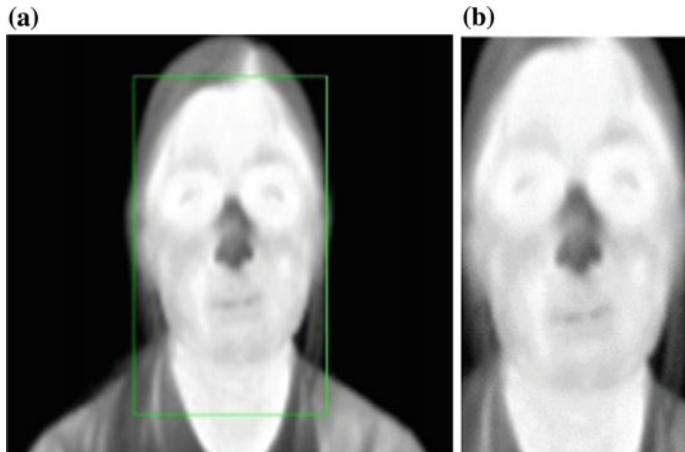


Fig. 6 **a** Original thermal image **b** Extracted face

4.2 FLSIFT Feature Extraction

After face localization, the features of faces are extracted using SIFT [9]. Extraction of feature using SIFTS have three steps: (1) Scale space key point selection, (2) Key point localization, (3) Orientation assignment. Scale space key point is selected, where local maxima and minima of difference-of-Gaussian function in the scale space is present. The convolution result of variable Gaussian function with the image gives scale space of an image. If $G(x, y, \sigma)$ is a variable Gaussian function and $I(x, y)$ is input image then scale space function $L(x, y, \sigma)$ is—

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (1)$$

With

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2 + y^2)/2\sigma^2} \quad (2)$$

The difference-of-Gaussian functions is derived as follows –

$$D(x, y, \sigma) = L(x, y, \sigma) - L(x, y, k\sigma) \quad (3)$$

with two nearby scales separated by a multiplicative factor (k).

At this stage many key points are obtained which either present in all scale space or in some of them. Final key points are selected which are present in all scale space, and a detail model created for location and scale determination. For invariance to rotation every key point assigned gradient orientation by the gradient

magnitude. The Eqs. (4) and (5) give the detail information about the gradient magnitude and gradient orientation.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (4)$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y))) \quad (5)$$

4.3 FLSIFT Descriptor

The computation of descriptor converts these key points into vector which is used further as a feature vector. A rectangular area of $(16 \times 16$ pixels) is centered on each key point; and then it is divided into 4×4 subregions which is characterized by 8-bin orientation histogram [9]. A 128 element descriptor vector is created using 8-bin orientation over 16 subregions.

In the proposed approach, the system localizes face in the thermal image and then computes descriptors for the face localized region. We call these descriptors as FLSIFT descriptors. The extracted descriptors for the facial images of all the subjects are stored and used for matching.

4.4 FLSIFT Descriptor Matching

The nearest neighbor distance metric is used for matching. Match is declared if and only if the Euclidean distance between the closed key point descriptor is less than 0.6 times the next closest key point descriptor. Figure 7 shows the FLSIFT feature matching between same subject and different subject.

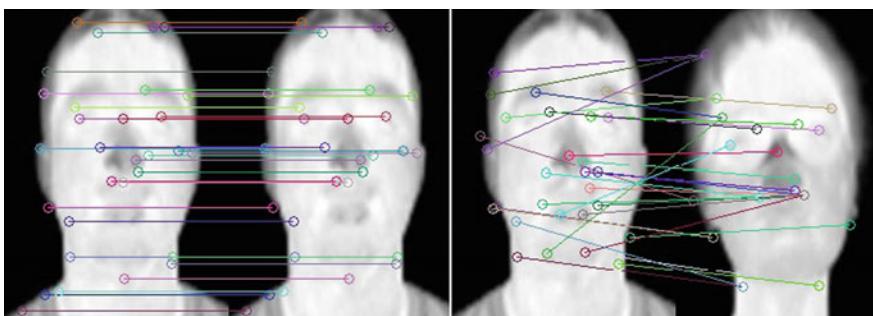


Fig. 7 FL-SIFT Feature Matching between same subject and different subject

5 Experimental Results and Discussion

The experiments are performed with following objectives, (1) To evaluate importance of FLSIFT over SIFT, (2) To compare results of proposed system with BVP [7] technique.

5.1 Recognition on ITFDB

From ITFDB dataset Images of randomly chosen 57 subjects for the training set and the images from the remaining 54 subjects are used for testing. The training set thus contains 109 images similarly; the testing set consists of 108 images. Further the testing set is divided into two parts gallery and probe.

For each subject, one complex background image is taken as gallery and one complex background image, and one simple background images are constitute the query set. For evaluating recognition experiment, the combination of train set and query set are tested as considering gallery set as trained set.

Using this configuration experimental results are reported in terms of receiver operating characteristics (ROC) curves. This experiment is carried out on ITFDB dataset mentioned above. Figure 8 shows the performance of proposed approach compared with SIFT operator on thermal face images from ITFDB. The ROC curves indicate that thermal image recognition does not get affected much due to background information in FLSIFT.

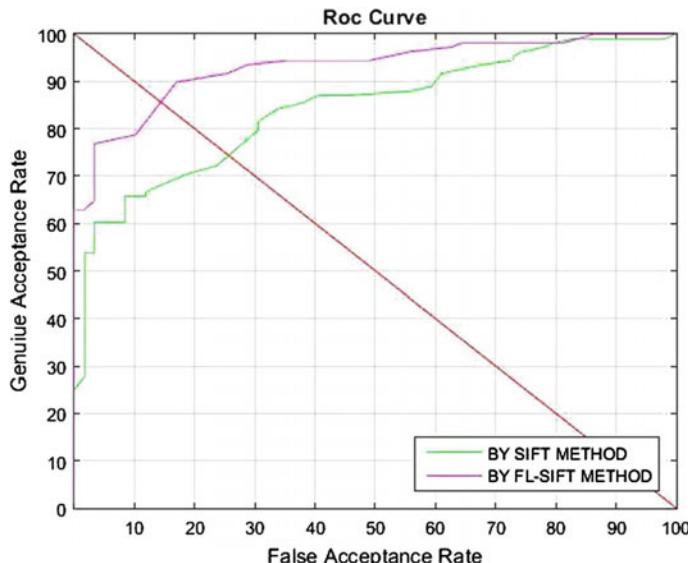


Fig. 8 ROC curve on ITFDB dataset

5.2 Identification

This experimentation deals with problem of face identification. For identification comparison of the proposed FLSIFT and BVP [7] technique is done on ITFDB dataset. The dataset consists of images with simple background as well as complex background. The comparative results on both the background are as follows:

Simple background.

In the proposed ITFDB dataset, there are 452 (113×4) images of 113 subjects with simple background, i.e. four images per subject. These images are divided into four subsets as, S-1, S-2, S-3, and S-4. To test the performance of FLSIFT and BVP technique [7], one subset is used for training (consisting of 113 images) and remaining sets are used for testing. The obtained results are as shown in Table 2.

Table 3 demonstrates the average identification accuracy for both the systems on the subsets S1-S4. The proposed FLSIFT gives 100 % recognition accuracy on both training and test sets. Whereas, recognition accuracy of BVP technique on test dataset is 79.28 and 100 % on trained dataset with signature.

By adding the four signature of same subject one unique template for individual subject is created. When all signature were tested using BVP considering this template as a training set we got 86.9 % recognition accuracy. It may be noted that with formation of templates for each face using four signatures of the same subject, performance of BVP technique is found to be improved as compared to its performance on signature of faces.

Table 2 Comparison over Different Training Set

Train set	Recognition technique	Test set				Average performance (Test set) (%)
		S-1 (%)	S-2 (%)	S-3 (%)	S-4 (%)	
S-1	BVP	100	85.84	73.45	66.37	75.22
	FLSIFT	100	100	100	100	100
S-2	BVP	85.84	100	90.26	79.64	85.24
	FLSIFT	100	100	100	100	100
S-3	BVP	79.64	87.61	100	80.53	82.60
	FLSIFT	100	100	100	100	100
S-4	BVP	69.91	73.45	78.76	100	74.04
	FLSIFT	100	100	100	100	100

Table 3 Comparison Result of Average Accuracy of Recognition

	BVP with template (%)	BVP with signature (%)	FLSIFT (%)
Train Set	86.9	100	100
Test Set	86.9	79.28	100

Table 4 Comparison Result of Accuracy of Recognition

	BVP [7] (%)	SIFT [9] (%)	FLSIFT (%)
Train Set	100	100	100
Test Set	5.55	81.48	98.14

Complex background.

Thermal images also have temperature gradient because of background. As discussed in Sect. 3, the background consists of different objects having different emissivity and reflection coefficients results in a disturbed thermal characteristic in the image. Experimental results show that FLSIFT method is sustainable with this type of complex background as well. The following Table 4. Shows the comparative results of FLSIFT, SIFT [9], and BVP systems.

The training set consists of 162 (54×3) images of 54 subjects with a combination of two images of simple background and one image from complex background per subject. During testing of the systems, a set of another 54 images with complex background is used.

Table 4 describes the detail comparison of average accuracy of the complex background for both the methods. It is observed that BVP technique performed satisfactorily on training dataset, however its performance on complex dataset is very poor. Note that proposed recognition rate of FLSIFT is 98.14 % even on complex background. To ensure the effect of face localization on performance, results without face localization are also reported in Table 4. This clearly shows importance of face localization.

6 Conclusion

This work evaluates the performance of existing thermal face recognition BVP technique. It is observed from experimental results that BVP performs poorly on complex thermal background. A dataset for thermal facial images is developed in Indian environmental conditions. To overcome the difficulty in face recognition in complex thermal background, a novel FLSIFT method is proposed. Its performance is compared empirically with BPV on simple and complex background. It may be concluded from this work that Human Thermal Face recognition is a good biometric for identification and proposed FLSIFT is found to be a suitable for the task.

References

1. Chen, X., Flynn, P.J. & Bowyer, K. W. PCA-based face recognition in infrared imagery: baseline and comparative studies. *2003 IEEE Int. SOI Conf. Proc. (Cat. No.03CH37443)* (2003).
2. Selinger, A. & Socolinsky, D. a. Face Recognition in the Dark. *2004 Conf. Comput. Vis. Pattern Recognit. Work.* 0–5 (2004).

3. Socolinsky, D. & Selinger, A. A Comparative Analysis of Face Recognition Performance with Visible and Thermal Infrared Imagery. *Proceedings. 16th Int. Conf. Pattern Recognition, 2002* **4**, 217–222 (2002).
4. Socolinsky, D.A. & Selinger, a. Thermal face recognition in an operational scenario. *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2004. CVPR 2004.* **2**, (2004).
5. Desa, S.M. & Hati, S. IR and visible face recognition using fusion of kernel based features. *2008 19th Int. Conf. Pattern Recognit.* 1–4 (2008).
6. Vogiannou, A. *et al.* Advances in Biometrics. *Adv. Biometrics* **5558**, pp. 838–846 (2009).
7. Guzman, A.M. *et al.* Thermal imaging as a biometrics approach to facial signature authentication. *IEEE J. Biomed. Heal. Informatics* **17**, 214–222 (2013).
8. Zhou, Q., Li, Z. & Aggarwal, J. K. Boundary extraction in thermal images by edge map. *Proc. 2004 ACM Symp. Appl. Comput. - SAC'04* 254 (2004).
9. Lowe D.G., Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, **60**(2), 91–110, (2004).
10. Perona, P. & Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 629–639 (1990).
11. Candocia, F. & Adjouadi, M. A similarity measure for stereo feature matching. *IEEE Trans. Image Process.* **6**, 1460–1464 (1997).
12. Indian Thermal Face Database: http://cvprlab-sggs.co.in/index_files/Page2141.htm.
13. Dhamecha, T.I., Nigam, A., Singh, R. & Vatsa, M. Disguise detection and face recognition in visible and thermal spectrums. *Biometrics (ICB), 2013 Int. Conf.* 1–8 (2013).
14. Dhamecha, T.I., Singh, R., Vatsa, M. & Kumar, A. Recognizing disguised faces: Human and machine evaluation. *PLoS One* **9**, (2014).
15. Dillencourt, M.B., Samet, H. & Tamminen, M. A general approach to connected-component labeling for arbitrary image representations. *J. ACM* **39**, 253–280 (1992).

Integrating Geometric and Textural Features for Facial Emotion Classification Using SVM Frameworks

Samyak Datta, Debasish Sen and R. Balasubramanian

Abstract In this paper, we present a fast facial emotion classification system that relies on the concatenation of geometric and texture-based features. For classification, we propose to leverage the binary classification capabilities of a support vector machine classifier to a hierarchical graph-based architecture that allows multi-class classification. We evaluate our classification results by calculating the emotion-wise classification accuracies and execution time of the hierarchical SVM classifier. A comparison between the overall accuracies of geometric, texture-based, and concatenated features clearly indicates the performance enhancement achieved with concatenated features. Our experiments also demonstrate the effectiveness of our approach for developing efficient and robust real-time facial expression recognition frameworks.

Keywords Emotion classification • Geometric features • Textural features • Local binary patterns • DAGSVMs

1 Introduction

This paper attempts to address the problem of enabling computers to recognize emotions from facial expressions in a fast and efficient manner. Emotion recognition is a challenging problem due to the high degree of variability in the emotions expressed through human faces. Extracting a subset of facial features that best captures this variation has been a long-standing problem in the Computer-Vision community. A basic expression recognition framework is expected to involve modules for detecting

S. Datta (✉) · R. Balasubramanian
Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee, India
e-mail: datta.samyak@gmail.com

D. Sen
Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology, Kharagpur, India

faces, deciding on an appropriate subset of features to best represent the face which involves a trade-off between accuracy of representation and fast computation and finally, classification of the feature vector into a particular emotion category.

In this paper, we propose a framework for performing fast emotion classification from facial expressions. Two types of features are extracted for each facial image frame: geometric and texture-based. Angles formed by different facial landmark points have been selected as geometric features which is a novel and speed optimized technique as compared to other expression recognition methods. Spatially enhanced, uniform pattern local binary pattern (LBP) histograms have been used as texture-based features. The hybrid feature vector for classification is then constructed by concatenating both the types of features. The concatenated features are able to capture both types of facial changes—high-level contortions of facial geometry (geometric features) and low-level changes in the face texture (texture-based features). In comparison with previous methods, our approach of using concatenated features results in enhanced performance. The classification module is based on support vector machines. One of the novelties of the work lies in the use of hierarchical SVM architectures to leverage the binary classification of SVMs to multi-class classification problems. The use of hierarchical SVMs results in much faster execution times than the traditional SVM-based multi-class classification approaches (such as one-vs-one SVMs) making the system suitable for real-time applications.

The remainder of the paper is structured as follows. Section 2 talks about the current state of the art in the field. Section 3 discusses the proposed architecture of the work in detail where both the feature extraction and classification phases of the emotion recognition system are explained. Section 4 consists of a discussion regarding the results obtained as a consequence of this work and finally and in Sect. 5, we conclude by discussing the relevance of our work in enhancing the state of the art.

2 Related Work

The state of the art in emotion classification can be broadly divided into two categories: (a) geometric feature based or (b) texture feature based.

Pantic and Rothkrantz [1] used a rule-based classifier on frontal-facial points to achieve an accuracy of 86 % on 25 subjects from the MMI database. Similar attempts by were made by Pantic and Patras in 2005 [2] where they tracked a set of 20 fiducial points and obtained an overall recognition rate of 90 % on the CK-database. Cohen et al. [3], 2003 extracted a vector of motion units using the PBVD tracker by measuring the displacement of facial points. More recently, Anwar et al. [4] in their 2014 paper use a set of eight fiducial points to achieve the state-of-the-art classification rates.

The texture-based methods involve techniques, such as local binary patterns (L.B.P.) or applying some image filters to either the entire facial image (global) or some parts (local). Zhang et al. [5] use LBP along with local fisher discriminant

analysis (LFDA) to achieve an overall recognition rate of 90.7 %. Although Gabor filters are known to provide a very low error rate, but it is computationally expensive to convolve a face image with a set of Gabor filters to extract features at different scales and orientations.

3 Proposed Architecture

In the proposed architecture, as shown in Fig. 1, both geometric and texture-based features have been used for classifying emotions. Two particular frames of interest from the extended Cohn-Kanade (CK+) database [6]—neutral and peak expression have been selected for each subject. The calculation of geometric features involves using both the frames whereas texture-based features only make use of the peak-expression image frame.

3.1 Geometric Features

Facial angles subtended by lines joining some key facial feature points have been selected as candidates for geometric features as shown in Fig. 2. Solely relying on facial angles as geometric features allows facial images of different sizes and orientations to be treated in a similar manner and removes the need for intra-face normalization of features.

The face detector proposed by Viola and Jones based on Haar cascades [7] has been applied to both the frames followed by the active shapes model (ASM) algorithm [8] to locate the 17 key facial points. Subsequently, 10 facial angles have been

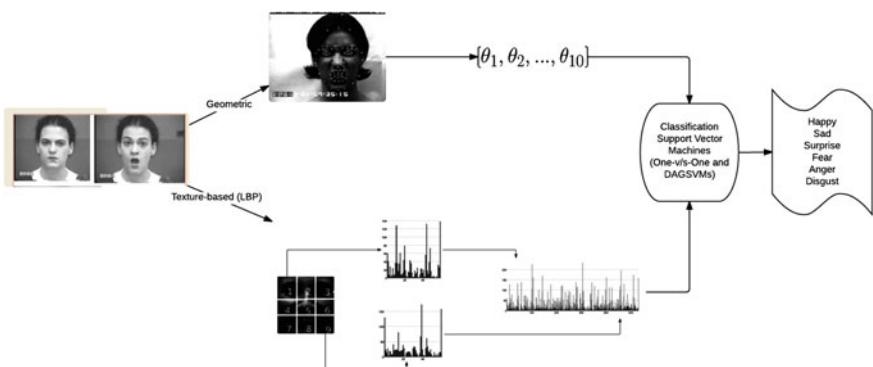
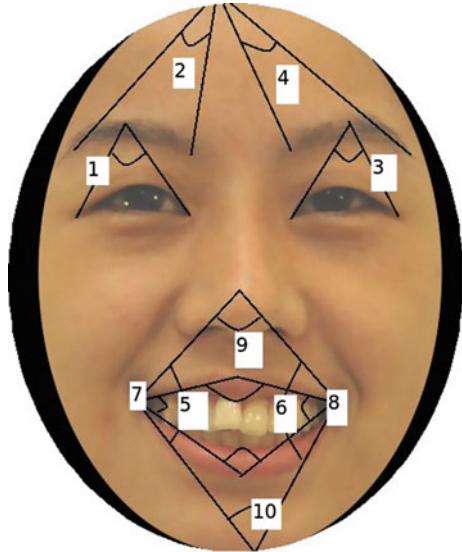


Fig. 1 A flowchart depicting the proposed architecture of our hybrid feature-based facial emotion classification system

Fig. 2 A frontal face image depicting the facial angles (numbered from 1 to 10) used as geometric features. These angles are computed from the lines joining the 17 key facial feature points as detected by A.S.M. algorithm



computed for each frame as shown in Fig. 2. The facial angles have been computed from the landmark points using the following set of basic coordinate geometry formulas.

Let $A(x_1, y_1), O(x_0, y_0)$ and $B(x_2, y_2)$ be three points in the two-dimensional Euclidean space. We define two vectors $\mathbf{OA} = (x_1 - x_0, y_1 - y_0)$ and $\mathbf{OB} = (x_2 - x_0, y_2 - y_0)$. The angle between \mathbf{OA} and \mathbf{OB} is given by:

$$\theta = \cos^{-1} \frac{\mathbf{OA} \cdot \mathbf{OB}}{|\mathbf{OA}| |\mathbf{OB}|} \quad (1)$$

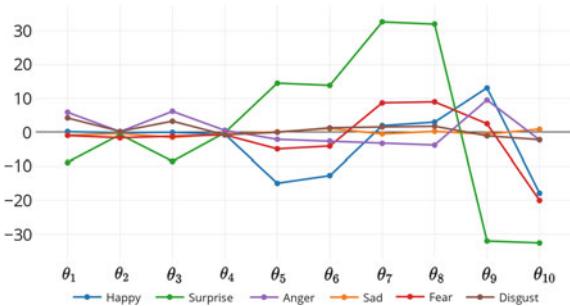
The difference in the values (in degrees) in the corresponding facial angles between the neutral and peak expressions serves as the 10 geometric features for the face which help capture the high-level distortions in the facial geometry across emotion classes.

3.2 Texture-Based Features

For texture-based feature extraction, local binary pattern histograms have been selected due to their simplicity, intuitiveness, and computational efficiency. The $LBP_{8,1}$ operator has been used in our experiments which essentially computes the LBP code for each pixel (x_c, y_c) using the expression:

$$LBP(x_c, y_c) = \sum_{n=0}^7 s(i_n - i_c) \cdot 2^n \quad (2)$$

Fig. 3 Variation of geometric features across the 6 emotion classes



where i_k is the gray-scale intensity value of the pixel with coordinates (x_k, y_k) and $s(x)$ is 1 if $x \geq 0$ and 0 otherwise.

A variant of the traditional LBP operator that has been used as part of this work is the uniform pattern LBP operator denoted by $LBP_{8,1}^{u2}$. Uniform patterns are those binary patterns that have at most 2 bit transitions when the 8-bit LBP code is interpreted as a circular string. For example, 11000010 is not a uniform pattern whereas 11110000 is (3 and 2 bit transitions respectively). Using only uniform patterns (all nonuniform pattern LBP codes are assigned to a single bin) brings down the number of histogram bins from 256 to merely 59. After computing the $LBP_{8,1}^{u2}$ codes for each pixel, a histogram of the LBP values is constructed.

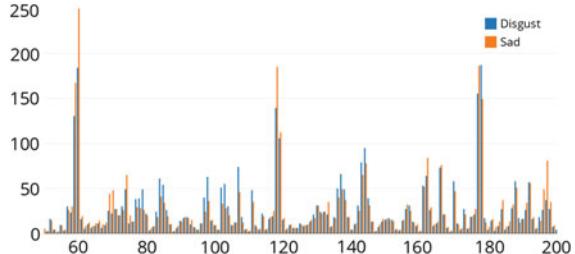
For computing texture-based features, all peak expression facial images are aligned and cropped to a uniform spatial resolution of 120×120 pixels and are then divided into nine equal sized blocks of 40×40 pixels each. The local $LBP_{8,1}^{u2}$ histogram is computed for each subimage and then concatenated into a global spatially enhanced uniform pattern LBP histogram for the facial image. The feature vector thus formed is of size $59 \times 9 = 531$.

3.3 Concatenated Features

The graph in Fig. 3 shows the variation in values of the 10 geometric features for each of the 6 emotion classes. From a visual inspection, it is evident that the facial feature angles do a good job in differentiating between classes such as “Happy,” “Surprise,” and “Fear.” However, “Disgust” and “Sad” are difficult to differentiate due to very low (almost nonexistent) interclass variance. The inability to completely capture variations between certain emotion classes arises due to the fact that these features only capture the high-level distortions in the facial geometry. Examples of such distortions would include the opening/closing of the mouth and widening of the eyes or curvature of lips. Hence, simply using geometric features is not sufficient to train a facial expression classifier with good discriminative powers.

On the other hand, the L.B.P. histograms contain information about the distribution of micro-patterns such as edges, flat areas and wrinkles which represent some

Fig. 4 Variation of LBP histogram features between “disgust” and “sad”



of the finer details of the face. To illustrate with an example, the histogram in Fig. 4 shows a comparative analysis between the spatially concatenated LBP histograms of the “disgust” and “sad” emotion classes. As noted earlier, geometric features were not able to capture the interclass variations between these two classes. However, through a simple visual inspection, we can see that there are significant differences in the values of the histogram bins (e.g., peaks near the bin numbers 60, 80 and 120) between the two classes. These differences help impart discriminative powers to the classifiers trained using these features.

Further, dividing the facial image into subimages and spatially concatenating their corresponding LBP histograms preserves the spatial texture-information while representing the global description of the face. The 531 uniform pattern $LBP_{8,1}^{u2}$ histogram features are concatenated with the 10 geometric facial feature angle features to form a combined, hybrid 541-dimensional feature vector. These hybrid vectors are used as a representative of the face for classification purposes.

3.4 Classification

Support vector machines (SVMs) are primarily binary classifiers that find the best separating hyperplane by maximizing the margins from the support vectors. Support vectors are defined as the data points that lie closest to the decision boundary. Some common SVM architectures, such as one-vs-one, one-vs-rest, and directed acyclic graph SVMs (DAGSVMs) [9] have been proposed to leverage the binary classification capabilities of an SVM classifier for multi-class classification problems.

In the one-vs-one scheme for an n -class classification problem, $\binom{n}{2}$ binary SVM classifiers are trained corresponding to each pair of classes. The test point is put across all $\binom{n}{2}$ SVMs and the winning class is decided on the basis of a majority vote. On the other hand, in a hierarchical SVM architecture such as DAGSVM, $\binom{n}{2}$ SVMs are trained, but the test point only has to go across $(n - 1)$ SVMs by traversing the directed graph as shown in Fig. 5. To evaluate the DAGSVM for a test point x , starting at the root node, a binary SVM is evaluated. Depending on the classification at this stage, the node is exited either via the left or the right edge. Then, the value of

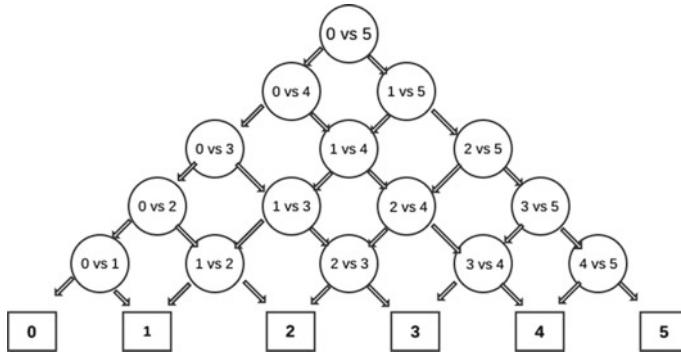


Fig. 5 Decision diagram for a 6-class DAGSVM

the binary SVM corresponding to the next node is evaluated until we reach one of the leaf nodes. The final classification is the class associated with the leaf node.

The statistical classification algorithm of SVMs is compared with an instance-based learning technique, the k-nearest neighbors (k-NN) classifier. The results are summarized in Tables 1 and 2. The motivation behind selection of the two classifiers lies in the fact that while SVMs construct an explicit model from the training data, instance-based techniques refrain from such generalizations. k-NN classifiers data points by comparing new problem instances with those seen in training.

Further, a comparison, in terms of both classification accuracy and execution times between two of the multi-class SVM frameworks: one-vs-one and DAGSVMs has been presented as part of this work. Since using DAGSVMs involves putting the

Table 1 A comparison of classification accuracies of different SVM architectures for different feature extraction techniques

Architecture	Classification accuracy (%)		
	Geometric	Texture (LBP)	Geometric + Texture
One-vs-One	78.15	88.52	91.85
DAGSVMs	76.67	86	89.26

Table 2 A comparison of classification accuracies of the k-NN algorithm for different values of k and feature extraction techniques.

k	Classification accuracy (%)		
	Geometric	Texture (LBP)	Geometric + Texture
3	73.528	64.7	69.93
5	75	67.65	68.29
7	75.29	66.01	67
9	76.76	63.4	64.37

test example through a lesser number ($n - 1$) of SVMs than one-vs-one SVMs ($\binom{n}{2}$), a drastic reduction in the execution time is expected.

4 Results and Discussions

The Cohn-Kanade extended (CK+) dataset consists of 593 image sequences (frames of a video shot) from 123 subjects, out of which 327 sequences are labeled as belonging to one of the seven emotion categories: happy, sad, surprise, fear, anger, disgust, and contempt. Labeled facial expression images (neutral and peak frames) for the six basic emotions—happy, sad, surprise, fear, anger, and disgust have been used in the tests. All the results are tenfold cross validated.

The classification accuracies for the various SVM architectures for geometric, texture-based, and hybrid features are summarized in Table 1. For benchmarking purposes, the classification accuracies for SVM-based classifiers have been compared with those of the k-Nearest neighbor algorithm for different values of k . The results are reported in Table 2.

It is clear that irrespective of the SVM classifier architecture used, there is a significant enhancement in the overall classification accuracies with our approach of using hybrid features in place of simply using geometric or LBP features. For example, a one-vs-one SVM classifier gives an overall recognition rate of 78.15 and 88.52 % with only geometric and LBP-based features respectively, which increases to 91.85 % when using a hybrid feature set.

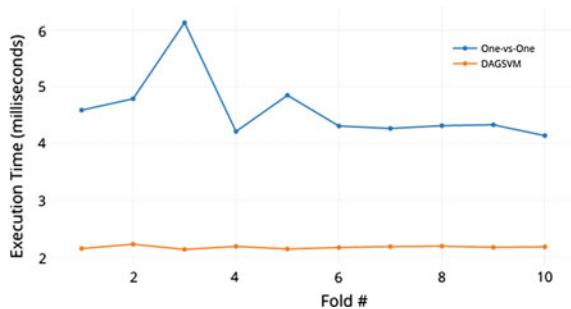
In the case of k-NN classifiers, there is a sharp decrease in the classification accuracies as we move from geometric to texture (LBP)-based or hybrid features due to the increase in the dimensionality of feature vectors. The geometric feature vector has 10 attributes which increases to 531 and 541 in the case of texture and hybrid features respectively. This demonstrates the inability of the k-NN classifier to work well in high-dimensional feature spaces. However, when the recognition rates of texture-based and hybrid features are compared, both of which are high dimensional, we see the hybrid feature vectors outperforming again as evident in Table 2.

The confusion matrices for the 6-class classification problem using our approach of hybrid-features is shown in Table 3 for both one-vs-one and directed acyclic graph SVMs (DAGSVMs). In both cases, it is evident that “happy” and “surprise” are easiest, whereas “sad” and “fear” are the most difficult to classify.

A comparison of the execution times of one-vs-one and DAGSVMs averaged over 30 test samples and tenfolds in Fig. 6 clearly shows the gain in computational efficiency in terms of time while using DAGSVMs. The reported execution times are on an Intel Core(TM) i3-2330M CPU with a clock speed of 2.20 GHz. Since the overall classification accuracy for DAGSVMs (89.26 %) is not significantly less than one-vs-one SVM classifiers (91.85 %), the reduction in time may render DAGSVM as a suitable candidate for real-time emotion classification problems.

Table 3 Confusion matrix for SVM classification using hybrid features

Actual class	Predicted class					
(a) One-vs-One SVMs						
	Happy	Sad	Surprise	Fear	Anger	Disgust
Happy	98.53	0	0	0	0	1.47
Sad	0	58.82	0	0	35.29	5.88
Surprise	0	0	95.83	1.43	0	2.86
Fear	6.25	6.25	6.25	75	6.25	0
Anger	0	5.71	0	0	91.43	2.86
Disgust	0	3.28	0	0	1.64	95.08
(b) DAGSVMs						
	Happy	Sad	Surprise	Fear	Anger	Disgust
Happy	100	0	0	0	0	0
Sad	0	50	15	0	35	0
Surprise	1.43	0	94.29	1.43	0	2.86
Fear	10.53	0	5.26	78.94	5.26	0
Anger	0	7.89	0	0	86.84	5.26
Disgust	0	1.72	1.72	0	6.89	89.65

Fig. 6 Comparison between the total execution time of one-vs-one (blue) and DAGSVMs (orange) for the classification of 30 test points

5 Conclusion

In this paper, we have presented a framework for fast emotion classification from facial expression.

Our experimental results show an enhanced performance when using a concatenated feature vector which is a combination of geometrical and texture-based LBP features. We also present a comparative analysis of two major multi-class, SVM-based architectures for classification, namely one-vs-one and DAGSVMs (hierarchical SVMs). Our results indicate that both the systems give almost equal performance. However, using hierarchical multi-class SVM architectures leads to increased efficiency in terms of computation time.

References

1. M. Pantic and J.M. Rothkrantz, *Facial Action Recognition for Facial Expression Analysis from Static Face Images*, IEEE Trans. Systems, Man and Cybernetics Part B, vol. 34, no. 3, pp. 1449–1461, 2004.
2. M. Pantic, I. Patras, *Detecting Facial Actions and their Temporal Segments in Nearly Frontal-view Face Image Sequences*, Proc. IEEE conf. Systems, Man and Cybernetics, vol. 4, pp. 3358–3363, Oct 2005.
3. I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, *Facial Expression Recognition From Video Sequences: Temporal and Static Modeling*, Computer Vision and Image Understanding, vol. 91, pp. 160–187, 2003.
4. A. Saeed, A. Al-Hamadi, R. Niese, and M. Elzobi, *Frame-Based Facial Expression Recognition Using Geometrical Features*, Advances in Human-Computer Interaction, vol. 2014, April 2014.
5. S. Zhang, X. Zhao and B. Lei, *Facial Expression Recognition Based on Local Binary Patterns and Local Fisher Discriminant Analysis*, WSEAS Transactions on Signal Processing, issue 1, vol. 8, pp. 21–31, Jan 2012.
6. P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and Matthews, *The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression*, IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010), 2010.
7. P. Viola, M.J. Jones, *Robust Real-Time Face Detection*, International Journal of Computer Vision, 2004.
8. S. Milborrow and F. Nicolls, *Active Shape Models with SIFT Descriptors and MARS*, International Conference on Computer Vision Theory and Applications (VISAPP), 2014.
9. J.C. Platt, N. Cristianini and J.S. Taylor, *Large Margin DAGs for Multiclass Classification*, Advances in Neural Information Processing Systems (NIPS), 1999.

Fast Non-blind Image Deblurring with Sparse Priors

Rajshekhar Das, Anurag Bajpai and Shankar M. Venkatesan

Abstract Capturing clear images in dim light conditions remains a critical problem in digital photography. Long exposure time inevitably leads to motion blur due to camera shake. On the other hand, short exposure time with high gain yields sharp but noisy images. However, exploiting information from both the blurry and noisy images can produce superior results in image reconstruction. In this paper, we employ the image pairs to carry out a non-blind deconvolution and compare the performances of three different deconvolution methods, namely, Richardson Lucy algorithm, Algebraic deconvolution, and Basis Pursuit deconvolution. We show that the Basis Pursuit approach produces the best results in most cases.

Keywords Basis pursuit • Richardson Lucy • Non-blind deconvolution • Compressed sensing • Deblurring

1 Introduction

The advent of digital optics has led to tremendous advancements in the camera technology. However, capturing clear images in dim light conditions still remains a critical problem in digital photography. To handle this problem, one has to carefully adjust the three parameters related to camera exposure, i.e., exposure time, aperture size, and sensor sensitivity (ISO). Using a low camera shutter speed, one can increase the exposure time and hence enhance the sensor illumination. However, long exposures in hand-held cameras generally lead to motion blur due to camera

R. Das (✉) · A. Bajpai · S.M. Venkatesan
Samsung R & D Institute, 2870, Phoenix Building,
Bagmane Constellation Business Park, Bengaluru, India
e-mail: rajshekhar.d@samsung.com

A. Bajpai
e-mail: anurag.bajpai@samsung.com

S.M. Venkatesan
e-mail: s.venkatesan@samsung.com

shake. Large aperture size, on the other hand, decreases the depth of field causing depth-dependent blur. High ISO setting corresponding to high gain not only boosts image contrast but also amplifies the sensor noise, resulting in a noisy image. The undesired blurry/noisy photograph so obtained also has a degrading effect in further process such as feature extraction or object recognition. Therefore, effective removal of motion blur holds significant importance in the area of computer vision.

It has been previously shown in [1] that if we have blurry/noisy image pairs instead of a single image, it is possible to recover high-quality images. Such methods exploit the sharp details of a noisy image and the correct color intensity and SNR of a blurry image to produce the optimum results. To do so, deblurring is modeled as a non-blind deconvolution process wherein the blur kernel is estimated prior to image deconvolution. The underlying assumption to this process is that the image scene is static in nature. In this paper, we particularly focus on the image deconvolution techniques. Image deconvolution is a standard example of ill-posed inverse problems which require suitable constraints or prior knowledge for successful recovery. In the recent past, there has been a considerable interest in compressive sensing techniques like basis pursuit [2–5], which incorporate sparsity priors instead of standard smoothing constraints. These priors enforce the sparseness of natural signals in some suitable domain. The most popular of them is the now celebrated convex L_1 prior, also known as basis pursuit denoising. We demonstrate that image deconvolution with noisy/blurry image pairs can be accommodated under such optimization schemes to give superior results than Algebraic [6] or RL [1] deconvolution approaches.

2 Previous Work

Most recent motion deblurring techniques can be categorized into two basic classes: Single-image deblurring and multi-image deblurring. Based on the fundamental assumptions of blurring process, they can be further distinguished into spatially invariant point spread function (PSF) estimation and spatially variant PSF estimation. In single-image deblurring, Fergus et al. [7] proposed a variational Bayesian approach to estimate the blur kernel by maximizing the marginal probability. Levin et al. [8] proposed an improved efficient marginal likelihood approximation. Detailed analysis of the issues of the MAP problem in motion deblurring was provided in [9]. Several methods [10–12] followed the line of altering the traditional MAP framework to estimate the blur kernel and latent image iteratively, introducing a two-phase (blur kernel and latent image) estimation. In those semi-blind or non-blind deconvolutions, they discourage the trivial delta kernel and sharp image solution in each phase by either explicitly reweighting strong edges [10], predicting sharp edges using different priors [11–15], or directly imposing normalized sparsity measures on image gradients [16]. The gradient sparsity prior was earlier used by a lot of work including Levin et al. [14], calculating iteratively reweighted least squares (IRLS) to regularize results for images exhibiting defocus

blur and motion blurred images. Krishnan et al. [16] analyze all the present priors and introduce a novel sharp favorite prior for deconvolution. Also, variable substitution schemes [15] were employed to constrain the deconvolution solution.

As for detailed recovery, Yuan et al. [17] proposed a multi-scale approach to progressively recover blurred details, while Shan et al. [10] introduced regularization based on high-order partial derivatives to reduce image artifacts. Meanwhile, Raskar et al. [18] coded the exposure to make the PSF more suitable for deconvolution. Jia [19] demonstrated how to use an object's alpha matte to better compute the PSF. However, all these methods above assume that the blur kernel is spatially invariant. Due to the deconvolution problem being severely ill-posed, more information is required and multi-image methods were proposed in a lot of literature. Ben-Ezra and Nayar [20] attached a video camera to a conventional high-resolution still camera to facilitate PSF estimation. The hybrid camera system was extended by Tai et al. [3] to compute general motion blur with optical flow. In [1], Yuan et al. took an additional noisy-unblurred image to form a noisy/blurred image pair, making PSF estimation robust. Assuming constant velocity camera motion, Li et al. [21] used a set of motion blurred frames to create a deblurred panoramic image. One of the more robust MAP-based frameworks which employ image pairs for deblurring is described in Tico et al. [36]. They use Weiner filter-based kernel estimation for an initial estimate, suggesting the practicality of the filter in determining simple kernels. In one of the recent works, Cho et al. [22] explicitly handle outliers based on the deconvolution process.

3 Problem Formulation

We have the following pair of images at our disposal—a noisy and underexposed image N (captured with high shutter speed and high ISO) and a blurred image B (captured with slow shutter speed and low ISO). The exposure difference between B and N is accounted for by pre-multiplying the noisy image by a factor of $\frac{ISO_R \Delta t_R}{ISO_N \Delta t_N}$ in the irradiance space. Under the assumption that the blur kernel is spatially invariant, motion blur can be modeled as follows:

$$B(\mathbf{x}) = (I \otimes K)(\mathbf{x}) + \eta(\mathbf{x}) \quad (1)$$

where I is the original image to be estimated from B and the blur kernel K . The term $\eta(\mathbf{x})$ represents zero mean, independent and identically distributed noise at every pixel $\mathbf{x} = (x, y)$. In practice, the variance of $\eta(\mathbf{x})$ is usually much smaller than the noise variance in N . In our case of non-blind deblurring, we first compute the blur kernel which is then followed by an image deconvolution with the now known kernel. Being iterative in nature, the kernel estimation can benefit from a good initial estimate. This is provided by a denoised version, N_D , of the noisy image. A sophisticated denoising program [23–25] can preserve most of the power



Fig. 1 a, d Noisy image. b, e Wavelet-based denoising. c, f NLM denoising

spectrums of the actual image to yield a fairly accurate estimate. In this paper, we have applied a fast non-local means (NLM) denoising algorithm [24] which generally produces better results than most other state-of-the-art algorithms like [23, 25]. Figure 1 demonstrates the superiority of the NLM denoising over wavelet-based denoising. The input parameters for the denoising programs have been manually adjusted to obtain the best balance between noise removal and detail preservation.

4 Kernel Estimation

The nature of the blur model allows it to be expressed as a simple system of linear equations given by

$$\mathbf{b} = \mathbf{A}\mathbf{k} \quad (2)$$

where \mathbf{b} and \mathbf{k} are the linearized vectors of B and K , respectively, and \mathbf{A} is the corresponding matrix form of I . The kernel estimation is modeled as a linear least



Fig. 2 **a** Estimated kernels at each level (rightmost being the finest) of the scale space. **b** True blur kernel

squares problem with Tikhonov regularization to stabilize the solution. Incorporating nonnegativity and energy preservation constraints for the blur kernel, the optimization problem can be written as

$$\begin{aligned} \min_k \|b - Ak\|^2 + \lambda^2 \|k\|^2 \\ \text{s.t. } k_i \geq 0, \sum_i k_i = 1 \end{aligned} \quad (3)$$

To solve this, we use Landweber's method integrated with hysteresis thresholding in scale space [1] with a minor modification of using a dynamic step size instead of a constant value (as mentioned in [1]). In the scale space implementation, we use $1/\sqrt{2}$ as the downsampling factor, with a kernel size of 9×9 in the coarsest level. In our case, we choose the lower and higher hysteresis thresholds as 0.23 and 0.28, respectively (Fig. 2).

5 Deconvolution

Having estimated the blur kernel K , we are now ready to reconstruct the original image through various deconvolution techniques. Consider the denoising of the noisy image which yields N_D with some loss in information. The loss in detail layer is represented as a residual image ΔI

$$I = N_D + \Delta I \quad (4)$$

It is important to note here that for deconvolution, we estimate ΔI instead of I from a residual form of the blur model by substituting Eq. (4) in Eq. (1) to give

$$\Delta B = \Delta I \otimes K \quad (5)$$

where $\Delta B = B - N_D \otimes K$ represents the residual blurred image. By the virtue of their relatively small magnitudes, the introduction of residual image quantities helps reduce the effect of Gibbs phenomena observed at the edges. This, in turn, dampens the ringing artifacts. Once ΔI has been estimated, the final image can simply be recovered from Eq. (4). We now discuss the mathematical formulations which govern the three different deconvolution approaches.

5.1 Richardson Lucy (RL)

The RL algorithm [26] is a Bayesian iterative method which imposes nonnegativity constraints on the pixel values. The residual image estimated after each iteration is given by

$$\Delta I_{n+1} = \left(K^* \frac{\Delta B + 1}{(\Delta I_n + 1) \otimes \mathbf{K}} \right) \cdot (\Delta I_n + 1) - 1 \quad (6)$$

Here, ' $*$ ' denotes correlation operator and ' \cdot ' denotes element-wise multiplication. Since all the residual images are normalized to a range of $[0, 1]$, they have been offset by a constant 1, i.e., $\Delta B \rightarrow \Delta B + 1$ and $\Delta I \rightarrow \Delta I + 1$. We restrict the number of RL iterations to about 15 so as to prevent ringing from excess iterations.

In the RL output, high-frequency information tends to get lost due to inevitable smoothing. To revive the fine scale detail layer I_D , a high-pass filter is applied to the output as shown below:

$$I_D = I - F(I) \quad (7)$$

where $F(\cdot)$ is a low-pass filtering operation. The filter used here is an edge preserving bilateral filter [27] expressed as

$$F(I(x); I_g) = \frac{1}{C_x} \sum_{\tilde{x} \in \omega(x)} G_d(x - \tilde{x}) G_r(I(x) - I_g(\tilde{x})).J_{\tilde{x}} \quad (8)$$

where G_d and G_r are the domain and range Gaussian kernels, respectively, with σ_d and σ_r as the corresponding standard deviations. Also, C_x denotes the normalization constant and $\omega(x)$ denotes the neighboring window. We typically choose $\sigma_d = 0.5$ and $\sigma_r = 1$.

5.2 Algebraic Deconvolution

Algebraic deconvolution is based on standard matrix inversions to reconstruct the original image. The convolution of image I with the blur kernel k can be written as

$$\mathbf{b} = \mathbf{K}\mathbf{i} \quad (9)$$

where \mathbf{b} and \mathbf{i} are the linearized vectors of B and I , respectively, and \mathbf{K} is the corresponding blur matrix. A naïve way to deconvolve would be to pre-multiply \mathbf{b} in Eq. (9) with the inverse of \mathbf{K} . However, in practical scenarios, the blurred image is usually corrupted with some noise \mathbf{n} , such that $\mathbf{b} = \mathbf{b}_{\text{exact}} + \mathbf{n}$. Now, if the SNR of the captured blurry image is low, then the naïve approach can lead to highly

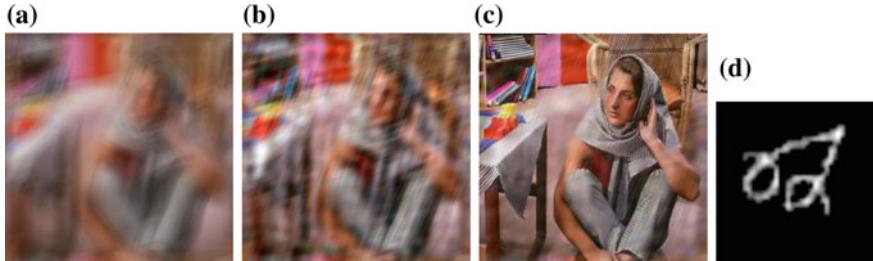


Fig. 3 **a** Blurry image. **b** Deconvolution without the residual image concept. **c** Deconvolution with the residual image concept. **d** Estimated kernel

degraded results. This is due to the dominance of the inverted noise term, $\mathbf{K}^{-1}\mathbf{n}$ over the actual image term $\mathbf{K}^{-1}\mathbf{b}_{\text{exact}}$. A detailed analysis of this phenomena has been provided in [6] based on the SVD of the blur matrix

$$K = U \sum V^T \quad (10)$$

where U and V are the orthogonal matrices, and \sum represents the singular matrix with σ_i as the i th singular value. By truncating \sum to a dimension $k < N$, one can reduce the noise perturbation in the deconvolved result. The truncation is done according to generalized cross-validation method (described in [6]). In our implementation, we choose a value of 0.4 for the regularization parameter. In Eq. (9), if we assume periodic extension of the original image beyond its boundary, then the blur matrix K can be shown to be block circulant with circulant blocks (BCCB) [6]. It is well known that the SVD of a BCCB matrix can be expressed in terms of the DFT matrix W as

$$K = W^H \Lambda W \quad (11)$$

where Λ is the singular value matrix. The implementation of the truncated SVD (TSVD) method primarily involves simple fft2 and iff2 operations which leads to significant speed up in the deblur process. The TSVD approach also suffers from the Gibbs phenomena. So replacing the actual images with their residual forms helps alleviate the results and suppress the ringing to some extent. This is demonstrated in Fig. 3.

5.3 Basis Pursuit Deconvolution (BPD)

Non-blind deconvolution can also be modeled as a convex optimization problem. However, the ill-posed nature of this inverse problem calls for a good regularization

scheme. The standard Tikhonov regularization, which imposes a quadratic penalty, assumes the images to be smooth and piecewise constant. This leads to a loss of high-frequency information in the image. Sophisticated regularization schemes based on shape parametrization [28–30] can improve the results but at the cost of increased computational complexity.

Basis pursuit (BP), on the other hand, is more robust than the Tikhonov method in terms of suppressing noise while preserving structure. Basis pursuit searches for signal representations in overcomplete dictionaries (in our case, the blur matrix K) by convex optimization: it obtains the decomposition that minimizes the l_1 norm of the coefficients occurring in the representation. BP is solved using linear programming techniques. Recent advances in large-scale linear programming have made it possible to solve BP optimization problem in nearly linear time. Thus, it is also computationally less intensive than the shape-based regularization methods.

Basis pursuit deconvolution is based on the assumption that natural images when resolved in some suitable domain can have a sufficiently sparse representation. This prior is exploited by a penalty function based on the l_1 norm of the residual image. The optimization statement then is given by

$$\min_{\Delta i} \|\Delta b - K\Delta i\|_2^2 + \lambda^2 \|\Delta i\|_1^2 \quad (12)$$

where Δb and Δi are the linearized vectors of the residual blurred image and the image to be reconstructed and λ is the regularization constant. In general, the assumption that Δi is sparse might not always hold true. However, it is well known that natural images have a sparse representation in the Fourier basis. Thus, to ensure the validity of the sparsity constraint, we estimate its Fourier coefficients instead. Let the Fourier coefficients of Δi be given by $\Delta i_c = W\Delta i$, where W represents the DFT matrix. Then, Eq. (12) can be rewritten as

$$\min_{\Delta i} \|\Delta b - K\Delta i_c\|_2^2 + \lambda^2 \|\Delta i\|_{c1}^2 \quad (13)$$

where $K = KW^H$ represents the modified blur matrix and W^H is the Hermitian conjugate of the DFT matrix. The above objective function can be minimized using split augmented Lagrangian shrinkage algorithm (SALSA) [31–33]. The SALSA algorithm is known to have a high convergence speed among all existing l_1 norm-based algorithms, enabled via variable splitting of minimization problem [34]. This convergence is achieved using an alternating direction method of multipliers (ADMM), which is based on augmented Lagrangian method (ALM) [31, 32]. The details of the algorithm are given in [31, 32].

The SALSA program in its standard form is slow for real-time applications. An efficient way of implementing would be to introduce SVD in the original blur matrix K (described below). The use of SVD to accelerate SALSA has been

previously proposed in [35]. Ours is similar to this method. However, the novelty in our implementation lies in the fact that we exploit the BCCB property of the blur matrix to obtain a spectral decomposition. This allows us to replace the standard SALSA with simple fft2 and ifft2 operations, thus resulting in a faster implementation as described below.

Algorithm Our proposed method based on SVD of the blur matrix

Aim: Estimation of Δi_c in eq. (10)

Input: Λ , Δb , λ , α , N_{it}

Output: Δi_c

Initialize: $d = 0$

$$1. \quad \Delta i_c = \Lambda \text{fft2}(\Delta b)$$

For $k = 1, 2, \dots, N_{it}$

$$2. \quad v = \text{soft}(\Delta i_c + d, \frac{5\lambda}{\alpha}) - d.$$

$$3. \quad \Delta i_c = (\frac{\Lambda}{\Lambda^2 + \alpha I}) \text{fft2}(\Delta b) + (\frac{\alpha}{\Lambda^2 + \alpha I}) v$$

$$4. \quad d = \Delta i_c - v$$

End

In the above algorithm, $\text{soft}(x, T) = \text{sign}(x) \cdot \max(|x| - T, 0)$. Once the frequency image Δi_c is estimated, the original image can be obtained by a simple inverse Fourier transform.

6 Results and Conclusions

In this section, we discuss the image deblurring results for various datasets. Figures 4 and 5 (cutout from the images in [1]) show the estimated blur kernel along with the deconvolved results corresponding to the three different approaches. The kernel sizes in Fig. 4 and Fig. 5 are 31×31 and 45×45 , respectively. The noisy image in Fig. 4 is captured using compact cameras, and hence is corrupted by very strong noise. Based on visual perception, one can see that BPD produces the sharpest of all images. Also, the encircled areas in the figures provide a visual comparison of the amount of ringing due to each method. Basis pursuit is most effective in suppressing the artifacts. We also carry out the deblurring experiment on a synthetic dataset (shown in Fig. 6). We blur the ground truth image with the kernel (e) and add the more appropriate ‘Poisson’ noise to obtain (a) and (b), respectively. We can observe that algebraic deconvolution has the worst performance in terms of ringing. This is strongly validated by the corresponding PSNR value. However, the algebraic method has a faster execution time than the RL algorithm. Basis pursuit, on the other hand, outperforms both these algorithms in terms of PSNR and has a

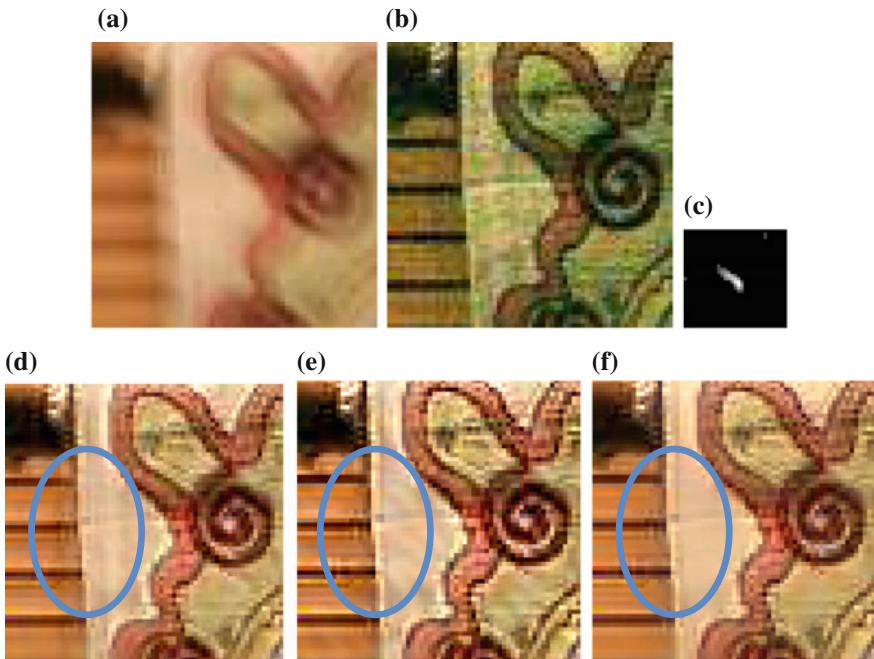


Fig. 4 Example A: **a** Blurred image. **b** Noisy image after exposure correction. **c** Estimated blur kernel. **d** RL output. **e** Algebraic deconvolution. **f** BPD output

generally faster execution time. In Fig. 6, the PSNR values of the deblurred images are 24.42 for RL method, 22.69 for algebraic method, and 26.40 for BPD method. Further, a quantitative comparison of the computational complexities of each of these methods is given in Table 1 which summarizes the execution times for each method with various resolutions.

The kernel estimation technique discussed in this paper has the surprising ability to tackle various nonlinear kernels with a fairly uniform performance across all datasets. However, we believe that for some highly complex kernels corresponding to sudden and large camera motions, the kernel estimation might not yield the best results. In such a scenario, having additional noisy images at various stages of signal integration might help guide the estimation process to produce superior results. Thus, one of the future directions for our work would be to formulate a robust optimization problem to handle such complexities.

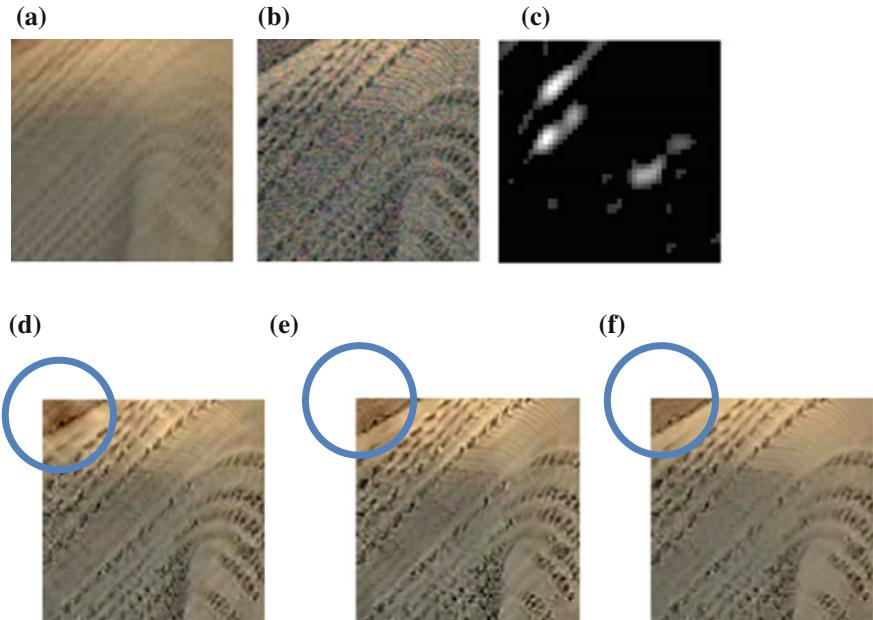


Fig. 5 Example B: **a** Blurred image. **b** Noisy image after exposure correction. **c** Estimated blur kernel. **d** RL output. **e** Algebraic deconvolution. **f** BPD output

Table 1 Comparison of the execution times for each method with various resolutions

Images with various resolutions	Execution times of the deconvolution methods (in s)		
	RL	Algebraic	BPD
Figure 3 (129×129)	1.24	1.22	1.18
Figure 4 (267×267)	1.70	1.60	1.40
Figure 5 (511×511)	6.63	4.75	5.09

Acknowledgments We profusely thank Prof. Phaneendra K. Yalavarthy for his innumerable suggestions on the algebraic and the basis pursuit methods and for his theoretical insights in these topics. We also thank the Multimedia team and Nitin for sharing the modified kernel estimation code based on Landweber's method but with dynamic step size.



Fig. 6 **a** Blurred image. **b** Input noisy image. **c** Denoised image after exposure correction in noisy image. **d** Estimated kernel. **e** Actual kernel. **f** RL output (**PSNR = 24.42**). **g** Algebraic deconvolution (**PSNR = 22.69**). **h** BPD output (**PSNR = 26.40**). **i** Original image

References

- Yuan, L., Sun, J., Quan, L., Shum H.-Y.: Image deblurring with blurred/noisy image pairs. *ACM Transactions on Graphics* 26, 3, 1 (2007)
- Candes, E., Wakin, M.: An introduction to compressive sampling. *IEEE Signal Process. Mag.* 25, 2, 21–30 (2008)
- Romberg, J.: Imaging via compressive sampling. *IEEE Signal Process. Mag.* 25, 2, 14–20 (2008)
- Cotter, S. F., Rao, B. D., Engan, K., Kreutz-Delgado, K.: Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Process.* 53, 7, pp. 2477–2488 (2005)
- Malouf, D., Cetin, M., Willsky, A.: A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Process.* 53, 8, 3010–3022 (2005)
- Hansen, P. C., Nagy, J. G., O’Leary, D. P.: *Deblurring Images: Matrices, Spectra, and Filtering*. Society for Industrial and Applied Mathematics, Philadelphia (2006)
- Fergus, R., Singh, B., Hertzmann, A., Roweis, S. T., Freeman, W. T.: Removing camera shake from a single photograph. *ACM Trans. Graph.* 25, 3(2006)
- Levin, A., Weiss, Y., Durand, F., Freeman, W. T.: Efficient marginal likelihood optimization in blind deconvolution. *CVPR* (2011)
- Levin, A., Weiss, Y., Durand, F., Freeman, W. T.: Understanding and evaluating blind deconvolution algorithms. *CVPR* (2009)
- Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. *ACM Trans. Graph.* 27(3) (2008)
- Cho, S., Lee, S.: Fast motion deblurring. *ACM Trans. Graph.* 28(5) (2009)
- Joshi, N., Szeliski, R., Kriegman, D. J.: Psf estimation using sharp edge prediction. *CVPR* (2008)
- Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. *ECCV* (1) (2010)

14. Levin, A., Fergus, R., Durand, F., Freeman, W. T.: Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph.* 26(3) (2007)
15. Krishnan, D., Fergus, R.: Fast image deconvolution using hyper-laplacian priors. *NIPS*, (2009)
16. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. *CVPR* (2011)
17. Yuan, L., Sun, J., Quan, L., Shum, H.-Y.: Progressive interscale and intra-scale non-blind image deconvolution. *ACM Trans. Graph.* 27(3) (2008)
18. Agrawal, A., Raskar, R.: Resolving objects at higher resolution from a single motion-blurred image. *CVPR* (2007)
19. Jia, J.: Single image motion deblurring using transparency. *CVPR* (2007)
20. Ben-Ezra, M., Nayar, S.: Motion-based Motion Deblurring. *IEEE Trans. PAMI* 26(6), 689–698, (2004)
21. Li, Y., Kang, S. B., Joshi, N., Seitz, S., Huttenlocher, D.: Generating sharp panoramas from motion blurred videos. *CVPR* (2010)
22. Cho, S., Wang, J., Lee, S.: Handling Outliers in NonBlind Image Deconvolution. *ICCV* (2011)
23. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans Image Processing* (2007)
24. Buades, A., Coll, B., Morel, J.M.: A non local algorithm for image denoising. *IEEE Computer Vision and Pattern Recognition*, 2, 60–65 (2005)
25. Portilla, J., Strela, V., Wainwright, M., Simoncelli, E. P.: Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans. on Image Processing* 12, 11, 1338–1351 (2003)
26. Richardson, W. H.: Bayesian-based iterative method of image restoration. *JOSA*, A 62, 1, 55–59 (1972)
27. Petschnigg, G., Agrawala, M., Hoppe, H., Szeliski, R., Cohen, M., and Toyama, K.: Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.* 23, 3, 664–672 (2004)
28. Boverman, G., Miller, E. L., Brooks, D. H., Isaacson, D., Fang, Q., Boas, D. A.: Estimation and statistical bounds for three-dimensional polar shapes in diffuse optical tomography. *IEEE Trans. Med. Imag.* 27, 6, 752–765 (2008)
29. Chen L. Y., Pan M. C., Pan, M. C.: Implementation of edge preserving regularization for frequency-domain diffuse optical tomography. *Appl. Opt.* 51, 43–54 (2012)
30. Chen, L. Y., Pan M. C., Pan, M. C.: Flexible near-infrared diffuse optical tomography with varied weighting functions of edge-preserving regularization. *Appl. Opt.* 52, 1173–1182 (2013)
31. Figueiredo, M., Bioucas-Dias, J., Afonso, M.: Fast frame-based image deconvolution using variable splitting and constrained optimization. *IEEE Workshop on Statistical Signal Processing* 109–112 (2009)
32. Afonso, M. V., Bioucas-Dias, J. M., Figueiredo, M. A. T.: Fast image recovery using variable splitting and constrained optimization. *IEEE Trans. Image Process.* 19, 9, 2345–2356, (2010)
33. Chambolle, A.: An algorithm for total variation minimization and applications. *J. Math. Imag. Vis.* 20, 89–97 (2004)
34. Lee, O., Ye, J. C.: Joint sparsity-driven non-iterative simultaneous reconstruction of absorption and scattering in diffuse optical tomography. *Opt. Exp.* 21, 26589–26604 (2013)
35. Prakash, J., Dehghani, H., Pogue, B. W., Yalavarthy, P. K.: Model-Resolution-Based Basis Pursuit Deconvolution Improves Diffuse Optical Tomographic Imaging. *IEEE Trans. on Medical Imaging* 33, 4, (2014)
36. Tico, M., Vehvilainen, M.: Estimation of motion blur point spread function from differently exposed image frames. *IEEE Signal Processing Conference*, (2006)

Author Index

A

- Agrawal, Subhash Chand, 285
Ahmed, Arif, 261, 273
Ahmed, Mushtaq, 149
Akula, Aparna, 297
Alom, Md. Zahangir, 409
Alpana, 109
Amudha, J., 69
Anbarasa Pandian, A., 239
Ansari, Zahir Ahmed, 345
Anuj, Akash, 581
Arivazhagan, S., 533
Arya, K.V., 57
Asari, Vijayan K., 409

B

- Bajpai, Anurag, 629
Balasubramanian, R., 239, 619
Banerjee, Subhashis, 249
Belhe, Swapnil, 399
Bhalla, Vandna, 475
Bhatnagar, Charul, 467
Bhuyan, M.K., 523

C

- Chaitanya, Myneppalli Siva, 119
Chakraborty, Soman, 79
Chaudhuri, Bidyut Baran, 367
Chaudhury, Santanu, 475
Chowdhury, Ananda S., 79
Cui, Zhenxing, 1

D

- Dahiya, Kalpana, 89
Das, Nibaran, 205
Das, Partha Pratim, 581
Das, Rajsekhar, 629
Datta, Samyak, 619
Deshmukh, Maroti, 149

- Devara, Vasumathi, 181
Diwakar, Manoj, 571
Dogra, Debi Prosad, 261, 273
Duvieubourg, Luc, 159

F

- Fasogbon, Peter, 159

G

- Gadi Patil, Akshay, 443
Gawande, Ujwala, 215
Geetha, S., 511
Ghosh, Ripul, 297
Girish, G.N., 133
Golhar, Yogesh, 215
Goswami, Chitrata, 205
Gupta, Manish Kumar, 399
Guru, D.S., 555

H

- Hajari, Kamal, 215
Halder, Chayan, 205

I

- Induja, P., 533
Iwahori, Yuji, 523

J

- Jadhav, Narendra, 335
Jalal, Anand Singh, 285, 467
Jalan, Ankit, 119
Javed, Mohammed, 367
Johnson, Bibin, 499
Joshi, Yashwant, 335

K

- Kandpal, Neeta, 227, 377
Kaushik, Brajesh Kumar, 23
Khan, Ajaze Parvez, 101

Khare, Sangeeta, 101
 Khare, Sudhir, 23
 Khemchandani, Reshma, 193
 Kumar, Ashish, 227
 Kumar, Avnish, 345
 Kumar Chauhan, Vinod, 89
 Kumar, Manoj, 419, 571, 467
 Kumar, Nikhil, 227
 Kumar, Satish, 297
 Kumar, Shashi, 523

M

Macaire, Ludovic, 159
 Majhi, Bansidhar, 171
 Majumdar, Arun Kumar, 581
 Mallick, Tanwi, 581
 Mane, Sayali, 593
 Manthalkar, Ramchandra, 335
 Manyam, Gorthi R.K.S.S., 499
 Mishra, Deepak, 309
 Mishra, Deepak Kumar, 389
 Mishra, Sonali, 171
 Mitra, Sushmita, 249
 Mohapatra, Subrajeet, 109
 Mukherjee, Anindita, 79
 Mukherjee, Somnath, 323
 Mysore, Sheshera, 399

N

Nagabhushan, P., 367
 Nagananthini, C., 487
 Nair, Madhu S., 453
 Nain, Neeta, 149
 Nandedkar, Abhijeet V., 607
 Nigam, M.J., 345

O

Obaidullah, Sk.Md., 205

P

Pachori, Shubham, 545
 Pandey, Saurabh, 323
 Patel, Diptiben, 431
 Porwal, Sudhir, 101
 Punithavathi, P., 511
 Purohit, Manoj, 23

R

Rahman, S.M.Mahbubur, 357
 Rajan, Jeny, 133
 Raman, Shanmuganathan, 431, 545, 443
 Rani, J. Sheeba, 499

Rao, T.J. Narendra, 133
 Rohit, Kumar, 309
 Roy, Kaushik, 205
 Roy, Partha Pratim, 261, 273

S

Sabui, Arko, 119
 Sadasivam, Sowndarya Lakshmi, 69
 Saini, Rajkumar, 261, 273
 Saira Banu, J., 593
 Sai Subrahmanyam Gorthi, R.K., 309
 Sakthivel, P., 11
 Santosh, K.C., 205
 Sa, Pankaj Kumar, 171
 Sardana, H.K., 297
 Savithri, G., 593
 Sen, Debasish, 619
 Sharma, Anuj, 89
 Sharma, Lokesh, 171
 Sharma, Sweta, 193
 Sidike, Paheding, 409
 Sil, Jaya, 79
 Singh, Abhijeet, 119
 Sinha, Aloka, 35, 47
 Singh, Himanshu, 23
 Singh, Kshitij, 545
 Singh, Manvendra, 23
 Singh, Navjot, 389
 Sojan Lal, P., 453
 Solanki, Pooja, 467
 Sonam, 419
 Sonane, Bhoomika, 431
 Srivastava, Anurag Kumar, 377
 Srivastava, Rajeev, 57

T

Taha, Tarek M., 409
 Thivya, K.S., 11
 Tiwari, Shailendra, 57
 Tripathi, Rajesh Kumar, 285

U

Uke, Shruti R., 607
 Ullah, Md. Azim, 357
 Uma Shankar, B., 249

V

Vaddella, Rama Prasad V., 181
 Vadlamudi, Loka Nadham Naidu, 181
 Veera, Viswanath, 119
 Venkatesan, Shankar M., 119, 629
 Verma, Gaurav, 35, 47

Vinay Kumar, N., [555](#)

Vinupriya, [593](#)

W

Wang, Feng, [1](#)

Y

Yang, Mingqiang, [1](#)

Yogameena, B., [487](#)

Z

Zacharias, Geevar C., [453](#)

Zeng, Wei, [1](#)