Advances in Intelligent Systems and Computing 460

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

Proceedings of International Conference on Computer Vision and Image Processing

CVIP 2016, Volume 2

Springer

Advances in Intelligent Systems and Computing

Volume 460

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

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 2

Springer

Editors
Balasubramanian Raman
Department of Computer Science
and Engineering
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

Sanjeev Kumar
Department of Mathematics
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 978-981-10-2106-0 ISBN 978-981-10-2107-7 (eBook)
DOI 10.1007/978-981-10-2107-7

Library of Congress Control Number: 2016952824

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 26 to 28 February, 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 a total of 253 paper submissions, 106 (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 composed of 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 5 plenary talks, and paper presentations in 8 oral and 3 poster sessions. Emphasis was given on 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

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.

Fingerprint Image Segmentation Using Textural Features

Reji C. Joy and M. Azath

Abstract Automatic Fingerprint Identification System (AFIS) uses fingerprint segmentation as its pre-processing step. A fingerprint segmentation step divides the fingerprint image into foreground and background. An AFIS that uses a feature extraction algorithm for person identification will tend to fail if it extracts spurious features from the noisy background area. So fingerprint image segmentation plays a crucial role in reliably separating ridge like part (foreground) from its background. In this paper, an algorithm for fingerprint image segmentation using GLCM textural feature is presented. Four block level GLCM features: Contrast, Correlation, Energy and Homogeneity are used for fingerprint segmentation. A linear classifier is trained for classifying per block of fingerprint image. The algorithm is tested on standard FVC2002 dataset. Experimental results show that the proposed segmentation method works well in noisy fingerprint images.

Keywords Fingerprint · Segmentation · GLCM features · Logistic regression · Morphological operations

1 Introduction

The increasing interest in security over the last years has made the recognition of people by means of biometric features receive more and more attention. Admission to restricted areas, personal identification for financial transactions, lockers and forensics are just a few examples of its applications. Though iris, face, fingerprint, voice, gait etc. can be used as a biometric characteristic, the most commonly used one is the fingerprint. The fingerprint sensors are relatively low priced and not much effort

R.C. Joy (✉)
Karpagam University, Coimbatore, India
e-mail: reji.c.j@vidyaacademy.ac.in

M. Azath
Karpagam University, Coimbatore, India
e-mail: mailmeazath@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_1

is required from the user. Moreover, fingerprint is the only biometric trait left by criminals during a crime scene.

A fingerprint is a sequence of patterns formed by ridges and valleys of a finger. A fingerprint verification system is carried out by four main steps: *acquisition*, *preprocessing*, *feature extraction* and *matching*. A fingerprint image is usually captured using a fingerprint scanner. A captured fingerprint image usually consists of two parts: the foreground and the background (see Fig. 1). The foreground is originated and acquired from the contact of a fingertip with the sensor [1]. A ridge is a curved like line segment in a finger and the region adjacent to two ridges is defined as a valley. A fingerprint matching is performed by extracting feature points from a preprocessed fingerprint image. Minutiae (ridge ending and bifurcation) and singular points (core and delta) are the most commonly used feature points. It is important that the features should be extracted only from the foreground part for an accurate matching. Therefore, fingerprint image segmentation plays a crucial role in the reliable extraction of feature points.

Several approaches are known in fingerprint image segmentation for years. Bazen and Gerez [1], used pixel based features like local mean, local variance and coherence for fingerprint segmentation. Then, a linear combination of these features is taken for segmentation. The coherence feature indicates how well the orientations over a neighborhood are pointing in the same direction. Since coherence measure is higher in the foreground than in the background, the combination of these features is used to characterize foreground and background. But this algorithm is not robust to noise and also it is costly since it is based on pixel features. Gabor features of the fingerprint image are used by Alonso-Fernandez et al. [2] for segmentation. It

Fig. 1 A fingerprint image showing the foreground and the background

is known that the Gabor response is higher in the foreground region than that in the background region. Chen et al. [3] proposed a feature called cluster degree (CluD) which is a block level feature derived from gray-level intensity values. CluD measures how well the ridge pixels are clustered in a block and they have stated that this measure will be higher in the foreground. Harris corner point features are used by Wu et al. [4] to separate foreground and background. The advantage of this approach is that it is translation and rotation invariant. It has been stated that the strength of a Harris point is much higher in the foreground area.

In this paper a fingerprint segmentation algorithm is presented. This proposal is made on the observation that a fingerprint image can be viewed as an oriented texture pattern of ridges and valleys. This paper proposes four block level GLCM features: Contrast, Correlation, Energy and Homogeneity for fingerprint segmentation. A linear classifier is trained for classifying each block of fingerprint image into foreground and background. Finally a morphological operator is used to obtain compact foreground clusters.

This paper is organized in four sections. Section 2 describes Gray-Level Co-occurrence Matrix (GLCM) and the proposed feature extraction method. Section 3 discusses the classification techniques and the proposed classifier. Section 4 gives experimental results and discussion and Sect. 5 draws the conclusion.

2 Fingerprint as Texture

In a fingerprint image, the flow of ridges and valleys can be observed as an oriented texture pattern [5]. Texture is a repeated pattern of local variations in image intensity. One of the important characteristics is that most textured images contain spatial relationships. Mutually distinct texture differs significantly in these relationships and can easily be discriminated by a joint distribution based on their co-occurrences and orientation channels.

2.1 Co-occurrence Matrix

Gray Level Co-occurrence Matrix (GLCM) is one of the popular methods that compute second-order gray-level features for textural image analysis. GLCM was originally proposed by Haralick [6, 7]. The co-occurrence matrix is used to measure the relative probabilities of gray-level intensity values that are present in an image. A co-occurrence matrix can be defined as a function $P(i, j, d, \theta)$ that estimates the probability of co-occurring two neighborhood pixel values in an image with a gray level value i and the other with gray level j for a given distance d and direction θ. Usually, the parameter value for the distance d is taken as $d = 1$ or 2 and the direction θ utilizes values of 0°, 45°, 90° and 135°. The following equations given by [7] are used to compute co-occurrence matrix:

$$P(i,j,d,0°) = \#\{((k,l),(m,n)) \in (L_r \times L_c) \times (L_r \times L_c) \Big| k - m = 0, |l-n| = d, I(k,l) = i, I(m,n) = j\} \tag{1}$$

$$P(i,j,d,45°) = \#\{((k,l),(m,n)) \in (L_r \times L_c) \times (L_r \times L_c) \Big| (k-m = d, l-n = -d) or (k-m = -d, l-n = d) \; I(k,l) = i, I(m,n) = j\} \tag{2}$$

$$P(i,j,d,90°) = \#\{((k,l),(m,n)) \in (L_r \times L_c) \times (L_r \times L_c) \Big| |k-m| = d, l-n = 0, I(k,l) = i, I(m,n) = j\} \tag{3}$$

$$P(i,j,d,135°) = \#\{((k,l),(m,n)) \in (L_r \times L_c) \times (L_r \times L_c) \Big| (k-m = d, l-n = d) or (k-m = -d, l-n = -d) \; I(k,l) = i, I(m,n) = j\} \tag{4}$$

where # denotes number of elements in the set, L_r and L_c be the spatial domains of the row and column dimensions.

2.2 GLCM Features

Haralick et al. [6] proposed 14 statistical features from each co-occurrence matrix computed by the Eqs. (1–4). However, in this paper we have used only 4 features that can successfully separate the foreground and background regions (experimentally determined) of the fingerprint images for fingerprint segmentation. These are

$$\begin{aligned} Contrast: \quad & f_1 = \sum_{i,j} |i-j|^2 p(i,j) \\ Correlation: \quad & f_2 = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j} \\ Energy: \quad & f_3 = \sum_{i,j} p(i,j)^2 \\ Homogeneity: \quad & f_4 = \sum_{i,j} \frac{p(i,j)}{1+|i-j|} \end{aligned} \tag{5}$$

where μ_i, μ_j are the means and σ_i, σ_j are the standard deviations of the row and column respectively and $p(i,j)$ is the probability of co-occurring i with j using the chosen distance d.

2.3 *Proposed Feature Extraction Method*

Computation of co-occurrence matrix is highly influenced by factors like the gray levels used and the distance (d). Since a digital image is represented by 256 gray level values, computing 256×256 co-occurrence matrices at all position is computationally complex. In addition, large variations in the intensity values also fail to capture the fingerprint image textural patterns. Usually, a gray-level quantized to its

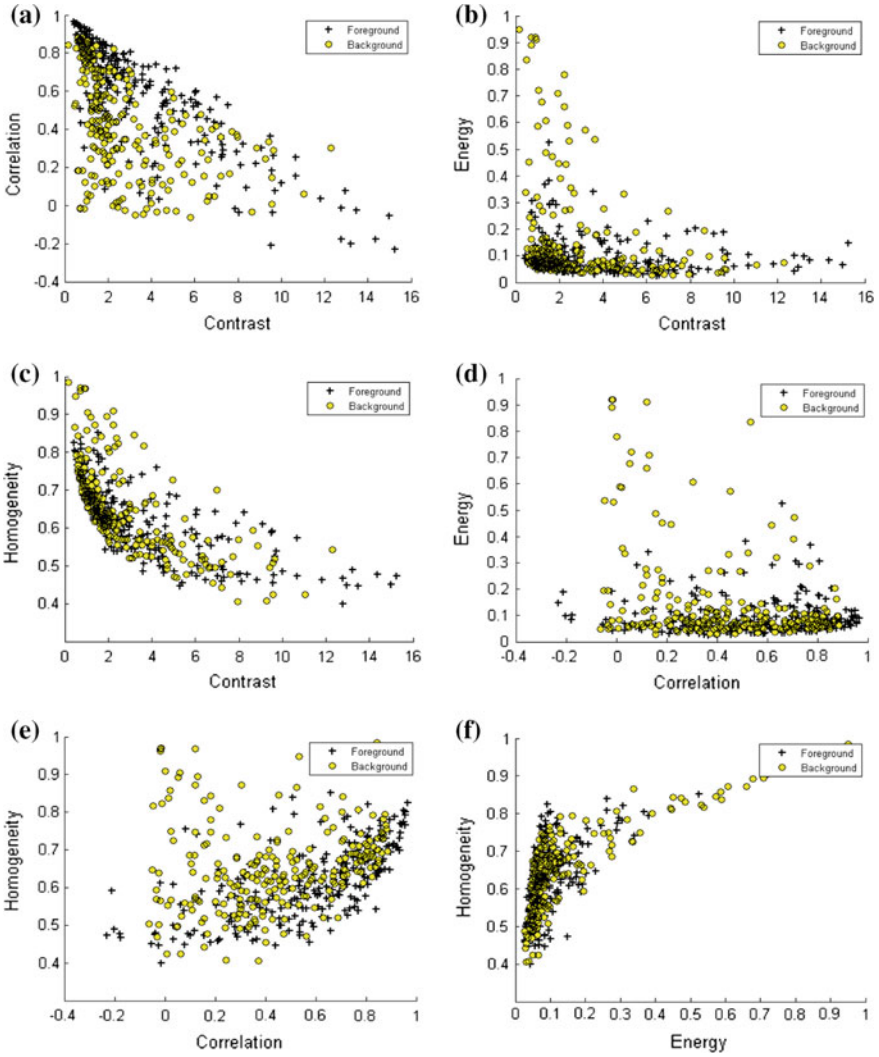


Fig. 2 Joint distributions of the combination of two block features for foreground and background of Db_2

lower intensity value will be considered for processing. Zacharias et al. [8] states that the co-occurrence matrix may use false textural information if not used an optimal value of the parameter distance (d) and would tend to extract false GLCM features.

Extracting GLCM features is performed as follows: Divide the input fingerprint image into a non-overlapping block size of $W \times W$. From each block compute co-occurrence matrix in 4 directions with a predefined set of parameters using equations (Eq. 1–4). From each co-occurrence matrix, 4 set of GLCM features are computed using the Eq. 5 to form a total of 16 features. To reduce the feature size of the classifier, we have used the summed up values of each feature with respect to 4 directions to form a 4 feature vector represented as [*Contrast Correlation Energy Homogeneity*]. In our work, we have used a block size $W = 15$, fingerprint image is quantized to contain only 8 gray levels and the distance d is taken as 2 for different parameters for computing a co-occurrence matrix.

To show the usefulness of these block features taken from the fingerprint, we have shown (see Fig. 2) the joint distribution of the combination of two block features taken from both the foreground and the background area. We have used standard FVC2002 Db2_b dataset [9] for testing our algorithm.

3 Classification

Fingerprint segmentation is a problem which divides the fingerprint image into foreground and background parts. Essentially this problem can be treated as a classification problem to classify a fingerprint image into two classes: class *foreground* and class *background*. For a classification problem there are two main approaches: supervised learning and unsupervised learning. In literature, many segmentation algorithms have been reported to have used supervised learning approach [1, 3] and unsupervised learning approach [10]. In this paper, we are using supervised learning approach since we already know what could be the features for the samples to classify them either to class foreground or class background. There are several supervised learning algorithms reported in the literature like linear and quadratic discriminant functions, neural networks, K-nearest neighbor, decision trees, support vector machines etc. [9]. However, in our work we have used a linear classifier called logistic regression as the segmentation classifier since it requires a low computational cost.

3.1 Logistic Regression

Let a variable y represents the class of a fingerprint sample then, $y = 0$ means that the sample belongs to class background and $y = 1$ means that the sample belongs to class foreground. Let x_j represents jth feature of the sample. A logistic regression model is defined as:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

where the function $g(z)$ is known as logistic function. The variable z is usually defined as

$$z = \theta^T \mathrm{x} = \theta_0 + \sum_{j=1}^{n} \theta_j x_j \tag{7}$$

where θ_0 is called the intercept term and $\theta_1, \theta_2, ..., \theta_n$ are called the regression coefficients of $x_1, x_2, ..., x_n$ respectively. Hence, Logistic regression is an attempt to find a formula that gives the probability $p(y = 1|x, \theta)$ that represents the class foreground. Since only two classes are considered, the probability of the sample representing the class background is therefore $1 - p(y = 1|x, \theta)$. A logistic regression is modeled as linear combination of

$$\eta = \log \frac{p}{1 - p} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{8}$$

where $\theta_0, \theta_1, ..., \theta_n$ are the optimal parameters that minimizes the following cost function.

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y_i \log p_i + (1 - y_i) \log(1 - p_i)] \tag{9}$$

$$\theta = \min_{\theta} J(\theta) \tag{10}$$

where m is the total number of samples and y_i is the assigned class for the sample x_i. The prediction of the given new sample data x is class label 1 if this criteria is true: $p(y = 1|x, \theta) \geq 0.5$.

3.2 Proposed Classifier

The proposed classifier is linear classifier which tests a linear combination of the features, given by:

$$z = \theta^T \mathrm{x} = \theta_0 + \theta_1 Contrast + \theta_2 Correlation + \theta_3 Energy + \theta_4 Homogeneity \tag{11}$$

where z is the value to be tested, $\theta = [\theta_0\ \theta_1\ \theta_2\ \theta_3\ \theta_4]^T$ is the weight vector and $\mathrm{x} = [1\ Contrast\ Correlation\ Energy\ Homogeneity]^T$ is the feature vector. Then, using the class ω_1 for the foreground, class ω_0 for the background and $\hat{\omega}$ for the assigned class, the following decision function is applied:

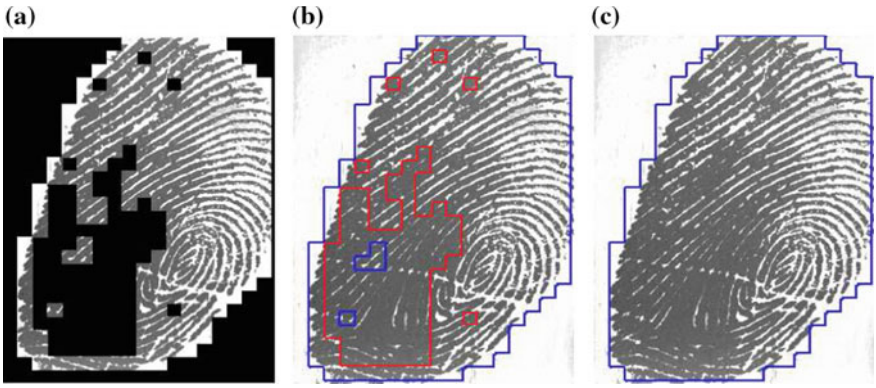


Fig. 3 **a** Before Morphology **b** Boundary showing the regions classified as foreground (*blue color*) and background (*red color*) **c** After Morphology

$$\hat{\omega} = \begin{cases} \hat{\omega}_1 \text{ if } g(z) \geq 0.5 \\ \hat{\omega}_0 \text{ if } g(z) < 0.5 \end{cases} \tag{12}$$

3.3 Morphological Operations

Segmenting foreground and background especially from a low quality fingerprint images is error prone (see Fig. 3a). Therefore, it may produce some spurious segmentations where regions of the class background may appear inside the class foreground and vice versa (see Fig. 3b). However, we can group these small regions that appear inside a larger region to form a meaningful clustered region. This fact can be seen from Table 2 where 24.5 % is misclassified when a morphological operations is not used. A number of post-processing methods can be applied to get more meaningful segmentation results. In our work, we have chosen morphological operations [11] as post-processing method to get an optimal classification estimate. The morphological operations work well to remove these small regions by adding it to form the part of a larger region, thus creating more compact clusters.

Many fingerprint segmentation methods [1, 3] are reported in the literature that uses morphological operations as post-processing. First, small isolated clusters that are incorrectly classified as foreground regions in the background area are removed using morphological open operation and regions that are incorrectly classified as background regions in the foreground area are removed based on morphological reconstruction [12] algorithm. Finally, all small regions outside the large region will be treated as background. Fig. 3c shows the segmented fingerprint image after the morphological operations.

4 Experimental Results

The segmentation algorithm was tested on FVC 2002 Db2_a standard dataset [9]. In order to quantitatively measure the performance of the segmentation algorithm, initially we have manually identified the foreground and the background blocks. Evaluation is done by comparing the manual segmentation with the segmentation results given by the classifier. The number of misclassification can be used as a performance measure.

The misclassification is given by:

$$\begin{aligned} p(\hat{\omega}_1|\omega_0) &= \frac{N_{be}}{N_b} \\ p(\hat{\omega}_0|\omega_1) &= \frac{N_{fe}}{N_f} \\ p_{error} &= \frac{p(\hat{\omega}_1|\omega_0) + p(\hat{\omega}_0|\omega_1)}{2} \end{aligned} \tag{13}$$

where N_{be} is number of background classification error, N_b is the total number of true background blocks in the image and $p(\hat{\omega}_1|\omega_0)$ gives the probability that a foreground block is classified as background. N_{fe} is the number of foreground classification error, N_f is the total number of true foreground blocks in the image and $p(\hat{\omega}_0|\omega_1)$ is the probability that a background block is classified as foreground. The probability of error p_{error} is the average of $p(\hat{\omega}_0|\omega_1)$ and $p(\hat{\omega}_1|\omega_0)$.

The proposed fingerprint segmentation algorithm has been trained on the 800 block samples from 80 fingerprint images in FVC Db2_b dataset [9] which consists of 10 equal number of background and foreground block samples and labeled correspondingly to get the optimal values of the regression coefficients. The regression coefficients of the trained result is

$$\theta^T = [\theta_0\, \theta_1\, \theta_2\, \theta_3\, \theta_4] = [-2.083 \;\; 0.2916 \;\; 1.5393 \;\; 1.4644 \;\; -2.0191] \tag{14}$$

We have used this vector for classifying the fingerprint image into foreground and background. Table 1 gives the test result of a 10 randomly selected images from Db2_a dataset with respect to the equation given in Eq. 13.

To quantify the effect of morphological operations as post-processing is quantified in Table 2 and it consolidates the total error probabilities before and after morphological operations on the same 10 fingerprint images. The results show that the overall misclassification rate is very less (Fig. 4).

The overall performance measure is also quantified as how well the foreground region contains the feature parts. Since the fingerprint singular points is one of the important feature for fingerprint classification and matching, the performance of the proposed method can be analyzed by finding how well the singular points of the fingerprint images are included by the segmentation algorithm. To conduct this analy-

Table 1 Result of the logistic regression classifier

Fingerprint	N_f	N_b	N_{fe}	N_{be}	$p(\hat{\omega}_1\|\omega_0)$	$p(\hat{\omega}_0\|\omega_1)$	p_{error}
Db2_a_04_8.tif	542	277	7	11	0.0397	0.0129	0.0263
Db2_a_10_1.tif	568	251	9	7	0.0278	0.0158	0.0218
Db2_a_14_3.tif	675	144	5	1	0.0069	0.0074	0.0071
Db2_a_26_1.tif	602	217	8	10	0.0460	0.0132	0.0296
Db2_a_32_5.tif	353	466	15	4	0.0085	0.0424	0.0254
Db2_a_49_2.tif	595	224	15	5	0.0223	0.0252	0.0237
Db2_a_51_1.tif	576	243	15	6	0.0246	0.0260	0.0253
Db2_a_61_6.tif	590	229	12	10	0.0436	0.0203	0.0319
Db2_a_72_1.tif	522	297	10	10	0.0336	0.0191	0.0263
Db2_a_77_2.tif	576	243	15	7	0.0288	0.0260	0.0274

Table 2 Effect of morphological operations as post-processing

	N_f	N_b	N_{fe}	N_{be}	$p(\hat{\omega}_1\|\omega_0)$	$p(\hat{\omega}_0\|\omega_1)$	p_{error}
Before morphology	5599	2591	111	71	0.2818	0.2083	0.2450
After morphology	5599	2591	47	34	0.0131	0.0083	0.0107

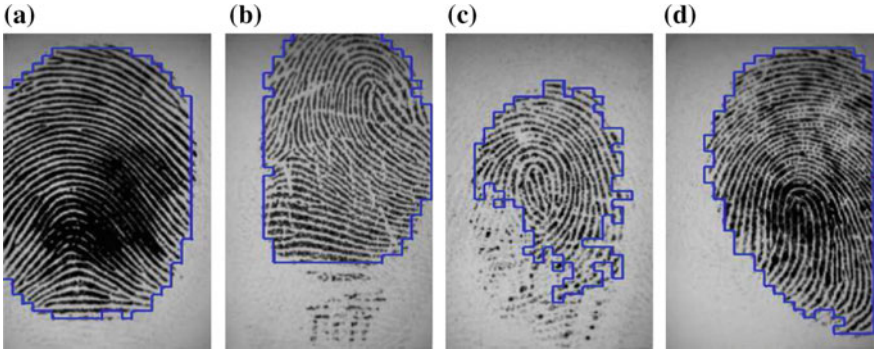


Fig. 4 Segmentation results of some fingerprint images from FVC2002 Db2_a dataset

sis, out of 800 images in the FVC2002 Db2_a dataset [9], we have excluded 11 images since either the singular point is absent or it is very near to the image border. Table 3 shows the test result of this analysis. It can be shown even when the segmentation of fingerprint image is moderate, the algorithm is able to include the singular point region in the foreground area for 98.6 % of images. This shows the efficiency of our proposed algorithm as singular point is very important for the subsequent stages of fingerprint identification.

Table 3 Performance of the proposed methods based on detected singular points

Total number of fingerprints in the dataset	800
Total number of fingerprints taken for testing	789
Successful inclusion of singular points in segmentation	778
Segmentation accuracy	98.6

5 Conclusion

Accurate segmentation of fingerprint images is important in any of the automatic fingerprint identification system as it is used to reliably separate ridge like part (foreground) from its background. Fingerprint segmentation is crucial since the overall performance of an automatic fingerprint identification system using fingerprint depends on this pre-processing step. It is critical to keep feature parts and discard the noisy parts of the fingerprint for the accurate performance. In this paper, a method for fingerprint segmentation is presented. Since a fingerprint image exhibits strong textural characteristics, this method uses four block level GLCM features, being the contrast, the correlation, the energy and the homogeneity, that measures the textural properties of the fingerprint image for segmentation. A logical regression classifier has been trained for segmentation which classifies each block to either foreground or background. Morphological operations are applied as post-processing to obtain more compact regions for reducing the overall classification errors.

References

1. Bazen, A. M., Gerez, S. H.: Segmentation of fingerprint images. In: ProRISC 2001 Workshop on Circuits, Systems and Signal Processing, (2001)
2. Alonso-Fernandez, F., Fierrez-Aguilar, J., Ortega-Garcia, J.: An enhanced Gabor filter-based segmentation algorithm for fingerprint recognition systems. In: Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA 2005), pp. 239–244, (2005)
3. Chen, X., Tian, J., Cheng, J., Yang, X.: Segmentation of fingerprint images using linear classifier. EURASIP Journal on Applied Signal Processing. 2004(4), pp. 480–494, (2004)
4. Wu, C., Tulyakov, S., Govindaraju, V.: Robust Point-Based Feature Fingerprint Segmentation Algorithm. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007, LNCS, vol. 4642, pp. 1095–1103, Springer, Heidelberg (2007)
5. Jain, A. K., Ross, A.: Fingerprint Matching Using Minutiae and Texture Features. In: Proceeding of International Conference on Image Processing, pp. 282–285, (2001)
6. Haralick, R. M., Shanmugan, K., Dinstein, J.: Textual features for image classification. IEEE Trans. Syst. Man. Cybern. Vol. SMC-3, pp. 610–621, (1973)
7. Haralick, R. M.: Statistical and Structural Approaches to Texture. In: Proceedings of IEEE, Vol. 67, No. 5, pp. 768–804, (1979)

8. Zacharias, G.C., Lal, P.S.: Combining Singular Point and Co-occurrence Matrix for Fingerprint Classification. In: Proceedings of the Third Annual ACM Bangalore Conference, pp. 1–6, (2010)
9. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of fingerprint recognition (Second Edition). Springer, New York (2009)
10. Pal, N.R., Pal, S.K.: A review on image segmentation techniques. Pattern Recognition. Vol. 26, No. 9, pp. 1277–1294, (1993)
11. Gonzalez, R. C., Wintz, P.: Digital Image Processing.2nd Edition. Addison-Wesley, (1987)
12. Soille, P., Morphological Image Analysis: Principles and Applications. Springer-Verlag, pp. 173–174, (1999)

Improved Feature Selection for Neighbor Embedding Super-Resolution Using Zernike Moments

Deepasikha Mishra, Banshidhar Majhi and Pankaj Kumar Sa

Abstract This paper presents a new feature selection method for learning based single image super-resolution (SR). The performance of learning based SR strongly depends on the quality of the feature. Better features produce better co-occurrence relationship between low-resolution (LR) and high-resolution (HR) patches, which share the same local geometry in the manifold. In this paper, Zernike moment is used for feature selection. To generate a better feature vector, the luminance norm with three Zernike moments are considered, which preserves the global structure. Additionally, a global neighborhood selection method is used to overcome the problem of blurring effect due to over-fitting and under-fitting during K-nearest neighbor (KNN) search. Experimental analysis shows that the proposed scheme yields better recovery quality during HR reconstruction.

Keywords Super-resolution · Zernike moment · Luminance norm · Manifold learning · Global neighborhood selection · Locally linear embedding

1 Introduction

Visual pattern recognition and analysis plays a vital role in image processing and computer vision. However, it has several limitations due to image acquisition in the unfavorable condition. Super-resolution (SR) technique is used to overcome the limitations of the sensors and optics [1]. Super-resolution is a useful signal processing

D. Mishra (✉) · B. Majhi · P.K. Sa
Department of Computer Science and Engineering,
National Institute of Technology, Rourkela 769008, India
e-mail: deepasikhame@gmail.com

B. Majhi
e-mail: bmajhi@nitrkl.ac.in

P.K. Sa
e-mail: pankajksa@nitrkl.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_2

technique to obtain a high-resolution (HR) image from an input low-resolution (LR) image. In this work, we have modeled a learning based super-resolution approach to generate a HR image from a single LR image.

The problem of learning based SR was introduced by Freeman et al. [2] called example-based super-resolution (EBSR). In their work, a training set has been used to learn the fine details that correspond to the region of low-resolution using the one-pass algorithm. Later, Kim et al. [3] extended their formulation by considering kernel ridge regression which combines the idea of gradient descent and matching pursuit. Afterward, Li et al. [4] have proposed example-based single frame SR using support vector regression (SVR) to illustrate the local similarity. However, due to lack of similarities in local geometry and neighborhood preservation, aliasing effect is generated during HR reconstruction. To preserve the neighborhood information, a neighbor embedding based SR (SRNE) was introduced by Chang et al. [5]. Thereafter, in [6–10] an extended neighbor embedding based SR is used by considering different feature selection methods. Chan et al. [8] have proposed a neighbor embedding based super-resolution algorithm through edge detection and feature selection (NeedFS), where a combination of luminance norm and the first-order gradient feature is introduced for edge preservation and smoothening the color region. To preserve the edge, Liao et al. [9] have proposed a new feature selection using stationary wavelet transform (SWT) coefficient. Mishra et al. [10] have emphasized on neighborhood preservation and reduction of sensitivity to noise. Therefore, they have proposed an incremental feature selection method by combining the first-order gradient and residual luminance inspired by image pyramid. Gao et al. [11] have proposed a method to project the original HR and LR patch onto the jointly learning unified feature subspace. Further, they have introduced sparse neighbor selection method to generate a SR image [12]. Bevilacqua et al. [13] have introduced a new algorithm based on external dictionary and non-negative embedding. They have used the iterative back-projection (IBP) to refine the LR image patches and a joint K-means clustering (JKC) technique to optimize the dictionary. In [14], a new Zernike moment based SR has been proposed for multi-frame super-resolution. Due to orthogonality, rotation invariance, and information compaction of Zernike moment, they have formulated a new weight value for HR image reconstruction.

However, in practice, preserving the fine details in the image is inaccurate in embedding space, which is still an open problem. For better local compatibility and smoothness constraints between adjacent patches, a better feature selection is necessary. Hence, we have proposed a new feature selection method inspired by Zernike moment [15]. In our work, a feature vector has been generated by the combination of three Zernike moments and luminance norm. In addition, a global neighborhood selection method is used to generate the K value for neighborhood search to overcome the problem of over-fitting and under-fitting. The proposed approach is verified through the different performance measures. The experimental results indicate that proposed scheme preserves more fine details than the state-of-the-art methods.

The remainder of the paper is organized as follows. Section 2 describes the problem statement. Section 3 presents an overall idea about Zernike moment. Section 4 discusses the proposed algorithm for single image super-resolution using Zernike moment. Experimental results and analysis are discussed in Sect. 5 and the concluding remarks are outlined in Sect. 6.

2 Problem Statement

In this section, the objective of single image super-resolution problem is defined and formulated. Let us consider a set of n low-resolution images of size $M \times N$. Theoretically each low-resolution image can be viewed as a single high-resolution image of size $DM \times DN$ that has been blurred and down sampled by a factor of D. A particular low-resolution image X_l is represented as

$$X_l = DB(X_h)\,, \tag{1}$$

where X_h is a $DM \times DN$ high-resolution image, B is 5×5 Gaussian blur kernel and D is the down sampling factor. In the proposed scheme, we consider a neighbor embedding approach to generate a SR image for a given LR image. Hence, a set of LR and its corresponding HR training image is required to find out a co-occurrence relationship between LR and HR patches.

3 Background

In the field of image processing and pattern recognition, moment-based features play a vital role. The use of the Zernike moments in image analysis was introduced by Teague [15]. Zernike moments are basically projections of the image information to a set of complex polynomials, that from a complete orthogonal set over the interior of a unit circle, i.e. $\sqrt{x^2 + y^2} \leq 1$.

The two-dimensional Zernike moments of an image intensity function $f(x, y)$ of order n and repetition m are defined as

$$Z_{nm} = \frac{n+1}{\pi} \int \int_{\sqrt{x^2+y^2} \leq 1} f\,(x, y)\, V^*_{nm}\,(x, y)\, dxdy\,, \tag{2}$$

where $\frac{n+1}{\pi}$ is a normalization factor. In discrete form Z_{nm} can be expressed as

$$Z_{nm} = \sum_x \sum_y f\,(x, y)\, V^*_{nm}\,(x, y) \;\;, \sqrt{x^2 + y^2} \leq 1\,. \tag{3}$$

The kernel of these moments is a set of orthogonal polynomials, where the complex polynomial V_{nm} can be expressed in polar coordinates (ρ, θ) as

$$V_{nm}(\rho, \theta) = R_{nm}(\rho) e^{-jm\theta}, \tag{4}$$

where $n \geq 0$ and $n - |m|$ is an even positive integer.

In (4), $R_{nm}(\rho)$ is radial polynomial and is defined as

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{(-1)^s (n-s)! r^{n-2s}}{s! \left(\frac{n+|m|}{2} - s\right)! \left(\frac{n-|m|}{2} - s\right)!}. \tag{5}$$

The real and imaginary masks are deduced by a circular integral of complex polynomials. On the whole, edge detection is conducted at the pixel level. At each edge point, orthogonal moment method is used to calculate accurately gradient direction. Mostly, the higher-order moments are more sensitive to noise. Therefore, first three 2nd order moments has been employed for feature selection. The real and imaginary 7×7 homogeneous mask of M_{11} and M_{20} should be deduced by circular integral of V_{11}^* and V_{20}^* [16]. Hence, three Zernike moments are $Z_{11}R$, $Z_{11}I$ and Z_{20}.

4 Neighbor Embedding Based SR Using Zernike Moment

In this section, a new feature selection method is proposed using Zernike moments for neighbor embedding based super-resolution. The feature vector is generated by combining the three Zernike moments with luminance norm. Moreover, neighborhood size for K-nearest neighbor (KNN) search is generated by global neighborhood selection [17]. The overall block diagram of the proposed scheme is shown in Fig. 1.

4.1 Neighbor Embedding Based SR

To perform neighbor embedding based SR, luminance component of each image is split into a set of overlapping patches. $X_L = \left\{x_l^t\right\}_{t=1}^T$ is the training LR image and $X_H = \left\{x_h^s\right\}_{s=1}^S$ is the corresponding HR image. To preserve the inter-patch relationship between the LR and HR patch, if the patch size of LR image is $s \times s$ then the patch size of corresponding HR image will be $fs \times fs$, where f is the magnification factor. The input LR image $Y_L = \left\{y_l^t\right\}_{t=1}^T$ and expected HR image $Y_H = \left\{y_h^s\right\}_{s=1}^S$ pair should have same number of patches.

In training process, for each LR patch K-nearest neighbors search among all training LR patches and the optimal reconstruction weight vector W_t calculated by minimizing the local reconstruction error as

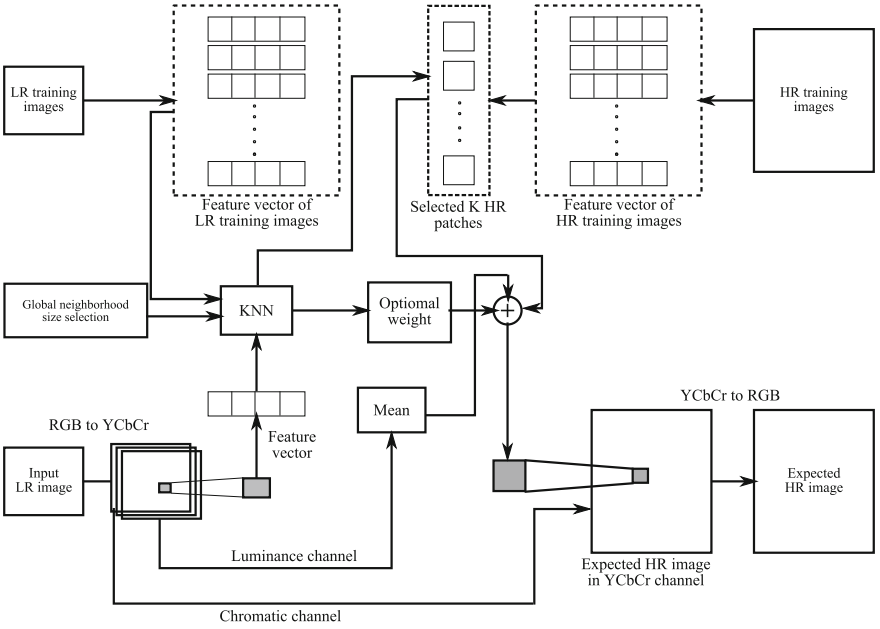


Fig. 1 Block diagram of proposed scheme

$$\varepsilon^t = \min \left\| y_l^t - \sum_{x_l^s \in N_t} w_{ts} x_l^s \right\|^2 , \tag{6}$$

subject to two constrains, i.e., $\sum_{y_l^s \in N_t} w_{ts} = 1$ and $w_{ts} = 0$ for any $y_l^s \notin N_t$. This is generally used for normalization of the optimal weight vector, where N_t is the set of neighborhood of y_l^t in training set X_L.

The local Gram matrix G_t plays an important role to calculate the weight w_t associated to y_l^t, which is defined as

$$G_t = \left(y_l^t 1^T - X\right)^T \left(y_l^t 1^T - X\right) , \tag{7}$$

where one's column vectors are considered to match the dimensionality with X. The dimension of X is $D \times K$, where its columns represent the neighbors of y_l^t. The optimal weight vector W_t for y_l^t having the weights of each neighbors w_{ts} are reordered by s. The weight is calculated as

$$w_{t=} \frac{G_t^{-1} 1}{1^T G_t^{-1} 1} , \tag{8}$$

After solving w_t efficiently, the high-resolution target patch y_h^t is computed as follows:

$$y_h^t = \sum_{x_l^s \in N_q} w_{ts} x_h^s \tag{9}$$

Then the HR patches are stitched according to the corresponding coordinates by averaging the overlapping regions. The detailed procedure of the proposed scheme is given in Algorithm 1.

Algorithm 1 Neighbor embedding based SR using Zernike feature

Input : Training LR image $X_L = \{x_l^t\}_{t=1}^T$ and HR image $X_H = \{x_h^s\}_{s=1}^S$,
Testing LR image $Y_L = \{y_l^t\}_{t=1}^T$,
Patch size – s, Up sampling size – f.
Output : Expected HR image.

1. Split X_L and Y_L into patches of size $s \times s$ with overlapping by one pixel.
2. Split X_H into patches of size $fs \times fs$ with overlapping by $f \times 1$ pixels accordingly.
3. Concatenate the three Zernike moments of X_L, X_H and Y_L with its corresponding luminance norm for feature vector.
4. For each testing LR patch $y_l^t \, \epsilon \, Y_L$.
   (a) Find N_t by K-nearest neighbors among all training patches using Euclidean distance. Here, K is calculated by global neighborhood selection.
   (b) Compute optimal reconstruction weights of y_l^t by minimizing the local reconstruction error.
   (c) Compute the high-resolution embedding y_h^s using (9).
5. To generate expected HR image enforce inter-patch relationships among the expected HR patches by averaging the feature values in overlapped regions between adjacent patches.

4.2 Zernike Moment Based Feature Selection

In this section, an efficient feature selection method for neighbor embedding based super-resolution method is proposed. In [5, 7, 8], several features are used for better geometry preservation in the manifold. But, consistency in structure between the neighborhood patches embedding still is an issue. To overcome the problem like sensitivity of noise, recovery quality, and neighborhood preservation among the patches, Zernike moment feature descriptor is used as appropriate feature selection. Due to robustness to noise and orthogonal properties of Zernike moment, a perfect representation of information is done. Basically, the features are selected from the luminance channel because it is sensitive to the human visual system. Luminance norm is also considered as a part of the features because it represent the global structure

of the image. For each pixel, there are four components of a feature vector *i.e.*, $[LN, Z_{11}R, Z_{11}I, and\, Z_{20}]$. As the learning based SR perform on the patch, feature vector of each patch size is $4p^2$, where p is the patch size.

4.3 Global Neighborhood Selection

Choosing the neighborhood size for locally linear embedding has great influence on HR image reconstruction because the neighborhood size K determines the local and global structure in the embedding space. Moreover, fixed neighborhood size leads to over-fitting or under-fitting. To preserve the local and global structure, the neighbor embedding method search a transformation. Hence, global neighborhood selection method is used. The reason for global neighborhood selection is to preserve the small scale structures in manifold. To get the best reconstructed HR image, well representation of high dimensional structure is required in the embedding space.

This method has been introduced by Kouropteva et al. [17], where Residual Variance is used as a quantitative measure that estimate the quality of the input-output mapping in embedding space. The residual variance [18] is defined as

$$\sigma_r^2(d_X, d_Y) = 1 - \rho^2_{d_X, d_Y}\,, \tag{10}$$

where ρ is the standard linear correlation coefficient, takes over all entries of d_X and d_Y matrices; The element of d_X and d_Y matrices having size $m \times m$ represents the Euclidean distance between pair of patches in X and Y. According to [17] lower is the residual variance better is the high dimensional data representation. Hence, optimal neighborhood size $K = (k_{opt})$ computed by hierarchical method as

$$k_{opt} = \arg\min_k (1 - \rho^2_{d_X, d_Y})\,. \tag{11}$$

The overall mechanism of global neighborhood selection is summarized in Algorithm 2

Algorithm 2 Neighborhood selection

Input : All patches.
Output : Neighborhood size K.

1. Set k_{max} as the maximal possible value of k_{opt}.
2. Calculate the reconstruction error $\epsilon = \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{N} w_{ij} x_{ij} \right\|$ for each $k\epsilon[1, k_{max}]$.
3. Find all minimum of $\epsilon(k)$ and corresponding k's which compose the set of s of initial candidate.
4. For each $k\epsilon s$ compute residual variance.
5. Compute $K = (k_{opt})$ using (11).

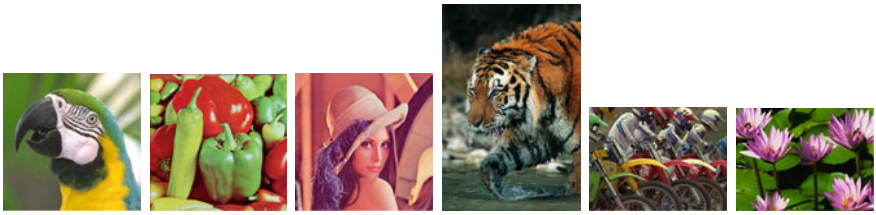
5 Experimental Results

5.1 Experimental Setting

To validate the proposed algorithm, simulations are carried out on some standard images of different size like Parrot, Peppers, Lena, Tiger, Biker, and Lotus. In this experiment, a set of LR and HR pairs are required for training. Hence, LR images are generated from the ideal images by blurring each image using (5×5) Gaussian kernel and decimation using $3:1$ decimation ratio in each axis. A comparative analysis has been made with respect to two performance measures, namely, pick signal to noise ratio (PSNR) and feature similarity index (FSIM) [19]. The value of FSIM lies between 0 to 1. The larger value of PSNR and FSIM indicates better performance.

5.2 Experimental Analysis

To evaluate the performance of the proposed scheme, we compare our results with four schemes namely, Bicubic interpolation, EBSR [2], SRNE [5], and NeedFS [8].



(a) Parrot (b) Peppers (c) Lena (d) Tiger (e) Biker (f) Lotus

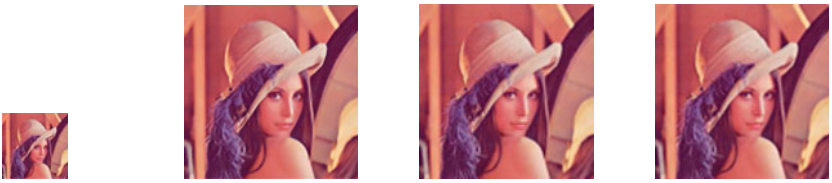
Fig. 2 Test images

Table 1 PSNR and FSIM results for test images with 3× magnification

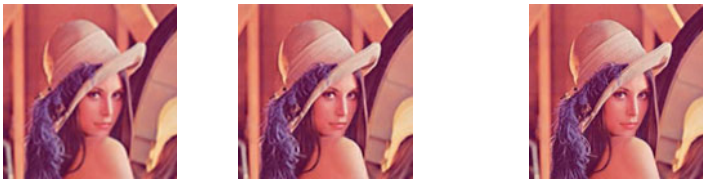
Images	Bicubic	EBSR [2]	SRNE [5]	NeedFS [8]	Proposed
Parrot	27.042	28.745	29.623	31.764	32.135
	0.8340	0.8458	0.8511	0.8603	0.8693
Peppers	28.756	29.137	30.969	32.111	33.249
	0.8397	0.8469	0.8582	0.8725	0.8839
Lena	29.899	30.117	31.826	33.026	34.762
	0.8527	0.8702	0.8795	0.8889	0.9023
Tiger	24.549	25.771	26.235	27.909	28.423
	0.8239	0.8394	0.8403	0.8519	0.8604
Biker	25.009	26.236	27.169	28.669	29.973
	0.8331	0.8481	0.8537	0.8601	0.8715
Lotus	26.829	27.787	28.979	30.276	31.862
	0.8338	0.8501	0.8637	0.8756	0.8904

(a) $Z_{11}R$ (b) $Z_{11}I$ (c) Z_{20}

Fig. 3 Three Zernike moments of *Lena* image



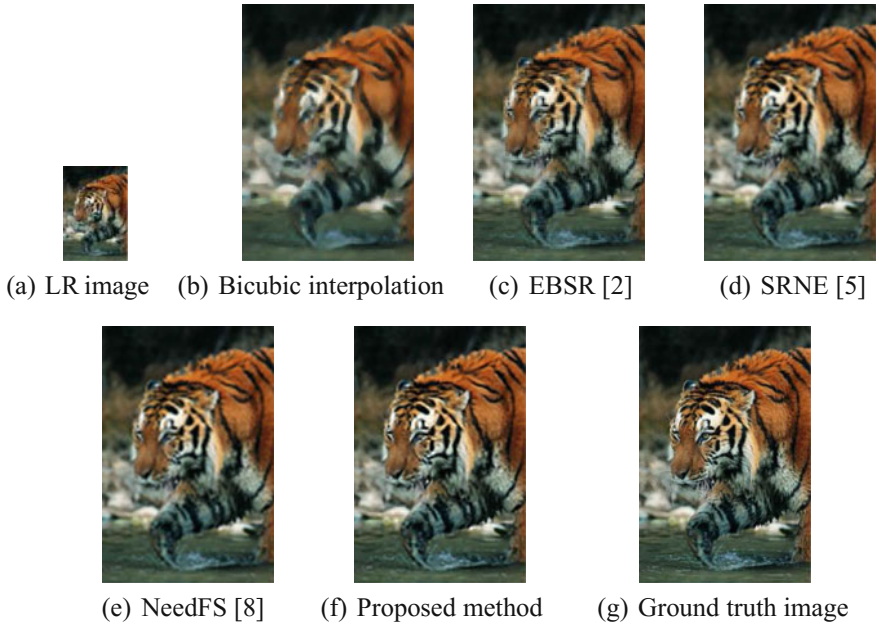
(a) LR image (b) Bicubic interpolation (c) EBSR [2] (d) SRNE [5]



(e) NeedFS [8] (f) Proposed method (g) Ground truth image

Fig. 4 Comparison of SR results(3×) of *Lena* image

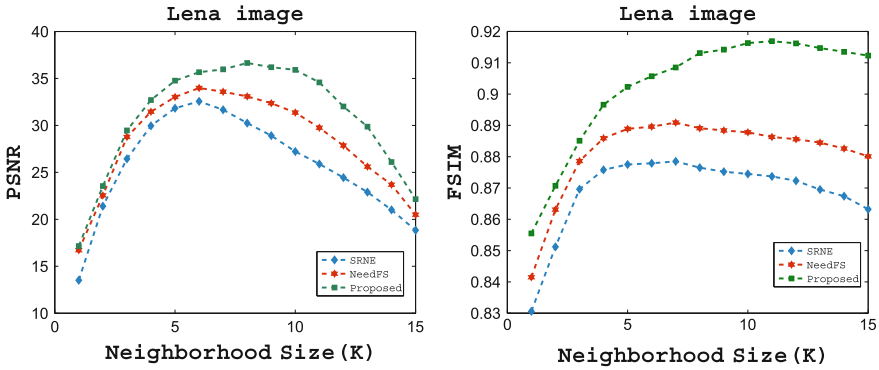
The test images are shown in Fig. 2. Table 1 lists the PSNR and FSIM values for all test images. The 1st row and 2nd row in the table indicates PSNR and FSIM values respectively. The features generated by Zernike moment for Lena image are shown in Fig. 3. The visual comparison for *Lena* and *Tiger* image are shown in Figs. 4 and 5 respectively. To validate the performance of the proposed scheme, we compare the results with state-of-the-art approaches with different K value. In SRNE [5], the K value is fixed which leads to blurring effect in the expected HR image; whereas in NeedFS [8] two different K values are provided according to the patches having edge. In our scheme, the K value lies between 1 to 15. Due to global neighborhood selection, our method gives a better results in terms of both PSNR and FSIM as shown in Fig. 6. It shows the graph is increased gradually between the K value 5 to 9. However, it gives only good results for a certain K value in the state-of-the-arts.



(a) LR image (b) Bicubic interpolation (c) EBSR [2] (d) SRNE [5]

(e) NeedFS [8] (f) Proposed method (g) Ground truth image

Fig. 5 Comparison of SR results(3×) of *Tiger* image



(a) PSNR (b) FSIM

Fig. 6 PSNR and FSIM comparison of *Lena* image

6 Conclusion

In this paper, we have proposed a new feature selection method for neighbor embedding based super-resolution. The feature vector is generated by combining three Zernike moments with the luminance norm of the image. The global neighborhood size selection technique is used to find the K value for K-nearest neighbor search.

Both qualitative and quantitative comparison of the proposed method is carried out with the state-of- the-art methods. The results show that the proposed method is superior to the other methods in terms of PSNR and FSIM values. However, for texture based image edge preservation is still an issue that will be addressed in our future work.

References

1. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine **20**(3) (2003) 21–36
2. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer Graphics and Applications **22**(2) (2002) 56–65
3. Kim, K.I., Kwon, Y.: Example-based learning for single-image super-resolution. In: Proceedings of the 30th DAGM Symposium on Pattern Recognition. (2008) 456–465
4. Li, D., Simske, S.: Example based single-frame image super-resolution by support vector regression. Journal of Pattern Recognition Research **1** (2010) 104–118
5. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 1. (2004) 275–282
6. Chan, T.M., Zhang, J.: Improved super-resolution through residual neighbor embedding. Journal of Guangxi Normal University **24**(4) (2006)
7. Fan, W., Yeung, D.Y.: Image hallucination using neighbor embedding over visual primitive manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition. (June 2007) 1–7
8. Chan, T.M., Zhang, J., Pu, J., Huang, H.: Neighbor embedding based super-resolution algorithm through edge detection and feature selection. Pattern Recognition Letters **30**(5) (2009) 494–502
9. Liao, X., Han, G., Wo, Y., Huang, H., Li, Z.: New feature selection for neighbor embedding based super-resolution. In: International Conference on Multimedia Technology. (July 2011) 441–444
10. Mishra, D., Majhi, B., Sa, P.K.: Neighbor embedding based super-resolution using residual luminance. In: IEEE India Conference. (2014) 1–6
11. Gao, X., Zhang, K., Tao, D., Li, X.: Joint learning for single-image super-resolution via a coupled constraint. IEEE Transactions on Image Processing **21**(2) (2012) 469–480
12. Gao, X., Zhang, K., Tao, D., Li, X.: Image super-resolution with sparse neighbor embedding. IEEE Transactions on Image Processing **21**(7) (2012) 3194–3205
13. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.L.A.: Super-resolution using neighbor embedding of back-projection residuals. In: International Conference on Digital Signal Processing. (2013) 1–8
14. Gao, X., Wang, Q., Li, X., Tao, D., Zhang, K.: Zernike-moment-based image super resolution. IEEE Transactions on Image Processing **20**(10) (2011) 2738–2747
15. Teague, M.R.: Image analysis via the general theory of moments. Journal of the Optical Society of America **70** (1980) 920–930
16. Xiao-Peng, Z., Yuan-Wei, B.: Improved algorithm about subpixel edge detection based on zernike moments and three-grayscale pattern. In: International Congress on Image and Signal Processing. (2009) 1–4
17. Kouropteva, O., Okun, O., Pietikinen, M.: Selection of the optimal parameter value for the locally linear embedding algorithm. In: International Conference on Fuzzy Systems and Knowledge Discovery. (2002) 359–363

18. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (2000) 2323–2326
19. Zhang, L., Zhang, D., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing **20**(8) (2011) 2378–2386

Target Recognition in Infrared Imagery Using Convolutional Neural Network

Aparna Akula, Arshdeep Singh, Ripul Ghosh, Satish Kumar and H.K. Sardana

Abstract In this paper, deep learning based approach is advocated for automatic recognition of civilian targets in thermal infrared images. High variability of target signature and low contrast ratio of targets to background makes the task of target recognition in infrared images challenging, demanding robust adaptable methods capable of capturing these variations. As opposed to the traditional shallow learning approaches which rely on hand engineered feature extraction, deep learning based approaches use environmental knowledge to learn and extract the features automatically. We present convolutional neural network (CNN) based deep learning framework for automatic recognition of civilian targets in infrared images. The performance evaluation is carried on infrared target clips obtained from 'CSIR-CSIO moving object thermal infrared imagery dataset'. The task involves four categories classification one category representing the background and three categories of targets -ambassador, auto and pedestrians. The proposed CNN framework provides classification accuracy of 88.15 % with all four categories and 98.24 % with only three target categories.

Keywords Thermal infrared imaging · Deep learning · Target recognition · Convolutional neural network

1 Introduction

With the advancement of computer technology and availability of high end computing facilities, research in the area of recognition is gaining momentum across wide range of applications like defense [1, 2], underwater mine [3], face recognition, etc. Target recognition is a crucial area of research from security point of view. Generalized recognition system consists of two stages, feature extraction stage

A. Akula (✉) · A. Singh · R. Ghosh · S. Kumar · H.K. Sardana
CSIR-Central Scientific Instruments Organisation (CSIR-CSIO),
Chandigarh 160030, India
e-mail: aparna.akula@csio.res.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_3

followed by a classifier stage. The feature extraction stage takes the detected target region and performs computation to extract information in the form of features. This information is fed to the classifier which categorizes the target to the most relevant target class. The performance of recognition algorithms is highly dependent on the extracted features. Imaging systems which capture data in visible spectrum fail to perform during night time and under dark conditions. It needs strong artificial illumination to capture data [4, 5]. Thermal infrared imaging systems which work in the infrared band of the electromagnetic spectrum sense the heat released by the objects above absolute zero temperature and form an image [6], thereby capable of working in no light conditions. The heat sensing capability of thermal infrared imaging make it superior over visible imaging [7, 8]. However, variability of target infrared signatures due to a number of environment and target parameters, pose challenge to researchers working towards development of automated recognition algorithms [9]. In this paper we present a robust recognition framework for target recognition in infrared images.

2 Related Work

The recent trends in recognition show researchers employing neural network based approaches. These approaches learn from experience. Similar to human brain system, the neural networks extract the information from the external environment. These approaches have been widely applied in character recognition [10], horror image recognition [11], face recognition [12] and human activity recognition [13]. These methods are providing better performances than the classical methods [14]. We can be broadly classify these methods into shallow learning and deep learning methods.

Commonly used learning based classifiers such as support vector machine (SVM), radial basis function neural network (RBFNN), k nearest neighbor method (k-NN), modular neural network (MNN) and tree classifier are the shallow learning methods that considers hand engineered features using some of the commonly used methods local binary pattern [15], principal component analysis, shift invariant feature transform (SIFT) [16] and histogram of oriented gradients (HOG) [17]. The feature selection is time consuming process and we need to specifically work towards identifying and tuning features that are robust for particular application. On the other hand, deep learning based methods employs learning based feature extraction using hierarchical layers, providing a single platform for feature extraction and classification [18]. Convolution neural network which is a deep learning method has shown state of the art performances in various applications such as MNIST handwriting dataset [19], Large Scale Visual Recognition Challenge 2012 [20], house number digit classification [21]. Convolution neural network is also shown to provide more tolerance to variable conditions such as pose, lightning, surrounding clutter [22].

The work in this paper is aimed at presenting an infrared target recognition framework by employing a deep learning approach. The rest of the paper is as follows; Sect. 3 describes an overview of convolution neural network, Sect. 4 provides a brief about the experimental data, Sect. 5 describes the proposed convolutional neural network design and Sect. 6 presents the results and analysis.

3 An Overview of Convolution Neural Network

Convolution neural networks (CNN) are feed forward neural networks having hierarchy of layers. They combine the two stages of recognition, feature extraction and classification stages in a single architecture. Figure 1 shows a representative architecture of deep convolution neural network. Feature extraction stage consists of convolution layers and subsampling layers. Both layers have multiple planes which are called feature maps. Typically the networks may have multiple feature extraction stages. A feature map is obtained by processing the input image or previous layer image with the kernel. The operation may be the convolution (as in convolution layer) or subsampling (averaging or pooling, as for subsampling layer). Each pixel in a feature map is a neuron. A neuron in a convolution layer receives the weighted sum of inputs (convolved result of the local receptive field of the input image with kernel) from the previous neuron and a bias, which is then passed

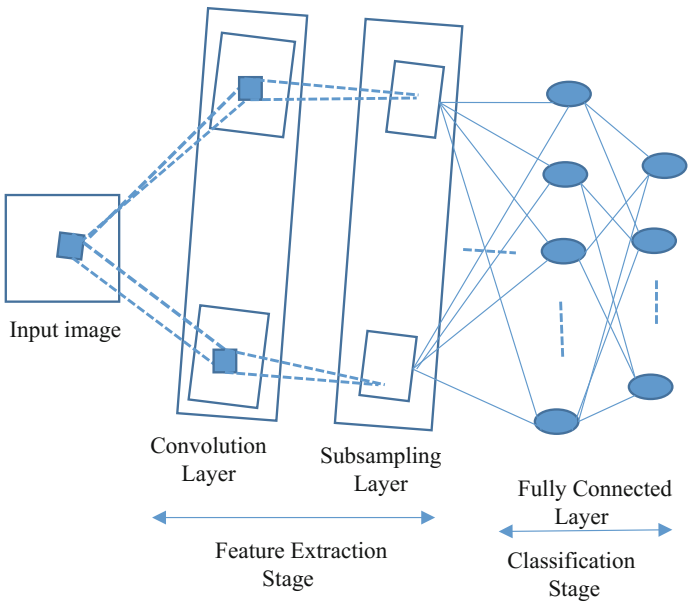


Fig. 1 Deep convolution neural network architecture

through a linear activation function. A neuron in subsampling layer receives the weighted value of averages or maximum value from the local region and a bias, which is then passed to non-linear activation function. The last stage is the fully connected layer representing the classification stage which is similar to multilayer perceptron model (MLP). It has hidden and output layer. A neuron in the fully connected layer has weighted sum and a bias as the input and has a non-linear activation function. The neurons of the output layer correspond to the one of the categories of target objects. CNN architecture has three main properties, weight sharing, spatial subsampling and local receptive field. The neurons of the next layer feature map use the same kernel to extract the same information from all parts of the image. Similarly, other kernels extract other information. Weight sharing reduces the number of free parameters. The kernel performs the computation on the local regions. The spatial or temporal subsampling reduces the size of the data and makes the network invariant to small changes [23].

The number of feature maps for first convolutional layer might vary from 3 to 1600. For a given binary image, number of possible binary feature map is given as 2^{r^2} where r is the receptive field width or kernel width. Number of useful feature map (lower bound on u) is:

$$u = h(r) + s \tag{1}$$

$$h(r) = \begin{cases} \frac{r^2+1}{2}, \; if \; r \; is \; odd \\ \frac{r^2}{2} + 1, if \; r \; is \; even \end{cases} \tag{2}$$

And s is minimum of h, s is 1. For subsequent feature map

$$u_l = v u_{l-1} \tag{3}$$

where v is a constant either 2 or 4 [24].

The number of fully connected hidden layer neuron is 2/3 (or 70 % to 90 %) of the size of the input layer. The number of hidden neurons can be added later on depending upon the complexity of the data [25].

4 Experimental Dataset

'CSIR-CSIO Moving Object Thermal Infrared Imagery Dataset' is used for validating the performance of the proposed deep learning framework [26]. Detected target regions obtained from the moving target detection method presented in [9] are used to train and test the system. It was observed that the detection algorithm presented some false positives where background was also detected as moving target. To handle this, in this work, we have designed a four class classifier system, one class representing background and three classes representing targets—ambassador, auto-rickshaw and pedestrian.

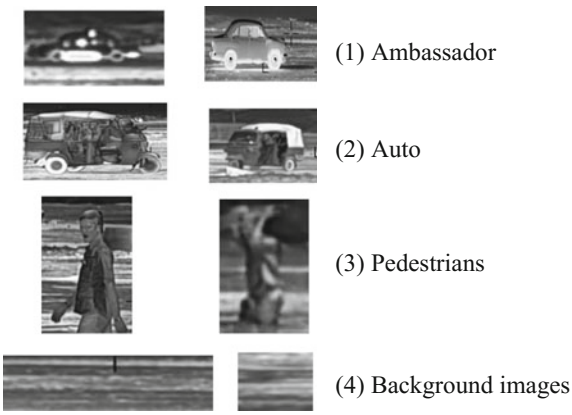


Fig. 2 Experimental dataset

Some representative images of the experimental dataset are shown in Fig. 2. We have a total of 378 images in four categories having 108 ambassadors, 108 autos, 82 pedestrians and 80 background images having different resolution. The total dataset is divided into training and testing datasets, 80 % for training and 20 % for testing. The training dataset has 302 images (89 ambassador, 89 autos, 63 pedestrians and 61 background images) and testing dataset has 76 images (19 images of each category). The data has high inter-class variability due to variations in scale, pose and temperature of targets.

5 Proposed System Design

The CNN architecture adapted from the LeNet-5 architecture [23] as shown in Fig. 1 is designed. We have used ten layers in the architecture. Before training of the neural network, all the experimental dataset is resized into a square of 200 × 200. Also the input image pixel values are normalized to zero mean and 1 standard deviation according to Eq. (4). The normalization improves the convergence speed of the network [27].

$$x(new) = \frac{x(old) - m}{sd} \tag{4}$$

where $x(new)$ is the preprocessed pixel value, $x(old)$ is the original pixel value, m is the mean value of pixel from the image and sd is the standard deviation of pixels. $x(new)$ is the new zero mean and 1 standard deviation value. First the original image is resized to 200 × 200 and then normalized.

The feature maps of each layer are chosen according to the Eqs. (1) and (3). The first layer is the dummy layer having sampling rate 1. This is just for the symmetry of architecture. The second layer is convolutional layer (C2) with 6 feature maps.

The size of kernel for this layer is 3 × 3. The size of each feature map in C2 layer is 198 × 198. The third layer is subsampling layer (S3) with 6 feature maps. The subsample rate is 2. The connections from C2-layer to S3 layer are one to one. The size of each feature map is 99 × 99. The fourth layer is convolutional layer (C4) with 12 feature map each of size 96 × 96. The kernel size is 4 × 4. The connections from S3 to C4 layer are random. The fifth layer is subsampling layer (S5) with 12 feature maps. The subsample rate is 4. The connections from C4 layer to S5 layer are one to one. The size of each feature map is 24 × 24 in S5 layer. The sixth layer is convolutional layer (C6) with 24 feature map each of size 20 × 20. The kernel size is 5 × 5. The connections from S5 to C6 layer are random. The seventh layer is subsampling layer (S7) with 24 feature maps. The subsample rate is 4. The connections from C6 layer to S7 layer are one to one. The size of each feature map is 5 × 5 in S7 layer. The eighth layer is convolutional layer (C8) with 48 feature map each of size 1 × 1. The kernel size is 5 × 5. The connections from S7 to C8 layer are random. The fully connected layer has random number of hidden neurons which are varied while performing simulation. The output layer has 4 neurons corresponding to four categories.

The kernel or the weights and bias for convolution layers are initialized randomly. The weights of subsampling layer are initialized with unit value and zero bias. The activation function for convolutional layer neurons is linear and scaled bipolar sigmoidal for all other neurons. Scaled activation function is given in Eq. (5).

$$F(n) = 1.7159 * \left(\frac{2}{(1 + \exp(-1.33*\mathrm{n}))} - 1 \right) \tag{5}$$

where n is the weighted output of neuron. $F(n)$ is the output obtained after applying activation function. The number of neurons in the output layer are equal to the number of categories. The neural network is trained in such a way that the true category neuron corresponds to +1 and others to –1. The bipolar sigmoidal function is scaled to 1.7159 and the slope control constant is set to 1.33 as used by [23]. The scaling improves the convergence of the system.

The network is trained with Stochastic Diagonal Levenberg-Marquardt method [28]. At each kth learning iteration, free parameter w_k is updated according to Eq. (6) stochastic update rule.

$$w_k(\mathrm{k}+1) \leftarrow w_k(\mathrm{k}) - \in_k \frac{\partial E^P}{\partial w_k} \tag{6}$$

where $w_k(\mathrm{k}+1)$ the weight at $k+1$ iteration, E^p is the instantaneous loss function for pth pattern, $\in_k$, the step size or adaptive learning constant

$$\in_k = \frac{\eta}{\mu + h_{kk}} \tag{7}$$

where, η is the constant step size which is controlled by the second order error term μ is the hand-picked constant, and h_{kk} is the estimate of the second order derivative of loss function w.r.t the connection weights u_{ij} as given in Eq. (8).

$$h_{kk} = \sum \frac{\partial^2 E}{\partial u_{ij}^2} \tag{8}$$

In the proposed architecture, we used the constant step size $\eta = 0.0005$, $\mu = 0.001$, stopping Criterion is average root mean square error per epoch less than or equal to 0.09 or number of epochs exceeds more than 50 and the loss function is average root mean square error. The smaller values of μ and η prevent the step size from becoming very large [23].

6 Results and Analysis

The development environment of the proposed method is MATLAB® R2014a on a 64 bit, Intel® Core™ i5 CPU 650 @3.20 GHz processor with 2 GB RAM configuration. The performance of the trained system is validated with the test dataset. Also, the significance of the number of hidden neurons in fully connected layer is studied by varying them from 28 to 36 while keeping the rest of the architecture constant.

Figure 3 is the convergence plot between average root mean square error (RMSE) per epoch versus number of epochs when number of hidden neurons in fully connected layer are 32. The network is converging after 10 epochs. The generalization of trained system is verified with the test dataset. Table 1 gives the confusion matrix. It can be observed that the network is classifying all of the ambassador and auto targets accurately and one of the pedestrians is misclassified. However around 40 % of the background images are misclassified and falling into ambassador category.

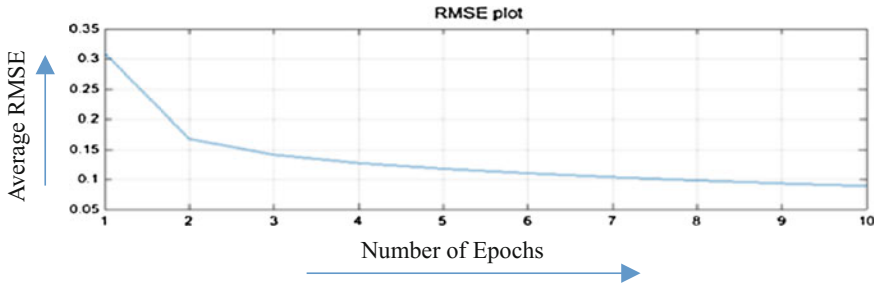


Fig. 3 Convergence plot

Table 1 Confusion matrix

		Predicted class			
		Background images	Ambassador	Auto	Pedestrians
Actual class	Background images	11	8	0	0
	Ambassador	0	19	0	0
	Auto	0	0	19	0
	Pedestrians	0	0	1	18

Table 2 Different Architecture Results

Sr. No.	Neurons in fully connected layer	Training time (In epoch)	Accuracy with background images (%)	Accuracy without background images (%)
1	28	11	86.84	96.49
2	32	10	88.15	98.24
3	36	10	85.53	96.49

Table 2 shows the performance results of different architectures obtained after varying the hidden neurons in fully connected layer. The accuracy is measured with and without background images. It can be observed that accuracy is better with 32 hidden neurons architecture. It was observed that the network with 28 units has underfitting problem while the network with 36 neurons has overfitting problem. The optimal number of hidden units seems to be 32 for the proposed system. The proposed system with all the architectures is classifying all ambassador and auto targets accurately and pedestrians with around 90 % accuracy. The misclassified pedestrians are falling into auto category. Background images are classified with accuracy around 57 %. The misclassified background images are either falling to ambassador category or auto category. Misclassification of background images are primarily because of very low resolution and lack of any distinctive features as compared to objects. Also, the background information is part of the detected objects leading to categorise them into one of the target classes. The deep architecture is extracting the features locally. As more deep the architecture is, it would result in invariant features. The subtle differences of non-uniform intensity of objects and background might result in the very similar features as that of ambassador and auto objects. However, the uniform intensity targets (pedestrians) have noticeable differences from the background, so background images are not falling into this category.

7 Conclusion and Future Scope

The preliminary results reported in this work, demonstrate that deep learning based automatic feature extraction and classification system can accurately classify civilian targets in infrared imagery. The proposed system could classify the

non-uniform intensity-vehicle and uniform intensity-pedestrian targets with an accuracy of 98.24 % in different experimental conditions. It is observed that the accuracy is reduced in case of background images. The rationale behind this might be due to the low resolution of background images and also due to the correlation between the non-uniform intensity images and background as the target images contain information of the background as well. In future work the work shall be extended to analyze its performance for more complex data set such as physically occluded, scaled or rotated test data set. Also the system can be designed to categorize more number of target classes. The real time implementation and optimization of proposed design in terms of processing is also planned.

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.

References

1. J. G. Verly, R. L. Delanoy, and D. E. Dudgeon, "Machine Intelligence Technology for Automatic Target Recognition," *The Lincoln Laboratory Journal*, vol. 2, no. 2, pp. 277–310, 1989.
2. A. Arora, P. Dutta, S. Bapat, V. Kulathumani, H. Zhang, V. Naik, V. Mittal, H. Cao, M. Demirbas, M. Gouda, Y. Choi, T. Herman, S. Kulkarni, U. Arumugam, M. Nesterenko, A. Vora, and M. Miyashita, "A line in the sand: a wireless sensor network for target detection, classification, and tracking," *Computer Networks*, vol. 46, no. 5, pp. 605–634, 2004.
3. D. Kraus and A. M. Zoubir, "Contributions to Automatic Target Recognition Systems for Underwater Mine Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 1, pp. 505–518, 2015.
4. S. G. Narasimhan and S. K. Nayar, "Shedding light on the weather," *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003 Proceedings*, vol. 1, 2003.
5. S. K. Nayar and S. G. Narasimhan, "Vision in bad weather," *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 820–827, 1999.
6. J. M. Lloyd, *Thermal imaging systems*. Springer Science & Business Media, 2013.
7. M. Vollmer, J. A. Shaw, and P. W. Nugent, "Visible and invisible mirages: comparing inferior mirages in the visible and thermal infrared," *in journal of applied optics*, vol. 54, no. 4, pp. B76–B84, 2014.
8. Y. Fang, K. Yamada, Y. Ninomiya, B. Horn, and I. Masaki, "Comparison between infrared-image-based and visible-image-based approaches for pedestrian detection," *IEEE IV2003 Intelligent Vehicles Symposium Proceedings (Cat No03TH8683)*, pp. 505–510, 2003.
9. A. Akula, R. Ghosh, S. Kumar, and H. K. Sardana, "Moving target detection in thermal infrared imagery using spatiotemporal information.," *Journal of the Optical Society of America A, Optics, image science, and vision*, vol. 30, no. 8, pp. 1492–501, 2013.
10. M. Khayyat, L. Lam, and C. Y. Suen, "Learning-based word spotting system for Arabic handwritten documents," *Pattern Recognition*, vol. 47, no. 3, pp. 1021–1030, 2014.
11. B. Li, W. Hu, W. Xiong, O. Wu, and W. Li, "Horror Image Recognition Based on Emotional Attention," in *Asian Conference on Computer Vision (ACCV)*, 2011, pp. 594–605.
12. S. Z. Li, L. Zhang, S. Liao, X. X. Zhu, R. Chu, M. Ao, and H. Ran, "A Near-infrared Image Based Face Recognition System," *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pp. 455–460, 2006.

13. V. Elangovan and A. Shirkhodaie, "Recognition of human activity characteristics based on state transitions modeling technique," p. 83920 V–83920 V–10, May 2012.
14. B. Li, R. Chellappa, R. Chellappa, Q. Zheng, Q. Zheng, S. Der, S. Der, N. Nasrabadi, N. Nasrabadi, L. Chan, L. Chan, L. Wang, and L. Wang, "Experimental evaluation of FLIR ATR approaches—A comparative study," *Computer Vision and Image Understanding*, vol. 84, pp. 5–24, 2001.
15. T. Ahonen, A. Hadid, M. Pietikäinen, S. S. Member, and M. Pietika, "Face description with local binary patterns: application to face recognition.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–41, 2006.
16. A. P. Psyllos, C. N. E. Anagnostopoulos, and E. Kayafas, "Vehicle logo recognition using a sift-based enhanced matching scheme," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 322–328, 2010.
17. O. Déniz, G. Bueno, J. Salido, and F. De La Torre, "Face Recognition Using Histograms of Oriented Gradients," *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598–1603, 2011.
18. I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning — A New Frontier in Artificial Intelligence Research," *IEEE Computational Inteligence Magazine*, vol. 5, no. November, pp. 13–18, 2010.
19. D. Ciresan, "Multi-column Deep Neural Networks for Image Classification," in *Computer Vision and Pattern Recognition, IEEE*, 2012, pp. 3642–3649.
20. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.
21. P. Sermanet, S. Chintala, and Y. LeCun, "Convolutional neural networks applied to house numbers digit classification," *Proceedings of International Conference on Pattern Recognition ICPR12*, pp. 10–13, 2012.
22. Y. LeCun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2004, vol. 2, pp. 97–104.
23. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2323, 1998.
24. J. L. Chu and A. Krzy, "Analysis of Feature Maps Selection in Supervised Learning Using Convolutional Neural Networks," in *Advances in Artificial Intelligence*, Springer International Publishing, 2014, pp. 59–70.
25. S. Karsoliya, "Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture," *International Journal of Engineering Trends and Technology*, vol. 3, no. 6, pp. 714–717, 2012.
26. A. Akula, N. Khanna, R. Ghosh, S. Kumar, A. Das, and H. K. Sardana, "Adaptive contour-based statistical background subtraction method for moving target detection in infrared video sequences," *Infrared Physics & Technology*, vol. 63, pp. 103–109, 2014.
27. S. Ioffe and C. Szegedy, "Batch Normalization : Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv preprint* arXiv:150203167v3, 2015.
28. H. Yu and B. M. Wilamowski, "Levenberg-Marquardt training," *Industrial Electronics Handbook, vol 5—Intelligent Systems*, pp. 12–1 to 12–18, 2011.

Selected Context Dependent Prediction for Reversible Watermarking with Optimal Embedding

Ravi Uyyala, Munaga V. N. K. Prasad and Rajarshi Pal

Abstract This paper presents a novel prediction error expansion (PEE) based reversible watermarking using 3×3 neighborhood of a pixel. Use of a good predictor is important in this kind of watermarking scheme. In the proposed predictor, the original pixel value is predicted based on a selected set, out of the eight neighborhood of a pixel. Moreover, the value of prediction error expansion (PEE) is optimally divided between current pixel and top-diagonal neighbor such that distortion remains minimum. Experimental results show that the proposed predictor with optimal embedding outperforms several other existing methods.

Keywords Reversible watermarking · Prediction error expansion · Optimized embedding · Watermarking scheme

1 Introduction

Multimedia has become more popular in modern life with rapid growth in communication systems. Digital media is widely being used in various applications and is being shared in various forms with sufficient security. But advancements of signal processing operations lead to malicious attacks and alterations of these contents. Watermarking is a technique to protect these contents or the ownership of the contents against such malicious threats. A watermark is embedded in the multimedia content to achieve the said purpose. Later, the same watermark is extracted to

R. Uyyala (✉) · M.V.N.K. Prasad · R. Pal
Institute for Development and Research in Banking Technology, Hyderabad, India
e-mail: uyyala.ravi@gmail.com

M.V.N.K. Prasad
e-mail: mvnkprasad@idrbt.ac.in

R. Pal
e-mail: iamrajarshi@yahoo.co.in

R. Uyyala
SCIS, University of Hyderabad, Hyderabad, India

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_4

establish the ownership or genuineness of the content. Moreover, in case of reversible watermarking, the original cover media is also restored back from the watermarked content along with extraction of watermark [1].

Several approaches for reversible watermarking can be found in the literature. A difference expansion (DE) based reversible watermarking scheme is proposed in [2], where the differences between pairs of pixels are expanded to insert the watermark. Later, several extensions of this method have been proposed. For example, watermark is embedded in quad of adjacent pixels in [3]. The differences between one pixel with three neighboring pixels are used to insert the watermark. Moreover, this difference expansion technique has been extended to vectors [4] instead of pairs of pixels. A high difference value between adjacent pixels may cause the overflow/underflow problems during the embedding. The intelligent pairing of pixels to tackle the overflow/underflow conditions has been exploited in [5].

In another variation of these difference expansion based approaches, prediction error is expanded to embed the watermark bits. Basically, the difference between an original pixel value and a predicted value (based on context of the pixel) is termed as prediction error. Watermark is hidden in the expansion of prediction error. Closeness between the original and predicted pixel values is the strength of this approach. It reduces the distortion and overflow/underflow cases as compared to previous difference expansion based approaches. Several approaches of predicting a pixel value have been adapted in this kind of techniques. For example, upper, left, and upper-left neighbors of a pixel is used in [6] to predict the value for the pixel. On the contrary, bottom, right, and bottom-right neighbors of a pixel is used to embed data in [7]. A simple average of four neighboring pixels predicts the center pixel value in [8]. Bilinear interpolation technique using four diagonal neighbors estimates the center pixel value in [9]. Few other notable approaches include utilizing of weights with median based prediction in [10] and adaptive prediction in [11].

Recently, it has been found that a gradient based predictor [12] outperforms several other methods. Gradients on four directions (horizontal, vertical, diagonal and anti diagonal) are estimated using 4×4 neighborhood. Four predictors estimate the value of the concerned pixel in each of these four directions. Finally, the weighted average of two such predicted values, as obtained for the directions of two smallest gradients, is considered to predict the pixel value. The weights are chosen as proportional to the gradient magnitude in the corresponding directions.

The work in [13] considers the uniformity of the neighborhood for predicting a pixel value. If four neighbors (two horizontal and two vertical neighbors) are uniform, then a simple average of these four pixel values computes the predicted value for the pixel. Otherwise, the neighbors are grouped into two different pairs (horizontal and vertical). The central pixel value is computed as the average of the more homogeneous group among these two groups. Unlike none of the above approaches, in [14], mathematical complexity of a predictor has been reduced by using a predictor for group of pixels, not for each individual pixel.

Similar to the work in [13], the proposed work presents a novel predictor based on the uniformity of the pixels in a neighborhood. But unlike the approach in [13], it considers eight-neighborhood of a pixel. The neighbors are grouped into four different pairs (horizontal, vertical, diagonal and anti diagonal). A strategy has been devised to consider some of these groups to predict the center pixel value. Less diversity among the pairs of pixels in each group and closeness between average values among these groups decide which of these can be considered for prediction. Coupled with an optimal embedding scheme, this proposed prediction error based reversible watermarking scheme outperforms not only the four-neighbor based method [13], but also scores better than the recent gradient based method in [12].

The outline of the paper is as follows: The proposed prediction scheme based on a select of the 8-neighborhood is depicted in Sect. 2. Proposed optimal embedding scheme is explained in Sect. 3. Extraction of watermark is described in Sect. 4. Experimental results are presented in Sect. 5. Finally, the conclusion is drawn in Sect. 6.

2 Proposed Selected Context Based Prediction Using Eight Neighborhood

The proposed uniformity based prediction scheme considers 8-neighborhood of a pixel as shown in Fig. 1. According to this method, the eight pixels in the neighborhood is divided into four groups containing the pair of horizontal, vertical, diagonal, and anti-diagonal neighbors. Diversity of the pair of pixels in a group has been measured by considering the absolute difference of the pixel value in the group. Let d_h, d_v, d_d, and d_a denote the diversity of the pixels in horizontal, vertical, diagonal, and anti-diagonal groups. Hence,

$$\begin{aligned} d_h &= |x_{m,n-1} - x_{m,n+1}| \\ d_v &= |x_{m-1,n} - x_{m+1,n}| \\ d_d &= |x_{m-1,n-1} - x_{m+1,n+1}| \\ d_a &= |x_{m-1,n+1} - x_{m+1,n-1}| \end{aligned} \tag{1}$$

Moreover, averages of the neighbors in any particular direction (horizontal, vertical, diagonal, and anti diagonal) are computed as:

$x_{m-1,n-1}$	$x_{m-1,n}$	$x_{m-1,n+1}$
$x_{m,n-1}$	$x_{m,n}$	$x_{m,n+1}$
$x_{m+1,n-1}$	$x_{m+1,n}$	$x_{m+1,n+1}$

Fig. 1 8-neighborhood of a pixel $x_{m,n}$

$$
\begin{aligned}
a_h &= \lfloor \frac{x_{m,n-1} + x_{m,n+1}}{2} \rfloor \\
a_v &= \lfloor \frac{x_{m-1,n} + x_{m+1,n}}{2} \rfloor \\
a_d &= \lfloor \frac{x_{m-1,n-1} + x_{m+1,n+1}}{2} \rfloor \\
a_a &= \lfloor \frac{x_{m-1,n+1} + x_{m+1,n-1}}{2} \rfloor
\end{aligned} \tag{2}
$$

Only homogeneous (less diverse) groups are considered for predicting the center pixel values. Hence, the group of pixels having the least diversity is considered for estimating the current pixel. Let four diversity values in Eq. 1 are sorted in non-decreasing order and let these be denoted as d_1, d_2, d_3, and d_4 (while d_1 is the smallest of these four values). Moreover, let the averages in these four directions (as computed in Eq. 2) be sorted in non-decreasing order of the diversities in respective directions and let the sorted values be a_1, a_2, a_3, and a_4. Here, the a_i corresponds to the direction having diversity d_i. Basically, these average values act as a predicted value in their respective directions. At first, the predicted value a_1 according to the least diverse group (with diversity value d_1) is considered to predict the central value. Additionally, the predictions in other directions are considered, only if the predicted (average) values in those directions are also close enough to the value a_1. A threshold T decides the closeness of these average values to the value a_1 (The value of T is assumed to be 1 for our experiments). To focus on the groups of less diverse pixels, closeness of these average values have been tested iteratively, starting with second least diverse group. This complete algorithm is mentioned in Algorithm 1, where the iteration has been broken down using if-else constructs for three other groups (apart from the least diverse group). Ultimately, if predicted (i.e.,average) values of all four groups are similar enough, then average of all four prediction values (i.e., average of individual groups) predicts the center pixel value.

3 Proposed Optimal Embedding Scheme

Before introducing the proposed scheme, basic principle of difference expansion is discussed here. Let x be a original pixel and $x^{'}$ be the estimated value of original pixel as computed on the selected 8-neighborhood (N^x) of x using the procedure in Sect. 2. The prediction error PE is computed as $\text{PE} = x - x^{'}$. Let, further w be the watermark bit which is to be added into the cover image pixel. The original pixel value x will be replaced by watermarked pixel value X, which is computed as

$$
X = x^{'} + 2 \times (x - x^{'}) + w = 2 \times x - x^{'} + w = x + x - x^{'} + w = x + PE + w \tag{3}
$$

This means that the estimated error and the watermark information are directly added to the pixel intensity of the cover image. At detection, the estimation of the

Algorithm 1 Predicting the Original Pixel from 8-Neighbor Context

Compute the diversities in various directions [d_h, d_v, d_d, and d_a] using equation 1.
Compute the averages of two neighbors of various directions [a_h, a_v, a_d, and a_a] using equation 2.
Sort the diversity values in non-decreasing order and store them as [d_1, d_2, d_3, and d_4].
Sort the average values in non-decreasing order of diversities in respective direction and store them as [a_1, a_2, a_3, and a_4].
// Difference (D) between average of first two groups having minimum diversity
$D = |a_1 - a_2|$
if $D \geq T$ **then**
    // The second group is not considered.
    // Estimated value is the average of first group.
    $x^{'} = a_1$
else
    // Difference between average of first and third groups having minimum diversity
    $D = |a_1 - a_3|$
    if $D \geq T$ **then**
        // Third group is not considered, only first and second groups are considered.
        // Estimated value is the average of first and second groups
        $x^{'} = \lfloor \frac{a_1 + a_2}{2} \rfloor$
    else
        // Difference between average of first and fourth groups having minimum diversity.
        $D = |a_1 - a_4|$
        if $D \geq T$ **then**
            // Fourth group is not considered, only first, second, and third groups are considered.
            // Estimated value is the average of first, second and third groups
            $x^{'} = \lfloor \frac{a_1 + a_2 + a_3}{3} \rfloor$
        else
            // Estimated value is the average of all groups
            $x^{'} = \lfloor \frac{a_1 + a_2 + a_3 + a_4}{4} \rfloor$
        end if
    end if
end if

watermarked pixel, should not be changed so that the prediction error is generated from the watermarked pixels. Based on the estimated value of the watermarked pixel, the prediction error is determined. Then, the watermark bit w is taken as the least significant bit of the $X - x^{'}$, namely

$$w = (X - x^{'}) - 2 \times (\lfloor \frac{X - x^{'}}{2} \rfloor) \tag{4}$$

Then, the original cover image pixel x is computed as

$$x = \frac{X + x^{'} - w}{2} \tag{5}$$

Let $PW = PE + w$. From the above discussion (specifically, Eq. 3) it can be observed that the entire value of PW is added into the current pixel. Hence, the distortion in this pixel is PW.

This paper, similar to [15], proposes to bring down this amount of distortion by breaking it into two parts. A part of PW is added to the top-diagonal pixel ($x_{m-1,n-1}$). This part d of PW is computed as

$$d = \lfloor (L \times PW + 0.5) \rfloor. \tag{6}$$

where, $0 \leq L \leq 1$. After inserting some optimal value d into the top-diagonal pixel, the remaining amount ($PW1 = PW - d$) is added to its original pixel value.

$$x = x + PW1 \tag{7}$$

Now the context has been modified due to addition of value d into top-diagonal pixel. The modified context N_d^x can be written as

$$N_d^x = f(N^x, d). \tag{8}$$

Hence, the new predicted value $X^{'}$ is estimated based on the modified context N_d^x using the procedure in Sect. 2. Now, in order to extract the watermark data and original cover image pixel, the difference between the watermarked pixel and the estimated value must remain same. To prevent the change in this difference, the new watermarked value X_d has to be computed as follows.

$$X_d = X^{'} + X - x^{'} \tag{9}$$

To obtain the optimal fraction value L, the optimization of the embedding error is determined in [15] based on the minimum square error (MSE).

$$MSE = (x - X_d)^2 + \sum_{i}^{j} (N^x(i,j) - N_d^x(i,j))^2 \tag{10}$$

where x and X_d are original and watermarked pixel values. Moreover, N_x and N_x^d are original and modified context. The above equation in proposed case can be rewritten as

$$MSE = (x - X_d)^2 + (x_{m-1,n-1} - (x_{m-1,n-1} + d))^2. \tag{11}$$

Moreover, as X_d is the new watermarked pixel value, the top-diagonal and current pixels are modified during embedding. Based on the above equation the minimum value of d is obtained as $1/2PW$. Equivalently $L = 0.50$. Thus, the function f in Eq. 8 splits the data between the present pixel and its context, whereas L controls the optimal fraction of d to be embedded in the top-diagonal pixel. The optimal embedding is used in the pixel locations where the prediction error (PE) falls within

a range of threshold ($-T_E$, T_E). This controls the embedding distortion by embedding the low values of prediction error (PE). If PE falls in between $-T_E$ and T_E, then the watermark bit is embedded into the prediction error. If the value of PE is greater than T_E, then an amount of T_E is added to the pixel. If PE is less than $-T_E$, then an amount of T_E is subtracted from the pixel.

Thus, this embedding process (as depicted in Algorithm 2) is repeated for every pixel in the image in a raster scan sequence.

Algorithm 2 Embedding the Watermark bit in a pixel

Input: Cover image pixel x, Predicted pixel value $x^{'}$, Watermark bit w, L=0.5, Threshold Value T_E
$PE = x - x^{'}$
if ($PE >= -T_E$**and**$PE <= T_E$) **then**
    // embedding of watermark is done
    $PW = PE + w$
    $X = x + PW$
    $d = \lfloor (L \times PW + 0.5) \rfloor$
    $PW1 = PW - d$
    // $PW1$ is added to current pixel.
    $x = x + PW1$
    // d is added to diagonal pixel.
    $x_{m-1,n-1} = x_{m-1,n-1} + d$
    The estimated value on the modified context N_d^x using the prediction scheme mentioned in the Sect. 2 is $X^{'}$.
    // The value of x is changed to X_d
    $X_d = X^{'} + X - x^{'}$
else
    // The pixel values are shifted without embedding the watermark.
    if $PE < -T_E$ **then**
        // Subtracting an amount of T_E from x.
        $X_d = x - T_E$
    else
        if $PE > T_E$ **then**
            // Adding an amount of T_E to x
            $X_d = x + T_E$
        end if
    end if
end if

4 Extraction of Watermark

As the embedding is carried out in raster scan order, the extraction is performed in opposite order, from lower right to top left. The estimated value $X^{'}$ is computed using the predictor in Sect. 2 from the context of a pixel in watermarked image.

Then, the reversibility of the modified scheme is immediately follows. The amount of prediction that can be recovered at detection is

$$PE1 = X_d - X' \tag{12}$$

The embedded watermarked data w is computed as follows:

$$w = (PE1) - 2 \times (\lfloor \frac{PE1}{2} \rfloor) \tag{13}$$

The optimal fraction value d can be computed as

$$d = \lfloor (L \times PW + 0.5) \rfloor \tag{14}$$

where, PW is computed as follows:

$$PW = (PE1 + w)/2 \tag{15}$$

The original context can be recovered by inverting the function f at detection as follows.

$$N^x = f^{-1}(N_d^x, d) \tag{16}$$

Basically, d is subtracted from top-diagonal neighbor and added to current pixel

$$X_{d(m-1,n-1)} = X_{d(m-1,n-1)} - d \tag{17}$$

Then, after generating the original context from function f^{-1} and after computing PW, the original pixel x can be computed as follows;

$$x = X_d - PW + d. \tag{18}$$

This extraction procedure is shown in Algorithm 3. If prediction error ($PE1$) falls between $-2T_E$ and $2T_E$ then, the watermark data and original data will be extracted. If the prediction error ($PE1$) is less than $-2T_E$ then, the same amount of value, which has been subtracted at embedding, should be added. If prediction error ($PE1$) is greater than $2T_E$ then, the same amount is subtracted so that the original data is recovered.

As the extraction process works on a sequence of pixels which is reverse of the embedding sequence, the extracted sequence of watermark bits is reversed to get the original watermark data.

Algorithm 3 Extracting the Watermark bit from a Pixel

Input: The watermarked pixel is $= X_d$
Estimate the value $X^{'}$ for the watermarked image pixel using the predictor in Algorithm 1
The Prediction Error $PE1 = X_d - X^{'}$
if $PE1 <= 2T_E$ **and** $PE1 >= -2T_E$ **then**
    The watermark bit (w)=$(PE1) - 2(\lfloor (PE1)/2 \rfloor)$
    The prediction error expansion $PW = (PE1 + w)/2$
    d=$\lfloor (L \times PW + 0.5) \rfloor$
    $X_{d(m-1,n-1)} = X_{d(m-1,n-1)} - d$
    The original pixel (x)=$(X_d - PW + d)$
else
    if $PE1 > 2T_E$ **then**
        $x = X_d - T_E$
        // Subtracted an amount of T_E from watermarked values
    else
        if $PE1 < -2T_E$ **then**
            $x = X_d + T_E$
            // Added an amount of T_E to watermarked values
        end if
    end if
end if

5 Experimental Results

In this section, experimental results for the proposed reversible watermarking based on eight neighborhood with optimal embedding are presented. Standard four test images (Lena, Barbara, Mandrill, and Boat) of size 512×512 pixels are considered for evaluation. These images are shown in Fig. 2. Peak-signal-to-noise ratio (PSNR) between the cover image and the watermarked image is used as evaluation metric. It quantifies the distortion in the watermarked image due to the watermark embedding. The outcome of the proposed method is compared with the outcomes of the extended gradient based selective weighting (EGBSW) [12] and rhombus average [13]. The results are compared for various embedding rates. The proposed method outperforms both of these methods at various embedding rates as it can be observed from the values in Table 1. Higher PSNR value indicates better result. The comparison among these methods using PSNR value at various embedding rates has also been plotted in Fig. 3. Moreover, it has also been observed that the original image can be perfectly restored back after extracting the watermark.

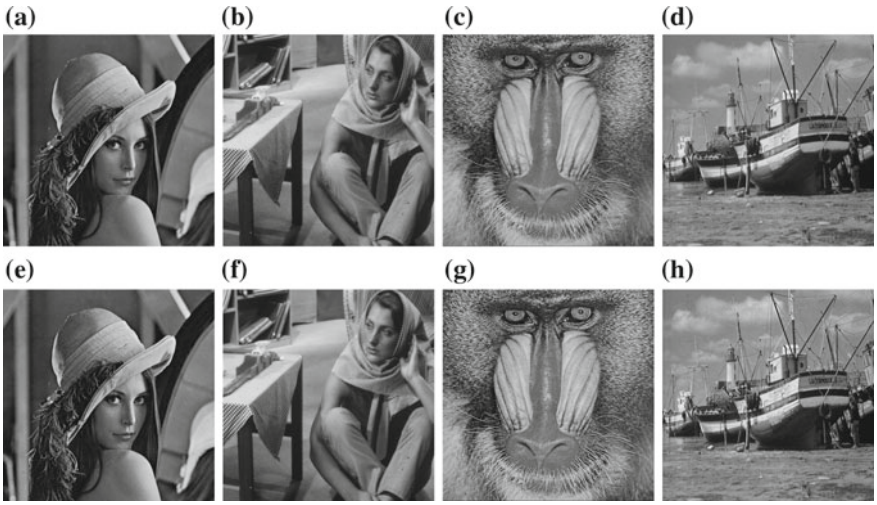


Fig. 2 Test Images (Cover Images (**a**–**d**), Watermarked Images (**e**–**h**) with embedding rate of 0.8 bpp): Lena, Barbara, Mandrill, and Boat

Table 1 PSNR at various embedding rates

Test image	Bit-rate (bpp)	Rhombus average [13] (dB)	EGBSW [12] (dB)	Proposed predictor (dB)
LENA	0.2	50.03	50.2	52.30
	0.4	45.11	45.32	46.14
	0.6	41.17	41.45	41.50
	0.8	38.14	38.51	38.98
MANDRILL	0.2	41.31	41.33	45.65
	0.4	34.84	35	38.22
	0.6	30.51	30.92	33.32
	0.8	26.3	27.36	28.04
BARBARA	0.2	48.7	48.94	52.31
	0.4	42.54	42.99	45.67
	0.6	38.29	38.9	40.69
	0.8	33.75	34.66	34.76
BOAT	0.2	46.70	46.94	48.13
	0.4	40.54	40.66	41.08
	0.6	35.29	35.60	36.92
	0.8	31.75	31.80	32.99

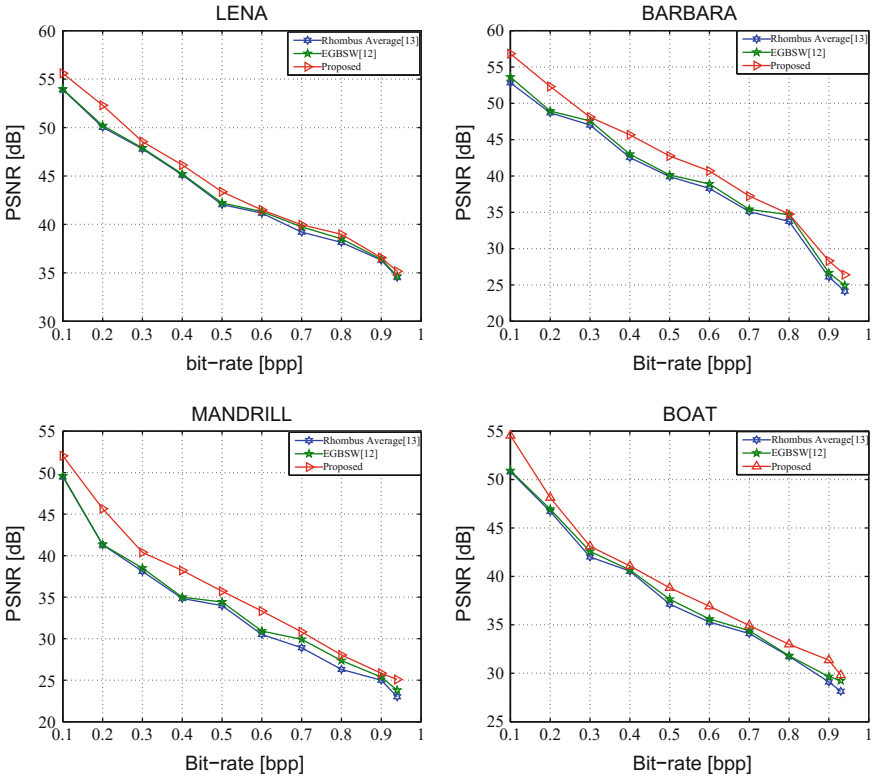


Fig. 3 Embedding rate versus PSNR comparison

6 Conclusion

A novel reversible watermarking based on a selection of context with optimal embedding is proposed in this paper. The proposed approach produces promising result as it delivered a better prediction scheme based on a selection of context pixels depending on the diversity and average value of neighboring pixel pairs in a direction. The optimal embedding procedure also helps to bring down the distortion further. It has been experimentally observed that the proposed scheme outperforms the some state of art techniques as in gradient based approach [12] and rhombus context based approach [13].

References

1. A. Khan, A. Siddiqa, S. Munib, S. A. Malik, A recent survey of reversible watermarking techniques, Information Sciences, (279) (2014) 251–272.
2. J. Tian, Reversible data embedding using a difference expansion, IEEE Transactions on Circuits and Systems for Video Technology, 13 (8) (2003) 890–896.
3. S. Weng, Y. Zhao, J.-S. Pan, R. Ni, A novel reversible watermarking based on an integer transform, in: IEEE International Conference on Image Processing, Vol. 3, 2007, pp. 241–244.
4. A. Alattar, Reversible watermark using the difference expansion of a generalized integer transform, IEEE Transactions on Image Processing, 13 (8) (2004) 1147–1156.
5. M. Tejawat, R. Pal, Detecting tampered cheque images using difference expansion based watermarking with intelligent pairing of ixels, in: International Conference on Advanced Computing, Networking and Informatics, 2015, pp. 631–641.
6. D. Thodi, J. Rodriguez, Expansion embedding techniques for reversible watermarking, IEEE Transactions on Image Processing, 16 (3) (2007) 721–730.
7. X. Li, B. Yang, T. Zeng, Efficient reversible watermarking based on adaptive prediction-error expansion and pixel selection, IEEE Transactions on Image Processing, 20 (12) (2011) 3524–3533.
8. F. L. Chin, L. C. Hsing, Adjustable prediction-based reversible data hiding, Digital Signal Processing 22 (6) (2012) 941–953.
9. H. Wu, J. Huang, Reversible image watermarking on prediction errors by efficient histogram modification, Signal Processing, 92 (12) (2012) 3000–3009.
10. R. Naskar, R. Chakraborty, Reversible watermarking utilising weighted median-based prediction, IET Image Processing, 6 (5) (2012) 507–520.
11. S. P. Jaiswal, O. C. Au, V. Jakhetiya, Y. Guo, A. K. Tiwari, K. Yue, Efficient adaptive prediction based reversible image watermarking, in: Proceedings of the 20th International Conference on Image Processing, 2013, pp. 4540–4544.
12. I. C. Dragoi, D. Coltuc, I. Caciula, Gradient based prediction for reversible watermarking by difference expansion, in: Proceedings of the 2nd ACM Workshop on Information Hiding and Multimedia Security, 2014, pp. 35–40.
13. C. Dragoi, D. Coltuc, Improved rhombus interpolation for reversible watermarking by difference expansion, in: Proceedings of the 20th European Signal Processing Conference 2012, pp. 1688–1692.
14. I. C. Dragoi, D. Coltuc, On local prediction based reversible watermarking, IEEE Transactions on Image Processing, 24 (4) (2015) 1244–1246.
15. D. Coltuc, Improved embedding for prediction based reversible watermarking, IEEE Transactions on Information Forensics and Security, 6 (3) (2011) 873–882.

Cancelable Biometrics Using Hadamard Transform and Friendly Random Projections

Harkeerat Kaur and Pritee Khanna

Abstract Biometrics based authentication increases robustness and security of a system, but at the same time biometric data of a user is subjected to various security and privacy issues. Biometric data is permanently associated to a user and cannot be revoked or changed unlike conventional PINs/passwords in case of thefts. Cancelable biometrics is a recent approach which aims to provide high security and privacy to biometric templates as well as imparting them with the ability to be canceled like passwords. The work proposes a novel cancelable biometric template protection algorithm based on Hadamard transform and friendly random projections using Achlioptas matrices followed by a one way modulus hashing. The approach is tested on face and palmprint biometric modalities. A thorough analysis is performed to study performance, non-invertibility, and distinctiveness of the proposed approach which reveals that the generated templates are non-invertible, easy to revoke, and also deliver good performance.

Keywords Cancelable biometrics · Hadamard transform · Random projections · Non-invertible

1 Introduction

Biometrics based authentication is a significant component of current and emerging identification technologies. Typical examples are physical and online access control systems in government organizations, banks, and other commercial uses. There are various security and privacy issues stemming from the widespread usage of biometric systems that needs to be addressed. Ratha et al. [1] identified eight points at

H. Kaur (✉) · P. Khanna
PDPM Indian Institute of Information Technology,
Design and Manufacturing, Jabalpur, Madhya Pradesh, India
e-mail: harkeerat.kaur@iiitdmj.ac.in

P. Khanna
e-mail: pkhanna@iiitdmj.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_5

which a generic biometric system can be attacked. However, amongst many identified issues, stolen biometric scenario where an imposter is able to spoof by providing a stolen biometric sample of the genuine user, is the current threat to deal with. Database attacks leads to permanent template compromise, where an attacker uses the stored biometric data to obtain illegitimate access. As biometric data is being increasingly shared among various applications, cross matching of different databases may be performed to track an individual. Unlike passwords or PINs, biometric templates cannot be revoked on theft. Biometric template are permanently associated with a particular individual and once compromised, it will be lost permanently. Moreover, the same template is stored across different application databases which can be compromised by cross-matching attack. Template data once compromised for one application renders it compromised and unsafe for all other applications for entire lifetime of the user. The concept of cancelable biometrics addresses these concerns. Instead of original biometrics, it uses its transformed versions for storing and matching purposes. In case of any attack, the compromised template can be revoked and new transformed versions can be easily generated.

The objective of this work is to generate biometric templates which can canceled like passwords while at the same time provide non-repudiation and perform like generic biometric templates. Cancelability is achieved by first subjecting the image to Hadamard transformation (HT) and then projecting it on a random matrix whose columns are independent vectors having values –1, +1, or 0 with probabilities 1/6, 1/6, and 2/3, respectively. The sample is then subjected to inverse HT followed by a one-way modulus hashing on the basis of a vector computed in Hadamard domain. The organization of the paper is as follows. A formal definition of cancelable biometrics and related works is provided in Sect. 2. It is followed by the proposed template transformation approach explained in Sect. 3. The experimental results are covered in Sect. 4, and finally the work is concluded in Sect. 5.

2 Cancelable Biometrics

Cancelable biometrics is an important template protection scheme which is based on intentional and systematic repeated distortions of biometric data to protect user specific sensitive information. Unlike other protection schemes like cryptosystems and steganography, the original biometric is never revealed and system operates on transformed data. The biometric data is transformed using some user-specific parameters and transformed template is registered. At authentication, the query template is distorted using similar constraints, thereafter matched with the reference template. Same biometric can be enrolled differently by changing the transformation function and/or parameters for its use in different applications. This prevents cross matching attacks and leads to increase in overall security, privacy, and non-linkability of biometric data. *Biometric salting* and *non-invertible transformation* are two main template transformation approaches.

Teoh et al. (2004) proposed BioHashing which salts biometric features by projecting them on user-specific random matrices (Random Projection) followed by thresholding to generate binary codes. BioHashing becomes invertible if the binary codes and user-specific random matrices are compromised and pre-image attack can be simulated to recover the original data [2]. Sutcu et al. (2005) proposed a nonlinear transformation based salting technique known as robust hashing [3]. The technique is non-invertible but the hashed templates tend to compromise on discriminability. Teoh et al. (2006) proposed BioPhasoring which iteratively mixes biometric features with user-specific random data in a non-invertible fashion without losing discriminability [4]. To address the invertibility of Biohashing, Teoh and Yaung (2007) proposed salting techniques which involve Multispace Random Projections (MRP) [5]. Further, Lumini et al. (2007) combined Multispace Random Projections, variable thresholding, and score level fusions to enhance performance [6].

Non-invertible transformations are many-to-one functions that easily transform biometric data into a new mapping space. Ratha et al. (2007) generated non-invertible cancelable fingerprint templates by distorting minutiae features using Cartesian, polar, and surface folding transformation functions [7].Tulyakov et al. (2005) distorted minutiae features using polynomial based one way symmetric hash functions [8]. Ang et al. (2005) generated cancelable minutiae features using key dependent geometric transformation technique [9]. Bout et al. (2007) generated revocable biometric based identity tokens from face and fingerprint templates by using one way cryptographic functions. The technique separates data into two parts, such that the integer part is used for encryption and the fractional part is used for robust distance computations [10]. Farooq et al. (2007) and Lee et al. (2009) extracted rotation and translation invariant minutiae triplets to generate cancelable bit string features [11].

Each of the above mentioned approaches have their own advantages and disadvantages. BioHashing and other salting techniques are effective but are subjective to invertibility. Also their performance degrades considerably in stolen token scenario. Non-invertible transforms tends to compromise discriminability of transformed biometric in order to achieve irreversibility which degrades the performance. It is imperative to maintain a balance between non-invertiblity, discriminability, and performance for a cancelable biometric technique. This work is motivated towards designing a transformation approach such that the templates are easy to revoke, difficult to invert, and maintains performance in stolen token scenario.

3 Template Transformation

Along with the basics of Hadamard transform and Random Projection, proposedtemplate transformation technique is discussed here.

3.1 Hadamard Transform

Hadamard transform (HT) is non-sinusoidal and orthogonal transformation which offers significant computational advantage over Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT). It decomposes an arbitrary input signal into a set of Walsh functions. Walsh functions are orthogonal, rectangular and can be generated using Kroneckers product of the Hadamard matrices. Hadamard matrix of order n is the $N \times N$ matrix, where $N = 2^n$, generated by the iteration rule given as

$$H_n = H_1 \otimes H_{n-1} \tag{1}$$

$$H_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2}$$

Since the elements of the Hadamard matrix (H_n) are real containing only +1 and –1, they are easy to store and perform computations. H_n is an orthogonal matrix. HT has good energy packing properties, but it cannot be considered a frequency transform due to its non-sinusoidal nature. The sign change along each row of H_n is called sequence which exhibits characteristics like frequency. HT is fast as its computation requires only simple addition and subtractions operations. It can be performed in $O(Nlog_2N)$ operations. For a 2-D vector I of dimensions $N \times N$ where $N = 2^n$, the forward and inverse transformations are performed using Eqs. 3 and 4, respectively.

$$F = (H_n \times I \times H_n)/N \tag{3}$$

$$I = (H_n \times F \times H_n)/N \tag{4}$$

3.2 Random Projection

Random projection is a widely used dimensional reduction technique based on Johnson and Lindenstrauss lemma (JL lemma). JL lemma states that a set of d points in a high dimensional Euclidean space can be mapped down onto a k-dimensional subspace ($k \geq O(logd/\epsilon^2)$, where $0 < \epsilon < 1$), such that the distances between any two points before and after projection is approximately preserved [12]. The effect to which pair-wise distances between points before and after projection are preserved depends upon the projection vectors. The essential property of the projection matrix R used in JL lemma is that its column vectors $r_i \in R$ are required to be orthogonal to each other. Gram Schmidt orthogonalization process is a technique that is usually applied to transform the columns of a random vector into orthogonal ones. Achieving orthogonality is computationally expensive.

To reduce the computation costs of dimensionality reduction algorithms various variants and improvements are proposed by researchers [13]. In a research on

approximating nearest-neighbor in a high dimensional Euclidean space, Indyk and Motwani claimed that column vectors of projection matrix need not to be orthogonal to each other while using random projections [14]. They proved that projection on a random matrix whose column entries are independent random variables with the standard normal distribution having zero mean and unit variance is a distance preserving mapping with less computation cost. Dasgupta proposed a similar construction of random projection matrix in which each row is also rescaled to a unit vector and proved its distance preserving ability using elementary probabilistic techniques [15]. Achlioptas replaced the Gaussian distribution with a computationally inexpensive and upto three times faster projection matrix, A, whose columns are independent vectors defined as [16]

$$A(i,j) = \sqrt{3}\begin{cases} +1, \text{with probability } 1/6; \\ 0, \text{ with probability } 2/3; \\ -1, \text{with probability } 1/6. \end{cases} \tag{5}$$

This allows computation of projected data using simple addition and subtraction operations and is well suited for database environments. Detailed proofs and deeper insights about the distance preservation property of projections using Achiloptas matrix can be found in [13, 16, 17].

3.3 Proposed Transformation Algorithm

A raw biometric grayscale image I is acquired and preprocessed by applying histogram equalization for illumination enhancement followed by extracting region of interest. For the sake of applying HT the dimensions of preprocessed image are kept of the order of power of 2, here $N \times N$ pixels, $N = 128$. Image I^H is obtained by applying forward HT to the preprocessed image using Eq. 3, where size of H_n is $N \times N, N = 128$. A user specific random matrix R of dimensions $d \times k$ is generated using Eq. 5 where $d = k = 128$. Randomness is introduced by projecting the forward HT image I^H on the matrix R as

$$I^{RP} = I^H \times R/\sqrt{k} \tag{6}$$

The column wise mean of the projected image matrix I^{RP} is calculated and stored in a vector M, $M \in R^k$. The elements of vector M are transformed as

$$M(j) = \max\{256, \text{abs}(\lfloor M(j) \rfloor) + 1\} \tag{7}$$

where *abs* is absolute value function. Exploiting the energy compaction property of HT, the coefficients confining to the upper left triangle which gives the basic details of the image are retained and rest are discarded by equating them to zero. On the

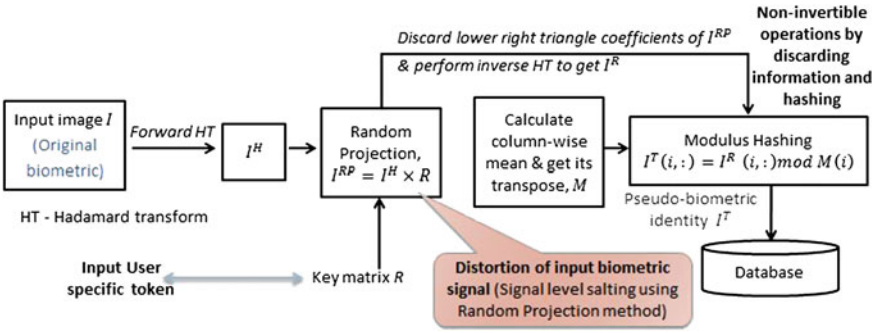


Fig. 1 Block diagram of the proposed approach

resultant, inverse HT is performed using Eq. 4 to obtain I^R. Modulus for each ith row of I^R is separately calculated using vector M.

$$I^T(i,:) = I^{RP}(i,:) \bmod M(i) \tag{8}$$

where i varies from 1 to N and the total number of rows and columns being k and N respectively. After computing the transformed template I^T, the vector M is discarded. Overall I^T can be written as

$$I^T = (H_n \times \left((H_n \times I \times H_n) \times R\right) \times H_n) \bmod M \tag{9}$$

Approximate fractional values of the elements of I^T towards positive infinity. Since the maximum value of modulus is 256, the resultant transformed template after approximation possess integral values between 0 to 255. Figure 1 depicts the block diagram of the proposed approach. Discriminative features are extracted from the transformed template I^T using Linear Discriminant Analysis (LDA). Matching is performed by calculating Euclidean distances between the extracted feature vectors of reference and query biometric templates [18, 19].

4 Experimental Results and Discussion

4.1 Databases Used for Experimentation

The performance is evaluated on two different biometric modalities, i.e., face and palmprint. To study the functional performance of the proposed system on face modality, three different standard face databases– ORL, Extended Yale Face Data-

base B, and Indian face are used. ORL is an expression variant database consisting of 40 subjects with 10 images per subject capturing different facial expressions [20]. Extended YALE face database is an illumination variant database containing 64 near frontal images for 38 subjects under various illumination conditions [21]. Out of it only 10 images per subject having uniform illumination are selected. The Indian face database is a collection of 61 subjects, 39 males and 22 females with 11 images per subjects collected by IIT Kanpur for different orientation of face, eyes, and emotions on face [22]. For each database, 3 images are randomly selected for training database and 7 images for test database. CASIA and PolyU palmprint databases are used to study the functional performance of the proposed system on palmprint image templates. CASIA contains 5,239 palmprint images of left and right palms of 301 subjects thus a total 602 different palms [23]. PolyU database includes 600 images from 100 individuals, with 6 palmprint images from each subject [24]. For palmprint databases, per subject 2 images for training and 4 images for testing purposes are randomly selected after extracting the region of interest [25].

4.2 Performance Evaluation on Face and Palmprint Image Templates

The performance is determined using Equal Error Rates (EER) and Decidability Index (*DI*). Decidability Index (*DI*) is defined as the normalized distance between means of Genuine (μ_G) and Imposter distributions (μ_I). *DI* index measures the confidence in classifying patterns for a given classifier. The value is either positive or negative, according to the score assigned to that pattern. Higher values of *DI* indicate better decidability while classifying genuine and imposter populations. *DI* is calculated as

$$DI = \frac{|\mu_G - \mu_I|}{\sqrt{(\sigma_G^2 + \sigma_I^2)/2}} \tag{10}$$

Matching performance of the proposed approach is evaluated in case of stolen token scenario [18, 19]. It represents the worst case scenario when an attacker is always in possession of users' specific data or tokens. The stolen token scenario is simulated by assigning same random matrix R to each subject in the database. Matching is performed on transformed templates which are generated using same R for each subject in every database. It is expected that the system performance must not regress when it operates on transformed templates under stolen token scenario. Matching performance is also evaluated on the original (untransformed) templates using LDA which is used for comparison. Table 1 provides results for matching performance on conventional untransformed biometric templates and transformed templates using proposed approach under stolen token scenario for various databases. The ROC curves are shown in Fig. 2.

Table 1 Matching performance for face and palmprint templates

Modality	Database	Original		Proposed	
		EER (%)	DI	EER (%)	DI
Face	ORL	5.42	3.133	6.90	3.438
	Indian Face	11.11	2.398	11.53	2.352
	YALE	10.52	2.612	8.91	2.921
Palmprint	PolyU	0.62	7.569	0.56	8.735
	CASIA	2.34	5.083	2.50	4.183

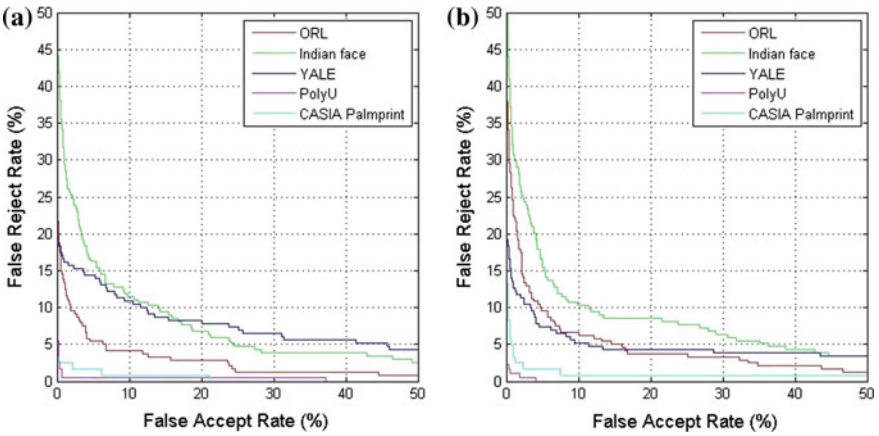


Fig. 2 ROC curves for matching performance **a** original domain **b** transformed domain

It can be observed that the matching performance, i.e., EER of proposed approach under stolen token scenario is comparable to non-cancelable based technique. The experimental results validate that the proposed approach transforms biometric templates while effectively preserving their discriminability and meets the performance evaluation criteria of cancelability under stolen token scenario. The genuine and imposter populations in transformed domain is well distributed. DI values obtained from genuine and imposter mean and variance in stolen token scenario are sufficiently high which indicate good separability among transformed templates. The performance in case of legitimate key scenario, when each subject is assigned different random matrix R results in nearly 0 % EER for all modalities and databases.

4.3 Invertibility Analysis

Consider the scenario, when the transformed template I^T and projection matrix R are available simultaneously. The inverse operation (decryption) requires the projection

of I^T over the inverse of random matrix R^{-1} as

$$I^{inv\_proj} = H_n \times ((H_n \times I^T \times H_n) \times R^{-1}) \times H_n \tag{11}$$

The next step requires an attacker to have the exact values over which modulus is computed for each row, i.e., the mean vector M, which is discarded immediately after transformation. Hence, it cannot be inverted. Yet, we consider a scenario where the exact vector M is approximated by the attacker using intrusion or hill climbing attacks. Then the inverse template should be computed as

$$I^{rec}(i,:) = I^{inv\_proj}(i,:) mod M(i) \tag{12}$$

To decrypt the information, inverse or psuedo-inverse of matrix R needs to be computed. However, for lossless recovery from encrypted data, the matrix R should be selected such that the elements of inverse matrix R^{-1} posses non-negative integral values. In our case the key space is restricted to random Achlioptas matrices comprising of +1, 0, or –1. It is possible to compute inverse or psuedo-inverse of these matrices but the inverted matrices are always found to possess non-integral and negative values. It makes the recovery of information very noisy on decryption and does not reveal original template.

4.4 Distinctiveness Analysis

To evaluate distinctiveness, ten different transformed templates corresponding to the same biometric are generated for each database by changing user-specific parameter (random projection matrix R). Mutual information content between each pair of transformed templates, C_r, is calculated using Eq. 13.

$$C_r(I_1, I_2) = \frac{\sum\sum(I_1 - \bar{I}_1)(I_2 - \bar{I}_2)}{\sqrt{(I_1 - \bar{I}_1)^2 + (I_2 - \bar{I}_2)^2}} \tag{13}$$

where $\bar{I}_1$,$\bar{I}_2$ represents the mean of templates I_1, I_2, respectively. The correlation index (CI) is the mean of all collected C_r values. Table 2 provides CI values between transformed templates for different modalities and databases. For example, average value of $I = 0.121$ means that two templates generated from the same biometric sample using different random matrices share 12.1 % of mutual information and are different to each other by 87.9 %. It is observed from Table 2 that CI values are low. This indicates that the proposed approach offers good revocability and diversity.

Table 2 Correlation index values for different databases

Modality	Face			Palmprint	
Database	ORL	Indian face	YALE	PolyU	CASIA
CI	0.121	0.132	0.112	0.095	0.134

5 Conclusion

The proposed approach successfully meets an important requirement of achieving good recognition rates in transformed domain and addresses stolen token scenario. Non-invertiblity being an important requirement is also ascertained without giving up on performance. The compaction of energy using HT before random projection allows mean vector M to coincide for templates belonging to same user. This way templates belonging to the same user are tend to have similar M. The ability to generate various transformed templates by changing the transformation parameter is evaluated in distinctiveness analysis which supports revocability and diversity.

References

1. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. IBM systems Journal **40** (2001) 614–634
2. Lacharme, P., Cherrier, E., Rosenberger, C.: Preimage attack on biohashing. In: International Conference on Security and Cryptography (SECRYPT). (2013)
3. Sutcu, Y., Sencar, H.T., Memon, N.: A secure biometric authentication scheme based on robust hashing. In: Proceedings of the 7th workshop on Multimedia and security, ACM (2005) 111–116
4. Teoh, A.B.J., Ngo, D.C.L.: Biophasor: Token supplemented cancellable biometrics. In: Control, Automation, Robotics and Vision, 2006. ICARCV'06. 9th International Conference on, IEEE (2006) 1–5
5. Teoh, A., Yuang, C.T.: Cancelable biometrics realization with multispace random projections. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **37** (2007) 1096–1106
6. Lumini, A., Nanni, L.: An improved biohashing for human authentication. Pattern recognition **40** (2007) 1057–1065
7. Ratha, N., Connell, J., Bolle, R.M., Chikkerur, S.: Cancelable biometrics: A case study in fingerprints. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Volume 4., IEEE (2006) 370–373
8. Tulyakov, S., Farooq, F., Govindaraju, V.: Symmetric hash functions for fingerprint minutiae. In: Pattern Recognition and Image Analysis. Springer (2005) 30–38
9. Ang, R., Safavi-Naini, R., McAven, L.: Cancelable key-based fingerprint templates. In: Information Security and Privacy, Springer (2005) 242–252
10. Boult, T.E., Scheirer, W.J., Woodworth, R.: Revocable fingerprint biotokens: Accuracy and security analysis. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–8
11. Farooq, F., Bolle, R.M., Jea, T.Y., Ratha, N.: Anonymous and revocable fingerprint recognition. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE (2007) 1–7

12. Dasgupta, S., Gupta, A.: An elementary proof of the johnson-lindenstrauss lemma. International Computer Science Institute, Technical Report (1999) 99–006
13. Matoušek, J.: On variants of the johnson–lindenstrauss lemma. Random Structures & Algorithms **33** (2008) 142–156
14. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, ACM (1998) 604–613
15. Dasgupta, S.: Learning mixtures of gaussians. In: Foundations of Computer Science, 1999. 40th Annual Symposium on, IEEE (1999) 634–644
16. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ACM (2001) 274–281
17. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2001) 245–250
18. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. Neural Networks, IEEE Transactions on **13** (2002) 1450–1464
19. Connie, T., Teoh, A., Goh, M., Ngo, D.: Palmprint recognition with pca and ica. In: Proc. Image and Vision Computing, New Zealand. (2003)
20. ORL face database: (AT&T Laboratories Cambridge) http://www.cl.cam.ac.uk/.
21. Yale face database: (Center for computational Vision and Control at Yale University) http://cvc.yale.edu/projects/yalefaces/yalefa/.
22. The Indian face database: (IIT Kanpur) http://vis-www.cs.umas.edu/.
23. CASIA palmprint database: (Biometrics Ideal Test) http://biometrics.idealtest.org/downloadDB/.
24. PolyU palmprint database: (The Hong Kong Polytechnic University) http://www4.comp.polyu.edu.hk/biometrics/.
25. Kekre, H., Sarode, T., Vig, R.: An effectual method for extraction of roi of palmprints. In: Communication, Information & Computing Technology (ICCICT), 2012 International Conference on, IEEE (2012) 1–5

A Semi-automated Method for Object Segmentation in Infant's Egocentric Videos to Study Object Perception

Qazaleh Mirsharif, Sidharth Sadani, Shishir Shah, Hanako Yoshida and Joseph Burling

Abstract Object segmentation in infant's egocentric videos is a fundamental step in studying how children perceive objects in early stages of development. From the computer vision perspective, object segmentation in such videos poses quite a few challenges because the child's view is unfocused, often with large head movements, effecting in sudden changes in the child's point of view which leads to frequent change in object properties such as size, shape and illumination. In this paper, we develop a semi-automated, domain specific method, to address these concerns and facilitate the object annotation process for cognitive scientists, allowing them to select and monitor the object under segmentation. The method starts with an annotation of the desired object by user and employs graph cut segmentation and optical flow computation to predict the object mask for subsequent video frames automatically. To maintain accurate segmentation of objects, we use domain specific heuristic rules to re-initialize the program with new user input whenever object properties change dramatically. The evaluations demonstrate the high speed and accuracy of the presented method for object segmentation in voluminous egocentric videos. We apply the proposed method to investigate potential patterns in object distribution in child's view at progressive ages.

Q. Mirsharif (✉) · S. Shah
Department of Computer Science, University of Houston, Houston 77204, USA
e-mail: Qazaleh.mirsharif@gmail.com

S. Shah
e-mail: Sshah@central.uh.edu

S. Sadani
Department of Electronics & Communication,
Indian Institute of Technology, Roorkee, India
e-mail: Sidharthsadani@gmail.com

H. Yoshida · J. Burling
Department of Psychology, University of Houston,
126 Heyne Building, Houston, TX 77204-5022, USA
e-mail: Yoshida@uh.edu

J. Burling
e-mail: Jmburling@uh.edu

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_6

Keywords Child's egocentric video · Cognitive development · domain specific heuristic rules · Head camera · Object perception · Object segmentation · Optical flow

1 Introduction

Infants begin to learn about objects, actions, people and language through many forms of social interactions. Recent cognitive research highlights the importance of studying the infant's visual experiences in understanding early cognitive development and object name learning [1–3]. The infant's visual field is dynamic and characterized by large eye movements and head turns owing to motor development and bodily instabilities which frequently change the properties of their visual input and experiences. What infants attend to and how their visual focus on objects is structured and stabilized during early stages of development has been studied to understand the underlying mechanism of object name learning and language development in early growth stages [1, 4–6].

Technological advancement allows researchers to have access to these visual experiences that are critical to understanding the infant's learning process first hand [1, 7]. Head cameras attached to the child's forehead enables researchers to observe the world through child's viewpoint by recording their visual input [2, 8]. However, it becomes very time consuming and impractical for humans to annotate objects in these high volume egocentric videos manually.

As discussed in [9], egocentric video is an emerging source of data and information, the processing of which poses many challenges from a computer vision perspective. Recently, computer vision algorithms have been proposed to solve the object segmentation problem in such videos [10, 11]. The nuances of segmentation in egocentric videos arise because the child's view is unfocused and dynamic. Specifically, the egocentric camera, (here, the head camera) is in constant motion, rendering the relative motion between object and background more spurious than that from a fixed camera. In addition, the random focus of a child causes the objects to constantly move in and out of the view and appear in different sizes, and often the child's hand may occlude the object. Examples of such views are shown in Fig. 1a, b and c. Finally, if the child looks towards a light source, the illumination of the entire scene appears very different, as shown in Fig. 1d.

In this paper, we develop an interactive and easy to use tool for segmentation of objects in child's egocentric video that addresses the above problems. The method enables cognitive scientists to select the desired object and monitor the segmentation process. The proposed approach exploits graph cut segmentation to model object and background and calculate optical flow between frames to predict object mask in following frames. We also incorporate domain specific heuristic rules to maintain high accuracy when object properties change dramatically.

The method is applied to find binary masks of objects in videos collected by placing a small head camera on the child as the child engages in toy play with a

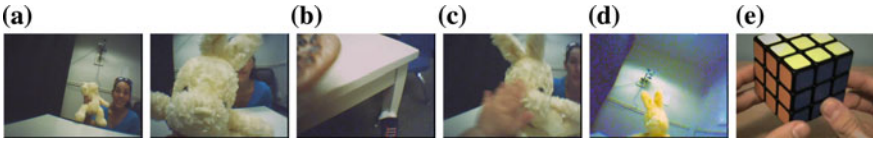


Fig. 1 **a** Size variation. **b** Entering and leaving. **c** Occlusion. **d** Illumination variation. **e** Orientation variations

parent. The object masks are then used to study object distribution in child's view at progressive ages by generating heat maps of objects for multiple children. We investigate the potential developmental changes in children's visual focus on objects.

The rest of the paper is organized as follows: We describe the experimental setup and data collection process in Sect. 2. The semi automated segmentation method is explained in detail in Sect. 3. Results are presented and discussed in Sect. 4. Finally Sect. 5 will conclude the present study and highlights the main achievements and contributions of the paper.

2 The Experimental Setup

A common approach to study infant's cognitive developmental process is to investigate their visual experience in a parent-child toy play experiment. In the current study we extract the infant's perspective by placing a head camera on the child and recording the scene as it engages in tabletop toy play sitting across from the mother. Each play session is around 5 min where the mother plays with toys of different colors and sizes including bunny, carrot, cup, cookie and car. A transparent jar was initially among the toys, but was removed from the study as the proposed method does segment such objects accurately. The mother plays with the toys one by one and names the toys as she attempts to bring the infant's attention to that toy. Multiple toys may sit in the view at the same time. The videos are recorded at progressive ages including 6, 9, 12, 15 and 18 months. In this paper we have used 15 videos which consist of three videos from each age. From each video, we extracted approximately 9500 image frames. The image resolution is 480 to 640 pixels.

3 The Proposed Approach

In this section we explain our proposed method in detail in three main steps namely, initialization and modeling of the object and background, object mask prediction for next frame, and performing a confidence test to continue or restart the program. The flow diagram for the method is shown in Fig. 2. We use a graph based segmentation approach [12, 13] to take user input, for initialization and also when recommended

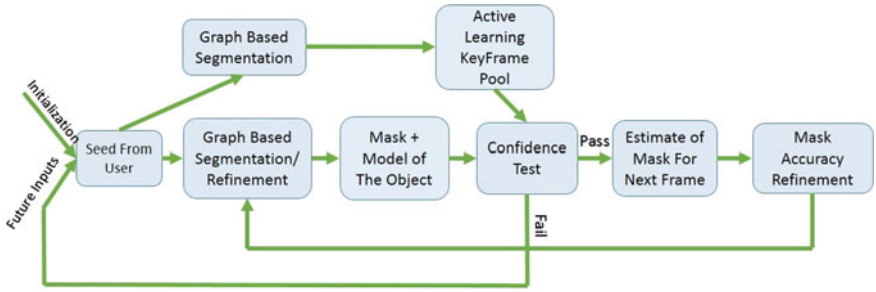


Fig. 2 Flow diagram of the proposed approach

by the confidence mechanism (Sect. 3.3). For each user input we model the object and save the features as a KeyFrame in an active learning pool, which is used as ground truth data. We then use optical flow [14] to estimate the segmentation mask for the next frame and subsequently refine it to obtain the final prediction mask (Sect. 3.2). The obtained segmentation result is then evaluated under our confidence test (Sect. 3.3). If the predicted mask is accepted by the test, the system continues this process of automatic segmentation. When the confidence test fails, the system quickly goes back and takes a new user input to maintain the desired accuracy in segmentation.

3.1 Initialization, User Input and Object Model

We initialize the segmentation algorithm with a user seed manually encompassing the object in a simple polygon. We then use an iterative Graph-Cut [12] based segmentation approach to segment the desired object. Given that we work with opaque objects this method suits well to our requirement of accuracy. We calculate the minimum cut of the graph whose nodes are each of the pixels. Each of the nodes (pixels) are given a foreground f_i and background prior b_i. For the initial frame this prior is calculated from the user seed, and for each subsequent frame the prior is calculated from the model of the foreground and background of the previous frame. In our approach we use a Multivariate Gaussian Mixture Model (GMM) of the RGB components in the image, as our model for the object as well as the background.

Along with the prior term we use a normalized smoothness cost term S_{ij}, which is a penalty term if two adjacent pixels have different assignments. As mentioned in [12], a normalized gradient magnitude obtained from the sobel operator is used for the same. We run the Graph-Cut algorithm [12] iteratively until the number of changes in the pixel assignments between two consecutive iterations falls below a certain acceptable threshold. The Multivariate GMM is updated in each iteration using kmeans clustering. The mask along with the Multivariate GMM define the combined model of the object and are the starting point of the prediction of the mask for the next frame.

3.2 Segmentation Prediction for Next Frame

We begin our prediction with the calculation of dense optical flow between the previous frame and the current frame. Since the two frames are consecutive frames of an egocentric video, we can assume there isn't a drastic change in the characteristics of the foreground or the background. Using optical flow we predict pixel to pixel translation of the mask. This initial calculation provides a starting estimate of the mask.

$$\begin{aligned}(x,y)_{CurrentMask} &= (x,y)_{PreviousMask} + v(x,y)\\ (x,y) &\in Pixel\ Coordinates\ of\ Foreground\ ,\\ v(x,y) &= Optical\ Flow\ of\ Pixel(x,y)\end{aligned} \tag{1}$$

Some refinement in this mask is required to maintain the accuracy for the following reasons. Firstly, the pixel to pixel transformation using optical flow is not a one to one but a many to one transformation i.e. many pixels in the previous frame may get translated to the same pixel in the current frame. Secondly, if part of the object is just entering the frame from one of the boundaries, optical flow by itself cannot predict if the boundary pixels belongs to the object or not. Lastly, under occlusion, as is the case in many frames when the mother is holding the object or the child's hands are interacting with the object, flow estimates the mask fairly well at the onset of occlusion but fails to recover the object once the occlusion has subsided thus leading to severe under segmentation (Fig. 3)

To refine this initial estimate of the mask, we define a region of uncertainty around this initial estimate of the mask, both inwards and outwards from the mask. If the object happens to be near one of the edges of the frame, we define areas along the right, left, top, or bottom edges as part of the uncertain region based on the average flow of the object as well as the local spatial distribution of the object near the edges. We then input this unlabeled, uncertain region into the earlier Graph-Cut stage to label these uncertain pixels as either foreground or background. This helps in obtaining a more accurate, refined segmentation mask (Fig. 4).

This segmentation result is now compared against the learnt ground truth from the user, stored in the active learning KeyFrame Pool based on a confidence test

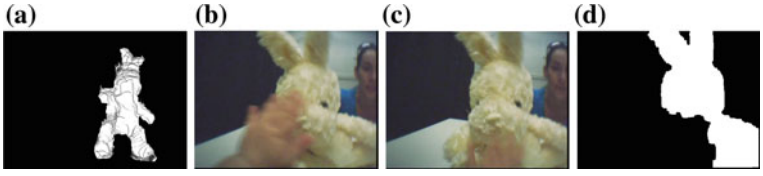


Fig. 3 Need for refinement. **a** Error due to optical flow. **b–c** Occlusion of object by the child's hand in successive frames. **d** Predicted mask

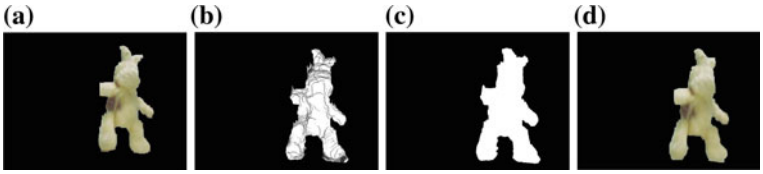


Fig. 4 Steps in segmentation prediction. **a** User seed. **b** First estimate of next mask using optical flow. **c** The uncertain region. **d** Final mask using graph cut (Bunny moved downwards, legs expanded)

(explained in the next section). If the segmentation result is accepted by the confidence test, we go on to predict the mask for the next frame using this mask. If the result fails the confidence test, we go back and take a new input from the user.

3.3 Confidence Test, Learning Pool, Keyframe Structure

We define Keyframes as those frames in which the segmentation mask has been obtained from user input. This represents the ground truth data. For each keyframe we save the following parameters:

- Size of The Segmentation Mask in Pixels
- Multivariate GMM parameters for foreground and background (Centers, Covariances, Weights)
- The amount of acceptable error in the centers of the Modes of the GMM

These three parameters are utilized to test the validity of the predicted segmentation mask. Firstly, we check if the size of the mask has increased or decreased beyond a certain fractional threshold of the size of the mask in the most recent keyframe. This test is introduced as a safeguard to maintain the reliability of segmentation because when the child interacts with the object, brings it closer or moves it away, the object may suddenly go from occupying the entire frame to just a small no. of pixels. We've implemented to flag 39 % variation, but anywhere between 25–50 % can be chosen depending on the sensitivity of the results required.

$$Size_{CurrentFrame} \leq 0.61 * Size_{KeyFrame} \quad (or)$$
$$Size_{CurrentFrame} \geq 1.39 * Size_{KeyFrame} \; ; Confidence = False \tag{2}$$

Secondly, we look to flag under segmentation, mostly during recovery of the entire segment after occlusion. To detect under segmentation we assume that the background cluster of the current frame moves closer or is similar to the foreground cluster of the most recent keyframes to account for the presence of the incorrectly labelled foreground pixels as background. We hypothesize that the background clus-

ter that moves closer to the foreground does so very slightly as it still must account for the background pixels but the length of the projection of the eigen vectors along the line joining the centers of this background—foreground cluster pair increases.

$$\begin{aligned} D_{BFiKey} &= \min_j ||\boldsymbol{BC_{i,Key}} - \boldsymbol{FC_{j,Key}}||^2 \\ D_{BFiCurr} &= \min_j ||\boldsymbol{BC_{i,Curr}} - \boldsymbol{FC_{j,Curr}}||^2 \end{aligned} \tag{3}$$

From the above two distances, the first one being for the keyframe clusters and the second for the current frame, we can find which background cluster moved closer to which foreground cluster. After which we calculate the projections of the eigen vectors of the Background Cluster that moved closest to the Foreground Cluster onto the line joining the centers of these two clusters, in the keyframe and the current frame.

$$P_{Key} = \sum_{k=1}^{3} ||\boldsymbol{E.Vec_{k,Key}} \cdot \boldsymbol{CC_{j,Key}}|| \, , P_{Curr} = \sum_{k=1}^{3} ||\boldsymbol{E.Vec_{k,Curr}} \cdot \boldsymbol{CC_{j,Curr}}|| \tag{4}$$

If this increase is greater than 25 % then we can reliably conclude under segmentation.

Lastly, the object may appear differently in different orientations or in different illumination conditions, for which we compare the GMM for the predicted segment against all the Models in the Keyframe. We do this by checking if the average distance between corresponding centers in the two segments is within the acceptable error for that GMM, and if so, are the difference in weights of these corresponding centers also within a certain acceptable threshold. The latter threshold is set manually depending on the number of modes and the desired sensitivity. The former is obtained using standard deviation of the RGB channels

$$\begin{aligned} \textit{Thresh For Avg Dist For Center } i &= (\sum_{k=1}^{3} Std_{i,k})/3 \\ \textit{where} \quad k &\in (R, G, B \textit{ the } 3 \textit{ Channels}) \end{aligned} \tag{5}$$

If this criteria is not met then we take a new user input and it becomes a keyframe in the learning pool.

4 Results and Discussions

We use the method to extract multiple objects from videos and compare the resulting object masks with their corresponding manual annotation provided by experts. Further, the performance of the algorithm in terms of run time and amount of user

Table 1 Performance measures (*Note* The above values are average over 300 frames)

Object	Total time (s)	Optical flow time (s)	Processing time (s)	Image size (px)	% Area	User input (per 300 frames)	Accuracy (%)
Bunny	8.7002	8.0035	0.6967	199860	65.00	11	97.76
Cup	8.1156	7.6534	0.4622	47261	15.38	9	98.43
Carrot	7.9513	7.5180	0.4333	32080	10.44	9	99.71
Car	7.9237	7.6320	0.2917	10872	3.54	10	99.35
Cookie	7.9421	7.6145	0.3276	27653	9.00	13	98.10

interaction, for each of these objects is reported in Table 1. The run time of the method consists of the time taken for optical flow calculation and the processing time required by the proposed method. We see clearly that over a large number of frames the average total time would easily outperform the time required in manual segmentation. We also observe that the processing time varies directly with the (image size) area occupied by the object. Lastly we observe how frequently the algorithm requires user input. We let the method run for a large number of frames (300 frames, a subset of the entire video) and count the number of user input requests. It is important to note that we have set the method to take user input every 50 frames, so even in case of no errors we would take 6 user inputs. Hence the additional user inputs required, due to uncertainty in prediction, are 5, 3, 3, 4, 7 on average, respectively. In any case this is a significant reduction in the amount of user involvement as only 3 % of the frames require user input on average.

As with any automated approach to segmentation, user interaction and processing time is reduced, what is traded off is the accuracy of the automated segmentation as compared to manual segmentation. To evaluate this we take 30 randomly picked frames and have them manually annotated by 5 different people. We calculate the DICE similarity between automated masks versus the manually annotated masks and then the DICE similarity between the manually segmented masks for each of the frames for each pair of people. The mean and standard deviations of which are noted in Table 2. We see that, on average, we lose only 3.21 % accuracy as compared to manual segmentation. Note: DICE similarity is measured as the raio of twice the no. of overlap pixels to the sum of the no. of pixels in each mask.

In our application, under segmentation is not tolerable but slight over segmentation is. We see that our approach doesn't undersegment any worse than manual segmentation would, as can be seen from the recall values in the two columns. On the other hand, our algorithm consistently, but not excessively (as we see from the DICE measurements), oversegments the object, as can be seen from the precision values in the two columns. Thus we see that the proposed approach significantly reduces time and user interaction with little loss in accuracy as compared to manual segmentation. Note: Precision is the proportion of mask pixels that overlap with manual segmentation and Recall is the proportion of the manual segmentation pixels that are part of the predicted mask.

Table 2 DICE similarity coefficient and precision and recall values

'DICE'	Our algorithm versus manual	Manual versus manual
Mean	0.9248	0.9569
Std	0.043	0.0167
'Precision and recall'	Our algorithm versus manual	Manual versus manual
Precision	0.8746	0.9532
Recall	0.9567	0.9734

We use the results obtained from segmentation to investigate how objects are distributed in the child's view at progressive ages. We look to study potential regularities in object movement patterns and concentration in child's view with age. To visualize the areas where infants focus and fixate on objects we plot the heat maps of object for each video. To see which locations have been occupied by objects most recently, we give each pixel of the object a weight W_i for each object mask and accumulate the results in time. The final output stores the following values for each image pixels:

$$P_{xy,Output} = \sum_{i=1}^{L} W_i * P_{xy,ObjectMask}\ , W_i = i/L \tag{6}$$

where i is the frame number and L is the total number of frames in video (usually around 9500).

From the heat maps, we can see that object movement in 6 months infants is large and does not follow any specific pattern. The object concentration region changes across infants and their visual focus on objects are not stabilized. However after 9 months, the object distribution pattern becomes more structured and the object concentration area moves down toward the middle-bottom part of the visual field. This region seems to be the active region where child interacts with object most of the time. For 18 months old children, object movements increase and the pattern change across the children. However the concentration point is still in the bottom area very close to child's eyes. The results might be aligned with previous psychological hypothesis which discusses an unfocused view in 6 month old infants and increasing participation of child in shaping his visual field with age [15]. 18 month old infants are able to make controlled moves and handle objects. This study is still at early stages and more investigation of other factors such as who is holding the object is required to discover how child's visual focus of attention is shaped and stabilized over the early developmental stages and who is shaping the view at each stage. The results demonstrate a developmental trend in child's visual focus with physical and motor development which might support a controversial psychological hypothesis on existence of a link between physical constraint and language delay in children suffering from autism (Fig. 5).

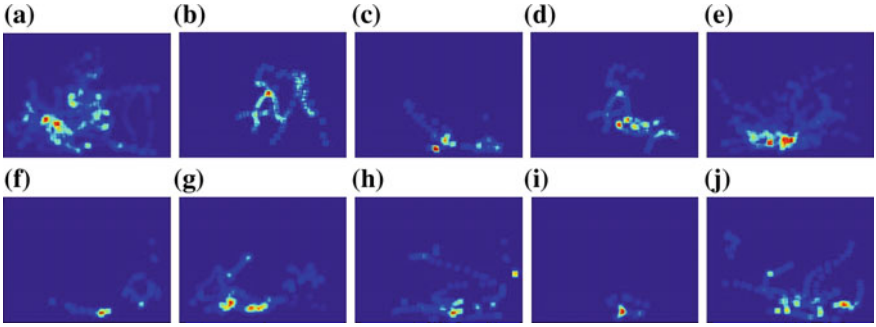


Fig. 5 Heatmaps of objects for infants at progressive ages (**a**, **b**) 6 months infants (**c**, **d**) 9 months infants (**e**, **f**) 12 months infants (**g**, **h**) 15 months infants (**i**, **j**) 18 months infants

5 Conclusions

We proposed a semi-automated method for object segmentation in egocentric videos. The proposed method uses domain specific rules to address challenges specific to egocentric videos as well as object occlusion and allows researchers to select the desired object in the video and control the segmentation process. The method dramatically speeds up the object annotation process in cognitive studies and maintain a high accuracy close to human annotation. We applied the method to find object masks for a large number of frames and studied object movement and concentration patterns in child's visual field at progressive ages. We found a developmental trend in child's visual focus of attention with age and movement of active region toward middle bottom region of child's visual field. The current study is one step toward understanding the mechanism behind early object name learning and cognitive development in humans.

References

1. Pereira, A.F., Smith, L.B., Yu, C.: A bottom-up view of toddler word learning. Psychonomic bulletin & review 21(1), 178–185 (2014)
2. Pereira, A.F., Yu, C., Smith, L.B., Shen, H.: A first-person perspective on a parent-child social interaction during object play. In: Proceedings of the 31st Annual Meeting of the Cognitive Science Society (2009)
3. Smith, L.B., Yu, C., Pereira, A.F.: Not your mothers view: The dynamics of toddler visual experience. Developmental science 14(1), 9–17 (2011)
4. Bambach, S., Crandall, D.J., Yu, C.: Understanding embodied visual attention in child-parent interaction. In: Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on. pp. 1–6. IEEE (2013)
5. Burling, J.M., Yoshida, H., Nagai, Y.: The significance of social input, early motion experiences, and attentional selection. In: Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on. pp. 1–2. IEEE (2013)

6. Xu, T., Chen, Y., Smith, L.: It's the child's body: The role of toddler and parent in selecting toddler's visual experience. In: Development and Learning (ICDL), 2011 IEEE International Conference on. vol. 2, pp. 1–6. IEEE (2011)
7. Yoshida, H., Smith, L.B.: What's in view for toddlers? Using a head camera to study visual experience. Infancy 13(3), 229–248 (2008)
8. Smith, L., Yu, C., Yoshida, H., Fausey, C.M.: Contributions of Head-Mounted Cameras to Studying the Visual Environments of Infants and Young Children. Journal of Cognition and Development (just-accepted) (2014)
9. Bambach, S.: A Survey on Recent Advances of Computer Vision Algorithms for Egocentric Video. arXiv preprint arXiv:1501.02825 (2015)
10. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3137–3144. IEEE (2010)
11. Ren, X., Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis. In: Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on. pp. 1–8. IEEE (2009)
12. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. Pattern Analysis and Machine Intelligence, IEEE Transactions on 26(9), 1124–1137 (2004)
13. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23(11), 1222–1239 (2001)
14. Horn, B.K., Schunck, B.G.: Determining optical flow. In: 1981 Technical Symposium East. pp. 319–331. International Society for Optics and Photonics (1981)
15. Yoshida, H., Burling, J.M.: Dynamic shift in isolating referents: From social to self-generated input. In: Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on. pp. 1–2. IEEE (2013)

A Novel Visual Secret Sharing Scheme Using Affine Cipher and Image Interleaving

Harkeerat Kaur and Aparajita Ojha

Abstract Recently an interesting image sharing method for gray level images using Hill Cipher and RG-method has been introduced by Chen [1]. The method does not involve pixel expansion and image recovery is lossless. However, use of Hill Cipher requires a 2×2 integer matrix whose inverse should also be an integer matrix. Further, to extend the method for multi-secret sharing, one requires higher order integer matrices. This needs heavy computation and the choice of matrices is also very restricted, due to integer entry constraints. In the present paper we introduce an RG-based Visual Secret Sharing Scheme (VSS) scheme using image interleaving and affine cipher. Combined effect of image interleaving and affine transformation helps in improving the security of the secret images. Parameters of the affine cipher serve as keys and the random grid and encrypted image form the shares. No one can reveal the secret unless the keys and both the shares are known. Further, as opposed to the method in [1], the present scheme does not require invertible matrix with integer inverse. The scheme is also extended for multi-secret sharing.

Keywords Image interleaving · Pixel correlation · Affine cipher

1 Introduction

Enormous growth in multimedia communication and its applications in various fields such as banking, military surveillance, medical imaging, biometric etc. have led to significant development in the field of information security and cryptography. Whereas research on plaintext cryptography for secure information communication has been matured and standard protocols exist for such applications, research

H. Kaur (✉) · A. Ojha
PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, Madhya Pradesh, India
e-mail: harkeerat.kaur@iiitdmj.ac.in

A. Ojha
e-mail: aojha@iiitdmj.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_7

on secure visual and multimedia information communication is of relatively recent origin. One of the first visual secret sharing (VSS) scheme was proposed by Naor and Shamir [2] in 1994, in which a secret image was transformed to n meaningless image shares and decryption was performed by stacking k out of n shares ($k \leq n$), so that the revealed image could be recognized by human visual system (HVS) only. This brought a revolution in the field of image information security and many researchers started working on VSS schemes. The method proposed by Naor and Shamir required a comprehensive codebook to encode black and white pixels of secret image into shares. Moreover, each pixel was transformed to more than one (sub) pixels. Pixel expansion and low contrast of the reconstructed image were the major drawback of the scheme. Nonetheless, the innovative idea of VSS attracted the attention of a large number of researchers in the field and numerous VSS schemes have been proposed and implemented in various applications over the last two decades. Most of these methods are focused one or more of the following aspects (i) minimizing pixel expansion (ii) contrast enhancement (iii) multi-secret sharing (iv) reversible data hiding (v) probabilistic visual secret sharing schemes (vi) color image cryptography and (vii) image authentication and repairing. A detailed account of VSS schemes may be found in [2].

In recent years many improved VSS techniques have been proposed which are focused on enhancing the security, visual quality and/or minimizing pixel expansion and other important aspects. Kafri and Keren [3] introduced the concept of random grids (RG) as a low-cost and simple technique for encryption of images [3]. Compared to previous schemes, RG-based visual cryptography scheme does not require any code book and lossless recovery of the secret is possible without any pixel expansion. These important features attracted the attention of researchers working in the field and several techniques have been proposed based on random grids. Shyu [4] has devised a simple VSS scheme for gray scale and color images using random grids by generating two shares of an image. For a long period, people proposed only (2, 2) RG-based VSS schemes, until the first (k, n)—threshold scheme was proposed by Chen and Tsao in the year 2011 [5]. However, the scheme suffered from low contrast problem in the decrypted image. To address this issue, very recently an improved (k, n)—threshold RG-VSS has been proposed by Guo et al. [6]. Their claims are validated by experimental results and contrast quality is significantly improved. Wu and Sun [7] have also presented a VSS scheme that encodes a given secret image into n random grids using the concept of general access structure. While the qualified set of participants can decrypt the secret using HVS, it remains unrevealed to others. Furthermore, a cheating prevention method is also devised to enhance the security.

Wei-Kuei Chen [1] has recently proposed an interesting image sharing method for gray level images using Hill Cipher and RG-method. He first employs Hill Cipher to divide the secret image into two sub-images. Then the concept of random grid is applied to construct the share images. The method does not involve pixel expansion and image recovery is lossless. However, use of Hill Cipher requires a 2×2 integer matrix whose inverse should also be an integer matrix. Further, to extend the method for multisecret sharing, one requires higher order integer matrices. This

needs heavy computation and the choice of matrices is also very restricted, due to integer entry constraints. In addition, it has been observed in [14] that a small guess of diagonal entries of the 2×2 matrix reveals the secret, especially when the matrix is diagonally dominant. This motivated us to study image encryption using other affine transformations. A RG-based VSS scheme is proposed in the present paper for multi-secret sharing using image interleaving and affine cipher. Combined effect of image interleaving and affine transformation helps in improving the security of the secret images. Given secret image is divided into four sub-images and all the sub-images are packed within each other using interleaving. To enhance the security, an affine cipher is applied on the resulting image followed by XOR operation with a given random grid of the same size. Decryption is performed by applying the same operations in the reverse order. Parameters of the affine cipher serve as keys and the random grid and encrypted image form the shares. No one can reveal the secret unless the keys and both the shares are known. Further, as opposed to the method by Chen [1], the present scheme does not require invertible matrix with integer inverse. The scheme is also extended for multi-secret sharing. Rest of the paper is organized as follows. Section 2 introduces the preliminaries. Section 3 is devoted to the proposed encryption and decryption process. In Sect. 4, experimental results are presented. The proposed technique is also compared with the Hill Cipher based encryption method proposed in [1].

2 Preliminary

Image interleaving is a well-known concept in computer graphics and image processing. The scheme is briefly discussed in the following section.

2.1 Image Interleaving

Interleaving is an approach of permutation in which pixels from alternate rows/columns of an image are inserted between rows/columns of another or the same image. The output image is a regular interwoven pattern of all the input pixels. Thus, the adjacent pixel values become no longer remain correlated in contrast to the original image, in which adjacent pixels are mostly highly correlated. In the present paper, an interleaving scheme is proposed that takes as input, a $2m \times 2n$ image $S = (S_{i,j})_{i=1,j=1}^{2m,2n}$. It first breaks the image into four sub-images or quadrants of equal size. The sub-images $S_1 = (S_{i,j})_{i=1,j=1}^{m,n}$, $S_2 = (S_{i,j})_{i=1,j=n+1}^{m,2n}$, $S_3 = (S_{i,j})_{i=m+1,j=1}^{2m,n}$, $S_1 = (S_{i,j})_{i=m+1,j=n+1}^{2m,2n}$ are then interleaved as follows.

S_1 and S_2 are interleaved by inserting the jth column of S_1 after the $n + j$th column of S_2 for $j = 1, ..., n$. Let the resulting image be denoted by $S_{1,2}$. Similarly, S_3 and S_4 are interleaved to create $S_{3,4}$. This completes the column interleaving. Similar

operation is then performed row wise on the two images $S_{1,2}$ and $S_{3,4}$. Let the final interleaved image be denoted by $S_{1,2,3,4}$. The image is then transformed to another intermediate image using affine cipher discussed in the next section.

2.2 *Affine Cipher*

Affine cipher is a one-to-one mapping, where a given plaintext is transformed to a unique cipher-text [8, 9]. In affine cipher, relationship between the plaintext and the cipher text is given by the following equations

$$C = (K_1 P + K_0) \bmod N \tag{1}$$

$$P = K_1^{-1}(C - K_0) \bmod N \tag{2}$$

where C stands for the cipher text, P for the plaintext and $N = 255$ for gray scale images. The key K_0 and K_1 are selected in the range [0, 255], so that decryption can be ensured. The function $C = (K_1 \mathrm{P} + K_0) mod N$ defines a valid affine cipher if K_1 is relatively co-prime to 256, and K_0 is an integer between 0 and 255 (both values inclusive).

3 Proposed Method

3.1 *Image Encryption*

A stepwise process consisting of interleaving and ciphering is described as follows:

Step 1. *Divide Image*: Partition the secret image S into four equal sub-images (quadrants) (Fig. 1a).

Step 2. *Interleave Image*: Let the columns of the first quadrant S_1 of the image be labeled as $1, ..., \lfloor n/2 \rfloor, \lfloor n/2 \rfloor + 1, ..., n$ and that of the second quadrant S_2 be denoted as $n + 1, ..., \lfloor 3n/2 \rfloor, \lfloor 3n/2 \rfloor + 1, ..., 2n$. Perform the image interleaving of and by arranging the columns in the sequence: $1, \lfloor n/2 \rfloor + 1, n + 1, \lfloor 3n/2 \rfloor + 1, 2, \lfloor n/2 \rfloor + 2, \lfloor n + 2 \rfloor, \lfloor 3n/2 \rfloor + 2, ..., \lfloor n/2 \rfloor, n, \lfloor 3n/2 \rfloor, 2n$.

Label the output as $S_{1,2}$. Similarly generate $S_{3,4}$. Finally use column interleaving between $S_{1,2}$ and $S_{3,4}$ to generate $S_{1,2,3,4}$. Figure 1b, c exhibit the results of interleaving upper quadrants S_1 and S_2 ($S_{1,2}$) and lower quadrants S_3 and S_4 ($S_{3,4}$). The resulting images shown in Fig. 1b, c are again interleaved row wise to generate the final interleaved image $S_{1,2,3,4}$ (Fig. 1d).

Step 3. *Apply Affine Cipher*: Choose the appropriate keys k_0, k_1 and apply the affine cipher on each of the pixels of the image using Eq. (1). After all the pixels are transformed, the intermediate encrypted image is generated and labeled as (Fig. 1e).

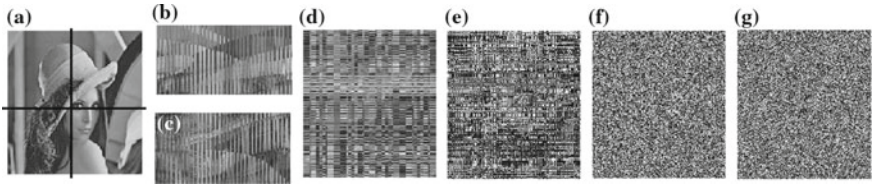


Fig. 1 Encryption process for Lena image **a** divided image, **b** & **c** column wise interleaving of the *upper* and *lower* two quadrants respectively, **d** row wise interleaving **b** and **c**, **e** affine cipher, **f** random grid R, **g** XORed image

Step 4. *XOR Operation* Generate a random grid of the same size as that of the original image ($2m \times 2n$) and perform XOR operation to get the final encrypted image $I = E \oplus R$.

Figure 1g shows the final encrypted image obtained after following the above steps. It may be worthwhile to mention here that the security property is satisfied here, since the image cannot be revealed without the keys, and the random grid. Further, image interleaving makes it too difficult to decipher the image content even if the parameters are known. The decryption process is discussed in the next section.

3.2 Image Recovery

The recovery of the image requires the following inputs. The encrypted image I, the random grid R and the secret keys K_0, K_1. The first step involves XOR operation of the random grid with the encrypted image I, to obtain the interleaved and affine encrypted image $E = I \oplus R$ (Fig. 6a). We now apply the inverse affine cipher (2) to each pixel of the image I. This gives us the intermediate image (Fig. 6b) which is then de-interleaved to obtain by applying the reverse process of the process explained in Step 2 of the encryption process. This reveals the original image (Fig. 6c).

4 Extention to Multi-secret Sharing

The scheme can easily be extended to share multiple secrets by considering the original image to be composed of more than one secret image. Number of secret images can be even or odd. For example to share four secret images of the same size, a new image is composed with four quadrants consisting secret images (refer Fig. 4) and for sharing three secrets any two quadrants can be selected to fit in the first secret image and, second and third secret images can be fitted in the remaining two quadrants (refer Fig. 5). Hence, the number of images per quadrant can be adjusted as required to share multiple secrets without any change in the encryption and recovery procedures.

5 Experimental Results and Discussion

For performing experimentation we use MATLAB version 7 as the platform supported by Windows 7. To analyze the performance of the proposed method, tests were performed on three different kinds of images—a dense image (Lena), a sparse image (Dice) and a text image as shown in Figs. 1, 2 and 3. Figures 4 and 5 demonstrate extension of the proposed method for sharing multiple secret images. In each of figures shown below part (a) depicts the secret image (grayscale, size 400×400) to be shared divided into four equal quadrants, part (b) depicts the column wise interleaving of upper two quadrants, part (c) depicts the column wise interleaving of lower two quadrants, part (d) depicts the result obtained after row wise interleaving of part (b) and part (c). The finally interleaved image as shown in part (d) successfully dissolves all the four quadrants by altering horizontal and vertical correlations between the pixels without any change in the original dimensions. The process is followed by performing encryption using affine cipher. Result after encryption is shown in part (e). Part (f) shows the random grid of the same size as that of the original image (400×400) obtained using a random number generating function. Part (e) depicts the final encrypted image obtained after performing pixel wise XOR operation between the interleaved image and Figs. 4 and 5 show similar operations for multiple secrets. Results obtained after the first step of recovery process followed by decryption of affine cipher and de-interleaving operations are shown in Figs. 5, 6 and 7 for Lena, Dice and Text Image respectively.

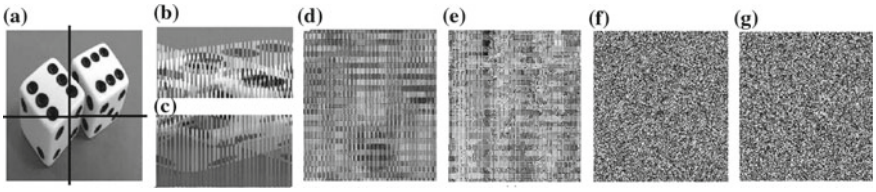


Fig. 2 Encryption process for Dice image **a** divided image, **b** & **c** column wise interleaving of the *upper* and *lower* two quadrants respectively, **d** row wise interleaving **b** and **c**, **e** affine cipher, **f** random grid R, **g** XORed image

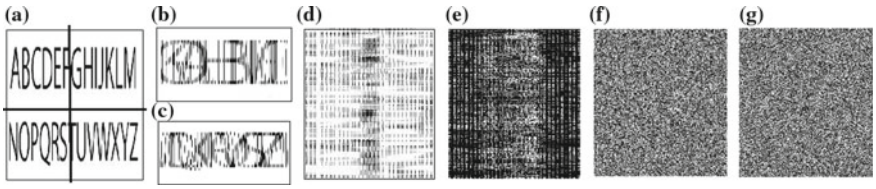


Fig. 3 Encryption process for Iext image **a** divided image, **b** & **c** column wise interleaving of the *upper* and *lower* two quadrants respectively, **d** row wise interleaving **b** and **c**, **e** affine cipher, **f** random grid R, **g** XORed image

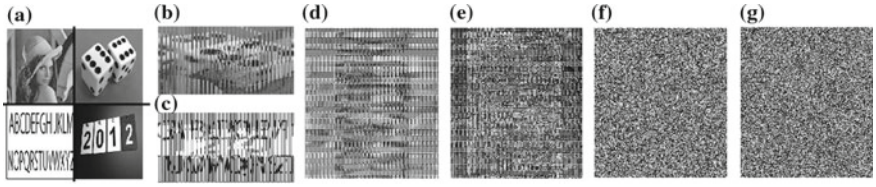


Fig. 4 Encryption process for four images **a** divided image, **b** & **c** column wise interleaving of the *upper* and *lower* two quadrants respectively, **d** row wise interleaving **b** and **c**, **e** affine cipher, **f** random grid R, **g** XORed image

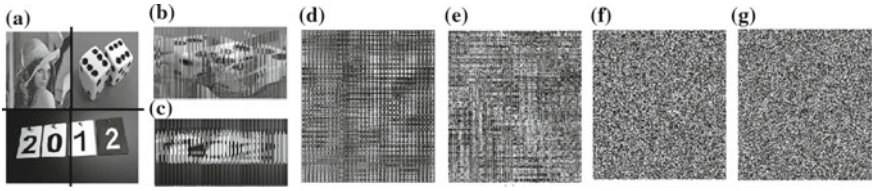


Fig. 5 Encryption process for three images **a** divided image, **b** & **c** column wise interleaving of the *upper* and *lower* two quadrants respectively, **d** row wise interleaving **b** and **c**, **e** affine cipher, **f** random grid R, **g**XORed image

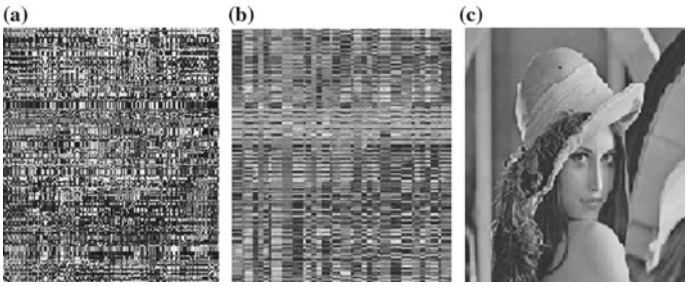


Fig. 6 Recovery of Lena image: **a** XOR operation with R, **b** decrypting affine cipher, **c** de-interleaved image

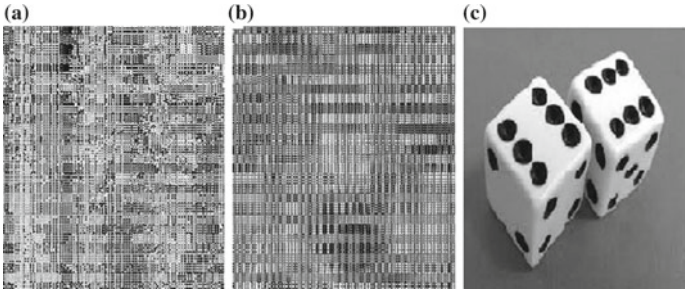


Fig. 7 Recovery of Dice image: **a** XOR operation with R, **b** decrypting affine cipher, **c** de-interleaved image

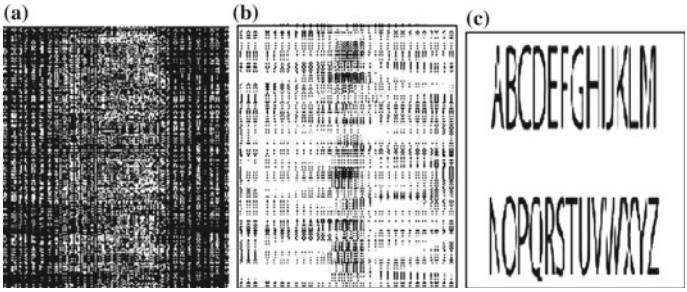


Fig. 8 Recovery of Text image: **a** XOR operation with R, **b** decrypting affine cipher, **c** de-interleaved image

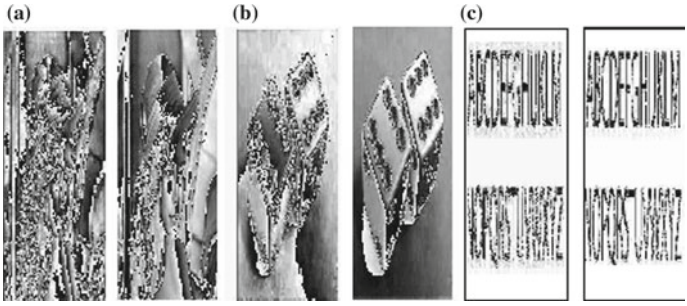


Fig. 9 Recovery results obtained after XOR operation for Chens method: **a** & **b** subshares for Lena image, **c** & **d** subshares for Dice image, **e** & **f** subshares for Text image

The experimental results show that interleaving operations in the proposed method provides extra layer of security. The main reason for interleaving is to protect the information even if the encryption keys are compromised. The proposed scheme provides a three layer security for sharing a secret image. The random grid technique provides security at first level and also renders a noisy appearance to the image. In case of Chens scheme the security of the shared information is breached if the random grid is available as the sub-images recovered after decryption tends to reveal the secret as shown in Fig. 9. As compared to the above results the recovered shares after XOR operation using the proposed method are much secured as shown in Figs. 6, 7 and 8.

The Hill cipher algorithm used by Chen [1] suffers from the problem of limited key space of matrices which have integral inverse. To overcome the problem, at the second level affine cipher is used to encrypt the information. Various techniques in literature suggest it is possible to partially obtain the information if the attacker guesses the keys or has partial known secret keys. To deal with such attacks, interleaving provides an extra third layer of security.

Even if the random grid and the encrypting keys are compromised, the revealed information will be the interleaved image and hence will not provide any guess about

Table 1 Comparison results for a few multi-secret sharing schemes

Author(s)	No. of secrets	Contrast	Pixel expansion	Recovery
Shyu et al. [10]	2	1/4	2	Lossy
	4	1/8	8	Lossy
Wu and Chen [11]	2	1/4	4	Lossy
Feng et al. [12]	2	1/6	6	Lossy
	4	1/12	12	Lossy
Hsu et al. [13]	2	1/2	4	Lossy
Proposed	2	1	1	Lossless
	4	1	1	Lossless

the original secret image. A comparison of the proposed scheme with some of the recent multiple secret sharing schemes is provided in Table 1.

6 Conclusion

In this paper, a novel secret sharing scheme is presented which is based on affine cipher, image interleaving and random grids. The scheme provides a solution to the security flaws observed in Hill cipher-based method introduced in [1]. As opposed to the Hill cipher-based method [1] where two layers of security is proposed, the scheme provides three layers of security. Further, the matrix used in Hill cipher-based method is required to have an integer inverse matrix, which is a major constraint not only in the construction, but also in extending the method to multi-secret sharing. The proposed method is easily extended to multi-secret sharing without making any major modifications. The scheme provides lossless recovery and is also not having any pixel expansion issues. Numerical results demonstrate robustness of the method.

References

1. Chen, W.K.: Image sharing method for gray-level images. Journal of Systems and Software **86** (2013) 581–585
2. Naor, M., Shamir, A.: Visual cryptography. In: Advances in Cryptology EUROCRYPT'94, Springer (1995) 1–12
3. Kafri, O., Keren, E.: Encryption of pictures and shapes by random grids. Optics letters **12** (1987) 377–379
4. Shyu, S.J.: Image encryption by multiple random grids. Pattern Recognition **42** (2009) 1582–1596
5. Chen, T.H., Tsao, K.H.: Threshold visual secret sharing by random grids. Journal of Systems and Software **84** (2011) 1197–1208
6. Guo, T., Liu, F., Wu, C.: k out of k extended visual cryptography scheme by random grids. Signal Processing **94** (2014) 90–101

7. Wu, X., Sun, W.: Improved tagged visual cryptography by random grids. Signal Processing **97** (2014) 64–82
8. William, S., Stallings, W.: Cryptography and Network Security, 4/E. Pearson Education India (2006)
9. De Palma, P., Frank, C., Gladfelter, S., Holden, J.: Cryptography and computer security for undergraduates. In: ACM SIGCSE Bulletin. Volume 36., ACM (2004) 94–95
10. Shyu, S.J., Huang, S.Y., Lee, Y.K., Wang, R.Z., Chen, K.: Sharing multiple secrets in visual cryptography. Pattern Recognition **40** (2007) 3633–3651
11. Chen, L., Wu, C.: A study on visual cryptography. Diss. Master Thesis, National Chiao Tung University, Taiwan, ROC (1998)
12. Feng, J.B., Wu, H.C., Tsai, C.S., Chang, Y.F., Chu, Y.P.: Visual secret sharing for multiple secrets. Pattern Recognition **41** (2008) 3572–3581
13. Hsu, H.C., Chen, T.S., Lin, Y.H.: The ringed shadow image technology of visual cryptography by applying diverse rotating angles to hide the secret sharing. In: Networking, Sensing and Control, 2004 IEEE International Conference on. Volume 2., IEEE (2004) 996–1001
14. Bunker, S.C., Barasa, M., Ojha, A.: Linear equation based visual secret sharing scheme. In: Advance Computing Conference (IACC), 2014 IEEE International, 2014 IEEE (2014) 406–410

Comprehensive Representation and Efficient Extraction of Spatial Information for Human Activity Recognition from Video Data

Shobhanjana Kalita, Arindam Karmakar and Shyamanta M. Hazarika

Abstract Of late, human activity recognition (HAR) in video has generated much interest. A fundamental step is to develop a computational representation of interactions. Human body is often abstracted using *minimum bounding rectangles* (MBRs) and approximated as a set of MBRs corresponding to different body parts. Such approximations assume each MBR as an independent entity. This defeats the idea that these are *parts* of the *whole* body. A representation schema for interaction between entities, each of which is considered as set of related rectangles or what is referred to as *extended objects* holds promise. We propose an efficient representation schema for extended objects together with a simple recursive algorithm to extract spatial information. We evaluate our approach and demonstrate that, for HAR, the spatial information thus extracted leads to better models compared to CORE9 [1] a compact and comprehensive representation schema for video understanding.

1 Introduction

Human activity recognition (HAR) deals with recognition of activities or interactions that include humans within a video [2]. This involves automated learning of interaction models, i.e. generalized descriptions of interactions. These models are then used for HAR in video [3]. Figure 1 shows an example of a *kick* activity from the Mind's Eye dataset.[1] By learning an interaction model for the activity, one is expected to recognize the *kick* activity in any video thereafter.

[1] http://www.visint.org.

S. Kalita (✉) · A. Karmakar · S.M. Hazarika
Biomimetic and Cognitive Robotics Lab, Computer Science and Engineering,
Tezpur University, Tezpur 784028, India
e-mail: kalitas@tezu.ernet.in

A. Karmakar
e-mail: arindam@tezu.ernet.in

S.M. Hazarika
e-mail: smh@tezu.ernet.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_8

Fig. 1 *Kick* activity from the Mind's Eye dataset

Qualitative spatio-temporal reasoning (QSTR) has been used for description of interactions—the interrelations between humans and objects involved in an activity. Use of *qualitative features* for description of video activities abstracts away noise and video specific details generating conceptually stronger models [1]. In QSTR approaches for HAR, humans are often abstracted as bounding boxes; this also abstracts away a lot of interaction details involving body parts. Human interactions are better described when the human body is viewed as a collection of parts [4]. However, most works that use such a part-based model of the human body view body parts as independent entities; this is counter-intuitive to the notion of body-parts being *part* of a *whole* body. Herein lies our motivation of developing a representation for extended objects. We define *extended objects* as entities having multiple components, approximated as a set of discrete rectangles.

Existing representation models for extended objects are either too restricted or too rich for HAR [5, 6]. CORE9 is a compact representation for various spatial aspects of a pair of rectangle objects [1] but is ineffective when discussing relations of extended objects. This paper proposes an extension of CORE9 to deal with extended objects, focusing on the topological and directional aspects.

2 QSTR for Activity Recognition in Video

Knowledge Representation and Reasoning (KR& R) is concerned with how symbolic *knowledge*, instead of quantitative information, can be used in an automated system for reasoning. KR& R methods in the area of video understanding are gaining popularity because they lead to more conceptual and generic models [7]. Logic-based learning has been used for learning models of video events as first-order logic formulae [3]. Human activities have also been described using first-order logic predicates [2] in a grammar based approach for HAR. It is worth noting, that most KR& R formalisms use qualitative abstractions of space-time for representation of knowledge pertaining to activities in the video [3].

Qualitative Reasoning within KR&R is concerned with capturing common-sense knowledge of the physical world through qualitative abstractions. Given appropriate reasoning techniques, the behaviour of physical systems can be explained without having to fall back on intractable or unavailable quantitative models [8]. QSTR provides formalisms for capturing common-sense spatial and temporal knowledge. Such

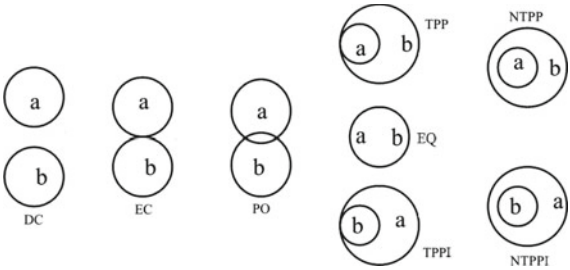


Fig. 2 The base relations of RCC8

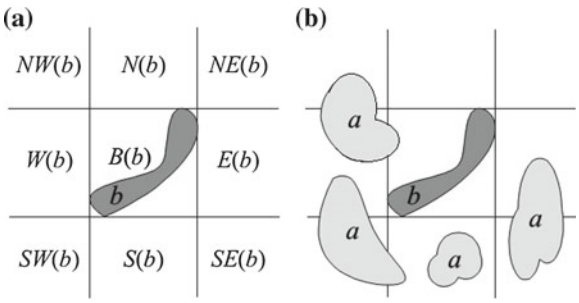


Fig. 3 **a** Cardinal directions of *b*, **b** *a* B:S:SW: W:NW:E:SE *b*

formalisms, notable for the ability to capture interactive information, are often used for description of video activities [3]. Topology and direction are common aspects of space used for qualitative description. Topology deals with relations unaffected by change of shape or size of objects; it is given as Region Connection Calculus (RCC-8) relations: Disconnected (DC), Externally Connected (EC), Partially Overlapping (PO), Equal (EQ), Tangential Proper Part (TPP) and its inverse (TPPI), and Non-Tangential Proper Part (nTPP) and its inverse (nTPPI) [9] (Fig. 2). Directional Relations are one of the 8 cardinal directions: North (N), NorthEast (NE), East (E), SouthEast (SE), South (S), SouthWest (SW), West (W), NorthWest (NW)—or as a combination [6]. Figure 3 shows cardinal direction relations for extended objects.

2.1 CORE9

CORE9 is a compact and comprehensive representation schema that allows qualitative topological, directional, size, distance and motion relations to be obtained by maintaining the *state* of nine *cores* of the *region of interest* (RoI) [1].

Given objects A and B, the RoI and nine cores are obtained as shown in Fig. 4. The state information (SI) of $core_{i,j}$(A, B) is $state_{i,j}$(A, B) and can have values: (i) AB if the core is a part of A ∩ B (ii) A if the core is a part of A – B (iii) B if

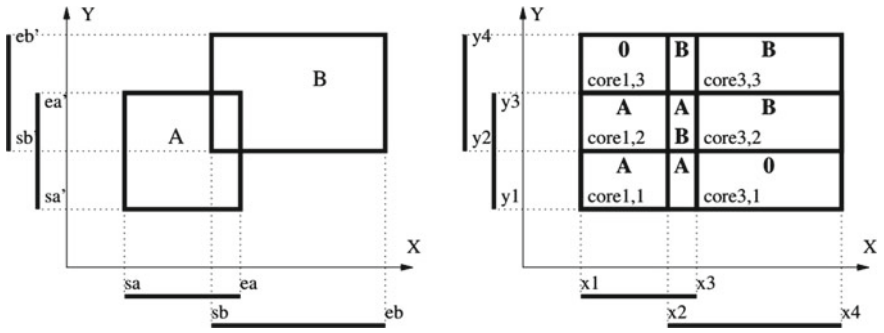


Fig. 4 The RoI and 9 cores of CORE9 [1]

the core is a part of B – A (iv) $\square$ if the core is not a part of A or B (v) ϕ if the core is only a line segment or point. The state of objects A and B in Fig. 4 is the 9-tuple $[A, A, \phi, A, AB, B, B, \phi, B, B]$. From this SI it is possible to infer that the RCC-8 relation between A and B is PO, because there is at least one core that is part of both A and B.

2.2 CORE9 for Extended Objects

In CORE9, objects are assumed to be single-piece; whereas, by definition, extended objects have multiple components/pieces. Within CORE9, there are two ways to deal with extended objects:

a. Approximate whole entity as single MBR: We write this as CORE9_W. The problem with CORE9_W is it cannot distinguish between configurations shown in Fig. 5; this is because all components of A and B are abstracted away with a single MBR.
b. Treat components as individual single-piece entities approximated using a single MBR: We write this as CORE9_C. The problem with CORE9_C is it fails to recognize the relation between entities A and B as a whole. Additionally, it computes all ${}^{m+n}C_2$ relations (where m and n are the number of components in A and B respectively). All intra-entity component relations are included despite being relatively uninteresting, especially for human *interactions*; for example in a *kick* activity it is interesting to note the sequence of relations between one person's foot with the object or another person's body part rather than the relations between the person's foot and his/her own hand.

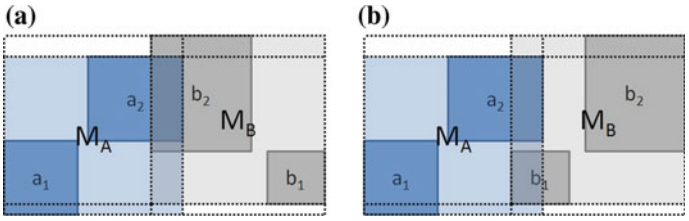


Fig. 5 Indistinguishable to CORE9$_W$

3 Extended CORE9

Consider a pair of extended objects, say A and B, such that $a_1, a_2, ..., a_m$ are m components of A and $b_1, b_2, ..., b_n$ are n components of B, i.e., $A = \bigcup_{i=1}^{m} a_i$ and $B = \bigcup_{i=1}^{n} b_i$. The MBR of a set of rectangles, MBR$(a_1, a_2, ..., a_n)$, is defined as the axis-parallel rectangle with the smallest area covering all the rectangles. To extract binary spatial information between A and B, we first obtain MBRs of A and B, i.e. MBR(A) = MBR$(a_1, a_2, ..., a_m)$ and MBR(B) = MBR$(b_1, b_2, ..., b_n)$. In Fig. 5, M_A is MBR(A), where extended object $A = a_1 \cup a_2$; similarly M_B is MBR(B). The nine cores of the extended objects A and B is obtained from MBR(A) and MBR(B) as defined in [1]. For each of the nine cores we store an *extended state information* (ESI) which tells us whether a particular core has a non-empty intersection with any of the components of the extended objects.

For each $core_{xy}(A, B)$ $x, y \in \{1...3\}$ of the whole MBRs of A and B, the ESI, $\sigma_{xy}(A, B)$, is defined as,

$$\sigma_{xy}(A,B) = \bigcup_{k=1}^{m}(a_k \cap core_{xy}(A,B)) \cup \bigcup_{k=1}^{n}(b_k \cap core_{xy}(A,B)) \tag{1}$$

and can have the following values:

$\{a_{i_1}, a_{i_2}, ...a_{i_k}, b_{j_1}, b_{j_2}, ...b_{j_k}\}$, where $i_1, i_2, ..., i_k \in \{1...m\}$, $j_1, j_2, ...j_k \in \{1...n\}$
if $\bigcup_{k=1}^{m}(a_k \cap core_{xy}(A,B)) \cup \bigcup_{k=1}^{n}(b_k \cap core_{xy}(A,B)) \neq \phi$

$\square = \{\phi\}$, if $\bigcup_{k=1}^{m}(a_k \cap core_{xy}(A,B)) \cup \bigcup_{k=1}^{n}(b_k \cap core_{xy}(A,B)) = \phi$

ϕ, if $core_{xy}(A,B) = \text{NULL}$

The ESI of the the objects, $\sigma(A, B)$, is defined as-

$$\sigma(A,B) = \begin{bmatrix} \sigma_{1,3}(A,B) & \sigma_{2,3}(A,B) & \sigma_{3,3}(A,B) \\ \sigma_{1,2}(A,B) & \sigma_{2,2}(A,B) & \sigma_{3,2}(A,B) \\ \sigma_{1,1}(A,B) & \sigma_{2,1}(A,B) & \sigma_{3,1}(A,B) \end{bmatrix} \tag{2}$$

In a human interaction, *component relations* are the relations between body parts of one person with body parts of another person/object. The overall relation between the interacting person(s)/object is obtained as a function of these component rela-

tions; we term this as *whole-relation*. Using the ESI, we are interested in computing the whole relations and inter-entity component relations.

3.1 Component Relations

Component relations are the relations between parts of one entity with parts of the other entity. We give a general recursive algorithm, Algorithm 1, to find all inter-entity component relations, $R(a_i, b_j)$. We focus on topological relations expressed as RCC8 relations [9] and directional relations expressed as cardinal directions [6]; the algorithm is valid for both topological and directional relations. The algorithm takes advantage of the fact that when two components are completely in different cores, the topological relation between two such components can be immediately inferred to be DC (a_1 and b_1 in Fig. 5a). On the other hand, the directional relation between two such components in different cores can be inferred following the cardinal directions:

$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{xz}(A,B) \wedge y > z \rightarrow a_i N b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{xz}(A,B) \wedge y < z \rightarrow a_i S\, b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zy}(A,B) \wedge x > z \rightarrow a_i E\, b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zy}(A,B) \wedge x < z \rightarrow a_i W b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zw}(A,B) \wedge x < z \wedge y < w \rightarrow a_i SW b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zw}(A,B) \wedge x > z \wedge y < w \rightarrow a_i SE b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zw}(A,B) \wedge x < z \wedge y > w \rightarrow a_i NW b_j$$
$$\forall core_{xy}(A,B), a_i \in \sigma_{xy}(A,B) \wedge b_j \in \sigma_{zw}(A,B) \wedge x > z \wedge y > w \rightarrow a_i NW b_j$$

An Illustrative Example: Consider the extended objects A and B ($A = a_1 \cup a_2 \cup a_3$ and $B = b_1 \cup b_2 \cup b_3$) as shown in Fig. 6. In the highest level of recursion, level 0 in Fig. 7, ESI of A and B will be:

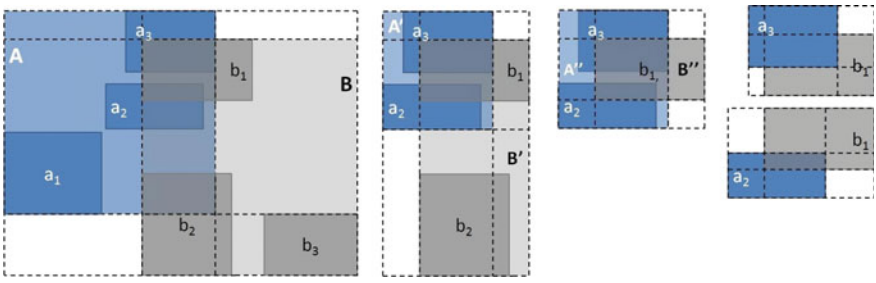


Fig. 6 The objects in the first three levels of recursion and the base case

Algorithm 1: *boolean* $Rel(\sigma(A,B))$, Recursive algorithm to find $R(a_i,b_j) \forall i \in \{1...m\}, \forall j \in \{1..n\}$

Input: $\sigma(A,B)$
Output: *boolean*
begin
  if $\forall core_{xy}(A,B),\ \sigma_{xy}(A,B) \cap \{a_i,b_j\} = \phi$ **then**
    $R(a_i,b_j) = \Psi$ ▷ $\Psi \in \mathfrak{R}$, where $\mathfrak{R}$ is a set of spatial relations
  else if $\forall core_{xy}(A,B),\ \exists \sigma_{xy}(A,B) - \{a_i,b_j\} = \phi$ **then**
    Compute $R(a_i,b_j)$ using CORE9 SI
  else if $\forall core_{xy}(A,B), \exists \sigma_{xy}(A,B) - \{a_i,b_j\} = \{a_{i_1},...a_{i_k},b_{j_1},...b_{j_k}\} \neq \phi$ **then**
    $A' = MBR(a_{i_1},a_{i_2},...a_{i_k})$
    $B' = MBR(b_{j_1},b_{j_2},...b_{j_k})$
    $newr \leftarrow Rel(\sigma(A',B'))$ ▷ Recursively find relations using the ESI
    if $newr = FALSE$ **then**
      forall the $i \in \{i_1,i_2,...i_k\}\ \varsigma\ j \in \{j_1,j_2,...j_k\}$ **do**
        Compute $R(a_i,b_j)$ using CORE9 SI
  else
    return *FALSE* ▷ no new relations are computed
  return *TRUE* ▷ at least one new relation is computed

$$\sigma(A,B) = \begin{bmatrix} \{a_3\} & \{a_3\} & \square \\ \{a_1,a_2,a_3\} & \{a_2,a_3,b_1,b_2\} & \{b_1,b_2\} \\ \square & \{b_2\} & \{b_2,b_3\} \end{bmatrix}$$

From this ESI, we can infer $R(a_1,b_1)$, $R(a_1,b_2)$, $R(a_1,b_3)$, $R(a_2,b_3)$, $R(a_3,b_3)$ are DC. Rest of the relations are recursively obtained from new objects $A' = a_2 \cup a_3$ and $B' = b_1 \cup b_2$ (where $a_2,a_3,b_1,b_2 \in core_{22}(A,B)$) as shown in Fig. 6; this happens at level 1 in Fig. 7. The ESI of A' and B' will be:

$$\sigma(A',B') = \begin{bmatrix} \{a_3\} & \{a_3\} & \square \\ \{a_2,a_3\} & \{a_2,a_3,b_1\} & \{b_1\} \\ \square & \{b_2\} & \{b_2\} \end{bmatrix}$$

From this ESI, we further infer that $R(a_2,b_2), R(a_3,b_2)$ are DC. For the rest of the relations we recursively compute $A'' = a_2 \cup a_3$ and $B'' = b_2$ (where $a_2,a_3,b_2 \in core_{22}(A',B')$) as shown in Fig. 6; this is level 2 in Fig. 7.

$$\sigma(A'',B'') = \begin{bmatrix} \{a_3\} & \{a_3\} & \square \\ \{a_2,a_3\} & \{a_2,a_3,b_1\} & \{b_1\} \\ \{a_2\} & \{a_2\} & \square \end{bmatrix}$$

At this stage, no new information is obtained using the ESI; hence CORE9 SI is used to infer $R(a_3,b_1)$ and $R(a_2,b_1)$ as PO. This is the base case and level 3 in Fig. 7. The

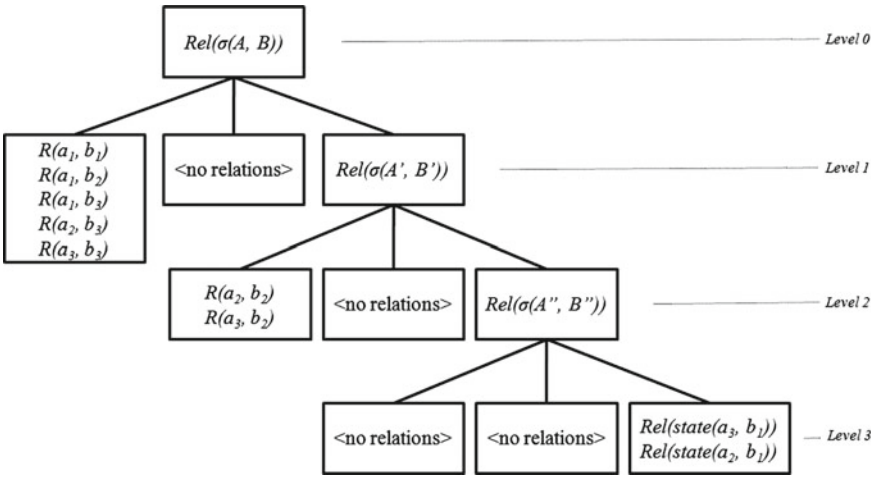


Fig. 7 The tree of recursive calls by Algorithm 1 on A and B of Fig. 6

recursive algorithm ensures that the number of computations is minimal for a given pair of extended objects.

Theorem 1 *Extended CORE9 is linear in the number of overlappings between components of the two objects.*

Proof Algorithm 1 computes only the most important (from the point of HAR) $m \times n$ opportunistically, requiring a computation only if there is an overlap between components. When components belong to different cores it is possible to immediately infer the spatial relation between them using ESI. In the worst case, if all components of A overlap all components of B the number of computations required would be mn. ■

For extended objects, A and B (say m and n components), CORE9 could use either of the two variants CORE9_W and CORE9_C for representation. In CORE9_C, the number of computations is quadratic in the total number of components of A and B, i.e. $O((m+n)^2)$. Note that CORE9_W has constant number of computations, this is at the expense of information loss (as detailed in Sect. 2.2). Whereas number of computations required to obtain all relations using ExtCORE9 is linear in the number of overlaps between components of A and B.

3.2 Whole-Relations

We derive whole relations between the extended objects, for both topological and directional aspects, from the component relations computed previously. The topological whole relation between A and B ($R(A, B)$), is obtained as follows:

if $\forall a_i, b_j, i \in \{1...m\}, j \in \{1...n\}, R(a_i, b_j) = \rho$ **then** $R(A, B) = \rho$
where $\rho \in \{PO, DC, EC, EQ, TPP, nTPP, TPPI, nTPPI\}$
if $\forall a_i, b_j, i \in \{1...m\}, j \in \{1...n\}, R(a_i, b_j) = EQ \vee TPP \vee nTPP$
then $R(A, B) = TPP$
if $\forall a_i, b_j, i \in \{1...m\}, j \in \{1...n\}, R(a_i, b_j) = EQ \vee TPPI \vee nTPPI$
then $R(A, B) = TPPI$
else $R(A, B) = PO$

The set of relations computed using Extended CORE9 include the set of inter-entity component relations and the whole relation; we write this as ExtCORE9$_W$. For comparison, we also consider a variant that includes only the component-wise relations without the whole-relations; we write this as ExtCORE9$_C$.

4 Human Activity Classification: Experimental Results

For classification we use Latent Dirichlet Allocation (LDA) which is a generative probabilistic model for discrete data [10]. Although LDA was originally designed for clustering text corpora into topics, it has been effectively used for clustering other forms of discrete data in the field of computer vision and image segmentation. Specifically, it has been used for human action classification [11, 12] and scene categorization [13] within the field of computer vision.

To evaluate effectiveness of ExtCORE9$_W$ and ExtCORE9$_C$ against CORE9 for HAR, we compare the respective activity classification results. For experimentation, we follow an approach similar to what was adopted by [12]. We extract the qualitative topological and directional relations amongst the interacting entities for each activity sequence. The qualitative relations thus obtained are treated as a bag of words describing the activity within a video; each video is treated as a document and the activity classes as topics that LDA is to model.

We experimented on short video sequences from the challenging Mind's Eye dataset. For our experiments we choose 10 videos each for five different actions: *approach, carry, catch, kick* and *throw*. We use the keyframes of the videos[2] and manually label the humans and objects involved in each of the keyframes. We evaluated ExtCORE9$_W$ using an implementation of LDA with Gibbs Sampling [14]. The videos are clustered into K different topics/activities; here K = 5. The LDA parameters α and β were set 10 and 0.1 respectively.

Table 1 shows clustering results obtained using features through ExtCORE9$_W$. Each cluster is assigned the activity class corresponding to the highest number of examples. We compute the *precision*, *recall* and *f1-scores*. The results are shown in Table 2. *Recall* values for activities *approach*, *catch* and *throw* are low because of the nature of these activities. In all of these activities, there is one entity that exists in the scenes throughout the video while the other entity either arrives in the scene at some later point in the video or leaves the scene midway through.

[2] We use I-frames obtained using the tool *ffmpeg* as keyframes, http://www.ffmpeg.org.

Table 1 Clustering results for ExtCORE9$_W$

	Class 0	Class 1	Class 2	Class 3	Class 4
Approach	2	2	**3**	0	3
Catch	**4**	1	1	3	1
Carry	0	0	0	**10**	0
Kick	1	**7**	1	0	1
Throw	1	0	1	4	**4**

Table 2 Quantitative evaluation

	Precision	Recall	F1-score
Approach	0.5	0.3	0.375
Catch	0.5	0.4	0.44
Carry	0.58	1	0.73
Kick	0.7	0.7	0.7
Throw	0.44	0.4	0.42

Similar LDA clustering experiments were performed using topological and directional features obtained using (a) CORE9$_W$ (b) CORE9$_C$ (c) ExtCORE9$_W$ and (d) ExtCORE9$_C$. A comparison of the f-measures is given in Fig. 8. For all activities used in the experimentation, the qualitative features obtained using ExtCORE9$_W$ provide a much better feature set for the activity compared to that obtained from CORE9$_W$. This is because CORE9$_W$ fails to recognize many interesting interaction details at the component level.

ExtCORE9$_W$ performs better for most activities compared to CORE9$_C$. In case of CORE9$_C$, even though all interaction details involving components are considered, a lot of unimportant intra-entity component-wise relations are incorporated as well, while losing out on the more interesting inter-entity whole relations. An interesting

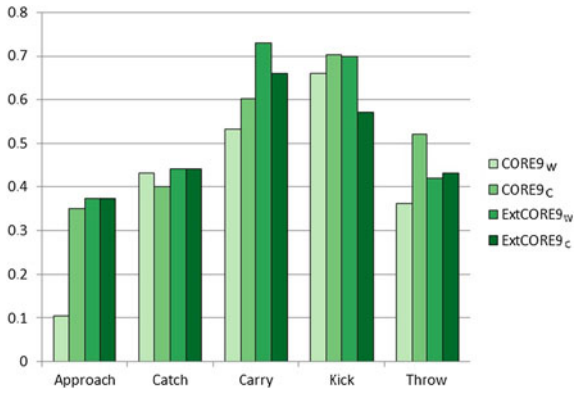


Fig. 8 F1-scores of **a** CORE9$_W$, **b** CORE9$_C$, **c** ExtCORE9$_W$, **d** ExtCORE9$_C$

result is seen in case of the activity *throw* where CORE9$_C$ achieves the best performance. We believe, this is because of the nature of the *throw* activity in which entities tend to be overlapping for the most part in the beginning and move apart suddenly; the entity being thrown is no longer in the scene and there is less evidence within the feature set of the activity regarding the moving apart phase. However, CORE9$_C$ utilizing the full set of inter-entity and intra-entity component relations as features provide a better description.

A similar case is seen in case of ExtCORE9$_C$. For most activity classes, performance results of ExtCORE9$_W$ is better; this emphasizes the importance of the inter-entity whole relations as computed by ExtCORE9$_W$. However, for the activity class *throw* ExtCORE9$_C$ performs marginally better. This shows that for activities in which there is less evidence of interaction amongst entities, using the inter-entity whole-relation only aggravates the classification results.

5 Final Comments

The part-based model of the human body obtained during tracking is easily seen as an extended object. ExtCORE9$_W$ leads to better interaction models by focusing on component-wise relations and whole relations of these extended objects. A recursive algorithm is used to opportunistically extract the qualitative relations using as few computations as possible. ExtCORE9$_W$ assumes components are axis-aligned MBRs. For single-component objects that are not axis-aligned, more accurate relations can be obtained [12]. Adapting ExtCORE9$_W$ such that objects and components are not axis-aligned is part of an ongoing research.

References

1. Cohn, A.G., Renz, J., Sridhar, M.: Thinking inside the box: A comprehensive spatial representation for video analysis. In: Proc. 13th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR2012). pp. 588–592. AAAI Press (2012)
2. Aggarwal, J., Ryoo, M.: Human activity analysis: A review. ACM Computing Surveys 43(3), 16:1–16:43 (Apr 2011)
3. Dubba, K.S.R., Bhatt, M., Dylla, F., Hogg, D.C., Cohn, A.G.: Interleaved inductive-abductive reasoning for learning complex event models. In: ILP. Lecture Notes in Computer Science, vol. 7207, pp. 113–129. Springer (2012)
4. Kusumam, K.: Relational Learning using body parts for Human Activity Recognition in Videos. Master's thesis, University of Leeds (2012)
5. Schneider, M., Behr, T.: Topological relationships between complex spatial objects. ACM Trans. Database Syst. 31(1), 39–81 (2006)
6. Skiadopoulos, S., Koubarakis, M.: On the consistency of cardinal directions constraints. Artificial Intelligence 163, 91 – 135 (2005)
7. Chen, L., Nugent, C., Mulvenna, M., Finlay, D., Hong, X.: Semantic smart homes: Towards knowledge rich assisted living environments. In: Intelligent Patient Management, vol. 189, pp. 279–296. Springer Berlin Heidelberg (2009)

8. Cohn, A.G., Hazarika, S.M.: Qualitative spatial representation and reasoning: An overview. Fundam. Inform. 46(1-2), 1–29 (2001)
9. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: Proc. of 3rd Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92). pp. 165–176. Morgan Kauffman (1992)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
11. al Harbi, N., Gotoh, Y.: Describing spatio-temporal relations between object volumes in video streams. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
12. Sokeh, H.S., Gould, S., J, J.: Efficient extraction and representation of spatial information from video data. In: Proc. of the 23rd Int. Joint Conf. on Artificial Intelligence (IJCAI'13). pp. 1076–1082. AAAI Press/IJCAI (2013)
13. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 524–531 (2005)
14. Phan, X.H., Nguyen, C.T.: GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) (2007)

Robust Pose Recognition Using Deep Learning

Aparna Mohanty, Alfaz Ahmed, Trishita Goswami, Arpita Das, Pratik Vaishnavi and Rajiv Ranjan Sahay

Abstract Current pose estimation methods make unrealistic assumptions regarding the body postures. Here, we seek to propose a general scheme which does not make assumptions regarding the relative position of body parts. Practitioners of Indian classical dances such as *Bharatnatyam* often enact several dramatic postures called *Karanas*. However, several challenges such as long flowing dresses of dancers, occlusions, change of camera viewpoint, poor lighting etc. affect the performance of state-of-the-art pose estimation algorithms [1, 2] adversely. Body postures enacted by practitioners performing Yoga also violate the assumptions used in current techniques for estimating pose. In this work, we adopt an image recognition approach to tackle this problem. We propose a dataset consisting of 864 images of 12 *Karanas* captured under controlled laboratory conditions and 1260 real-world images of 14 *Karanas* obtained from Youtube videos for Bharatnatyam. We also created a new dataset consisting of 400 real-world images of 8 Yoga postures. We use two deep

A. Mohanty (✉) · R.R. Sahay
Department of Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: aparnamhnty@gmail.com

R.R. Sahay
e-mail: rajivsahay@gmail.com

A. Ahmed · T. Goswami · A. Das
Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Kolkata, India
e-mail: alfazahmed@gmail.com

T. Goswami
e-mail: trstgsm@gmail.com

A. Das
e-mail: dasarpita1993@gmail.com

P. Vaishnavi
Sardar Vallabhai National Institute of Technology Surat, Surat, India
e-mail: pratik18v@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_9

learning methodologies, namely, convolutional neural network (CNN) and stacked auto encoder (SAE) and demonstrate that both these techniques achieve high recognition rates on the proposed datasets.

Keywords Pose estimation · Deep learning · Convolutional neural network · Stacked auto encoder

1 Introduction

The current state-of-the-art pose estimation methods are not flexible enough to model horizontal people, suffers from double counting phenomena (when both left and right legs lie on same image region) and gets confused when objects partially occlude people. Earlier works on pose estimation [1, 2], impose a *stick-man* model on the image of the body and assume that the head lies above the torso. Similarly, shoulder joints are supposed to be higher than the hip joint and legs. However, these assumptions are unrealistic and are violated under typical scenarios shown in this work. As an example, we show how one state-of-the-art approach [1] fails to estimate the pose correctly for an image taken from the standard PARSE dataset as shown in Fig. 1a, b. The images of Indian classical dance (ICD) and Yoga too have such complex configuration of body postures where current pose estimation methods fail as shown in Fig. 1c, d. There exists a set of 108 dance postures named *Karanas* in the original Natya Shastra enacted by performers of Bharatnatyam. Yoga too is popular as a system of physical exercise across the world. Several challenges such as occlusions, change in camera viewpoint, poor lighting etc. exist in the images of body postures in dance and Yoga. The proposed ICD and Yoga dataset have such complex scenarios where the head is not necessarily above the torso, or have horizontal or overlapping people, twisted body, or objects that partially occlude people. Hence, we also tested Ramanan et al. approach [1] on the proposed dataset and the results are depicted in Fig. 1c, d. The results of another recent technique using tree models for pose estimation proposed by Wang et al. [2] on our dataset are also reported in Sects. 6.1 and 6.3.

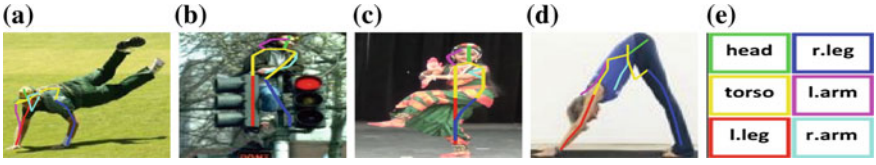


Fig. 1 Failure of state-of-the-art approach [1] on few images from PARSE [3] and our datasets. **a** and **b** represents failure results of [1] on PARSE dataset. **c** and **d** represent failure results of [1] on our ICD and Yoga datasets. Failure cases emphasise the lacuna of approach in [1] to model horizontal people as in (**a**) and it's inability to handle partially occluded people as shown in (**b**). The color assignment of parts is depicted in (**e**)

Deep learning has recently emerged as a powerful approach for complex machine learning tasks such as object/image recognition [4], handwritten character recognition [5] etc. The ability of deep learning algorithms to not rely on the hand crafted features to classify the images motivated us to use it for pose identification in typical situations wherein pose estimation algorithms such as [1, 2] fail due to unrealistic assumptions. Since there is no publicly available dataset on ICD we created our own dataset containing images of twelve dance postures collected in laboratory settings and a dataset of fourteen poses from videos on Youtube. We also created a dataset of eight Yoga poses to show the efficacy of a trained CNN model and a SAE in identifying postures dance and Yoga.

Because of limited labeled data we used data augmentation and transfer learning. We used a pre-trained model which is trained with a large dataset such as MNIST [5]. Interestingly, we observe a significant reduction in time taken to train a pre-trained network on our datasets and also improvements in accuracy.

2 Prior Work

There are several works in literature pertaining to the identification of poses. Mallik in their work in [6] tried to preserve the living heritage of Indian classical dance. However, unlike our work, they do not identify body postures of the dancer. To classify ICD a sparse representation based dictionary learning technique is proposed in [7]. In the literature there are very few significant works addressing the problem of recognition of postures in ICD but a vast literature on general pose identification of humans exists. Initial works for 2D pose estimation in the images/video domains is [8]. Entire human shapes have been matched in [9].

Discriminatively trained, multi scale, deformable part based model for pose estimation is proposed in [10]. This idea is also used for object detection in [11]. Felzenszwalb et al. [12] describe a statistical framework for representing the visual appearance of objects composed of rigid parts arranged in a deformable configuration. A generic approach for human detection and pose estimation based on the pictorial structures framework in [13] is proposed by Andriluka et al.

Very recently, a deep learning approach using CNNs has been used for estimating pose in [14] but it does not deal with complex datasets like ICD and Yoga as in this work. Recently several models which incorporated higher order dependencies while remaining efficient in [15] have been proposed. A state-of-the-art method for pose estimation using tree models is given in [2]. A new hierarchial spatial model that can capture an exponential number of poses with a compact mixture representation is given in [16]. Still images were used for estimating 2D human pose by Dantone et al. by proposing novel, nonlinear joint regressors in [17]. A method for automatic generation of training examples from an arbitrary set of images and a new challenge of joint detection and pose estimation of multiple articulated people in cluttered sport scenes is proposed by Pischchulin et al. [18]. Eichner et al. [19] are capable of estimating upper body pose in highly challenging uncontrolled images, without prior

knowledge of background, clothing, lighting, or the location and scale of the person. A learning based method for recovering 3D human body pose from single images and monocular image sequences is given by [20]. An efficient method to accurately predict human pose from a single depth image is proposed by Shotton et al. [21].

3 Deep Learning Framework for Pose Identification

3.1 CNN: Architecture

The general architecture of the proposed CNN is shown in Fig. 2a. Apart from the input and the output layers, it consists of two convolution and two pooling layers. The input is a 32×32 pixels image of a dance posture.

As shown in Fig. 2a, the input image of 32×32 pixels is convolved with 10 filter maps of size 5×5 to produce 10 output maps of 28×28 in layer 1. The output convolutional maps are downsampled with max-pooling of 2×2 regions to yield 10 output maps of 14×14 in layer 2. The 10 output maps of layer 2 are convolved with each of the 20 kernels of size $5 \times 5 \times 10$ to obtain 20 maps of size 10×10. These maps are further downsampled by a factor of 2 by max-pooling to produce 20 output maps of size 5×5 of layer 4. The output maps from layer 4 are concatenated to form a single vector while training and fed to the next layer. The quantity of neurons in the final output layer depends upon the number of classes in the database.

3.2 Stacked Auto Encoder (SAE): Architecture

Auto-encoder is an unsupervised learning approach for dimensionality reduction. Auto-encoder can be used for encoding which can then be followed by a decoder

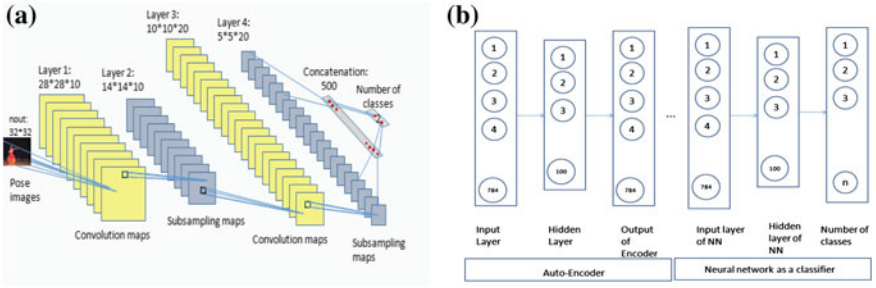


Fig. 2 **a** Architecture of the proposed CNN model used for pose and Yoga classification. **b** Detailed block diagram of the proposed SAE architecture used for pose and Yoga classification

(neural network) to classify the input [22]. The architecture of the proposed SAE is shown in Fig. 2b. In SAE the image inputs are fed to the hidden layer to extract features as seen in Fig. 2b. Then the features are fed to the output layer of SAE to reconstruct back the original input. Output of the last layer is treated as input to a classifier. We use a neural network as a classifier, training it to map the features extracted to the output labels. We used an input of 784 nodes followed by a hidden layer of 100 nodes before 784 number of output nodes. This SAE is followed by a neural network having 784 input nodes, 100 hidden nodes and output nodes identical to the number of classes as shown in Fig. 2b.

4 Data Augmentation

It has been shown in [4] that data augmentation boosts the performance of CNNs. We performed data augmentation for the Yoga dataset so as to increase the number of labeled data. We did not augment the synthetic and Youtube based pose databases since the number of images was substantial. The Yoga pose database has only 8 classes with 50 images per class. We performed data augmentation of the training data by five-fold cropping and resizing images to original size. Of the 50 images per class we used 40 images per class for training which we augmented 5 times to 200 images per class. The test images were 10 per class. Hence, we obtained a total of 1600 training images and 80 test images for all 8 classes.

5 Transfer Learning

Because of limited labeled training data, the proposed CNN is pre-trained from randomly initialized weights using MNIST [5] which contains 50,000 labeled training images of hand-written digits. The CNN is trained for 100 epochs with this data yielding an MSE of 0.0034 and testing accuracy of 99.08 % over 10,000 images. The convereged weights of the trained network are used to initialize the weights of the CNN model to which our dataset of dance poses and Yoga dataset were given as input. We obtained much faster convergence during training with a pre-trained network and improved accuracies on the test data.

6 Experimental Results

6.1 CNN Model: Pose Data

Training Phase: Synthetic Case The onstrained database used for training the proposed CNN architecture described in subsection 3.1 consists of 864 images which were captured using Kinect camera originally at 640×480 pixels resolution. We used images of 12 different poses as shown in Fig. 3a enacted 12 times by 6 different

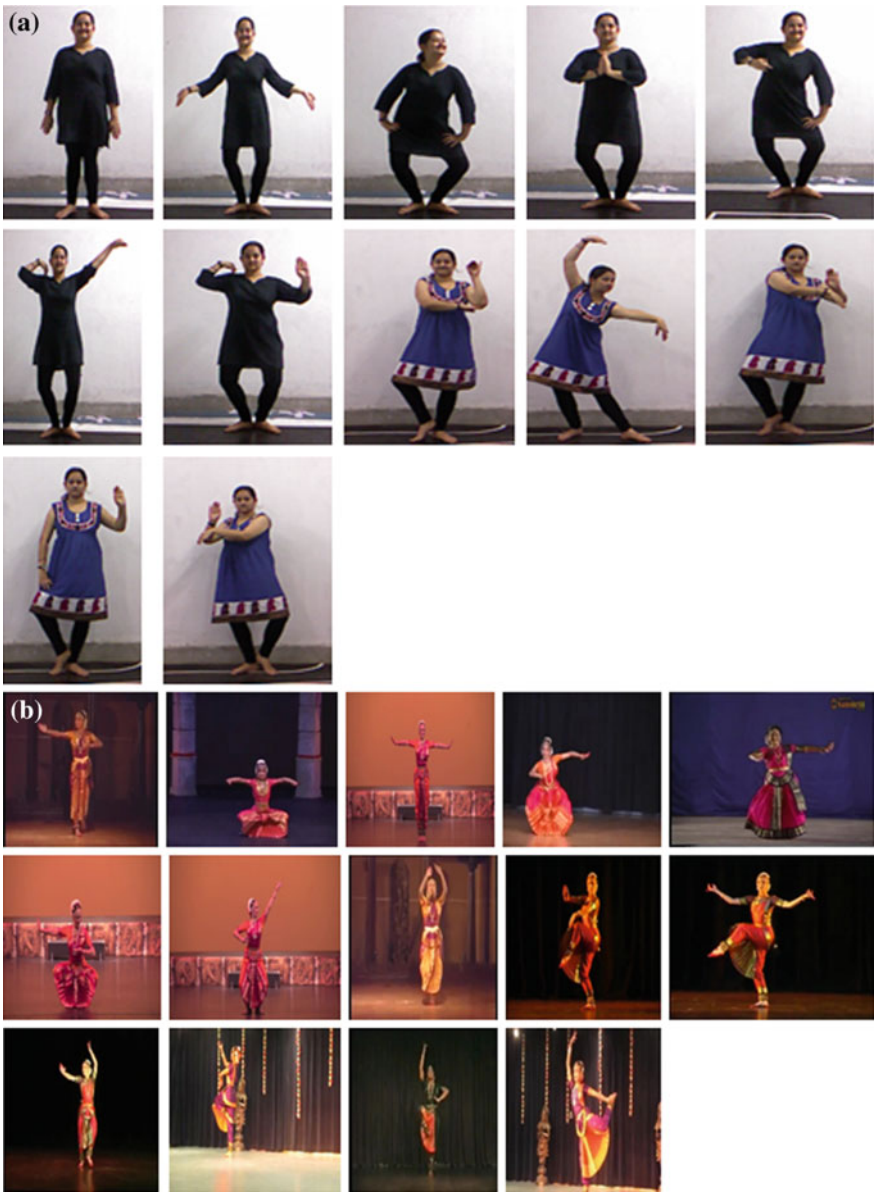


Fig. 3 **a** A snapshot of twelve *Karanas* collected in constrained environment. **b** A snapshot of fourteen Karanas extracted from Youtube videos

volunteers. The training set is composed of 10 images of each pose enacted by 6 different persons leading to a total of 720 photos. The test set is made up of the rest

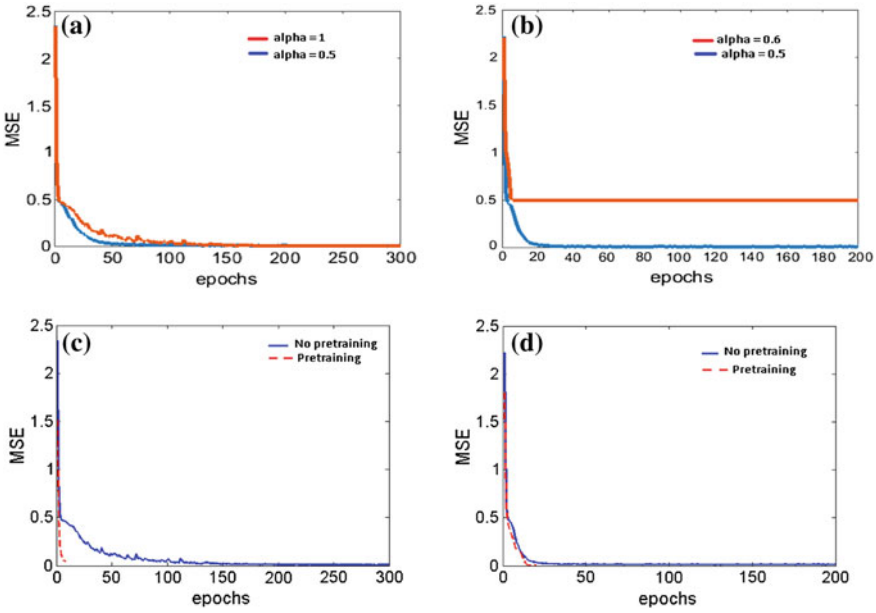


Fig. 4 **a** Mean squared error (MSE) versus epochs for the CNN trained on the synthetic pose dataset. **b** MSE versus epochs plot for pose data from Youtube videos. **c** Effect of pre-training on synthetic pose data using a CNN pre-trained with MNIST data. **d** Effect of pre-training on real world pose data using a pre-trained model

of 144 images. There is no overlap between the training and the test datasets. All images are down-sampled to 32×32 pixels before feeding to the CNN.

The weights of the proposed CNN are trained by the conventional back-propagation method using the package in [23]. The total number of learnable parameters in the proposed CNN architecture is 6282. We have chosen batch size as 4 and constant learning rate $\alpha = 1$ throughout all the layers. The network is trained for 300 epochs using random initialization of weights on a 3.4 GHz Intel Core i7 processor with 16 GB of RAM. The variation of the mean square error (MSE) versus epochs during the training phase is shown in Fig. 4a and the final MSE during training is 1.53 % in 300 epochs. Interestingly, by using a pre-trained MNIST model to initialize the weights of a CNN we could achieve an MSE of 1.74 % in only 15 epochs as represented in Fig. 4c.

Testing Phase For the testing phase, we give as input to the trained CNN model images from the test dataset. Given a test image, the output label with maximum score is chosen at the output layer of the CNN. The accuracy for 144 images is 97.22 %. By using transfer learning we could achieve an improved accuracy of 98.26 % in only 15 epochs as compared to 97.22 % with 300 epochs in case of random initialization of weights as shown in Table 1.

Table 1 Performance of the proposed CNN method on our proposed pose dataset of synthetic pose (ICD), real-world pose (ICD) and Yoga dataset

Data	No. of classes	Training set	Testing size	α	Batch	Epochs	MSE	Proposed approach (%)	Transfer learning (%)
Synthetic pose (ICD)	12	720	144	0.5	5	300	0.0153	97.22	98.26 (15 epochs)
Real-world pose (ICD)	14	1008	252	0.5	4	200	0.0258	93.25	99.72 (2 epochs)
Yoga data	8	1600	80	0.5	5	500	0.0062	90.0	

Training Phase: Real-World Data We downloaded some dance videos from the Youtube. The extracted frame is then re-sized to 100×200 pixels. We created a dataset of such real-world images for 14 different poses performed by 6 different dancers extracting 15 frames per pose for each dancer. A snapshot of the 14 postures is depicted in Fig. 3b. To create the training set, we used 12 frames per pose for each of the 6 performers leading to 1008 images. The testing set consisted of the rest 252 images. Similar to the synthetic case, there is no overlap between the training and testing sets. All images were further re-sized to 32×32 pixels before feeding to the CNN.

The CNN model was trained for 200 epochs using random intial weights with batch size as 4 and constant learning rate $\alpha = 0.5$ throughout all the layers. The variation of the mean square error (MSE) versus epochs during the training phase is shown in Fig. 4b. By using a pre-trained MNIST model to initialize the weights of a CNN we could achieve an MSE of 0.37 % in only 2 epochs as represented in Fig. 4d. The first layer filter kernels for an image from the Youtube pose dataset (in Fig. 5a) is shown in Fig. 5b and the convolved outputs at the first layer are shown in Fig. 5c.

Testing Phase The test set containing 252 images is input to the trained CNN and yields an overall accuracy of 93.25 %. By using transfer learning we could achieve an improved accuracy of 99.72 % in only 2 epochs as compared to 93.25 % with 200 epochs for the random initialization of weights as shown in Table 1. The existing state-of-the-art methods for pose estimation [1, 2] work well for the standard datasets, but fail to perform on our proposed dataset due to the complexity in our dataset involving illumination, clothing and clutter. The failure cases of the state-of-the-art approaches [1, 2] on the proposed dataset is shown in Fig. 6a, b. The strong performance of the proposed CNN architecture shows that it is an apt machine learning algorithm for identifying dance postures *(Karanas)*.

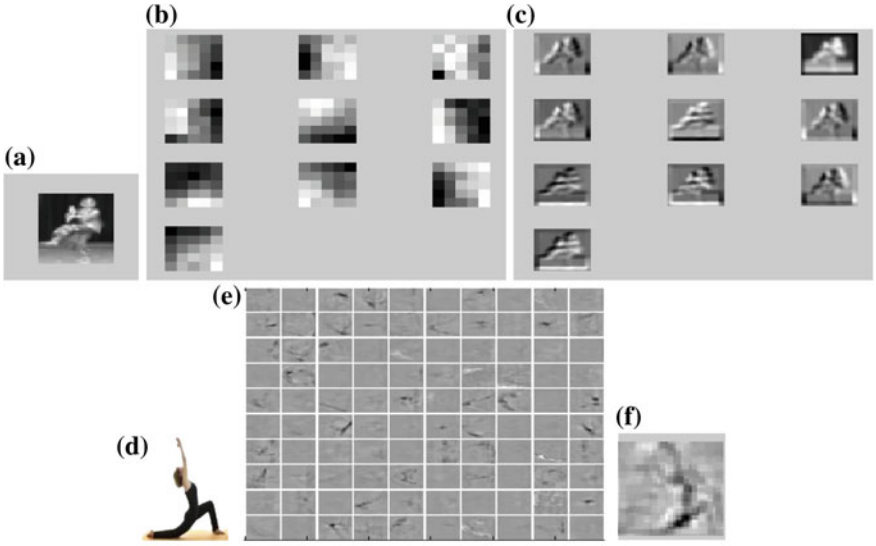


Fig. 5 **a** Original input image to a CNN. **b** First layer filter kernels in the Youtube pose dataset of a CNN. **c** First layer convolved output in Youtube pose dataset. **d** The input Yoga pose. **e** First layer filters of the SAE for the Yoga data. **f** The reconstructed output of the SAE for the Yoga pose in (**d**)

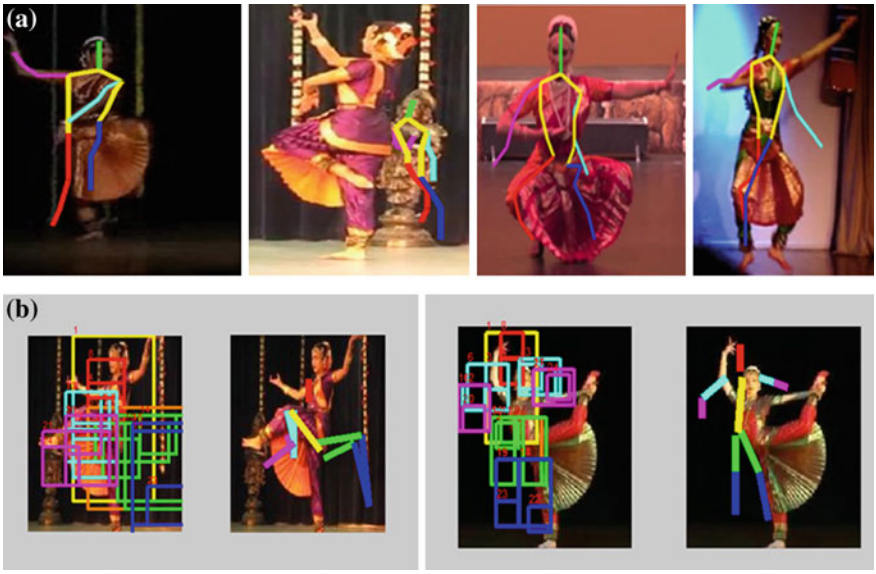


Fig. 6 Comparision with state-of-the-art: **a** Some images of Karanas from our proposed dataset where the approach proposed by Ramanan et al. [1] fails. **b** Failure results of Wang et al. [2] due to the complexity of our dataset with regard to illumination, clutter in the background, clothing etc.

Table 2 Performance of the proposed SAE method on our proposed pose dataset of synthetic pose, real-world pose and yoga dataset

Data	No. of	Training classes	Testing set	α, Batch size, Epochs of auto-encoder	α, Batch size, Epochs of neural network	Testing accuracy (%)
Synthetic pose (ICD)	12	720	144	0.5, 4, 1000	0.5, 4, 1000	86.11
Real-world pose (ICD)	14	1008	252	0.5, 4, 200	0.5, 4, 200	97.22
Yoga data	8	1600	80	0.09, 5, 500	0.09, 5, 500	70.0

6.2 SAE Model: Pose Dataset

As explained earlier, our SAE consisting of three layer along with a neural network is used to classify images of both ICD and Yoga data-sets. The accuracy obtained by using a stacked auto encoder for the synthetic pose data is 86.11 % and for the real-world pose data is 97.22 %. The details of using the stacked auto encoder is reported in Table 2.

6.3 CNN Model: Yoga Dataset

Training Phase We downloaded 50 images per class for 8 Yoga postures and re-sized them to 100×100 pixels. A snapshot of these 8 Yoga postures is depicted in Fig. 7. To create the training set, we used 40 images per pose. The testing set consisted of the rest 10 images per pose. There is no overlap between the training and testing sets. Then we performed data augmentation by cropping successively and resizing to original size. All images were further re-sized to 32×32 pixels before feeding to the CNN. The CNN model was trained for 500 epochs from random initial weights with batch size as 5 and constant learning rate $\alpha = 0.5$ throughout all the layers.

Testing Phase The test set containing 80 images as input to the trained CNN and yields an overall accuracy of 90 %. The existing state-of-the-art methods for pose estimation [1, 2] fail to perform on our proposed dataset due to poor illumination,

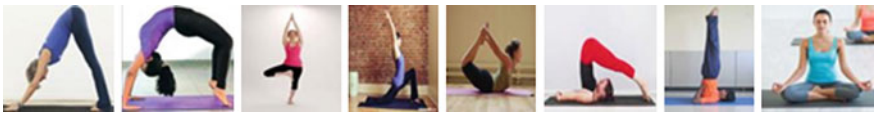


Fig. 7 A snapshot of ten Yoga poses from our dataset

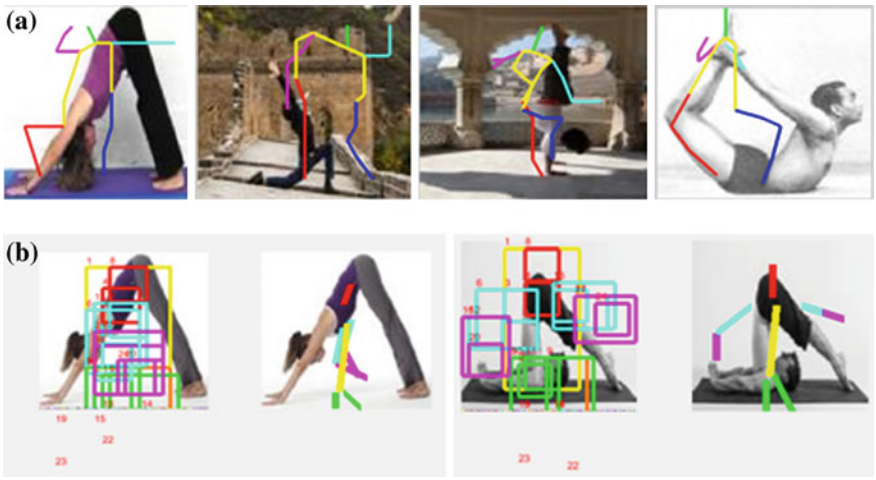


Fig. 8 A snapshot of Yoga poses extracted from our proposed dataset **a** where the state-of-the-art approach proposed by Ramanan et al. [1] fails and **b** where Wang et al. [2] fails due to the complexity associated with our dataset i.e. twisted body, horizontal body etc

clothing on body parts and clutter in the background. Importantly, note that the assumption of the head being above the torso always is not satisfied for these images. The failure of the existing state-of-the-art methods of [1, 2] on the complex Yoga dataset is represented in Fig. 8a, b respectively.

SAE Model: Yoga Data Auto encoders were stacked as in case of pose data to use it for initialising the weights of a deep network which was followed by a neural network to classify the poses. An image depicting Yoga posture is input to the SAEs is shown in Fig. 5d. The 100 filters in the first layer of the SAE is shown in Fig. 5e. The reconstructed output of the SAE for the Yoga dataset for a single Yoga posture is shown in Fig. 5f. The accuracy obtained by using a stacked auto encoder for the Yoga dataset is 70 %. The details regarding the proposed SAE is reported in Table 2.

7 Conclusions

The state-of-the-art approaches [1, 2] are not robust enough for estimating poses in conditions such as bad illumination, clutter, flowing dress, twisted body commonly found in the images in the proposed datasets of ICD and Yoga. Hence, a deep learning framework is presented here to classify the poses which violate the assumptions made by state-of-the-art approaches such as the constraint that the head has to be above the torso which is not necessarily maintained in ICD or Yoga. The proposed CNN and SAE models have been demonstrated to be able to recognize body postures to a high degree of accuracy on both ICD and Yoga datasets. There are several chal-

lenges in the problem addressed here such as occlusions, varying viewpoint, change of illumination etc. Both ICD and Yoga has various dynamic poses which we aim to classify by analyzing video data in our future work.

References

1. Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011, pp. 1385–1392.
2. F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *CVPR*, 2013, pp. 596–603.
3. D. Ramanan, "Learning to parse images of articulated bodies," in *NIPS*, 2006, pp. 1129–1136.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
5. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.
6. A. Mallik, S. Chaudhury, and H. Ghosh, "Nrityakosha: Preserving the intangible heritage of indian classical dance," *Journal on Computing and Cultural Heritage (JOCCH)*, vol. 4, no. 3, p. 11, 2011.
7. S. Samanta, P. Purkait, and B. Chanda, "Indian classical dance classification by learning dance pose bases," in *IEEE Workshop on Applications of Computer Vision (WACV)*, 2012, pp. 265–270.
8. J. O'Rourke and N. Badler, "Model-based image analysis of human motion using constraint propagation," *IEEE Trans. PAMI*, vol. 2, no. 6, pp. 522–536, Nov 1980.
9. G. Mori and J. Malik, "Estimating human body configurations using shape context matching," in *ECCV*, ser. 02, 2002, pp. 666–680.
10. P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008, pp. 1–8.
11. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
12. P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Intnl. Jrnl. Comp. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.
13. M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009, pp. 1014–1021.
14. A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014, pp. 1653–1660.
15. L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *CVPR*, 2013, pp. 588–595.
16. Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 256–269.
17. M. Dantone, J. Gall, C. Leistner, and L. Van Gool, "Human pose estimation using body parts dependent joint regressors," in *CVPR*, 2013, pp. 3041–3048.
18. L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *CVPR*, 2012, pp. 3178–3185.
19. M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *IJCV*, vol. 99, no. 2, pp. 190–214, 2012.
20. A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. PAMI*, vol. 28, no. 1, pp. 44–58, 2006.

21. J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.
22. J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *NIPS*, 2012, pp. 341–349.
23. R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, 2012. [Online]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox

A Robust Scheme for Extraction of Text Lines from Handwritten Documents

Barun Biswas, Ujjwal Bhattacharya and Bidyut B. Chaudhuri

Abstract Considering the vast collection of handwritten documents in various archives, research studies for their automatic processing have major impact in the society. Line segmentation from images of such documents is a crucial step. The problem is more difficult for documents of major Indian scripts such as Bangla because a large number of its characters have either ascender or descender or both and the majority of its writers are accustomed in extremely cursive handwriting. In this article, we describe a novel strip based text line segmentation method for handwritten documents of Bangla. Moreover, the proposed method has been found to perform efficiently on English and Devanagari handwritten documents. We conducted extensive experimentations and its results show the robustness of the proposed approach on multiple scripts.

Keywords Handwritten document analysis · Line segmentation · Piecewise projection profile · Connected component

1 Introduction

Segmentation of text lines from offline handwritten document images is a major step in Optical Character Recognition (OCR) task. The difficulties arise mainly due to various idiosyncracies of the writing styles of its writers such as widely varying inter-line distance, presence of irregular skew, touching and overlapping text-lines etc. Line segmentation of offline handwritten documents is more challenging than

B. Biswas (✉) · U. Bhattacharya · B.B. Chaudhuri
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute,
Kolkata 108, India
e-mail: barun.isical@gmail.com

U. Bhattacharya
e-mail: ujjwal@isical.ac.in

B.B. Chaudhuri
e-mail: bbc@isical.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_10

its online counterparts due to the availability of less information. Although several studies [1–5] of this problem can be found in the literature, it still remains an open field of research. That is why several handwriting text line segmentation contests have been held recently in conjunction with a few reputed conferences [6].

In this article, we present a novel and simple method based on certain divide and conquer strategy for line segmentation of textual documents irrespective of the script. We simulated the proposed approach on a standard dataset and the recognition results are comparable with the state-of-the-art approaches.

The remaining part of this paper is organized as follows: Sect. 2, provides a brief survey of the existing works. The proposed approach has been described in Sect. 3. Results of our experimentation have been provided in Sect. 4. Conclusion is drawn in Sect. 5.

2 Previous Works

In a number of reports [1, 2, 7] of line segmentation of handwritten documents, existing approaches have been categorized into a number of classes. These are projection profile-based [3, 8], smearing-based [1, 9], Hough transform-based [10], thinning-based [11] approaches.Among the other proposed methods Löthy et al. [12] used a hidden Markov model while LiE and Zheng [13] used a boundary growing approach. Yin and Liu [14] used the variational Bayes method, Brodic and Milivojevic [15] proposed the use of an adapted Water flow algorithm and Papavassiliou et al. [3] used the binary morphology algorithm. Dinh et al. [16] employed a voting based method for this text line segmentation problem.

3 Proposed Segmentation Strategy

The proposed scheme is based on a divide and conquer strategy. Where the input document is first divided into several vertical strips and text lines in each strip are identified. Next, the individual text lines of a strip are associated with the corresponding text lines (if any) of the adjacent strip to the right side. This association process starts from the two consecutive leftmost strips of the document and is terminated at the two consecutive rightmost strips. Finally, the text lines of the entire document get segmented. The overall flow of the process is shown in Fig. 1. Specific strategies are employed for (i) consecutive text lines which vertically overlap within a strip or (ii) touching texts in vertically adjacent lines.

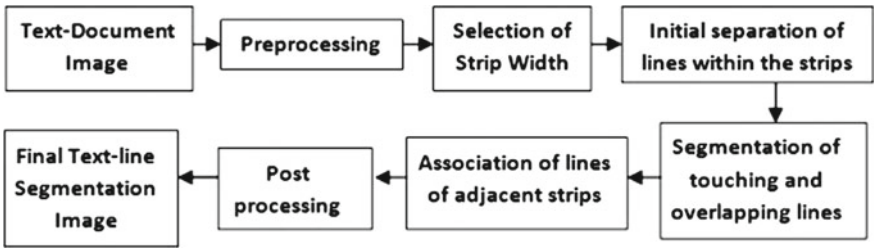


Fig. 1 Block diagram of proposed line segmentation method

3.1 Preprocessing

The input raw image is first subjected to a few preprocessing operations. These include mean filtering with window size 3×3 followed by binarization using a recently proposed method [17]. The minimum bounding rectangle of the binarized image is computed for further processing. The text portions of the processed image is black against white background.

3.2 Estimation of Strip Width

Since neither the words in a text line nor the consecutive text lines in a handwritten document are expected to be properly aligned, the binarized document is first vertically divided into several strips. The width of these strips are intelligently estimated so that it is not be too small or too large. The horizontal projection profile plays a major role in the proposed approach. If the width is too large, the separation between two consecutive lines inside a strip often may not be signalled by its horizontal projection profile. On the other hand, if the width is too small, the horizontal projection profile may frequently indicate false line breaks. Here we estimate the width of the vertical strips as follows.

Step 1: Divide the document into a few (here, 10) vertical strips of equal width.
Step 2: Compute horizontal projection profile of each strip.
Step 3: Identify the connected components of horizontal projection profile in each strip.
Step 4: Decide the horizontal segment bounded by the upper and lower boundaries of each such connected component as a text line inside the strip.
Step 5: Compute the average height (H_{avg}) of all such text lines in the input document.
Step 6: Similarly, compute the average height (LS_{avg}) of the gaps between two consecutive text lines.
Step 7: Obtain the Strip Width estimate $S_w = 3 * (H_{avg} + LS_{avg})$.

Next, we obtain certain rough estimation of text lines in each individual strip of width S_w as described in the following section.

3.3 Initial Separation of Lines Within the Strips

Horizontal projection profile of a document provides significant information for the line segmentation as long as the lines are free from skew, touching, overlapping etc. However, a handwritten documents rarely follow such an ideal structure. It contains touching or overlapping characters/words between consecutive lines. Such documents are often affected by skewed or curved lines, touching lines, vertically overlapping words etc. For efficient segmentation of text lines we divide the input document image into a number of vertical strips of width S_w and obtain a rough segmentation of text lines inside each strip. This is achieved by executing the following steps.

Step 1: Divide the input image into vertical strips of equal width S_w.
Step 2: Compute an image P which consists of horizontal projection profiles of all vertical strips of input document.
Step 3: Obtain connected components of P and compute the average height (H_{avg}) of these connected profile components.
Step 4: Initially, ignore all profile components of P whose height is less than $\frac{H_{avg}}{3}$.
Step 5: Obtain horizontal line segments at the top and bottom of the remaining profile components inside respective vertical strips.
Step 6: Thus, a rough segmentation of text lines is obtained.

In Fig. 2 portion illustrations of different stages of initial separation of text lines are shown. Next, we concentrate on associating text lines in one strip to its neighboring strip.

3.4 Segmentation of Touching and Overlapping Lines

The segmentation of touching and overlapping lines is described in a stepwise fashion as follows. Here, we scan all the vertical strips of thebibliography input document from left to right. We start at the leftmost strip and stop at the rightmost one. In each vertical strip, we scan from top to bottom.

Step 1: Consider the next strip and verify its initial segmented lines from top to bottom until there is no more strip.
Step 2: If the height of the next line in the current strip is less than $2H_{avg}$, then we accept it as a single line and move to the next line in the strip until we reach the bottom of the strip when we go to Step 1. If the height of a line exceeds the above threshold, we move to the next step (Step 3).

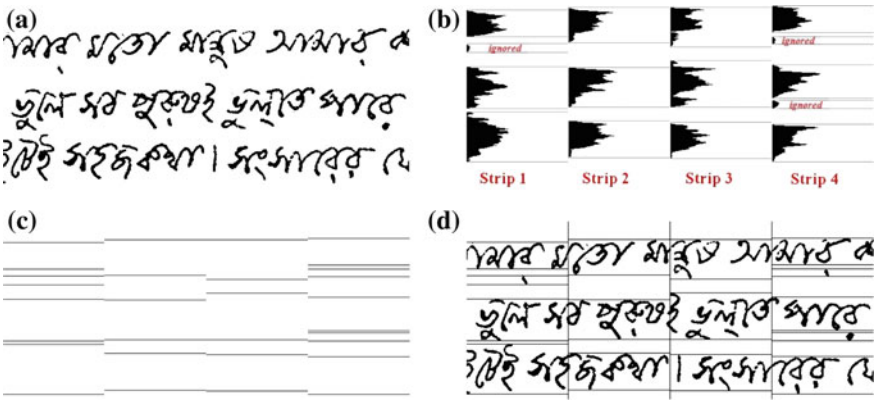


Fig. 2 Initial separation of text lines: **a** Part of a handwritten manuscript of poet Rabindranath Tagore, **b** *horizontal* line segments are drawn at the *top* and *bottom* of each profile component inside individual *vertical* strips, **c** estimated line segments barring the line of small height, **d** initial *vertical* strip-wise separation of text lines of the image shown in (**a**)

Step 3: Find the connected components in the current segmented line and if the height of all such components are less than ($2H_{avg}$), we move to Step 4. Otherwise, we decide that this component consists of touching characters of two vertically consecutive lines. We use the projection profile component of this line and find its minimum valley around the middle of the region. We segment the component at a point where the horizontal straight line segment through this valley intersects the component. As illustrated in Fig. 3 the initial line is now segmented into two lines above and below this horizontal straight line segment. Next, move to Step 2.
Step 4: It is a case of vertically overlapping lines and the leftmost strip of Fig. 3a shows an example. Here we find the valley region at the middle of the projection profile of the current segmented line. Usually, in similar situations, a small contiguous part of the profile can be easily identified as the valley instead of a single valley point. We consider the horizontal line through the middle of this valley region and the components, major parts of which lie above this horizontal line are considered to belong to the upper line and other components are considered to belong to the lower line. Figure 3c illustrates this and the segmentation result is shown in Fig. 3d. Next, move to Step 2.

3.5 Association of Lines of Adjacent Strips

The lines of individual strips have already been identified in Sect. 3.4. Now, it is required to associate the lines in a vertical strip with the corresponding lines of the adjacent strips, if any. Here, at any time we consider a pair of adjacent strips. We

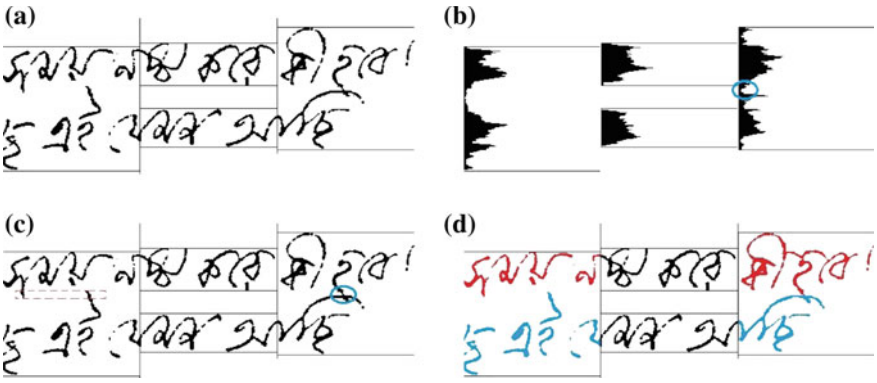


Fig. 3 Illustration of segmentation of *vertically* overlapping and touching lines: **a** Part of a handwritten manuscript; its *leftmost* and *rightmost* strips respectively contain vertically overlapping and touching lines, **b** the minimum valley around the *middle* region of the *horizontal* projection profile component corresponding to the touching line is identified and shown by a *blue* circle, **c** the overlapping region around the valley of its projection profile is shown by a dotted dark *red* colored rectangle and the segmentation site of the touching component is shown by a *blue* colored oval shape, **d** the initial line of each of the *leftmost* and *rightmost* strips is now segmented into two separate lines

start with the left most two strips and finish at the rightmost pair. The strategy used here is described below in a stepwise fashion.

Step 1: Set $i = 1$.
Step 2: Consider the pair of i-th and $(i + 1)$-th strips until there is no more strip.
Step 3: Consider the next line of $(i + 1)$-th strip. If there is no more line, increase i by 1 and go to Step 2, else move to the next Step.
Step 4: If the current line consists of no component which has a part belonging to a line of the i-th strip, then move to the next Step. Otherwise, associate the current line of $(i + 1)$-th strip with the line of i-th strip which accommodates a part of one of its components. If there are more than one such component common to both the strips and they belong to different lines of i-th strip, then we associate the present line with the line of the i-th strip corresponding to the larger component. Go to Step 3.
Step 5: We associate the current line with the line of the i-th strip having the maximum vertical overlap. If there is no such line in the i-th strip, then we look for similar overlap with a another strip at further left. If any such strip is found at the left, then the two lines are associated and otherwise, the current line is considered as a new line. Go to Step 3.

The above strategy of association of text lines in adjacent strips is further illustrated in Fig. 4 using part of the document shown in Fig. 2.

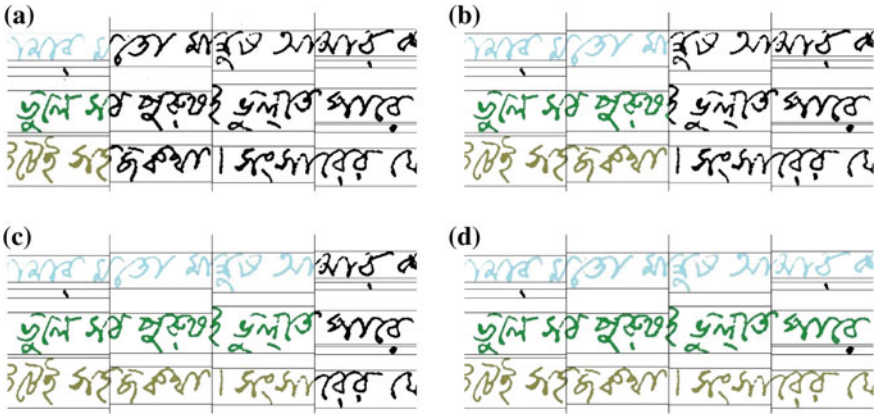


Fig. 4 Association of text lines of adjacent strips: **a** Initial segmentation into different lines of the *left* strip is shown in color, **b** lines of initial segmentation of the 2nd strip are associated with the lines of 1st strip, **c** lines of initial segmentation of the 3rd strip are associated with the lines of 2nd strip, **d** segmented lines of the 4th strip are associated with the lines of 3rd strip

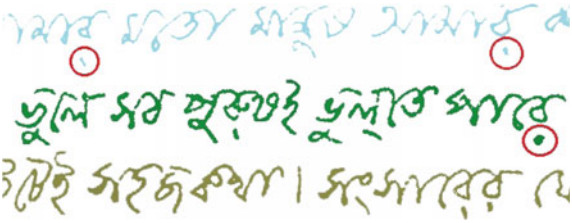


Fig. 5 Final segmentation result (after postprocessing) of the example of Fig. 2. These parts are highlighted by *red* circles

3.6 Postprocesing

During initial segmentation of text lines in individual strips described in Sect. 3.3, we ignore all text components with their profile height less than $\frac{H_{avg}}{3}$. Here, we consider the above components and associate them to the lines nearest to them. In the example shown in Fig. 2, there were 3 such small components which are now associated with their respective lines. The final segmentation result of this example is shown in Fig. 5.

4 Experimental Results

We simulated the proposed algorithm on the dataset of ICDAR 2013 Handwriting Segmentation Contest [6]. This data set contains 150 English and Greek (Latin Languages) and another 50 Bengali (Indian Language) handwritten document images. We used the evaluation tool provided by the ICDAR 2013 Line Segmentation Contest for comparative studies of the proposed algorithm with the methods participated

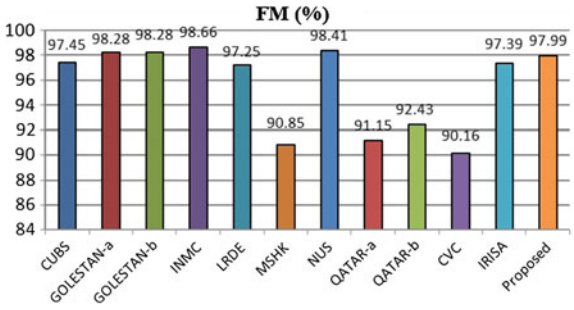


Fig. 6 Comparative performance evaluation result of the proposed method provided by the ICDAR 2013 line segmentation contest

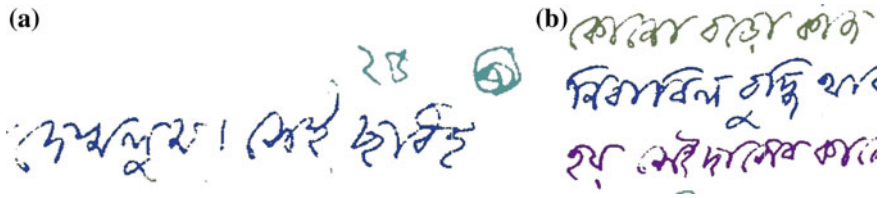


Fig. 7 Two peculiar scenarios where the algorithm worked perfectly

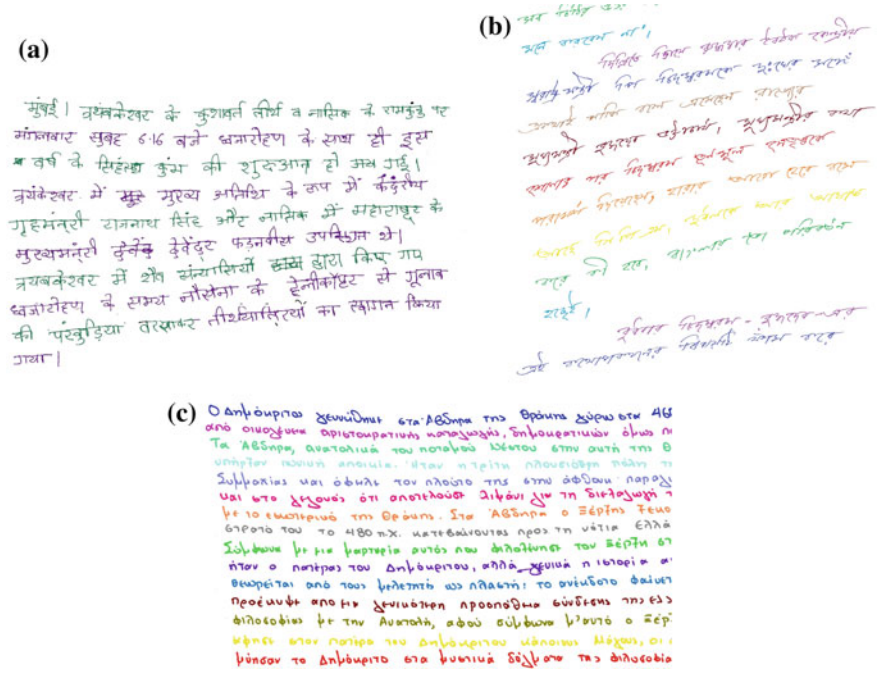


Fig. 8 A few line segmented handwritten documents of different scripts by the proposed algorithm

in this contest. The result of this comparison is shown in Fig. 6. The comparison is provided by the Performance Metric (PM) defined in terms of the measures Detection Rate (DR) and Recognition Accuracy (RA) as follows:

$$FM = \frac{2(DR \times RA)}{(DR + RA)}, DR = \frac{o2o}{N}, RA = \frac{o2o}{M},$$

where $o2o$ is the number of one-to-one matches between result image and ground truth, N and M are respectively the counts of ground truth result elements.

From Fig. 6, it can be seen that the accuracy of the proposed method on ICDAR 2013 dataset is 97.99 %. Examples of a few difficult situations where the proposed algorithm performed efficiently are shown in Fig. 7.

In Fig. 8, we show some more results of line segmentation on Devanagari, Bengali, Greek and English handwritten documents

5 Conclusions

In this article, we presented a novel method based on a simple strategy for line segmentation of handwritten documents of different Indian scripts. Its performance on "ICDAR 2013 Line Segmentation Contest" dataset is quite impressive. The method works equally efficiently on different types of scripts and can handle various peculiar situations of handwritten manuscripts. The only situation where we observed consistent failure of the present algorithm is the use of a caret to insert a line just top of another line of the input document. A few examples of such situations are shown in Fig. 9.

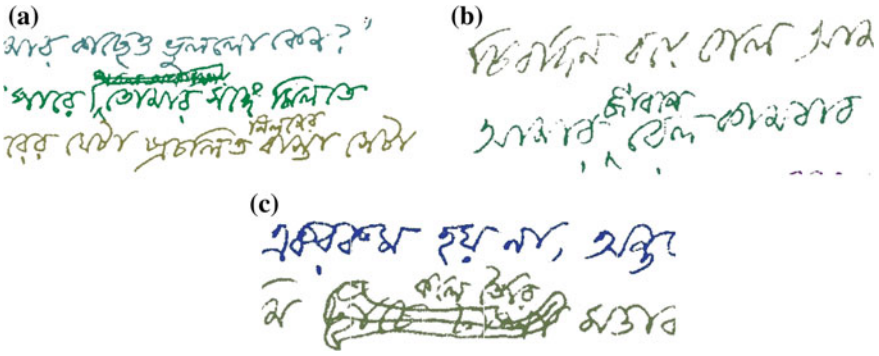


Fig. 9 A few failures of the proposed algorithm

References

1. Mullick, K., Banerjee, S., and Bhattecharya, U.: An Efficient Line Segmentation Approach for Handwritten Bangla Document Image. Eighth International Conference on Advences in pattern Recognition (ICAPR), 1–6 (2015)
2. Alaei, A., Pal, U., and Nagabhushan, P.: A New Scheme for Unconstrained Handwritten Text-Line Segmentation. Pattern Recognition. 44(4), 917–928, (2011)
3. Papavassiliou, V., Stafylakis, T., Katsouros, V., Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition. 147, 369–377 (2010)
4. Shi, Z., Seltur, S., and Govindaraju, V.: A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines. Proceedings of 10th International Conference on Document Analysis and Recognition, 176–180, (2009)
5. Louloudis, G., Gatos, B., and Halatsis, C: Text Line and Word Segmentation of Handwritten Documents. Pattern Recognition, 42(12):3169–3183, (2009)
6. Stamatopoulos, N., Gatos, B., Louloudis, G, Pal, U., Alaei, A.: ICDAR 2013 Handwritten Segmentation Contest. 12th International Conference on Document Analysis and Recognition, 14021–1406 (2013)
7. Likforman-Sulem, L., Zahour, A., and Taconet, B.: Text Line Segmentation of Historical Documents: a Survey. International Journal of Document Analysis and Recognition: 123–138, (2007)
8. Antonacopoulos, A., Karatzas, D.: Document Image analysis for World War II personal records, International Workshop on Document Image Analysis for Libraries. DIAL, 336–341 (2004)
9. Li, y., Zheng, Y., Doermann, D., and Jaeger, S.: A new algorithm for detecting text line in handwritten documents. International Workshop on Frontiers in Handwriting Recognition, 35–40 (2006)
10. Louloudis, G. Gatos, B., Pratikakis, I., Halatsis, K., Alaei, A.: A Block Based Hough Transform Mapping for Text Line Detection in Handwritten Documents. Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, 515–520 (2006)
11. Tsuruoka, S., Adachi, Y., and Yoshikawa, T.: Segmentation of a Text-Line for a Handwritten Unconstrained Document Using Thinning Algorithm, Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition:505–510, (2000)
12. Luthy, F., Varga, T., and Bunke, H.,: Using Hidden Markov Models as a Tool for Handwritten Text Line Segmentation. Ninth International Conference on Document Analysis and Recognition. 9, 630–632 (2007)
13. Lie, Y., Zheng, Y.: Script-Independent Text Line Segmentation in Freestyle Handwritten Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(8), 1313–1329 (2008)
14. Yin, F., Liu, C: A Variational Bayes Method for Handwritten Text Line Segmentation. International Conference on Document Analysis and Recognition. 10, 436–440 (2009)
15. Brodic, D., and Milivojevic, Z.: Text Line Segmentation by Adapted Water Flow Algorithm. Symposium on Neural Network Applications in Electrical Engineering. 10, 225–229 (2010)
16. Dinh, T. N., Park, J., Lee, G.: Voting Based Text Line Segmentation in Handwritten Document Images. International Conference on Computer and Information Technology. 10, 529–535 (2010)
17. Biswas, B., Bhattacharya, U., and Chaudhuri, B.B.: A Global-to-Local Approach to Binarization of Degraded Document Images. 22nd International Conference on Pattern Recognition, 3008–3013 (2014)

Palmprint Recognition Based on Minutiae Quadruplets

A. Tirupathi Rao, N. Pattabhi Ramaiah and C. Krishna Mohan

Abstract Palmprint recognition is a variant of fingerprint matching as both the systems share almost similar matching criteria and the minutiae feature extraction methods. However, there is a performance degradation with palmprint biometrics because of the failure of extracting genuine minutia points from the region of highly distorted ridge information with huge data. In this paper, we propose an efficient palmprint matching algorithm using nearest neighbor minutiae quadruplets. The representation of minutia points in the form of quadruplets improves the matching accuracy at nearest neighbors by discarding scope of the global matching on false minutia points. The proposed algorithm is evaluated on publicly available high resolution palmprint standard databases, namely, palmprint benchmark data sets (FVC ongoing) and Tsinghua palmprint database (THUPALMLAB). The experimental results demonstrate that the proposed palmprint matching algorithm achieves the state-of-the-art performance.

Keywords Palmprint recognition · k-Nearest neighbors · Minutiae and quadruplets

1 Introduction

Due to the growing demand of human identification for many ID services, biometrics has become more attracting research area. Fingerprint recognition system is more convenient and accurate. Palmprints can be considered as a variant of fingerprints which shares the similar feature extraction and matching methodology. Palm

A.T. Rao (✉) · N.P. Ramaiah · C.K. Mohan
Visual Learning and Intelligence Lab (VIGIL), Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, Hyderabad 502205, India
e-mail: tirupathi.avula@gmail.com

N.P. Ramaiah
e-mail: n.p.in@ieee.org

C.K. Mohan
e-mail: ckm@iith.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_11

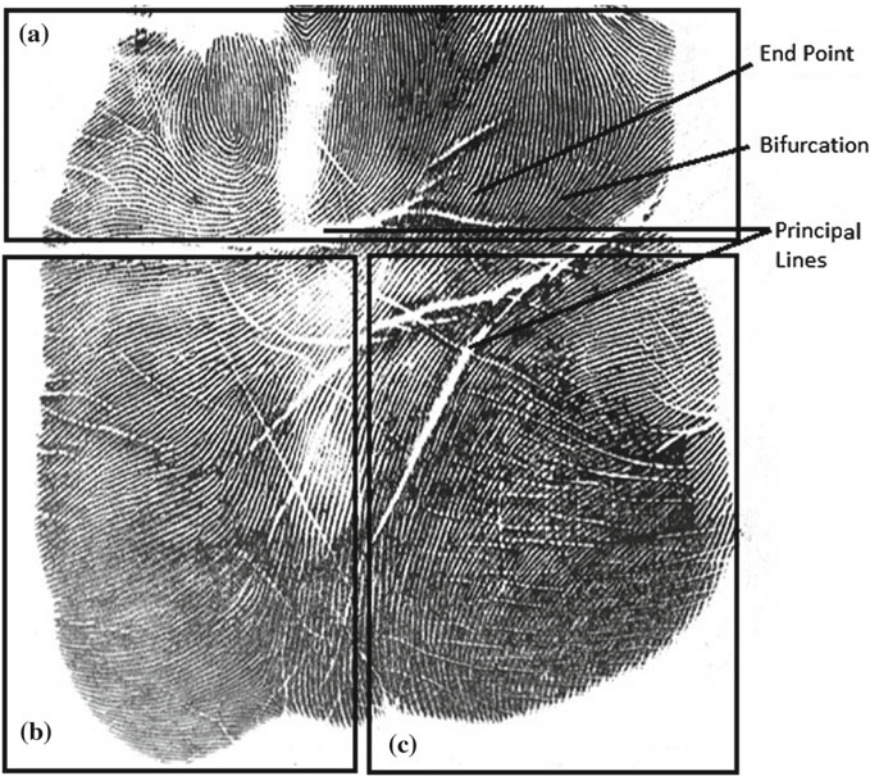


Fig. 1 Palm print image. **a** Interdigital, **b** hypothenar, **c** thenar

consists friction ridges and flexion creases as main features. Due to folding of the palm the flexion creases will be formed. The palmprint is having three regions namely hypothenar, thenar and interdigital (see Fig. 1).

In many instances, examination of hand prints like fingerprint and palmprint was the method of differentiating illiterate people from one another as they are not able to write. The first known automated palmprint identification system (APIS) [1] developed to support palmprint identification is built by a Hungarian company. There are mainly two different approaches in palmprint matching on high resolution palmprints, namely, minutiae based [2], ridge feature based [3]. Minutiae based palmprint matching methods find number of minutiae matches between the input palmprint (probe) and the enrolled palmprint (gallery). This is the most popular and widely used approach. In ridge feature-based palmprint matching, features of the palmprint ridge pattern like local ridge orientation, frequency and shape are extracted for comparison. These features may be more reliable for comparison in palmprint of low-quality images than minutiae features. The matching of the palmprint matching algorithm is correct when there are genuine matches (true accepts) and genuine rejects (true non-matches). The matching is wrong when there are impostor matches (false accepts) and impostor non matches (false rejects).

In this paper, a hybrid palmprint matching algorithm is proposed based on *k*-nearest neighbor and minutiae quadruplets. The rest of the paper is organized as follows: the related work is discussed in Sect. 2. In Sect. 3, palmprint feature extraction is presented. The proposed palmprint matching algorithm is presented using the representations of quadruplets in Sect. 4. Experimental setup and results for the proposed algorithm are discussed in Sect. 5. Conclusions are given in Sect. 6.

2 Related Work

Palmprint recognition research mainly concentrates on low-resolution palmprint images that are captured through low cost cameras [4–8]. These images are usually captured in a contactless way, the quality is very low. With such low quality, matching will be based on minor and major creases, as the ridges can not be observed. In [9, 10], researchers tried to explicitly extract and match major creases. In [11], Jain and Feng et al. proposed a palmprint recognition based on minutiae by segmenting the palmprint into multiple sub regions to achieve the acceptable accuracy. In [12], Dai and Zhou et al. proposed a multi-feature based palmprint recognition system, where multiple features including orientation field, density map, major creases and minutia points are extracted to get higher accuracy. There are few problems in large-scale palmprint applications. Few of the important problems are skin distortion, diversity of different palm regions and computational complexity.

The existing minutiae based palmprint techniques depend on segmentation to reduce the time complexity. The enrolled palmprints are divided into equal segments. As some segments in palmprints are very distinctive, it is possible to discard many non-mated reference palmprints by comparing the distinctive segments. In this paper, a minutiae based matching with out segmenting the palmprint is proposed to improve the accuracy of matching.

3 Feature Extraction

Palmprint feature extraction is challenging problem, in extracting of robust features. Palmprint image quality is low due to the wide creases (principal lines) present and more number of thin creases. The size of palmprint image is very large. Full fingerprint at 500 dpi is about 256 kB, where as a full palmprint at 500 dpi is about 4 megapixels. A palmprint feature extraction that is robust enough to deal with the average quality is not easy to design. The following are the steps involved in feature extraction of palmprint:

1. *Smooth*: Smoothing is a simple and frequently used image processing operation to reduce the noise of the image.

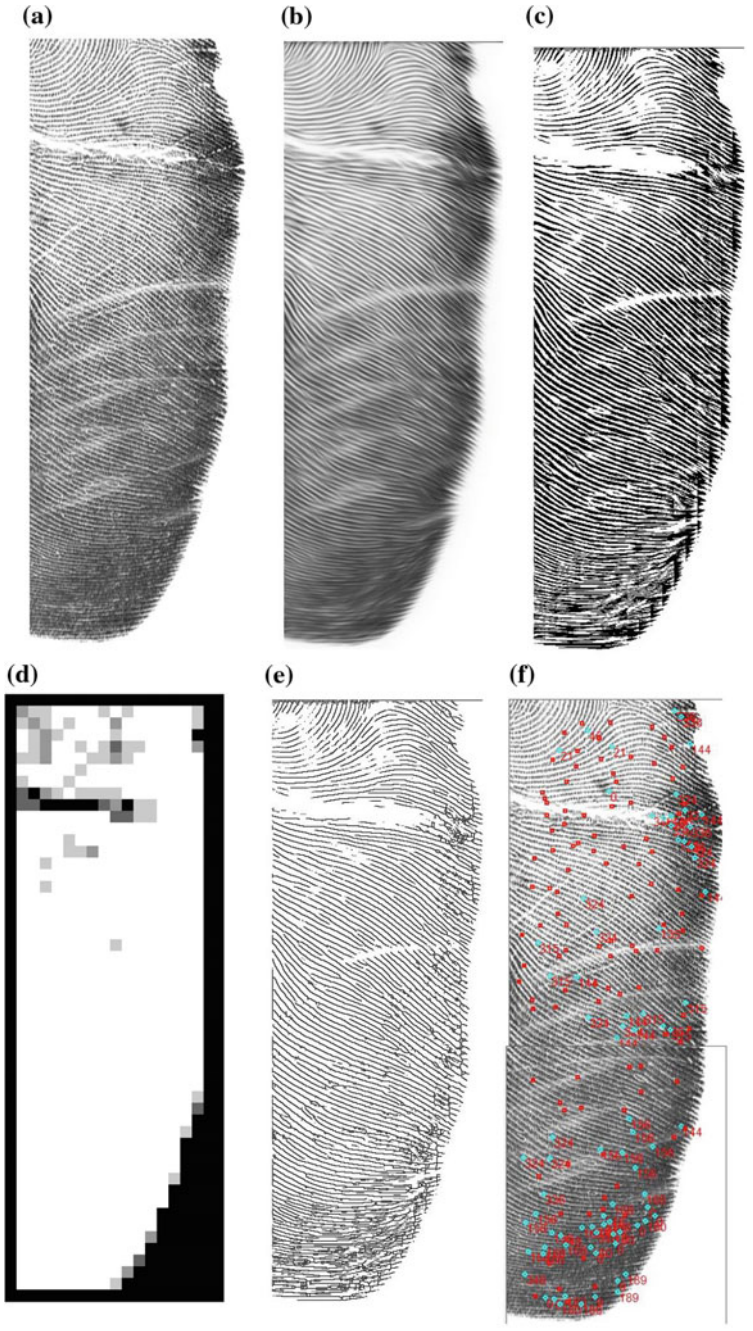


Fig. 2 Various stages of palmprint feature extraction: **a** original image, **b** smoothed image, **c** binarized image, **d** quality map of image, **e** thinned image, **f** minutiae interpolation

2. *Binarization*: In this step, the image is converted to complete black and white pixels from gray scale.
3. *Thinning*: Binarized image ridges are converted to one pixel thickness. Which will be useful for extracting of minutiae.
4. *Minutiae Extraction*: Minutiae extraction is done on thinned image. When traversing pixel by pixel on a thinned image where the pixel find one neighbor is End point and three neighbors is bifurcation point.
5. *Spurious Minutiae Removal*: This is final stage of feature extraction, where the spurious minutiae due to ridge cuts, border minutiae, bridges, lakes are removed.

The Fig. 2 shows the various phases involved in feature extraction and their corresponding output images of palmprint.

4 Proposed Palmprint Matching Algorithm

In this section, the proposed palmprint matching algorithm is explained. The quadruplet details are given first and then the *k*-nearest neighbor matching and global minutia matching using quadruplets is described.

4.1 Quadruplets

Let *A* be the set of palmprint minutiae and the *n*-quadruplets can be computed as follows: The *k*-nearest neighbors from the set *A* are computed for all $m \in A$ in order to find all *n*-quadruplets which have *m* and three of its nearest minutiae which is tolerant to the low quality. Figures 3 and 4 illustrate the sample quadruplet representation of minutiae points. In Fig. 3, *ab*, *bc*, *cd*, *ad*, *bd* and *ac* are euclidean distances between each pair of minutiae. Each minutia point has mainly 3 characteristics *x*, *y*, *Direction*. Figure 4 illustrates each minutiae pair features for matching, *ab* is the Euclidean distance, *b* is direction at *B*, *a* is direction at minutiae *A*.

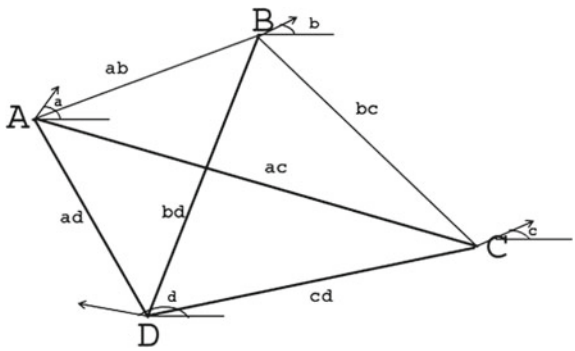


Fig. 3 Quadruplet representation of minutiae

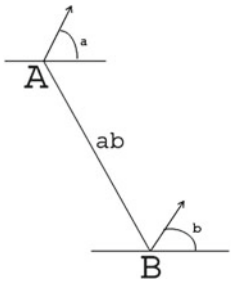


Fig. 4 Characteristics of minutiae pair

4.2 k-Nearest Neighbor Matching Algorithm

This step finds the similar mates from query template and probe template using *k*-nearest neighbor local minutiae matching techniques. G and P are the palmprint minutiae feature vectors. The proposed Minutiae-based method considers mainly three features from each minutia m = *x*, *y*, θ, where *x*, *y* is location θ is direction. Let $G = m_1, m_2, \ldots m_m, m_i = x_i, y_i, \theta_i, i = 1, \ldots, m$ and $P = m_1, m_2, \ldots m_n, m_i = x_i, y_i, \theta_i, i = 1, \ldots, n$, where *m* and *n* denote the number of minutiae in gallery and probe template respectively. Equations (1) and (2) denote the Euclidean distance and angle of minutiae a and b, respectively.

$$Dist_{ab} = \sqrt{(X_a - X_b)^2 - (Y_a - Y_b)^2} \tag{1}$$

$$Dir_{ab} = \arctan \frac{(Y_a - Y_b)}{(X_a - X_b)} \tag{2}$$

A minutia m_j in P and a minutia m_i in G are considered to be mate, when the $k/2$ number of minutia in k-nearest neighbors (*KNN*) are similar using the Eq. (3).

$$\sum_{k=1}^{KNN} P_i, \quad \sum_{l=1}^{KNN} G_j \tag{3}$$

$$Dist_p^{ik} - Dist_G^{jl} < DistThr, \quad Dir_p^{ik} - Dir_G^{jl} < DirThr$$

4.3 Computing Match Score Using Quadruplets

This step considers short listed minutiae from *k*-nearest neighbor stage. Each minutiae pair as reference pair for finding quadruplet. The following three conditions should be considered to determine whether the two minutia in quadruplet are matched in order to overcome the tolerance to distortions and rotations. In order to qualify a

quadruplet as mate, the four edges of each quadruplet should satisfy following three conditions:

1. The Euclidean distance between two minutiae < *DistThr*.
2. The difference between minutia directions < *DirThr*.
3. Minutiae relative direction with edge < *RelThr*.

5 Experimental Setup and Results

The experiments are conducted on the standard palmprint benchmark data sets FVC ongoing competition test data [13] and Tsinghua university [12, 14, 15]. The test data consists of 10 people, 1 palm of 8 instances. Tsinghua university data set consists of 80, people 2 palms of 8 instances. These experiments were carried out on Intel core i3 machine with 4 GB ram and 1.70 GHz processor. Figures 5 and 6 show the ROC curves over the standard databases FVC Ongoing and Tsinghua THUPALMLAB data sets.

Table 1 shows that the databases with number of persons, genuine and impostor attempts. Table 2 shows that the standard databases, nearest neighbors considered and Equal Error Rate (EER). The accuracy of the algorithm is good when nearest

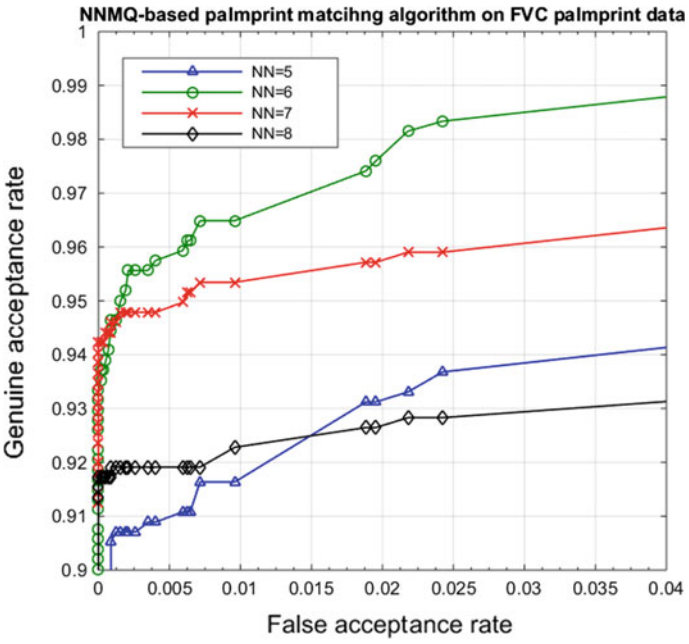


Fig. 5 ROC on standard FVC test data

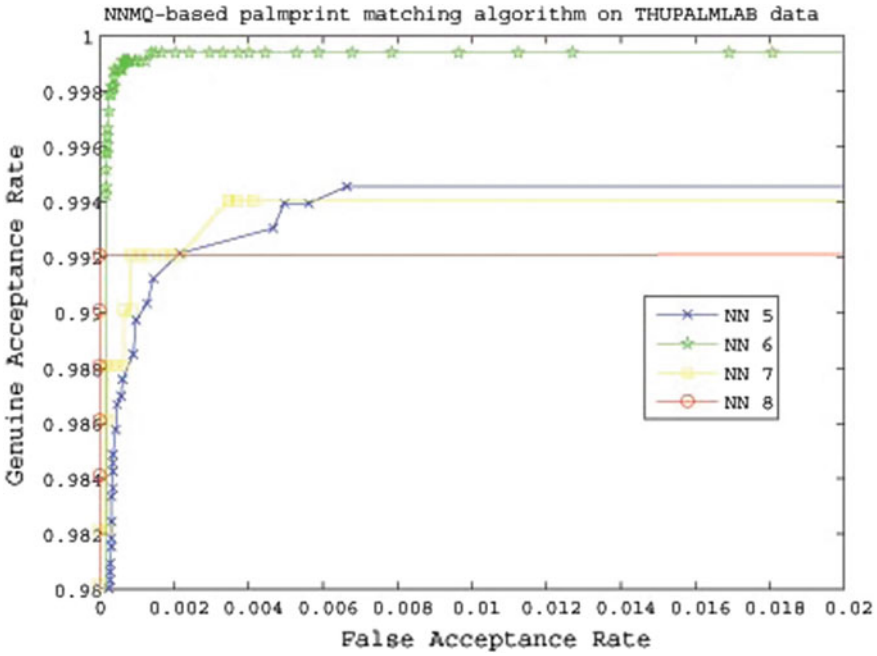


Fig. 6 ROC on Tsinghua THUPALMLAB data

Table 1 Databases

Data set	No of persons	Instances	Genuine	Impostor
FVC ongoing	10	8	280	2880
THUPALMLAB	80	8	4480	20000

Table 2 EER on FVC and THUPALMLAB data sets

# of NNs	EER % on FVC ongoing	EER % on THUPALMLAB
5	4.56	0.56
6	3.87	0.12
7	4.30	0.69
8	5.08	0.83

neighbors considered 6 with the two databases. The *DistThr* is 12, *DirDiff* is 30 and *RelDiff* is 30 considered in all the experiments. Table 3 shows that the standard databases, nearest neighbors considered, space and time taken for each verification.

The proposed algorithm achieved 0.12 % of EER on THUPALMLAB data set where as, the EER of [11, 12] on THUPALMLAB data set are 4.8 and 7 % respectively.

Table 3 Space and times taken on FVC and THUPALMLAB data sets

# of NNs	Space (Kb) FVC data	Time (ms)	Space (Kb) THUPALAMLAB	Time (ms)
5	32524	2089	32480	2535
6	32572	2347	32564	4017
7	32488	2388	32512	4435
8	32536	2711	32504	5154

6 Conclusion

The existing minutiae based matching algorithms have few limitations which are mainly based on segmentation. The accuracy of these algorithms is affected with different qualities of the palmprint regions. The proposed palmprint matching algorithm used the new representation of minutiae points using quadruplets and the matching is done with out segmenting the palmprint. The experiments have proved that the proposed matching algorithm achieves very good accuracy over existing standard data sets. The proposed algorithm on FVC palm test data have achieved EER 3.87 % and on THUPALMLAB data set achieved EER of 0.12 %.

Acknowledgements We are sincerely thankful to FVC and Tsinghua university for providing data sets for research. The first author is thankful to Technobrain India Pvt Limited, for providing support in his research.

References

1. FBI: https://www.fbi.gov/about-us/cjis/fingerprints_biometrics/biometric-center-of-excellence/files/palm-print-recognition.pdf
2. Liu N, Yin Y, Zhang H: Fingerprint Matching Algorithm Based On Delauny Triangulation Net. In: Proc. of the 5th International Conference on Computer and information Technology, 591–595 (2005)
3. Jain A, Chen Y, Demirkus M: Pores and Ridges: Fingerprint Matching Using level 3 features. In Proc. of 18th International Conference on Pattern Recognition (ICPR'06), 477–480 (2006)
4. Awate, I. and Dixit, B.A.: Palm Print Based Person Identification. In Proc. of Computing Communication Control and Automation (ICCUBEA), 781–785 (2015)
5. Ito, K. and Sato, T. and Aoyama, S. and Sakai, S. and Yusa, S. and Aoki, T.: Palm region extraction for contactless palmprint recognition. In Proc. of Biometrics (ICB), 334–340 (2015)
6. George, A. and Karthick, G. and Harikumar, R.: An Efficient System for Palm Print Recognition Using Ridges. In Proc. of Intelligent Computing Applications (ICICA), 249–253 (2014)
7. D. Zhang, W.K. Kong, J. You, and M. Wong: Online Palmprint Identification. IEEE Trans. Pattern Analysis and Machine Intelligence 25(9), 1041–1050 (2003)
8. W. Li, D. Zhang, and Z. Xu: Palmprint Identification by Fourier Transform. Pattern Recognition and Artificial Intelligence 16(4), 417–432 (2002)
9. J. You, W. Li, and D. Zhang: Hierarchical Palmprint Identification via Multiple Feature Extraction. Pattern Recognition 35(4), 847–859 (2002)

10. N. Duta, A.K. Jain, and K. Mardia: Matching of Palmprints. Pattern Recognition Letters 23(4), 477–486 (2002)
11. A.K. Jain and J. Feng: Latent Palmprint Matching. IEEE Trans. Pattern Analysis and Machine Intelligence 31(6), 1032–1047 (2009)
12. J. Dai and J. Zhou: Multifeature-Based High-Resolution Palmprint Recognition. IEEE Trans. Pattern Analysis and Machine Intelligence 33(5), 945–957 (2011)
13. B. Dorizzi, R. Cappelli, M. Ferrara, D. Maio, D. Maltoni, N. Houmani, S. Garcia-Salicetti and A. Mayoue: Fingerprint and On-Line Signature Verification Competitions at ICB 2009. In Proc. of International Conference on Biometrics (ICB), 725–732 (2009)
14. THUPALMLAB palmprint database. http://ivg.au.tsinghua.edu.cn/index.php?n=Data.Tsinghua500ppi
15. Dai, Jifeng and Feng, Jianjiang and Zhou, Jie: Robust and efficient ridge-based palmprint matching. IEEE Trans. Pattern Analysis and Machine Intelligence 34(8), 1618–1632 (2012)

Human Action Recognition for Depth Cameras via Dynamic Frame Warping

Kartik Gupta and Arnav Bhavsar

Abstract Human action recognition using depth cameras is an important and challenging task which can involve highly similar motions in different actions. In addition, another factor which makes the problem difficult, is the large amount of intra class variations within the same action class. In this paper, we explore a Dynamic Frame Warping framework as an extension to the Dynamic Time Warping framework from the RGB domain, to address the action recognition with depth cameras. We employ intuitively relevant skeleton joints based features from the depth stream data generated using Microsoft Kinect. We show that the proposed approach is able to generate better accuracy for cross-subject evaluation compared to state-of-the-art works even on complex actions as well as simpler actions but which are similar to each other.

Keywords Human action recognition · Depth-camera · Skeleton information · Dynamic frame · Warping · Class templates

1 Introduction

The problem of human action recognition is an important but challenging one, and has applications in various domains such as automated driving systems, video retrieval, video surveillance (for security purposes), elderly care, and human-robot interactions. Traditionally, research in action recognition was based on video sequences from RGB cameras or complex motion capture data. Despite many research efforts and many encouraging advances, achieving good accuracies in recognition of the human actions, is still quite challenging. With the recent advent

K. Gupta (✉) · A. Bhavsar
School of Computing & Electrical Engineering,
Indian Institute of Technology Mandi, Mandi, Himachal Pradesh, India
e-mail: kk1153@gmail.com; kartik_gupta@students.iitmandi.ac.in

A. Bhavsar
e-mail: arnav@iitmandi.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_12

of low-cost depth sensors such as Microsoft Kinect, the advantages of depth sensors (such as illumination and color invariance), have been realized to better understand and address the problems such as gesture recognition, action recognition, object recognition etc.

In general, the problem of action recognition using depth video sequences involves two significant questions. The first is about effective representation of RGBD data, so as to extract useful information from RGBD videos of complex actions. The second question concerns developing approaches to model and recognize the actions represented by the suitable feature representation.

For the video representation, we use an existing approach of skeleton joints representation, that of Eigen Joints [1]. The advantage of using this is that most of the existing works are mainly on video level features but with Eigen Joints feature representation, we are able to work with frame level features which provides us more information and flexibility to work with.

Unlike the traditional RGB camera based approaches, the classification algorithm for the depth stream should be robust enough to work without even huge amount of training data, and handle large intra-class variations. In this respect, we explore a recently proposed work on Dynamic frame warping framework for RGB based action recognition [2], for the task of depth based action recognition. This framework is an extension to Dynamic time warping framework to handle the large amount of intra class variations which cannot be captured by normal Dynamic time warping algorithm. Unlike in [2], we do not use RGB features, but the skeleton joint features mentioned above.

With the best of our knowledge, such a dynamic frame warping framework on depth data has not been attempted till now. Such an adaptation from the technique proposed in [2], for action recognition in depth videos, is not obvious as depth data for action recognition brings with it its own challenges.

For instance, skeleton features involved in RGBD sequences are often inaccurate with respect to the joint positions and involve some amount of noise. Furthermore, the complexity in the actions is further enhanced in the case of 3D action recognition as it involves more information available for a single frame which is needed to be captured by a good classifier. Also, some actions in depth representations are typically more similar to each other which makes the problem harder. With more subjects performing same action in different environments in some different ways, it becomes evidently important to come up with a more robust technique to deal with high intra-class variations. Our experiments involve different subsets of data, which highlight the above mentioned cases of complex actions and similar actions, and demonstrate superior performance of the proposed approach over the state-of-the-art.

We also note that, in conjunction with frame-level features, such a framework has another advantage over discriminative models like Support Vector Machines (SVM). It can be further extended as a dynamic programming framework, which can work for continuous action recognition rather than isolated action recognition. Continuous action recognition involves unknown number of actions being performed with unknown transition boundaries in a single video sequence. (While, in this work, we

do not consider the problem of continuous action recognition, we believe that it is interesting to note this aspect of the proposed framework).

We propose that this approach enables us to develop such a robust depth based action recognition system which can capture articulated human motion and can distinguish between actions with similar motions using skeleton joints representation without a large corpus of training examples. Our results clearly state that the proposed approach clearly outperforms some of the existing methods for the cross subject tests done on MSR-Action3D dataset [3] consisting of both complex actions, and actions with similar motion. A good performance in such tests are important requirement for real world applications.

1.1 Related Work

With the advent of real-time depth cameras, and availability of depth video datasets, there is now a considerable work on the problem of human action recognition from RGBD images or from 3D positions (such as skeleton joints) on the human body.

Li et al. [3] developed a bag of words model using 3D points for the purpose of human action recognition using RGBD data. They used a set of 3D points from the human body to represent the posture information of human in each frame. The evaluation of their approach on the benchmark MSR-Action3D dataset [3] shows that it outperforms some state of the art methods. However, because the approach involves a large amount of 3D data, it is computationally intensive.

Xia et al. [4] proposed a novel Histogram of 3D Joint Locations (HOJ3D) representation. The authors use spherical coordinate system to represent each skeleton and thus also achieve view-invariance, and employ Hidden Markov models (HMMs) for classification.

In the work, reported in [1], the authors proposed an Eigen Joints features representation. which involves pairwise differences of skeleton joints. Their skeleton representation consists of static posture of the skeleton, motion property of the skeleton, and offset features with respect to neutral pose in each frame. They use a Naive Bayes classifier to compute video to class distance. An important advantage with this representation is that it involves frame level features which not only captures temporal information better but also has an adaptability to continuous action recognition framework. Moreover, these features are also simple and efficient in their computation.

The approaches reported in [5–7] also have been shown to perform well on the MSR Action 3D dataset. However, these works use video level features instead of frame level features as we use in our work. We reiterate that with frame level features, this work can be extended to a continuous action recognition module, which is difficult with video level features.

In [2], the dynamic frame warping (DFW) framework was proposed to solve the problem of continuous action recognition using RGB videos. Like the traditional DTW, this framework has the ability to align varying length temporal sequences.

Moreover, an important advantage of this approach over DTW is that it can better capture intra-class variations, and as a result, is more robust.

Our proposed approach also uses the Eigen Joint feature representation, but in a modified Dynamic time warping framework as proposed in [2]. The major advantage with such a dynamic programming framework is it can work with frame level features, so it can arguably understand the temporal sequences of frames better than Naive Bayes nearest neighbour classifier such as in [1]. In addition, we demonstrate that it is able to work without a large amount of training data required as in case of HMM (such as in [4]), as also indicated in [2].

Our paper is divided into subsequent sections. In Sect. 2 we explain our approach in depth, describing the Eigen Joints representation technique and the DFW algorithm. We show the experimental evaluations and their comparisons, with consideration in Sect. 3. We provide our conclusions in Sect. 4.

2 Proposed Approach

We now elaborate on our proposed approach in the following two subsections. We first discuss the representation of 3D video using Eigen Joints, followed by the modified DTW approach for recognition.

2.1 3D Video Representation

As mentioned earlier, we employ the Eigen Joints features [1] which are based on the differences of skeleton joints. The overall Eigen Joints feature characterizes three types of information in the frames of an action sequence, including static posture, motion property, and overall dynamics.

The three dimensional coordinates of 20 joints can be generated using human skeletal estimation algorithm proposed in [8], for all frames: $X = \{x_1, x_2, \ldots, x_{20}\}$, $X \in \mathfrak{R}^{3\times 20}$. Based on the skeletal joint information three types of pair-wise features are computed.

Differences between skeleton joints for the current frame: These features capture the posture of skeleton joints within a frame:
$f_{cc} = \{x_i - x_j | i, j = 1, 2, \ldots, 20; i \neq j\}$.

Skeleton joint differences between the current frame-c and its previous frame-p: These features take into the account the motion from previous to the current frame:
$f_{cp} = \{x_i^c - x_j^p | x_i^c \in X_c; x_j^p \in X_p\}$.

Skeleton joint differences between frame-c and frame-i (initial frame which contains neutral posture of the joints): These features capture the offset of an intermediate posture with respect to a neutral one:
$f_{ci} = \{x_i^c - x_j^i | x_i^c \in X_c; x_j^i \in X_i\}$.

The concatenation of the above mentioned feature channels forms the final feature representation for each frame: $f_c = [f_{cc}, f_{cp}, f_{ci}]$. Feature rescaling is used to scale the feature in the range $[-1, +1]$ to deal with the inconsistency in the coordinates. In each frame, 20 joints are used which result in huge feature dimension i.e. $(190 + 400 + 400) \times 3 = 2970$ as these differences are along three coordinates after feature rescaling, which gives us f_{norm}. Finally, PCA is applied over the feature vectors reduce redundancy and noise from f_{norm} where we use leading 128 eigen vectors to reduce the dimensionality.

Such a feature representation on the depth videos is much more robust (in terms of invariances) than ordinary color based features on RGB counterparts of such videos, and also provide structural information in addition to Spatio-temporal interest points.

2.2 *Machine Modeling of Human Actions via the DFW Framework*

Rabiner and Juang [9], Mueller [10] proposed Dynamic time warping (DTW) framework to align two temporal sequences $P_{1:T_P}$ and $Q_{1:T_Q}$ of unequal lengths. In this algorithm, the frame-to-frame assignments helps to match two temporal sequences:

$$A(P, Q) = \{(l_1, l'_1), (l_i, l'_i), \ldots, (l_{|A|}, l'_{|A|})\} \tag{1}$$

where $1 \leq l_i \leq T_P$ and $1 \leq l'_i \leq T_Q$ are indices of the frames of P and Q sequences, respectively.

The DTW algorithm finds the best alignment possible between the two temporal sequences P and Q. Each match between the elements of P and elements of Q gives a distance between that match while finding the best alignment. By passing through the best alignment path from $(1, 1)$ to (T_P, T_Q), the matched distances are accumulated to come up with an overall DTW distance between the two temporal sequences as done in Eq. (2).

$$\text{DTW}(P, Q) = \frac{1}{|A|} \sum_{i=1}^{|A|} d(p_{l_i}, q_{l'_i}). \tag{2}$$

As a variant to the traditional DTW algorithm, Dynamic frame warping i.e. DFW framework was introduced in [2]. This concept of DFW involves two main components: Action template represented by Y^l and Class template represented by $\tilde{Y}^l$ for each action class l. Here, the closest match to all the training samples of class l is found, $X^l_{i^*} \in \{X^l_n\}_{n=1}^{N_l}$. The closest match for each class l is defined as the action template of class l. Solving minimization in (3), yields the index of the sequence which is selected as the action template of each class:

$$i^* = \underset{i}{\operatorname{argmin}} \sum_{j \neq i} \mathrm{DTW}\left(X_i^l, X_j^l\right) \tag{3}$$

Finally, denoting the action template of class l as Y^l, each training example X_j^l is aligned with Y^l using the above Eq. (3). This provides the class template:

$$\tilde{Y}^l = \left(\tilde{y}_1^l, \ldots \tilde{y}_{t'}^l, \ldots, \tilde{y}_{T_{Y^l}}^l\right), \tag{4}$$

with the length equal to length of Y^l, which constitutes of a sequence of metaframes. Each metaframe $\tilde{y}_{t'}^l$ is set of frames from the training sequences, which show a closest match with a corresponding frame of Y^l.

Having computed the class template and action template, using the training dataset, the problem of recognition is as follows: Given a test sequence, find the distance between the sequence of test frames and sequence of metaframes, $\tilde{Y}^l$. The frame-to-frame distance $d(x_t, y_{t'})$ which was considered in the DTW distance in Eqs. (2) and (3), is now replaced by a new frame-to-metaframe distance measure $\tilde{d}(z_t, \tilde{y}_{t'}^l)$.

Now, a test frame $z_t \in \mathbb{R}^K$ can be considered as a linear combination of the training frames which are elements of a metaframe, $\tilde{y}_{t'}^l$. As, typically, only a small number of training frames within a metaframe, would be similar to the test frame, a sparse solution is computed for the weights in the linear combination $w_{t'}$, by solving the following Eq. (5) from [11],

$$\overline{w}_{t'} = \underset{w}{\operatorname{argmin}} \|z_t - \tilde{y}_{t'}^l w\|_2 + \gamma \|w\|_1. \tag{5}$$

Finally, the frame-to-metaframe distance has been expressed in following Eq. (6), reproduced from [2]. This can be solved using [12].

$$\tilde{d}\left(z_t, \tilde{y}_{t'}^l\right) = \min_w \left\| z_t - \frac{\tilde{y}_{t'}^l w}{\|\tilde{y}_{t'}^l w\|_2} \right\|_2^2 s.t. \sum_{i=1}^{|S|} w_i = 1. \tag{6}$$

We reiterate that the existing work using Dynamic frame warping framework on human action recognition is only on the RGB videos with color based features. Our approach extends this work to depth sequences with arguably more robust features which provide structural information apart from the spatio-temporal interest points. As indicated above, we believe that such an exploration of the Dynamic frame warping framework with skeleton joint features is important from the aspects of complex actions, large intra-class variability (actions performed by different subject or in different environments) and performance under noisy joint features, which are more specific to depth videos.

3 Experiments and Results

We evaluate our work on a benchmark MSR-Action3D dataset [3] with cross subject evaluation settings as used in previous works. We show detailed comparisons of our work with the existing Eigen Joints approach [1], Bag of 3D Points [3] and HOJ3D [4]. Figure 1 shows some of the frames of different actions from the MSR-Action3D dataset.

3.1 MSR-Action3D Dataset

MSR-Action3D Dataset is captured by depth sensor, Microsoft Kinect and is action recognition dataset. This dataset comprises of 20 human actions. Each action was performed by ten subjects for two or three times. The dataset in total contains 567 action sequences. Out of these 10 action sequences have been stated as too noisy either because there is no skeleton joints tracked or they have very noisy predictions. The frame rate is 15 frames per second and resolution is 640×480. The twenty actions are relevant mainly as gaming actions. Background has been already cleaned in this dataset. 20 skeleton joints positions are already known for each frame of the action sequences courtesy Shotton et al. [8].

3.2 Experimental Settings

We evaluate proposed approach on the cross subject evaluation as used in other works to make comparisons more accurate. So, we use actions sequences of 5 subjects i.e. 1, 3, 5, 7 and 9 for training of each action and rest 5 i.e. 2, 4, 6, 8 and 10 for testing. These tests are done separately on three subsets of the total 20 actions as listed in the table depicted in Fig. 2.

Fig. 1 Frames depicting MSR-Action3D dataset out of the 20 actions performed (Reproduced from Li et. al. [3])

Subset 1	Subset 2	Subset 3
Horizontal wave (HoW)	High wave (HiW)	High throw (HT)
Hammer (H)	Hand Catch (HC)	Forward Kick (FK)
Forward punch (FP)	Draw X (DX)	Side kick (SK)
High throw (HT)	Draw tick (DT)	Jogging (J)
Hand clap (HC)	Draw circle (DC)	Tennis swing (TSw)
Bend (B)	Hands wave (HW)	Tennis serve (TSr)
Tennis serve (TSr)	Forward kick (FK)	Golf swing (GS)
Pickup throw (PT)	Side boxing (SB)	Pickup throw (PT)

Fig. 2 Partitioning of the MSR-Action3D dataset as done in the previous works

These three subsets have been made exactly as they have been used in the previous works to reduce computational complexity. Another important reason for this partitioning is based on the fact that Subset 3 contains very complex but distinctive actions in terms of human motion where as Subset 1 and 2 contains simpler actions but with quite similar human motion (i.e. with more overlap). More specifically, a complex action (in Subset 3) is an action which is a combination of multiple actions (e.g. *bend* and *throw* constitute *pickup and throw*) and also actions like *jogging* involving periodic repetitive motion.

3.3 Comparison with Existing Approaches

The recognition rates of the proposed approach for each subset of the MSR-Action3D dataset is illustrated by means of three confusion matrices for respective subset in Fig. 3. It is clearly visible that the proposed approach works very well on the complex actions of Subset 3 as the joint motion is quite different in all the actions in this subset. The recognition accuracy is relatively low in Subset 1 and 2 as they mainly comprise of action which have quite similar motions. Another reason for somewhat reduced recognition rate in Subset 1 and 2, is that they contain sequences with noisier skeleton information in training as well as testing sets.

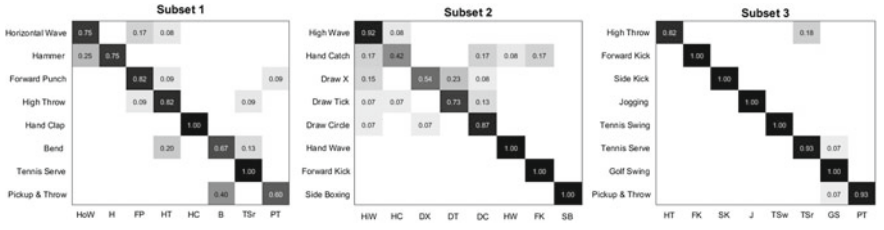


Fig. 3 Confusion matrix of our proposed approach in different subsets under cross subject evaluation settings. Each element of the matrix gives the recognition results

Table 1 Depicting recognition rates in %age, of our approach in comparison to the state of the art techniques for all the subsets of MSR-Action3D dataset

	3D Silhouettes [3]	HOJ3D [4]	EigenJoints [1]	Ours
Subset 1	72.9	87.98	74.5	80.18
Subset 2	71.9	85.48	76.1	82.14
Subset 3	79.2	63.46	96.4	96.40

Having discussed our absolute performance, the relative comparisons with state-of-the-art approaches show very encouraging results, thus highlighting the efficacy of the proposed approach. The comparison with 3D Silhouettes, HOJ3D and Eigen Joints for the cross-subject evaluation for the respective subsets of MSR-Action3D dataset is as listed in Table 1. Clearly, for all the subsets of data the proposed approach outperforms 3D Silhouettes and Eigen Joints (except in subset 3 where Eigen Joints performs similar to ours). This highlights that our frame level features within the dynamic frame warping framework is able to handle inter and intra-class variations better. For subset 3, which consists of complex actions, our comparative performance is very high as compared to the HMM based classification of HOJ3D. Note that the HMM based classification [4] also uses skeleton-joint based features. Considering this, along with the difference in performance for Subset 3, it is apparent that the training data is not sufficient for the HMM based approach in case of complex actions. Thus, the result for Subset 3 clearly indicates our approach can perform well even with less number of training samples.

Finally, Table 2 compares the overall performance of proposed approach with various state-of-the-art human action recognition approaches, and clearly shows that as a whole, the proposed approach gives better results than all the other existing approaches with simpler feature representation. This indicates that the proposed approach can better handle the trade-off between interclass variation, intraclass variation and noisy features.

Table 2 Comparisons of recognition rates in %age of our approach to some of state of the art techniques on the MSR-Action3D dataset with cross subject evaluation (5 subjects for training and 5 subjects for testing) settings

Method	Accuracy
DTW [10]	54
HMM [15]	63
3D Silhouettes [3]	74.67
HOJ3D [4]	78.97
HOG3D [14]	82.78
EigenJoints [1]	83.3
HDG [13]	83.70
Ours	86.24

4 Conclusion and Future Work

We present a modified Dynamic time warping framework known as Dynamic frame warping by Kulkarni et al. [2] for the 3D action recognition which is more robust to intra class variations, and can perform well even with low training data. The framework employs straightforward feature representation based on skeleton joints which can incorporate spatio-temporal information. It is clearly evident from the action recognition results on the MSR-Action3D dataset, that our approach can outperform state of the art techniques. The results also suggest that this framework is able to better capture the motion information of skeleton joints which is necessary to discriminate between similar actions, and can model complex actions well. This framework works well even with the frame level features, and has an important advantage with a possibility of extension for continuous action recognition.

References

1. X. Yang, & Y. Tian. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1), 2014, pp. 2–11.
2. K. Kulkarni, G. Evangelidis, J. Cech, & R. Horaud. Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 112(1), 2015, pp. 90–114.
3. W. Li, Z. Zhang, & , Z. Liu. Action recognition based on a bag of 3d points. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010)*, 2010, pp. 9–14.
4. L. Xia, C. C. Chen, & J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2012)*, 2012, pp. 20–27.
5. J. Wang, Z. Liu, Y. Wu, & J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 2012. pp. 1290–1297.
6. O. Oreifej, & Z. Liu. HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, 2013. pp. 716–723.
7. C. Chen, K. Liu & N. Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*. 2013. pp.1–9.
8. J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, & R. Moore. Real-time human pose recognition in parts from single depth images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, 2011, pp. 116–124.
9. L. Rabiner, & B. H. Juang. *Fundamentals of speech recognition*. Salt Lake: Prentice hall 1993.
10. M. Mueller. Dynamic time warping. *Information retrieval for music and motion*, Berlin: Springer 2007, pp. 6984.
11. S. S. Chen, D. L. Donoho, & M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1), 1998, pp. 33–61.
12. G. D. Evangelidis, & E. Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10), 2008, pp. 1858–1865.
13. H. Rahmani, A. Mahmood, D. Q. Huynh, & A. Mian. Real time human ation recognition using histograms of depth gradients and random decision forests. *IEEE Winter Conference on Applications of Computer Vision (WACV 2014)*, 2014, pp. 626–633.

14. A. Klaser, M. Marszaek, & C. Schmid. A spatio-temporal descriptor based on 3D-gradients. *British Machine Vision Conference (BMVC 2008)*, 2008, pp. 275:1–10.
15. F. Lv, & R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. *European Conference on Computer Vision (ECCV 2006)*, Springer Berlin Heidelberg 2006, pp. 359–372.

Reference Based Image Encoding

S.D. Yamini Devi, Raja Santhanakumar and K.R. Ramakrishnan

Abstract This paper describes a scheme to encode an image using another *reference image* in such a way that an end user can retrieve the encoded image *only* with the reference image. Such encoding schemes could have a potential application in secure image sharing. The proposed scheme is simple and similar to fractal encoding; and a key feature is it simultaneously performs compression and encryption. The encoding process is further speeded up using PatchMatch. The performance in terms of encoding time and PSNR is examined for the different encoding methods.

Keywords Fractal encoding · Reference-based encoding · PatchMatch

1 Introduction

Images can be securely shared using well known techniques in steganography and cryptography [2, 7, 8]. Since raw image data are huge in size, in general, they have to be compressed before sharing. A natural way to share images securely would be to compress them first, followed by scrambling or encrypting. On the other end, a user would need to decrypt and decode in order to retrieve the encoded image. Essentially, in the above scheme, the image encoding and decoding is a 2-step process. In this paper, we propose an image encoding scheme, similar to fractal encoding [3], which simultaneously achieves compression and encryption in a single step.

Few earlier works have attempted to do secure image sharing using fractal codes in two separate steps: compression and encryption. Shiguo lian [5] encrypts some of the fractal parameters during fractal encoding to produce the encrypted and encoded

S.D.Y. Devi (✉) · K.R. Ramakrishnan (✉)
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
e-mail: yamini@ee.iisc.ernet.in

K.R. Ramakrishnan
e-mail: krr@ee.iisc.ernet.in

R. Santhanakumar
Airbus Group Innovations, Airbus Group India, Bangalore, India

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_13

data. In [6], Jian Lock et al. by modifying Li et al. [4], *actually* perform compression and symmetric key encryption in a 2-step process wherein fractal codes of an image is multiplied with a equal sized Mandelbrot image generated through an equation. Before the multiplication, the fractal code and Mandelbrot image matrices are permuted. The product matrix is transmitted along with some few other parameters. Decryption is done through inverse permutation and some matrix manipulations.

In contrast to the above approaches, we propose a single step reference-based image encoding wherein an image is encoded using another "reference" image in such a way that decoding is possible only by the receiver or a user having the same reference image. In other words, the reference image serves like a key for secure image sharing. To begin with in Sect. 2, we give a brief overview of fractal encoding and decoding, followed by a description of our proposed reference-based approach and highlight important differences between the two encoding methods. In Sect. 3 we describe how the PatchMatch algorithm [1] is used in order to reduce encoding time. Experiments and results are provided in Sect. 4 along with some insights about the functioning of PatchMatch in encoding. Finally, Sect. 5 concludes the paper.

2 Reference-Image Based Image Encoding

Since the proposed reference-based encoding is very similar to fractal encoding, for the sake of completeness, in what follows we provide a brief description of fractal image encoding and decoding. For more elaborate details on the theoretical and implementation aspects, please see [3]. Let f be the image to be encoded. Fractal encoding is essentially an inverse problem where an Iterated Function System (**IFS**) W is sought such that the fixed point of W is the image to be encoded f. An approximate solution to this problem is to partition f into M disjoint sub-blocks ("range blocks") f_i such that $f = \cup_{i=1}^{M} f_i$; for each f_i, we find a block of twice the size ("domain block") elsewhere in the same image which best matches to f_i after resizing, spatial transformation (isometry) and intensity modifications (brightness and contrast). The mapping between f_i and its corresponding matching block is a contractive mapping w_i. This process of searching self-similar block can be seen as having two copies of the original image; one is called the range image and the other is called the domain image. The range image contains non-overlapping range blocks f_i, and the domain image contains overlapping domain blocks of twice the size D_j. The relationship between f_i and the best matching block in D_j is represented through the contractive mapping w_i. It can be shown that $W = \cup_{i=1}^{M} w_i$ is also a contractive mapping, and in practice, the fixed point of W is very close to the original f which allows to encode the image f as the parameters of $w_i, i = 1, \ldots, M$. Unlike the encoding, fractal decoding is a much simpler process. Since the encoder is based on the concept of contractive mappings, *starting from any initial image*, application of the contractive mappings $w_i, i = 1, \ldots, M$ repeatedly will converge to the same fixed point which will be an approximation to the encoded image. Image compression is achieved because the

image is entirely encoded in terms of the parameters of the contractive mappings w_i, and the number of the parameters is the same $\forall w_i, i = 1, \ldots, M$. Note that an interesting consequence of fractal encoding is that the size of the encoded data depends only on the number of sub-blocks M; therefore all images with the same number of partitioned sub-blocks will have the encoded data of equal size.

In contrast to the fractal encoding wherein the original image can be recovered from the encoded data *starting from a arbitrary initial image*, our proposed scheme must encode the given image in such a way that the original image (or an approximation) can be recovered only from a *specific reference image*.

2.1 Proposed Encoding Scheme

Unlike fractal encoding described above where the same image is used as domain and range images, in our proposed scheme, we use the reference image as the domain image, and the original image as the range image. For encoding the non-overlapping range blocks, the best matching block among the overlapping domain blocks of equal size in the domain image (reference) is found. By the above two simple modifications, the proposed encoding method is sufficient to meet our objective. In other words, for decoding, the same domain image used for encoding is *necessary as a key* to recover the original (range) image. To keep the reference-based encoder simple, while searching for a matching domain block, no spatial transformations are applied to the domain blocks. Given a range block f_i and a candidate domain block we find only the optimal "contrast" and "brightness" which will bring the domain block and range block "closer". Equations for computing optimal values of contrast and brightness are given in [3], p. 21. Although the encoded data has compression properties similar to a fractal encoder, the encoding quality depends on finding suitable domain blocks in the reference image for the range blocks. The block diagrams showing a comparison between reference-based and fractal encoding are shown in Fig. 1, and the pseudo-code for the reference-based encoding is given in Algorithm 1.

Algorithm 1: Reference-based image encoding

Data: Range image I, Reference image I_1
Result: ImageCodes
for $i = 1 : M$ *(No: of Range blocks)* **do**
  for $k = 1 : N$ *(No: of Domain blocks)* **do**
    From $I_1(D_k)$ pick a domain block
    Find s,o (intensity modifying parameters contrast and brightness)
    $[s, o] \leftarrow \min_{s,o} ||f_i - T(D_k; s, o)||$
  end
  Best matching domain block $D_k = \arg\min_{D_k} \left\{ \min_{s,o} ||f_i - (D_k; s, o)|| \right\}$
end

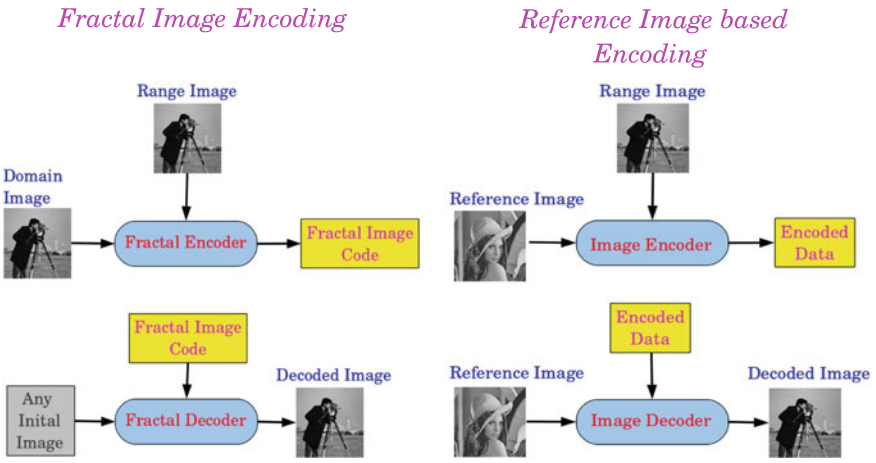


Fig. 1 Fractal versus reference-based image encoding

Given the reference-based image codes, *decoding involves a single step* wherein the mappings w_i are applied on the reference image used during encoding. Note that unlike fractal decoding, *No* iterative process is involved in the our proposed method. In a fractal encoder, the domain block is twice the size of the range block which is important to make the mapping contractive because contractivity is a necessary requirement for the nature of fractal decoding. The pseudo-code for the reference-based decoding is given in Algorithm 2.

Algorithm 2: Reference-based image decoding

Data: ImageCodes: $\{(p(D_i), s_i, o_i)\}_{i=1}^{M}$
Result: Decoded image
Initialization: Domain image I_1;
for $i = 1 : M$ *(No: of Range blocks)* **do**
  From $p(D_i)$ crop domain block: $D_i = I_1(p(D_i))$
  intensity modification with (s_i, o_i)
  DecodedImage$(p(f_i)) \leftarrow T(D_i; s_i, o_i)$
end

Note $\cup_{i=1}^{M} f_i$ covers the entire image

The search process in the above encoding algorithm compares a range block with every candidate domain block to find the best match. Since this makes the encoding time very long, in the next section, we propose a way to reduce the search time by finding an *approximate* matching block using PatchMatch.

3 PatchMatch Reference Encoder

PatchMatch [1] is a iterative randomized algorithm for finding the best match between image patches across two different images A and B. Initially, for a patch centred at (x, y) in image A, a random matching patch (nearest neighbor) is assigned at a offset $f(x, y)$ or $\mathbf{v}$ in image B. Let $D(\mathbf{v})$ denote the error distance between the patches at (x, y) in A and at $((x, y) + \mathbf{v})$ in B. In every iteration, the algorithm refines the nearest neighbor for every patch of image A in two phases: (1) propagation and (2) random search.

Propagation Phase: Assuming the neighboring offsets of (x, y) are likely to be the same, the nearest neighbor for a patch at (x, y) is improved by "propagating" the known offsets of neighbors of (x, y) only if the patch distance error $D(\mathbf{v})$ improves. Propagation of these offsets use different neighbors during odd and even iterations: the updated offset $\mathbf{v}$ in odd iterations is, $\mathbf{v} = \arg\min\{D(f(x, y)), D(f(x - 1, y)), D(f(x, y - 1))\}$, and in even iterations, $\mathbf{v} = \arg\min\{D(f(x, y)), D(f(x + 1, y)), D(f(x, y + 1))\}$.

Random Search: If $\mathbf{v}_0$ is the updated offset after the propagation phase, further improvement is done by constructing a window around $\mathbf{v}_0$ and searching through a sequence of offsets $\{\mathbf{u}_i\}$ at an exponentially decreasing distance from $\mathbf{v}_0$. Let w be the maximum search distance around $\mathbf{v}_0$, and $\mathbf{R}_i$, a sequence of uniformly distributed random points in $[-1, 1] \times [-1, 1]$. The sequence of random offsets is given by $\mathbf{u}_i = \mathbf{v}_0 + w\alpha^i\mathbf{R}_i$, where $\alpha = 1/2$, and $i = 1, 2, \ldots$. The number of random offsets searched is determined by w with the condition that the last search radius $w\alpha^i$ is less than 1 pixel. In this phase, the updated offset is $\mathbf{v} = \arg\min\{D(\mathbf{v}_0), D(\mathbf{u}_1), D(\mathbf{u}_2), \ldots\}$; in other words, the offset is only updated if the patch distance error reduces.

In order to reduce the search time for finding a matching domain block for a range block, PatchMatch is modified in the following ways and incorporated in the reference encoder.

- As an initial step, every range block f_i in the original image (A) is randomly assigned a domain block in domain image (reference or B). An image code is generated for this range block; consisting of the domain block position, range block position, scale and offset. In contrast to PatchMatch where for all patches (overlapping) in A, matching patches in image B are found, in our modification, matching domain blocks need to be found only for range blocks which are non-overlapping.
- Block matching in PatchMatch involves computing Euclidean distance between blocks and finding the block in B with the least Euclidean distance. We, however, change the distance error by finding two parameters for intensity modification, contrast (scale) and brightness (offset), which minimizes the matching error. The expressions for optimal brightness and contrast are the same as in the reference coder (and fractal encoder [3]).

Apart from the above two differences, block matching is similar to PatchMatch, including two phases of propagation and random search to improve the matching domain block for a range block iteratively. The pseudo-code for the PatchMatch reference encoder is given in Algorithm 3.

Algorithm 3: Reference-based encoding with PatchMatch

Data: Given Image A and B(Reference Image)
Result: Decoded image
Initialization: Randomly assign D_i to f_i and compute $[s_i, o_i]$
for $k = 1 : MaxIterations$ **do**
  for $k = 1 : M$ *(No: of Range blocks)* **do**
    Propagation: update if neighbors are better (D_i, s_i, o_i) Random Search: search around $v_o = D_i$ for random matching domain blocks
    Update if candidates are better based on RMS error
  end
end

The decoding process is the same as described in Sect. 2.1 wherein the original image is recovered in a single step by applying the image codes on the same domain image that was used for encoding.

4 Experiments and Results

Reference Encoding: To illustrate a typical result of the reference-based encoding, we have taken the "cameraman" as the range image (Fig. 2a), and "lena" as the domain (reference) image (Fig. 2b). The range block size is chosen to be 8×8, and for finding a matching domain block an exhaustive process is used. The decoded result shown in Fig. 2c indicates a fair reconstruction of the original image.

PatchMatch Encoding: The "lena" image (Fig. 3a) is PatchMatch encoded with range blocks of size 8×8, and using "cameraman" as the reference. The decoded result in Fig. 3b is noticeably different in the boundaries when compared with the original. The same information is reflected in Fig. 3c which shows the matching distance of all the range blocks.

Fig. 2 Reference-based image encoding: **a** range image, **b** reference (domain) image, **c** decoded image

Fig. 3 PatchMatch: **a** original image, **b** decoded image, **c** matching distance of range blocks

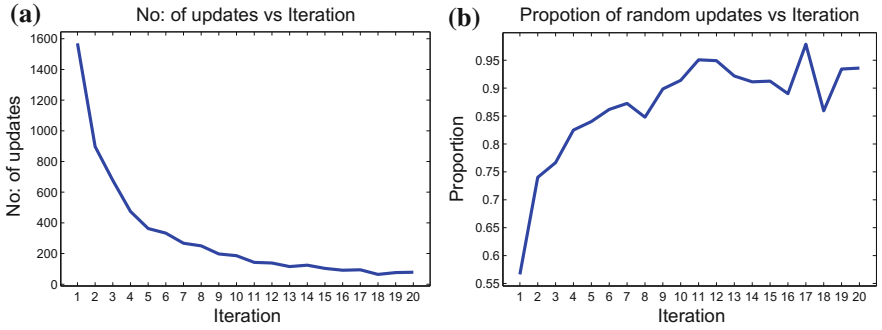


Fig. 4 **a** No: of updates versus iteration, **b** random update versus iteration

To better understand the functioning of PatchMatch in the encoding, we plot two graphs: (1) No: of updates vs iteration in Fig. 4a, and (2) No: of random updates versus iteration in Fig. 4b. As mentioned earlier, in every iteration of PatchMatch, the nearest neighbor (matching domain block) for a range block gets refined either by a spatial neighbor or from a random search. An interesting observation in Fig. 4a, b is that beyond 15 iterations there are significantly less updates from the neighborhood information, and relatively more updates take place in the random search.

The PatchMatch encoder is tested on a synthetic image having many uniform regions (Fig. 5a) with range blocks of size 4×4. The decoded result (Fig. 5b) and the block matching distance (Fig. 5c) show that the range blocks with large matching distance are on the region boundaries. Matching domain blocks for the uniform regions are more likely to be found in first few iterations of the propagation phase (Fig. 6a), but after certain number of iterations most of the updates come from the random search phase (Fig. 6b) for finding matches for range blocks on image boundaries. More study is required to determine the maximum number of iterations for efficiently finding matching blocks.

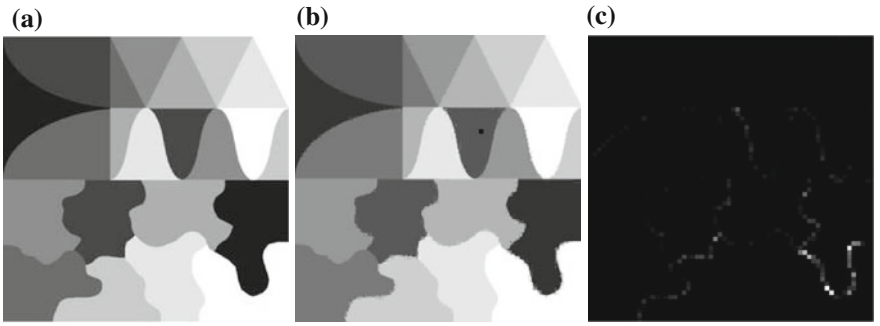


Fig. 5 PatchMatch encoding of an image with many uniform regions: **a** original image, **b** decoded image, **c** matching distance

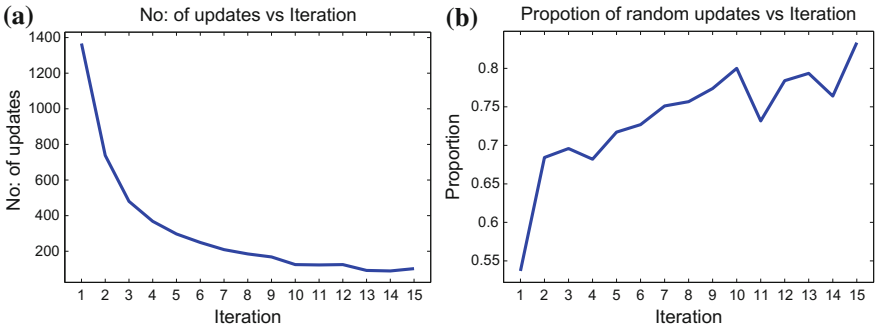


Fig. 6 **a** No of updates versus iteration for image with uniform region, **b** random updates versus iteration

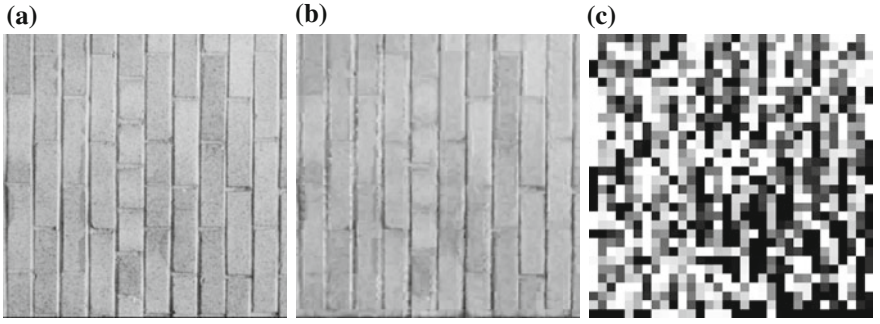


Fig. 7 **a** Original image, **b** decoding using reference image, **c** decoding without the same reference image

For the texture image in Fig. 7a the decoded result using "lena" as reference is shown in Fig. 7b. When an image with all pixels as 255 is used as reference while decoding, the resulting output shown in Fig. 7c indicates that the decoder fails if the appropriate reference is not used.

The tables below summarize the key differences between three types of image encoders: (1) fractal, (2) reference image-based and (3) PatchMatch-based. Table 1 gives the major differences between the original PatchMatch algorithm and the reference-based encoder, while Table 2 highlights some of the key differences between fractal encoding and PatchMatch encoder. Table 3 compares the performance of the above three techniques in terms of PSNR and the encoding time. Fractal encoding results in the best PSNR, but with the highest encoding time compared to reference-based and PatchMatch encoders. Since the reference-based image encoder does not use spatial transforms while searching for matching blocks, the encoding time is less in comparison to fractal encoder. For the "lena" image (Fig. 8a), the decoded result from the three different encoders are shown in Fig. 8b–d.

Table 1 Comparison of PatchMatch algorithm and reference-based encoding

PatchMatch	Reference-based encoder
Best match needed at every (x, y)	Best match needed only at $f_i, i = 1, 2, \ldots M$
Random search and neighbor propagation	Exhaustive search for each f_i
Has no intensity modification	Block matching finds optimal contrast and brightness
Convergence of nearest neighbor is fast	Search is very time consuming

Table 2 Comparison of fractal image encoding and PatchMatch encoder

Fractal encoder	Patchmatch encoder
Range and domain images are the same	Range and domain images are different
Exhaustive search for block matching	Randomized search for block matching
Search time depends on no: of domain blocks	Search time depends on image size
No specific image needed for image decoding	Requires same reference image (domain) for decoding

Table 3 Performance comparison of encoding techniques on "lena" image (256×256) with 4×4 range blocks

	Fractal	Reference	PatchMatch
Iteration	NA	NA	15
PSNR	41.75	38.59	34.90
Time (sec)	2560	1350	963

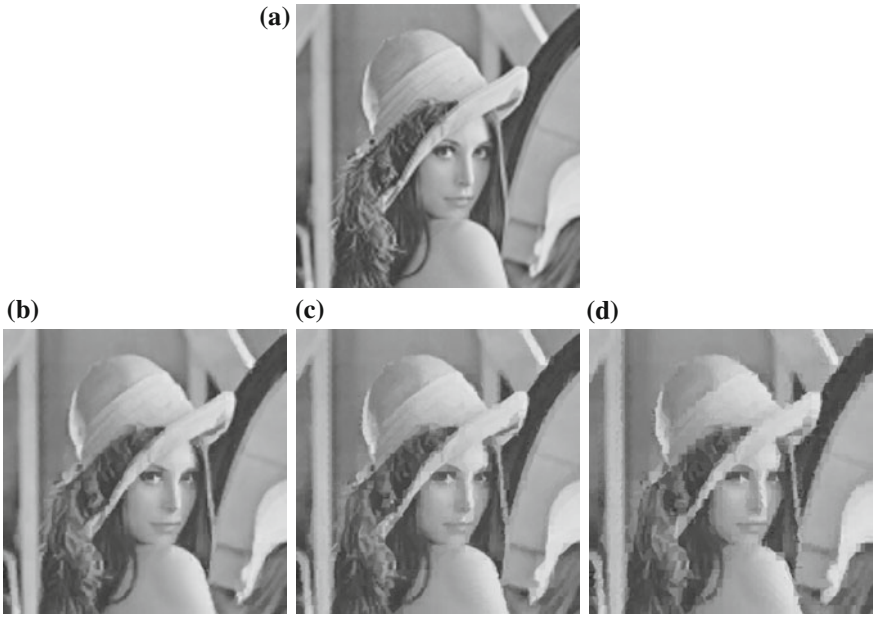


Fig. 8 Comparison of different encoding techniques: **a** original image, **b** fractal, **c** reference-based, **d** PatchMatch-based

5 Conclusions

In this paper we have proposed methods to share images in a secure way using a single step encoding process which combines compression and encryption. Results show that the user will be able to decode a meaningful image only when the "reference image" is given. The proposed work, though not similar to fractal encoding, adopts some of the key features of fractal encoding, a lossy compression technique. PatchMatch has been leveraged in order to speed up the encoding process. However, the results obtained are inferior in comparison to the reference-based encoder. One future direction to improve the PSNR could be selecting an "appropriate" reference image from an image set.

References

1. Connelly Barnes, Ei Shechtman, Adam Finkelstein, and Dan B Goldman. PatchMatch: A Randomized Correspondence Algorithm for Structured Image Editing. pages 24:1–24:11. ACM, 2009.
2. Chin-Chen Chang, Min-Shian Hwang, and Tung-Shou Chen. A New Encryption Algorithm for Image Cryptosystems. *Journal of Systems and Software*, 58(2):83–91, 2001.
3. Yuval Fisher. *Fractal Image Compression: Theory and Application*. Springer Verlag, 1995.

4. Xiaobo Li, Jason Knipe, and Howard Cheng. Image Compression and Encryption Using Tree Structures. *Pattern Recognition Letters*, 18(11):1253–1259, 1997.
5. Shiguo Lian. Secure fractal image coding. *CoRR*, abs/0711.3500, 2007.
6. A. J. J. Lock, Chong Hooi Loh, S.H. Juhari, and A Samsudin. Compression-Encryption Based on Fractal Geometric. In *Computer Research and Development, 2010 Second International Conference on*, pages 213–217, May 2010.
7. Debasis Mazumdar, Apurba Das, and Sankar K Pal. MRF Based LSB Steganalysis: A New Measure of Steganography Capacity. In *Pattern Recognition and Machine Intelligence*, volume 5909, pages 420–425. Springer Berlin Heidelberg, 2009.
8. Ren Rosenbaum and Heidrun Schumann. A Steganographic Framework for Reference Colour Based Encoding and Cover Image Selection. In *Information Security*, volume 1975, pages 30–43. Springer Berlin Heidelberg, 2000.

Improving Face Detection in Blurred Videos for Surveillance Applications

K. Menaka, B. Yogameena and C. Nagananthini

Abstract Performance of face detection system drops drastically when blur effect is present in the surveillance video. Motivated by this problem, the proposed method deblurs facial images to detect and improve faces degraded by blur in the scenario like banks, ATMs where sparse crowd is present. Prevalent Viola Jones technique detect faces, but fails in the presence of blur. Hence, to overcome this, first the target frame is decomposed using Discrete Wavelet Transform(DWT) into LL, LH, HL and HH bands. The LL band is processed using Lucy-Richardson's algorithm which removes blur using Point Spread Function (PSF). Then the super enhanced de-blurred frame without ripples is given into Viola-Jones algorithm. It has been observed and validated experimentally that, the detection rate in the Viola Jones algorithm has been improved by 47 %. Experimental results illustrate the effectiveness of the proposed algorithm.

Keywords Face detection · DWT · Lucy-Richardson's algorithm · De-blurring · Viola-Jones

1 Introduction

Most of the available surveillance cameras are of low resolution. Hence, face detection is the most challenging task in surveillance systems than the normal face detection in the photo images. Challenges faced by face detection often involve low resolution images, blurred version of images, occlusion of facial features such as beards, moustaches and glasses, facial expressions like surprised, crying and poses

K. Menaka (✉) · B. Yogameena · C. Nagananthini
Department of ECE, Thiagarajar College of Engineering, Madurai, India
e-mail: ece.menaka@gmail.com

B. Yogameena
e-mail: b.yogameena@gmail.com

C. Nagananthini
e-mail: nagananthiniece2010@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_14

like frontal and side view, illumination and poor lighting conditions such as in video surveillance cameras image quality and size of image as in passport control or visa control, complex backgrounds also make it extremely hard to detect faces. To detect and recognize face in an intensity image, holistic methods which use (Principal Component Analysis) PCA [1] can recognize a person by comparing the characteristics of face to those of known individuals, FLDA and LBP (Local Binary Pattern) [2] which summarizes local structures of images efficiently by comparing each pixel with its neighboring pixels can be used. These methods work well under low resolution but fails in case when there is a large variation in pose and illumination. The alternate approach is feature based approach which uses Gabor filter and Bunch graph method. These methods work well under ideal condition and the disadvantages are difficult in computation and automatic detection.

Nowadays, face detection and improving its detection rate in surveillance video plays a major role in recognizing the face of individuals to identify the culprits involved in crime scenes. The various approaches used to detect faces in surveillance video are based on Crossed Face Detection Method that instantly detects low resolution faces in still images or video frames [3], Howel and Buxter method, Gabor wavelet analysis and the most commonly used is Viola-Jones detector [4]. Face recognition can be done by using algorithms such as reverse rendering and Exemplar algorithm. These methods for face detection and recognition works well under low resolution and cluttered background and needs super enhancement techniques.

2 Related Work

From the literature review, it is found that there is not much research concentrated on improving face detection rate in surveillance videos for face recognition to authenticate the person specifically. There are so many factors that affect the efficacy and credibility of any surveillance videos such as blur, occlusion, masking, illumination and other environmental factors included. This paper is designed with respect to scenarios where blur is a major concern. Though, there is little research going on in removal of noise and overcoming illumination changes, not much is focused on blur removal. Hence, it is vital to develop a face detection algorithm that is robust to blur which will help the smart surveillance system to recognize the person.

As per the survey related to the topic, the face detection and recognition in blurred videos is extremely difficult. It sometimes provides inaccurate detection rate in presence of blur. Here are some of the approaches used for de-blurring in face recognition and detection method. In [5], a combination of image-formation models and differential geometric tools is used to recover the space spanned by the blurred versions. Joint blind image restoration and recognition approach is to jointly de-blur and recognition of face image. This approach is based on sparse representation to solve challenging task of face recognition from low quality image in blind setting [6].

Laplacian Sharpening filter can be used if the frame is degraded by Gaussian blur only but performs poorly in the presence of noise. The box blur can be removed by using Poisson Map algorithm but it is slow as it involves more number of mathematical operations. The most important Gaussian blur can be removed by means of Optimized Richardson-Lucy algorithm which does not concern the type of noise affecting the image. In [7], corrupted image has been recovered using Modified Lucy Richardson algorithm in the presence of Gaussian blur and motion blur. This paper shows the efficacy of Modified Lucy Richardson method compared to Wiener filter, Constraint Least Square method and Lucy Richardson algorithms. The main contributions towards this work are

- From the literature, the DWT method had been simply used to decompose the still image [8] into high frequency parts and low frequency parts. But this proposed work has been applied for surveillance video to extract the low frequency part for the further process. Also, it has been applied for a specific scenario such as ATM, institute and banks. This enhances the person identification in surveillance video.
- The Lucy-Richardson algorithm was applied to the still images in the earlier papers whereas the same Lucy-Richardson algorithm has been applied for surveillance video which is used as a de-blurring descriptor.
- Performance on improvement of face detection in surveillance video has been analyzed by applying the L-R algorithm to selective DWT component.

3 Proposed Method

At present, all the surveillance videos are almost rendered useless. The volumes of data captured, stored and time stamped doesn't helps to solve crimes. The major problem is blurred video. Thus, the proposed method involved removing blur in the surveillance videos. The video is first converted into required number of frames that has the target person for detection. The target frames are first fed as input to the most rampant face detection algorithm, called the Viola Jones algorithm, in the computer vision field of research. Though it can be trained for object class detection, the problem of face detection is the primary motivation. Then, the same frames are taken and a series of preprocessing techniques are applied. The blurred frame is transformed using DWT. The foremost reason for using DWT it captures both frequency and location information (location in time) whereas Fourier transforms is temporal resolution. DWT is used to extract the LL band information from the target frame. This contains the most varied smooth information. The proposed method takes this LL-band information and performs de-blurring in that image.

The deblurring algorithm chosen is Lucy-Richardson's (L-R). The Richardson–Lucy algorithm, also known as Lucy Richardson deconvolution, is an iterative procedure for recovering a latent image that has been blurred by a known PSF.

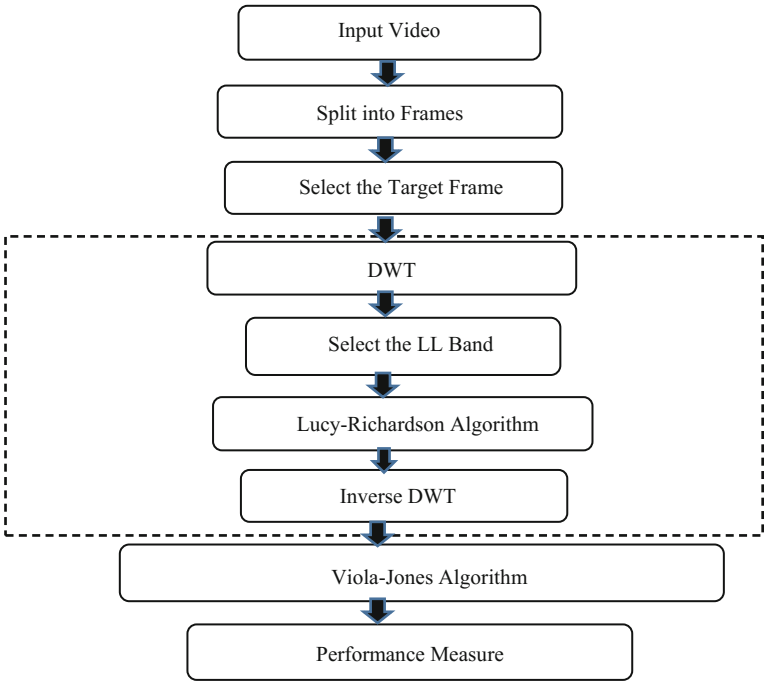


Fig. 1 Flow chart of the process

Thus the effectiveness of de-blurring and reconstruction is increased by 47 %. After de-blurring the enhanced LL band image along with its other counterparts such as LH, HL and HH are reconstructed using Inverse DWT. Thus a perfectly reconstructed image, with enhanced features is obtained. Now the perfectly reconstructed image is taken and processed using Viola-Jones algorithm. Then the performance is obtained for the proposed algorithm in terms of detection rate and is compared with Viola Jones algorithm. The flow chart of the process is given in Fig. 1.

3.1 Methodology

3.2 Discrete Wavelet Transform

The discrete variant of the wavelet transform is Discrete Wavelet Transform. The image is processed by DWT using an appropriate bank of filters and this transformed images involve D levels based on tree structure. Based on the criteria of extracting strings of image samples, it follows two approaches. The first approach involves generating the string by queuing image lines and then executing decomposition on D levels, after which the D strings are generated by queuing the

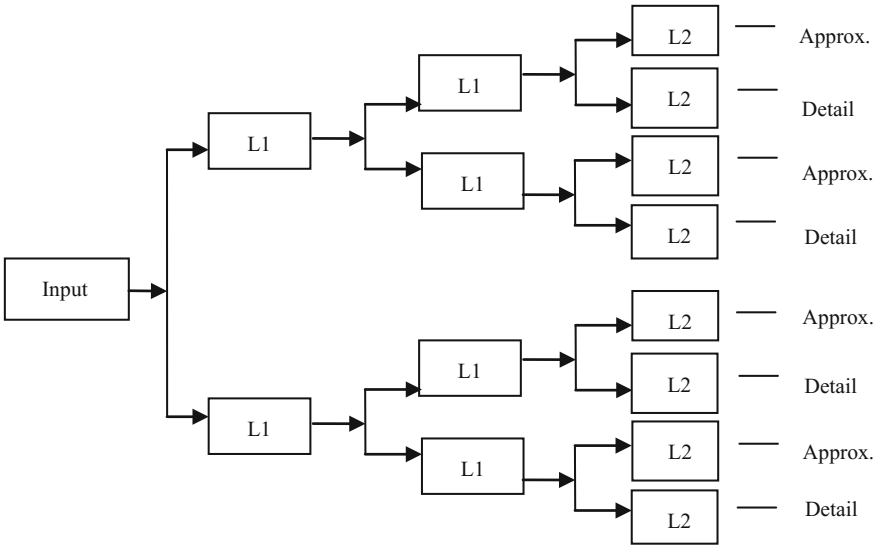


Fig. 2 Bank of filters iterated for the 2D-DWT standard

columns from the sub-images found and decomposition is again done on each string. The simplified version of resulting decomposition extended up to the third level, is shown in Fig. 2.

3.3 Lucy Richardson Deblurring Algorithm

Since the nonlinear iterative methods often yield results better than those obtained with linear methods, Lucy Richardson (L-R) algorithm which is a nonlinear iterative restoration method is chosen. The L-R algorithm arises from maximum likelihood formulation in which image is modelled with poison statistics. Maximizing the likelihood function of the model yields an equation is satisfied when following iteration converges:

$$f_{k+1}(x,y) = f_k(x,y)[h(-x,-y) * g(x,y)/(h(x,y) * f_k(x,y))] \quad (1)$$

Based on the size and complexity of PSF matrix, good solution is obtained. Hence, the specific value for the number of iterations is difficult to claim. The algorithm usually reaches a stable solution in few steps with a small PSF matrix which makes the image smoother. The computational complexity is increased when increasing the number of iterations which will result in amplification of noise and the ringing effect is also produced. Hence, by determining the optimal number of

iterations manually for every image a good quality of restored image is obtained. The optimal number is obtained by alternating one decomposition by rows and another one by columns, iterating only on the low-pass sub-image according to the PSF size.

3.4 Improvisation of L-R Method

In the proposed method, the DWT of degraded image is taken. The target frame is decomposed into four sub-bands by DWT: three high frequency parts (HL, LH and HH) and one low frequency part (LL). The high frequency parts may contain the fringe information while the low frequency part may contain strength of target frame. These low frequency parts are more stable. Therefore, Lucy Richardson algorithm is applied to LL sub-band. The steps involved in this process are

(1) A non-blurred image f(x, y) is chosen.
(2) Gaussian or Motion Blur is added to produce blurred image.
(3) Gaussian noise is added to the blurred image to produce degraded image.
(4) The degraded image is decomposed into four sub-bands LL, HL, LH and HH by using DWT.
(5) Apply L-R algorithm to the LL sub-band to produce the restored low frequency band (LL Modified by LR).
(6) For the remaining sub-bands (HL, LH and HH) apply the threshold.
(7) Finally, the restored image is obtained by applying inverse DWT to restored low frequency band (LL modified by LR) and high frequency bands (HL, LH and HH).

3.5 Viola-Jones Algorithm: Face Detection

The features used in the framework of detection involve the addition of pixels of an image within rectangular areas. These features resemble the Haar basis functions, which have been employed in the field of image-based object detection. Subtracting the total sum of the pixels inside clear rectangles from the total sum of the pixels inside shaded rectangles gives the value of given feature. It is because in a feature, each rectangular area is always next to at least one other rectangle. Followed that using six array references, any two-rectangle feature can be computed, using eight any three-rectangle feature, and using just nine array references any four-rectangle feature can be calculated. The estimation of the strong classifiers is not fast enough to run in real-time. For this reason, based on the order of complexity, these strong classifiers are put together in a cascade form. Each consecutive classifier is trained only on those selected samples which pass through the preceding classifiers. No further processing is performed if at any step in the cascade, a classifier discards the

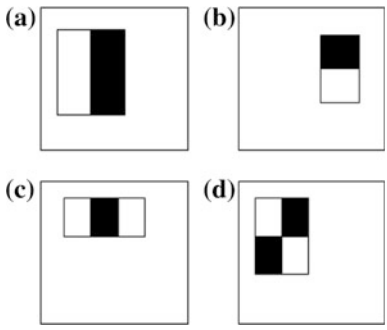


Fig. 3 Haar features. **a**, **b** indicate two-rectangle features, **c** indicates a three-rectangle feature and **d** indicates a four-rectangle feature

sub-window under inspection and it continues to search for the next sub-window. Hence, the cascade has the degenerate tree structure. In the case of faces, to obtain approximately 0 % of false negative rate and 40 % of false positive rate, the first classifier in the cascade called the attentional operator uses only two features. The effect of this single classifier decreases half the amount of times the entire cascade is evaluated. Some Haar features are given in the Fig. 3.

A threshold value is obtained from Haar features using the equation,

$$\text{Threshold Value} = \sum(\text{pixels in white area}) - \sum(\text{pixels in black area}) \quad (2)$$

If this threshold value is above a certain level, then the corresponding area is detected as face and if it is below that level it is found to be a non-face region. Thus, to match the false positive rates typically achieved by other detectors, each classifier can get away with having surprisingly poor performance.

4 Results and Discussion

Experimentations are carried out on a bench mark Umist [9] dataset and college surveillance video with 16 min duration, with the resolution of 811 × 508, frame rate of 11 frames/sec and 11339 numbers of frames. The camera is mounted at appropriate location on the wall to focus the intended area and focus the passing objects.

First, Viola Jones face detection is applied on both the datasets to test the performance. It is found that the accuracy of detection decreases in the presence of blur. Hence, the frames are de-blurred and the same Viola Jones algorithm is applied. After deblurring the detection rate is improved. Figure 4 shows the bench mark dataset which is used as a default database. Viola Jones face detection algorithm is applied for an image to check its performance. Out of 78 faces in the dataset (50 profile faces and 28 frontal faces), Viola Jones the most popular algorithm is able to detect only 31 frontal faces and 28 profile faces with a total of

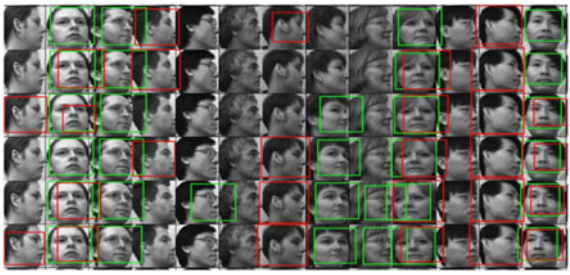


Fig. 4 Face detection using Viola Jones Bench mark Umist dataset with Image size of 1367 × 652

45 faces (12 faces are detected both as frontal and profile faces). The accuracy rate is just 58 %.

TCE college dataset is used to test the performance of Viola Jones in real time surveillance video, in which there are two frontal, non-occluded faces which could have been detected by Viola Jones. But it fails to detect faces which is shown in Fig. 5. The reason is the presence of low resolution and blur in the video, which is the case in most real time applications. So L-R method is applied to the frames and then it is seen that Viola Jones algorithm is able to detect the non-occluded, de-blurred faces.

Also, manually Gaussian blur is introduced to the bench mark data set to test the performance of Viola Jones. Gaussian blur with filter size of 20 and standard deviation of 3.5 is introduced. It is seen that the performance suddenly reduces. Out of 78 faces in that dataset, the observation is that it has detected only 30 faces which is shown in Fig. 6. (5 faces detected as both frontal and profile face) with a minimum accuracy of 38 %.

Fig. 5 **a–d** Face detection using Viola Jones Dataset: College dataset Frame size: 704 × 576 and **e–h** Face detection in de-blurred frame Dataset: College dataset Frame size: 704 × 576

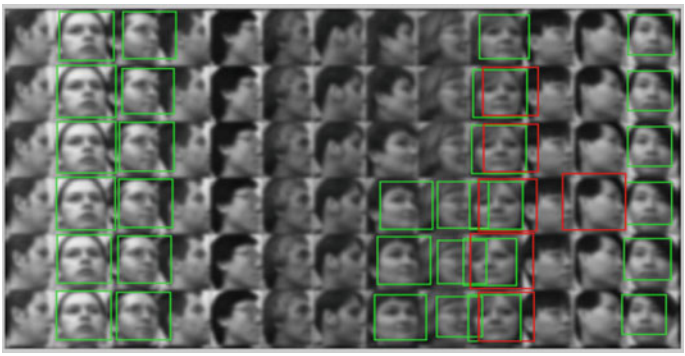


Fig. 6 Face detection in blurred image: Umist Image size: 1367 × 652

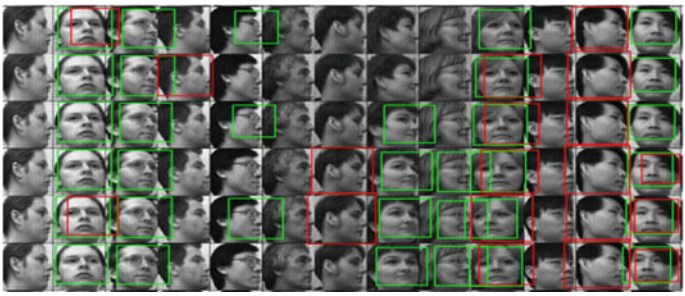


Fig. 7 Face detection in de-blurred image: Bench mark Umist dataset Image size: 1367 × 652

In this bench mark dataset, it is seen that after deblurring the accuracy has greatly increased and it has reached nearly the ideal image which was not blurred. Out of 78 faces in the dataset it has detected 30 frontal faces and 28 profile faces with a total of 43 faces which is shown in Fig. 7 (10 faces detected as both frontal and profile face). The accuracy percentage is increased to 55 %

The hit rate or face detection rate for above dataset is calculated using the following:

$$\text{Detection rate} = \frac{\text{Total number of faces correctly detected}}{\text{Total number of faces available}} \times 100$$

Performance on surveillance video has been analyzed and the results are shown in Fig. 8.

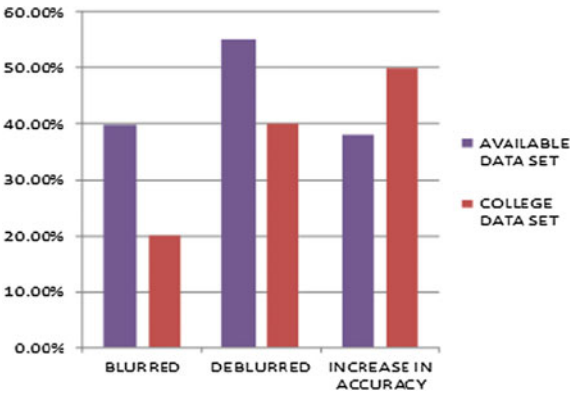


Fig. 8 Performance comparison of Viola Jones method and the proposed method

5 Conclusion

Thus the face detection accuracy in surveillance video can be greatly improved by preprocessing the frames initially i.e. deblurring the frames. Subsequently, existing Viola Jones system is applied for face detection. It has been proved that the proposed method increases the detection accuracy over 47 %, when compared to that of Viola Jones algorithm. The future work includes building a fully automated face recognition system invariant to blur and illumination. Also, estimation of PSF can be automated thereby removing blur for all frames in a video completely. Other parameters such as noise, occlusion can also be taken into consideration for robust face detection and recognition system can be built for a smart surveillance system.

Acknowledgements 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. Also, it has been supported by UGC under Major Research Project Scheme entitled, Intelligent Video Surveillance System for Human Action Analysis with reference F.No.41-592/2012(SR).

References

1. Turk, Matthew, and Alex Pentland: Eigenfaces for recognition. In: Journal of cognitive neuroscience, vol. 3, Issue 1, pp. 71–86. (1991).
2. Di Huang, Caifeng Shan; Ardabilian, M., Yunhong Wang: Local Binary Patterns and Its application to Facial Image Analysis: A Survey. In: IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, vol. 41, Issue 6, pp. 765–781. IEEE (2011).
3. Amr El, Maghraby Mahmoud, Abdalla Othman, Enany Mohamed, El Nahas, Y.: Hybrid Face Detection System using Combination of Viola - Jones Method and Skin Detection. In: International Journal of Computer Applications, vol. 71, Issue 6, pp. 15–22. IJCA Journal (2013).
4. Yi-Qing Wang: An Analysis of the Viola-Jones Face Detection Algorithm. In Image Processing On Line. vol. 2, pp. 1239–1009, (2013).

5. Raghuraman Gopalan, Sima Taheri, Pavan Turaga, Rama Chellappa: A blur robust descriptor with applications to face recognition. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, Issue 6, pp. 1220–1226, IEEE (2013).
6. Haichao Zhang, Jianchao Yang, Yanning Zhang, Nasser M. Nasrabadi and Thomas S. Huang.: Close the Loop: Joint Blind Image Restoration and Recognition with Sparse Representation Prior. In: ICCV Proceedings of IEEE international conference on computer vision., pp. 770–777, Barcelona, Spain (2011).
7. Swati Sharma, Shipra Sharma, Rajesh Mehra: Image Restoration using Modified Lucy Richardson Algorithm in the Presence of Gaussian and Motion Blur. In: Advance in Electronic and Electric Engineering. vol. 3, Issue 8, pp. 1063–1070, (2013)
8. Harry C. Andrews, Hunt, B.R.: Digital Image Restoration. Book Digital Image Restoration, Prentice Hall Professional Technical Reference. (1977).
9. D. Graham. The UMIST Face Database, 2002. URL http://images.ee.umist.ac.uk/danny/database.html. (URL accessed on December 10, 2002).

Support Vector Machine Based Extraction of Crime Information in Human Brain Using ERP Image

Maheshkumar H. Kolekar, Deba Prasad Dash and Priti N. Patil

Abstract Event related potential (ERP) is a non-invasive way to measure person's cognitive ability or any neuro-cognitive disorder. Familiarity with any stimulus can be indicated by the brain's instantaneous response to that particular stimulus. In this research work ERP based eye witness identification system is proposed. Electroencephalogram (EEG) signal was passed through butterworth band-pass filter and EEG signal was segmented based on marker. EEG segments were averaged and ERP was extracted from EEG signal. Grey incidence degree based wavelet denoising was performed. ERP was converted to image form and structural similarity index feature was extracted. Radial basis function kernel based support vector machine classifier was used to classify a person with or without crime information. The observed accuracy of proposed approach was 87.50 %.

Keywords ERP · EEG · ERP image · Support vector machine classifier

1 Introduction

Crime rate has increased a lot in past 2–3 years all over the world. As per national crime record bureau, India [1] cognizable crime in India has increased steadily from 1953 till date. During 2012 a total of 6041559 cognizable crimes comprising of 2387188 penal code crime and 3654371 special and local laws crime were reported.

This research work is funded by Centre on Advanced Systems Engineering, Indian Institute of Technology, Patna.

M.H. Kolekar (✉) · D.P. Dash
Indian Institute of Technology Patna, Bihta, India
e-mail: mahesh@iitp.ac.in

D.P. Dash
e-mail: dpdash.srf14@iitp.ac.in

P.N. Patil
Netaji Subhash Institute of Technology, Amhara, Patna, India
e-mail: priti2008@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_15

Conviction rate is fairly low in 2012 in many states of India. A technique which can identify the concealed crime information in brain will be helpful to validate the authenticity of information. Techniques available today target physiological body parameters such as heart rate, electordermal activity, blood oxygen saturation etc. for validation of crime information. Psychological parameters include emotion assessment, voice change during test and questionnaires [2]. Charlotte [3] used eye fixation as a parameter to detect concealed information in brain. They found that more fixation duration to concealed information compared to non-target pictures. Uday Jain [4] used facial thermal imaging for crime knowledge identification. Maximum classification rate achieved was 83.5 %. Brain wave frequency change was also explored for crime knowledge detection. Maximum match overview achieved was 79 %. Wavelet analysis was also used for decomposition of EEG signal and analyzing the frequency band activity changes for crime knowledge detection. EEG signal was recorded during question answer session and beta frequency band of EEG signal showed maximum variation [5]. Vahid et al. [6] used ERP as an tool for crime knowledge identification. Features extracted were morphological features, frequency features and wavelet features. Linear discriminant analysis was used for classification purpose. Farwell and Richardson had conducted a study to detect brain response to stimulus in four different fields like real life event, real crime with substantial consequence, knowledge unique to FBI agent and knowledge unique to explosive expert [7]. Paradigm for this work was designed based on the work done by Farwell and Richardson. Present research work is unique in addressing the problem of false eye witness identification, feature used extracted from ERP image is unique and data collection method is modified version of Farwell and Richardson paradigm according to the

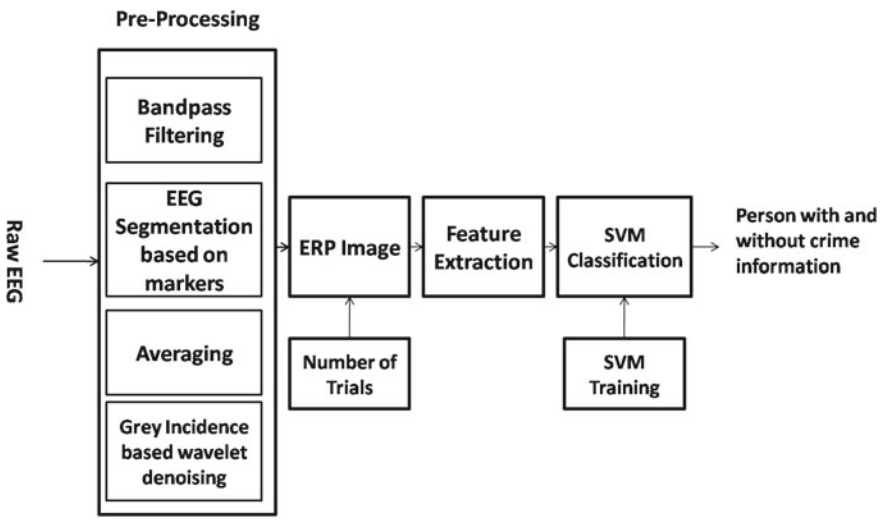


Fig. 1 Block diagram of image processing approach

requirement of the problem. Research work presented in this paper focuses on novel combination of existing technique for classifying ERP response as crime related or not related more efficiently (Fig. 1).

2 Data Collection

Eye witness claims to have information about crime. Authors have designed the test to verify the authenticity of the eye witness. In this research work stimulus related and not related to crime are shown to the participants. False eye witness can identify target (crime not related) stimulus but will fail to recognize stimulus from that particular crime. Total 10 participants aging 18–22 years participated in the research work voluntarily. Crime topic was selected not known to anyone. They were divided into two groups called information present group given information about crime and information absent group having no information about crime.

Nexus 10 neuro-feedback system was used for data collection from Cz and Pz electrode position. 10–20 electrode placement system was followed for electrode positioning. Participants were asked to observe the stimulus and to give response if they recognize the stimulus or not by pressing the laptop right arrow key. A total of 3 set of picture stimulus were shown, crime instruments, crime place and victim name. Each stimulus set consists of 20 images, 16 target and 4 probe stimulus. Target stimulus is the stimulus not from that particular crime but related to some other similar crime and probe stimulus is the stimulus from that particular crime. Five trials were performed per image set.

3 Proposed Approach

3.1 Pre-processing

Bandpass Filtering, Segmentation and Averaging

Signal was recorded with sampling frequency of 256 Hz. Fourth order Butterworth band pass filter with cut-off frequency (0.2–35 Hz) was used to keep data up to beta band and remove others. EEG signal was segmented based on marker and averaged to extract ERP signal.

Grey Incidence Degree (GID) Based Wavelet Denoising

Grey system theory can be applied to problems involving small samples and poor information. In real world similar situation arises many a time and so grey theory

application is popular. Grey incidence degree based wavelet denoising method is used to denoise the signal without losing the originality of the signal. The grey similarity is calculated and threshold is selected based on the similarity between the approximate and detailed coefficients [8].

Definition 1 $x_i = x_1(1), x_1(2), \ldots, x_i(n)$ be the behavioral time sequence and let D_1 be the sequence operator, then-
$x_1 \times D_1 = (x_1(1) \times d1, x_1(2) \times d2, \ldots, x_i(n) \times d1)$
Where
$x_i(k) \times d1 = x_i(k) \div x_i(1)\ , x_i(1) \neq 0, k = 1, 2, \ldots n$
Then D_1 is called initial value operator and $x_i D_1$ is a mapping of x_i under the D_1.

Theorem 1 *Let the approximate time sequence* $x_o = (x_o(1), x_o(2), \ldots x_o(n))$ *and* $x_i = (x_i(1), x_i(2), \ldots x_i(n))$ *i = 1, 2, … n and* $\xi \in (0, 1)$ *define*

$\gamma(x_o(k), x_i(k)) =$

$$\frac{\min_i \min_k \mid x_o(k) - x_i(k) \mid + \xi \max_i \max_k \mid x_o(k) - x_i(k) \mid}{\mid x_o(k) - x_i(k) \mid + \xi \max_i \max_k \mid x_o(k) - x_i(k) \mid} \tag{1}$$

and

$$\gamma(x_o, x_i) = \frac{1}{n} \sum_{k=1}^{n} \gamma(x_o(k), x_i(k)) \tag{2}$$

then $\gamma(x_o, x_i)$ *is a degree of grey incidence between* x_o *and* x_i*, where* ξ *is known as distinguishing coefficient.*

Wavelet is a waveform of limited duration having average value of zero. Wavelet is an efficient tool for analyzing local characteristic of non stationary EEG signal. Discrete wavelet transform of a signal $x(t)$ is represented by

$$c_j(k) = \sum_m h_0(m - 2k) * c_{j+1}(m) \tag{3}$$

$$d_j(k) = \sum_m h_1(m - 2k) * c_{j+1}(m) \tag{4}$$

cj is the approximate coefficient and dj is the detail coefficient. h_0 and h_1 represents low and high pass filter respectively.

Filtering is done by the scaled and shifted version of the wavelet. In this study daubechies wavelet of order 4 was used for decomposition of EEG signal [8]. Here daubechies wavelet was selected because of similarity of the waveform with that of P300 ERP component. The original ERP was decomposed into different level of high and low frequency component. The level of decomposition was selected according

to the frequency band of EEG signal. Signal was decomposed up to 6 levels. Noisy components were removed by setting a hard threshold. Threshold is selected based on absolute and detail coefficients GID value [8].

$$Threshold = \sigma \times \gamma \times \sqrt{2 \times (\log l)} \tag{5}$$

3.2 ERP Image

Event related potential varies in latency and amplitude across each trial. So it is difficult to select a common window to differentiate the signal based on amplitude. ERP represented in image form can be helpful to find common activation region in each trial. Let s be the total average trial set for each block defined as $s = s1, s2, s3$. ERP image was constructed for each block, crime image, crime place and victim name stimulus set by plotting time in x axis, trial in y axis for each block response [9] and color represents amplitude variation of signal. Figures 6, 7, 8 and 9 represents ERP images of crime related and not related stimulus.

3.3 Feature Extraction

Structural Similarity Index (SSIM)

Structural similarity index algorithm assess three terms between two images x and y, luminescence l(x, y), contrast c(x, y) and structure s(x, y). Mathematically it can be defined as-

$$l(x, y) = \frac{2 \times \mu_x \times \mu_y + c_1}{{\mu_x}^2 + {\mu_y}^2 + c_1} \tag{6}$$

$$c(x, y) = \frac{2 \times \sigma_x \times \sigma_y + c_2}{{\sigma_x}^2 + {\sigma_y}^2 + c_2} \tag{7}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x \times \sigma_y + c_3} \tag{8}$$

where $c_1 = (k_1 \times l)^2$, $c_2 = (k_2 \times l)^2$, $c_3 = \frac{c_2}{2}$ μ_x and μ_y are the mean value of image x and y. σ_x and σ_y represents the variance of x and y and σ_{xy} represents the co-variance between image x and y. l is the dynamic range of pixel value. $k_1 \ll 1$ and $k_2 \ll 1$ are scalar constant. The constants c_1, c_2 and c_3 provides spatial masking properties and ensure stability when denominator approaches zero. combining 3 terms

$$SSIM(x, y) = [l(x, y) \times c(x, y) \times s(x, y)] \tag{9}$$

As there is variation in ERP response for probe and target stimulus, similarity comparison between two response is an good approach to differentiate person with or without crime knowledge. In this paper image similarity was calculated between the brain response to probe and target stimulus. Window of 200 ms was selected and similarity index was calculated for each 200 ms block. For comparison between different group 400–600 ms window was selected as maximum variation can be observed in that window.

ERP Signal Features

ERP signal features were extracted to compare its effectiveness with that of image processing approach. Extracted features are power spectral entropy, Energy, Average ERP peak and Magnitude square coherence. Some features equations are given below.

$$H = -\sum_{i=1}^{n} p_i \times ln(p_i) \tag{10}$$

where p_i is the amplitude of power spectral density.

$$coh(f) = \frac{|\, p_{xy}(f)\,|^2}{p_{xx}(f) \times p_{yy}(f)} \tag{11}$$

where p_{xy} is the cross spectral density of $x(t)$ and $y(t)$ and p_{xx} and p_{yy} is the auto spectral density of x(t) and y(t) respectively.

3.4 Support Vector Machine Classifier

In this research work Support Vector Machine (SVM) classifier is used for classification because this is a two class problem and SVM is one of the efficient and most used classifier in Neuroscience research. The idea of SVM is to find an optimal hyperplane which can separate the two classes. Support vectors are the points closest to the separating hyperplane. Margin of both classes is found out by finding the line passing through the closest point [10, 11]. Margin length is set to be $\frac{2}{\|w\|}$, Where w is the line perpendicular to the margin.

When the class cannot be linearly separated, optimal hyperplane can be found out by allowing some error in linear separation or converting the data into linearly separable set by transforming it to a higher dimension. The function used for converting the data in to a higher dimension is called the kernel function. w is the vector perpendicular to the hyperplane and b is the bias. The equation $wx + b = 0$ represents the hyperplane and $wx + b = 1$ and $wx + b = -1$ represents the margin of separation of two class.

$$L = \begin{cases} \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \alpha_n \alpha_m y_n y_m x_n^t x_m - \sum_{n=1}^{N} \alpha_n, \\ \alpha_n \geq 0, \\ y \times \alpha = 0 \end{cases} \tag{12}$$

$k(x_n, x_m)$ is the matrix after transforming data in to higher dimension. The function mentioned above has to be minimized to get values of α. The condition to be followed is shown above.The value of w and b is found by

$$w = \sum_{n=1}^{N} \alpha_n \times k(x_n, x^{'}) \times y_n \tag{13}$$

$$b = y_n - \sum_{\alpha_n > 0} (\alpha_n \times y_n \times k(x_n, x^{'}) \tag{14}$$

Here kernel based SVM classifier was used. Radial basis function (RBF) kernel was used to classify data into person with or without crime knowledge. Kernel mathematically represented as

$$K(x_n, x^{'}) = \exp(-\gamma \parallel x - x^{'} \parallel^2) \tag{15}$$

Based on the kernel, the decision function is modified as

$$d(x) = sign(\sum_{\alpha_n > 0} \alpha_n \times y_n \times k(x_n, x^{'}) + b) \tag{16}$$

Entire feature set is divided into information present group designated as $+1$ and information absent group designated as -1. Five participants from information present group and five from information absent group features were given as input to classifier for training purpose. Coherence and Maximum coherence feature resulted in lower accuracy compared to features such as power spectral entropy, energy and average peak amplitude. Structural similarity extracted from ERP images was used as feature vector. Accuracy obtained with structural similarity index is 87.50 %.

4 Results

Wavelet denoising using grey incidence degree based threshold approach denoise the signal keeping the originality of the signal intact. Universal threshold based wavelet denoising resulted in more smoothed signal. Since in ERP interpretation, amplitude plays a critical role GID based wavelet denoising is proposed in this research. Familiarity of stimulus is indicated if the person has higher P300 ERP activation to probe stimulus. For exact assessment of difference in P300 ERP component both ERP

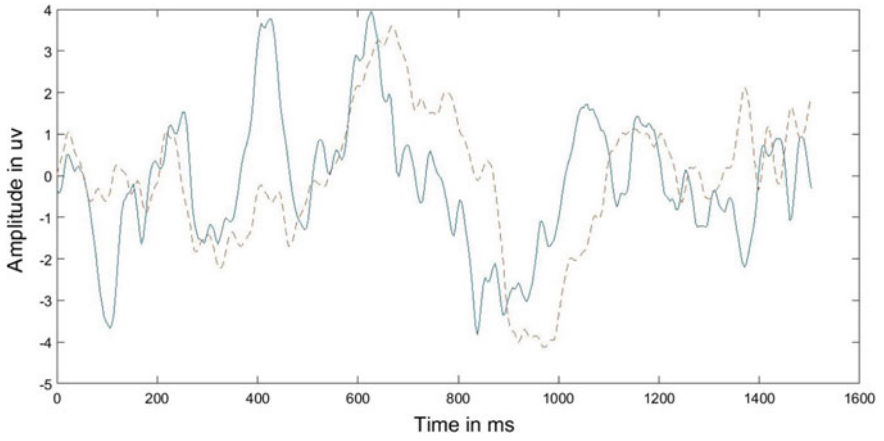


Fig. 2 Grand average ERP of person with crime information for Cz electrode (*dotted line* represents ERP for stimulus not related to crime, *solid line* represents ERP for stimulus related to crime)

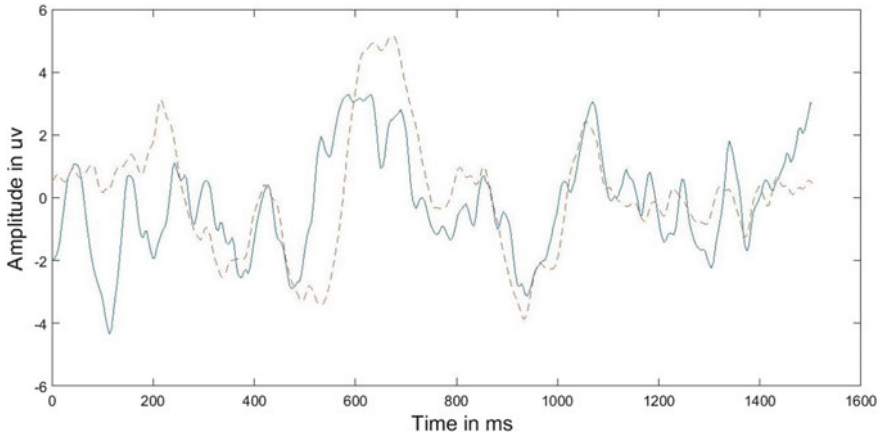


Fig. 3 Grand average ERP of person with crime information for Pz electrode (*dotted line* represents ERP for stimulus not related to crime, *solid line* represents ERP for stimulus related to crime)

component for person with and without crime knowledge were subtracted. Figures 2, 3, 4 and 5 represents comparison between subtracted probe and target response of both person with and without crime knowledge. More clear response can be seen from pz electrode. Figures 6, 7, 8 and 9 shows ERP image for probe and target stimulus for person with and without crime information. It is observed that ERP image of person with crime information has higher activation for probe stimulus as compared to target stimulus. It is also observed that ERP image of person without crime information showed equal activation for both stimulus.

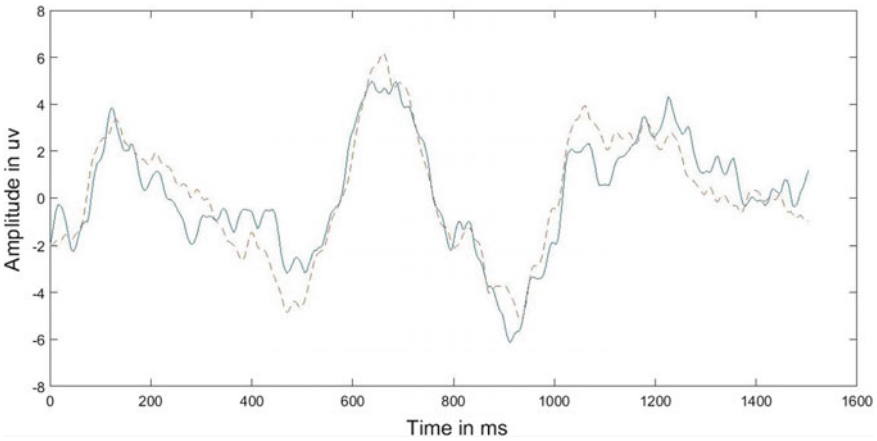


Fig. 4 Grand average ERP of person without crime information for Cz electrode (*dotted line* represents ERP for stimulus not related to crime, *solid line* represents ERP for stimulus related to crime)

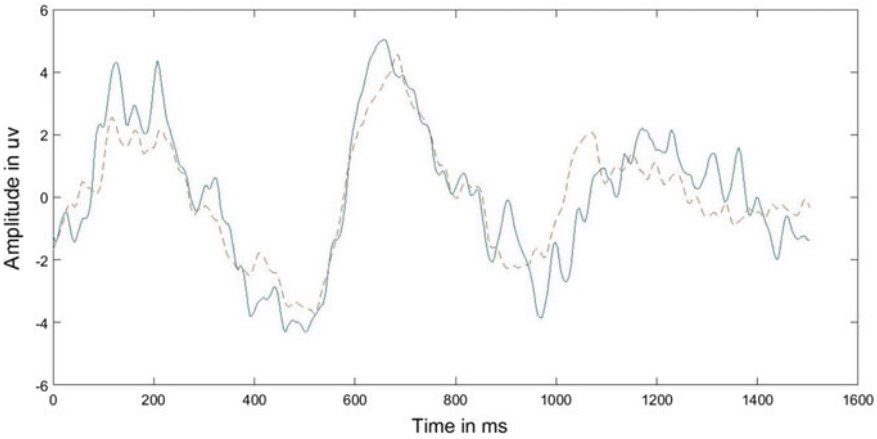


Fig. 5 Grand average ERP of person without crime information for Pz electrode (*dotted line* represents ERP for stimulus not related to crime, *solid line* represents ERP for stimulus related to crime)

The research work presented is unique in data collection method and has unique combination of existing technique to analyze the ERP signal for detection of validity of eye witness. Grey incidence degree based wavelet denoising method is more effective compared to wavelet denoising with universal threshold. Participants with crime knowledge had higher brain activation to crime related stimulus compared to participants without crime information. Data was classified into information present

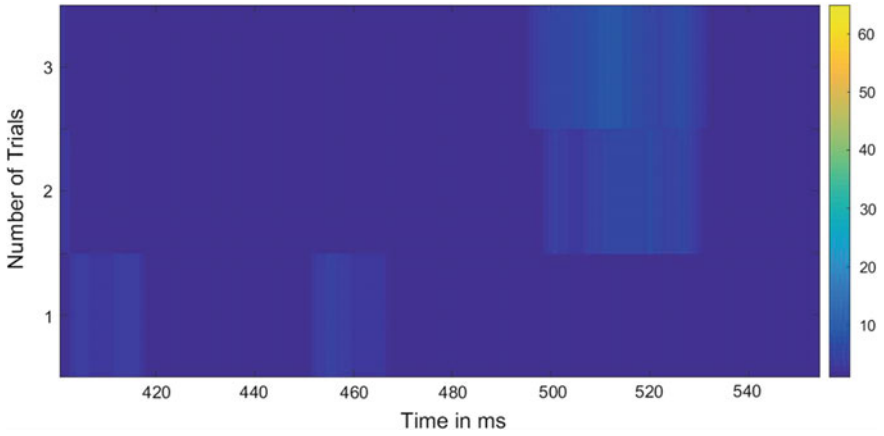


Fig. 6 ERP of person with crime information in image form for crime related stimulus (Pz electrode) (x axis- time, y axis- number of trials. Three trials are ERP responses to crime instrument, crime place and victim name, *color* represents amplitude variation)

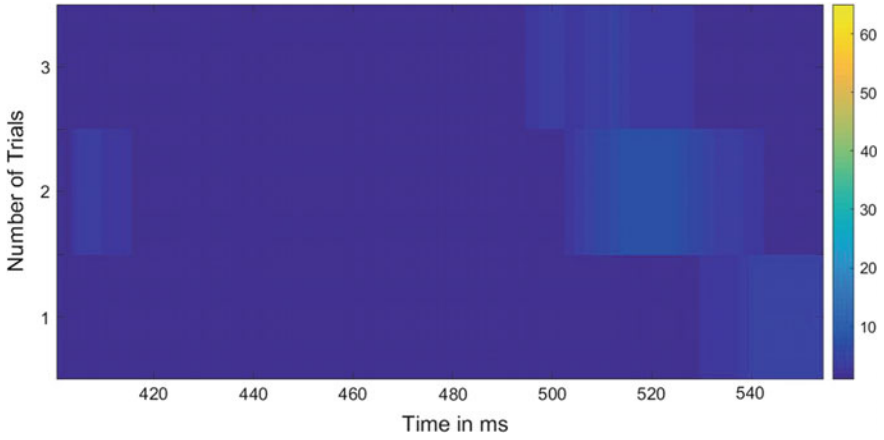


Fig. 7 ERP of person with crime information in image form for stimulus not related to crime (Pz electrode) (x axis- time, y axis- number of trials. Three trials are ERP responses to crime instrument, crime place and victim name, *color* represents amplitude variation)

and absent group by extracting Structural similarity index of ERP image resulted in maximum accuracy of 87.50 % which is significantly high in this type of research work (Table 1).

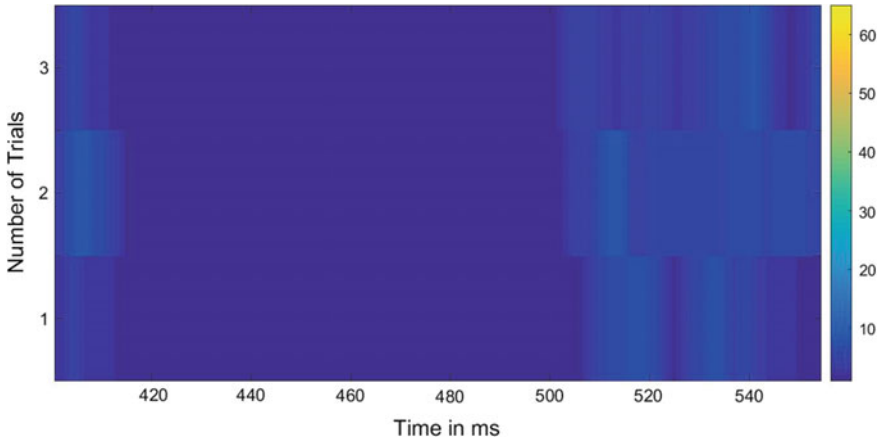


Fig. 8 ERP of person without crime information in image form for crime related stimulus (Pz electrode) (x axis- time, y axis- number of trials-Three trials are ERP responses to crime instrument, crime place and victim name, *color* represents amplitude variation)

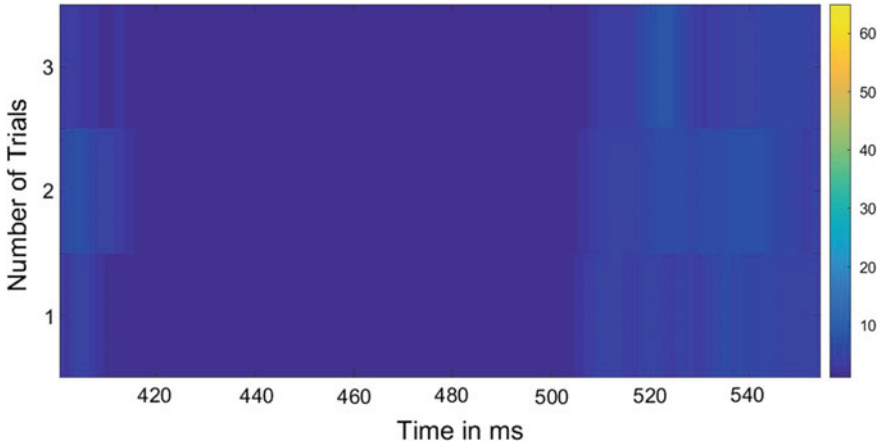


Fig. 9 ERP of person without crime information in image form for stimulus not related to crime (Pz electrode) (x axis- time, y axis- number of trials-Three trials are ERP responses to crime instrument, crime place and victim name, *color* represents amplitude variation)

Table 1 Classification performance of SVM classifier for different features

Sl.no	Features	Classification accuracy (%)
1	PSE, energy, average peak amplitude	64
2	Maximum coherence	52
3	Structural similarity index	87.50

5 Conclusion and Future Scope

In this paper a support vector machine classifier based crime information detection system is proposed. In preprocessing grey scale index based wavelet denoising is used which proved to be more efficient as compared to wavelet denoising with universal threshold. In feature extraction Structural similarity of ERP image was used which resulted in maximum accuracy. The results obtained are encouraging and can be extended further for criminal identification. In future more rigorous research work will be conducted taking number of subjects and testing probability based classifier like Hidden Markov Model [12, 13] in research work. ERP visualization of real time brain activation can give practical input of how the person is responding to each stimulus.

References

1. United Nations, World crime trends and emerging issues and responses in the field of crime prevention and criminal justice, Report, vol. 14, pp. 12–13, (2014)
2. Wang Zhiyu, Based on physiology parameters to design lie detector, International Conference on Computer Application and System Modeling, vol. 8, pp. 634–637, (2010)
3. Schwedes, Charlotte, and Dirk Wentura, The revealing glance: Eye gaze behavior to concealed information, Memory & cognition vol. 40.4, pp 642–651, (2012)
4. Rajoub, B.A., Zwiggelaar R., Thermal Facial Analysis for Deception Detection, IEEE Transactions on Information Forensics and Security, vol. 9, pp. 1015–1023, (2014)
5. Merzagora, Anna Caterina, et al. Wavelet analysis for EEG feature extraction in deception detection. Engineering in Medicine and Biology Society, (2006)
6. Abootalebi, Vahid, Mohammad Hassan Moradi, and Mohammad Ali Khalilzadeh. A new approach for EEG feature extraction in P300-based lie detection. Computer methods and programs in biomedicine, vol. 94.1, pp 48–57, (2009)
7. Lawrence A Farwell, Drew C Richardson, and Graham M Richardson, Brain fingerprinting field studies comparing p300-mermer and p300 brainwave responses in the detection of concealed information, Cognitive neurodynamics, vol. 7, pp. 263–299, (2013)
8. Wei Wen-chang, Cai Jian-li, and Yang Jun-jie, A new wavelet threshold method based on the grey incidence degree and its application, International Conference on Intelligent Networks and Intelligent Systems, pp. 577–580, (2008)
9. Ming-Jun Chen and Alan C Bovik, Fast structural similarity index algorithm, Journal of Real-Time Image Processing, vol. 6, no. 4, pp. 281–287, (2011)
10. A. Kumar and M.H. Kolekar, Machine learning approach for epileptic seizure detection using wavelet analysis of EEG signals, International Conference on Medical Imaging, m-Health and Emerging Communication Systems, pp. 412–416, (2014)
11. Maheshkumar H Kolekar, Deba Prasad Dash, A nonlinear feature based epileptic seizure detection using least square support vector machine classifier, IEEE Region 10 Conference, pp. 1–6, (2015)
12. Maheshkumar H. Kolekar, S. Sengupta, Semantic Indexing of News Video Sequences: A Multimodal Hierarchical Approach Based on Hidden Markov Model, IEEE Region 10 Conference, pp. 1–6, (2005)
13. Maheshkumar H Kolekar and Somnath Sengupta. Bayesian Network-Based Customized Highlight Generation for Broadcast Soccer Videos., IEEE Transactions on Broadcasting, vol. 61, no. 2, pp. 195–209, (2015)

View Invariant Motorcycle Detection for Helmet Wear Analysis in Intelligent Traffic Surveillance

M. Ashvini, G. Revathi, B. Yogameena and S. Saravanaperumaal

Abstract An important issue for intelligent traffic surveillance is automatic vehicle classification in traffic scene videos, which has great prospective for all kinds of security applications. Due to the number of vehicles in operation surpassed, occurrence of accidents is increasing. Hence, the vehicle classification is an important building block of surveillance systems that significantly impacts reliability of its applications. It helps in classifying the motorcycles that uses public transportation. This has been identified as an important task to conduct surveys on estimation of people wearing helmets, accident with and without helmet and vehicle tracking. The inability of police power in many countries to enforce helmet laws results in reduced usage of motorcycle helmets which becomes the reason for head injuries in case of accidents. This paper comes up with a system with view invariant using Histogram of Oriented Gradients which automatically detects motorcycle riders and determines whether they are wearing helmets or not.

Keywords Background subtraction · Histogram of Oriented Gradients (HOG) · Center-Symmetric Local Binary Pattern (CS-LBP) · K-Nearest Neighbor (KNN)

M. Ashvini (✉) · G. Revathi · B. Yogameena
Department of ECE, Thiagarajar College of Engineering, Madurai, India
e-mail: ashvinimano@gmail.com

G. Revathi
e-mail: rev.gsa@gmail.com

B. Yogameena
e-mail: b.yogameena@gmail.com

S. Saravanaperumaal
Department of Mechanical, Thiagarajar College of Engineering, Madurai, India
e-mail: sfpmech@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_16

1 Introduction

Recently, detecting and classifying moving objects from video sequences has become active research topics. They are used in various circumstances nowadays. Segmenting and classifying four moving objects such as bicycles, motorcycles, pedestrians and cars with view invariant in a video sequence is a challenging task. The object can be detected both in motion as well as in rest position depending on the application. Despite of its significance, classification of objects in wide scenario surveillance videos is challenging because of the following reasons. As the capability of conventional surveillance cameras [1] is limited, Region of Interest (ROI) in videos may be of low resolution. As a result, the information supplied by these regions is very limited. Also, the intra class variation for each category is very huge. Objects have diverse appearances and they may vary significantly because of lighting, different view angles and environments. The potential for object classification in real time application is great and so its performance has to be improved.

However, the above mentioned issues reduce the accurate working of object classification algorithms. Helmets are essential for motorcyclists' security from deadly accidents. The inability of police power in many countries to enforce helmet laws results in reduced usage of motorcycle helmets which becomes the reason for head injuries in case of accidents. The goal of this work is to develop an integrated and automated system approach for identifying motorcycle riders who are not wearing a helmet.

2 Related Work

Motorcycles have always been a very significant focus for traffic monitoring research which is based on computer vision. This requires some sort of camera calibration which can greatly affect the accurate working of a traffic monitoring system. Chiu and Ku et al. [2, 3] developed algorithms for detecting occluded motorcycles using the pixel ratio, visual width and visual length based on assumption that motorcycle riders always wear helmets. Anyway, these surveys do not focus on detecting helmets but used as a cue to identify a motorcycle. For the studies focusing on helmet detection, Liu et al. [4] proposed a technique to find a full-face helmet which used circle that fits on a Canny edge image. Wen et al. [5, 6] introduced similar techniques that detect helmets based on Circle Hough Transform. These techniques are used in surveillance systems in banks and at ATM machines. These algorithms are compatible with full-face helmets that have extractable circles and circular arcs. But, these papers do not focus on different view angles.

3 Proposed Method

3.1 Methodology

The first and foremost step of the system is to detect and to extract any moving object in a scene. This involves extracting shape features using Histogram of Oriented Gradients. By using K-Nearest Neighbor (KNN) classifier, the extracted object is classified as a motorcycle or other objects. Subsequently, with the help of background subtraction, foreground frame is extracted and the rider heads are extracted. The features are derived from it for further classification. Finally, KNN classifier is used which classifies whether the extracted head is wearing a helmet or not wearing a helmet. Features used here are circularity of the head region, average intensity and hues of each head quadrants. Figure 1 gives the overview of the proposed method.

3.2 Vehicle Detection

Vehicle classification is the process by which the vehicles are detected in the frame and are classified as object of interest. Vehicle identification can be done based on different parameters like shape, motion, color and texture. The color or texture based classification does not yield much information about the detected vehicle in traffic surveillance [7]. Hence, shape based classification are used for vehicle detection. After the vehicles are detected in a video sequence, the next step is to identify whether the vehicle is motorcycle or not.

3.3 Motorcycle Classification

Shape-based feature extraction. Different descriptors for information about shapes in motion regions such as point representation, blob and box are available for classifying the moving objects. Image and scene-based object parameters such

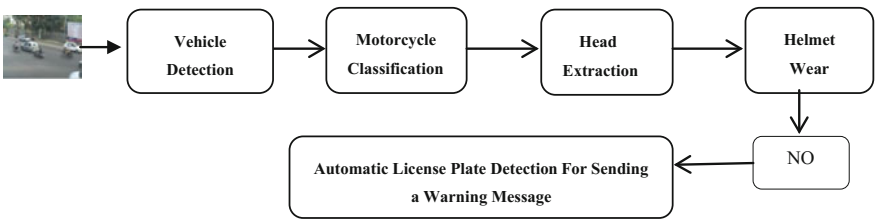


Fig. 1 Methodology of the proposed method

as apparent aspect ratio of the blob bounding box and image blob area are given as input features. Classification of objects for each blob is performed at every frame and results are stored in the form of histogram.

Feature extraction using HOG. To detect the shapes of objects in image processing and computer vision technique Histogram of oriented gradients (HOG) is a feature descriptor used [8]. The HOG descriptor technique counts the occurrences of gradient orientation in particular or localized portions of an image detection window. The HOG features are used to detect the objects like humans and vehicles. To capture the overall shape of the object it is used. For instance, in the below visualization of the features using HOG technique (Fig. 2), the outline of the motorcycle is prompt. The HOG is computed as follows: The magnitude of gradient is

$$|G| = \sqrt{I_x^2 + I_y^2} \tag{1}$$

The orientation of the gradient is given by

$$\theta = arctan \frac{I_x}{I_y} \tag{2}$$

where I_x, I_y, are image derivatives.

Motorcycle classification using classifier named K-Nearest Neighbor. For classification of motorcycle, K-Nearest Neighbor is used. Based on closest training examples, KNN method for classifying the objects is used. Euclidean distance is used and is given by

$$d(p, q) = \sqrt{\sum_i (p_i - q_i)^2} \tag{3}$$

where p_i and q_i are the nearest pixels. In this paper, HOG feature vector is taken as an input to the KNN classifier for further classification of the vehicle such as motorcycle or not.

Fig. 2 Extracted HOG features for an input frame

Head Extraction
Therefore, from the previous step, if it is detected as a motorcycle, the following procedure has to be followed. If the frame contains motorcycle, background subtraction is done to obtain the foreground which is exactly motorcycle for further analysis. The region of interest, here, the motorcyclist's head portion will be detected immediately. Followed that, the features are used to classify whether the head portion wears helmet or not. Again, KNN classifier is used for this classification.

Background subtraction and morphological operation. Video sequence may be separated into background and foreground. If the background data is removed from the video frame, then the necessary data left out is considered as foreground which contains the object of interest. Better accuracy can be achieved, if the background is already known. For example, in stationary surveillance cameras as in road traffic monitoring, the background is always constant. The road remains in the same position with respect to the camera. The background subtraction method is subtracting the current from the background frame which helps to detect moving objects with simple algorithm. The foreground extracted image will have certain unnecessary information because of shadows and illumination changes. So, the morphological operations (opening, closing) are performed to remove these changes. Closing operations performs the enlargement of boundaries of the foreground (bright) regions in the image and shrinks the background color holes in such regions. The opening operation is performed to remove foreground pixels which occurs due to the illumination changes.

3.4 Helmet Wear or not Detection by KNN Classifier

The heads of motorcyclist are in the upper part of motorcycle blob. Hence, the Region of Interest (ROI) for extracting the heads of motorcyclist is at the top 25 % of the height of a motorcycle blob. The totals of 4 different features are derived from the four quadrants of head region.

1. Feature 1: Arc circularity
2. Feature 2: Average intensity
3. Feature 3: Average hue
4. Feature 4: Texture feature extraction

Arc circularity. The similarity measures between arc and a circle is given by

$$c = \frac{\mu_r}{\sigma_r} \tag{4}$$

where σ_r, μ_r is the standard deviation and mean of the distance r from the head centroid to the head contour. These features are extracted because the head portion which contains a helmet is more circular than a head without a helmet, which reflects in high circularity of head contour.

Average intensity. The feature to be extracted is average intensities. It is denoted as μ, computed individually from a gray scale image as

$$\mu_I = \frac{1}{N}\sum_{i=0}^{N-1} I_i \tag{5}$$

where I_i is an intensity of the *i*th pixel, N is the pixel count in the head. These features are employed since the intensity on the top and the back of the head without helmet are mostly dark. Here, the assumption is made according to Indian scenario. These features are normalized with the help of maximum gray scale intensity.

Average hue of Head portion. Average hue of face is another important feature which is computed exclusively by:

$$\mu_H = \frac{1}{N}\sum_{i=0}^{N-1} H_i \tag{6}$$

where H_i is the hue of *i*th pixel and N is the pixel count in the head. These features are applied because a large portion of his/her face is covered by their helmet and it also varies with the average hue value. Additionally, a rider without a helmet has certain average hue of skin color.

Texture feature extraction by Centre Symmetric-Local Binary Pattern (CS-LBP). The Centre Symmetric Local Binary Patterns (CS-LBP) are devised which compares the center-symmetric pairs of pixels. This reduces the number of comparisons for the similar number of neighbors. For 8 neighbors, only 16 different binary patterns are produced by CS-LBP. So, CS-LBP is used as a texture feature for helmet detection even under different illumination changes [9].

$$CS-LBP_{P,R}(\mathrm{c}) = \sum_{i=0}^{\left(\frac{P}{2}\right)-1} s\left(gi - gi + \left(\frac{P}{2}\right)\right) 2^i \tag{7}$$

where gi and $gi + \left(\frac{P}{2}\right)$ represents the gray values pixels of center-symmetric pairs P which is equally spaced on a circle of radius R.

Classification using K-Nearest Neighbor classifier. Here the head is classified based on the majority vote of its neighbors either the motorcyclists are "wearing a helmet" or "not wearing". These neighbors are taken from the head part where the correct classification is known and labeled. For helmet classification, the Standard deviation, hue, average intensity and CS-LBP texture features are calculated. These features are given as an input to the KNN classifier. At last, the classifier output displays as 'helmet detected' or 'no helmet'. This will help to warn the particular motorcyclist who does not wear helmet or to take a survey on motorcyclists with/without helmet for the authorities. Thus, it may help to reduce deadly accidents due to this issue.

4 Results and Experiments

The proposed system involves motorcycle detection and helmet wear or not classification. These experiments are tested separately where the results of each test are independent and also there is no propagation of error from previous algorithm. The proposed system performed KNN classifications with approximately 100 training frames and 20 testing frames. Each experiment has different view angle and resolution in Table 1.

The sample frames of benchmark datasets are shown in Fig. 3.

The first step of the proposed method is to extract HOG features from the given current image which is shown in Fig. 4.

Subsequently, the extracted HOG feature vectors are given as an input to the KNN classifier to detect and classify whether it is a motorcycle or not in Fig. 5.

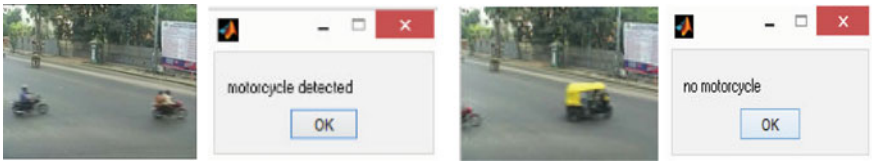
Table 1 Bench mark datasets with specifications

Datasets	Mirpur	TCE	IIT	Bangalore	IISC
Year	2010	2014	2015	2015	2015
Frames total	1309	20,641	1119	2624	200
Resolution	640 × 360	704 × 288	360 × 238	704 × 576	704 × 576
Place (outdoor)	Outdoor	Outdoor	Outdoor	Outdoor	Outdoor
View angle	Back view	Front view	Side view	Side view	Side view
Frame type	.jpg	.jpg	.jpg	.jpg	.jpg
Frames \seconds (fps)	25	24	30	25	30

Fig. 3 The sample frames of benchmark datasets. **a** TCE dataset. **b** IIT dataset. **c** Mirpur dataset. **d** IISC. **e** Bangalore dataset

Original frame (Frame no 38) HOG extracted frame (Frame no 38)

Fig. 4 Output of HOG feature extraction



Original frame (Frame no 38) Original frame (Frame no 1)

Fig. 5 Motorcycle detection and verification using KNN classifier

Background frame Current frame (Frame no 38) Foreground extraction

Fig. 6 Foreground extraction using background subtraction

Fig. 7 Morphological closing and opening

Morphological closing (Frame no 38) Morphological opening (Frame no 38)

Helmet wear or not Detection. Followed that, the foreground image is extracted using background subtraction. The morphological operation (closing, opening) is performed to enlarge the foreground pixels in Figs. 6 and 7.

The head region is extracted. Subsequently, the Standard deviation, hue, average Intensity and CSLBP texture features are extracted and are given as a input to the KNN classifier. It detects whether the motorcyclist wears helmet or not shown in Fig. 8.

TCE DATASET

Helmet detection in TCE dataset is shown in Fig. 9. The features extracted from TCE dataset is listed in Tables 2 and 3.

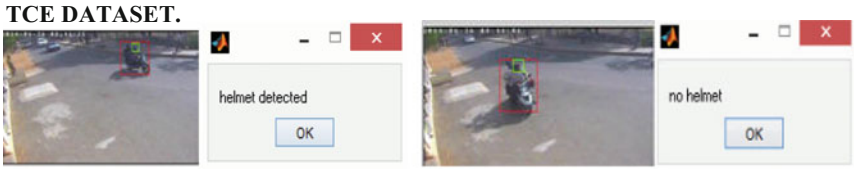
To differentiate the motorcycles and bicycles is the challenging task because they have similar features/characteristics. SVM [10] along with LBP proved to be robust. Compared to other features it is found from the Table 4 that SVM with LBP gives the better accuracy.

The accuracy of the motorcycle recognition algorithm is given by 95 % and the detail is shown in the below confusion matrix in Table 5.

Head extraction (Frame no 38) IISC dataset

Helmet detection using KNN Side view

Fig. 8 Helmet classification using KNN classifier for IISC dataset



Frame no 756 Front view

Frame no 272 Front view

Fig. 9 Motorcyclist wears helmet or not classification using KNN classifier for TCE dataset

Table 2 Features for various frames with helmet

TCE dataset frames	Arc circularity	Average intensity	Average hue	Texture feature using CS-LBP
FRAME 753	0.0206	63.5074	0.5466	$3.145e^{04}$
FRAME 2770	0.0060	59.1046	0.5140	$3.1289e^{04}$
FRAME 3614	0.0030	92.5694	0.4054	$3.1304e^{04}$
FRAME 3616	0.0138	65.7077	0.5102	$3.1127e^{04}$
FRAME 4094	0.2346	64.6910	0.3520	$3.127e^{04}$

Table 3 Features for various frames without helmet

TCE dataset frames	Arc circularity	Average intensity	Average hue	Texture feature using CS-LBP
FRAME 254	0.0031	102.7370	0.7049	$3.1471e^{04}$
FRAME 435	0.0282	98.5443	0.6401	$3.138e^{04}$
FRAME 2015	0.1649	97.7045	0.4791	$3.137e^{04}$
FRAME 2394	0.0030	96.5230	0.4054	$3.134e^{04}$
FRAME 3894	0.0361	77.3631	0.6062	$3.142e^{04}$

Table 4 Accuracy of features obtained using SVM classifier

Image database	Accuracy
HAAR	0.9226
HOG	0.9482
LBP	0.9763
SURF	0.9719

Table 5 Confusion matrix for detection of motorcycle

Original	Predicted	Class
	Motorcycle (%)	Others
Motorcycle	95	9
Others	3	96

Table 6 Confusion matrix for detection of helmet

Actual	Predicted	Class
	With helmet (%)	Without helmet (%)
With helmet	89	9
Other	10	87

The classification of helmet algorithm with manually cropped head images as inputs is 89 %. The results are shown in Table 6.

From the confusion matrix, the accuracy of motorcycle classification and helmet detection is inferred.

5 Conclusion

From the survey of various shape based feature extraction algorithms which is robust to different view angles; it is found that HOG descriptor provides better results when compared with other algorithms such as SURF, SIFT, LBP and RIFT. The proposed method makes use of HOG descriptor with KNN classifier for motor cycle classification under challenging environment (different view angles) in intelligent traffic surveillance system. After motorcycle detection, head region is detected with the help of the features such as Arc circularity, Average intensity, Average Hue and CS-LBP texture features. Finally, these features are used to detect the motorcyclist wears helmet or not even under different illumination changes which are usual in real time. The proposed algorithm helps to segment motorcycles on public roads which act as an important task. This can be used to motivate the two wheeler riders to wear helmets via providing awareness. This can estimate accident with and without helmet, speed computation and vehicle tracking. The future work is that average intensity feature can be extracted for motorcycle rider not wearing helmet with white hair or bald as intensity varies. The proposed method can be extended further for automatic license plate detection for sending a warning message when the motorcyclist without helmet is detected.

References

1. Zhaoxiang Zhang, Yunhong Wanga: Automatic object classification using motion blob based local feature fusion for traffic scene surveillance. In: International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, Issue. 5, pp. 537–546. SP Higher Education Press (2012).
2. C.C. Chiu, M.Y. Ku, and H.T. Chen: Motorcycle Detection and tracking system with occlusion segmentation. In: WIAMIS'07 Proceedings of the Eight International Workshop on Image Analysis for Multimedia Interactive Services, pp. 32. IEEE Computer Society Washington, DC, USA (2007).
3. M.Y. Ku, C.C. Chin, H.T. Chen and S.H. Hong: Visual Motorcycle Detection and Tracking Algorithm. In: WSEAS Trans. Electron, vol. 5, Issue. 4, pp. 121–131, IEEE (2008).
4. C.C. Liu, J.S. Liao, W.Y. Chen and J.H. Chen: The Full Motorcycle Helmet Detection Scheme Using Canny Detection. In: 18th IPPR Conf, pp. 1104–1110. CVGIP (2005).
5. C.Y. Wen, S.H. Chiu, J.J. Liaw and C.P. Lu: The Safety Helmet Detection for ATM's Surveillance System via the Modified Hough transform. In: Security Technology, IEEE 37th Annual International Carnahan Conference, pp. 364–369. IEEE (2003).
6. C.Y. Wen: The Safety Helmet Detection Technology and its Application to the Surveillance System. In: Journal of Forensic Sciences, Vol. 49, Issue. 4, pp. 770–780. ASTM international, USA (2004).
7. Damian Ellwart, Andrzej Czysewski: Viewpoint independent shape-based object classification for video surveliance. In: 12th International Workshop on Image Analysis for Multimedia Interactive Services, TU Delft; EWI; MM; PRB, Delft, The Netherlands (2011).
8. Chung-Wei Liang, Chia-Feng Juang: Moving object classification using local shape and HOG features inwavelet-transformed space with hierarchical SVM classifiers. In: Applied Soft Computing, vol. 28, Issue. C, pp. 483–497. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands (2015).
9. Zhaoxiang Zhang, KaiqiHuang, B., YunhongWanga, MinLi: View independent object classification by exploring scene consistency information for traffic scene surveillance. In: Journal Neurocomputing, Vol. 99, pp. 250–260. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands (2013).
10. C. Tangnoi, N. Bundon, V. Timtong, and R. Waranusast: A Motorcycle safety helmet detection system using svm classifier. In: IET Intelligent Transport System, vol. 6, Issue. 3, pp. 259–269. IET (2012).

Morphological Geodesic Active Contour Based Automatic Aorta Segmentation in Thoracic CT Images

Avijit Dasgupta, Sudipta Mukhopadhyay, Shrikant A. Mehre and Parthasarathi Bhattacharyya

Abstract Automatic aorta segmentation and quantification in thoracic computed tomography (CT) images is important for detection and prevention of aortic diseases. This paper proposes an automatic aorta segmentation algorithm in both contrast and non-contrast CT images of thorax. The proposed algorithm first detects the slice containing the carina region. Circular Hough Transform (CHT) is applied on the detected slice to localize ascending and descending aorta (circles with lowest variances) followed by a morphological geodesic active contour to segment the aorta from CT stack. The dice similarity coefficients (DSC) between the ground truth and the segmented output were found to be 0.8845 ± 0.0584 on LIDC-IDRI dataset.

Keywords Cardiovascular diseases · Computed tomography · Aorta segmentation · Computer-based automated segmentation · Active contour

1 Introduction

Cardiovascular diseases (CVDs) were the primary cause for death of around 788,000 people in 2010 in western countries [1]. Moreover, cardiovascular disease related deaths in eastern countries are growing at an alarming rate [2]. Aortic abnormalities

A. Dasgupta (✉) · S. Mukhopadhyay · S.A. Mehre
Computer Vision and Image Processing Laboratory,
Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology Kharagpur, West Bengal 721302, India
e-mail: avijit.dasgupta@iitkgp.ac.in

S. Mukhopadhyay
e-mail: smukho@ece.iitkgp.ernet.in

S.A. Mehre
e-mail: shrikant.mehre@gmail.com

P. Bhattacharyya
Institute of Pulmocare & Research, Kolkata 700156, West Bengal, India
e-mail: parthachest@yahoo.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_17

such as calcification, dissection etc., are the most common cardiovascular diseases. Thus, the detection and analysis of aorta is of medical importance. At present, the available imaging modalities for manifestation of CVDs are lung computed tomography, cardiac computed tomography, magnetic resonance (MR) etc. Aortic abberivations can be identified in the thoracic CT image which is the widely used non-invasive imaging technique. The manual annotation and assessment of those CT images could be tedious and inherently inaccurate even for highly trained professionals. To obviate such difficulties, an automated aorta quantification system is of utmost importance which requires accurate automatic aorta localization and segmentation.

Automated assessment of aorta has been reported and evaluated on both contrast-enhanced and non-contrast-enhanced cardiac CT and MR images [3–10]. Multiple atlas based aorta segmentation for low-dose non-contrast CT has been proposed by Ivsgum et al. [7]. However, this method uses multiple registrations of images which are manually labelled to get the final segmented output. Kurkure et al. [4] first proposed an automated technique to localize and segment aorta from cardiac CT images using dynamic programming. The authors of [4] formulated an entropy based cost function in [5] for improved automatic segmentation of aorta. An automatic aorta detection in non-contrast cardiac CT images using bayesian tracking algorithm has been proposed by Zheng et al. [9]. Kurugol et al. [10] first reported an aorta segmentation algorithm in thoracic CT images using 3D level set approach. Xie et al. [8] reported an automated aorta segmentation in low-dose thoracic CT image which makes use of pre-computed anatomy label maps (ALM). However, the ALM may not be always available with CT images.

Inspired by the works done previously towards the automation of aorta quantification, in this paper we propose an automated active contour based two stage approach for aorta segmentation in CT images of thorax. In the first stage, a suitable slice is chosen automatically to find the seed points for active contour. It is experimentally found that the slice in which trachea bifurcates, aorta (both ascending and descending) takes nearly circular shape. So the slice in which trachea bifurcation occurs is detected using image processing and taken as the suitable slice to localize aorta. After the suitable slice is chosen, two seed points (center of two circles) among the circles detected by CHT having lowest variances are selected automatically as the ascending aorta and descending aorta. In the second stage, the aortic surface is determined by upward and downward segmentation of ascending and descending aorta. This segmentation algorithm builds upon morphological geodesic active contour [11, 12].

The key contributions of the proposed algorithm are the following: a fully automatic algorithm for locating and segmenting the aortic surface using morphological geodesic active contour. The proposed algorithm can be seamlessly applied to the contrast-enhanced as well as non-contrast enhanced thoracic CT images. Unlike other methods [4, 5], the algorithm proposed in this paper does not need any prior knowledge of the span of thoracic CT slices to be processed. The proposed algorithm automatically finds and segments the start and end of aorta from thoracic CT images. Results produced by the proposed algorithm is compared with annotations prepared by experts for quantitative validations.

The rest of the paper is organized as follows: Sect. 2 presents the detailed description of the proposed technique. Section 3 provides quantitative and qualitative results and finally Sect. 4 concludes the paper.

2 Methodology

The input to the proposed algorithm is a 3D thoracic CT volume $I \in \mathbb{R}^{m \times n \times p}$. Note that, the algorithm does not need any assumption regarding the span of the thoracic CT volume for aorta segmentation. Figure 1 shows the block diagram of the proposed algorithm.

2.1 Localization of Aorta from 3D CT Volume

In order to localize the ascending and descending aorta, a suitable slice from the axial 3D CT volume I needs to be determined where the ascending and descending aorta are well defined in terms of its geometry. It has been analytically found that in human anatomy the axial slice in which the trachea bifurcates (carina), the ascending and descending aorta are almost circular in shape. However, accurate detection of trachea bifurcation is not needed. A margin of ± 2 slices does not incur any loss in performance.

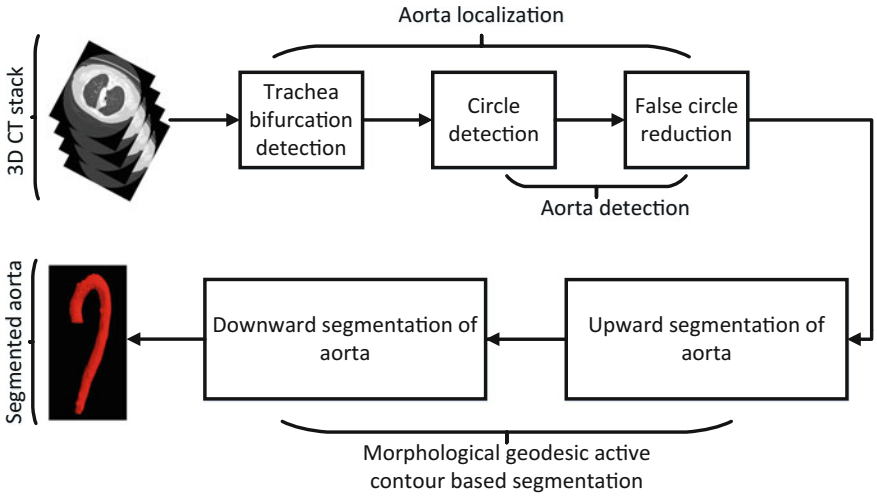


Fig. 1 The block diagram of the proposed automated aorta segmentation method

Trachea Bifurcation Detection

In order to detect the carina location, the CT image stack is first median filtered with a window of size 3×3. Then the stack is thresholded at -700 HU which retains the lower intensity air filled regions (including surrounding air) in the CT image stack. A morphological binary erosion operation has been done with a disk type structuring element of radius 1 pixel on the thresholded CT stack I. The main purpose of this operation is to remove all the undesirable small regions present in the CT slices. In order to remove the surrounding air and detect bifurcation, connected component analysis is done on the binary eroded CT stack and trachea region is extracted from the labelled connected components. Let the preprocessed CT stack now is represented by $\hat{I} \in \mathbb{Z}_2^{m \times n \times p}$.

Once the preprocessing is done, trachea needs to be located. Regions with area between 100 and 400 pixels and circularity ($\frac{Perimeter^2}{4 \times \pi \times Area}$) between 1 and 2 are considered to be the trachea in preprocessed stack $\hat{I}$. Once the initial location of the trachea is detected, it is tracked using the centroid in next subsequent slices. This method is applied progressively until either the circularity becomes greater than 2 or the connectivity is lost.

Circle Detection Using CHT and False Positive Reduction

Once the slice containing the carina region is detected Canny edge detector [13] is applied followed by the CHT [14] for circle detection. Despite being a robust and powerful method, the CHT often suffers from noise inherently present in the CT images, as a result it produces false object boundaries along with the true circular objects. To remove false objects, we construct circles with radius 8–12 pixels using the centers detected by the CHT and choose two circles having lowest variances which are considered to be the two seed points of ascending and descending aorta. We construct two circles of radius 12 pixels using these two seed points which act as the initial contour for Morphological Geodesic Active Contour (MGAC) to segment ascending and descending aorta from each of the slices of the CT stack.

2.2 Morphological Geodesic Active Contour Based Aorta Segmentation

Active contour based segmentation methods are being used in medical image processing research for years now. Geodesic active contour (GAC) is one of the most popular contour evolution methods [15, 16]. GAC tries to separate foreground (object) and background with the help of image intensity and gradient. GAC solves a partial differential equation (PDE) to evolve the curve towards the object boundary. Let $u : \mathbb{R}^+ \times \mathbb{R}^2 \rightarrow \mathbb{R}$ be an implicit representation of C such that

$C(t) = \{(x, y)|u(t, (x, y)) = 0\}$. The curve evolution equation of GAC can be represented in implicit form as

$$\frac{\partial u}{\partial t} = g(I)|\nabla u|\Big(\nu + div\big(\frac{\nabla u}{|\nabla u|}\big)\Big) + \nabla g(I).\nabla u, \tag{1}$$

where, ν is the balloon force parameter, $div\big(\frac{\nabla u}{|\nabla u|}\big)$ is the curvature of the curve and the stopping function $g(I)$ is defined as follows: $g(I) = \frac{1}{\sqrt{1+\alpha|\nabla G * I|}}$. Typically the value α is set to 0.15. It attains minima at the boundary of the object, thus, reducing the velocity of the curve evolution near the border.

The GAC contour evolution equation comprises of three forces: (a) Balloon force, (b) Smoothing force and (c) Attraction force. However, solving PDEs involves computationally expensive numerical algorithms.

In this paper morphological operators are used to solve the PDE of GAC as proposed in [11, 12]. Let the contour at nth iteration is represented by $u^n(x)$. The balloon force $(g(I)|\nabla u|\nu)$ can be solved using a threshold θ, binary erosion (E_h) and dilation (D_h) operations for $(n + 1)$th iteration as

$$u^{n+\frac{1}{3}}(x) = \begin{cases} (D_h u^n)(x), & \text{if } g(I)(x) > \theta \text{ and } \nu > 0, \\ (E_h u^n)(x), & \text{if } g(I)(x) > \theta \text{ and } \nu < 0, \\ u^n(x), & \text{otherwise.} \end{cases} \tag{2}$$

The attraction force $(\nabla g(I).\nabla u)$ can be solved very easily from intuition. The main purpose of attraction force is to attract the curve C towards the edges. Mathematically we can discretize this force as

$$u^{n+\frac{2}{3}}(x) = \begin{cases} 1, & \text{if } \nabla u^{n+\frac{1}{3}} \nabla g(I)(x) > 0, \\ 0, & \text{if } \nabla u^{n+\frac{1}{3}} \nabla g(I)(x) < 0, \\ u^{n+\frac{1}{3}}(x), & \text{otherwise.} \end{cases} \tag{3}$$

In order to solve the smoothing term $(g(I)|\nabla u|div\big(\frac{\nabla u}{|\nabla u|}\big))$ Alvarez et al. [11, 12] defined two morphological operators, *sup-inf* (SI_h) and *inf-sup* (IS_h). In binary images, both SI_h and IS_h operators look for small straight lines (3 pixels long) in four possible directions (see Fig. 2). If no straight line is found, the pixel is made inactive and active respectively. The difference between SI_h and IS_h is that, the first one operates on active pixels (i.e. pixels having values 1) and the second one operates on inactive pixels (i.e. pixels having values 0).

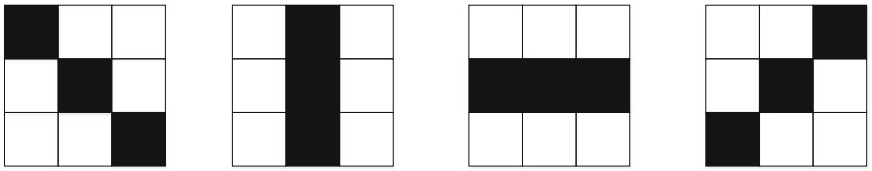


Fig. 2 The structuring elements B for the 2D discrete operator $SI_h \circ IS_h$

It can be proved that, the mean curvature ($div(\frac{\nabla u}{|\nabla u|})$) can be obtained using the composition of these operators ($SI_h \circ IS_h$). So, the smoothing force with smoothing constant μ can be written as

$$u^{n+1}(x) = \begin{cases} ((SI_h \circ IS_h)^{\mu} u^{n+\frac{2}{3}})(x), & \text{if } g(I)(x) > \theta, \\ u^{n+\frac{2}{3}}(x), & \text{otherwise.} \end{cases} \tag{4}$$

This morphological geodesic active contour is applied which is a composition of the three forces (balloon force, attraction force followed by smoothing force) to each of the slices of the CT volume.

First, the proposed algorithm segments ascending and descending aorta in upward direction from carina region followed by downward direction. The algorithm stops automatically if either the area of aorta changes significantly with respect to the previous slice (2 times for upward segmentation and 1.5 times for downward segmentation) or the difference between mean intensities of the segmented regions of consecutive slices is greater than 50 HU for upward direction only.

3 Results and Discussion

The proposed algorithm was applied on 30 (26 contrast enhanced and 4 non-contrast enhanced) randomly selected cases taken from the widely used LIDC-IDRI public dataset [17]. On an average the dataset contains 187 slices per CT scan of 512×512 resolution with a spacing of $0.5469 - 0.8828$ mm in x, y direction and $0.6250 - 2.5$ mm in z direction. To make CT data isotropic in all directions (x, y, z) each CT stack was resampled before further processing as suggested by [18].

The proposed methodology was evaluated by following the same technique as described in [8]. Each of the thirty cases have 26 images manually annotated (5 images for ascending aorta, 3 images for aortic arch, 10 images for descending aorta and 8 images for all three parts). In total, the proposed methodology was evaluated on 780 images as compared to [8] which was evaluated on 630 images.

It was observed from the data, that the images were acquired mainly using two types of CT machines—(a) GE Light Speed Plus and (b) GE Light Speed 16. For the first case the value of θ of Eq. 2 was chosen as 50th percentile of $g(I)$ for upward

ascending aorta segmentation and 55th percentile of $g(I)$ for downward ascending aorta and whole descending aorta (both in upward and downward direction). The values of standard deviation, lower threshold and higher threshold in Canny edge detection algorithm are 0.5, 100 and 400 respectively. For second case, 45th percentile of $g(I)$ for upward segmentation and 55th percentile of $g(I)$ for downward segmentation of both ascending and descending aorta as the value of parameter θ. The values of standard deviation, lower threshold and higher threshold in Canny edge detection algorithm are 0.2, 200 and 400 respectively. The values of smoothing parameter μ and the balloon force parameter ν were set to 2 and 1 respectively for all experiments.

Figure 3 shows the result of each steps involved in localization of ascending and descending aorta which are marked in green and red respectively in Fig. 3c. Figure 4 shows 3D visualization of two correctly segmented aorta as well as one partially segmented aorta where ascending aorta could not be segmented near the heart region.

The quality of the segmentation was evaluated in terms of Dice Similarity Coefficient (DSC). The DSC is defined as $\frac{2|GT \cap S|}{|GT|+|S|}$, where, GT and S represent the groundtruth image and segmented image respectively and $|A|$ denotes the total number of active pixels in an image A.

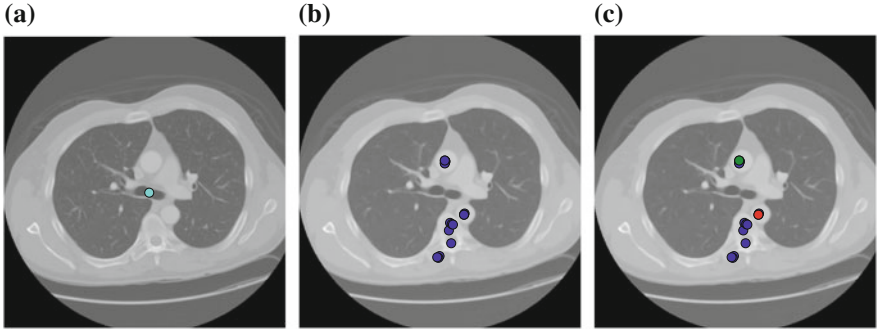


Fig. 3 Aorta localization: **a** Trachea bifurcation detection, **b** Circle detection using CHT, **c** Detected ascending and descending aorta (*green* and *red point*) after false circle reduction

Fig. 4 3D visualization of accurately segmented (*left and middle*) aorta and an aorta (*right*) in which segmentation stopped early in heart region

Table 1 Quantitative evaluation of our proposed algorithm in terms of average of dice similarity coefficient (DSC) for (a) Whole aorta (b) Ascending aorta (c) Descending aorta (d) Aortic arch

Statistics	DSC			
	Whole aorta	Ascending aorta	Descending aorta	Aortic arch
Mean	0.8845	0.8926	0.9141	0.8327
Std. dev. σ	0.0584	0.0639	0.0223	0.1192

Table 1 shows the quantitative results of the algorithm tested on 30 cases. By visual inspection, in some of the cases the proposed algorithm were inaccurate due to the heart region where the aorta is occluded by the organs having similar intensity values.

Note that, [5, 6, 10] used their private datasets. Although Xie et al. [8] used randomly sampled public databases (LIDC-IDRI [17], VIA-ELCAP [19]), results of the proposed algorithm could not be compared with them as the exact case ids are not known and the groundtruths are not publicly available.

4 Conclusion

A novel fully automated aorta segmentation algorithm has been developed for analyzing the aorta from thoracic CT images. The algorithm proposed in this paper does not need any prior information regarding the span of the CT scan. Aorta can be localized and segmented without any user intervention. It employs CHT on the slice in which trachea bifurcates (carina region) to localize circular regions. CHT generates many false positives from which two circles having lowest variances have been considered as the ascending and descending aorta.

The algorithm was tested on 30 randomly sampled cases from LIDC-IDRI dataset. In some cases the proposed algorithm fails to stop segmenting aorta near heart region due to adjacent regions with similar intensities. More work will be needed to develop an algorithm to address this issue. Future work should also involve to test the algorithm on a large number of test cases and release ground truths for benchmarking aorta segmentation.

References

1. National Heart Lung and Blood Institute, "Disease statistics," in *NHLBI Fact Book, Fiscal Year 2012*. NHLBI, 2012, p. 35.
2. R Gupta, P Joshi, V Mohan, KS Reddy, and S Yusuf, "Epidemiology and causation of coronary heart disease and stroke in India," *Heart*, vol. 94, no. 1, pp. 16–26, 2008.

3. Shengjun Wang, Ling Fu, Yong Yue, Yan Kang, and Jiren Liu, "Fast and automatic segmentation of ascending aorta in msct volume data," in *2nd International Congress on Image and Signal Processing, 2009. CISP'09.* IEEE, 2009, pp. 1–5.
4. Uday Kurkure, Olga C Avila Montes, Ioannis Kakadiaris, et al., "Automated segmentation of thoracic aorta in non-contrast ct images," in *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI*. IEEE, 2008, pp. 29–32.
5. Olga C Avila-Montes, Uday Kukure, and Ioannis A Kakadiaris, "Aorta segmentation in non-contrast cardiac ct images using an entropy-based cost function," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2010, pp. 76233J–76233J.
6. Olga C Avila-Montes, Uday Kurkure, Ryo Nakazato, Daniel S Berman, Debabrata Dey, Ioannis Kakadiaris, et al., "Segmentation of the thoracic aorta in noncontrast cardiac ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 5, pp. 936–949, 2013.
7. Ivana Išgum, Marius Staring, Annemarieke Rutten, Mathias Prokop, Max Viergever, Bram Van Ginneken, et al., "Multi-atlas-based segmentation with local decision fusionapplication to cardiac and aortic segmentation in ct scans," *IEEE Transactions on Medical Imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.
8. Yiting Xie, Jennifer Padgett, Alberto M Biancardi, and Anthony P Reeves, "Automated aorta segmentation in low-dose chest ct images," *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 211–219, 2014.
9. Mingna Zheng, J Jeffery Carr, and Yaorong Ge, "Automatic aorta detection in non-contrast 3d cardiac ct images using bayesian tracking method," in *Medical Computer Vision. Large Data in Medical Imaging*, pp. 130–137. Springer, 2014.
10. Sila Kurugol, Raul San Jose Estepar, James Ross, and George R Washko, "Aorta segmentation with a 3d level set approach and quantification of aortic calcifications in non-contrast chest ct," in *Annual International Conference of the IEEE on Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2012, pp. 2343–2346.
11. L. Alvarez, L. Baumela, P. Henriquez, and P. Marquez-Neila, "Morphological snakes," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 2197–2202.
12. Pablo Marquez-Neila, Luis Baumela, and Luis Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 2–17, 2014.
13. John Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.
14. Dana H Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.
15. Vicent Caselles, Ron Kimmel, and Guillermo Sapiro, "Geodesic active contours," in *Fifth International Conference on Computer Vision,*. IEEE, 1995, pp. 694–699.
16. Vicent Caselles, Ron Kimmel, and Guillermo Sapiro, "Geodesic active contours," *International journal of computer vision*, vol. 22, no. 1, pp. 61–79, 1997.
17. Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al., "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.
18. William J Kostis, Anthony P Reeves, David F Yankelevitz, Claudia Henschke, et al., "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical ct images," *IEEE Transactions on Medical Imaging*, vol. 22, no. 10, pp. 1259–1274, 2003.
19. ELCAP Public Lung Image Database, http://www.via.cornell.edu/lungdb.html

Surveillance Video Synopsis While Preserving Object Motion Structure and Interaction

Tapas Badal, Neeta Nain and Mushtaq Ahmed

Abstract With the rapid growth of surveillance cameras and sensors, a need of smart video analysis and monitoring system is gradually increasing for browsing and storing a large amount of data. Traditional video analysis methods generate a summary of day long videos but maintaining the motion structure and interaction between object is of great concern to researchers. This paper presents an approach to produce video synopsis while preserving motion structure and object interactions. While condensing video, object appearance over spatial domain is maintained by considering its weight that preserve important activity portion and condense data related to regular events. The approach is tested in the context of condensation ratio while maintaining the interaction between objects. Experimental results over three video sequences show high condensation rate up to 11 %.

Keywords Video analysis · Synopsis · Activity analysis · Object detection · Motion structure

1 Introduction

These days the enormous amount of cameras are installed around the world and the information produced by these devices are abundant enough for humans to extract knowledge present in videos. A lot of data mining effort is needed to process surveillance videos for browsing and retrieval of a specific event. Video synopsis condenses video by showing activities simultaneously that happened at the different time in a

T. Badal (✉) · N. Nain · M. Ahmed
Computer Science and Engineering Department, Malaviya National Institute of Technology, Jaipur, India
e-mail: tapasbadal@gmail.com

N. Nain
e-mail: nnain.cse@mnit.ac.in

M. Ahmed
e-mail: mahmed.cse@mnit.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_18

video sequence. Apart from summarizing information in a video, it also serves as a valuable tool for an application like activity analysis, augmented reality, crowd analysis, and many more knowledge extraction based applications.

Although the existing approach of video synopsis works well in condensing activities present in video over space, they do not preserve interaction between objects. While going through the various surveillance videos, it is observed that the object interaction in video possesses vital information such as information exchange, accidents, and theft.

This paper presents an approach of condensing the activities in surveillance video while preserving the interaction between the objects. The spatio-temporal tubes form a distinction between objects by separating their motion structure that helps in producing semantic information about the distinct object. The semantic information about moving objects present in a video not only helps in generating a summary of the activities in a video, but it also avoids spatial overlap between objects while synopsis of a video.

The segmented object trajectories are store as spatio-temporal tubes which are set of 3D tuples (x_i, y_i, t). Tube represents region belonging to an object i in frame t. Tubes store the motion structure of individual moving objects so that they can arrange over the spatial domain to generate video synopsis. Segmenting the motion structure of different objects helps in separating important activities and producing the synopsis of those activities only. We refer object motion structure and tube interchangeably in this paper. These tubes are further used to generate video synopsis by arranging them over space. It can also use for activity-based video indexing.

While arranging the spatio-temporal tubes over spatial domain, the time shift between tubes may destroy interaction between objects as shown in Fig. 1c for object 4 and 5. Proposed methodology keeps this interaction in synopsis video by merging the interacting tubes and consider them as a single tubeset as illustrated in Fig. 1d.

A method is propose for generating video synopsis, condensing as much information as possible while preserving the interaction between the object. The object interaction is maintained by finding the interaction point between two objects and merging their tubes. Energy minimization method is used for arranging the objects tubes over spatial domain. While arranging the tubeset over space, the cost is normalized using the length of tubes so that the participation of tube in cost function is proportional to the tube size. The remaining sections of this paper are structured as follows. Section 2 covers the related work explaining techniques given in the literature for generating video summarization. Section 3 describes important algorithms used for segmentation of moving object trajectories. In Sect. 4 outlines the proposed approach for preserving object interaction and generating video synopsis. Section 5 describes the experimental results of the proposed methods and the conclusion is given in Sect. 6.

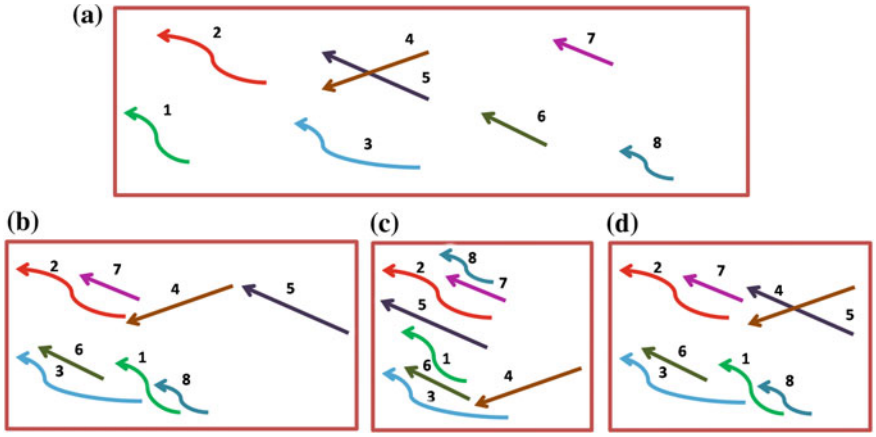


Fig. 1 Representation of different video synopsis approaches **a** Original video, **b** Synopsis video with time shift only, **c** Synopsis with time as well as space shift, **d** Synopsis using proposed approach preserving interaction between object 4 and 5

2 Related Work

Techniques of video motion analysis in literature are mainly put in order into two groups: Static method generates the short account of all activities in original videos in the form of image, and dynamic method produces summary as a video which is based on the content of original video.

In the static method, each shot is represented by key frames, which are selected to generate a representative image. Some of the examples of static image based summarization are video mosaic in which video frames are found using region of interest which are joined with each other to form a resulting video. Another form is video collage in which single image is generated by arranging region of interest on a given canvas. Storyboards and narratives are some more basic form of image based summarization.However, static methods produce an abstract of video in shorter space, but it does not take care of time-limited dependent relations between notable events. Also, summary in the form of video is more appealing than watching static images.

An example of the dynamic method is video synopsis. Video synopsis gets activities in the video into smaller space viewing part in both spatial and time-limited measures and generate a compact video that makes it easier to browse faster. Video synopsis presents a few limiting conditions as it has need of greatly sized memory area to keep background image of a scene along with segmented moving object trajectories. Although video synopsis saves memory in the final video it lost the information related to interaction occurs between objects when they come into proximity; Also the length of the synopsis video decides about the pleasing effect of the final video. Other examples of dynamic methods are video fast-forward, video skim-

ming, space-time video montage method, video narrative where selected frames are arranged to form an extremely condensed video.

The overall framework of generating video synopsis using energy minimization is given by Pritch et al. [1]. Fu et al. [2] measure sociological proximity distance to find an interaction between objects. Lee et al. [3] proposed method to generate video synopsis by discovering important object from the egocentric video.

3 Segmentation of Moving Object Tubes

The activity in a video is considered as motion structure of moving objects. Tubes of individual objects are segmented out by separating the motion structure of distinct objects using multiple moving object detection and tracking method. Automatic moving object detection and tracking aims at identification and segmentation of outline of the moving objects. Numerous approaches have been proposed in the literature for foreground segmentation such as temporal differencing [4], optical flow [5, 6], Gaussian mixture model (*GMM*) [7, 8], Codebook model [9, 10] etc. Segmentation of moving object trajectory using data association approach is proposed in [11].

This section gives a short description of approach applied for moving object detection and tracking in this paper. Moving object detection is implement using temporal difference combined background subtraction. An effective and most popular approach for background subtraction *GMM* is given by Stauffer and Grimson [7] that produces the promising output for slow moving objects and lighting change.

In *GMM* background is represented by multiple surfaces which is represented by a mixture of multiple Gaussians for each pixel. The recent history of a pixel is represented by a mixture of K Gaussian distributions at time t as $X_1, \ldots, X_t$. The probability of occurrence of the present pixel is calculated as follows:

$$P(X_t) = \sum_{i=1}^{k} \omega_{i,t} \times \eta \left\{ X_t, \mu_{i,t}, \sum_{i,t} \right\} \tag{1}$$

where K is used to denote the number of Gaussian distributions taken as 3–5 depending on the available memory, $\omega_{i,t}$, $\mu_{i,t}$ and $\sum_{i,t}$ represents the estimated weight, mean value and the covariance matrix assign to the ith Gaussian in the model at time t respectively. The parameter η is a Gaussian probability density defined as follows:

$$\eta(X_t, \mu, \sum) = \frac{1}{(2\Pi)^{(\frac{n}{2})} \mid \sum^{\frac{1}{2}} \mid} e^{\frac{-1}{2}(X_t - \mu_t)^T \sum^{-1}(X_t - \mu_t)} \tag{2}$$

Temporal differencing (*TD*) segments a moving object in a video by taking the difference between corresponding pixels of two or more consecutive frames as given in Eq. 3.

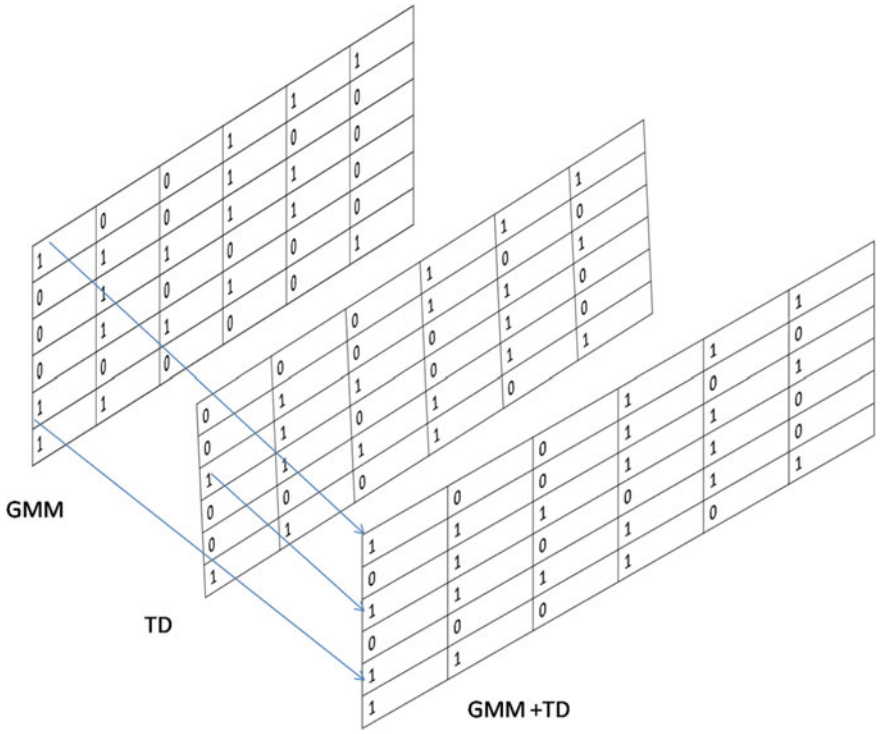


Fig. 2 *GMM* combined temporal differencing for foreground segmentation

$$I_A(i,j) = I_{n-1}(i,j) - I_n(i,j) \tag{3}$$

$$TD(i,j) = \begin{cases} 1 & \text{if } I_A(i,j) \geq Th, \\ 0 & \text{else} \end{cases} \tag{4}$$

The primary consideration for this technique is on how to determine the suitable threshold value. As both *GMM* and *TD* give the result as a mask of moving object(i.e. a binary image), we simply perform binary OR between each pixel of both mask images as shown in Fig. 2.

$$Mask = GMM \bigoplus TD. \tag{5}$$

The result of mask image and foreground segmented image using background subtraction combined temporal differencing is shown in Fig. 3.

Although above method of foreground segmentation is sufficient in most of the real-world situation, in some challenging situation it may assign pixels to the foreground which do not consider as moving objects like tree leaves, water surface, etc. An efficient system needs to eliminate those falsely detected pixels also labeled as

Fig. 3 Result of foreground segmentation

noise. Median Filter is used to removing background pixels detected as foreground as given in Eq. 6.

$$M(i,j) = median\{f(i-k, j-l), (k,l \in w)\} \tag{6}$$

where k and l defined the size of window w centred around the pixel (i,j) for taking median. After segmenting, to manage the identity of each foreground region a label is assigned to each of them that is require to maintain until an object is present in the video. The assignment of a label to an object in between the frame sequence is also termed as tracking. Tracking is used to separate array tubes of distinct moving object and generate their motion structure.

Due to miss detection or noise produced by object detection phase, the system generates the fragmented trajectory of an object that makes object tracking a challenging task. Yilmaz et al. [14] give the complete description of object tracking. An object is a label with an assignment to the detection by calculating the distance between the centroid of current detection to the predicted location computed by Kalman filter [15]. A detected region is assigned a label by calculating the minimum distance between the centroids of current detections and predicted centroid value computed by Kalman filter [15] as given in Eq. 7.

$$x_i = \boldsymbol{\Phi}_i x_{i-1} + w_{i-1} \tag{7}$$

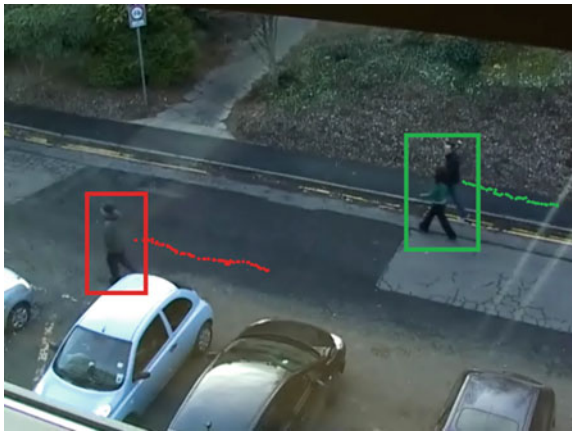


Fig. 4 Result of tracking

where x_i represent the location value at current step and x_{i-1} is at prior step, $\boldsymbol{\Phi}_i$ represents the state transition matrix relates the states between previous and current time steps. The w_i is a random variable used to represent the normally distributed process noise. Figure 4 shows the result of tracking superpixel area across the frame sequences. The motion structure of a moving object also denoted as tube in this paper is represented by a three tuples structure. In video analysis, tubes of distinct moving objects are the primary processing component. In video analysis, tubes of different moving objects are the primary processing element. Each tube A_i is a union of bounding boxes belongs to object i represented by b_i from frame j to frame k as given in Eq. 8.

$$A_i = \bigcup_{f=j}^{k} T_{(i,f,b_i)} \tag{8}$$

where the object i is tracked between frame number j to frame number k.

Background updation strategy: As in video synopsis the tubes are arranged over background changes are needed to synchronize. The background updation strategy is given in Eq. 9 which reflected changes immediately.

$$B_k(i,j) = \begin{cases} I_{k-1} & \text{if pixel(i,j) not belongs to motion region,} \\ B_{k-1}(i,j) & \text{otherwise} \end{cases} \tag{9}$$

where $B_k(i,j)$ are the pixels belongs to kth background, $B_{k-1}(i,j)$ pixels belongs to $(k-1)$th background and I_{k-1} is $(k-1)$th frame of the original video sequence.

4 Merge Interacting Object Tubes

The difference between the spatial overlapping tubes is considered here as the indication of intersection between the objects in an original video. We find the interaction between tubes by measuring the difference between tubes as given below in Eq. 10.

$$I_t(i,j) = \begin{cases} 0 & \text{if } d_t(i,j) > k, \\ k - d_t(i,j) & \text{otherwise} \end{cases} \tag{10}$$

where $d_t(i,j) = T_i^t - T_j^t$ used to compute the distance between tube i and j at time t and constant k is used for considering the minimum distance for interaction. Here K is taken as 5 which means that the tubes are considered as interaction if they are 5 pixels apart. The tubes having $I_t(i,j)$ other than 0 is merge and form a tubeset.

5 Video Synopsis by Energy Minimization

Energy minimization [1, 2, 16] is widely used and popular technique for generating video synopsis. The objective of energy minimization has defined a function that assigns cost for all possible solutions and finds solution with the lowest cost. While shifting the input pixel to synopsis pixel with time shift M we formulate energy functions to assign cost to activity loss and occlusion as follows:

$$E(M) = E_a(M) + \alpha E_o(M) \tag{11}$$

where $E_a(M)$ and $E_o(M)$ are used to represents the activity loss and occlusion across the frames respectively. α is used to assign a relative weight of occlusion. The activity loss of an object is the difference between pixels belongs to object tubes in input video and synopsis video. The occlusion cost represents the area that is shared by tubes in a frame in synopsis video.

Activity cost: As the length of an object tube depends on upon its appearance in the video it does not participate equally in synopsis video too. While calculating the activity loss weighted average of pixels belongs to input video and synopsis video is considered as given in Eq. 12.

$$E_a(i) = \frac{\sum_{t=SFrame_i}^{EFrame_i} ((x_i, y_i, t)_o - (x_i, y_i, t)_s)}{Length_i} \tag{12}$$

where $(x_i, y_i, t)_o$ and $(x_i, y_i, t)_s$ represent super pixel region belongs to object i in original video and synopsis video respectively.

Collision cost: While condensing the activity in an even shorter video it is required to share some pixel between tubes. Collision cost is computed by finding

Table 1 Notation

Symbol	Description
N	Number of frames in original video
S	Number of frames in synopsis video
Q	Set of tubes
K	Number of tubes
(x_i, y_i, t)	Pixel area of tube i at frame t
$SFrame_i$	Tube i starting frame
$EFrame_i$	Tube i ending frame
$Length_i$	Tube i length as in

the total number of pixels belongs to an object in consecutive frames share space in synopsis video as given in Eq. 13.

$$C_i = \begin{cases} 1 & if(x_i, y_i, t)_s = (x_i, y_i, t+1)_s, \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

It is also used to allow the user in defining the number of pixels in tubes that can overlap. Collision cost is normalized with the length of object tube for equal participation of each object as given in Eq. 14.

$$E_o(i) = \frac{\sum_{j \in Q} \sum_{t=1}^{S} C_i}{Length_i} \tag{14}$$

The higher value of E_o results in pixel overlapping between two objects that can affect the smoothness of object appearance. Moreover, the smaller value keeps objects well separated, but it generates a longer synopsis video. The procedure of to summarize the activity in the for of synopsis is explain in Algorithm 1. Table 1 used to explain the notations used in this paper.

Algorithm 1 Video synopsis algorithm.

Input: An array of K tubes as $A_i = (x_i, y_i, t)$ where $t = SFrame_i, \ldots, EFrame_i$.
Output: Synopsis video S with minimum energy $\sum_{n \in Q} E_n$.
Initialization: $S \longleftarrow \phi$;
1. Merge tube having interaction using Eq. 10.
2. Arrange tubes in ascending order according to their Length.
3. Process each tubes T_i from ordered list.
4. Find space and time shift for each tube in synopsis video.
5. $S \longleftarrow S \cup T_i$.with $E(M) \in min\{E(M)\}$
6. End.

Table 2 A summary of video description and result

Video	Frames	Tube	CR (%)
Video5 [17], 640 × 480, 25 fps	10,900	17	18
Person, 640 × 360, 30 fps	420	13	16
Car, 640 × 360, 29 fps	493	7	11

Fig. 5 Resulting synopsis frame for testing videos **a** Video5, **b** Person, **c** Car

6 Experimental Evaluation

The performance analysis of proposed approach is done on a number of publicly available datasets. Table 2 cover the description of these videos and condensation rate (*CR*) as well. *CR* denotes the percentage at which synopsis video compress the original video. The condensation ratio we get is between 10 and 20 for typical video having 10–17 tubes with no activity loss and allow occlusion of total 400 pixels between the tubes. Figure 5 shows qualitative result of synopsis generated through the proposed method. We can further reduce the condensation ratio by allowing some activity loss and an increase in pixel overlapping.

7 Conclusion

An approach to generate video synopsis of surveillance video is present in this paper. While existing methods condense activity in the video using shifting of object tubes over time and space, they do not maintain interaction between the objects. The interaction between object tubes was calculated using distance measure after that these tubes were merged. The cost function for activity loss and occlusion is computed for solving energy minimization problem. A weighted average of the cost function is taken to make participation of each object. The experimental result generated by proposed approach over the different video sequences provide promising results. In future, the synopsis comprising specific activities will be produced. We would extend this task by applying the methodology for a crowded video where segmentation of object tubes is still a challenging task.

References

1. Rav, A., Alex, R., Pritch, Y., Peleg, S.: Making a Long Video Short: Dynamic video synopsis. In: Computer Vision and Pattern Recognition, IEEE, pp. 435–441 (2006)
2. Fu, W., Wang, J.,Gui, L., Lu, H., Ma, S.: Online Video Synopsis of Structured Motion. In: Neurocomputing, Vol. 135.5, pp. 155–162 (2014)
3. Lee, Y., Ghosh, J., Grauman, K.: Discovering Important People and Objects for Egocentric Video Summarization. In: Computer Vision and Pattern Recognition (CVPR) pp. 1346–1353 (2012)
4. Liyuan, L., Huang, W., Irene, Y., Tian, Q.: Statistical Modeling of Complex Backgrounds for Foreground Object Detection. In: IEEE Transaction on Image Processing, IEEE Vol. 13.11, pp. 1459–1472 (2004)
5. Horn, Berthold, K., Schunck, Brian, G.: Determining Optical Flow. In: Artificial Intelligence, 17, pp. 185–203 (1981)
6. Suganyadevi, K., Malmurugan N., Sivakumar R.: Efficient Foreground Extraction Based On Optical Flow And Smed for Road Traffic Analysis. In: International Journal Of Cyber-Security And Digital Forensics. pp. 177–182 (2012)
7. Stauffer, C., Eric, W., Grimson, L.: Learning Patterns of Activity Using Real-Time Tracking. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, pp. 747–757 (2012)
8. Karasulu, B.: Review and Evaluation of Well-Known Methods for Moving Object Detection and Tracking in Videos. In: Journal Of Aeronautics and Space Technologies, 4, pp 11–22 (2012)
9. Kim, K., Chalidabhongse, T.H., Harwood, D., Davis, L.: Real-Time Foreground-Background Segmentation using Codebook Model. In: Real Time Imaging 11, Vol. 3, pp 172–185 (2005)
10. Badal, T., Nain, N., Ahmed, M., Sharma, V.: An Adaptive Codebook Model for Change Detection with Dynamic Background. In: 11th International Conference on Signal Image Technology & Internet-Based Systems, pp. 110–116. IEEE Computer Society, Thailand (2015)
11. Badal, T., Nain, N., Ahmed, M.: Video partitioning by segmenting moving object trajectories. In: Proc. SPIE 9445, Seventh International Conference on Machine Vision (ICMV 2014), vol 9445, SPIE, Milan, pp. 94451B–94451B-5 (2014).
12. Chen, W., Wang, K., Lan, J.: Moving Object Tracking Based on Background Subtraction Combined Temporal Difference. In: International Conference on Emerging Trends in Computer and Image Processing (ICETCIP'2011) Bangkok, pp 16–19 (2011)
13. Bastian, L., Leonardis, A., Schiele, B.: Robust Object Detection with Interleaved Categorization and Segmentation. In: International Journal of Computer Vision (IJCV), Vol. 77, pp. 259–289 (2008)
14. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. In: Acm computing surveys (CSUR). ACM, Vol. 38 pp. 4–13 (2006)
15. Fu, Z., Han, Y.: Centroid Weighted Kalman Filter for Visual Object Tracking. In: Elsevier Journal of Measurement, pp. 650–655 (2012)
16. Pritch, Y., Alex, R., Peleg, S.: Nonchronological Video Synopsis and Indexing. In: IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 30, NO. 11, pp. 1971–1984 (2008)
17. Blunsden, S., Fisher, R.: The BEHAVE Video Dataset: Ground Truthed Video for Multi-Person Behavior Classification. In: Annals of the BMVA, Vol. 4, pp. 1–12 (2010)

Face Expression Recognition Using Histograms of Oriented Gradients with Reduced Features

Nikunja Bihari Kar, Korra Sathya Babu and Sanjay Kumar Jena

Abstract Facial expression recognition has been an emerging research area in last two decades. This paper proposes a new hybrid system for automatic facial expression recognition. The proposed method utilizes histograms of oriented gradients (HOG) descriptor to extract features from expressive facial images. Feature reduction techniques namely principal component analysis (PCA) and linear discriminant analysis (LDA) are applied to obtain the most important discriminant features. Finally, the discriminant features are fed to the back-propagation neural network (BPNN) classifier to determine the underlying emotions from expressive facial images. The Extended Cohn-Kanade dataset (CK+) is used to validate the proposed method. Experimental results indicate that the proposed system provides the better result as compared to state-of-the-art methods in terms of accuracy with the substantially lesser number of features.

Keywords Face expression recognition · Histograms of oriented gradients · Principal component analysis · Linear discriminant analysis · BPNN

1 Introduction

Facial expression recognition is a standout amongst the most effective, regular, and prompt means for individuals to impart their feelings, sentiments and desires [1]. Facial expressions contain non-verbal communication cues, which helps to identify the intended meaning of the spoken words in face-to-face communication.

N.B. Kar (✉) · K.S. Babu · S.K. Jena
Department of Computer Science and Engineering, National Institute of Technology, Rourkela 769008, India
e-mail: nikunjakar@gmail.com

K.S. Babu
e-mail: ksathyababu@nitrkl.ac.in

S.K. Jena
e-mail: skjena@nitrkl.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_19

According to Mehrabian [2] in a face-to-face conversation, the spoken words carry 7 %, the voice intonation carry 38 %, and the face carry 55 % of the effect of the message. Therefore, the face expressions play a vital modality in human-machine interface.

Facial expression recognition with high accuracy stays to be troublesome and still speaks to a dynamic exploration area. Face appearance investigation utilized as a part of various applications which covers wide areas like robotics, influence delicate human computer interface (HCI), telecommunications, behavioural science, video games, animations, psychology, automobile safety, affect touchy music jukeboxes and TVs, educational software, and many more [3].

Facial expression recognition system consists of three major steps: (1) face acquisition, (2) facial data extraction and representation, and (3) facial expression detection [4]. The face acquisition step finds the face area from an input image or an image sequences. Head finder, head tracking, and pose estimation techniques are applied to handle head motion in face emotion recognition system. Facial features in emotion recognition system are classified into two broad categories, (1) geometric features and (2) appearance based features. Geometric features represented by shape and location of facial components including mouth, eyes, eye-brows, and nose. Image filters such as Gabor wavelets [5] are applied to the whole or part of the face to extract appearance features. Local binary pattern (LBP), local ternary pattern (LTP), and histograms of oriented gradients (HOG) [6] descriptors are also used as appearance features. Face emotion recognition system is also used hybrid approaches (combination of geometry and appearance features) gives promising results. Finally, a classifier is utilized to detect emotion from facial features.

The proposed method at first crops and resize the face region detected by Viola and Jones face detection algorithm [7]. Secondly, HOG is employed to extract features. Third, PCA+LDA technique is applied to reduce the dimension of the feature vector. Finally, the reduced features are fed to BPNN, where steepest descent is suggested to find the optimal weights of the BPNN. The complexity of classification algorithm is very low due to a lesser number of features. The main aim of this work is to keep high recognition accuracy with lesser number of features.

The remainder of the paper is organised as follows: Sect. 2 gives an overview of the related work. Section 3 portrays the working procedure of each step of the proposed method. Section 4 presents the experimental results and discussions. Finally, Sect. 5 concludes the paper.

2 Related Work

In recent years, a diverse of researchers suggested many face expression recognition methods. Appearance based feature extraction techniques like Gabor wavelets, LBP, LTP, and HOG are used to extract features from the whole face image or the part of the face image, without any prior knowledge of face expressions. Therefore, the size of the feature vector is the whole face or the component of the face. Geometric based

features represent facial components with a set of fiducial points. The drawback of this approach is that the fiducial points must be set manually, which includes a complex procedure. In this technique, recognition accuracy increases, with the increase in face feature points.

Tsai et al. [8] proposed a novel face emotion recognition system using shape and texture features. Haar-like features (HFs) and self-quotient image (SQI) filter is used to detect the face area from the image. Angular radial transformation (ART), discrete cosine transform (DCT) and Gabor filters(GF) are used for feature extraction. The model proposed by Tsai et al. adopts ART features with 35 coefficients, SQI, Sobel, and DCT with 64 features, GF features with 40 texture change elements. A SVM classifier was employed to classify images into eight categories including seven face expressions and non-face. Chen et al. [9] proposed hybrid features that include facial feature point displacement and local texture displacement between neutral to peak face expression images. The resultant feature vector contains 42-dimensional geometric features and 21-dimensional texture features. A multiclass SVM was deployed to recognise seven facial expressions. Valster and Pantic [10] located 20 facial fiducial points using a face detector based on Gabor-feature-based boosted classifier. These fiducial points can be tracked through a series of images using particle filtering with factorized likelihoods. Action unit (AU) recognition can be done with a combination of GentleBoost, SVM, and hidden Markov models.

Hsieh et al. [11] used six semantic features, which can be acquired using directional gradient operators like GFs and Laplacian of Gaussian (LoG). Active shape model (ASM) is trained to detect the human face and calibrates facial components. Later, Gabor and LoG edge detection is used to extract semantic features. Chen et al. [12] applied HOG to face components to extract features. The face components are eyes, eye-brows, and mouth. Initially, the HOG features are extracted from each of these components, and they are concatenated to have a feature vector of size 5616. Gritti et al. [6] proposed HOG, LBP, and LTP descriptors for facial representations. Their experiments reveal that HOG, LBP-Overlap, LTP-Overlap descriptors result in 18,954 features, 8437 features, and 16,874 features respectively. A linear SVM with ten-fold cross validation testing scheme was used in their recognition experiments.

The literature review reveals that local features like HOG, LBP, LTP, and Gabor wavelets have been used to represent the face with quite a large number of features, which slows down the face expression recognition process. Thus, there is a scope to reduce the feature vector size which in turn minimize the computational overhead.

3 Proposed Work

The proposed technique includes four principal steps: preprocessing of facial images, feature extraction, feature length reduction, and classification. Figure 1 shows the block diagram of the proposed system. Detail description of each block of the proposed system is given below.

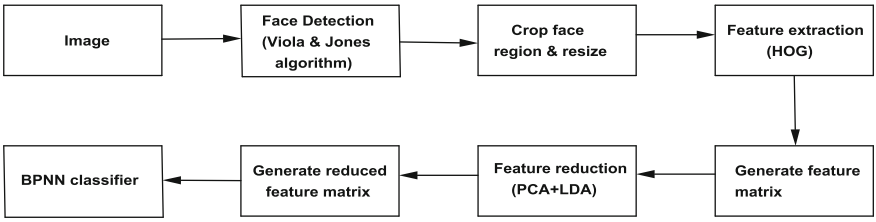


Fig. 1 Block diagram of proposed system for classification of face expression images

3.1 Preprocessing

At first, the face images are converted to a gray scale image. Then the contrast of the image was adjusted so that 1 % of the information is immersed at low and high intensities. After the contrast adjustment, the face is detected using popular Viola and Jones face detection algorithm [7]. The detected face region is cropped from the original face image. Then the cropped face is reshaped to an image of size 128×128.

3.2 Feature Extraction

Dalal and Higgs [13] proposed HOG descriptor for human detection. After that it has been widely used for various computer vision problems like pedestrian detection, face recognition, and face expression recognition. In HOG, images are represented by the directions of the edges they contain. Gradient orientation and magnitudes are computed by applying gradient operator across the image for HOG features.

Initially, the image is divided into a number of cells. A local 1-D histogram of gradient directions over the pixel are extracted for each cell. The image is represented by combining histograms of each cell. Contrast-normalization of the local histograms is necessary for better invariance to illumination, shadowing, etc. So, local histograms are combined over a larger spatial region, called blocks by using the result of normalization of cells within the block. The feature length increases when the blocks are overlapping. The normalized blocks are combined to represent HOG descriptor.

3.3 Feature Reduction

LDA has been effectively connected to different classification problems like face recognition, speech recognition, cancer detection, multimedia information retrieval etc. [14]. The fundamental target of LDA is to discover projection F that boosts the proportion of between class scatter S_b to within class scatter S_W.

$$arg \max_{F} \frac{\left|FS_b F\right|}{\left|FS_W F\right|} \tag{1}$$

For a very high dimensional data, the LDA algorithm faces various challenges. In our proposed work, the HOG descriptor is used to extract features from the pre-processed face image. First, it divides the image into 16×16 blocks each with 50 % overlap. Each block contains 2×2 cells each of size 8×8. As a whole we get $15 \times 15 = 225$ blocks. For each cell, it computes the gradient orientation with nine bins that are spread over (0°–180°) (signed gradient). That implies the feature vector of $225 \times 4 \times 9 = 8100$ dimensions.

The scatter matrices are of size $8100 \times 8100 \cong 63M$. It is computationally challenging to handle such enormous matrices. The matrices are always singular because the number of samples needs to be at least 63M, with the goal that they will non-degenerate. This problem is known as small sample size (SSS) problem as the size of the sample set is smaller than the dimension of the original feature space. To dodge these issues, another rule is utilized before LDA approach to reducing the dimension of the feature vector. PCA is used to reduce the dimension and S_W is no longer degenerate. After that LDA methodology can continue with no inconvenience.

$$RFV = (F_{PCA})_{LDA} \tag{2}$$

where, F is the feature vector to be reduced and RFV is the reduced feature vector after applying the PCA+LDA criterion. PCA is a method that is utilized for applications, for example, dimensionality reductions, lossy information pressure, highlight extraction, and information representation [15]. PCA is utilized to reduce the dimension to $X - 1$, where X is the total number of samples and the feature vector is of size $X \times (X - 1)$. Then, LDA is employed to decrease the dimension further. This method creates a reduced feature vector. The feature vector alongside a vector holding the class names of all samples is fed as an input to the classifier.

3.4 Classification

Artificial neural network with back-propagation (BP) algorithm has been used in solving various classification and forecasting problems. Despite the fact that BP convergence is moderate, yet it is ensured. A BPNN with one input, one hidden and one output layer is presented. The network employed sigmoid neurons at the hidden layer and linear neurons at the output later. The training samples are introduced to the network in batch mode. The network configuration is $I_M \times H_Y \times Z$, i.e., M number of features, Y number of hidden neurons, and Z number of output neurons, which indicate emotions. The network structure is depicted in Fig. 2.

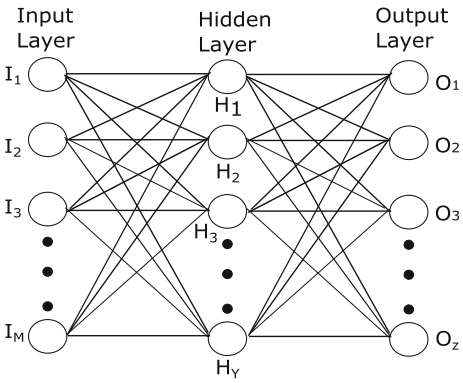


Fig. 2 A three layer BPNN used for face expression recognition

The input layer consist of 6 neurons as per the six features are selected after applying PCA+LDA standard. The number of hidden neurons Y can be calculated as per the Eq. 3,

$$Y = \frac{(M + Z)}{2} \tag{3}$$

The back-propagation algorithm with Steepest descent learning rule is most frequently used training algorithm for classification problems, which is also utilized in this work. Back-propagation learning consists of two phases, namely forward pass and backward pass [16]. Initially, the input features are presented to the input nodes and its output propagates from one layer to other layers of the network. All the network weights and biases are fixed during the forward pass.

The difference of the actual output from the desired output treated as an error signal. In backward pass, the weights and biases are updated by passing the error signal backward to the network. The learning performance is measured by root mean square error ($RMSE$).

4 Experiments

The examinations are carried out on a PC with 3.40 GHz Core i7 processor and 4 GB of RAM, running under Windows 8 working framework. The proposed system is simulated utilizing Matlab tool. The summary of the proposed scheme is presented in Algorithm 1.

Algorithm 1 Proposed face expression recognition system

Require: Face expression images
X: Total number of images
N: Total number of features
M: Number of reduced features
Ensure: Emotion class of the test face image
Step 1: Face detection using Viola and Jones algorithm and resize the image
Loop on $i \leftarrow 1$ to X
Read the face expression image
Detect the face using Viola and Jones algorithm and then cropped the detected face
Resize the face image to 128×128
End Loop
Step 2: Features extraction using HOG
Loop on $i \leftarrow 1$ to X
Read the 128×128 size gray scale image
Get the HOG features in a 1-by-N vector
Put the HOG features in a matrix $A(X \times N)$
End Loop
Step 2: PCA+LDA feature reduction
Choose a desired dimension $M \ll N$
Apply PCA to reduce the dimension of feature vector to $X - 1$
Perform LDA to further reduce the size of the feature vector to M
Get the reduced feature vector $RFV(X \times M)$
Step 3: Classification using BPNN
Design the neural network using feed forward back propagation algorithm
Create a new dataset T using RFV and a target vector C
Train the BPNN classifier
Input the test images to the network and classify

4.1 Dataset

One standard dataset, CK+ [17] was used to validate the proposed method. In CK+ dataset, the facial expression of 210 adults is captured. The members were 18–50 years old, among them 69 % female, 81 % Euro-American, 13 % Afro-American, and 6 % different gatherings. CK+ contains 593 posed facial expressions from 123 subjects. Among 593 posed facial expressions, 327 were labeled with seven basic emotion categories. We have selected 414 images from the CK+ dataset, which includes 105 neutral images and 309 peak expressive images for the experiment. However, we have excluded the contempt face expression images from the dataset. The preprocessed sample images from CK+ of the experiment are shown in Fig. 3.

Fig. 3 Preprocessed sample images from CK+ dataset of the experiment

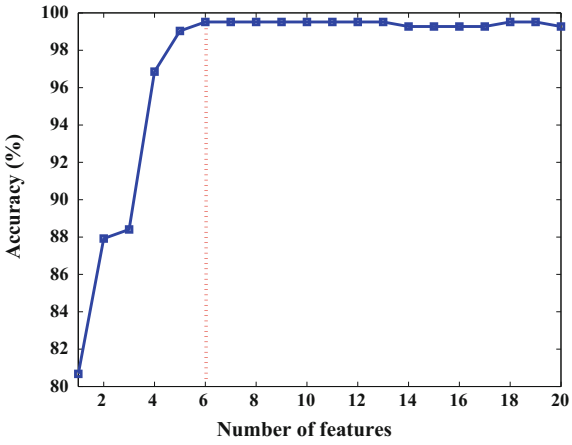


Fig. 4 Performance evaluation with respect to number of features

4.2 Results and Discussion

The feature extraction stage is implemented by HOG with following settings: signed gradient with nine orientation bins, which are evenly spread over 0°–180° a cell size of 8×8, 4 number of cells in a block, 50 % block overlap, L2-Hys (L2-norm followed by clipping) block normalization, and $[-1, 0, 1]$ gradient filter with no smoothing.

All the images in the dataset are of size 640×490. After face detection, the images are cropped and resized to 128×128. Then the features are extracted using HOG with above-mentioned settings. The extracted feature dimension of each image is 1×8100 (225 blocks $\times$ 4 cells in each block $\times$ 9 bins = 8100). Then PCA+LDA approach is used to reduce the dimension of the feature vector from 8100 to 6. Figure 4 depicts the plot between accuracy and the number of features. It is observed that with only 6 features, the proposed system achieves highest classification accuracy. The six features with a target vector containing all class labels are combined to form a resultant dataset.

The resultant dataset is fed to the BPNN classifier. The network consists of three layers, with six nodes (represents the six features) in the input layer, six nodes in the hidden layer, and seven nodes (represents seven emotions, i.e. anger, disgust, fear, happy, sad, surprise, neutral) in the output layer. The training error for the dataset is

Table 1 Fold-wise performance analysis of the proposed system

Folds	Training instances	Testing instances	Correctly classified	Incorrectly classified	Accuracy (%)
Fold 1	331	83	83	0	100
Fold 2	331	83	82	1	98.79
Fold 3	331	83	82	1	98.79
Fold 4	331	83	83	0	100
Fold 5	332	82	82	0	100
			Average accuracy		99.51

Table 2 Confusion matrix of the proposed approach

Emotions	Disgust	Surprise	Anger	Happy	Sad	Fear	Neutral
Disgust	100	0	0	0	0	0	0
Surprise	0	98.8	0	0	0	0	1.2
Anger	0	0	97.78	0	0	0	2.22
Happy	0	0	0	100	0	0	0
Sad	0	0	0	0	100	0	0
Fear	0	0	0	0	0	100	0
Neutral	0	0	0	0	0	0	100

0.0419. Table 1 shows the result of 5-fold stratified cross-validation (CV) procedure. The confusion matrix of the proposed facial recognition system is given in Table 2.

The comparative analysis of proposed method with state-of-art methods in terms of classification accuracy and number of features are listed in Table 3. All the existing methods have been validated on the same dataset. It is evident that the proposed scheme yields higher classification accuracy with the lesser number of features compared to other methods. The use of these features reduces computational overhead. In addition, it makes the classifier task more feasible. For a test image, the execution time for preprocessing and feature extraction is 0.133 s, whereas for feature reduction and classification it is 0.008 s and 0.001 s respectively. During time analysis, the time needed for training the classifier is not considered.

5 Conclusion

This paper proposes a hybrid system for facial expression recognition. At first, the face is detected using Viola and Jones face detection algorithm. In order to maintain a uniform dimension of all the face images, the detected face region is cropped and resized. Then the system introduced HOG to extract features from the preprocessed face image. A PCA+LDA approach is harnessed to select most significant features

Table 3 Performance comparison of proposed method with state-of-the-art methods

References	Facial features	Classifier	Feature length	Accuracy (%)
Gritti et al. [6] ′08	LBP-Overlap	Linear SVM	8437	92.9
Gritti et al. [6] ′08	HOG	Linear SVM	18,954	92.7
Tsai et al. [8] ′10	SQI+Sobel+DCT+ ART+GF	Nonlinear SVM	179	98.59
Valstar and Pantic [10] ′12	Facial points	SVM and HMM	20	91.7
Saeed et al. [18] ′14	Geometrical features	Nonlinear SVM	8	83.01
Hsieh et al. [11] ′15	Semantic features	ASM+ Nonlinear SVM	6	94.7
Proposed method	**HOG+PCA+LDA**	**Nonlinear SVM**	**6**	**99.27**
	HOG+PCA+LDA	**BPNN**	**6**	**99.51**

from the high dimensional HOG features. Finally, BPNN classifier has been used to build an automatic and accurate facial expression recognition system. Simulation results show the superiority of the proposed scheme as compared to state-of-the-art methods on CK dataset. The proposed scheme achieves recognition accuracy of 99.51 % with only six features. In future, other machine learning techniques can be suggested to enhance the performance of the system. In addition, contempt face images of CK+ dataset can be taken into consideration.

References

1. Tian, Y.l., Brown, L., Hampapur, A., Pankanti, S., Senior, A., Bolle, R.: Real world real-time automatic recognition of facial expressions. In: Proceedings of IEEE workshop on Performance Evaluation of Tracking and Surveillance (PETS) (2003)
2. Pantic, M., Rothkrantz, L.J.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1424–1445 (2000)
3. Bettadapura, V.: Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722 (2012)
4. Tian, Y.L., Kanade, T., Cohn, J.F.: Facial expression analysis. In: Handbook of face recognition, pp. 247–275. Springer (2005)
5. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Fully automatic facial action recognition in spontaneous behavior. In: 7th International Conference on Automatic Face and Gesture Recognition. pp. 223–230. (2006)
6. Gritti, T., Shan, C., Jeanne, V., Braspenning, R.: Local features based facial expression recognition with face registration errors. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–8. (2008)
7. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. I–511. (2001)
8. Tsai, H.H., Lai, Y.S., Zhang, Y.C.: Using svm to design facial expression recognition for shape and texture features. In: International Conference on Machine Learning and Cybernetics (ICMLC). vol. 5, pp.2697–2704. (2010)

9. Chen, J., Chen, D., Gong, Y., Yu, M., Zhang, K., Wang, L.: Facial expression recognition using geometric and appearance features. In: 4th International Conference on Internet Multimedia Computing and Service. pp. 29–33. (2012)
10. Valstar, M.F., Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 42(1), 28–43 (2012)
11. Hsieh, C.C., Hsih, M.H., Jiang, M.K., Cheng, Y.M., Liang, E.H.: Effective semantic features for facial expressions recognition using svm. Multimedia Tools and Applications pp. 1–20 (2015)
12. Chen, J., Chen, Z., Chi, Z., Fu, H.: Facial expression recognition based on facial components detection and hog features. In: International Workshops on Electrical and Computer Engineering Subfields (2014)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR). vol. 1, pp. 886–893. IEEE (2005)
14. Yu, H., Yang, J.: A direct LDA algorithm for high-dimensional datawith application to face recognition. Pattern recognition 34(10), 2067–2070 (2001)
15. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
16. Haykin, S., Network, N.: A comprehensive foundation. Neural Networks 2 (2004)
17. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 94–101. (2010)
18. Saeed, A., Al-Hamadi, A., Niese, R., Elzobi, M.: Frame-based facial expression recognition using geometrical features. Advances in Human-Computer Interaction (2014)

Dicentric Chromosome Image Classification Using Fourier Domain Based Shape Descriptors and Support Vector Machine

Sachin Prakash and Nabo Kumar Chaudhury

Abstract Dicentric chromosomes can form in cells because of exposure to radioactivity. They differ from the regular chromosomes in that they have an extra centromere where the sister chromatids fuse. In this paper we work on chromosome classification into normal and dicentric classes. Segmentation followed by shape boundary extraction and shape based Fourier feature computation was performed. Fourier shape descriptor feature extraction was carried out to arrive at robust shape descriptors that have desirable properties of compactness and invariance to certain shape transformations. Support Vector Machine algorithm was used for the subsequent two-class image classification.

Keywords Cytogenetic image analysis · Shape-based classification · Fourier shape descriptor

1 Introduction

Chromosomes contain most of the DNA of an organism. During cell division, chromosomes are replicated so that they may be passed on to the daughter cells. Cell division is broadly of two types—mitosis and meiosis. In mitosis, a single parent cell produces two daughter cells that are genetically identical. The general cellular repair and replenishment in humans occurs by the process of mitosis. Meiosis is a specialized form of cell division which leads to genetic diversity. Cell division is organised into various stages. The mitotic cell division broadly consists of many stages such as interphase, prophase, prometaphase, metaphase, anaphase and telophase.

During exposure to radiation, either of the tissue sample or of the organism, cytogenetic changes can occur. One such manifestation is the creation of dicentric chromosomes—which are characterised by the presence of two centromeres rather

S. Prakash (✉) · N.K. Chaudhury
Institute of Nuclear Medicine & Allied Sciences, Delhi, India
e-mail: sachin.inmas@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_20

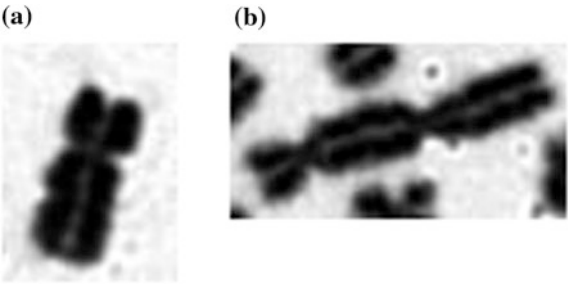


Fig. 1 **a** Single normal metaphase chromosomes: the four distinctly visible sections are each a chromatid. The location where they are fused together is the centromere. **b** A dicentric chromosome in its neighbourhood characterised by the presence of two centromeres

than one (Fig. 1). Telescoring is the technique of counting such, and other aberrant chromosomes (and at multiple geographical locations if needed say in the case of a radiological accident or attack) and in the process assessing the actual and effective radiation damage that the organism has suffered irrespective of what might be assessable by just the exposure dose information.

Although chromosomes are miniscule subcellular structures, metaphase is one phase of cell division where the chromosomes are at a very condensed stage and are hence quite conducive to microscopic imaging. The input images were acquired from Metafer 4 slide scanning microscope at our institute. This is an automated imaging and analysis system comprising of a Microscope (model Axio Imager M2, from Zeiss Germany) and DC score software for metaphase slide analysis. Some examples of one of the input microscope images are depicted in (Fig. 2) (scaled down to fit page). Our aim was to come up with a method to classify the constituent individual chromosome images into normal and dicentric chromosomes.

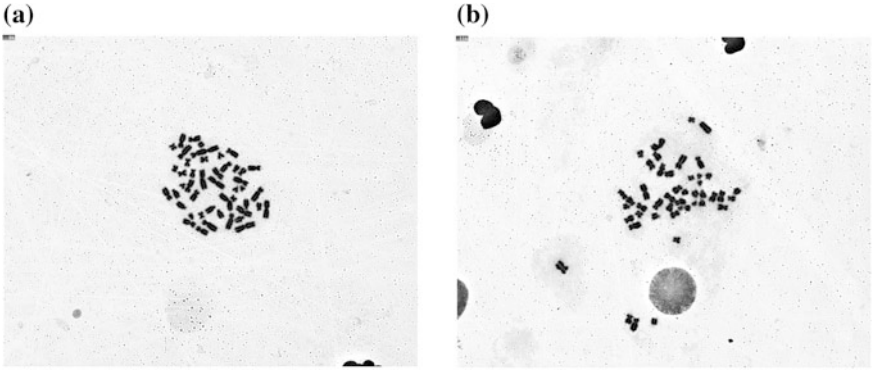


Fig. 2 **a** A good quality chromosomes image. **b** An image with a nucleus in the background and other debris from surrounding

2 Method

2.1 Preprocessing of Images

Chromosome image classification is an ongoing area of research globally and many good studies have been reported in literature [1–3]. The images were first preprocessed using gross morphological operations to extract the region of interest from the large input images. The ROI images contain primarily the chromosomes with some background artefacts such as the slide medium or in some cases speckles of dust. Specialized Segmentation techniques as surveyed in [2] have been applied on chromosome images as evident in the literature. But in the current paper the main focus is on shape based identification. In order to extract the individual chromosomes, the Otsu segmentation based on thresholding was applied followed by cleaning of the resultant image by means of morphological operations to remove small objects and holes. The result was individual binary images of chromosomes, some examples of which are depicted in (Fig. 3).

It also may happen that during microscope slide preparation of adequate precautions are not followed then the resulting slide will have overlapping chromosomes. Although some researchers have addressed this issue [3–5] it was not addressed in the current study for the reason that in such cases in metaphase images, even the dicentricity at crossovers is visually very difficult to discern and such cases are as such discarded. The issue of touching chromosomes [6] shall be taken up subsequently.

2.2 Shape Extraction and Feature Calculation

Shape has been shown to be an important component in the visual scene understanding by humans. In the extracted chromosomes the distinguishing feature in the dicentric chromosomes are two additional constrictions along their body length. Further, the chromosomes exhibit variability in size and shapes even amongst normal chromosomes. Another observation is that the chromosomes are manifested

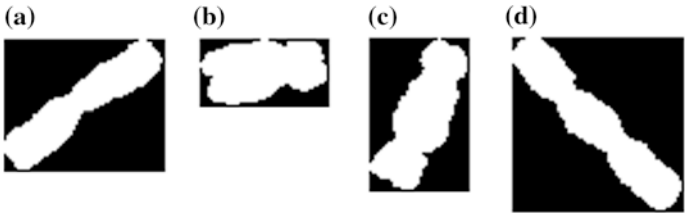


Fig. 3 Extracted individual chromosomes **a**, **b** are normal; **c**, **d** are dicentric

differently oriented. Shape feature methods are widely used over and above segmentation in image analysis [7]. The shape features for our use had to be robust to scale, rotation and boundary start point. Various methods of shape representation and feature extraction are found in literature such as those based on geometry [8, 9]. Fourier based shape descriptor was adopted as many desirable invariance properties are achievable in the Fourier space representation [10–12]. Some researchers have also used contour re-sampling to counter the variation in object size [13, 14] and to sample a fixed number of points. The method used by [15] was adapted to arrive at the Fourier Shape Descriptor as detailed below.

After boundary extraction, the shape is represented by its boundary contours g(x (t), y(t)) = f. The boundary can be taken to represent a periodic signal with a period of 2π when we take the function f as defined in Eq. (1).

$$f: [0, 2\pi] \rightarrow \Re^2 \tag{1}$$

Now, the function g can be identity function or any other mapping of the boundary contour and when this function f is expanded into Fourier series, the Fourier coefficients can be taken as an approximation of the shape. The boundary can be represented either as a collection of contour points as in [10, 15], or as in a multidimensional representation as in [12] or as a scalar as approached in this paper.

The complex domain representation of the boundary can be achieved by computing z(t) as shown in Eq. (2).

$$z(t) = x(t) + i \cdot y(t) \tag{2}$$

The Fourier expansion of the function z is as in Eq. (3).

$$z(t) = \sum_{k=0}^{N-1} a_k \exp(2\pi ikt/N) \tag{3}$$

The coefficients are given by the following Eq. (4).

$$a_k = \frac{1}{N} \sum_{k=0}^{N-1} z(t) \exp(-2\pi ikt/N) \tag{4}$$

A shape signature can be calculated from the 2-dimensional contour points rather than working directly on the boundary contour points. The shape signature used was the centroid distance function. The centroid of the shape boundary is given by the Eq. (5).

$$(x_0, y_0) = \frac{1}{N} \sum_{k=0}^{N-1} (x_k, y_k) \tag{5}$$

The Centroid Distance Function is computed as shown in Eq. (6).

$$r(t) = \sqrt{(x(t) - x_0)^2 + (y(t) - y_0)^2} \tag{6}$$

By its very nature this shape signature is rotation invariant. The Fourier transform of the function is computed as given by the Eq. (7) (Fig. 4):

$$a_k = \frac{1}{N} \sum_{t=0}^{N-1} r(t) \exp(-2\pi ikt/N) \tag{7}$$

And further invariance of scale and boundary start point is arrived at by phase normalization as depicted by Eq. (8):

$$anew_k = \left| \exp(is\alpha_k - ik\alpha_s) \frac{|a_k|}{|a_0|} \right| \tag{8}$$

$$\alpha_k = \arg(a_k)$$

Here, 0 is the index of the coefficient which is actually the mean of the data points while 's' is the index of the coefficient with the second largest magnitude. Further the log scale representation of these coefficients was taken and a subset of the coefficients was used as the final Fourier Boundary Descriptors set. The cardinality of the subset was kept the same for each chromosome. This was done to include only the foremost relevant discerning Fourier features and also to make the

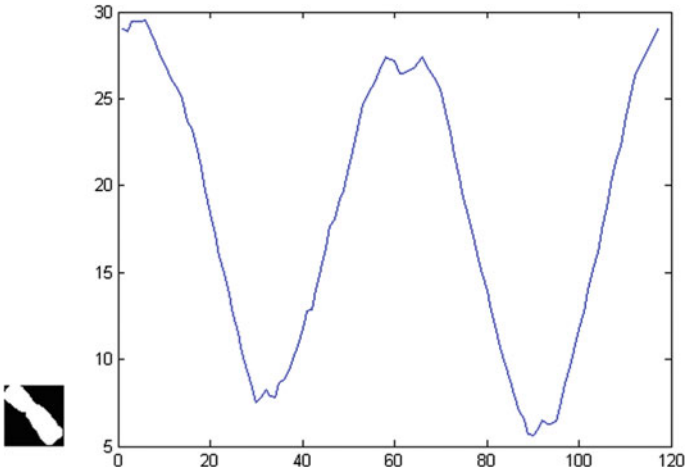


Fig. 4 A chromosome and its CDF

final feature vector uniform in size irrespective of the chromosome. After experimentation, 50 features comprised from the first 25, sparing the mean and the last 25 produced acceptable results on classification accuracy.

2.3 Image Classification

Support Vector Machine [16] is a robust machine learning algorithm for classification and a popular supervised learning method in use. SVM has been employed for cytogenetic image classification [17]. In our work, two-class image classification was carried out using the custom designed feature set as described above.

3 Results

A total of 141 chromosome images were used in this study. Out of these, 105 were the binary extracted images of normal chromosomes and 36 were the binary extracted images of dicentric chromosomes. The image dataset was randomly divided into training and testing datasets. At all times, 70 normal chromosome images and 20 dicentric chromosome images were used for the training phase and the remaining images were used for the test classification.

Linear and radial basis function kernels were tried. The best accuracy of 90.1961 % in classification was achieved by using a Linear Kernel with SVM in which 46 of the 51 test images were correctly classified. Matlab along with Libsvm [18] was used for implementation of the work.

References

1. Boaz Lerner, "Toward A Completely Automatic Neural-Network-Based Human Chromosome Analysis", IEEE Trans. On Systems, Man and Cybernetics, vol. 28, no. 4, Aug 1998.
2. W. Yan, D. Li, "Segmentation Algorithms of Chromosome Images", 3rd International Conference on Computer Science and Network Technology, 2013.
3. Graham C. Charters and Jim Graham, "Disentangling Chromosome Overlaps by Combining Trainable Shape Models With Classification Evidence", IEEE Trans. on Signal Processing, vol. 50, no. 8, Aug 2002.
4. G. Agam, Its'hak Dinstein, "Geometric Separation of Partially Overlapping Nonrigid Objects Applied to Automatic Chromosome Classification", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, no. 11, Nov 1997.
5. A. S. Arachchige, J. Samarabandu, J. H. M. Knoll, and P. K. Rogan "Intensity Integrated Laplacian-Based Thickness Measurement for Detecting Human Metaphase Chromosome Centromere Location", IEEE Trans. on Biomedical Imag, vol. 60, no. 7, Jul 2013.

6. Shervin Minaee, Mehran Fotouhi, Babak Hossein Khalaj, "A Geometric Approach to Fully Automatic Chromosome Separation", Signal Processing in Medicine and Biology Symposium (SPMB), 2014 IEEE.
7. Denshang Zhang, Guojun Lu, "Review of shape representation and description techniques", 2003 Pattern Recognition Society.
8. E. Poletti, E. Grisan, A. Ruggeri, "Automatic classification of chromosomes in Q-band images", 30th Annual Intl. IEEE EMBS Conf. Aug 2008.
9. H Ling, D.W. Jacobs, "Shape Classification Using the Inner-Distance", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 29, no. 2, Feb 2007.
10. D. Zhang, G. Lu, "A Comparative Study on Shape Retrieval Using Fourier Descriptors with Different Shape Signatures", Monash Univ.
11. D Zhang, G Lu, "Study and evaluation of different Fourier methods for image retrieval", Image Vis. Comput. 23, 33–49, 2005.
12. I. Kuntu, L. Lepisto, J. Rauhamaa, A. Visa, "Multiscale Fourier Descriptor for Shape Classification", Proc of the 12th Intl. Conf. on Image Analysis and Processing, 2003.
13. Volodymyr V. Kindratenko, Pierre J. M. Van Espen, Classification of Irregularly Shaped Micro-Objects Using Complex Fourier Descriptors, Proc of 13th Intl Conf on Pattern Recognition, vol.2, pp. 285–289, 1996.
14. J. Mataz, Z Shao, J. Kitter, "Estimation of Curvature and Tangent Direction by Median Filtered Differencing", 8th Int. Conf. on Image Analysis and Processing, San Remo, Sep 1995.
15. Christoph Dalitz, Christian Brandt, Steffen Goebbels, David Kolanus, "Fourier Descriptors for Broken Shapes", EURASIP Journ. On Advances in Signal Proc, 2013.
16. Cortes, C.; Vapnik, V. (1995). "Support-vector networks", Machine Learning 20 (3): 273. doi:10.1007/BF00994018.
17. Christoforos Markou, Christos Maramis, Anastasios Delopoulos, "Automatic Chromosome Classification using Support Vector Machines".
18. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

An Automated Ear Localization Technique Based on Modified Hausdorff Distance

Partha Pratim Sarangi, Madhumita Panda, B.S.P Mishra and Sachidananda Dehuri

Abstract Localization of ear in the side face images is a fundamental step in the development of ear recognition based biometric systems. In this paper, a well-known distance measure termed as modified Hausdorff distance (MHD) is proposed for automatic ear localization. We introduced the MHD to decrease the effect of outliers and allowing it more suitable for detection of ear in the side face images. The MHD uses coordinate pairs of edge pixels derived from ear template and skin regions of the side face image to locate the ear portion. To detect ears of various shapes, ear template is created by considering different structure of ears and resized it automatically for the probe image to find exact location of ear. The CVL and UND-E database have side face images with different poses, inconsistent background and poor illumination utilized to analyse the effectiveness of the proposed algorithm. Experimental results reveal the strength of the proposed technique is invariant to various poses, shape, occlusion, and noise.

Keywords Biometrics · Skin-color segmentation · Hausdorff distance · Ear localization · Ear varification

1 Introduction

In recent years, ear biometric has gained much attention and become an emerging area for the new innovation in the field of biometrics. Ear biometrics is achieving popularity as unlike face, ears are not affected by aging, mood, health, and posture. Automatic ear localization in the 2D side face image is a difficult task and the per-

P.P. Sarangi (✉) · B.S.P. Mishra
School of Computer Engineering, KIIT University, Bhubaneswar, India
e-mail: ppsarangi@gmail.com

M. Panda
Department of MCA, Seemanta Engineering College, Jharpokharia, India

S. Dehuri
Department of ICT, FM University, Balasore, India

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_21

formance of ear localization influences the efficiency of the ear recognition system. Ear can be used as a biometric trait for human recognition that has been first documented by the French Criminologist, Alphonse Bertillon [1]. More than a century ago, Alfred Iannarelli [2] has demonstrated a manual ear recognition system. He had examined more than ten thousand ears and observed that no two ears are exactly identical. The first technique for ear detection is introduced by Berge et al. [3]. It depends on building neighborhood graph from the edges of the ear. But its main disadvantages are first user interaction needs for contour initialization and second system is not enabled to discriminate the true ear edge and non-ear edge contours. Choras [4] used geometric feature for contour detection but this approach also suffered from the same problem of the selection of erroneous curves. Hurley et al. [5] have proposed a force field technique to detect ear. Although, reported algorithm has been tested on small background ear images, the results are established to be rather encouraging. Alvarez et al. [6] have detected ear from 2D face image using ovoid and active contour (snake) model. In this algorithm an initial approximated ear contour is needed to execute for ear localization. Ansari et al. [7] have described an ear detection technique that depends on outer ear helices edges. Therefore, it may fail when the outer helix edges are not clear. Yuan and Mu [8] have applied skin-color and contour information to detect ear. Proposed method assumes elliptical shape for ear and search the ellipse on the edges of the side face to get the location of the ear. But, considering elliptical ear shape may not be proper for all the individuals and cannot be used for detecting the ear universally. Sana et al. [9] have suggested ear detection technique based on template matching. In this work, different size of ear templates is maintained to locate ears in side face at different scales. In real world applications, ear occurs in various sizes and the off-line templates are not suitable to manage all the cases. Islam et al. [10] have proposed a cascaded AdaBoost based ear detection technique. The results of this approach are found to be rather promising for small database but it needs more training time for the large set of images. Prakash et al. [11, 12] have proposed an efficient distance transform and template based technique for ear localization. This approach is not efficient for illumination variations and noisy images. In recent paper of Prakash et al. [13] have used edge map of the side face image and convex hull criteria to construct a graph of connected components and largest connected component is the ear region. Here experimental results depend on quality of the input image and proper illumination condition.

This paper presents a new efficient scheme for automatic ear localization from side face images based on similarity measure of ear edge template and skin regions of side face using modified Hausdorff distance. As Hausdorff distance measure does not depends on pixels intensity and again ear template is a representative of different shape ears, the proposed approach is invariant to illumination variations, poses, shape, and occlusion. The remainder of this paper is organized as follows. Section 2 presents skin-color segmentation and the Hausdorff distance. Then Sect. 3 describes the proposed ear detection technique. Experimental results are discussed in Sect. 4 and finally, we draw some conclusions in Sect. 5.

2 Technical Background

This section briefly describes certain elementary methodologies, essential to build up the proposed ear localization technique. Section 2.1 defines a skin color model using samples from different color images, which is employed to compute skin-likelihood and used for skin segmentation. Section 2.2 concisely defines modified Hausdorff distance which is used for similarity measure between binary images.

2.1 Skin-Color Detection

The proposed approach includes color based skin segmentation method to detect only skin area of the side face images. Since, the ear is a part of the skin region, the search space for localizing the ear is reduced by excluding non-skin region. Skin color model suggested in [14] can be used for segmenting skin region from non-skin region. In our work, YCbCr color space [15] has been used to represent images. YCbCr color space is used to exclude luminance from blue and red colors then skin colors are separated in a small region. For this reason, skin color model exploits YCbCr color space. In skin detection method first, an image from RGB color space is converted to YCbCr color space then likelihood of each pixels are computed using Gaussian model $N(\mu, \Sigma)$. Each pixel of the image is represented using a color vector $c = (Cb, Cr)^T$ and the likelihood P (r, b) value can be calculated as follows:

$$P(r,b) = \frac{1}{\sqrt{2\pi|\Sigma|}} exp[\frac{1}{2}(c-\mu)\Sigma^{-1}(c-\mu)^T] \quad (1)$$

The skin-likelihood values are obtained using Eq. (1) to convert gray image into skin-likelihood image. Then skin region is segmented from non-skin region using skin segmentation. Skin segmentation is performed by thresholding the skin-likelihood image to obtain binary image. Finally, the binary side face image is dilated using morphological operator and multiplied with input color image to obtain skin segmented Image.

2.2 The Hausdorff Distance

The Hausdorff distance (HD) [16] is a promising similarity measure in many image matching applications. It measures the degree of resemblance between two binary images: the smaller the Hausdorff distance between edge point sets of two images the greater is the degree of similarity. Let Img and Tmp be the two sets of points in the input image and template, $Img = \{p_1, p_2, , p_{N_p}\}$ and $Tmp = \{q_1, q_2, , q_{N_q}\}$, with

each point p_i or q_j is the 2D pixel coordinates of the edge point extracted from the object of interest. The Hausdorff distance for the two point sets is defined as

$$d_h(Img, Tmp) = max(d(Img_i, Tmp), d(Tmp_j, Img)) \tag{2}$$

where d is a directed distance between two point sets Img and Tmp. It is used to compute the distance from p pixel p_i to all the points of the set Tmp, i.e.

$$d(p_i, Tmp) = min_{q_j \in Tmp}\{d(p_i, q_j)\} \tag{3}$$

Similarly, reverse distance from a pixel q_j to all the points of the set Img is computed as

$$d(q_j, Img) = min_{p_i \in Img}\{d(q_j, p_i)\} \tag{4}$$

The chessboard, city-block, or Euclidean distances are commonly used distance metric in 2D space. These metrics can be used to compute d between two points $p = (p_1, q_1)$ and $q = (p_2, q_2)$, i.e. $d_{chess}(p, q) = max(|p_1 - q_1|, |p_2 - q_2|)$, $d_{city}(p, q) = |p_1 - q_1| + |p_2 - q_2|$, $d_{euclidean}(p, q) = \sqrt{((p_1 - q_1)^2 + (p_2 - q_2)^2)}$. In this paper, the chessboard distance has been used to compute the Hausdorff distance between two edge point sets to replace the Euclidean distance for the improvement of the computation speed. It also does not require a one-to-one comparison between two images. Hence it is less affected by illumination variations than intensity based template matching technique. The Hausdorff distance technique is very sensitive to outlier points. The few outliers can perturb the distance to large value even through two objects are closer to each other, hence several modification to the conventional Hausdorff distance have been proposed for image matching. Dubuisson and Jain [17] reported a comparative study and revealed that modified Hausdorff distance (MHD) outperforms from other schemes for matching two images based on their edge points. This modified Hausdorff distance is given by

$$d_{mh}(Img, Tmp) = max(\frac{1}{N_p}\sum_{p_i \in Img} min_{q_j} d(p_i, q_j), \frac{1}{N_q}\sum_{p_i \in Tmp} min_{p_j} d(q_i, p_j)) \tag{5}$$

where N_p, N_q are edge points of image Img and Tmp.

3 Proposed Technique

Present technique includes four major parts: Skin-color segmentation, edge detection and pruning most of non-ear edges, comparing edge image patches of skin region and ear template to locate ear, finally validate true ear candidate using normalized cross correlation (NCC). Details of the proposed technique is illustrated in Fig. 1.

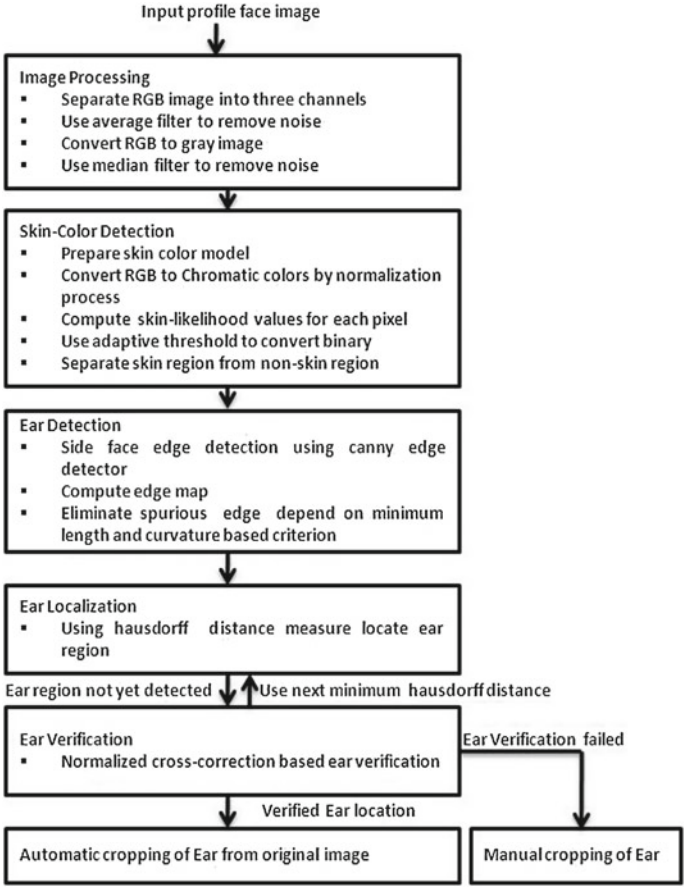


Fig. 1 Flow chart of the proposed technique

3.1 Skin-Color Segmentation

In this section, main objective is to detect skin region in a side face image. The first step of skin-color segmentation is to establish a skin color model which has been discussed in Sect. 2.1. By using an appropriate thresholding, gray scale image is segmented to binary image. However different people have different skin likelihood values, hence an adaptive thresholding process [14] is used to achieve the optimal threshold value. The optimal threshold is used to transform skin likelihood image to binary image. The binary image possesses holes due to presence of noise which is occupied using dilation morphological operation. Figure 2a shows a sample side face color image and its minimum edge set is obtained by processing with various intermediate stages. These are illustrated in Figs. 2b, c, d, e, and f.

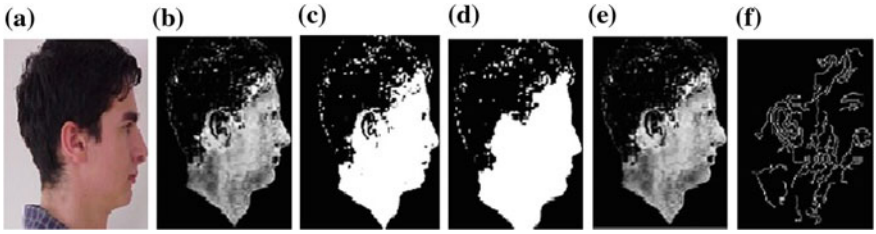


Fig. 2 Stages of skin segmentation **a** Input side face image, **b** gray scale image of skin-likelihood values, **c** Skin binary image with holes, **d** Skin binary image after dilation operation, **e** gray scale Skin image, **f** skin edge image

3.2 Edge Detection

In this work, Canny edge detector [18] is used to obtain edge map of the skin regions. In order to pruning non-ear edges various schemes have discussed. Figure 3 shows stages of pruning non-ear edges.

1. **Pruning spurious edges**
   The input edge image is checked to encounter edge junctions and accordingly edge lists are established. Subsequent, removing the edges whose length is shorter than threshold is represented as spurious edges. Generally, small spurious edges appear in the edge image due to noise or presence of hair. Let E be the set of all the edges present in the edge image and E_l be the set of edges present in set E after pruning edges whose edge length smaller than threshold as given by:

$$E = \{e|e \in Img, edgeimage\} \ \ and \ \ E_l = \{e|e \in Eandlength(e) > T_l\} \qquad (6)$$

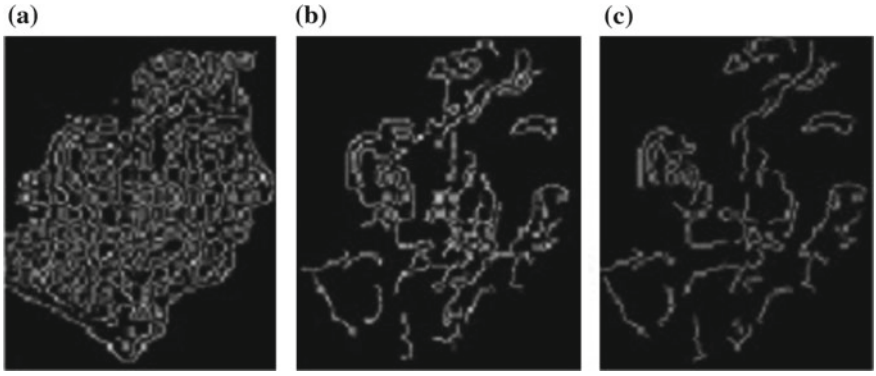


Fig. 3 Illustrates an example of edge pruning process in **a** side face edge image, **b** using minimum edge length and **c** using curvature measured in pixels

where Img is the side face skin edge image, length(e) represents the length of edge e and T_l is the threshold of allowable edge length.

2. **Piecewise linear representation of edges and pruning linear edges**
   Most of the non-ear edges are linear in nature, in this work linear edges are removed to reduce the number of edges from the edge image to speed up the matching operation. In the previous section, all the pixels present in an edge belong to the set E_l neither they are equally important or essential to represent the ear. In order to remove non-ear linear edges, line segments are approximated to the edge points. The linear edges are represented using only two pixels and non-linear edges are represented using line segments having more number of pixels. Fitting line segments algorithm receives edge points of set E_l and determines the locations of the maximum deviation for each edge and approximate lines between two points. Finally, a new set $E(l_s)$ is established for each edge is approximated using line segments. Hence the edges having two points these are non-ear edges and can be eliminated from the set E_{ls}. After pruning linear edges the set E_C is established which is defined as follows:

$$E_C = \{e | e \in E_{ls} \ \ and \ \ distance(e) > 2\} \tag{7}$$

   where function distance (e) counts number of times max deviation occurred and returns count value for each edge e.

3.3 Ear Localization

This section describes the ear localization which mainly comprises three subsections (1) ear template creation, (2) resizing ear template, and (3) edge based template matching using modified Hausdorff distance.

1. **Ear template generation**
   The main objective of ear template generation is to obtain an ear template which is a good representation of ear candidates available in the database. Surya Prakash et al. [11, 12] mentioned human ear can broadly be grouped into four kinds: triangular, round, oval, and rectangular. In this article taking above mentioned types of ear shapes into consideration a set of ear images are selected from the database manually off-line. The ear template Tmp is generated by averaging each pixel intensity values of all ear images and is defined as follows:

$$Tmp(i,j) = \frac{1}{N_{Img}} \sum_{k=1}^{n_{Img}} Img_k(i,j) \tag{8}$$

   where N_Img is the number of ear images selected manually for ear template creation and Imp_k is the K^{th} ear image. $Img_k(i,j)$ and Tmp (i, j) represent the $(i,j)^{th}$ pixel value of the K^{th} ear (Img_k) and template (Tmp) image respectively.

2. **Resizing ear template**
The ear template is resized into the same size of input ear image. Although Sana et al. [9] proposed template based approach ear localization where size of the template is fixed, but constant size template is not very often suitable to detect the ears of different size. Experimental evidence presented in Prakash et al. [11] that the shape of the ear is vary with respect to the shape of face image. Resizing ear template before template matching is defined as follows:

$$w_i^e = \frac{w_r^e}{w_r^f} * w_i^f \tag{9}$$

where w_r^f and w_i^f be the widths of the reference face image and the input ear image respectively. Similarly, w_r^e the width of the reference ear is same as the standard ear template width. Furthermore, to measure the height of the side face is difficult and incorrect because of extra skin pixels of neck person. In our work, an experiment has been made to find the relationship between the width and height of the ear using many cropped ear images. It is experiential that ratio between width and height of the human ear is greater than 0.5 (varies from 0.5 to 0.7) and is depended on the ratio between width of input side face and reference image. Hence, knowing the width of the ear using Eq. (9) and previous ratio value, height of the ear of the input side face image is estimated and found effective in most of the side face images given in Figs. 5 and 6. This feature allowed proposed approach fully automated for ear localization.
3. **Localization of ear using MHD**
After pruning the non-ear edges the set E_C established to define the new edge map of the side face image. The edge image of the ear template is compared with the same sized overlapping window of the input edge image using modified Hausdorff distance and the same process is repeated by moving the overlapping window over the skin region of the input edge image. Among all the distances of each block, the minimum Hausdorff distance block is selected and the corresponding region from the profile face is extracted. This region is expected as true ear and the claim is verified using the NCC criteria. When verification fails, then next minimum Hausdorff distance block is expected as the ear region and claim is verified again. This process carry on till ear is localized.
4. **Ear verification using NCC**
This section validates selected ear region as true ear candidate using normalized cross-correlation coefficient (NCC) method by verifying the correlation between the localized ear portion and ear template created off-line. The equation of NCC is written as

$$NCC = \frac{\sum_x \sum_y [Img(x,y) - \bar{I}mg][Tmp(x - x_c, y - y_c) - \overline{T}mp]}{\sqrt{\sum_x \sum_y [Img(x,y) - \bar{I}mg]^2} \sqrt{\sum_x \sum_y [Tmp(x - x_c, y - y_c) - \overline{T}mp]^2}} \tag{10}$$

where Img and Tmp are images of localized ear portion and ear template respectively. Similarly, $\overline{Img}$ and $\overline{Temp}$ is the average brightness values of the localized ear portion and ear template respectively. The NCC value lies between −1.0 and 1.0 when the NCC value is closer to 1 indicates better matching between localized ear and template. When the NCC value is typically above a predefined threshold the localization termed as true localization and otherwise, false localization.

4 Experimental Results and Discussion

We tested the proposed technique using two databases, namely CVL face database [19] and University of Notre Dame database (Collection E) [20]. CVL is library for image and data processing using graphics processing units (GPUs). This database contains 114 person's face images with 7 images from each subject. Images in the database are frontal view of the ears of both left and right for each person, hence total 456 side face images are contained in the database. In this work, 100 right side faces are used for experimentation. All images are of resolution of 640 × 480 with JPEG format captured by Sony Digital Mavica under uniform illumination, and with projection screen in background. The database contains 90 % of men and 10 % women side face images. Next, Collection E (UND-E) of University of Notre Dame database has been tested for the proposed scheme. It contains 462 side face images of 118 subjects and 2–6 samples per subject. In the Collection E database, images are captured on different days with various pose in dissimilar background and light conditions. The proposed approach was experimented on 100 side face images chosen from the CVL database of the right profile face images and results illustrated in Fig. 4. Similarly, Fig. 5 illustrates ear localization results of 462 profile face images of UND-E database. Authors in [11, 12] applied template matching techniques for ear localization based on pixel correspondences in turn, the performance of these

Fig. 4 Illustration of some ear localization outputs of the test images in the experiments using proposed approach based on modified Hausdorff distance

Fig. 5 Illustration of some ear localization outputs of the UND-E test images simulated using proposed approach based on modified Hausdorff distance

Table 1 Accuracy comparison of reported technique in [13] and proposed approach

Data set	# of Test images	Ear detection accuracy (%)	
		Reported technique in [13]	Proposed approach
CVL	100	88	91
UND-E	464	92.79	94.54

techniques largely depend on illumination variation and background condition in the side face image.

In this work, reported technique [13] has been implemented and tested because authors have used different datasets. Table 1 exhibits results of the reported technique in [13] and our proposed approach. In the experiment, for frontal ear images from two data bases where full ear structure is clearly visible as a separate larger size of connected component then ear is completely localized in [13] and proposed approach. However, the participation of noise edges with cluster of edges of ear portion leads wrong ear localization in [13] whereas our approach performs much better ear localization except few cases when more noise edges are present in connection with edges of ear portion. Next, as reported in [13] the larger size connected component is considered as true ear candidate for localization but sometimes there may exist a larger size of non-ear connected component. Similarly, it is also observed in the results of both database that localization of the ear was accurate in many side face images of different size and shape. On the other hand, effectiveness of our proposed approach entirely depends on successful detection of skin region of the side face. Figure 6 revealed some of the partial ear detection have occurred because of poor illumination and noise due to resemblance of hair color with skin color. The performance in terms of accuracy is described as:

$$Accuracy = \frac{Number\ \ of\ \ true\ \ ear\ \ detection}{Number\ \ of\ \ test\ \ sample} \times 100$$

Results in terms of accuracy obtained for two mentioned databases. The accuracy for CVL face database was found 91 and 94.54 % for Collection E database. It has observed that accuracy obtained for CVL face database is not found satisfactory

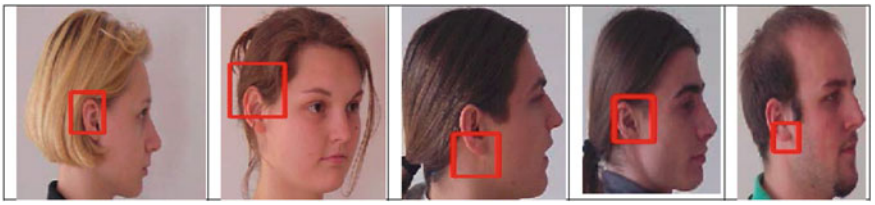


Fig. 6 Illustration of some partial ear detection in presence of less ear edges

because of poor illumination and similarity of background and hair color with skin color. Similarly, accuracy for the Collection E profile face database is encouraging even in case of partially occluded by hair in side faces but especially performance is poor because of similarity of hair color with side face of the test samples.

5 Conclusion

In this paper, we present an automated ear localization technique to detect ear from human side face images. The main contributions are two fold: first to separate the skin-color region from non-skin region using skin color model and second to locate ear within that skin-region using modified Hausdorff distance. Experiment is simulated on CVL and UND-E databases and extensive evaluations show that results are promising. The proposed approach is simple and robust for ear detection without any user intervention. This performance should encourage future research direction for ear localization method using variants of the Hausdorff distance measures.

References

1. A. Bertillon, La Photographie Judiciaire: Avec Un Appendice Sur La Classification Et L'Identification Anthropometriques, Gauthier-Villars, Paris, (1890).
2. A.V. Iannarelli, "Ear identification," in Proceedings of International Workshop Frontiers in Handwriting Recognition, Paramont Publishing Company, Freemont, California, (1989).
3. M. Burge and W. Burger, "Ear biometrics in computer vision," In Proceedings of ICPR, vol. 2, pp. 822–826, (2000).
4. Michal Choras, "Ear Biometrics Based on Geometrical Feature Extraction," Lecture Notes in Computer Science, pp. 51–61, (2004).
5. D. J. Hurley, M. S. Nixon, and J. N. Carter, "Force Field Feature Extraction for Ear Biometrics," Computer Vision and Image Understanding, vol. 98, pp. 491–512, (2005).
6. L. Alvarez, E. Gonzalez, and L. Mazorra "Fitting ear contour using an ovoid model," In Proceedings of ICCST, pp. 145–148, (2005).
7. S. Ansari and P. Gupta, "Localization of ear using outer helix curve of the ear," In Proceedings of ICCTA, pp. 688–692, (2007).
8. L. Yuan and Z.-C. Mu, "Ear detection based on skin-color and contour information," In Proceedings of ICMLC, vol. 4, pp. 2213–2217, (2007).

9. A. Sana, P. Gupta, and R. Purkait, "Ear biometric: A new approach," In Proceedings of ICAPR, pp. 46–50, (2007).
10. S.M.S. Islam, M. Bennamoun, and R. Davies, "Fast and fully automatic ear detection using cascaded adaboost," In Proceedings of IEEE Workshop on Applications of Computer Vision (WACV' 08), pp. 1–6, (2008).
11. Surya Prakash, J. Umarani, and P. Gupta, "Ear localization from side face images using distance transform and template matching," in Proceedings of IEEE Int'l Workshop on Image Proc. Theory, Tools and Application, (IPTA), Sousse, Tunisia, pp. 1–8, (2008).
12. Surya Prakash, J. Umarani, and P. Gupta, "A skin-color and template based technique for automatic ear detection," in Proceedings of ICAPR, India, pp. 213–216, (2009).
13. Surya Prakash, J. Umarani, and P. Gupta, "Connected Component Based Technique for Automatic Ear Detection," in Proceedings of the 16th IEEE Int'l Conference of Image Processing (ICIP), Cairo, Egypt, pp. 2705–2708, (2009).
14. J. Cai, and A. Goshtasby, "Detecting human faces in color images," Image and Vision Computing, 18(1), pp. 63–75, (1999).
15. G. Wyszecki and W.S. Styles, "Color Science: Concepts and Methods, Quantitative Data and Formulae," second edition, John Wiley & Sons, New York (1982).
16. D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, "Comparing images using the Hausdorff distance," IEEE Trans. Pattern Anal. Mach. Intell. 850–863, (1993).
17. M.P. Dubuisson and A.K. Jain, "A modified Hausdorff distance for object matching," In ICPR94, Jerusalem, Israel, pp. A:566–568, (1994).
18. J. Canny, "A computational approach to edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence 8(6), 679–698, (1986).
19. Peter Peer, "CVL Face Database," Available: http://www.lrv.fri.uni-lj.si/facedb.html.
20. University of Notre Dame Profile Face Database, Collection E, http://www.nd.edu/~cvrl/CVRL/DataSets.html.

Sclera Vessel Pattern Synthesis Based on a Non-parametric Texture Synthesis Technique

Abhijit Das, Prabir Mondal, Umapada Pal, Michael Blumenstein and Miguel A. Ferrer

Abstract This work proposes a sclera vessel texture pattern synthesis technique. Sclera texture was synthesized by a non-parametric based texture regeneration technique. A small number of classes from the UBIRIS version: 1 dataset was employed as primitive images. An appreciable result was achieved which solicits the successful synthesis of sclera texture patterns. It is difficult to get a huge collection real sclera data and hence such synthetic data will be useful to the researchers.

Keywords Sclera · Synthesis · Pattern · Texture · Biometrics

1 Introduction

Sclera is the white region with blood vessel patterns around the eyeball. Recently, as with other ocular biometric traits, sclera biometrics has gained in popularity [1–11]. Some recent investigations performed on multi-modal eye recognition (using iris and sclera) show that iris information fusion with sclera can enhance the biometric applicability of iris biometrics in off-angle or off-axis eye gaze. To

A. Das (✉) · M. Blumenstein
Institute for Integrated and Intelligent Systems, Griffith University, Queensland, Australia
e-mail: abhijit.das@griffithuni.edu.au

M. Blumenstein
e-mail: m.blumenstein@griffith.edu.au

P. Mondal · U. Pal
Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India
e-mail: prrabirmondal@gmail.com

U. Pal
e-mail: umapada@isical.ac.in

M.A. Ferrer
IDeTIC, University of Las Palmas de Gran Canaria, Las Palmas, Spain
e-mail: mferrer@dsc.ulpgc.es

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_22

establish this concept, it is first necessary to assess the biometric usefulness of the sclera trait independently (Fig. 1).

Moreover the research conducted on this subject is very limited and has not been extensively studied across a large proportion of the population. Therefore, to-date the literature related to sclera biometrics is still in its infancy and little is known in regard to its usefulness in personal identity establishment for large populations. It can be inferred from recent developments in the literature that a number of independent research efforts have explored the sclera biometric and serval dataset have been proposed, which are either publicly available or are proprietary datasets. Efforts were also made to nurture the various challenges which reside in processing the sclera trait towards personal identity establishment. It can be noted from the literature that the datasets developed are with a limited population of a maximum of 241 individuals. Growing the population is a tough task and moreover it also depends on the availability of volunteers. On some occasions, they may be available for a particular session and may not be available in the next session, which again can bring inconsistency to the dataset. These types of instances can be found in the datasets proposed in the literature. Hence establishing this trait for a larger population and generating a larger representative dataset is an open research problem.

In the biometric literature for larger population data collection, data synthesis is a proposed solution. Data synthesis refers to the artificial regeneration of traits by means of some computer vision or pattern synthesis based regeneration technique. Several such techniques have been proposed in the literature for iris biometrics [12] (texture pattern on the eye ball employed for biometrics). In order to mitigate the above mentioned problem in sclera biometrics, similar to the iris biometric, we propose a sclera biometric synthesis technique. The sclera contains the white area of the eye along with vessel patterns that appear as a texture pattern. Therefore we have applied texture regeneration theory for this application.

The organization of the rest of the paper is as follows: The concept of our proposed method is presented in Sect. 2. In Sect. 3 our experimental results along with the dataset are described, as well as a preliminary discussion on our experiments. Conclusions and future scope are presented in Sect. 4.

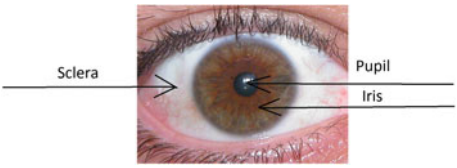


Fig. 1 Colour image of an eye consisting of pupil, iris and sclera area

2 Proposed Technique

Non-parametric Texture Synthesis using Efros and Leung's algorithm [13] is applied in this work as a remarkably simple approach. First, initialize a synthesized texture with a 3 × 3 pixel "seed" from the source texture. For every unfilled pixel which borders some filled pixels, find a set of patches in the source image that most resemble the unfilled pixel's filled neighbors. Choose one of those patches at random and assign to the unfilled pixel a color value from the center of the chosen patch, and repeat until the texture is achieved.

The detailed steps of our implementation for this method is as follows:

Step 1: Take a primitive Image (PI).
Step 2: A pixel is chosen randomly from PI. Let the pixel be p(x, y) as shown in (Fig. 2a).
Step 3: A matrix (SI), having dimensions the same as PI is created and initialized with zeros.
Step 4: Seed Size = 3.
Step 5: A 3 × 3 block (as seed size = 3) is chosen from SI in such a way that its top left most element's position is the middle element's position of SI (as described in Fig. 2b).
Step 6: Similarly in Fig. 3a. a 3 × 3 block is selected from PI in such a way that its top left most pixel is p(x, y) which was obtained in Step 3.
Step 7: The 3 × 3 block of SI is replaced by the 3 × 3 block of PI, (Fig. 3b).
Step 8: Another matrix (SI-1) is created with the same dimensions as SI and all its elements are initialized with zeros.
Step 9: A 3 × 3 block is obtained from (SI-1) by using the same method described in step 6, and the block is initialized with ones (Fig. 4a).
Step 10: Another matrix SI-2 is formed by dilating SI-1 by 1 pixel. So the neighbors of the block are assigned with ones (Fig. 4b).
Step 11: A column-wise neighborhood operation is performed on SI-2 as follows: By sliding a 3 × 3 window over every element of SI-2 and the corresponding element is replaced by the value obtained by summing all the elements residing in the window (Fig. 5).

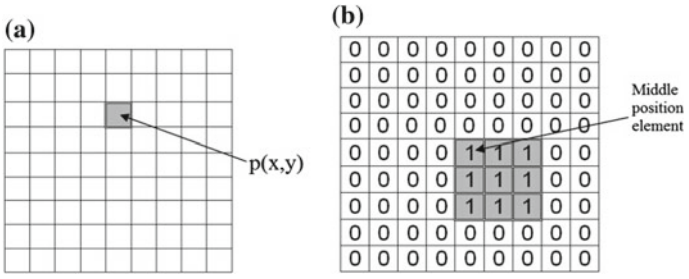


Fig. 2 **a** Primitive Image (PI). **b** SI Matrix

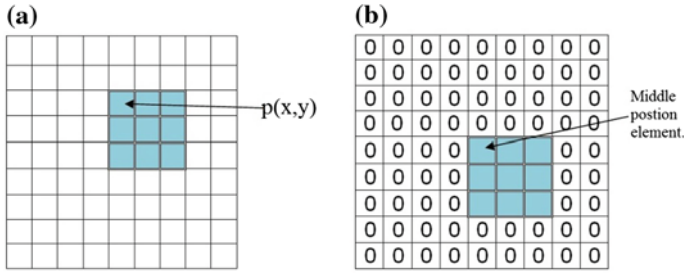


Fig. 3 **a** 3 × 3 block is selected from Primitive Image(PI), **b** SI Matrix with replaced 3 × 3 block from PI

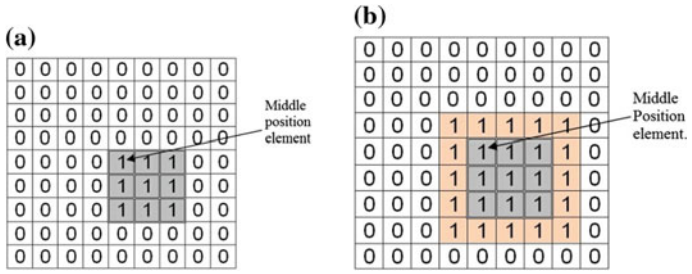


Fig. 4 **a** SI-1 Matrix. **b** SI-2 Matrix

Step 12: The positions of neighbor elements are listed after taking their values in descending order (Fig. 6).

Step 13: Elements of SI having the same position as that obtained from Step_12 are considered.

Step 14: A window (we consider here window size is 3 × 3 but in our experiments we consider it as 39 × 39) is placed on every considered element in such a way that the elements should be the middle element of the window (Fig. 7).

Step 15: After placing the window on every element of SI, the elements within the window are matched position-wise with the corresponding elements of PI, and an element is randomly chosen where the Match Error < Max Match Error threshold. We considered 0.1 as the value of Max Error Threshold.

Step 16: Let element q(x1, y1) of SI be an element that satisfies step_15, then we assign,

$$\mathrm{SI}(\mathrm{R}, \mathrm{C}) = \mathrm{PI}(\mathrm{x2}, \mathrm{y2}),$$

Where,

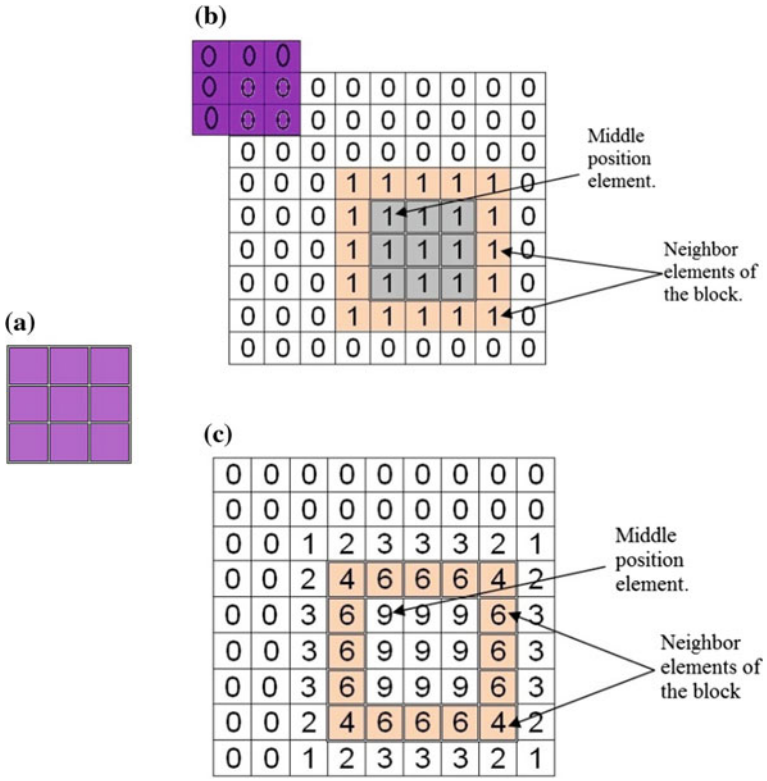


Fig. 5 **a**: 3 × 3 Window, **b** Column-wise neighborhood operation on SI-2 matrix with window, **c** SI-2 matrix after column-wise neighborhood operation

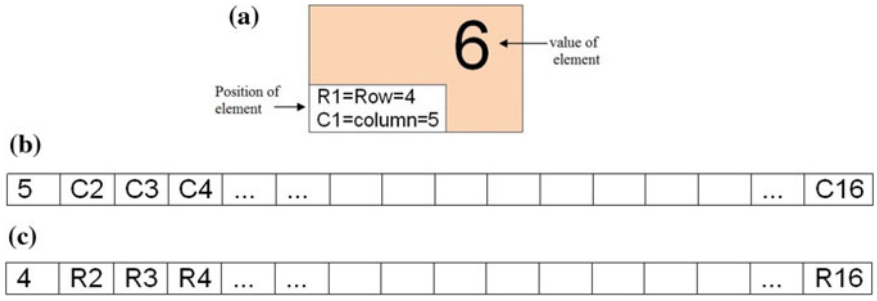


Fig. 6 **a** A neighbor element of an SI-2 matrix along with its row/column position, **b**, **c** Respective row, column position list of neighboring elements in the SI-2 matrix

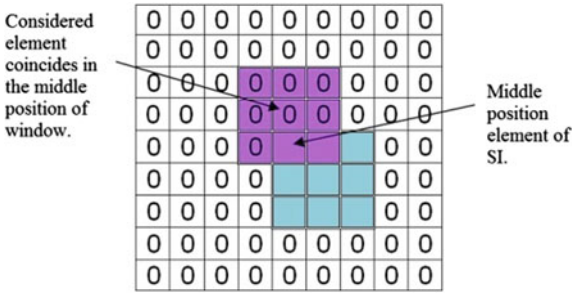


Fig. 7 SI matrix with middle pixel matched

$$x2 = (x1 + (\text{window size}/2)) \text{ and}$$
$$y2 = (y1 + (\text{window size}/2))$$

Step 17: SI-1(R, C) = 1.
Step 18: Step_13–Step_17 are iterated until all the elements of SI have the same position obtained from Step_11.
Step 19: Step_10–Step_18 are performed until all elements of SI-1 are assigned with 1.
Step 20: Finally we get the SI as the synthesis image of PI.

3 Experimental Details

Details about the implementation and the setup of the experiments are discussed here, before presenting the detailed results.

3.1 Dataset

In order to evaluate the performance of the proposed method, the UBIRIS version 1 database [14] was utilized in these experiments. This database consists of 1877 RGB images taken in two distinct sessions (1205 images in session 1 and 672 images in session 2), from 241 identities and images are represented in the RGB color space. The database contains blurred images and images with blinking eyes. Both high resolution images (800 × 600) and low resolution images (200 × 150) are provided in the database. All the images are in JPEG format. A few examples from session 1 are given below in Fig. 8.

For our experiments, the first 10 identities from session 1 were considered. 5 samples from each identity were selected and their sclera vessel patterns were manually cropped as shown below in Fig. 9. For each real image, a synthesized

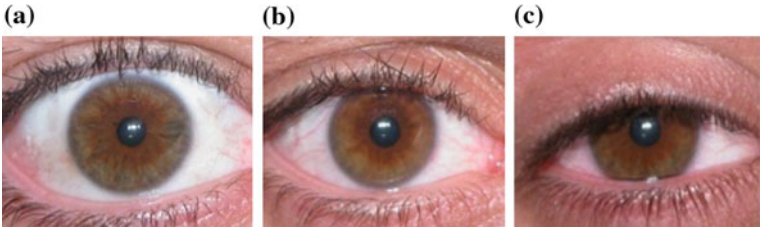


Fig. 8 Different quality eye images of Session 1 from UBIRIS version 1

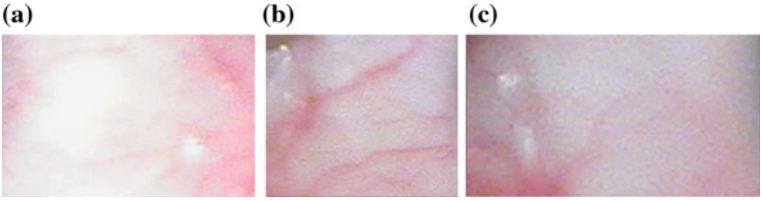


Fig. 9 Manually cropped vessel patterns from eye images of session 1 UBIRIS version1

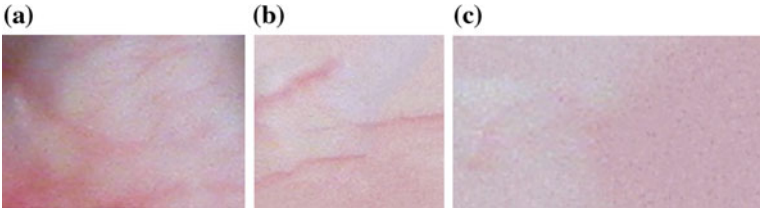


Fig. 10 Synthetic sclera vessel patterns from eye images of session 1 from UBIRIS version 1

image was generated. The synthesised images generated from the corresponding images in Fig. 9 as primitive images are shown in Fig. 10.

3.2 Experimental Setup

In order to establish the usefulness of the synthesized sclera vein patterns as a biometric trait, we observed the outcomes of some experiments. We undertook to determine the accuracy of the trait as a biometric trait. In order to feature the vein pattern and classification, the feature extraction and classification technique proposed in [5] were employed. We performed two sets of imposter and genuine experiments. In the first set we trained the system with 3 primitive images and tested it with the rest of the 2 primitive images. Maintaining the same protocol, the experiments for synthetic images was performed. In another set of experiments, the

system was trained with 5 synthetic images and tested with 5 primitive images and vise-versa.

For the first set of experiments, scores 10 * 2 for FRR and 10 * 9 * 2 scores for FAR statistics were obtained, whereby 10 * 5 scores for FRR and 10 * 9 * 5 scores for FAR statistics were obtained for the second set of experiments.

3.3 Results

The results of the experiments performed are described in this subsection in Table 1.

The first set of experiments i.e. with the primitive image (as training and test sets) or synthetic image (as training and test sets) was performed to establish the usefulness of the synthetically-generated pattern as a biometric trait. The second set of experiments was performed to find the inter-class variance in between each class of synthetic images. In both the scenarios, the desired result is achieved which can be depicted from the above table and the ROC curve in Fig. 11.

Table 1 The results of the experiments performed

Train set	Test set	Identification accuracy in %	Verification accuracy in %
Primitive image	Primitive image	100	100
Synthetic image	Synthetic image	100	100
Primitive image	Synthetic image	31	90
Synthetic image	Primitive image	48	75

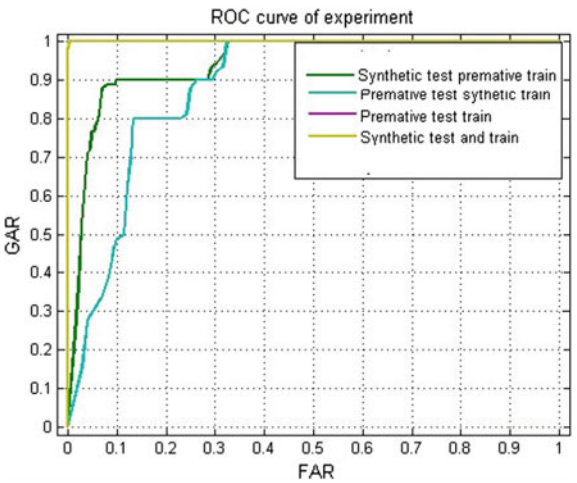


Fig. 11 ROC curve of the experiments

3.4 Discussion

This work is an initial investigation on sclera pattern synthesis. Here the images from the original (or the primitive) version are cropped manually and used for sclera vessel pattern synthesis. Although satisfactory results were achieved in the proposed experimental setup, the time complexity of the implementation is found to be high. The experiments were performed on a cluster server machine in a Linux environment using Matlab 2015. Therefore it can be easily assumed that it will take more time to generate the total vessel patterns of the eye. Future efforts to minimize this time will be an open research area for this field.

4 Conclusions and Future Scope

In this work we proposed a non-parametric texture synthesis method to generate synthetic sclera vessel patterns. We employed a set of images from the UBIRIS version 1 dataset as primitive images. Appreciable results were achieved in the experiments, which indicates satisfactory synthesis using the method.

Future scope of our work will concentrate on synthesization of the sclera vessel patterns without amanual cropping technique to generate the primitive images and to reduce the time complexity of the implementation.

References

1. Derakhshani, R., A. Ross, A., Crihalmeanu, S.: A New Biometric Modality Based on Conjunctival Vasculature. Artificial Neural Networks in Enineering (2006) 1–6.
2. Das, A., Pal, U., Ballester, M., F., A., Blumenstein, M.: A New Method for Sclera Vessel Recognition using OLBP. Chinese Conference on Biometric Recognition, LNCS 8232 (2013) 370–377.
3. Das, A., Pal, U., Ballester, M., F., Blumenstein, M.: Sclera Recognition Using D-SIFT. 13th International Conference on Intelligent Systems Design and Applications (2013) 74–79.
4. Das, A., Pal, U., Blumenstein M., Ballester, M., F.: Sclera Recognition—A Survey. Advancement in Computer Vision and Pattern Recognition (2013) 917 –921.
5. Das, A., Pal, U., Ballester, M., F., Blumenstein, M.: Fuzzy Logic Based Sclera Recognition. FUZZ-IEEE (2014) 561–568.
6. Das, A., Pal, U., Ballester M., F., Blumenstein, M.: Multi-angle Based Lively Sclera Biometrics at a Distance. IEEE Symposium Series on Computational Intelligence (2014) 22–29.
7. Das, A., Pal, U., Ballester, M., A., F., Blumenstein, M.: A new efficient and adaptive sclera recognition system. Computational Intellig. in Biometrics and Identity Management. IEEE Symposium (2014) 1–8.
8. Das, A., Pal, U., Blumenstein, M., Ballester, M., A., F.: Sclera Segmentation Benchmarking Competition (2015). http://www.ict.griffith.edu.au/conferences/btas2015.
9. Crihalmeanu, S., Ross., A.: Multispectral scleral patterns for ocular biometric recognition. Pattern Recognition Letters, Vol. 33 (2012) 1860–1869.

10. Zhou, Z., Du, Y., Thomas, N., L., Delp, E., J.: Quality Fusion Based Multimodal Eye Recognition. IEEE International Conference on Systems, Man, and Cybernetics (2012) 1297–1302.
11. Das, A., Kunwer, R.., Pal, U., M. A. Ballester, M., A., F., Blumenstein, M.: An online learning-based adaptive biometric system. Adaptive Biometric Systems: Recent Advances and Challenges (2015) 73–95.
12. Galbally, J., Ross, A., Gomez-Barrero, M., Fierrez, J., Ortega-Garcia, J.,: Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms, Computer Vision and Image Understanding, Vol. 117, n. 10, (2013) 1512–1525.
13. Efros, A., Leung, T.,: Texture Synthesis by Non-Parametric Sampling. In Proceedings of International Conference on Computer Vision, (1999), 1033–1038.
14. Proença, H., Alexandre, L., A.: UBIRIS: A noisy iris image database, Proceed. of ICIAP 2005 —Intern. Confer. on Image Analysis and Processing, 1: (2005) 970–977.

Virtual 3-D Walkthrough for Intelligent Emergency Response

Nikhil Saxena and Vikas Diwan

Abstract After various cases of terrorist-attacks and other emergency situations across the globe; the need towards the development of virtual 3D walkthrough for the important premises are progressively on the rise (Lee, Zlatanova, A 3D data model and topological analyses for emergency response in urban areas, [1]). In contrast to the conventional 2D layout of the premises; the 3D modeling adds another dimension to make a quick and intelligent emergency response to such situations. Modern 3D modeling and game development tools have given the capability to rapid development of such applications with near real-time rendering capacity. In this paper, we examine the potential of using virtual 3D walkthrough for the important installations that aim at facilitating the security personnel and decision-makers to effectively carryout their training, strategic and operational task in case of emergency or otherwise.

Keywords Computer graphics · 3D walkthrough · Emergency response · Safety and security

1 Introduction

Virtual reality technology, has introduced a new spatial metaphor with very interesting applications on intelligent navigation, social behavior over virtual worlds, full body interaction, virtual studios, etc. [2]. With the evolution of modern graphics processors, memory bandwidth capabilities and advanced optimization techniques now it's possible to add realism in real time 3D graphics.

With the advent of science and technology the form of threat is changing its face and so in today's scenarios planning and training is of paramount importance.

N. Saxena (✉) · V. Diwan
Bhabha Atomic Research Centre, Mumbai 400085, India
e-mail: nikhils@barc.gov.in

V. Diwan
e-mail: vikasd@barc.gov.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_23

In case of any such adverse situation the first and foremost thing is to protect human life and hence the evacuation is the first and most crucial logical step. The local security authorities should not only be familiar with the architectural topologies of the campus and its buildings along with all possible entry/exit points including emergency and make shift entries but should also be able to communicate the same to outside agencies in the minimum and most efficient manner.

Generally, the important installations all across the globe are being protected from fire, terrorist attacks and other dire conditions by the means of various physical protection systems such as different types of fire hydrants, CCTV cameras, security personnel, logging and access control mechanisms for incoming or outgoing individuals or vehicles. These systems seems to be reasonable measures for combating the emergency situations but when the access of these systems also gets forbidden due to the conditions, it becomes very difficult to look for the way out.

In case of the hostile conditions like full/partial restriction to normal entry to the premise; the usual practice and many a times the only option for the security personnel and decision-makers is to prepare the further course of action for restoring the state to the normalcy by referring 2D plan-elevation layouts of the targeted premise. Sometimes the security personnel are the outside agencies which are called for the rescue operations and are unaware of the topology of the campus. In such context, referring 2D plan-elevation layouts do not really give much of its insight; many a times it may lead to false presumptions.

Let us understand this problem through a scenario. Assume few people working on a floor consisting of corridors and rooms in some building of the campus. Figure 1 shows its 2D layout where three people in red (or marked 1), blue (or marked 2) and yellow (or marked 3) are shown working and it looks that these

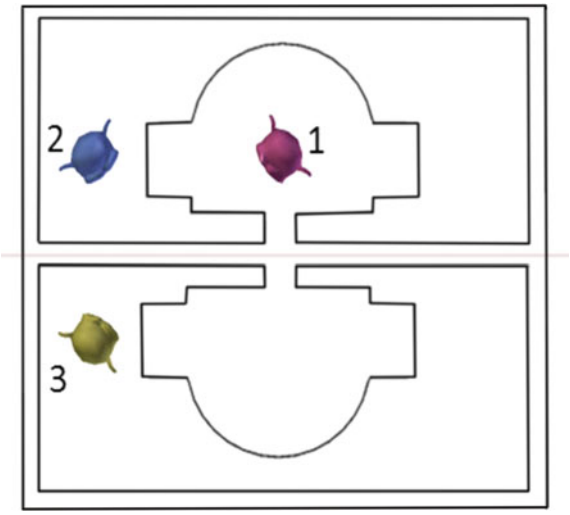


Fig. 1 2D Plan layout with 3 characters seems to be in talking range

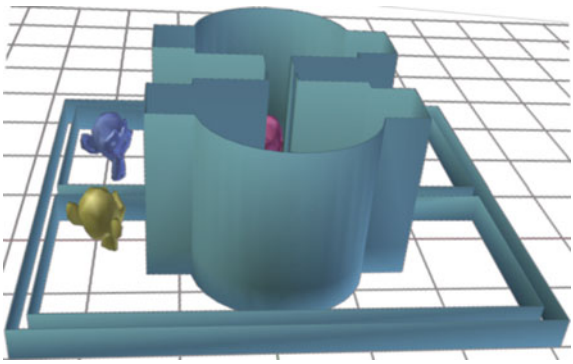


Fig. 2 3D layout depicting actual situations

people can observe each other but Fig. 2 shows the actual setup where red person cannot observe the other two persons. These figures in which loss of relevant information in 2D due to its deficiency of one dimension in comparison to 3D are good case of explaining better usability of 3D layouts over 2D layouts:

The usability of 3D over 2D layouts surely places a requirement for building a 3D solution for the security personnel and decision-makers to effectively carryout their training, strategic and operational task in case of emergency or otherwise. Our basic objective of developing 3D walkthrough is one such solution provided by the virtual reality which, keeping natural calamities, sabotage and terrorist threats in today's global scenario, is a need of an hour.

2 Previous and Related Work

There are various papers on applications of virtual reality ranging from different physical, biological and chemical simulation systems to 3D walkthroughs.

The paper [3] researches on the virtual natural landscape walkthrough technology, realizes natural landscape walkthrough by using Unity 3D, describes some common methods of making sky, landform, trees, flowers and water in the virtual walkthrough. The main focus of this paper was on rendering natural shapes than man-made objects which have convention methods of modeling. Natural shapes are highly relies on particle systems and other optimized techniques like bill-boarding etc.

The paper [4] introduces how to realize the driving training simulation system by using computer software aiming at the car driving training. The paper presents that their system can partly replace the actual operation training, designers hope to improve training efficiency and reduce the training cost. The main focus of this paper was simulation besides modeling. It also uses physics library for simulating the interaction between rigid bodies and hence enhances the scope of training efficiency at lower cost.

Another paper [5], emphasis was brought on the creation of worlds that represented real places or buildings, where the user could be able to access various kinds of information interacting with objects or avatars and travel at the same time in the virtual space. That paper presented the architecture and implementation of a virtual environment based on the campus of Guangxi University of Technology using Java language for the user-interfaces, VRML for the 3D scripting and visualization and HTML for all other multimedia pages (2D visualization) of the system. The main focus of this paper was 3D walkthrough inside a campus. This paper has presented the web based and platform independent solution but has not talked much about vegetation and terrain of the campus.

By and large, the 3D walkthrough applications were developed to focus on either the larger scenery with exteriors of the campus buildings using 3D modeling or only the realistic interiors using panorama photographs. Also, we have not come across any paper for its relevance on security perspectives. Therefore, in our approach; we presented the both interiors and exteriors with near actual surroundings.

3 Methodology

We have developed the hybrid approach of creating the virtual walkthrough for the premise where not only buildings' exterior facades but also interiors of the buildings including the corridors and its rooms are given due importance by modeling the entire architectural composition. The inputs for the development work are the

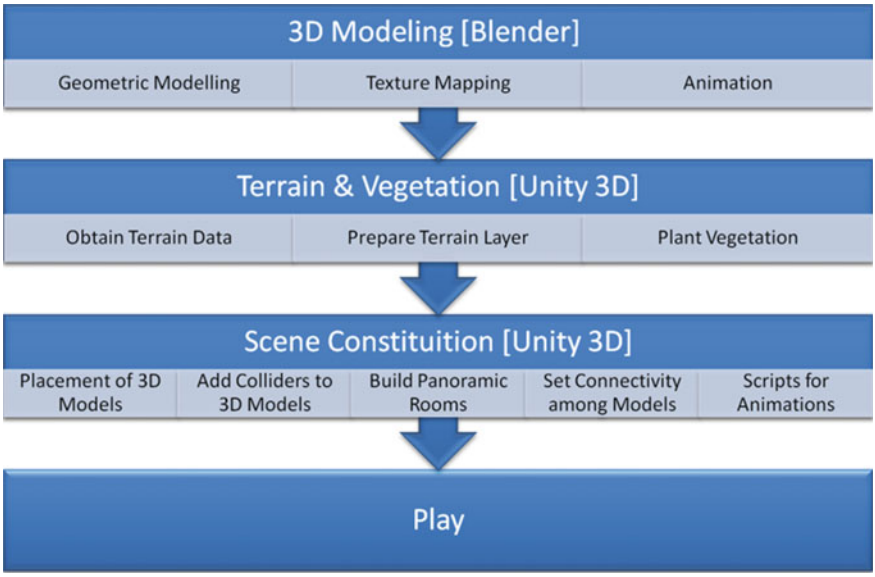


Fig. 3 Flow diagram

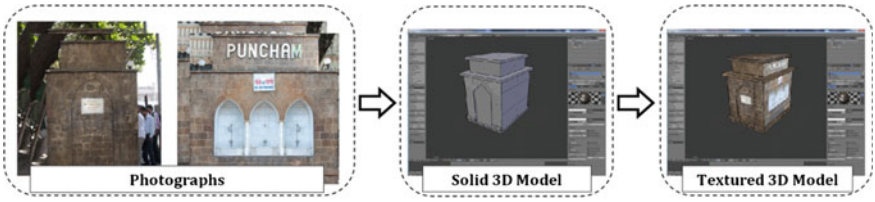


Fig. 4 Work-flow of 3D model development

architectural 2D drawings/layouts, high-definition photographs and 360° panoramic photographs. Figure 3 illustrates the approach for the development of 3D walk-through application.

Real world 3D entities like buildings, gates/doors and various interior-objects are modeled as per their true dimensions and wrapped with actual textures (derived through various high-definition photographs) using Blender software. The general work-flow of the modeling is shown in Fig. 4.

Player can also walk on the corridors of the building which may have doors for entries to its corresponding rooms. One such corridor, modeled and textured using Blender, is also shown in Fig. 5.

The rooms in the buildings are developed using 360° panoramic photographs taken through specialized fish-eye lenses. The entry to the room is controlled by a java script which detects the approach of the player towards its door via collision detection feature provided in Unity 3D. That script works as per the following flow diagram in Fig. 6.

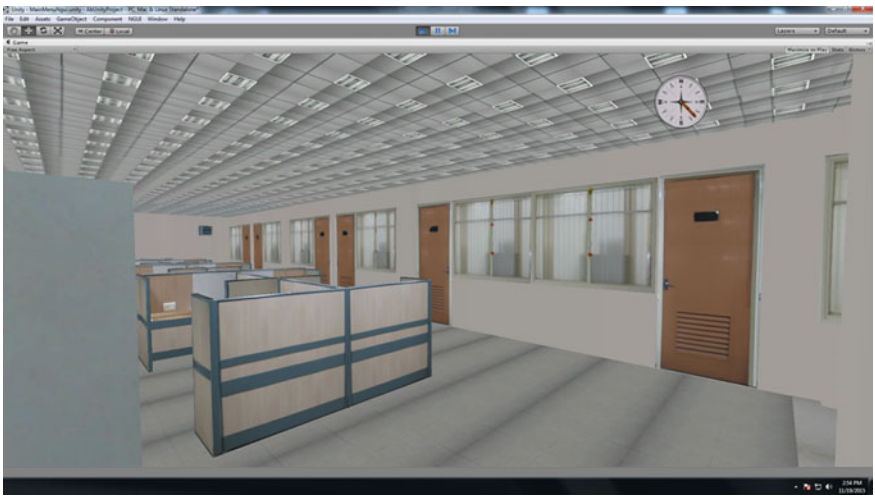


Fig. 5 3D textured building interior

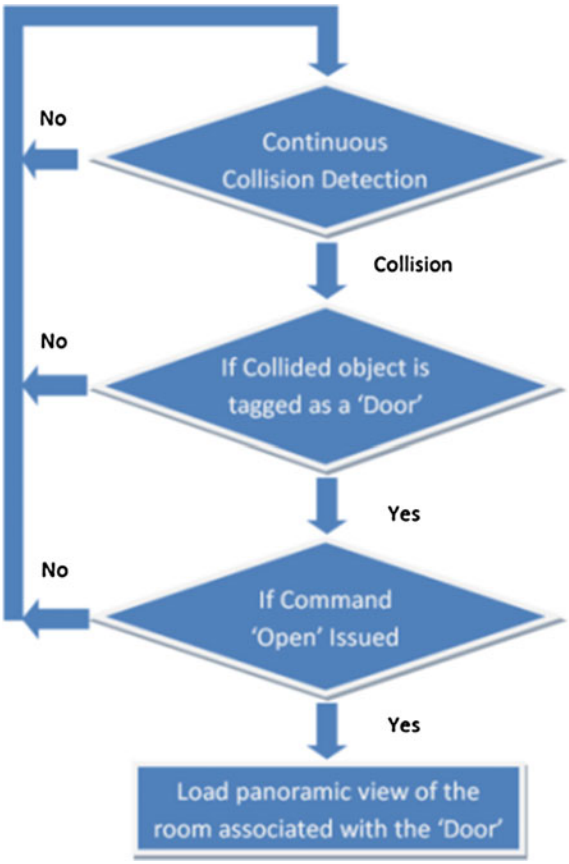


Fig. 6 Flow diagram of script for entry into the Room

These models along with the surrounding vegetation are placed on terrain at their actual geographical positions (obtained from Google maps) in Unity 3D Scene. This scene consists of the panoramic rooms in the buildings of the campus along with its major electrical, mechanical and security related work and equipment housed in the respective buildings. It also has the first person controller by which the end-user can navigate throughout the virtual plant by using just the mouse and keyboard/joystick.

3.1 Development Framework

The software and hardware configurations for development system of the 3D walkthrough application are described under the following subsections.

Software

Modern 3D modeling tools are user-friendly and rich in features which make the development task of 3D models (of almost all physical entities) reasonably smoother. In our case, we used the Blender software which is a free and open source 3D animation suite, cross-platform and runs equally well on Linux, Windows and Mac computers. Its interface uses OpenGL to provide a consistent experience [6].

Recent game development tools have given the capability to rapid development of virtual applications with near real-time rendering powers. The Unity engine is far and away the dominant development platform for creating games and interactive 3D and 2D experiences like training simulations and medical and architectural visualizations, across mobile, desktop, web, console and other platforms [7]. Unity comes in two variants one is free with some limited features and other one is paid with comprehensive features, called Unity-pro [8]. Free version of Unity 3D is selected for development of our 3D walkthrough.

The built-in material and script in Unity 3D can support most virtual reality application requirements, and numerous three-dimensional file formats, such as: 3DS MAX, Maya, Cinema-4D, making virtual scene production convenient and quick; DirectX and OpenGL have highly optimized rendering graphics pipeline, it makes the simulation result like true and the built-in physics engine can better support designing of driving simulation [4]. With the software and hardware performance improving, developers can truly focus on representing the scene when making landscape walkthrough by using Unity 3D [3].

Hardware

There is no constraint on going for particularly one hardware configuration but systems with powerful GPU, higher capacity RAMs (CPU/GPU) are better candidates. Our 3D walkthrough application is developed on system having the hardware configuration as per Table 1.

Though the above table lists a workstation of higher configuration but the application is tested on different systems having comparatively lower configurations too without much performance deviation.

Table 1 Hardware Configuration of development system

CPU	Intel Xeon 2.40 GHz
CPU RAM	8 GB
GPU	NVIDIA QUADRO FX 5800
GPU RAM	4 GB
Screen resolution	1920 × 1200

3.2 Weighing up with Google Earth

As Google earth, a popular virtual globe, map and geographical information application is extensively known and used throughout the world, it is important as well as interesting to compare our approach vis. a vis. Google earth approach. Google started with a textured based globe where large amount of satellite data was stitched and textured on globe modeled which is regularly updated [9]. Then in last three years basic models of the 3D buildings appeared on the big cities and finally Google street view, which is actual panoramic view of the street, taken around every 10 m or so are provided to the user. User has the option to switch between the 3D model and panoramic view. With this option Google has taken the hybrid approach of rendering the models of Google earth.

Our approach has also taken the hybrid path albeit with following differences:

- The interiors of the buildings can be accessed and corridors and rooms of the buildings can be navigated like a person actually visiting inside the building.
- Navigation is more like walking running and jumping than moving left, right, up down or rotating, which is closer to actual navigation.
- Collision detection is provided to all objects and buildings where entry/exit is permitted only through gates and windows or any other open place. In Google earth one can just cross the walls of building.
- Dynamic sensing of objects is provided in our application so that in future based on this intelligence scenarios can be generated for training.

Obviously there are stark differences in our approach though both Google earth and we have taken the hybrid approach of rendering 3D models. This is due to different aims and goals of the application, and also we have luxury to not have to handle such a vast amount of world data.

4 Results

The application has been tested on various systems having different hardware and software configurations. Figure 7 displays one scene from the application rendering on one of the testing workstation.

The systems under tests are observed to render 40–60 frames per second depending upon its hardware/software configurations which is considered to be good rendering performance by any standard.

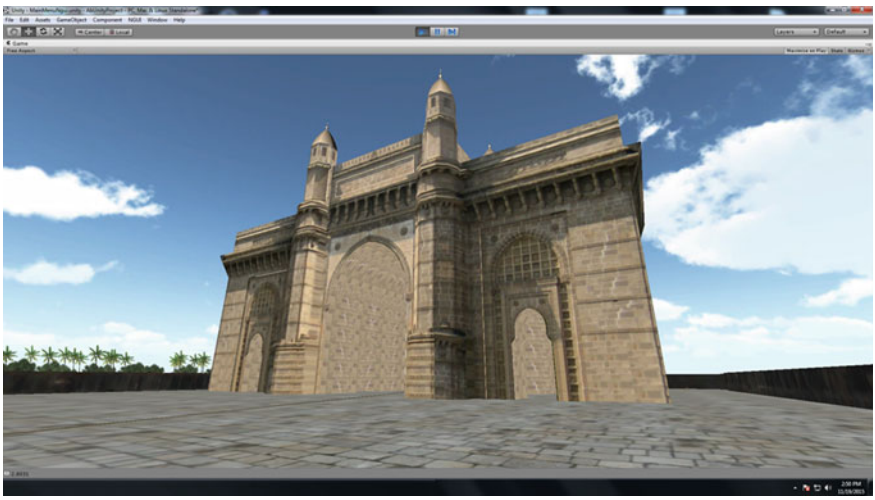


Fig. 7 Scene in 3D walkthrough application

5 Conclusion and Future Work

In this paper, we described the development of virtual tour product that may be used as a supplement to an actual visit of the building and its campus. In cases where an actual visit is inconvenient or prohibited this may act as a substitute. The basic objective of developing 3D walkthrough is to facilitate the offsite security response forces to respond to any emergency situation for enabling them to effectively carryout their operational tasks or evacuation planning in such times.

Besides evacuation planning as a usage of this application; many intelligent scenarios, based on the experiences, feedbacks and actual events, can be built in the subsequent versions of this 3D walkthrough. Building models rendered during different lighting conditions, different time of the day, different seasons, virtual flood, fire, hostage conditions, radioactivity released conditions etc. can be simulated and worth of the training can be increased with the experience and technology improvement.

References

1. J. Lee, S. Zlatanova, 'A 3D data model and topological analyses for emergency response in urban areas'.
2. Nikos A, Spyros V and Themis P, 'Using virtual reality techniques for the simulation of physics experiments', Proceeding of the 4th Systemics, Cybernetics and Informatics International Conference, Orlando, Florida, USA, 2000, pp 611.

3. Kuang Yang, Jiang Jie, Shen Haihui, 'Study on the Virtual Natural Landscape Walkthrough by Using Unity 3D', IEEE International Symposium on Virtual Reality Innovation 2011 19–20 March, Singapore.
4. Kuang Yang, Jiang Jie, 'The designing of training simulation system based on unity 3D', Int. Con. on Intelligent Computation Technology and Automation (ICICTA), 2011, Issue Date: 28–29 March 2011.
5. Ziguang Sun, Qin Wang, Zengfang Zhang, 'Interactive walkthrough of the virtual campus based on vrml'.
6. http://www.blender.org/about/.
7. http://unity3d.com/public-relations.
8. http://unity3d.com/unity/licenses.
9. http://www.gearthblog.com/blog/archives/2014/04/google-earth-imagery.html.

Spontaneous Versus Posed Smiles—Can We Tell the Difference?

Bappaditya Mandal and Nizar Ouarti

Abstract Smile is an irrefutable expression that shows the physical state of the mind in both true and deceptive ways. Generally, it shows happy state of the mind, however, 'smiles' can be deceptive, for example people can give a smile when they feel happy and sometimes they might also give a smile (in a different way) when they feel pity for others. This work aims to distinguish spontaneous (felt) smile expressions from posed (deliberate) smiles by extracting and analyzing both global (macro) motion of the face and subtle (micro) changes in the facial expression features through both tracking a series of facial fiducial markers as well as using dense optical flow. Specifically the eyes and lips features are captured and used for analysis. It aims to automatically classify all smiles into either 'spontaneous' or 'posed' categories, by using support vector machines (SVM). Experimental results on large UvA-NEMO smile database show promising results as compared to other relevant methods.

Keywords Posed · Spontaneous smiles · Feature extraction · Face analysis

1 Introduction

People believe that human face is the mirror/screen showing internal emotional state of the human body as and when it responds to the external world. This means that, what an individual thinks, feels or understands, etc., deep inside the brain, get imitated into the outside world through its face [7]. Facial smile expression undeniably plays a huge and pivotal role [1, 11, 25] in understanding social interactions within a community. People often give smile imitating the internal state of the body. For example, generally, people smile when they are happy or when sudden humorous

B. Mandal (✉) · N. Ouarti
Visual Computing Department, Institute for Infocomm Research, Singapore, Singapore
e-mail: bmandal@i2r.a-star.edu.sg

N. Ouarti
e-mail: nizarouarti@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_24

things happen/appear in front of them. However, people are sometimes forced to pose smile because of the outside pressure or external factors. For example, people would pose a smile even when they don't understand the joke or the humor. Sometimes people would also pose a smile even when they are reluctantly or unwillingly do or perform something in front of their bosses/peers [6].

Therefore being able to identify the type of smiles of individuals would give affective computing a deeper understanding of the human interactions. A large amount of research in psychology and neuroscience studying facial behavior demonstrate that spontaneous deliberately displayed facial behavior has differences both in utilized facial muscles and their dynamics as compared to posed ones [8]. For example, spontaneous smiles ar smaller in amplitude, longer in duration, slower in onset and offset times than posed smiles [3, 8, 22]. For humans, capturing such subtle facial movements is difficult and we often fail to distinguish between them. It is not surprising that in computer vision, algorithms developed for classifying such smiles usually fail to generalize to the subtlety and complexity of human posed and spontaneous affective behaviors [15, 25].

Numerous researchers asserted that dynamic features such as duration and speed of the smile play a part in differentiating the nature of the smile [11]. A spontaneous smile usually take longer time to reach from onset to apex and then offset as compared to a posed smile [5]. As for non-dynamic features, the aperture size of the eyes is found to be a useful clue and is generally of a higher value when extracted from a spontaneous smile as compared to a posed one. On the other hand, the symmetry in (or the lack of) movement of spontaneous and posed smiles do not produce significant distinction in identifying them and is therefore not much useful [21]. In [22] a multi-modal system using geometric features such as shoulder, head and inner facial movements are fused together and GentleSVM-sigmod is used to classify the posed and spontaneous smiles. He et al. in [10] proposed a technique for feature extraction and compared the performance using geometric and facial appearance features. Appearance based features are computed by recording statistics of overall pixel values of the image, or even using edge detection algorithm such as Gabor Wavelet Filter. Their comprehensive study shows that geometric features are generally more effective in detecting posed from spontaneous expressions [10].

A spatiotemporal method involving both natural and infrared face videos to distinguish posed and spontaneous expressions is proposed in [20]. Using temporal space and image sequences as volume, they extended the complete local binary patterns texture based descriptor into the spatiotemporal features to classify posed and spontaneous smiles. Dibeklioglu et al. in [4] used the dynamics of eyelid movements, distance measures and angular features in the changes of the eye aperture. Using several classifiers they have shown the superiority of eyelid movements over the eyebrows, cheek and lip movements for smile classification. Later in [5], they used dynamic characteristics of eyelid, cheek and lip corner movements for classifying posed and spontaneous smiles. Temporal facial information is obtained in [13] through segmenting the facial expression into onset, apex and offset which cover the entire duration of the smile. They reported good classification performance by using a combination of features extracted from the different phases.

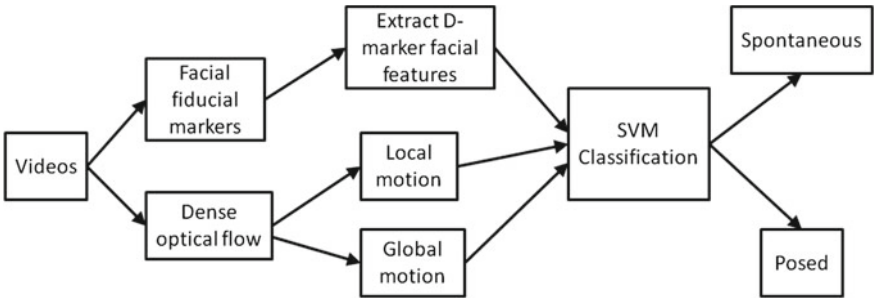


Fig. 1 Block diagram of the proposed system

The block diagram of our proposed method is shown in Fig. 1. Given smile video sequences of various subjects, we apply the facial features detection and tracking of the fiducial points over the entire smile video clip. Using D-markers, 25 important parameters (like duration, amplitude, speed acceleration, etc.) are extracted from two important regions of the face: eyes and lips. Smile discriminative features are extracted using dense optical flow along the temporal domain from the global (macro) motion and local (micro) motion of the face. All these information are fused and support vector machine (SVM) is then used as a classifier on these parameters to distinguish posed and spontaneous smiles.

2 Feature Extraction from Various Face Components

We use the facial tracking algorithm developed by Nguyen et al. in [17] to obtain the fiducial points on the face. The 21 tracking markers each are labeled and placed following the convention as shown in Fig. 2a. The markers are manually annotated in the first frame of each video by user input and thereafter it automatically tracks the remaining frames of the smile video, it is of good accuracy and precision as compared to other facial tracking software [2]. The markers are placed on important facial feature points such as eyelids and corner of the lips for each subject. The convention followed in our approach for selecting fiducial markers are shown in Fig. 2a.

2.1 Face Normalization

To reduce inaccuracy due to the subject's head motion in the video that can cause change in angle with respect to roll, yaw and pitch rotations, we use the face normalization procedure described in [5]. Let l_i represents each of the feature points used to align the faces as shown in Fig. 2. Three non-collinear points (eye centers and nose tip) are used to form a plane ρ. Eye centers are defined as $c_1 = \frac{l_1+l_3}{2}$ and

(a) Frame #1 (b) Frame #30 (c) Frame #58 (d) Frame #72

Fig. 2 **a** Shows the tracked points on the 1st frame, **b** shows the tracked points on 30th frame, **c** shows the tracked points on 58th frame and **d** shows the tracked points on 72nd frame on one subject. (Best viewed when zoomed in.)

$c_2 = \frac{l_4+l_6}{2}$. Angles between the positive normal vector N_ρ of ρ and unit vectors U on X (horizontal), Y (vertical), and Z (perpendicular) axes give the relative head pose as follows:

$$\theta = arccos \frac{U.N_\rho}{\|U\|\|N_\rho\|}, \text{ where } N = \overrightarrow{l_g c_2} \times \overrightarrow{l_g c_1}. \tag{1}$$

$\overrightarrow{l_g c_2}$ and $\overrightarrow{l_g c_1}$ denote the vectors from point l_g to points c_2 and c_1, respectively. $\|U\|$ and $\|N_\rho\|$ represents the magnitudes of U and N_ρ vectors respectively. Using the human face configuration, (1) can estimate the exact roll (θ_z) and yaw (θ_y) angles of the face with respect to the camera. If we start with the frontal face, the pitch angles (θ'_x) can be computed by subtracting the initial value. Using the estimated head pose, tracked fiducial points are normalized with respect to rotation, scale and translation as follows:

$$l'_i = [l_i - \frac{c_1 + c_2}{2}]R_x(-\theta'_x)R_y(-\theta_y)R_z(-\theta_z)\frac{100}{\epsilon(c_1 + c_2)}, \tag{2}$$

where l'_i is the aligned point. R_x, R_y and R_z denote the 3D rotation matrices for the given angles. $\epsilon()$ is the Euclidean distance measure. Essentially (1) constructs a normal vector perpendicular to the plane of the face using three points (nose tip and eye centers), then calculate the angle formed between X, Y and Z axis with regards to the normal vector of face plane. Thereafter, (2) process and normalize each and every point of the frame accordingly and set the interocular distance to 100 pixels with the middle point acting as the new origin of the face center.

2.2 D-Marker Facial Features

In the first part of our strategy, we focus on extracting the subject's eyelid and lips features. We first construct a amplitude signal variable based on the facial feature markers on the eyelid regions. We compute the amplitude of eyelid and lip end movements during a smile using the procedure described in [21]. Eyelid amplitude signals are computed using the eyelid aperture (D_{eyelid}) displacement at time t, given by:

$$D_{eyelid}(t) = \frac{\kappa(\frac{l^t_1+l^t_3}{2}, l^t_2)\epsilon(\frac{l^t_1+l^t_3}{2}, l^t_2) + \kappa(\frac{l^t_4+l^t_6}{2}, l^t_5)\epsilon(\frac{l^t_4+l^t_6}{2}, l^t_5)}{2\epsilon(l^t_1, l^t_3)} \quad (3)$$

where $\kappa(l_i, l_j)$ denotes the relative vertical location function, which equals to -1 if l_j is located (vertically) below l_i on the face, and 1 otherwise. The equation above uses the markers for eyelids namely 1–6 as shown in Fig. 2, to construct the amplitude signal that calculate the eyelid aperture size in each frame t. The amplitude signal D_{eyelid} is then further computed to obtain a series of features. In addition to the amplitudes, speed and acceleration signal are also extracted by computing the second derivatives of the amplitudes.

Smile amplitude is estimated as the mean amplitude of right and left lip corners, normalized by the length of the lip. Let D_{lip}(t) be the value of the mean amplitude signal of the lip corners in the frame t. It is estimated as

$$D_{lip}(t) = \frac{\epsilon(\frac{l^t_{10}+l^t_{11}}{2}, l^t_{10}) + \epsilon(\frac{l^t_{10}+l^t_{11}}{2}, l^t_{11})}{2\epsilon(l^t_{10}, l^t_{11})} \quad (4)$$

where l^t_i denotes the 2D location of the ith point in frame t. For each video of our subject we are able to acquire a 25-dimensional feature vectors based on the eyelids markers and lip corner points. Onset phase is defined as the longest continuous increase in D_{lip}. Similarly, the offset phase is detected as the longest continuous decrease in D_{lip}. Apex is defined as the phase between the last frame of the onset and the first frame of the offset. The displacement signals of eyelids and lip corners could then be calculated using the tracked points. Onset, apex and offset phases of the smile are estimated using the maximum continuous increase and decrease of the mean displacement of the eyelids and lip corners. The D-Marker is then able to extract 25 descriptive features each for eyelids and lip corner, so a vector of 50 features are obtained from each frame (using two frames at a time). The features are then concatenated and passed through SVM for training and classification.

2.3 *Features from Dense Optical Flow*

In the second phase of the feature extraction, we use our own proposed dense optical flow [19] for capturing both global and local motions appearing in the smile videos. Our approach is divided into four distinct stages that are fully automatic and does not require any human intervention. The first step is to detect each frame in which the face is present. We use our previously developed face, integration of sketch and graph patterns (ISG) eyes and mouth detectors for face recognition on wearable devices and human-robot-interaction [14, 23]. So we get the region of interest (ROI) for the face (as shown in Fig. 3, left, yellow ROI) with 100 % accuracy on the entire UvA-NEMO smile database [5]. In the second step, we determine the area corresponding to the

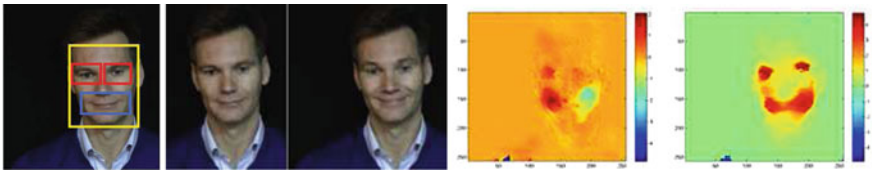


Fig. 3 *Left* Face, eyes and mouth detections. *Yellow ROI* for face detection, *red ROI* for eyes detection and *blue ROI* for mouth detection. *Middle* Two consecutive frames of a subject's smile video and *Right* their optical flows in x- and y-directions. (Best viewed in *color* and zoomed in.)

right eye, left eye in red ROI and mouth in blue ROI for which we get 96.9 % accuracy on the entire database.

In the third step, the optical flow is computed between the image at time t and at time $t + 1$ of the video sequence (see Fig. 3, middle). The two components of the optical flow are illustrated in Fig. 3, right, which shows the optical flow along the x-axis and the optical flow along the y-axis. Because we are using a dense optical flow algorithm, the time to process one picture is relatively important. To speed up the processing, we computed the optical flow only in the three ROI regions: right eye, left eye and mouth. The optical flow computed in our approach is a pyramidal differential dense algorithm that is based on the following constraint:

$$F = F_{smooth} + \beta F_{attach}, \tag{5}$$

where the *attach* term is based on thresholding method [24] and the regularization term (*smooth*) is based on the method developed by Meyer in [16], β is a weight controlling the ratio between the end attachment and the term control. Ouarti et al. in [19] proposed to use a regularization that do not use an usual wavelet but a non-stationary wavelet packet [18], which generalize the concept of wavelet for extracting optical flow information. We extend this idea for extracting fine grained information for both micro and macro motion variations in smile videos as shown in Fig. 4. Figure 5 shows the dense optical flows with spontaneous and posed smiles variations. In the fourth step, for each of the three ROIs, the median of the optical flow is determined that give a cue to the global motion of the area. An histogram is computed based on the optical flow that has 10 bins. The top three bins in term of cardinality are kept among all the bins. A linear regression is then applied to find the major axis of the point group for each of the three bins determined. In the end, for each ROI we obtain: the median value of the bin 1, the value of the bin 2 and the value of the bin 3. It also calculates the intercept and slope for points of bins 1, 2 and 3. These result in 60 features for each frame (using two consecutive frames in a smile video). SVM is then used on these features to classify the posed and spontaneous smiles.

The major advantage of this approach is that we can obtain useful smile discriminative features using a fully automatic analysis of videos, no marker are needed to be annotated by an operator/user. Moreover, rather than attempting to classify raw optical flow we design some processing to obtain a sparse representation of the optical

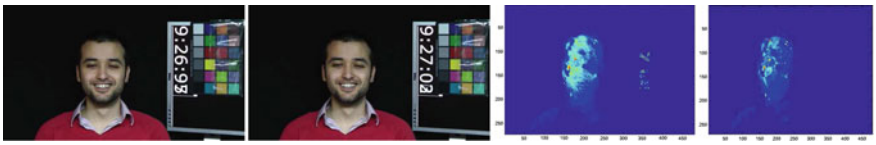


Fig. 4 Original images and their dense optical flows with their corresponding micro and macro motion variations of a subject. (Best viewed in *color* and zoomed in.)

Fig. 5 Original images and their dense optical flows with their corresponding spontaneous and posed smiles variations of a subject. (Best viewed in *color* and zoomed in.)

flow signal. This representation helps in classification by extracting only the useful information in low dimensions and speeds up the calculation of the SVM. Finally, information is not completely connected to the positioning of the different ROI knowing that this positioning may vary from one frame to another, it is dependent on the depth and highly variable depending on the individuals. Therefore a treatment which would be too closely related to the choice of the ROI would lead to non-consistent results.

3 Experimental Results

We test our proposed algorithm on UvA-NEMO Smile Database [5], it is the largest and most extensive smile (both posed and spontaneous) database with videos from a total of 400 subjects, (185 female, 215 male) aged between 8 to 76 years old, giving us a total of 1240 individual videos. Each video consists of a short segment of 3–8 s. The videos are extracted into frames at 50 frames per second. The extracted frames are also converted to gray scale and downsized to 480 $\times$ 270. In all the experiments, we split the database, in which 80 % is used as training samples and the remaining 20 % is used as testing samples. Binary classifier SVM with radial basis function as the kernel and default parameters as in LIBSVM [12], is used to form a hyperplane based on the training samples. When a new testing sample is passed into the SVM it uses the hyperplane to determine which class the new sample falls under. This process is then repeated 5 times using a 5-fold cross validation method. To measure the subtle differences in the spontaneous and posed smiles we compute the confusion matrices between the two smiles so as to find out how much accuracy we can obtain in using each of them in the actual and classified separately. The results from all 5 processes are averaged and shown in Tables 1, 2, 3, 4 and 5 and compared with other methods in Table 6.

Table 1 The overall accuracy (%) in classifying spontaneous and posed smiles using only the eyes features is 71.14 %. In bracket (·) shows accuracy using only the lips features as 73.44 %

Actual	Classified	
	Spontaneous	Posed
Spontaneous	60.1 (67.5)	39.9 (32.5)
Posed	17.5 (20.4)	82.5 (79.6)

Table 2 The overall accuracy (%) in classifying spontaneous and posed smiles using the combined features from eyes and lips is 74.68 %. (rows are gallery, columns are testing)

Actual	Classified	
	Spontaneous	Posed
Spontaneous	65.3	34.7
Posed	16.3	83.7

Table 3 The accuracy (%) in classifying spontaneous and posed smiles using our proposed X-directions dense optical flow is 59 %. In bracket (·) the accuracy using our proposed Y-directions is 63.8 %

Actual	Classified	
	Spontaneous	Posed
Spontaneous	57.8 (58.3)	42.2 (41.7)
Posed	39.8 (30.8)	60.2 (69.2)

Table 4 The accuracy (%) in classifying spontaneous and posed smiles using our proposed fully automatic system using X- and Y-directions of dense optical flow is 56.6 %

Actual	Classified	
	Spontaneous	Posed
Spontaneous	58.0	42.0
Posed	45.1	54.9

3.1 Results Using Parameters from the Facial Components

Table 1 and in bracket (·) show the accuracy rates in distinguishing spontaneous smiles from the posed ones using eyes and lips features respectively. The results show that the eye features play very crucial role in finding the posed smiles where as the lips features are important for spontaneous smiles. Overall we could obtain an accuracy of 71.14 and 73.44 % using eyes and lips features respectively. Table 2 shows the classification performance using combined features from eyes and lips. It is evident from the table that using these facial component features, pose smile can be classified better as compared to the spontaneous ones.

Table 5 The accuracy (%) in classifying spontaneous and posed smiles using our proposed fused approach comprising of both features from facial components and dense optical flow is 80.4 %

Actual	Classified	
	Spontaneous	Posed
Spontaneous	83.6	16.4
Posed	22.9	77.1

3.2 Results Using Dense Optical Flow

We use the features using dense optical flow as described in Sect. 2.3, the movement in both X- and Y-directions are recorded between every consecutive frames of each video. The confusion matrices are shown in Table 3, in bracket (·) and Table 4. It can be see from the tables that the performance of optical flow is lower as compared to the component based approach. However, the facial component based feature extraction method requires user initialization to find and track fiducial points, whereas the dense optical flow features are fully automatic. It does not require any user intervention, so it is more useful for practical applications like first-person-views (FPV) or egocentric views on wearable devices like Google Glass for improving real-time social interactions [9, 14].

3.3 Results Using Both Component Based Features and Dense Optical Flow

We combine all the features obtained from facial component based parameters and dense optical flow in to a single vector and apply SVM. Table 5 shows the confusion matrix using spontaneous and posed smiles. It can be seen that the performance of spontaneous smiles classification improved using features from dense optical flow. The experimental results in Table 5 show that both features from facial components and dense optical flows are important for improving the overall accuracy. Features from facial components (as shown in Table 2) are useful for encoding information arising from the muscle artifacts within a face, however, the regularized dense optical flow features helps in encoding fine grained information for both micro and macro motion variations in face smile videos. So combining them the overall accuracy has been improved.

3.4 Comparison with Other Methods

Correct classification rates (%) using various methods on UvA-NEMO are shown in Table 6. It is evident from the table that our proposed approach is quite competitive as compared to the other state-of-the-arts methodologies.

Table 6 Correct classification rates (%) on UvA-NEMO database

Method	Correct classification rate (%)
Pfister et al. [20]	73.1
Dibeklioglu et al. [4]	71.1
Cohn and Schmidt [21]	77.3
Eyelid Features [5]	85.7
Mid-level Fusion (voting) [5]	87.0
Ours Eye + Lips + dense optical flow	80.4

4 Conclusions

Differentiating spontaneous smiles from the posed ones is a challenging problem as it involves extracting subtle minute facial features and learning them. In this work we have analysed features extracted from facial component based parameters using fuducial points markers and tracking them. We have also obtained fully automatic features from dense optical flow on both eyes and mouth patches. It has been shown that the facial component based parameters give higher accuracy as compared to dense optical flow features for smile classification. However, the former requires initialization of the fiducial markers on the first frame and hence, it is not fully automatic. Dense optical flow has advantage that the features can be obtained without any manual intervention. Combining the facial components parameters and dense optical flow gives us highest accuracy for classifying the spontaneous and posed smiles. Experimental results on the largest UvA-NEMO smile database shows the efficacy of our proposed approach as compared to other state-of-the-arts methods.

References

1. Ambadar, Z., Cohn, J., Reed, L.: All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. Journal of Nonverbal Behavavior 33, 17–34 (2009)
2. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Incremental face alignment in the wild. In: CVPR. Columbus, Ohio, USA (2014)
3. Cohn, J., Schmidt, K.: The timing of facial motion in posed and spontaneous smiles. Intl J. Wavelets, Multiresolution and Information Processing 2, 1–12 (2004)
4. Dibeklioglu, H., Valenti, R., Salah, A., Gevers, T.: Eyes do not lie: Spontaneous versus posed smiles. In: ACM Multimedia. pp. 703–706 (2010)
5. Dibeklioglu, H., Salah, A.A., Gevers, T.: Are you really smiling at me? spontaneous versus posed enjoyment smiles. In: IEEE ECCV. pp. 525–538 (2012)
6. Ekman, P.: Telling lies: Cues to deceit in the marketplace, politics, and marriage. WW. Norton & Company, New York (1992)
7. Ekman, P., Hager, J., Friesen, W.: The symmetry of emotional and deliberate facial actions. Psychophysiology 18, 101–106 (1981)

8. Ekman, P., Rosenberg, E.: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System. Second ed. Oxford Univ. Press (2005)
9. Gan, T., Wong, Y., Mandal, B., Chandrasekhar, V., Kankanhalli, M.: Multi-sensor self-quantification of presentations. In: ACM Multimedia. pp. 601–610. Brisbane, Australia (Oct 2015)
10. He, M., Wang, S., Liu, Z., Chen, X.: Analyses of the differences between posed and spontaneous facial expressions. Humaine Association Conference on Affective Computing and Intelligent Interaction pp. 79–84 (2013)
11. Hoque, M., McDuff, D., Picard, R.: Exploring temporal patterns in classifying frustrated and delighted smiles. IEEE Trans. Affective Computing 3, 323–334 (2012)
12. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification (2010)
13. Huijser, M., Gevers, T.: The influence of temporal facial information on the classification of posed and spontaneous enjoyment smiles. Tech. rep., Univ. of Amsterdam (2014)
14. Mandal, B., Ching, S., Li, L., Chandrasekha, V., Tan, C., Lim, J.H.: A wearable face recognition system on google glass for assisting social interactions. In: 3rd International Workshop on Intelligent Mobile and Egocentric Vision, ACCV. pp. 419–433 (Nov 2014)
15. Mandal, B., Eng, H.L.: Regularized discriminant analysis for holistic human activity recognition. IEEE Intelligent Systems 27(1), 21–31 (2012)
16. Meyer, Y.: Oscillating patterns in image processing and in some nonlinear evolution equations. The Fifteenth Dean Jacquelines B. Lewis Memorial Lectures, American Mathematical Society (2001)
17. Nguyen, T., Ranganath, S.: Tracking facial features under occlusions and recognizing facial expressions in sign language. In: International Conference on Automatic Face & Gesture Recognition. vol. 6, pp. 1–7 (2008)
18. Ouarti, N., Peyre, G.: Best basis denoising with non-stationary wavelet packets. In: International Conferenc on Image Processing. vol. 6, pp. 3825–3828 (2009)
19. Ouarti, N., SAFRAN, A., LE, B., PINEAU, S.: Method for highlighting at least one moving element in a scene, and portable augmented reality (Aug 22 2013), http://www.google.com/patents/WO2013121052A1?cl=en, wO Patent App. PCT/EP2013/053,216
20. Pfister, T., Li, X., Zhao, G., Pietikainen, M.: Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In: ICCV Workshop. pp. 868–875 (2011)
21. Schmidt, K., Bhattacharya, S., Denlinger, R.: Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. Journal of Nonverbal Behavavior 33, 35–45 (2009)
22. Valstar, M., Pantic, M.: How to distinguish posed from spontaneous smiles using geometric features. In: In Proceedings of ACM ICMI. pp. 38–45 (2007)
23. Yu, X., Han, W., Li, L., Shi, J., Wang, G.: An eye detection and localization system for natural human and robot interaction without face detection. TAROS pp. 54–65 (2011)
24. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l1 optical flow. In: In Ann. Symp. German Association Patt. Recogn. pp. 214–223 (2007)
25. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. PAMI 31(1), 39–58 (2009)

Handling Illumination Variation: A Challenge for Face Recognition

Purvi A. Koringa, Suman K. Mitra and Vijayan K. Asari

Abstract Though impressive recognition rates have been achieved with various techniques under the controlled face image capturing environment, making recognition more reliable under uncontrolled environment is still a great challenge. Security and surveillance images, captured in open uncontrolled environments, are likely subjected to extreme lighting conditions like underexposed, and overexposed areas that reduce the amount of useful details available in the collected face images. This paper explores two different preprocessing methods and compares the effect of enhancement in recognition results using Orthogonal Neighbourhood preserving Projection (ONPP) and Modified ONPP (MONPP), which are subspace based methods. Note that subspace based face recognition techniques are highly sought after in recent times. Experimental results on preprocessing techniques followed by face recognition using ONPP and MONPP are presented.

Keywords Illumination variation · Dimensionality reduction · Face recognition

1 Introduction

Although face recognition algorithms are performing exceptionally well under controlled illumination environments, it is still a major challenge to reliably recognize a face under pose, expression, age and illumination variations. Illumination variation is most common while capturing face images. Lighting condition, camera position, face position all lead to change in illumination. Such illumination changes result in

P.A. Koringa (✉) · S.K. Mitra
DA-IICT, Gandhinagar 382007, Gujarat, India
e-mail: 201321010@daiict.ac.in

S.K. Mitra
e-mail: suman_mitra@daiict.ac.in

V.K. Asari
University of Dayton, Dayton, OH, USA
e-mail: vasari1@udayton.edu

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_25

a major source for recognition errors especially for appearance based techniques. The task of face recognition algorithm is to identify an individual accurately despite of such illumination variations. Two face images of same person can seem visually very different under various illumination intensities and directions in [1]. It has been shown that the variations in two face images of same person captured under different illumination conditions are larger than the face images of two different person, which makes face recognition under illumination variation a difficult task. To handle such cases, several approaches are used such as preprocessing and normalization, invariant feature extraction or face modeling [2]. In preprocessing based methods, several image processing techniques are performed on image to nullify illumination effects to some extent. Gamma correction, histogram equalization [3, 4] and logarithm transforms [5] are some of these image processing techniques.

In this paper, two different preprocessing techniques are applied to compensate illumination variation on face images taken under extreme lighting conditions and the recognition performance is compared using Orthogonal Neighbourhood Preserving Projection(ONPP) [6] and Modified ONPP(MONPP) [7]. ONPP and MONPP are mainly dimensionality reduction techniques which learn the data manifold using subspace analysis. Recently both ONPP and MONPP are efficiently used for face recognition task [6, 7]. Detailed experiments of preprocessing to nullify the illumination variation for face recognition have been performed on various benchmark face databases such as The extended Yale-B database [8] and CMU PIE face database [9]. Face recognition results of ONPP and MONPP are compared and presented.

In the next section, preprocessing techniques are explained in detail, followed by the dimensionality reduction algorithm MONPP explained in Sect. 3. Section 4 consists of experimental results followed by conclusion in Sect. 5.

2 Preprocessing

For better face recognition under uncontrolled and varying lighting conditions, the features useful for discrimination between two different faces need to be preserved. The shadows created in face images due to different lighting directions result in loss of facial features useful for recognition. A preprocessing method must increase the intensity in the areas those are under-exposed (poorly illuminated) and lower the intensity in the areas those are over-exposed (highly illuminated) simultaneously, while keeping the moderately illuminated area intact. Following two subsections discuss two different preprocessing techniques.

2.1 Locally Tuned Inverse Sine Nonlinear (LTISN) [10]

An enhancement technique for colour images proposed in [10] takes care of such extreme illumination conditions using a series of operations and a nonlinear inten-

sity transformation performed on images to enhance a colour image for better visual perception. The intensity transformation function is based on previous research suggested in [11].

In this paper, we have tested the nonlinear intensity transformation based on inverse sine function enhancement on grayscale face images having high illumination irregularities for better recognition. This nonlinear enhancement technique is a pixel by pixel approach where, the enhanced intensity value is computed using the inverse sine function with a locally tunable parameter based on the neighbourhood pixel values. The intensity range of the image is rescaled to [0 1], followed by a nonlinear transfer function (Eq. 1).

$$I_{enh}(x,y) = \frac{2}{\pi} sin^{-1}(I_n(x,y)^{\frac{q}{2}}) \tag{1}$$

where, $I_n(x,y)$ is the normalized intensity value at pixel location (x,y) and q is the locally tunable control parameter. In the darker area where intensity needs to be increased, the value of q should be less than 1, and the over bright area where intensity needs to be suppressed, the value of q should be greater than 1. Figure 1 shows the transformation function with the value of q ranging from 0.2 to 5 for intensity range [0 1]. The red curve shows transformation for q equal to 1, green curves show q less than 1, which enhances darker region of image and blue curve shows q greater than 1, which suppresses higher intensity in the over-exposed region of an image. The curve of the transformation function used for a pixel is decided by the value of q based on its neighbourhood.

The value of q is decided by the tangent function based on mean normalized intensity values, which is determined by averaging three Gaussian filtered smooth images. These smooth images are found using three different Gaussian kernels of size $M_i \times N_i$. Normalized Gaussian kernels are created as below:

$$kernel_i(n_1,n_2) = \frac{h_g(n_1,n_2)}{\sum_{n_1}\sum_{n_2} h_g(n_1,n_2)} \tag{2}$$

$$h_g(n_1,n_2) = e^{\frac{-(n_1^2+n_2^2)}{2\sigma^2}} \tag{3}$$

where, ranges of n_1 and n_2 are $[-\lfloor\frac{M_i}{2}\rfloor, \lfloor\frac{M_i}{2}\rfloor]$ and $[-\lfloor\frac{N_i}{2}\rfloor, \lfloor\frac{N_i}{2}\rfloor]$ respectively. Here, window size $M_i \times N_i$ is set to 6×6, 10×10 and 14×14, experimentally. Symbol i indicates which Gaussian kernel is being used and σ is to be $0.3(\frac{M_i}{2}-1)+0.8$

The Gaussian mean intensity at pixel (x, y) is calculated using

$$I_{M,i}(x,y) = \sum_{m=-\frac{M_i}{2}}^{\frac{M_i}{2}} \sum_{n=-\frac{N_i}{2}}^{\frac{N_i}{2}} I(m,n) kernel_i(m+x, n+y) \tag{4}$$

The mean intensity image I_{M_n} is then obtained by averaging these three filtered images. The mean intensity value is normalized to range [0 1] and based on intensity value in I_{M_n} at location (x, y), the tunable parameter q is determined using

$$q = \begin{cases} tan(\frac{\pi}{C_1} I_{M_n}(x,y)) + C_2 & I_{M_n}(x,y) \geq 0.3 \\ \frac{1}{C_3} ln(\frac{1}{0.3} I_{M_n}(x,y)) + C_4 & I_{M_n}(x,y) < 0.3 \end{cases} \tag{5}$$

where C_1, C_2, C_3 and C_4 are determined experimentally [10].

Transformation function for different values of q is presented in Fig. 1. As it can be seen from Fig. 1, extreme bright pixels having high mean intensity value nearer to 1, will be suppressed using high q values and dampen the high-intensity pixels. Extreme dark pixel with mean values nearer to zero will be enhanced with small q values and it will positively boost the low-intensity values.

2.2 DoG Based Enhancement [12]

This Preprocessing technique employs a series of operations including Gamma correction, Difference of Gaussian filtering and an equalization to enhance the over-lit or under-lit area of an image. Gamma correction is a nonlinear transformation that replaces gray-level i with i^{γ}, for $\gamma > 0$. It enhances the local dynamic range of the image in dark regions due to under-lit conditions, at the same time suppressing the range of bright regions and highlights due to over-lit conditions.

Gamma correction alone is not capable to remove the influence of all intensity variations, such as shadowing effects. Shading induced due to the facial surface structure is useful information for recognition, whereas the spurious edges generated due

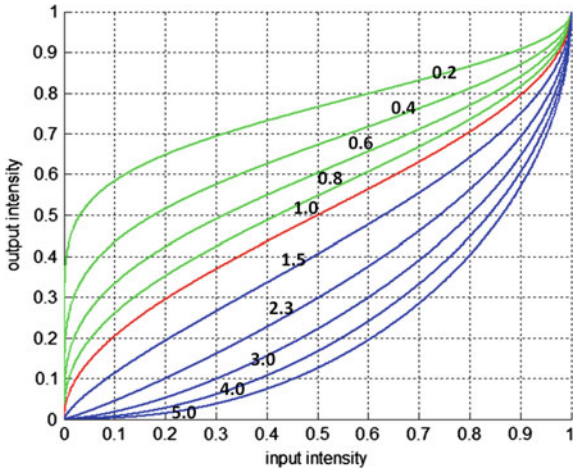


Fig. 1 Nonlinear inverse sine transformation with parameter q varying from 0.2 to 5 for intensity range [0 1]

to shading introduce false information for recognition. Band-pass filtering can help to retain useful information in face images while get rid of unwanted information or misleading spurious edge like features due to shadows. DoG filtering is a suitable way to achieve such a bandpass behavior. As DoG name suggests, it is basically a difference of $2D$ Gaussian filters G_{σ_1} and G_{σ_2} having different variances (Outer mask is normally 2−3 times broader than the inner mask). The inner Gaussian G_{σ_2} is typically quite narrow (usually variance $\sigma_2 \leq 1$ pixel essentially works as high pass filter), while the outer Gaussian G_{σ_1} is 2−4 pixels wider, depending on the spatial frequency at which low-frequency information becomes misleading rather than informative. Values for σ_1 and σ_2 are set to 1 and 2 respectively based on experiments carried out on face databases [12].

The DoG as an operator or convolution kernel is defined as

$$DoG \cong G_{\sigma_1} - G_{\sigma_2} = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_1} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{-\frac{x^2+y^2}{2\sigma_2^2}}\right)$$

Processed image still typically contains extreme values produced by highlights, small dark regions etc. Following approximation is used to rescale the gray values present in preprocessed image.

$$I(x,y) \leftarrow \frac{I(x,y)}{(mean(min(\tau, I(x,y))^{\alpha})^{\frac{1}{\alpha}}} \tag{6}$$

here, $I(x,y)$ is image intensity at (x,y) location, α is a strongly compressive exponent that reduces the influence of large values, τ is a threshold used to truncate large values after the first phase of normalization and the mean is average intensity value of the image. By default, values of α and τ are set experimentally as 0.1 and 10 respectively.

3 Face Recognition Using MONPP

Modified ONPP (MONPP) [7] is an extension of Orthogonal Neighborhood Preserving Projection (ONPP). ONPP is based on two basic assumptions. First assumption is that the linear relation exists in the local neighbourhood of manifold and thus any data point can be represented as a linear combination of its neighbors. The second assumption is that this linear relationship also holds true in the projection space. The later assumption gives rise to a compact representation of the data that can enhance the classification performance in the projection space. MONPP incorporates possible nonlinearity present in the manifold by introducing nonlinear Z-shaped weighing function for neighbour data points. It is empirically proved that MONPP results in better and stable representation of high dimensional data points. Thus, MONPP is applied for recognition.

MONPP is a two-step algorithm where, in the first step nearest neighbours for each data point are sought and the data point is expressed as a linear combination of these neighbors. In the second step, the data compactness is achieved through a minimization problem in the projected space.

Let $x_1, x_2, ..., x_n$ be the given data points in m-dimensional space ($x_i \in R^m$). So, the data matrix is $X = [x_1, x_2, ..., x_n] \in R^{m \times n}$. The basic task of subspace based methods is to find an orthogonal/non-orthogonal projection matrix $V_{m \times d}$ such that $Y = V^T X$, where $Y \in R^{d \times n}$ is the embedding of X in lower dimension as d is assumed to be less than m.

For, each data point x_i, nearest neighbors are selected in either of two ways: (1) k neighbors are selected by Nearest Neighbor (NN) technique where k is suitably chosen parameter, known as k nearest neighbors (2) neighbors could be selected which are within ϵ distance apart from the data point known as ϵ neighbors. Let $\mathcal{N}_{x_i}$ be the set of k nearest neighbors. In first step, data point x_i is expressed as a linear combination of its k neighbors. Let $\sum_{j=1}^{k} w_{ij} x_j$ be the linear combination of neighbors $x_j \in \mathcal{N}_{x_i}$ of x_i. The weight w_{ij} are calculated by minimizing the reconstruction errors i.e. error between x_i and the reconstruction of x_i using the linear combination of neighbours $x_j \in \mathcal{N}_{x_i}$.

$$\arg\min \mathcal{E}(W) = \frac{1}{2} \sum_{i=1}^{n} \parallel x_i - \sum_{j=1}^{k} w_{ij} x_j \parallel^2 \tag{7}$$

subject to $w_{ij} = 0$, if $x_j \notin \mathcal{N}_{x_i}$ and $\sum_{j=1}^{k} w_{ij} = 1$.

In traditional ONPP, closed form solution of Eq. 7 can be achieved by solving a least square problem, resulting in linear weights for each of nearest neighbours. MONPP incorporates nonlinear weighing scheme using following equation:

$$w_{i:} = \frac{Ze}{e^T Ze} \tag{8}$$

where, $e = [1, 1, ...1] \in \mathcal{R}^k$. $Z \in \mathcal{R}^{k \times k}$ and each elements of Z matrix is calculated using

$Z_{pl} = \mathcal{Z}(d_p; a, b) + \mathcal{Z}(d_l; a, b)\ for,\ \forall\, x_p, x_l\ \in \mathcal{N}_{x_i}$

here, $\mathcal{Z}(d_k; a, b)$ is computed using Z-shaped function [7], d_k is the distance between x_i and it's neighbor x_k (Here, Euclidean distance). Parameters a and b are set to 0 and maximum within-class distance respectively.

MONPP algorithm is summarized below:
Inputs: Dataset $\mathbf{X} \in R^{m \times n}$, number of reduced dimension d
Output: Lower dimension projection $\mathbf{Y} \in \mathbf{R}^{d \times n}$

1. Project training data on lower dimension space $\mathbf{R}^{n-c \times n}$ (in supervised mode) and $\mathbf{R}^{Rx \times n}$ (in unsupervised mode, where, Rx is rank of data matrix $\mathbf{X}$) using PCA.
2. Compute NN with class label information (in supervised mode) or using k-NN algorithm (in unsupervised mode).

3. Compute weight matrix $\mathbf{W}$, for each $\mathbf{x_j} \in \{\mathcal{N}x_i\}$ as given in Eq. (8)
4. Compute Projection matrix $\mathbf{V} \in \mathbf{R}^{m \times d}$ whose column vectors are eigen vectors corresponding to smallest d eigen values of matrix $\tilde{\mathbf{M}} = \mathbf{X}(\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W}^T)\mathbf{X}^T$
5. Compute Embedding on lower dimension by $\mathbf{Y} = \mathbf{V}^T\mathbf{X}$

4 Experiments and Results

The effect of LTISN and DoG based enhancement on recognition rates using ONPP and MONPP is compared with the recognition rates without any preprocessing and reported in this section. For unbiased results the experiments are carried out on 10 different realizations from Extended Yale-B [8] and CMU-PIE face database [9].

4.1 Extended YALE-B Face Database

The experiment is performed on 2432 frontal face images of 28 subjects each with 64 illumination condition. Images are resized to 60×40 to reduce computation. Figure 2 shows face images of a person with 24 different illumination direction along with preprocessed images using LTISN enhancement and DoG enhancement respectively.

Figure 3 left and right shows average recognition result of LTISN and DoG based enhancement techniques respectively, combined with ONPP and MONPP with varying number of nearest-neighbours(k) values 10, 15 and 20. The best recognition result achieved with MONPP + LNIST is 99.84 % at 110 dimension as listed in Table 1.

4.2 CMU-PIE Face Database

The experiment is performed on 42 frontal face images of 68 subjects with varying illumination. Figure 4left and right shows average recognition result of LTISN and DoG based enhancement techniques respectively, combined with ONPP and MONPP with varying number of nearest-neighbours(k) values 10, 15 and 20. MONPP + LNIST gives best recognition with 100 % accuracy at 90 dimension as given in Table 1.

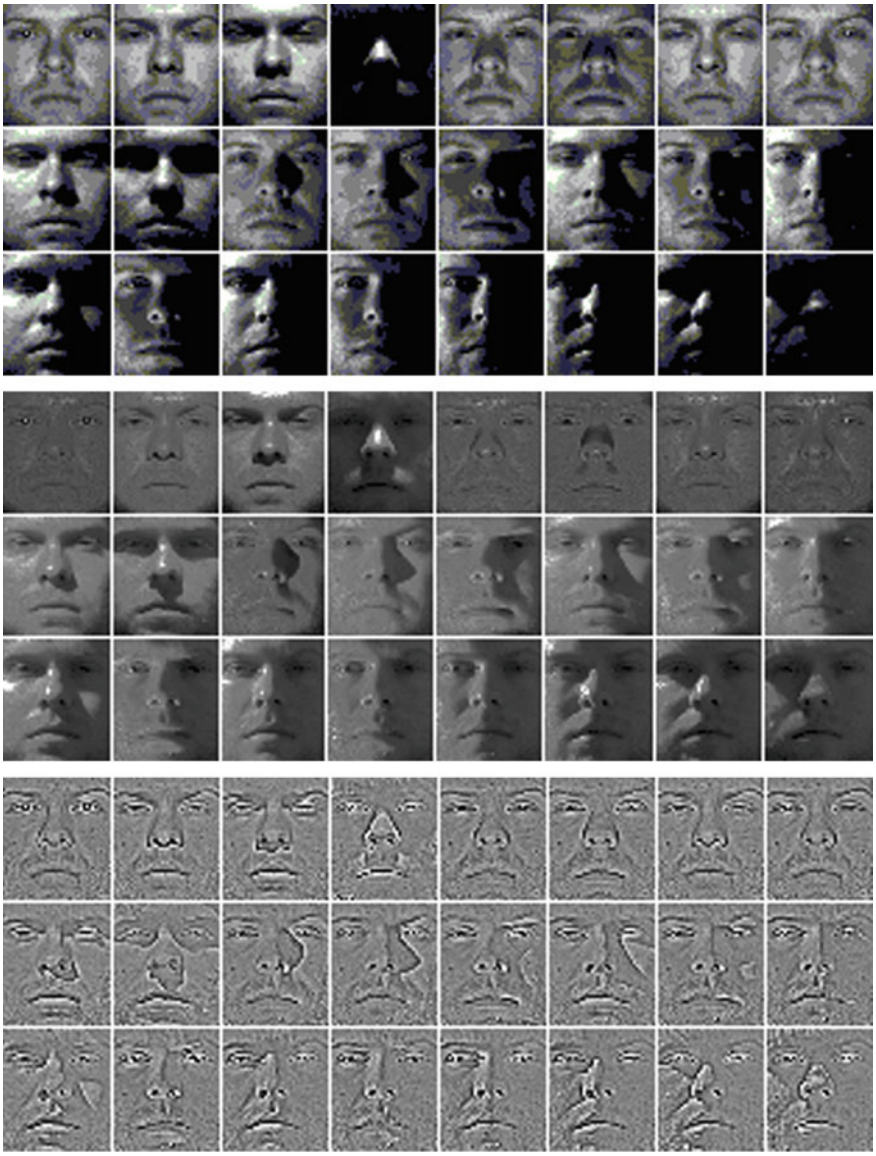


Fig. 2 Face images form Yale-B database (*left*), enhanced images using LTISN (*middle*), enhanced images using DoG (*right*)

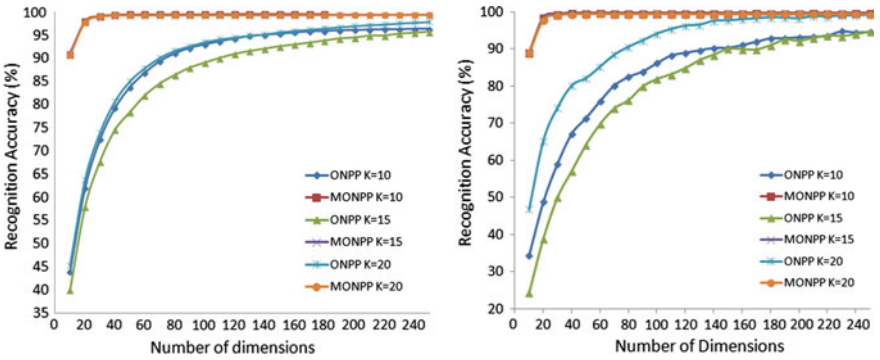


Fig. 3 Results of recognition accuracy (in %) using LNIST (*left*) and DoG (*right*) with ONPP and MONPP on extended Yale-B

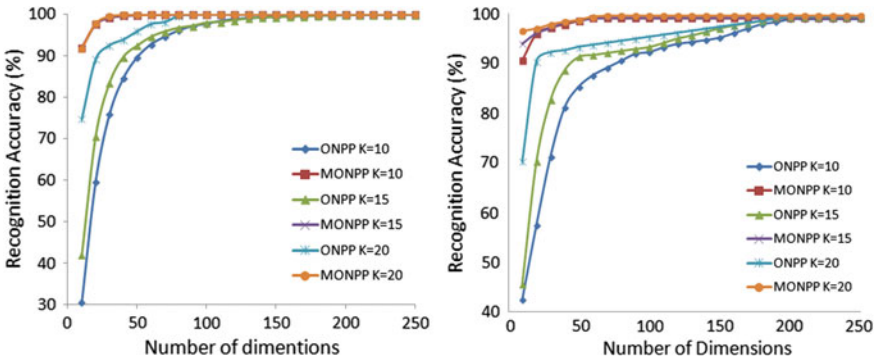


Fig. 4 Results of recognition accuracy (in %) using LNIST (*left*) and DoG (*right*) with ONPP and MONPP on CMU-PIE

Table 1 Comparison of performance of preprocessing techniques in the light of recognition score (in %) of ONPP and MONPP

Database	Extended Yale-B		CMU PIE	
	Average	Best (at dim)	Average	Best (at dim)
ONPP	92.51	93.42(140)	94.33	96.62(130)
ONPP + LTISN	97.84	99.10(250)	98.99	100(90)
ONPP + DoG	96.67	99.34(250)	98.63	99.95(160)
MONPP	94.07	95.80(120)	95.19	97.04(110)
MONPP + LTISN	**99.53**	**99.84(110)**	**99.95**	100(70)
MONPP + DoG	99.51	99.75(40)	99.07	100(50)

5 Conclusion

To handle illumination variations present in face images captured under uncontrolled environment, a robust preprocessing technique is highly sought. This paper mainly contributes to adopt Locally Tuned Inverse Sine Nonlinear (LTISN) transformation for grayscale face images to nullify the illumination variations present in the face database to improve recognition rate. The result of recognition along with LTISN as preprocessing is compared with that of another preprocessing technique called Difference of Gaussian (DoG). The classifier used in all technique is nearest neighbour applied on the coefficient obtained using ONPP and MONPP. In an earlier work, it was established that MONPP performs better than ONPP for face recognition. In the current proposal, it is also observed that LTISN based enhancement followed by MONPP outperforms ONPP with or without both DoG and LTISN enhancement techniques.

Acknowledgements The author acknowledges Board of Research in Nuclear Science, BARC, India for the financial support to carry out this research work.

References

1. Y Adini, Y Moses, and S Ullman. Face recognition: The problem of compensating for changes in illumination direction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):721–732, 1997.
2. W Chen, M J Er, and S Wu. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):458–466, 2006.
3. S M Pizer, E P Amburn, J D Austin, R Cromartie, A Geselowitz, T Greer, B Romeny, J B Zimmerman, and K Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
4. S Shan, W Gao, B Cao, and D Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In *Analysis and Modeling of Faces and Gestures. IEEE International Workshop on*, pages 157–164. IEEE, 2003.
5. M Savvides and BVK V Kumar. Illumination normalization using logarithm transforms for face authentication. In *Audio-and Video-Based Biometric Person Authentication*, pages 549–556. Springer, 2003.
6. E Kokiopoulou and Y Saad. Orthogonal neighborhood preserving projections: A projection-based dimensionality reduction technique. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2143–2156, 2007.
7. P Koringa, G Shikkenawis, S K Mitra, and SK Parulkar. Modified orthogonal neighborhood preserving projection for face recognition. In *Pattern Recognition and Machine Intelligence*, pages 225–235. Springer, 2015.
8. A S Georghiades, P N Belhumeur, and D J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.
9. T Sim, S Baker, and M Bsat. The cmu pose, illumination, and expression database. In *Automatic Face and Gesture Recognition. Proceedings. Fifth IEEE International Conference on*, pages 46–51. IEEE, 2002.

10. E Krieger, VK Asari, and S Arigela. Color image enhancement of low-resolution images captured in extreme lighting conditions. In *SPIE Sensing Technology + Applications*, pages 91200Q–91200Q. International Society for Optics and Photonics, 2014.
11. S Arigela and VK Asari. Self-tunable transformation function for enhancement of high contrast color images. *Journal of Electronic Imaging*, 22(2):023010–023010, 2013.
12. X Tan and B Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650, 2010.

Bin Picking Using Manifold Learning

Ashutosh Kumar, Santanu Chaudhury and J.B. Srivastava

Abstract Bin picking using vision based sensors requires accurate estimation of location and pose of the object for positioning the end effector of the robotic arm. The computational burden and complexity depends upon the parametric model adopted for the task. Learning based techniques to implement the scheme using low dimensional manifolds offer computationally more efficient alternatives. In this paper we have employed Locally Linear Embedding (LLE) and Deep Learning (with auto encoders) for manifold learning in the visual domain as well as for the parameters of robotic manipulator for visual servoing. Images of clusters of cylindrical pellets were used as the training data set in the visual domain. Corresponding parameters of the six degrees of freedom robot for picking designated cylindrical pellet formed the training dataset in the robotic configuration space. The correspondence between the weight coefficients of LLE manifold in the visual domain and robotic domain is established through regression. Autoencoders in conjunction with feed forward neural networks were used for learning of correspondence between the high dimensional visual space and low dimensional configuration space. We have compared the results of the two implementations for the same dataset and found that manifold learning using auto encoders resulted in better performance. The eye-in-hand configuration used with KUKA KR5 robotic arm and Basler camera offers a potentially effective and efficient solution to the bin picking problem through learning based visual servoing.

Keywords Computer vision · Manifold · Locally Linear Embedding (LLE) · Visual servoing · Auto-encoder

A. Kumar (✉) · S. Chaudhury
Department of Electrical Engineering, I.I.T, Delhi, India
e-mail: ashuk65@gmail.com

S. Chaudhury
e-mail: santanuc@ee.iitd.ernet.in

J.B. Srivastava
Department of Mathematics, I.I.T, Delhi, India
e-mail: jbsrivas@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_26

1 Introduction

Automation in manufacturing industry necessitated use of robotic systems in assembly line. In addition to the assembly related jobs, one of the significant tasks that these systems were employed was picking of parts/objects from a heap or a bin and placing them at the desired position in the assembly line. Use of robotic system for such repetitive and monotonous tasks was aimed at increasing efficiency and saving man hours towards more skilled tasks. Initial implementations of this system used mechanical vibratory feeders as sensors [1]. Such systems were highly customized, with almost no provisions for error correction and graceful degradation. In order to overcome these limitations, vision based systems were introduced, one of the earliest such systems used an overhead camera for sensing and a four degree of freedom robot with parallel jaw grippers [2]. The system used simple image processing techniques for clustering and localization of cylindrical pellets. Vision aided bin picking has been a subject of research as a classic vision problem. Significant class of approaches towards refinement of the solution are based on 2D representations (shape, appearance and pose) [3–5], 3D object recognition and localization [6, 7] and visual learning [5, 8]. Visual learning using PCA/Eigen-image analysis has been one of the most popular approaches. PCA based feature representation schemes are based on assumption of linearity of training data-set, hence it may not be suitable for the application which involve non-linear data environment. This paper proposes to apply LLE [9] and Autoencoder [10, 11] as two alternate ways of manifold learning for open loop system. The algorithms have been used for the visual data as well as the end-effector parameters of the robotic manipulator for efficient implementation of a vision based bin picking system.

Subsequent to this introduction, the outline of the paper is as follows:—Sect. 2 discusses the current state of art with reference to our problem, Sect. 3 presents a brief discussion on Locally Linear Embedding and Deep Learning, followed by the details of manifold learning in visual and robotic domain as employed in our scheme. The next two section present the methodology for implementation of the scheme with LLE and deep learning based algorithms. The results of simulation are discussed in Sect. 6. Finally Sect. 7 presents conclusions based on our work.

2 Review of Current State of Art

Typically parametric implementation algorithms for visual servoing are prone to errors due to calibration inaccuracies, image space errors and pose estimation inconsistencies. Learning based methods allow this task to be performed by making the system learn an action pertaining to the given visual input [12]. In the past, several approaches to learning based visual servoing have been made, besides the conventional reinforcement learning techniques, some of them employ fuzzy logic, neural networks, learning in manifold space etc. Object identification and Pose

estimation are some of the areas in which use of learning based techniques have been found to yield significant results. Identification based on appearance of an object in varying illumination and pose is an important problem in visual servoing. Towards this, manifold based representation of data plays a significant role in solution to this problem. One of the early implementation of this idea was by Nayar and Murase [5]. In these manifold based approaches a large data set of the images of object are created. Image features are extracted or alternatively images are compressed to lower dimensions. The dataset thus obtained is represented as a manifold. The test input (image of the unknown object) is projected onto the manifold after dimensionality reduction/feature extraction. Its position on the manifold helps identification and pose estimation. There are separate manifolds for object identification and pose estimation. A major limitation of this approach is large computational complexity. Interpolation between two points on a manifold is also an added complexity. A bunch based method with a shape descriptor model establishing low dimensional pose manifolds capable of distinguishing similar poses of different objects into the corresponding classes with a neural network-based solution has been proposed by Kouskourida et al. [13]. LLE was considered as a potentially appropriate algorithm for manifold learning in our problem as it discovers nonlinear structure in the high dimensional data by exploiting the local symmetries of linear reconstructions. Besides, use of deep learning algorithms with autoencoders is considered an efficient algorithm for non-linear dimensionality reduction. The two algorithms have been used as diverse techniques of manifold learning for open loop system and comparison of results.

3 Manifold Learning

Manifold Learning in Computer Vision has off late been regarded as one of the most efficient techniques for learning based applications that involves dimensionality reduction, noise handling, etc. The main aim in non-linear dimensionality reduction is to embed data that originally lies in a high dimensional space in a lower dimensional space, while preserving characteristic properties of the original data. Some of the popular algorithms for manifold modelling are Principal Component Analysis (PCA), Classical Multi Dimensional Scaling (CMDS), Isometric Mapping (ISOMAP), Locally Linear Embedding (LLE) etc. In this paper we have also considered Deep Learning Techniques as potentially efficient algorithm for manifold learning.

3.1 Manifold Modelling with LLE

Locally Linear Embedding (LLE), is an unsupervised learning algorithm [9]. It computes low dimensional, neighbourhood preserving embeddings of high

dimensional data. It attempts to discover nonlinear structure in high dimensional data by exploiting the local symmetries of linear reconstructions. It employs an Eigen vector method for feature representation. For any data set consisting of N real valued D dimensional vectors, X_i, it is assumed that each data point X_i and its neighbours lie on a locally linear patch of manifold, enabling use of Euclidean distance. LLE identifies K nearest neighbours of this data point to reconstruct the data points from its neighbours using weight W_{ij} (contribution of jth neighbour of data point i in its reconstruction). The basic algorithm for LLE involves following three steps:-

1. Computation of the (K) neighbours of each data point, X_i (D—dimensional) using Euclidean Distance.
2. Computation of the weights W_{ij} that best reconstruct each data point X_i from its neighbours, minimizing the cost function:

$$\varepsilon(W) = \sum_i \left| X_i - \sum_j W_{ij} X_j \right|^2 \tag{1}$$

3. Computation of the low dimensional vectors Y_i (d—dimensional, $d \ll D$) best reconstructed by the weights W_{ij}, minimizing the embedding cost function

$$\varnothing(Y) = \sum_i \left| Y_i - \sum_j W_{ij} Y_j \right|^2 \tag{2}$$

3.2 Manifold Modelling with Autoencoder

Autoencoder, an unsupervised learning algorithm belongs to the class of neural networks [10]. It applies backpropagation, setting the target values to be equal to the inputs. In a way, the network trains itself to simulate an identity function such that f the encoder function, with $h = f(x)$ the representation of x, and g the decoding function, with $\hat{x} = g(h)$ the reconstruction of x. Dimensionality reduction is achieved by reducing the number of nodes in the hidden layers (in between) the input and output layers. The training of auto encoder consists of following:

- Representation of training data at hidden layer h with a suitable encoding function f and decoding function g.
- Learning of the regularization function to prevent perfect reconstruction of the input data, keeping the representation as simple as possible.
- Regularization function must aim at keeping representations insensitive to inputs (Fig. 1).

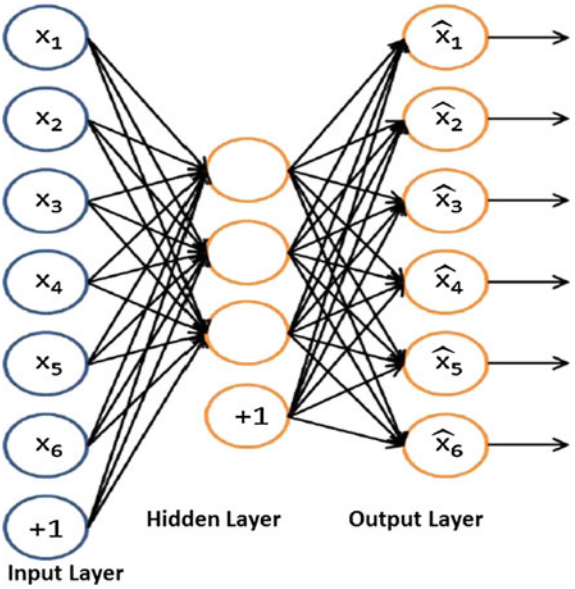


Fig. 1 Dimensionality reduction with autoencoders

4 Implementation Methodology: LLE

The most important aspect of automatic bin-picking is calculation of accurate parameters for the end-effectors of the robot system. In our experiment we have used Cartesian co-ordinate system for executing the task, alternatively the joint angle parameters could also be used.

4.1 Manifold Modelling in Visual Domain

The training image set in visual domain consists of N images, each normalized to size m × n. Every training image is converted to a D—dimensional (D = mn) vector. These N vectors (each of dimension D) are concatenated to form a D × N matrix. In order to find the K nearest neighbours for each of the N vectors, the Euclidean distance with each of the other data points is calculated. The data points pertaining to K shortest Euclidean distance are picked as the nearest neighbours. Reconstruction weights Wij are computed by minimizing the cost function given by Eq. 1, the two constraints for computation of W_{ij} are:-

$$W_{ij} = 0 \quad \text{if } j \notin \{K - \text{nearest neighbours}\} \tag{3}$$

$$\sum W_{ij} = 1 \tag{4}$$

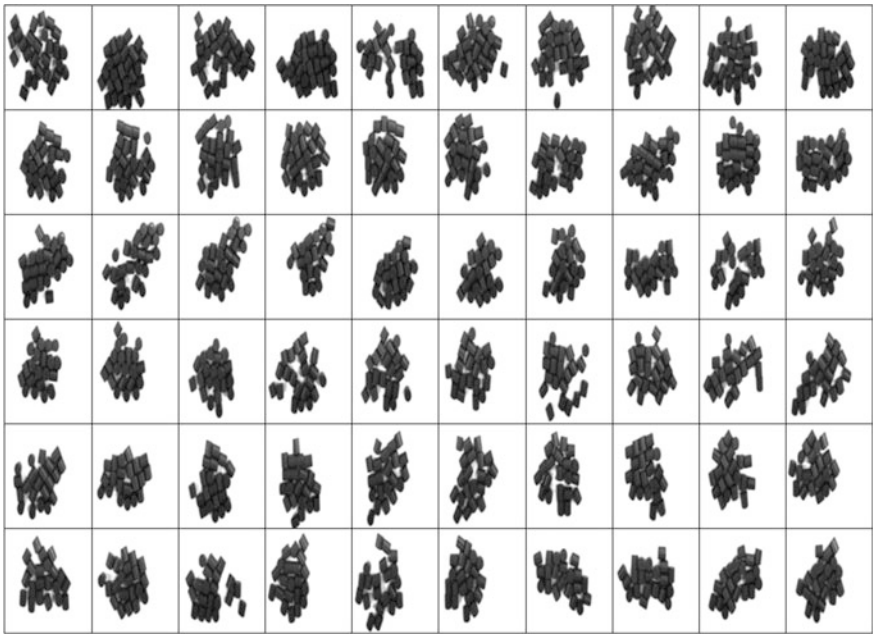


Fig. 2 Sample image data set of pellet clusters

Computation of neighbourhood indices yields a K × N matrix pertaining to N data elements. Weight coefficients also form a K × N matrix (Fig. 2).

4.2 Manifold Modelling in Robot Configuration Space

For every bin picking action, the six degrees of freedom robot parameters can be expressed as a six dimensional vector,

$$P = [X \quad Y \quad Z \quad A \quad B \quad C]^T \tag{5}$$

In Cartesian co-ordinate system (X, Y, Z are the usual three dimensional co-ordinates and A, B, C are the angular orientation of the end -effector with reference to the three axes). Robot end effector parameters can also be expressed in the joint angle system as

$$Q = [A1 \quad A2 \quad A3 \quad A4 \quad A5 \quad A6]^T \tag{6}$$

Here A1, A2……. A6 are the six joint angles for the six DOF robot system. In our experiment, with reference to N pellet-clusters expressed as n data points, there would be N different positions of the robot end-effector for picking up a pellet.

Hence in the robot domain the data set would consist of N, 6-dimensional vectors. These vectors in robot domain can also be mapped on to locally linear embedding in terms of K nearest neighbours as in the case of visual domain.

4.3 SVR Based Learning

The basic premise in this case is the correspondence between the selected data-point in visual domain and corresponding parameters in the robot domain. As shown in the flowchart in Fig. 3, in the learning phase of the algorithm, the K-nearest neighbours for each data-point in the visual domain would be applicable to the corresponding data-points in the robot domain. However the reconstruction weights in both domains could vary. Therefore reconstruction weights in the robot domain were computed based on the nearest neighbours of the visual domain. In order to establish correspondence between the reconstruction weights in the visual domain and in the robot domain, Support vector Regression (SVR) based learning algorithm [14] has been used. In the testing phase, the candidate image is mapped on to LLE manifold to find its K nearest neighbours and corresponding weights. The weight vector is then used for input for prediction in SVR algorithm for computation of corresponding reconstruction weights in the robot domain. Resultant end-effector co-ordinates of the robot domain pertaining to the test image is computed by applying these reconstruction weights on to the corresponding nearest neighbours in the training data-set of the robot domain.

5 Implementation Methodology: Auto Encoder

A typical encoding function for the autoencoder involves a deterministic mapping:

$$h = 1/\left(1 + e^{-(WX+b)}\right) \tag{7}$$

corresponding decoding function is given by

$$\hat{x} = 1/\left(1 + e^{-(W'X+b')}\right) \tag{8}$$

The objecting of training the autoencoder is to learn these parameters for representation of the input data through h and its reconstruction from h.

In our experiments with deep learning, as in the case of LLE, the input data consisted of 500 images. Each low resolution image was of the size of 39 × 33 pixels, thus forming a 1287 dimensional vector when transformed to the vector form. Thus the input layer for the feed forward neural network consisted of a 1287 dimensional vector. The end-effector co-ordinates of the robot were taken in a 2

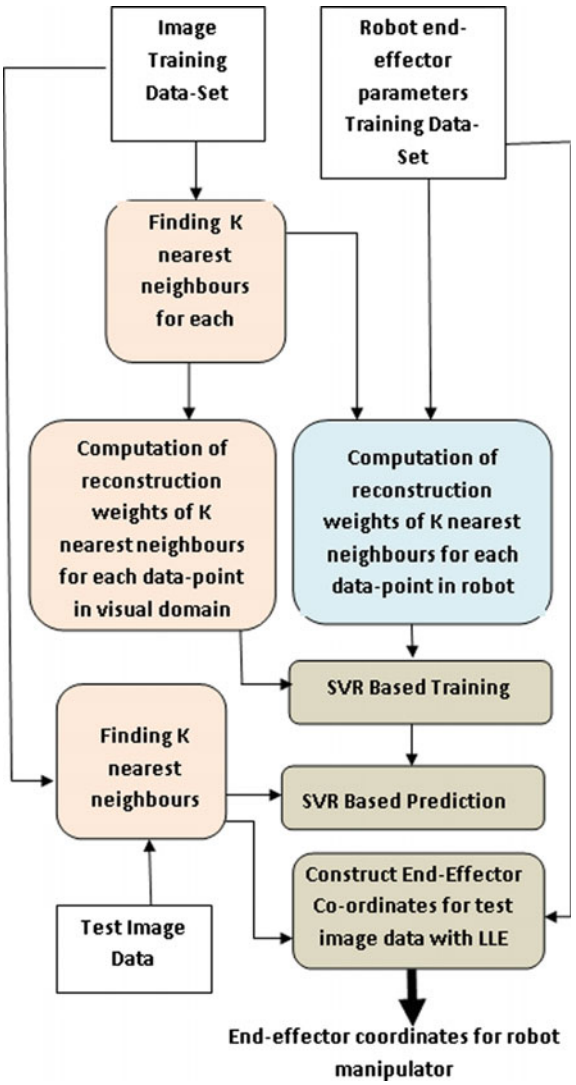


Fig. 3 Flow chart for LLE based implementation

dimensional space forming a 2D vector. The objective of manifold modelling through deep learning therefore was to learn the characteristics of data in 1287 dimensional space so as to find its correspondence with the data in 2 dimensional space. The data set of 500 images of pellet clusters and corresponding 500 end-effector coordinates were divided so as to have 400 data samples for training the auto encoders to learn the manifold and 100 data samples were used for testing the hypothesis. Dimensionality reduction was achieved through 3 hidden layers (learned using autoencoders) between the input and output layers of the feed forward neural network in the following manner.

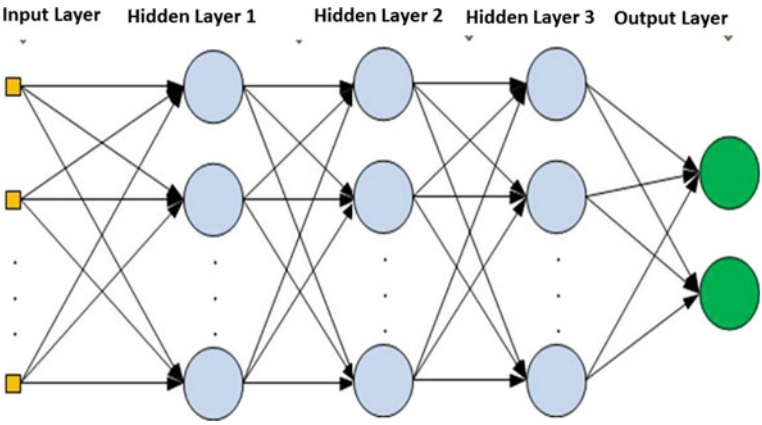


Fig. 4 Manifold modelling with deep learning in 5 layers

- Input Layer: 1287 Dimensions (Layer-1)
- First Hidden Layer: 500 Dimensions (Layer-2)
- Second Hidden Layer: 100 Dimensions (Layer-3)
- Third Hidden Layer: 10 Dimensions (Layer-4)
- Output Layer: 2 Dimensions (Layer-5)

Analysis of test results pertaining to end-effectors coordinates were done in terms of circular error probability (Fig. 4).

6 Results and Discussion

We have used 2000 images of the pellet clusters created for our experiments corresponding to our experimental setup with KUKA KR5 and Basler camera. The main objective of the algorithm was computation of X and Y co-ordinates in world coordinate system of robot for the centre of each pellet to be picked up. During manifold learning with LLE, in the data-set of 2000 images, 800 images were used for manifold modelling 800 were used for training and 100 data samples for testing. While working with deep learning 1600 data samples were used for learning the network and 400 samples for testing. The plots of robot end-effector coordinates pertaining to training and test data for LLE based manifold learning and autoencoders based manifold learning are presented at Fig. 5 and Fig. 6 respectively. Statistically the test data with LLE manifold had a localization accuracy of 89.6 % while the autoencoder based manifold learning had an accuracy of 91.7 %. Complexity of LLE algorithm ranges from cubic $O(DKN^3)$ to logarithmic $O(N log N)$ for various steps [9]. Computationally deep learning algorithm is less efficient compared to parametric algorithms [13]. Apparently the results of deep learning were more accurate due to better learning of the characteristics of data and its

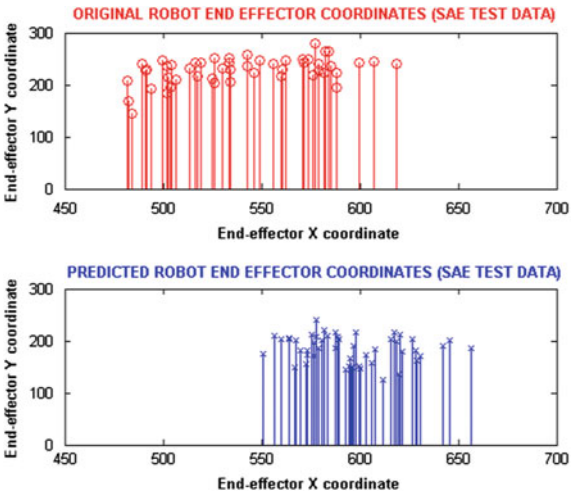


Fig. 5 Localization with LLE based learning

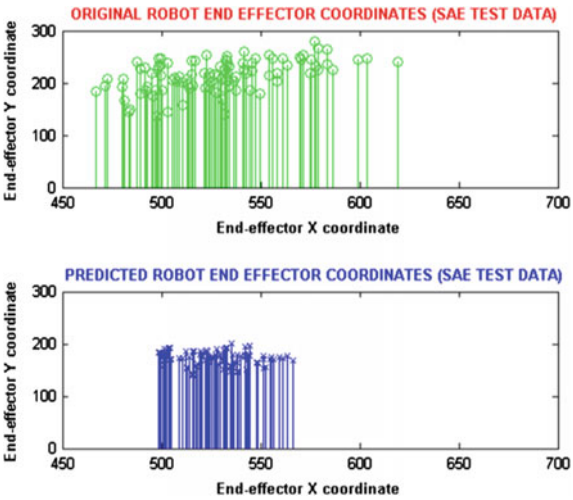


Fig. 6 Localization with deep learning based manifold modelling

correspondence in visual space as well as configuration space of robot end effectors through this algorithm. Deep learning Tool box [11] was used for implementing autoencoder based manifold learning. In this implementation, learning parameters were optimized iteratively for the best results. The number of nodes in the hidden layers and the number of hidden layers in the architecture of deep learning algorithm are significant parameters to be considered during manifold learning for non linear dimensionality reduction. This aspect was apparent while optimizing the parameters of deep learning in our experiments.

7 Conclusion

Application of manifold learning in automatic bin picking is a novel approach to visual servoing. Towards this end our work was based on two significant algorithms for manifold learning, LLE and Deep learning with autoencoders. Our experiments were primarily aimed at proving the concept. Further refinement of the algorithm with optimization of search algorithm for nearest neighbours in LLE, optimization of architecture and learning parameters in deep learning with more rigorous training data would result in much more improved results. It can subsequently find application in many other domains such as mobile robotics, vision based robotic tracking etc.

Acknowledgements We want to express our gratitude to Program for Autonomous Robotics at I. I.T. Delhi for allowing us to use their laboratory facilities. We would also like to thank Mr Manoj Sharma, Research Scholar, Department of Electrical Engg and Mr Riby Abraham, Research Scholar, Department of Mechanical Engg for their help in our work.

References

1. Ghita Ovidiu and Whelan Paul F. 2008. A Systems Engineering Approach to Robotic Bin Picking. *Stereo Vision, Book edited by: Dr. Asim Bhatti, pp. 372.*
2. Kelley, B.; Birk, J.R.; Martins, H. & Tella R. 1982. A robot system which acquires cylindrical workpieces from bins, *IEEE Trans. Syst. Man Cybern.*, vol. 12, no. 2, pp. 204–213.
3. Faugeras, O.D. & Hebert, M. 1986. The representation, recognition and locating of 3-D objects, *Intl. J. Robotics Res.*, vol. 5, no. 3, pp. 27–52.
4. Edwards, J. 1996. An active, appearance-based approach to the pose estimation of complex objects, Proc. of the IEEE Intelligent Robots and Systems Conference, Osaka, Japan, pp. 1458–1465.
5. Murase, H. & Nayar, S.K. 1995. Visual learning and recognition of 3-D objects from appearance, *Intl. Journal of Computer Vision*, vol. 14, pp. 5–24.
6. Ghita O. & Whelan, P.F. 2003. A bin picking system based on depth from defocus, *Machine Vision and Applications*, vol. 13, no. 4, pp. 234–244.
7. Mittrapiyanuruk, P.; DeSouza, G.N. & Kak, A. 2004. Calculating the 3D-pose of rigid objects using active appearance models, *Intl. Conference in Robotics and Automation*, New Orleans, USA.
8. Ghita, O.; Whelan, P.F.; Vernon D. & Mallon J. 2007. Pose estimation for objects with planar surfaces using eigen image and range data analysis, *Machine Vision and Applications*, vol. 18, no. 6, pp. 355–365.
9. Saul Lawrence K and Roweis Sam T. 2000, An Introduction to Locally Linear Embedding, https://www.cs.nyu.edu/~roweis/lle/papers/lleintro.pdf.
10. Deep Learning, An MIT Press book in preparation Yoshua Bengio, Ian Goodfellow and Aaron Courville, http://www.iro.umontreal.ca/~bengioy/dlbook,2015.
11. Deep Learning Tool Box, Prediction as a candidate for learning deep hierarchical models of data, Rasmus Berg Palm, https://github.com/rasmusbergpalm/DeepLearnToolbox.
12. Léonard, Simon, and Martin Jägersand. "Learning based visual servoing." Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on. Vol. 1. IEEE, 2004.

13. Rigas Kouskouridas, Angleo Amanatiadis and Antonios Gasteratos, “Pose Manifolds for Efficient Visual Servoing”, http://www.iis.ee.ic.ac.uk/rkouskou/Publications/Rigas_IST12b.pdf.
14. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
15. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, *11*, 3371–3408.

Motion Estimation from Image Sequences: A Fractional Order Total Variation Model

Pushpendra Kumar and Balasubramanian Raman

Abstract In this paper, a fractional order total variation model is introduced in the estimation of motion field. In particular, the proposed model generalizes the integer order total variation models. The motion estimation is carried out in terms optical flow. The presented model is made using a quadratic and total variation terms. This mathematical formulation makes the model robust against outliers and preserves discontinuities. However, it is difficult to solve the presented model due to the non-differentiability nature of total variation term. For this purpose, the Grünwald-Letnikov derivative is used as a discretization scheme to discretize the fractional order derivative. The resulting formulation is solved by using a more efficient algorithm. Experimental results on various datasets verify the validity of the proposed model.

Keywords Fractional derivative · Image sequence · Optical flow · Total variation regularization

1 Introduction

Motion estimation from a sequence of images is a key task in computer vision/image processing, and obtained a lot of attention from researchers. Motion is the movement of an object between two consecutive frames of an image sequence. Currently, most of the techniques that accurately estimate the motion field are based on variational methods. In these techniques, the motion is determined in terms of optical flow."Optical flow is the distribution of apparent velocity of movement of bright-

P. Kumar (✉)
Department of Mathematics, Indian Institute of Technology Roorkee,
Roorkee 247667, India
e-mail: pushpdma@iitr.ac.in

B. Raman
Department of Computer Science & Engineering,
Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: balarfma@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_27

ness patterns in an image sequence [10]". In general, optical flow can be illustrated as a two dimensional velocity vector which arises either due to the motion of the camera/observer or objects in the scene. It actively used in many vision applications such as robot navigation, surveillance, human-computer interaction, medical diagnosis, 3D reconstruction. Optical flow estimation is considered as an ill-posed problem. Therefore, further a priori assumption is required for accurate estimation. The researchers have proposed several variational models to determine the optical flow in the literature starting from the seminal work [10, 11].

In the past two decades, differential variational models are quite popular among the optical flow estimation techniques. The reason behind it is their advantages and simplicity in modeling the problem and quality of the estimated optical flow [4, 5, 14, 15, 19]. In order to improve the estimation accuracy of the optical flow models, different constraints have been imposed in the variational models and obtained an impressive performance. Recent models like [1, 4, 5] include the additional constraints such as convex robust data term and the gradient constancy assumption in order to make the model robust against outliers and reduce the local minima into global minima. Some of the variational models such as [4, 20] proposed the motion segmentation or parametric models to get a piecewise optical flow. Moreover, the variational models proposed in [9, 21, 22] are based on L_1, L_2 norms and total variation regularization (TV) terms. These models offer significant robustness against illumination changes and noise. All these models are based on integer order differentiation techniques. A modification of these differential variational models, which generalizes their differential from integer order to fractional order has obtained more attention from researchers. This can be categories into the class of fractional order variational model. The fractional order differentiation based methods are now quite popular in various image processing applications [7, 8, 16]. However, very rare attention has been given to use fractional order variational model in optical flow estimation problem. The first fractional order variational model for motion estimation was proposed by Chen et al. [8]. But, it is based on [10], which is more sensitive to noise.

The core idea of fractional order differentiation was introduced at the end of sixteen century and later published in the nineteenth [13]. Fractional order differentiation deals with differentiations of arbitrary order. The prominent difference between the fractional and integer order differentiations is that we can find the fractional derivatives even if the function is not continuous (as in case of images), whereas integer order derivative failed. Thus, fractional derivative efficiently provides the discontinuous information about texture and edges in the optical flow field [8, 13]. In some real life applications, fractional derivatives reduce the computational complexity [13].

2 Contribution

In this paper, we introduce a fractional order total variation model for motion estimation in the image sequences. The novelty of the proposed model is, it generalizes the existing integer order total variation models corresponding to the fractional order. The variational functional is formed using a quadratic and total variation terms. Due to the presence of fractional order derivatives, the proposed model efficiently handles texture and edges. The numerical implementation of fractional order derivative is carried out using Grünwald-Letnikov derivative definition. The resulting variational functional is decomposed into a more suitable scheme and numerically solved by an iterative method. The problem of large motion is solved by using a coarse to fine and warping techniques. The validity of the model is tested on various datasets, and compared with some existing models.

The rest of the paper is organized in the following sections: Section 3 describes the proposed fractional order total variation model followed by the minimization scheme. Section 4 describes the experimental datasets and evaluation metrics followed by the experimental results. Finally, paper is concluded in Sect. 5 with future work remarks.

3 Proposed Fractional Order Total Variation Model

In order to determine the optical flow $\mathbf{u} = (u, v)$, the model proposed by Horn and Schunck [10] minimizes the following variational functional

$$E(\mathbf{u}) = \int_{\Omega} \left[\lambda\,(r(u,v))^2 + (|\nabla u|^2 + |\nabla v|^2)\right] dxdy \tag{1}$$

where, $\lambda > 0$ is a regularization parameter and $r(u, v) = I_{2w} - I_1 + (\mathbf{u} - u_0)\nabla I_{2w}$. Here, $\nabla := (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$, $I_{2w} = I_2$ and u_0 is the existing estimate of the optical flow. The image frames I_1, $I_2 : \Omega \subset \mathbb{R}^3$. The first and second terms in (1) are known as optical flow constraint and smoothness term, respectively.

In order to obtain a more accurate flow field and improve the robustness of the model, a more robust new mathematical model of the energy functional (1) is given as [9]

$$E(\mathbf{u}) = \int_{\Omega} \left[\lambda\,(r(u,v))^2 + (|\nabla u| + |\nabla v|)\right] dxdy \tag{2}$$

The motivation of this work is to contain the minimization scheme computationally simpler, provides dense flow, and increase the robustness of the model against noise and outliers.

The proposed fractional order total variation model of the above variational functional (2) can be given as

$$E(\mathbf{u}) = \int_{\Omega} \left[\lambda\ (r(u,v))^2 + (|D^{\alpha}u| + |D^{\alpha}v|)\right] d\mathbf{X} \tag{3}$$

where, $D^{\alpha} := (D_x^{\alpha}, D_y^{\alpha})^T$ denotes the fractional order derivative operator [17] and $|D^{\alpha}u| = \sqrt{(D_x^{\alpha}u)^2 + (D_y^{\alpha}u)^2}$. The fractional order $\alpha \in \mathbb{R}^+$, when $\alpha = 1$, the proposed fractional order total variation model (3) take the form (2) [9]. In the similar way, when $\alpha = 2$, the derivative in (3) is reduced to the second order integer derivative. Thus, the proposed model (3) generalizes the typical total variation model (2) from integer to fractional order.

In order to minimize the proposed fractional order total variation model (3), it is decomposed into the following forms according to [6],

$$E_{TV-1} = \int_{\Omega} \left[\lambda\ (r(\hat{u},\hat{v}))^2 + \frac{1}{2\theta}(u-\hat{u})^2 + \frac{1}{2\theta}(v-\hat{v})^2\right] dxdy \tag{4}$$

$$E_{TV-u} = \int_{\Omega} \left[\frac{1}{2\theta}(u-\hat{u})^2 + |D^{\alpha}u|\right] dxdy \tag{5}$$

$$E_{TV-v} = \int_{\Omega} \left[\frac{1}{2\theta}(v-\hat{v})^2 + |D^{\alpha}v|\right] dxdy \tag{6}$$

where, θ is a small constant and work as a threshold between $(\hat{u}, \hat{v})$ and (u, v). For $TV-1$, (u, v) are considered as fixed and $(\hat{u}, \hat{v})$ have to determine. The variational functionals given in (5) and (6) are demonstrated in the same manner as image denoising model of Rudin et al. [18].

According to the Euler-Lagrange method, minimization of $(\hat{u}, \hat{v})$ of (4) results the following equations

$$\begin{aligned} (1 + 2\lambda\theta\,(I_{2w}^x)^2)\,\hat{u} + 2\lambda\theta\, I_{2w}^x I_{2w}^y\,\hat{v} &= u - 2\lambda\theta\ r_o I_{2w}^x \\ 2\lambda\theta\, I_{2w}^x I_{2w}^y\,\hat{u} + (1 + 2\lambda\theta\,(I_{2w}^y)^2)\,\hat{v} &= v - 2\lambda\theta\ r_o I_{2w}^y \end{aligned} \tag{7}$$

where, $r_o = I_t - u_o I_{2w}^x - v_o I_{2w}^y$ and $I_t = I_{2w} - I_1$.

Let D is the determinant of the above system of equations given in (7), then

$$D = 1 + 2\lambda\theta\,((I_{2w}^x)^2 + (I_{2w}^y)^2)$$

Solving this system of equations in (7) for $\hat{u}$ and $\hat{v}$, we get

$$D\hat{u} = (1 + 2\lambda\theta\,(I_{2w}^y)^2)u - 2\lambda\theta\,(I_{2w}^x I_{2w}^y)v \tag{8}$$

Similarly for $\hat{v}$

$$-D\hat{v} = 2\lambda\theta\,(I_{2w}^x I_{2w}^y)u - (1 + 2\lambda\theta\,(I_{2w}^x)^2)v + 2\lambda\theta\, r_o I_{2w}^y \tag{9}$$

After solving (8) and (9), we obtain the iterative expressions for $\hat{u}$ and $\hat{v}$ as

$$\hat{u} = \frac{(1 + 2\lambda\theta\,(I_{2w}^{y})^2)u - 2\lambda\theta\,(I_{2w}^{x} I_{2w}^{y})v}{D} \tag{10}$$

$$\hat{v} = \frac{2\lambda\theta\,(I_{2w}^{x} I_{2w}^{y})u - (1 + 2\lambda\theta\,(I_{2w}^{x})^2)v + 2\lambda\theta\, r_o I_{2w}^{y}}{-D} \tag{11}$$

This derivation composes a system of linear equations with respect to $2mn$ unknowns in $\hat{u}$ and $\hat{v}$. Here, m and n are the number of image pixels in x and y-directions, respectively. The solution of this system can be determined by any common numerical scheme such as a successive overrelaxation method, Gauss Seidel method or Jacobi method.

In order to minimize the flow field of (5), first we discretized the fractional derivative of order α of u using the Grünwald-Letnikov definition [12] as

$$D_x^{\alpha} u_{i,j} = \sum_{p=0}^{W-1} w_p^{(\alpha)} u_{i+p,j} \quad \text{and} \quad D_y^{\alpha} u_{i,j} = \sum_{p=0}^{W-1} w_p^{(\alpha)} u_{i,j+p} \tag{12}$$

where, W represents the window mask size and

$$w_p^{(\alpha)} = (-1)^p S_p^{\alpha} \text{ and } S_p^{\alpha} = \frac{\Gamma(\alpha + 1)}{\Gamma(p + 1)\,\Gamma(\alpha - p + 1)}$$

Here, $\Gamma(\alpha)$ represents the gamma function.

In order to determine the solution of (5) using a more suitable primal dual algorithm as described in [7], we reorder the (i, j) component of u and $\hat{u}$ in X and Y such that

$$X_{(j-1)n+i} = u_{i,j} \quad \text{and} \quad Y_{(j-1)n+i} = \hat{u}_{i,j} \tag{13}$$

Here, $X, Y \in \mathbb{R}^N$ and $N = n \times n$, n is the number of pixels in the image. Thus, the fractional order derivative D^{α} of u can be demonstrated in the following form for $q = 1, 2, \ldots, N$:

$$A_q^{(\alpha)} X = \begin{cases} (\sum_{p=0}^{W-1} w_p^{(\alpha)} X_{q+p}, \sum_{p=0}^{W-1} w_p^{(\alpha)} X_{q+np})^T \\ \text{if(q mod n)} \neq 0 \quad \text{and } q \leq N - n \\ (0, \sum_{p=0}^{W-1} w_p^{(\alpha)} X_{q+np})^T \\ \text{if(q mod n)} = 0 \quad \text{and } q \leq N - n \\ (\sum_{p=0}^{W-1} w_p^{(\alpha)} X_{q+p}, 0)^T \\ \text{if(q mod n)} \neq 0 \quad \text{and } q > N - n \\ (0, 0)^T \\ \text{if(q mod n)} = 0 \quad \text{and } q > N - n \end{cases} \tag{14}$$

where, $A_q^{(\alpha)} \in \mathbb{R}^{N\times 2}$. Thus, the discrete version of variational functional (5) can be written as

$$E_{TV-u} = \sum_{q=1}^{N} \|A_q^{(\alpha)} X\| + \frac{1}{2\theta}\|X - Y\|^2 \tag{15}$$

Now, the solution of the above formulation (15) is determined by using the following expression of the primal dual algorithm described in [7],

$$u^{p+1} = \frac{u^p - \tau_p div^\alpha d^{p+1} + \tau_p \frac{1}{\theta}\hat{u}}{1 + \frac{1}{\theta}\tau_p} \tag{16}$$

where, d is the solution of dual of (15). In the same way, we can determine the solution of (6). A summary of the proposed algorithm for estimating the optical flow is described in **Algorithm 1**.

Algorithm 1: Proposed algorithm
Step 1: **Input:** I_1, I_2, λ, α, θ and iterations
Step 2: Compute $\hat{u}$ and $\hat{v}$ from (7)
Step 3: Compute u and v from (5) and (6) using (16)
Step 4: **Output:** optical flow vector **u** = (u,v)

4 Experiments, Results and Discussions

4.1 Datasets

Experimental datasets play an important role in order to assess the performance of any optical flow model. In this paper, we analyzed the performance of the proposed fractional order total variation model on different datasets. The details about the datasets those are used in the experimental study are given as follows:

- **Middlebury training dataset**: Grove, RubberWhale, Urban and Venus [2]
- **Middlebury evaluation dataset**: Army and Mequon [2]

All the above datasets have different structures and properties such as texture, motion blur and shadow. The sample images from these datasets are shown in Figs. 1 and 2.

4.2 Evaluation Method

Performance measure:- For measuring the performance of an optical flow algorithm, numerous metrics have been there in the literature. We estimated the angu-

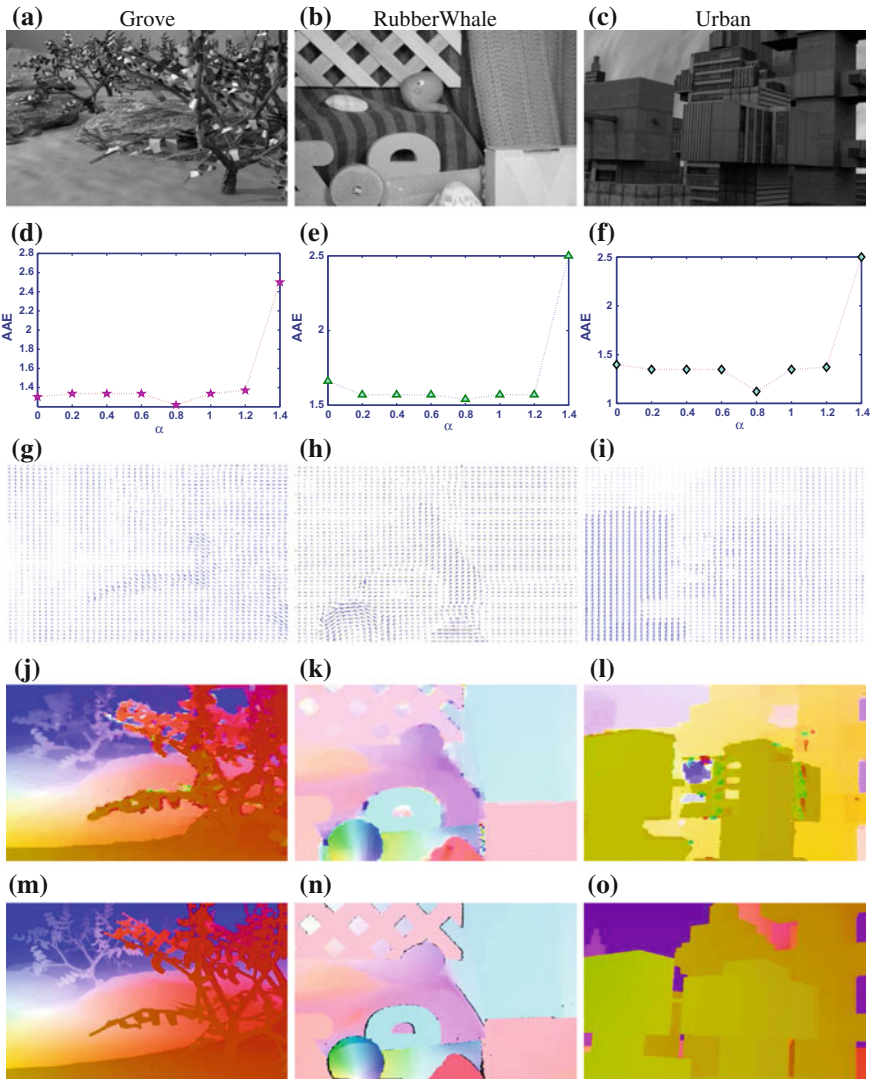


Fig. 1 Optical flow results: sample images (*first row*), optimal fractional order plots (*second row*), vector plots of the estimated optical flow (*third row*), estimated optical flow color maps (*fourth row*) and ground truth plots of the optical flow in *bottom row* [2]

lar error(*AE*) for evaluating the performance of the proposed algorithm. This *AE* is the angle between the correct flow vector $(u_c, v_c, 1)$ and the estimated flow vector $(u_e, v_e, 1)$ (see [3] for details). It is defined as,

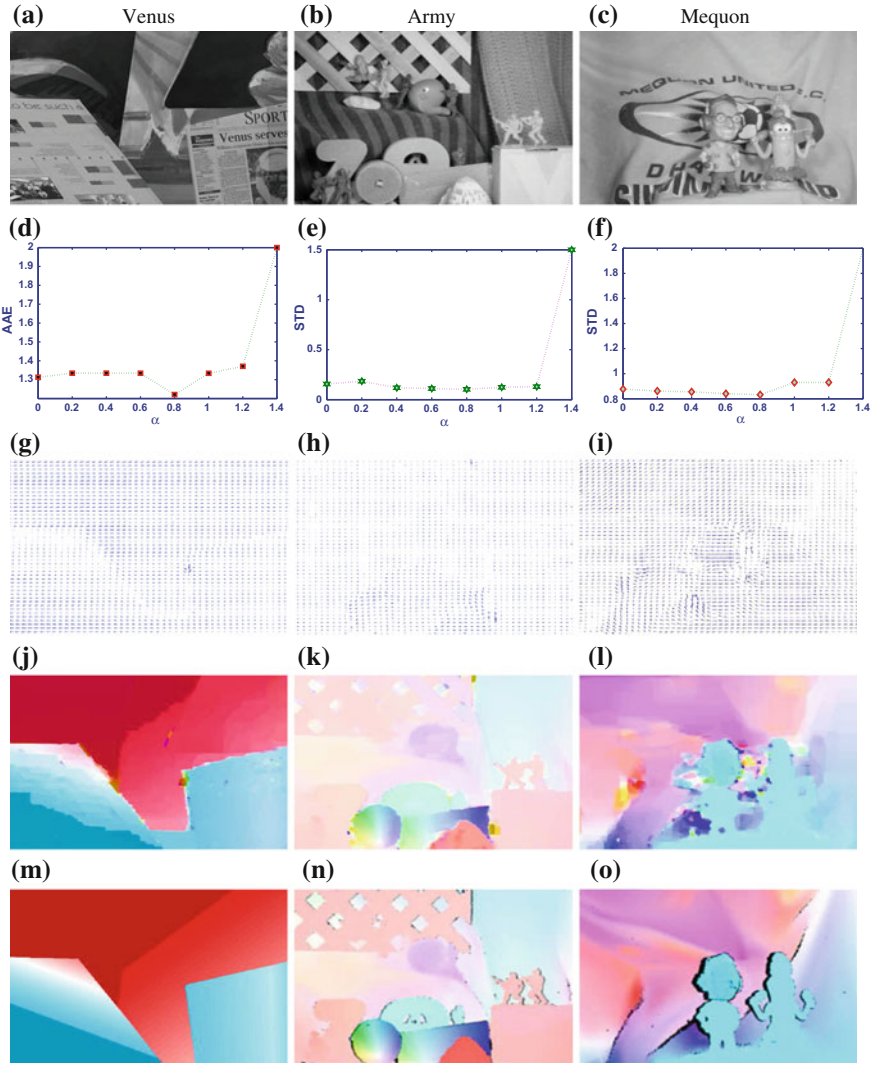


Fig. 2 Optical flow results: sample images (*first row*), optimal fractional order plots (*second row*), vector plots of the estimated optical flow (*third row*), estimated optical flow color maps (*fourth row*) and ground truth plots of the optical flow in *bottom row* [2]

$$AE = \cos^{-1}\left(\frac{u_c u_e + v_c v_e + 1}{\sqrt{(u_c^2 + v_c^2 + 1)(u_e^2 + v_e^2 + 1)}}\right) \tag{17}$$

Statistics:- The validity of the proposed model is also determined using the average angular error (AAE) and standard deviation (STD). These terms are briefly defined in [3].

4.3 Experimental Discussions

We performed experiments on different datasets for a detailed evaluation, analysis and comparisons of the proposed model with the existing models. The default values of the regularization parameter λ, α and θ are 100, 0.8 and 0.5, respectively. A window mask of size 3×3 is used to find the fractional derivatives. Both qualitative and quantitative results are given to demonstrate the validity of the proposed model. The qualitative performance has been illustrated in terms of color maps and vector plots of the flow field. In vector plots, the motion of a pixel/object between two image frames is represented by an arrow. In optical flow color plots, different color represents different directions and homogeneous region represents large displacement. This can be justified by the vector plots of the estimated flow fields. The quantitative performance is shown by the numerical results.

In the first experiment, we have estimated the statistical results for all datasets corresponding to different values of fractional order α of the proposed model. This relationship between statistical results and α are shown in Figs. 1 and 2. The smaller the values of statistical results, the higher is the estimation accuracy of the model. Therefore, the optimal value of fractional order α depends on statistical errors. The optimal value of fractional order is corresponds to the stable solution. Figures 1 and 2 indicate the optimal fractional order for all datasets.

In this experiment, we estimated the qualitative and quantitative results from the proposed model for all datasets corresponding to their fractional order. These datasets are of different dimensions and structure, and contain many image degrada-

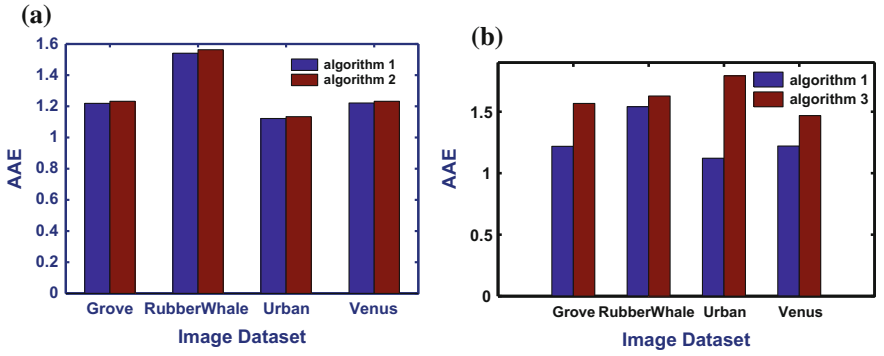


Fig. 3 Comparisons of quantitative results: **a** Proposed model (**Algorithm 1**) with model [10] (**Algorithm 2**), and **b** Proposed model (**Algorithm 1**) with total variation model [9] (**Algorithm 3**)

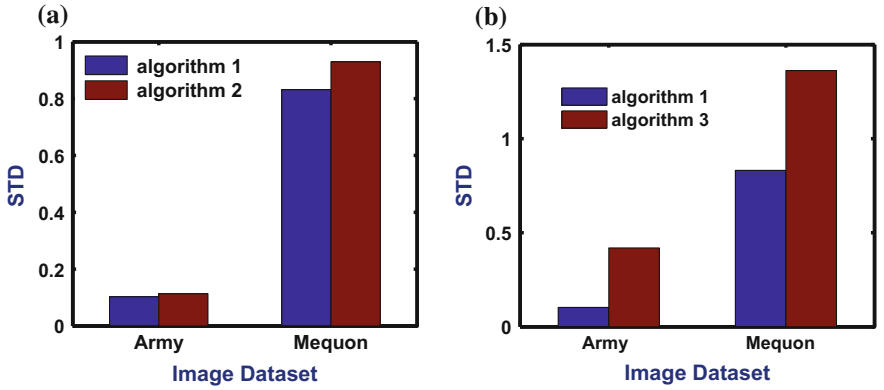


Fig. 4 Comparisons of quantitative results: **a** Proposed model (**Algorithm 1**) with model [10] (**Algorithm 2**), and **b** Proposed model (**Algorithm 1**) with total variation model [9] (**Algorithm 3**)

tion components such as texture, several independent motions and motion blur. The estimated color maps of the optical flow are compared with the ground truth color maps in Figs. 1 and 2. These results demonstrate that the proposed model efficiently handle textures and edges. The color maps of the optical flow are dense, which can be justified by the vector forms of the optical flow. The quantitative results of the model are compared with the total variation model [9] in Figs. 3 and 4. This comparison shows that the fractional order total variation model gives comparatively better results. Additionally, we compared our quantitative results with the model [10] in Figs. 3 and 4. This shows the significant out performance of the proposed model.

5 Conclusions and Future Work

A fractional order total variation model has been presented for motion estimation from image frames. For $\alpha = 1$, the proposed model generalizes the integer order total variation model [9]. The optimal fractional order for which the solution is stable has provided for each image sequence by graphs. Experimental results on different datasets validate that the proposed model efficiently handled texture and edges, and provides dense flow. As a future work, the proposed fractional order total variation model can be extended to the fractional order total variation-L_1 model.

Acknowledgements The author, Pushpendra Kumar gratefully acknowledges the financial support provided by Council of Scientific and Industrial Research(CSIR), New Delhi, India to carry out this work.

References

1. Alvarez, L., Weickert, J., Sánchez, J.: Reliable estimation of dense optical flow fields with large displacements. International Journal of Computer Vision 39(1), 41–56 (2000)
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision 92, 1–31 (2011)
3. Barron, J.L., Fleet, D.J., Beauchemin, S.: Performance of optical flow techniques. International Journal of Computer Vision 12, 43–77 (1994)
4. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise smooth flow. Computer Vision and Image Understanding 63(1), 75–104 (1996)
5. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. Computer Vision - ECCV 4, 25–36 (2004)
6. Chambolle, A.: An algorithm for total variation minimization and applications. Journal of Mathematical imaging and vision 20(1–2), 89–97 (2004)
7. Chen, D., Chen, Y., Xue, D.: Fractional-order total variation image restoration based on primal-dual algorithm. Abstract and Applied Analysis 2013 (2013)
8. Chen, D., Sheng, H., Chen, Y., Xue, D.: Fractional-order variational optical flow model for motion estimation. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 371(1990), 20120148 (2013)
9. Drulea, M., Nedevschi, S.: Total variation regularization of local-global optical flow. In: 14th International Conference on Intelligent Transportation Systems (ITSC). pp. 318–323 (2011)
10. Horn, B., Schunck, B.: Determining optical flow. Artificial Intelligence 17, 185–203 (1981)
11. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Seventh International Joint Conference on Artificial Intelligence, Vancouver, Canada. vol. 81, pp. 674–679 (1981)
12. Miller, K.S.: Derivatives of noninteger order. Mathematics magazine pp. 183–192 (1995)
13. Miller, K.S., Ross, B.: An introduction to the fractional calculus and fractional differential equations. Wiley New York (1993)
14. Motai, Y., Jha, S.K., Kruse, D.: Human tracking from a mobile agent: optical flow and kalman filter arbitration. Signal Processing: Image Communication 27(1), 83–95 (2012)
15. Niese, R., Al-Hamadi, A., Farag, A., Neumann, H., Michaelis, B.: Facial expression recognition based on geometric and optical flow features in colour image sequences. IET computer vision 6(2), 79–89 (2012)
16. Pu, Y.F., Zhou, J.L., Yuan, X.: Fractional differential mask: a fractional differential-based approach for multiscale texture enhancement. IEEE Transactions on Image Processing 19(2), 491–511 (2010)
17. Riemann, B.: Versuch einer allgemeinen auffassung der integration und differentiation. Gesammelte Werke 62 (1876)
18. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena 60(1), 259–268 (1992)
19. Schneider, R.J., Perrin, D.P., Vasilyev, N.V., Marx, G.R., Pedro, J., Howe, R.D.: Mitral annulus segmentation from four-dimensional ultrasound using a valve state predictor and constrained optical flow. Medical image analysis 16(2), 497–504 (2012)
20. Weickert, J.: On discontinuity-preserving optic flow. In: Proceeding of Computer Vision and Mobile Robotics Workshop (1998)
21. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic huber-l1 optical flow. In: BMVC. vol. 1, p. 3 (2009)
22. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Pattern Recognition, pp. 214–223. Springer (2007)

Script Identification in Natural Scene Images: A Dataset and Texture-Feature Based Performance Evaluation

Manisha Verma, Nitakshi Sood, Partha Pratim Roy and Balasubramanian Raman

Abstract Recognizing text with occlusion and perspective distortion in natural scenes is a challenging problem. In this work, we present a dataset of multi-lingual scripts and performance evaluation of script identification in this dataset using texture features. A 'Station Signboard' database that contains railway sign-boards written in 5 different Indic scripts is presented in this work. The images contain challenges like occlusion, perspective distortion, illumination effect, etc. We have collected a total of 500 images and corresponding ground-truths are made in semi-automatic way. Next, a script identification technique is proposed for multi-lingual scene text recognition. Considering the inherent problems in scene images, local texture features are used for feature extraction and SVM classifier, is employed for script identification. From the preliminary experiment, the performance of script identification is found to be 84 % using LBP feature with SVM classifier.

Keywords Texture feature · Local binary pattern · Script identification · SVM classifier · k-NN classifier

M. Verma (✉)
Mathematics Department, IIT Roorkee, Roorkee, India
e-mail: manisha.verma.in@ieee.org

N. Sood
University Institute of Engineering and Technology,
Panjab University, Chandigarh, India
e-mail: nitakshi.sood@gmail.com

P.P. Roy · B. Raman
Computer Science and Engineering Department, IIT Roorkee, Roorkee, India
e-mail: proy.fcs@iitr.ac.in

B. Raman
e-mail: balarfma@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_28

1 Introduction

The documents in multiple script environment, comprise mainly text information in more than one script. Script recognition can be done at different levels as page/paragraph level, text line level, word level or character level [8]. It is necessary to recognize different script regions of the document for automatic processing of such documents through Optical Character Recognition (OCR). Many techniques has been proposed for script detection in past [2]. Singhal et al. proposed a hand written script classification based on Gabor filters [11]. A single document may hold different kind of scripts. Pal et al. proposed a method for line identification in multilingual Indic script in one document [6]. Sun et al. proposed a method to locate the candidate text regions using the low level image features. Encouraging experimental results have been obtained on the nature scene images with the text of various languages [12]. A writer identification method, independent of text and script, has been proposed for handwritten documents using correlation and homogeneity properties of Gray Level Co-occurrence Matrices (GLCM) [1]. Several methods have been proposed on the identification technique that detects scripts, from document images using vectorization which can be implemented to the noisy and degraded documents [9]. Video script identification also uses the concept of text detection by studying the behavior of text lines considering the cursiveness and smoothness of the given script [7].

In the proposed work, a model is designed to identify words of Odia, Telugu, Urdu, Hindi and English scripts from a railway station board that depicts the name of place in different scripts. For this task, first a database of five scripts has been made using railway station board images. The presented method is trained to learn the distinct features of each script and then use k nearest neighbor or SVM for classification. Given a scene image of railways station, the yellow station board showing the name of the station in different scripts can be extracted. The railway station boards have the name of station written in different languages which includes English, Hindi, and any other regional language of that place. The image is first stored digitally in grayscale format and then it is further processed to have a script recognition accurately. It is a problem of recognizing script in natural scene images and as a sub problem it refers to recognizing words that appear on railway station boards. If these kind of scripts can be recognized, they can be utilized for a large number of applications.

Previously researchers have worked to identify text in natural scene images, but their scopes are limited to horizontal texts in the image documents. However, railway station boards can be seen in any orientation, and with perspective distortion. The extraction of region of interest, i.e. text data from the whole image is done in a semi-automatic way. Given a segmented word image, the aim is to recognize the script from it. Most of the script detection work have been done on binary images. In the conversion of grayscale to binary, the image can lose text information and hence detection process may affect. To overcome this issue, the proposed method is using grayscale images to extract features for images.

1.1 Main Contribution

Main contributions are as follows.

- An approach has been presented to identify perspective scene texts of random orientations. This problem appears in many real world issues, but has been neglected by most of the preceding works.
- For performance testing, we present a dataset with different scripts, which comprises texts from railway station scene images with a variety of viewpoints.
- To tackle the problem of script identification, texture features using local patterns have been used in this work.

Therefore, the main issue of handling perspective texts has been neglected by previous works. In this paper, recognition of perspective scripts of random orientations has been addressed in different natural scenes (such as railway station scene images).

Rest of the paper is structured as follows. Section 2 presents the data collection and scripts used in our dataset of scene images. Section 3 describes the texture features extracted from scene text images. The classification process is described in Sect. 4. Results obtained through several experiments are presented in Sect. 5. Finally, we conclude in Sect. 6 by highlighting some of the possible future extensions of the present work.

2 Data Collection

Availability of standard database is one of the most important issues for any pattern recognition research work. Till date no standard database is available for all official Indic scripts. Total 500 images are collected from different sources. Out of 500 script images, 100 for each script are taken.

Initially there were scenic images present in the database, which was further segmented into the yellow board pictures by selecting the four corner points of the desired yellow board. Further, these images were converted into grayscale and then into binary image using some threshold value. The binary image was then segmented into words of each script. For those images, which could not give the required perfect segments, vertical segmentation followed by horizontal segmentation have been carried out otherwise the manual segmentation for those script images has been done (Fig. 1).

Challenges: Script recognition is challenging for several reasons. The first and most obvious reason is that there are many script categories. The second reason is the viewpoint variation where many boards can look different from different angles. The third reason is illumination in which lighting makes the same objects look like different objects. The fourth reason is background clutter in which the classifier cannot distinguish the board from its background. Other challenges include scale, deformation, occlusion, and intra-class variation. Some of the images from database have been shown in Fig. 2.

Fig. 1 Data collection from station board scenic images

Fig. 2 Original database images

2.1 Scripts

Having said all that, when we look at country India, precisely, incredibly diverse India, where language changes from one region to another just as easily as notes of a classical music piece. In India, moving few kilometers north, south, east or for that matter west, there's a significant variation in language, both the dialect and the script change, not to mention the peculiar accents that occasionally adorn the language. Each region of this country is totally different from the rest, and this difference is for sure inclusive of the language too.

Narrowing the horizon and talking of Hindi, Punjabi, Bengali, Urdu and English, these languages have their own history, each equally unique and very ancient of course. Hindi, being in the Devanagari Script, Punjabi in the Gurumukhi Script, Bengali in the Bangla Script, Urdu in the Persian Script with a typical Nasta'liq Style, and English in the Roman Script, are very different in many ways despite a different script. But scripts mark an important component of studying variations in languages. The scripts decide to a much extent the development of a language. In the following section a brief outline about the English, Hindi, Odia, Urdu and Telugu languages is provided.

1. Roman Script: It is used to write English language which is an international language. This script is a descendant of the ancient Proto-Indo-European language family. About 328 million people in India use this language as a communication medium.
2. Devanagari Script: Hindi is the one of the most popular languages in India which uses this script. This language is under Indo-European language family. In India, about 182 million people mainly residing in northern part use this language as their communication medium.

3. Odia: Odia is language of Indian state Odisha and spoken by people of this state. Moreover, it is spoken in other Indian states, e.g., Jharkhand, West Bengal and Gujarat. It is an Indo-Aryan language used by about 33 million people.
4. Urdu Script: Urdu script is written utilizing Urdu alphabets in right-to-left order with 38 letters and no distinct letter cases, the Urdu alphabet is usually written in the calligraphic Nasta'liq script.
5. Telugu Script: Telugu script is utilized to write Telugu language and it is from the Brahmic family of scripts. Telugu is the language of Andhra Pradesh and Telangana states and spoken by people of these states alongwith few other neighboring states.

3 Feature Extraction

Features represent appropriate and unique attributes of an image. It is mainly important when image data is too large to process directly. Images in database are of different size and orientation, and hence feature extraction is crucial task of system to make a unique process for all images. Converting the input image into the set of features is called feature extraction [8]. In pattern recognition, many features have been proposed for image representation. There are mainly high level and low level feature which correspond to user and image perspective respectively. In low level features, color, shape, texture, etc. are most common features. Texture is an significant feature in images that can be noticed easily. In the proposed work, we have extracted texture features of image using local patterns. Local patterns work with the local intensity of each pixel in image, and transform the whole image into a pattern map.

Feature extraction is performed directly on images. After the pre-processing of the input script images, next phase is to carry out the extraction and selection of different features. It is a very crucial phase for the recognition system. Computation of good features is really a challenging task. The term "good" signifies the features which are good enough to capture the maximum variability among inter-classes and the minimum variability within the intra-classes and still computationally easy. In this work, LBP (local binary pattern), CS-LBP (center symmetric local binary pattern) and DLEP (directional local extrema pattern) features of both training and testing data for each script were extracted and studied. All three local patterns are extracted from grayscale version of original image. Brief description of each of the local pattern is given below:

3.1 Local Binary Pattern (LBP)

Ojala et al. proposed local binary pattern [5] in which, each pixel of the image is considered as a center pixel for calculation of pattern value. A neighborhood around

each center pixel is considered and local binary pattern value is computed. Formulation of LBP for a given center pixel I_c and neighboring pixel I_n is as follows:

$$\mathrm{LBP}_{P,R}(x_1, x_2) = \sum_{n=0}^{P-1} 2^n \times T_1(I_n - I_c) \tag{1}$$

$$T_1(a) = \begin{cases} 1 & a \geq 0 \\ 0 & \text{else} \end{cases}$$

$$H(L)\,|_{\mathrm{LBP}} = \sum_{x_1=1}^{m} \sum_{x_2=1}^{n} T_2(\mathrm{LBP}(x_1, x_2), L); \tag{2}$$

$$L \in [0, (2^P - 1)]$$

$$T_2(a_1, b_1) = \begin{cases} 1 & a_1 = b_1 \\ 0 & \text{else} \end{cases} \tag{3}$$

$LBP_{P,R}(x_1, x_2)$ computes the local binary pattern of pixel I_c, where number of neighboring pixels and the radius of circle taken for computation are denoted as P and R and (x_1, x_2) are coordinates of pixel I_c. $H(L)$ computes the histogram of local binary pattern map where $m \times n$ is the image size (Eq. 2).

3.2 Center Symmetric Local Binary Pattern (CSLBP)

Center-symmetric local binary patterns is modified form of LBP that calculated the pattern based on difference of pixels in four different directions. Mathematically, CSLBP can be represented as follows:

$$\mathrm{CSLBP}_{P,R} = \sum_{n=0}^{(P/2)-1} 2^n \times T_1(I_n - I_{n+(P/2)}) \tag{4}$$

$$H(L)\,|_{\mathrm{CSLBP}} = \sum_{x_1=1}^{m} \sum_{x_2=1}^{n} T_2(\mathrm{CSLBP}(x_1, x_2), L); \tag{5}$$

$$L \in [0, 5]$$

where I_n and $I_{n+(P/2)}$ correspond to the intensity of center-symmetric pixel pairs on a circle of radius R with number of neighboring pixels P. The radius is set to 1 and the number of neighborhood pixels are taken as 8. More information about CSLBP can be found in [3].

3.3 Directional Local Extrema Pattern (DLEP)

The Directional Local Extrema Patterns (DLEP) are used to compute the relationship of each image pixel with its neighboring pixels in specific directions [4]. DLEP has been proposed for edge information in 0°, 45°, 90° and 135° directions.

$$I'(_i) = I_n - I_c \,\forall \quad n = 1, 2, \ldots, 8 \tag{6}$$

$$D_\theta(I_c) = T_3(I'_j, I'_{j+4}) \,\forall \theta = 0°, 45°, 90°, 135° \quad \forall \quad j = (1 + \theta/45)$$

$$T_3(I'_j, I'_{j+4}) = \begin{cases} 1 & I'_j \times I'_{j+4} \geq 0 \\ 0 & \text{else} \end{cases} \tag{7}$$

$$\text{DLEP}_{pat}(I_c))\Big|_\theta = \{D_\theta(I_c);\, D_\theta(I_1);\, D_\theta(I_2); \ldots D_\theta(I_8)\}$$

$$\text{DLEP}(I_c)|_\theta = \sum_{n=0}^{8} 2^n \times \text{DLEP}_{pat}(I_c)|_\theta(n) \tag{8}$$

$$H(L)\,|_{\text{DLEP}(\theta)} = \sum_{x_1=1}^{m} \sum_{x_2=1}^{n} T_2(\text{DLEP}(x_1, x_2)\,|_\theta, L);$$
$$L \in [0, 511]$$

where $DLEP(I_c)|_\theta$ is the DLEP map of a given image and $H(L)\,|_{\text{DLEP}(\theta)}$ is histogram of the extracted DLEP map.

For all three features (LBP, CSLBP and DLEP) final histogram of pattern map work as a feature vector of image. Later, the feature vector for all scripts of training and testing data was made for experimental purpose.

4 Classifiers

After feature extraction, classifier is used to differentiate scripts into different classes. In the proposed work, script classification has been done using two well-known classifiers, i.e., k-NN and SVM classifier.

4.1 k-NN Classifier

Image classification is based on image matching, and it is calculated by feature matching. After feature extraction, similarity matching has been observed for testing image. Different distance measuring techniques are Canberra distance, Manhattan distance, Euclidean distance, chi-square distance, etc. In the proposed work, the best results were found from the Euclidean distance measure.

$$D(tr, ts) = \left(\sum_{n=1}^{L} \left| (F_{tr}(n) - F_{ts}(n))^2 \right| \right)^{\frac{1}{2}} \tag{9}$$

Distance measures of a testing image from each training image are computed and sorted. Based on sorted distances k nearest distance measures are selected as good matches.

4.2 SVM Classifier

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. A classification task usually involves separating data into training and testing sets. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. Though new kernels are being proposed by researchers, beginners may find in SVM books the following four basic kernel [13].

Linear

$$K(z_i, z_j) = z_i^T z_j \tag{10}$$

Polynomial

$$K(z_i, z_j) = (\gamma z_i^T z_j + r)^d, \gamma > 0 \tag{11}$$

Radial basis function (RBS)

$$K(z_i, z_j) = exp(-\gamma \|z_i - \quad z_j\|^2), \gamma > 0 \tag{12}$$

Sigmoid

$$K(z_i, z_j) = tanh(\gamma z_i^T z_j + r) \tag{13}$$

Here γ, r and d are kernel parameters.

5 Experimental Results and Discussion

5.1 Dataset Details

We evaluate our algorithm on the 'Station boards' data set. The results using texture-based features with k-NN and SVM classifier are studied in this section.

5.2 Algorithm Implementation Details

Initially the images were present in the form of station boards, out of which each word of different script was extracted. Later, different features were extracted out of these images such as CS-LBP, LBP and DLEP. These feature vectors of have been used to train and test the images together with support vector machine (SVM) or k nearest neighbour (k-NN) to classify the script type.

5.3 Comparative Study

We compare the results of SVM and k-NN for identification of 5 scripts. In the experiments, cross validation with 9:1 ratio has been adopted. Testing image set of 10 images and 90 training images have been taken for each script. During experiment, different set of testing images has been chosen and average result has been obtained from all testing sets. In each experiment, 50 images are used as test images and 450 images are total training images. We use multi-class SVM classifier with different kernels to get better results. In SVM classifier, Gaussian kernel with Radial Basis Function has given better performance than other kernels. The main reason for poor accuracy of k-NN is the less number of samples for training as it requires large training data base samples to improve the accuracy.

In k-NN, we found the distance between the feature vectors of training and testing data using various distance measures, whereas more computations are required in SVM such as kernel processing and matching feature vector with different parameter settings. The k-NN and SVM represent different approaches to learning. Each approach implies different model for the underlying data. SVM assumes there exist a hyper-plane separating the data points (quite a restrictive assumption), while k-NN attempts to approximate the underlying distribution of the data in a non-parametric fashion (crude approximation of parsen-window estimator).

The SVM classifier gave 84 % for LBP and 80.5 % for DLEP features whereas k-NN gave 64.5 % accuracy for LBP feature. Results for both k-NN and SVM are given in Table 1. The reason for poorer results for the k-NN as compared to SVM is that the extracted features lack in regularity between text patterns and these are not good enough to handle broken segments [10].

Some images are shown in Fig. 3 for which the proposed system has identified correctly. First and fourth images are very noisy and hard to understand. Second image is not horizontal and tilted with a small angle. Our proposed method worked well for these kind of images. Few more images have been shown in Fig. 4 for which

Table 1 Comparative results of different algorithms

Method	k-NN (%)	SVM (%)
LBP	64.5	84
CSLBP	54.5	57
DLEP	62.5	80.5

Fig. 3 Correctly identified images

Fig. 4 Wrong identified images

Table 2 Confusion matrix with LBP feature and SVM classification

Predicted/Actual	English	Hindi	Odia	Telugu	Urdu
English	90	7.5	0	0	2.5
Hindi	5	77.5	12.5	2.5	2.5
Odia	10	0	72.5	17.5	0
Telugu	0	7.5	10	82.5	0
Urdu	0	0	0	2.5	97.5

the system could not identify the accurate script. Most of the images in this category are very small in size. Hence, the proposed method does not work well for very small size images. Confusion matrix for all scripts used in this database is shown in Table 2. It shows that texture feature based method worked very well for English and Urdu scripts and average for other scripts.

6 Conclusion

In this work, we presented a dataset of multi-lingual scripts and performance evaluation of script identification in this dataset using texture features. A 'Station Signboard' database that contains railway sign-boards written in 5 different Indic scripts is used for texture-based feature evaluation. The images contain challenges like occlusion, perspective distortion, illumination effect, etc. Texture feature analysis has been done using well-known local pattern features that provide fine texture details. We implemented two different frameworks for image classification. With a proper learning process, we could observe that SVM classification outperformed k-NN classification. In future, we plan to include more scripts in our dataset. We hope that this work will be helpful for the research towards script identification in scene images.

References

1. Chanda, S., Franke, K., Pal, U.: Text independent writer identification for oriya script. In: Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. pp. 369–373. IEEE (2012)

2. Ghosh, D., Dube, T., Shivaprasad, A.P.: Script recognition–a review. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(12), 2142–2161 (2010)
3. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern recognition 42(3), 425–436 (2009)
4. Murala, S., Maheshwari, R., Balasubramanian, R.: Directional local extrema patterns: a new descriptor for content based image retrieval. International Journal of Multimedia Information Retrieval 1(3), 191–203 (2012)
5. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24(7), 971–987 (2002)
6. Pal, U., Sinha, S., Chaudhuri, B.: Multi-script line identification from indian documents. In: Proceedings of Seventh International Conference on Document Analysis and Recognition. pp. 880–884. IEEE (2003)
7. Phan, T.Q., Shivakumara, P., Ding, Z., Lu, S., Tan, C.L.: Video script identification based on text lines. In: International Conference on Document Analysis and Recognition (ICDAR). pp. 1240–1244. IEEE (2011)
8. Shi, B., Yao, C., Zhang, C., Guo, X., Huang, F., Bai, X.: Automatic script identification in the wild. In: Proceedings of ICDAR. No. 531–535 (2015)
9. Shijian, L., Tan, C.L.: Script and language identification in noisy and degraded document images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 30(1), 14–24 (2008)
10. Shivakumara, P., Yuan, Z., Zhao, D., Lu, T., Tan, C.L.: New gradient-spatial-structural features for video script identification. Computer Vision and Image Understanding 130, 35–53 (2015)
11. Singhal, V., Navin, N., Ghosh, D.: Script-based classification of hand-written text documents in a multilingual environment. In: Proceedings of 13th International Workshop on Research Issues in Data Engineering: Multi-lingual Information Management (RIDE-MLIM). pp. 47–54. IEEE (2003)
12. Sun, Q.Y., Lu, Y.: Text location in scene images using visual attention model. International Journal of Pattern Recognition and Artificial Intelligence 26(04), 1–22 (2012)
13. Ullrich, C.: Support vector classification. In: Forecasting and Hedging in the Foreign Exchange Markets, pp. 65–82. Springer (2009)

Posture Recognition in HINE Exercises

Abdul Fatir Ansari, Partha Pratim Roy and Debi Prosad Dogra

Abstract Pattern recognition, image and video processing based automatic or semi-automatic methodologies are widely used in healthcare services. Especially, image and video guided systems have successfully replaced various medical processes including physical examinations of the patients, analyzing physiological and bio-mechanical parameters, etc. Such systems are becoming popular because of their robustness and acceptability amongst the healthcare community. In this paper, we present an efficient way of infant's posture recognition in a given video sequence of Hammersmith Infant Neurological Examinations (HINE). Our proposed methodology can be considered as a step forward in the process of automating HINE tests through computer assisted tools. We have tested our methodology with a large set of HINE videos recorded at the neuro-development clinic of hospital. It has been found that the proposed methodology can successfully classify the postures of infants with an accuracy of 78.26 %.

Keywords HINE tests · Posture recognition · Skin segmentation · Hidden Markov model · Skeletonization

A.F. Ansari (✉)
Department of Civil Engineering, IIT Roorkee, Roorkee 247667, India
e-mail: abdulfatirs@gmail.com

P.P. Roy
Department of Computer Science & Engineering, IIT Roorkee, Roorkee 247667, India
e-mail: proy.fcs@iitr.ac.in

D.P. Dogra
School of Electrical Sciences, IIT Bhubaneswar, Bhubaneswar 751013, India
e-mail: dpdogra@iitbbs.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_29

1 Introduction

Computer vision guided automatic or semi-automatic systems are one of the major contributors in medical research. Images and videos recorded through X-Ray, Ultrasound (USG), Magnetic Resonance (MR), Electrocardiography (ECG), or Electroencephalography (EEG) are often analysed using computers to help the physicians in the diagnosis process. Above modalities are mainly used to understand the state of the internal structures of human body. On the other hand, external imaging systems or sensors can act as important maintenance or diagnostic utility. For instance, external imaging can be used in human gait analysis [1], infant [2] or old person monitoring systems [3], pedestrian detection [4], patient surveillance [5], etc.

Image and Video analysis based algorithms are also being used to develop automatic and semi-automatic systems for assistance in detection and diagnosis in medical examinations. Researches have, for instance, shown that experts conducting **Hammersmith Infant Neurological Examinations** (HINE) [6] can take the help of computer vision guided tools. HINE is used for assessment of neurological development in infants. These examinations include assessment of posture, cranial nerve functions, tone, movements, reflexes and behaviour.

Examinations are carried out by visually observing the reaction of the baby and assessing each test separately. Hence, these examinations often turn out to be subjective. Therefore, there is a need to automate some of the critical tests of HINE, namely adductors and popliteal angle measurement, ankle dorsiflexion estimation, observation of head control, testing of ventral and vertical suspension, and grading head movement during pulled-to-sit and lateral tilting to bring objectivity in the evaluation process. Significant progress has already been made in this context. For example, Dogra et al. have proposed image and video guided techniques to assess three tests of HINE set, namely **Adductors Angle Measurement** [7], **Pulled to Sit** [8], and **Lateral Tilting** [9] as depicted in Fig. 1.

However, existing solutions are not generic and hence cannot be used directly for the remaining set of tests. In this paper, we attempt to solve the problem of posture recognition of the baby which can be used in automating a large set of tests. We classify a given video sequence of HINE exercise into one of the following classes:

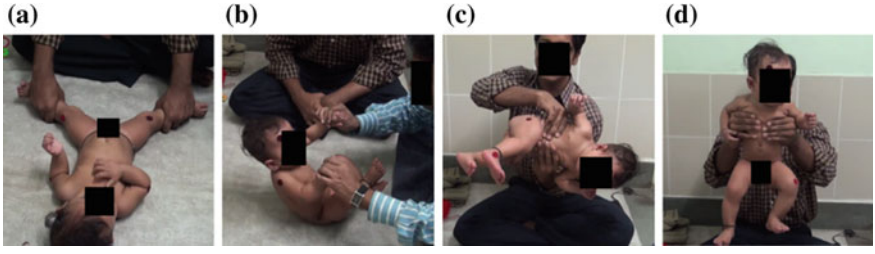


Fig. 1 Four exercises of HINE set, **a** Adductors Angle. **b** Pulled to Sit. **c** Lateral Tilting and **d** Vertical Suspension

Sitting (as in Pulled to Sit), **Lying along Y-axis** (as in Adductors Angle Measurement), **Lying along X-axis** and **Suspending** (as in Vertical Suspension Test).

The rest of the paper is organized as follows. Proposed method is explained in detail in Sect. 2. Results and performance of our system are demonstrated in Sect. 3. Conclusions and future work are presented in Sect. 4.

2 Proposed Methodology

In our approach, each frame (denoted by *I*) of a given video sequence is processed separately. After pre-processing, pixel based skin detection algorithm is used to detect the body of the baby. Colour based segmentation is used because it enables fast processing of frames and the process in not affected much by changes in geometry or illumination. A binary image is then generated. After generation of the binary image, certain morphological operations are performed on the image to make the body area of the infant more prominent. Noisy regions are removed based on the area of the blobs formed within the image frame. The remaining blobs in the binary image are then thinned to generate skeletons. Skeleton with the largest area (which is assumed to be the skeleton of the baby) is chosen for further processing. Features are then extracted from this skeleton to classify the given video sequence into one of the four classes. The whole process is depicted in the flowchart shown in Fig. 2.

2.1 Skin Color Based Segmentation

In the proposed method, we use pixel-color based skin detection method that can classify every pixel as a skin or a non-skin pixel. After testing with various color

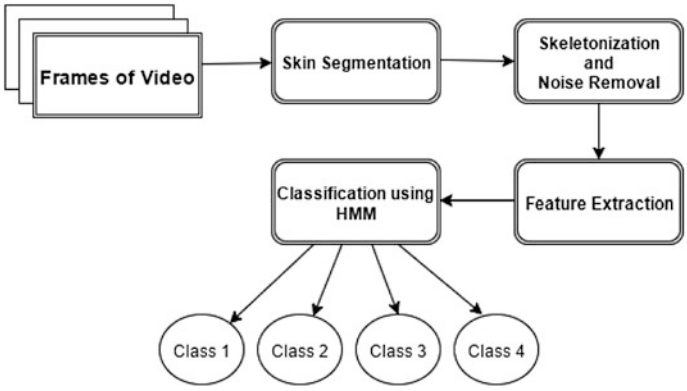


Fig. 2 Flowchart of the proposed methodology

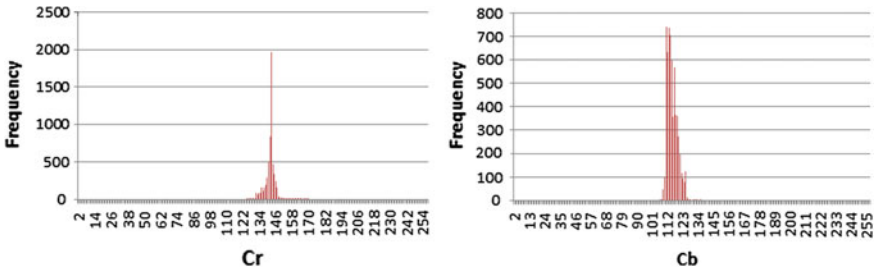


Fig. 3 Histograms of Cr and Cb values in skin region

spaces (e.g. HSV, normalized RGB, and YC_rC_b), we found YC_rC_b to be the most suitable for our application. This is because the luminance component (i.e. Y) doesn't affect the classification, and skins of different complexions can be detected using the same bounds in this color space.

The image is smoothed using a mean filter before conversion into YC_rC_b color space. Our algorithm relies on setting the bounds of C_r and C_b values explicitly which were tested rigorously on various types of human skins. Histograms of C_r and C_b components in skin pixels available in HINE videos are shown in Fig. 3. The YC_rC_b components of a pixel from RGB values can be obtained as:

$$\begin{aligned} Y &= 0.299R + 0.587G + 0.114B \\ C_r &= R - Y \\ C_b &= B - Y \end{aligned}$$

As the videos in our datasets were taken in constant illumination, fixed camera settings, and a good contrast in baby's body and background objects was maintained, we did not employ time-consuming classification and motion tracking algorithms to detect the body. Such methods would have been resource intensive and would have required lot of training datasets thereby defying the purpose of our algorithm. The output image of the above algorithm is a binary image (denoted by I_{seg}).

2.2 Skeletonization

I_{seg} comprises of blobs of skin regions (body of the baby) and of non-skin region. Morphological operations—erosion and dilation—were applied to make the body area prominent. The contours of the resulting binary image were determined. Blobs with size less than a threshold area A_T (experimentally set to 2000), were removed considering them spurious regions.

Fig. 4 Input image → skin segmentation → skeletonization

The output image was then skeletonized (thinned) as described by Guo and Hall [10]. Three thinning methodologies namely, morphological thinning, Zhang-Suen algorithm [11] and Guo-Hall algorithm [10] were tested. The thinning algorithm by Guo and Hall [10] provided best results with least number of redundant pixels within affordable time. Contours of the image generated after thinning were then determined and we saved the largest one (area wise) discarding others. This works perfectly for our analysis as the largest contour is the most probable candidate of the baby's skeleton. The image generated in this phase is denoted by I_{skel} (Fig. 4).

2.3 Feature Extraction

The junction points and end points from the baby's skeleton (I_{skel}) were then searched. For each white pixel, white pixels in its eight neighborhoods were counted. Depending on the number of white pixels (including the current pixel) in the neighborhood (denoted by count) a pixel was classified as a body point, junction point or end point.

- If count is equal to 2, then the pixel is classified as an end point.
- If count is greater than 4, then the pixel is classified as a junction point.
- In all other cases, the pixel is assumed as a normal body point.

Spurious junction points in the vicinity of actual junctions due to redundant pixels or skeleton defects may be detected by the above heuristics. These junctions were then removed by iterating over all the junction points and removing the ones whose Euclidean distance from any other junction point was less than a threshold D_T (experimentally set to 2.5).

The bounding box and center of mass of the baby's skeleton from I_{skel} was found out and the following 6 features were extracted from every frame of the video sequence.

1. **F1**: Width (marked W in Fig. 5) of the rectangle bounding the baby's skeleton.
2. **F2**: Height (marked H in Fig. 5) of the rectangle bounding the baby's skeleton.
3. **F3**: Aspect Ratio (W/H) of the rectangle bounding the baby's skeleton.
4. **F4**: The Euclidean distance (marked D in Fig. 5) between the farthest junction points in the baby's skeleton.

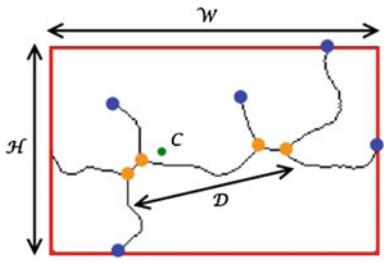


Fig. 5 Features for HMM training. Junctions shown in *orange*, end points in *blue* and centre of mass in *green*

5. **F5**: The normalized X-coordinate of the center of mass (marked C in Fig. 5) of the baby's skeleton.
6. **F6**: The normalized Y-coordinate of the center of mass (marked C in Fig. 5) of the baby's skeleton.

2.4 HMM Based Classification

After extraction of the 6 features from each frame of the video, we apply Hidden Markov Model (HMM) based sequential classifier for classification of the video sequence into one of the four classes namely *Sitting*, *Lying along X axis*, *Lying along Y axis* and *Suspending*. An HMM can be defined by initial state probabilities, state transition matrix $A = [a_{ij}]$, i, j = 1, 2 … N, where a_{ij} denotes the transition probability from state i to state j and output probability $b_j(O_K)$. The density function is written as $b_j(x)$ where x represents a k dimensional feature vector. The recognition is performed using Viterbi algorithm. For the implementation of HMM, the HTK toolkit was used.

3 Results and Discussions

3.1 Data Details

HINE tests were performed in the well-established setup of neuro-development clinic of the Department of Neonatology of SSKM Hospital, Kolkata, India, maintaining constant level of illumination. A homogenous background was maintained throughout for clarity and good contrast between the subject and the background. All clothes of the infant were removed before the commencement of the examinations and the tests were conducted by experts. A total of 25 videos of infants of age group of 3 to 24 months in different postures were used for classification in the experiment.

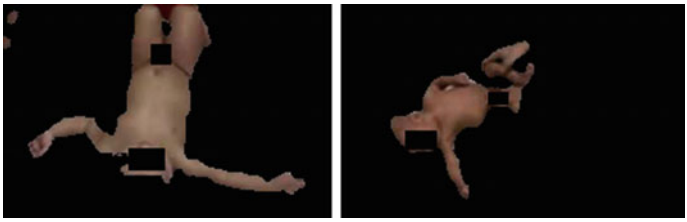


Fig. 6 Skin regions for babies with two different complexions

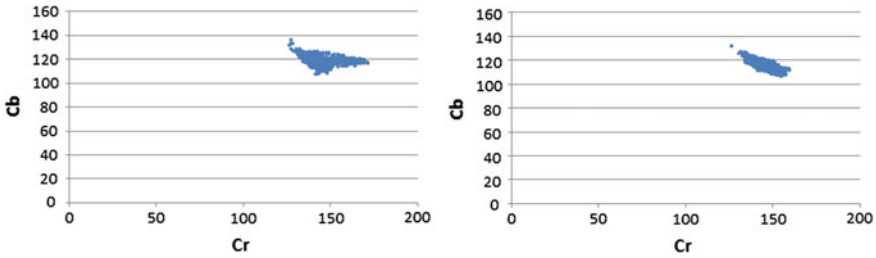


Fig. 7 The C_b and C_r values for two skin types shown in Fig. 6

3.2 *Results of Segmentation and Skeletonization*

The ranges that we found most suitable for a pixel to be classified as a skin pixel after testing on several HINE videos are, $C_r = [140, 173]$ and $C_b = [77, 127]$. Therefore, we slightly modified the bounds given by Chai and Ngan [12]. This range has been proven to be robust against different skin types present in our dataset (Fig. 6). In Fig. 7, we present the results of skin segmentation on babies with two different complexions along with their C_b versus C_r plots. It is evident from the plots that the values of C_b and C_r for skin pixels are indeed clustered within a narrow range.

The results of skin segmentation, morphological operations and skeletonization for each of the four classes, (a) Sitting, (b) Lying Y-axis, (c) Lying X-axis, and (d) Suspending, have been tabulated in Fig. 8.

3.3 *Results of Classification Using HMM*

Classification of videos was done after training using HMM. Testing was performed in Leave-one-out cross-validation (LOOCV) on videos of different exercises. The HMM parameters namely number of states and Gaussian Mixture

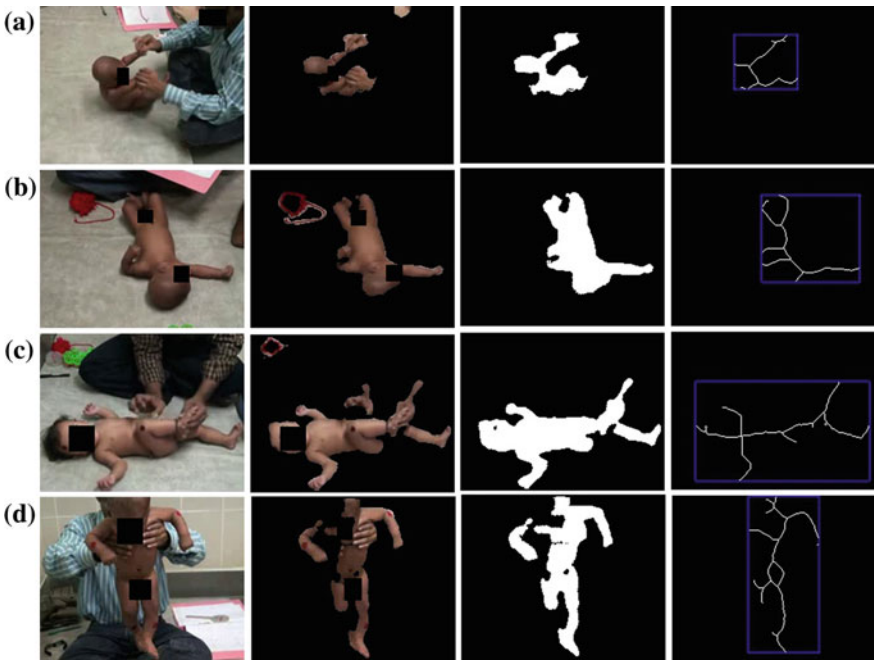


Fig. 8 Input frame → skin segmentation → morphological operations and noise removal → skeletonization

Models were fixed based on the validation set. Best results were achieved with 4 states and 4 Gaussian Mixture Models. Plots of variation of accuracy with number of states and number of Gaussian Mixture Models are shown in Fig. 9. An overall accuracy of 78.26 % was obtained from the dataset when compared against the ground truths (Table 1).

Challenges and Error Analysis: As there are multiple stages in the algorithm, the error propagates through these stages and often gets accumulated. Explicitly defining the boundaries of C_b and C_r for skin segmentation sometimes leads to spurious detection and no-detection of actual body region of the baby due to

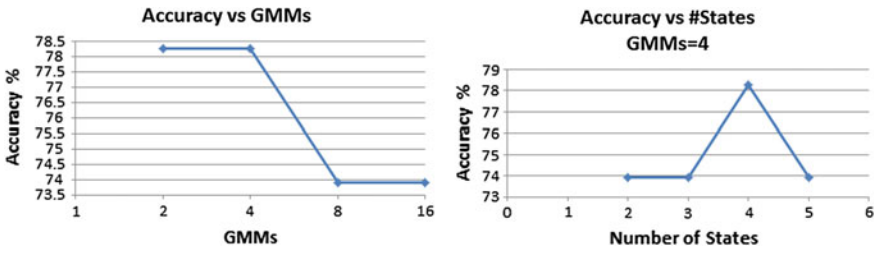


Fig. 9 Variation of accuracy with number of states and number of GMMs

Table 1 Confusion matrix for HMM Classifier

	Lying (X-axis) (%)	Lying (Y-axis) (%)	Sitting (%)	Suspending (%)
Lying (X-axis)	85.7	14.3	0	0
Lying (Y-axis)	20	80	0	0
Sitting	0	66.7	33.3	0
Suspending	0	0	0	100

interference caused by body parts and movements of the physician conducting the test. During the exercises, as the perspective changes with respect to the still camera, certain parts of the infant's body may get removed when blobs are removed based on area threshold. The thinning algorithm leads to skeleton defects and detection of extra junction points at times. These steps will add to the error in the feature extraction step. Classification using HMM requires a lot of training data to improve accuracy.

4 Conclusion and Future Scope

In this paper, we presented a novel approach for classification of a given HINE video sequence into one of the four classes, sitting, lying along Y-axis, lying along X-axis and suspending. After this classification, the HINE exercise performed in an individual video can be easily detected and sent for further analysis and automation of that specific exercise. The proposed algorithm is reasonably fast (averaging 3.5 frames per second) and efficient.

References

1. R. Zhang, C. Vogler, and D. Metaxas. Human gait recognition at sagittal plane. Image and Vision Computing, 25(3):321–330, 2007.
2. S. Singh and H. Hsiao. Infant Telemonitoring System. Engineering in Medicine and Biology Society, 25th International Conference of the IEEE, 2:1354–1357, 2003.
3. J. Wang, Z. Zhang, B. Li, S. Lee, and R. Sherratt. An enhanced fall detection system for elderly person monitoring using consumer home networks. IEEE Transactions on Consumer Electronics, 60(1):23–29, 2014.
4. P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. Ninth IEEE International Conference on Computer Vision: 734–741, 2003.
5. S. Fleck and W. Strasser. Smart camera based monitoring system and its application to assisted living. Proceedings of the IEEE, 96(10):1698–1714, 2008.
6. L. Dubowitz and V. Dubowitz and E. Mercuri. The Neurological Assessment of the Preterm and Full Term Infant. Clinics in Developmental Medicine, London, Heinemann, 9, 2000.
7. D. P. Dogra, A. K. Majumdar, S. Sural, J. Mukherjee, S. Mukherjee, and A. Singh. Automatic adductors angle measurement for neurological assessment of post-neonatal infants during

follow up. Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, 6744:160–166, 2011.
8. D. P. Dogra, A. K. Majumdar, S. Sural, J. Mukherjee, S. Mukherjee, and A. Singh. Toward automating hammersmith pulled-to-sit examination of infants using feature point based video object tracking. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 20 (1):38–47, 2012.
9. D. P. Dogra, V. Badri, A. K. Majumdar, S. Sural, J. Mukherjee, S. Mukherjee, and A. Singh. Video analysis of hammersmith lateral tilting examination using kalman filter guided multi-path tracking. Medical & biological engineering & computing, 52(9):759–772, 2014.
10. Z. Guo and R. W. Hall. Parallel thinning with two-subiteration algorithms. Communications of the ACM, 32(3):359–373, 1989.
11. T.Y. Zhang and C.Y. Suen. A Fast Parallel Algorithm for Thinning Digital Patterns. Communications of the ACM, 27(3):236–239, 1984.
12. D. Chai and K. Ngan. Face segmentation using skin-color map in videophone applications. IEEE Trans. on Circuits and Systems for Video Technology, 9(4):551–564, 1999.

Multi-oriented Text Detection from Video Using Sub-pixel Mapping

Anshul Mittal, Partha Pratim Roy and Balasubramanian Raman

Abstract We have proposes a robust multi oriented text detection approach in video images in this paper. Text detection and text segmentation in video data and images is a difficult task due to low contrast and noise from background. Our methodology focuses not only on spatial information of pixel but also optical flow of image data for detecting moving and static text. This paper provides an iterative algorithm with super resolution to reduce information into its fundamental unit, like alphabets and digits in our case. Proposed method performs image enhancement and sub pixel mapping Jiang Hao and Gao (Applied Mechanics and Materials. 262, 2013) [1] to localize text region and Stroke width Transformation Algorithm (SWT) Epshtein et al. (CVPR, 2010) [2] is used for further noise removal. Since SWT may include some non-text region, so SVM using HOM Khare et al. (A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video, 42(21):7627–7640, 2015) [3] as a descriptor is also used in Final text Selection, Components that satisfy is called a text region. Due to low resolution of images there is a text cluster to remove this text cluster, it is super resolved using sub pixel mapping and hence again passed through process for further segmentation giving an overall accuracy to around 80 %. Our proposed approach is tested in ICDAR2013 dataset in term of recall, precision and F-measure.

Keywords Multi-oriented text · Low resolution videos · Sub pixel mapping · Script independent text segmentation

A. Mittal (✉)
Department of Civil Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: anshulmittal71@gmail.com

P.P. Roy · B. Raman
Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee 247667, India
e-mail: proy.fcs@iitr.ac.in

B. Raman
e-mail: balarfma@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_30

1 Introduction

With the evolution of mobile devices and entry of new concept like augmented reality, text detection becomes trending in recent years. Increase of mobile and it's applications on mobile devices [4], including the Android platforms and iPhone, which can translate text into different languages in real time, has stimulated renewed interest in the problems. The most expressive means of communications is text, and can be embedded into scenes or into documents as a means of communicating information. The collection of huge amounts of street view data is one of the driving application.

To recognize the text information from scene image/video data we need to segment them before feeding to OCR. OCR typically achieves recognition accuracy higher than 99 % on printed and scanned documents [5], text detection and recognition in inferior quality and/or degraded data. Variations of text layout, chaotic backgrounds, illumination, different style of fonts, low resolution and multilingual content present a greater challenge than clean, well-formatted documents.

With huge number of applications of text detection, text segmentation becomes an important part. Multi-oriented low resolution text data which is still a problem. Method proposed in paper deals with segmenting multi- oriented, low resolution text data to its fundamental units.

2 Proposed Methodology

Problem of detecting text from images and Video Frames has been taken care by Chen et al. [6] The authors proposed connected component analysis based text detection but with the presence of low resolution multi-oriented and multi size variance, the recognition performance of text dropped to 0.46 [7]. The performance was increased with neural network classifier but due to non-ability of rotational invariance it loses its ability to detect multi lingual text. Inspired by these problem we have proposed an algorithm for detecting text and non-text pixel clusters and segmentation of text cluster to individual unit for recognition. As proposed by Khare et al. [3] instead of using HOG we have used HOM as a discriminator for detection possible text in frame as HOM uses both spatial and intensity values. HOM was modified as per requirement in a manner such that moment and rotation is marked at the centroid of individual block regardless of centroid of connected component. For low resolution videos we have used sub-pixel mapping based on CIE colorimetry [1]. Figure 1 shows the stages of our algorithm with each individual stage explained later.

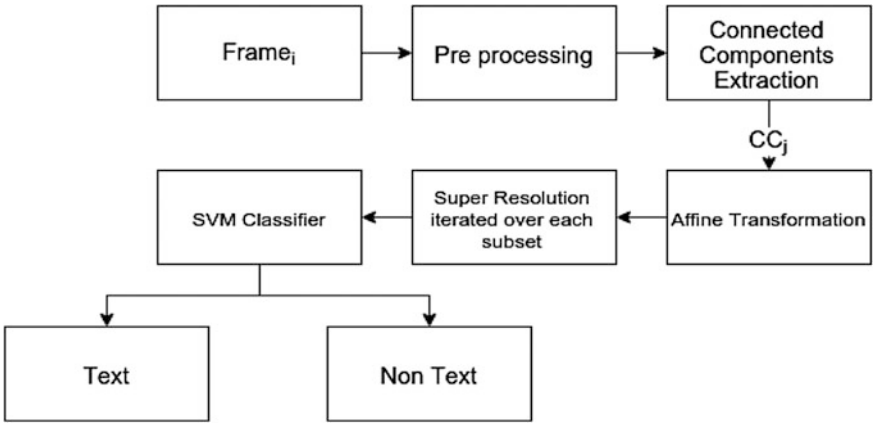


Fig. 1 Flow diagram of our proposed approach

2.1 Preprocessing

Edges are the physical property of any object in an image that distinguishes one object from other object. Text Objects in general have high contrast with the background (Other objects may also have high contract), So in segmentation considering edges increases accuracy. Enhancement of edges using morphological Operations Ensures that small gap of approximately 3px are closed. This Process ensures smooth contours (edges) around each objects. For a given Video, we Parse frames and enhance edges by sharpening the images for enhancing edges Fig. 2b.

Fig. 2 Output for each stage: **a** Input image. **b** Edge detection. **c** Stroke width transformation. **d** Output

2.2 Connected Component (CC) Detection

Contours are detected and Hierarchy is formed i.e. parent Contour, child contours are defined. For any contour bounding another contour than former contour is called parent and later is called child. To remove unnecessary noises Parent Contour having More than one child contour is removed, keeping children in the system. Since our primary target is to detect text in a video frame we define parameter like Solidity, Aspect ratio, Area of contour region. Contours within the threshold value were preserved and obscure contours are eradicated leaving only regions with high text object probability. Stroke Width Transform (SWT), as noted and used by Epshtein et al. [8] is a transform that can be used to differentiate between text and nontext with a very high level of confidence. Chen et al. [6] introduce a Stroke Width Operator, which uses the Distance Transform in order to compute SWT. The SWT value at any point in the image is the width of the stroke which the point is a part of with highest probability.

Since text in images will always have a uniform stroke this step removes noises to a large extent Fig. 2c.

2.3 Super Resolution of CC

In this Paper we have proposed a new way for segmenting low resolution video frames by super resolving of portion in masked image for segmentation of cluster.

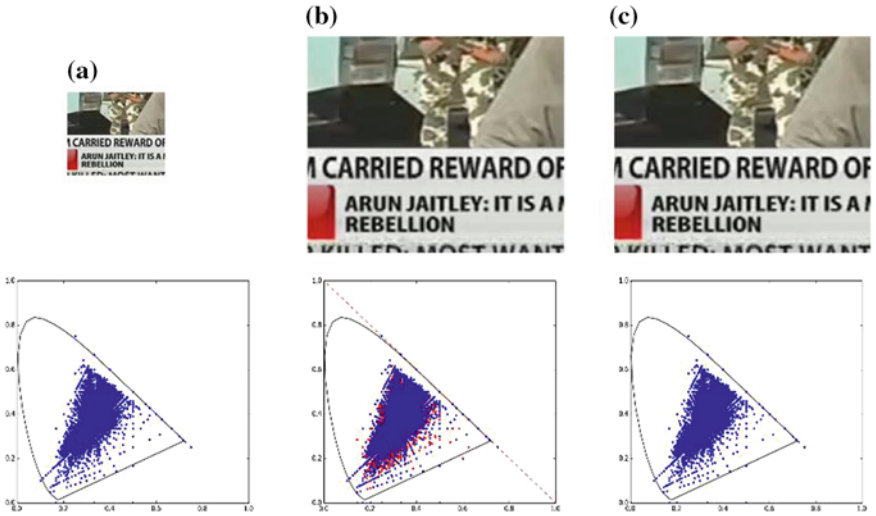


Fig. 3 Chromaticity plot for input and processes image showing error inclusion during up scaling and subsequent removal after processing: **a** Sample input. **b** Up scaled image. **c** Processed image

Fig. 4 Output of super resolution stage: **a** Input image (Text region from the first iteration). **b** Super resolved image. **c** Segmented image

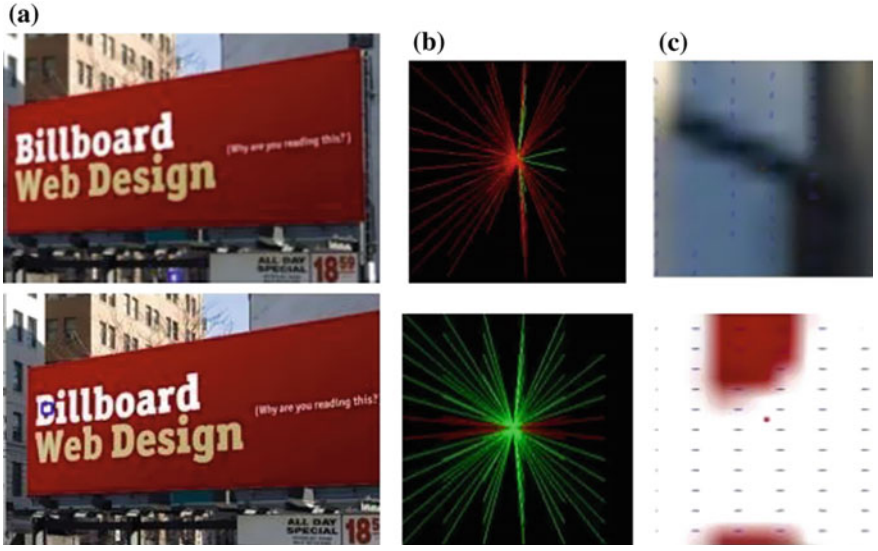


Fig. 5 Output for HOM classifier. **a** Input region. **b** Output values (*Green* = Positive *Red* = Negative). **c** Orientation of moments

Using CIE- XYZ colorimetry and converting it to xyY color space used by Jianghaoa et al. [1] we map each test pixel from the nearby mapped pixel for its closeness to original pixel and converting to original pixel with least distance to the test pixel and again segmenting this enhanced image till we get single segmented object. Each segmented part is Up scaled and super resolve using CIE XYZ-> xyY colorimetry based on Euclidean distance (Fig. 3). This ensures that no addition information i.e. Noise in not introduced. This increases the recall to very significant amount and hence improves our accuracy many fold. See Figs. 3 and 4.

Fig. 6 Output script independent text segmentation. **a** Input image. **b** Edge detection. **c** Stroke width transformation. **d** Initial text regions. **e** Output image

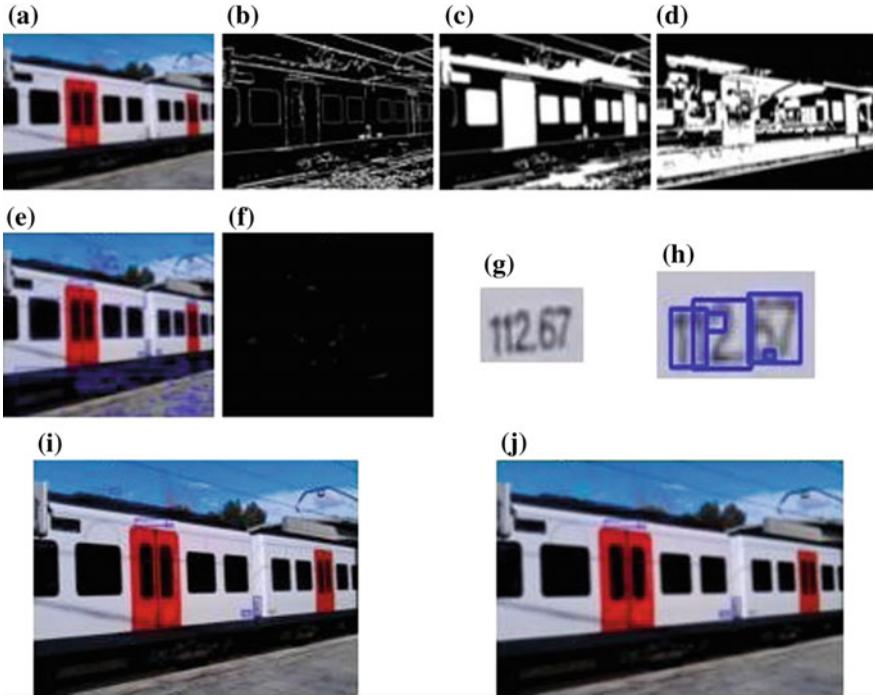


Fig. 7 Output summary for video text segmentation. **a** Initial input frame. **b** Edge detection and enhanced. **c** Possible text region and removing noises. **d** Separating background from foreground using MOG. **e** Intermediate output frame. **f** Possible text region from stroke width transform. **g** Super resolving text region. **h** Segmentation of super resolved masked image. **i** Text regions after removal of noises using SWT and SVM classifier. **j** Final output after SVM and SWT denoising with *green* color as stable text and *blue* color as moving text

Fig. 8 Comparison of results obtained from Chen et al. [6] 's method, images obtained by Stroke Width Transform [2] and our method. Images are of ICDAR datasets

2.4 Text/Non-text Classification

The components that survive the previous stage have very high probability of being a text. However, there are usually some common patterns that get past the previous stage of filtering by virtue of being extremely text-like. For example, arrows in sign boards, or repeating urban patterns Fig. 9 input 4. These components are discarded using a classifier trained to differentiate between text and non-text. The feature vector used in the proposed method for this consists of a scaled down version of the component (to fit in a constant small area), Orientation of Moments using HOM descriptor as proposed by Khare et al. [3], xy values (where (x, y) is the position of each foreground pixel), and aspect ratio. For our experimentation, we used a Support Vector Machine (SVM) classifier and Radial Basis Function (RBF) as kernel. Data set used was taken from ICDAR2013 [9]. Figure 5 Shows the orientation of moments for text as well as non-text clusters. Green line shows orientation towards centroid and Red line show orientation away from centroid.

Fig. 9 Output for multi-oriented script independent text segmentation. **a** Input image. **b** Text regions. **c** Output Image

3 Experimental Results

To evaluate the proposed method, we tested it on well-known ICDAR2013 Robust Reading competition [9] dataset. This data contains 24 videos captured at different rates and different situations. In addition, these videos contain text of different types such as different scripts, fonts, font size and orientations. Parameters for the different stages were trained from the ICDAR 2013 [9] training dataset. For evaluating the proposed algorithm guidelines given by ICDAR 2013 reading competition [10] were strictly followed. According to ICDAR 2013 [9] reading competition, the measures, namely, recall, precision and F-measure are referred as Wolf metric from the Wolf and Jolion (2006). For effectiveness of our algorithm we have compared our algorithm and some state-of-art algorithms as used by Khare et al. [3] as a

Table 1 Performance on ICDAR2013 with temporal information

Algorithm	Precision	Recall	F-factor
Our algorithm	0.75	0.7	0.72
Epshtein et al. [11]	0.69	0.65	0.7
Chen et al. [12]	0.67	0.69	0.679

Table 2 Performance in ICDAR2013 for segmentation in image

Algorithm	Precision	Recall	F-factor
Our algorithm	0.78	0.85	0.81

Table 3 Confusion matrix for SVM classifier

	Text (%)	Non-text (%)
Text	96	4
Non-text	16	84

benchmark summarized in Table 1 (Fig. 8). In our algorithm we are segmenting multi oriented text with orientation correction as explained in Sect. 2 we used ICDAR 2013 [9] dataset without temporal information for assessment of our segmentation algorithm, results are summarized in Table 2. See Fig. 9

Text/Non-Text Classifier stage is very important because it is use as a final filter for passing text candidates for further classification as moving and static text. Hence It's accuracy play a viable role and so does data used to train SVM classifier with RBF as kernel matrix and HOM as a descriptor with other features mentioned in Sect. 2 confusion matrix for text classification is summarized in Table 3.

Text Segmentation: High F value (from Tables 1and 2) indicates the efficiency of this algorithm to be independent of text orientation, illumination and size (Fig. 9) and also segmentation is not affected by low resolution of images and since images are always still we don't require temporal information so we can easily extend our algorithm to images as well (Fig. 6).

Sub Pixel Mapping: This step increases recall and hence improves our accuracy many fold. Sub Pixel Mapping is used to remove any noise generated because of the up-scaling of image and segmentation is thus efficiently done with high resolution image with better edge contrast (Fig. 7g, h).

4 Conclusion and Future Work

In this paper we have proposed a novel Algorithm for multi oriented text segmentation using iterative super resolution and Stroke Width Transform from low resolution images and video frame Figs. 8 and 9. Super Resolution of regions with high possibility of text regions increases the efficiency with marginal increase in time of completion (Tested on Ubuntu 15.1 core i5 6 GB RAM at 20 frame per

Table 4 Summary of complexity of algorithm

	Time complexity	Space complexity
Pre-processing	O(n)	O(n)
Super resolution	$O(n^2)$	$O(n^2)$
Over all	$O(n^2)$	$O(n^2)$

second) which is mainly dependent on complexity of scene. Segmentation Algorithm presented in paper is script independent. Since classifier in only used at the end stage verification, even for that SVM classifier can be easily trained this algorithm with minimal changes in parameters can be easily used to parse and segment any script e.g. Devanagari (Fig. 6). From our Experiments we have got accuracy up-to 80 % which surpasses the current state of art techniques. Since our algorithm segments objects of interest to its fundamental unit i.e. in our case alphabets and digits this gives an extra edge for Text recognition using OCR with connected component analysis with which words can be easily framed Complexity of algorithm for each step and over all complexity is summarized in Table 4.

It is evident from Table 4 that over all time complexity is dependent mainly on Super Resolution step of our proposed algorithm so there is still scope from improvement in the last step to speed up the processes however frame by frame capturing and no dependency of algorithm for the detection of text on previous frame except for the detection of moving text and static text gives us an edge for implementing it on real-time by adjusting fps of video feed.

With further experimentation with GPU and parallel processing fps as high as 40 fps was achieved.

There is still a scope for more accuracy by evolving classifier and using neural network for text classification. Using Grab cut recall can be further improved. Super resolution slows down the speed for text segmentation, objects in close affinity to text requires many iterations to get removed and sometimes may passes undetected. Text in general lies on the continuous surface so with the help of stereoscopic image processing we can extract continuous surface for removal on noises in the background. Text written in artistic style also poses a challenge as brush strokes and different font style makes it time consuming to segment.

References

1. Liu, Jiang Hao, and Shao Hong Gao. “Research on Chromaticity Characterization Methods of the Ink Trapping.” *Applied Mechanics and Materials*. Vol. 262. 2013.
2. B. Epshtein, E. Ofek, and Y. Wexler. “ Detecting text in natural scenes with stroke width transform.” In CVPR, 2010.
3. Vijeta Khare, Palaiahnakote Shivakumara, Paramesran Raveendran “A new Histogram Oriented Moments descriptor for multi-oriented moving text detection in video” Volume 42, Issue 21, 30 November 2015, Pages 7627–7640.
4. C. Liu, C. Wang, and R. Dai, “Text detection in images based on unsupervised classification of edge-based features,” in Proc. IEEE Int. Conf. Doc. Anal. Recognit., 2005, pp. 610–614.

5. J. J. Weinman, E. Learned-Miller, and A. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 10, pp. 1733–1746, Oct. 2009.
6. Chen H, Tsai S, Schroth G, Chen D, Grzeszczuk R, Girod B."Robust text detection in natural images with edge enhanced maximally stable extremal regions." Proceedings of International Conference on Image Processing. 2011:2609–2612.
7. J. Fabrizio, M. Cord, and B. Marcotegui, "Text extraction from street level images," in CMRT, 2009, pp. 199–204.
8. B. Epshtein, E. Ofek, and Y. Wexler. "Detecting text in natural scenes with stroke width transform". In CVPR, pages 2963–2970. IEEE, 2010.
9. D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez, S. Robles, J. Mas, D. Fernandez, J. Almazan, L.P. de las Heras, "ICDAR 2013 Robust Reading Competition", In Proc. *12Th International Conference of Document Analysis and Recognition*, 2013, IEEE CPS, pp. 1115–112.
10. Gomez, L., & Karatzas, D. (2014)." MSER-based real-time text detection and tracking". In Proceedings of ICPR (pp. 3110–3115).
11. X., Lin, K.-H., Fu, Y., Hu, Y., Liu, Y., & Huang, T.-S. (2011)." Text from corners: A novel approach to detect text and caption in videos." IEEE Transactions on Image Processing, 790–799.
12. Weihua Huang; Shivakumara, P.; Tan, C.L., "Detecting moving text in video using temporal information," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, vol., no., pp. 1–4, 8-11 Dec. 2008.

Efficient Framework for Action Recognition Using Reduced Fisher Vector Encoding

Prithviraj Dhar, Jose M. Alvarez and Partha Pratim Roy

Abstract This paper presents a novel and efficient approach to improve performance of recognizing human actions from video by using an unorthodox combination of stage-level approaches. Feature descriptors obtained from dense trajectory i.e. HOG, HOF and MBH are known to be successful in representing videos. In this work, Fisher Vector Encoding with reduced dimensions are separately obtained for each of these descriptors and all of them are concatenated to form one super vector representing each video. To limit the dimension of this super vector we only include first order statistics, computed by the Gaussian Mixture Model, in the individual Fisher Vectors. Finally, we use elements of this super vector, as inputs to be fed to the Deep Belief Network (DBN) classifier. The performance of this setup is evaluated on KTH and Weizmann datasets. Experimental results show a significant improvement on these datasets. An accuracy of 98.92 and 100 % has been obtained on KTH and Weizmann dataset respectively.

Keywords Human action recognition · Deep Belief Network · Dense trajectory features · Fisher Vector

1 Introduction

Human action recognition is an important field of computer vision research. It has found wide applications including video surveillance systems, patient monitoring systems, and a variety of systems that involving human-computer interfaces. Recognizing human actions from video is based on three basic steps. (i) Processing and

P. Dhar (✉)
Department of CSE, IEM Kolkata, India
e-mail: prithvirj95@gmail.com

J.M. Alvarez
NICTA, Sydney, Australia

P.P. Roy
Department of CSE, IIT, Roorkee, India

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_31

extracting features from the video. (ii) Aggregating the extracted feature descriptors and obtaining an encoding for each video, for the task of action localization. (iii) Training a system based on the encoding of the training videos and using it to classify the encoding of the test videos.

There exist many work on action recognition [1, 5, 6, 18]. We propose here an efficient pipeline to use the dense trajectory features for action recognition, by encoding each descriptor in the form of a super vector with reduced dimension and by using DBN to classify these encoding. We test our framework with possible combination of feature descriptors, to find the combination which corresponds to the best recognition accuracy. Our pipeline consists of all of the above mentioned stages. Here, we compare several approaches used in previous work for each stage, and use the best known approach in the pipeline. The contributions of this paper are:

1. An efficient pipeline, which contains a fusion of the best known stage-level approaches. Such stage-level approaches have never been amalgamated in any known previous experiment.
2. Usage of Fisher Vectors without including second order statistics, for the task of feature encoding. This helps to bring down the computation cost for the classifier, and achieve competitive recognition accuracy.

2 Related Work

In this section we review the most relevant approaches to our pipeline. The selection of related approaches for each stage of our pipeline is based on empirical results. We do the following comparisons in order to select the best known stage level approaches in our pipeline.

Local features vs Trajectory features: In this comparison, we tend to decide the type of features which are to be extracted from videos. Using local features has become a popular way for representing videos but, as mentioned in [20], there is a considerable difference between 2D space field and 1D time field. Hence, it would be unwise to detect interest points in a knotted 3D field.

KLT trajectory vs Dense trajectory: Having decided to include trajectory features in our framework, we now compare results obtained by KLT and Dense Trajectories, in previous experiments. In some of the previous works trajectories were often obtained by using KLT tracker which tracked sparse interest points. But in a recent experiment by Wang et al. [21], dense sampling proved to outperform sparse interest points, for recognizing actions.

Fisher Vector vs Other encoding approaches: For action recognition and localization Fisher Vectors have recently been investigated, and shown to yield state-of-the-art performance. In a recent work by Sun et al. [17], using Fisher Vectors produced better results as compared to Bag-of-Words on a test dataset which contained dissimilar videos of different quality. It was also established here that the usage of Fisher

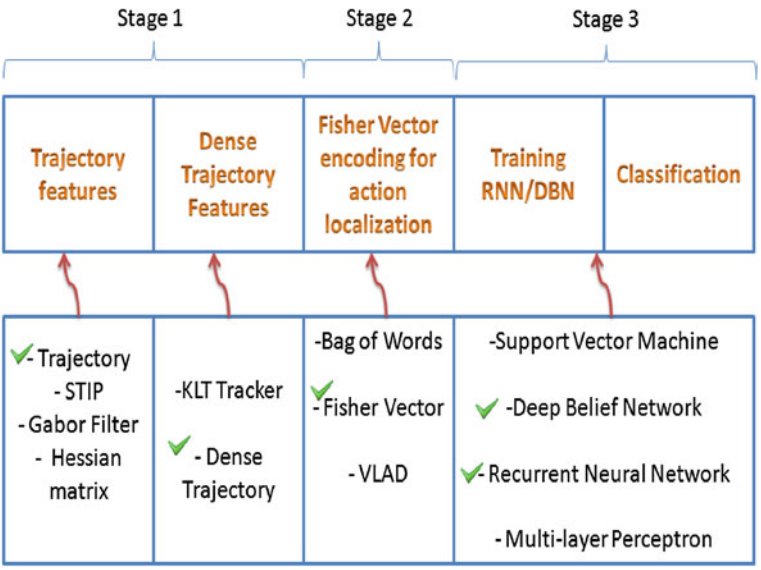


Fig. 1 For selecting an approach for a single stage of the pipeline, we compare several stage level approaches used in previous frameworks, and select the best known approach

Vector outperformed VLAD representation. Also in [13], a detailed analysis on the Hollywood2 dataset established the supremacy of FVs provide over BoV histograms since as lesser visual words are needed. Also computation time of FVs is much lesser than that of BOV histogram.

SVM vs neural network classifier with deep architecture: Referring to a work by Bengio and LeCun in [2], we can assume that deep architectures more efficiently represent high varying functions, in comparison to kernel machines. Also, the work establishes that architectures relying on local kernels can be very incompetent at denoting functions that have many fluctuations, i.e., functions which have several local optima. This has been a motivation to choose neural networks for classification task, over SVMs. The summary of process of selection of stage-level approaches for our framework is illustrated in Fig. 1.

As mentioned above, in the work by Wang et al. [20], dense trajectory features have been used. But for feature encoding, Bag of Words approach has been used, which only gives the number of local descriptors assigned to each Voronoi region. Other information like mean and variance of local descriptors, which is provided by Fisher Vector, is not utilized. The superiority of Fisher Vector over other encoding approaches has been discussed above. Also, in a work by Oneata et al. [13], Fisher Vectors have been used for action localization, but the features in consideration are only SIFT and MBH. Hence, the static information stored in HOG, and the motion information stored in HOF, have not been utilized at all. The effectiveness of Dense Trajectory features have also been discussed above.

This paper is organized as follows. In Sect. 3, we introduce the proposed framework for human action recognition. Here, we discuss in detail, all the stages of the pipeline, which includes feature extraction, feature encoding, training and classification. Finally, in Sect. 4, we present the experimental results of the proposed setup using different classifiers and different combination of features. Also, in this action, we compare our results with those of the previous work.

3 Proposed Framework

The detailed version of our pipeline is shown in Fig. 2. Here, firstly the low-level features are extracted and after that feature pre-processing is done. This is required for codebook generation. It is used for feature encoding. Here, a generative model is used to capture the probability distribution of the features. Gaussian Mixture Model (GMM) is used for this task. The feature encodings are obtained using Fisher Vectors, after which the obtained super-vectors are classified by DBN. The steps have been discussed in detail below.

3.1 Dense Trajectory Feature Extraction

In this experiment, we extract all feature points and their dense trajectory aligned feature descriptors from all the videos, using the public available code. Approximately, 3000 feature points are obtained from each video. Each feature point can be represented using these descriptors: HOG, HOF, MBHx, MBHy. HOG deploys the angular-binning of gradient orientations of an image cell. It saves static information of the image. HOF and MBH provide details regarding motion using optical flow. HOF quantifies the direction of flow vectors. MBH divides the optical flow into its horizontal and vertical components (MBHx and MBHy), and discretizes the deriva-

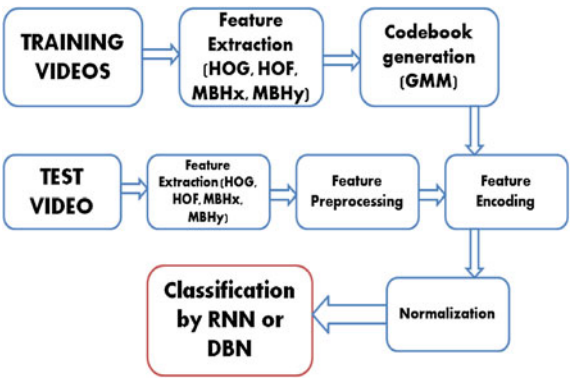


Fig. 2 Proposed framework of our system for action recognition. Please note that detailed information about LSTM-RNN was presented in Sect. 4.2

tives of every component. The dimensions of these descriptors are 96 for HOG, 108 for HOF and 192 for MBH (96 for MBHx and 96 for MBHy). After that we apply Principal Component Analysis (PCA) to each of these descriptors to half their dimension. The no. of features after PCA were selected mathematically. The dimensions are chosen so that almost 95 % of the total information was captured (i.e. we chose the minimum no. of dimensions so that cumulative energy content was just above or equal to 0.95). So, now the dimension of the HOG, HOF, MBHx and MBHy descriptors, which can represent any single feature point are 48, 54, 48 and 48. After this, the features are used to create fingerprints of each video. For this task we use the Fisher Vector (FV) encoding.

3.2 Fisher Vector Encoding

The Fisher Vector is a vector representation obtained by pooling video features. It is an extension of Bag of Visual words, which is also used to perform the same task. Unlike BOV, The Fisher Vector encoding uses soft clustering of data points, and uses higher order statistics of clusters, to pool the local features. Let $V = (x_1, \ldots, x_N)$ be an array of D dimensional feature vectors extracted from a video. Let $\theta = (\mu_k, \Sigma_k, \pi_k : k = 1, \ldots, K)$ be the mean, covariance and prior probability of a particular mode k of a Gaussian Mixture Model which fits the descriptors' distribution. The GMM assigns each point x_i to a mode k in the mixture with a weight value given by the posterior probability:

$$q_{ik} = \frac{\exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)\right]}{\sum_{t=1}^{K} \exp\left[-\frac{1}{2}(x_i - \mu_t)^T \Sigma_k^{-1}(x_i - \mu_t)\right]} \tag{1}$$

For each mode k, we consider the mean and covariance deviation vectors

$$u_{jk} = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}}, v_{jk} = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} \left[\left(\frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{2}$$

where j=1, 2, …, D covers all the dimensions. The conventional FV of video V is obtained by stacking the mean vectors and then the covariance vectors for all the modes in the mixtures:

$$\phi(V) = \left[\ldots u_k \ldots v_k \ldots \right]^T$$

In this experiment, we aim to form a separate Fisher Vector Encoding for each descriptor. In each case, we train a Gaussian Mixture Model (GMM) to obtain the codebook. For Fisher encoding, we choose the number of modes as k = 64 and sample a subset of 64,000 features from the training videos to get estimation of the GMM. After the training is finished the next task is to encode the feature descriptors into one fisher Vector per video. We choose to include only the derivatives with respect

to Gaussian mean (i.e. only u_k) in the FV, which results in an FV of size D $*$ 64, where D is the dimensionality of the feature descriptor. Including only the first order statistics in the Fisher Vector limits its dimension to half of the dimension of a conventional Fisher Vector. Thus, we conduct FV encoding for each kind of descriptor independently and the resulting super vectors are normalized by power and L2 normalization. Finally, these normalized Fisher Vectors are concatenated to form a super vector which represents the motion information for the training video. Hence we have one super vector representing for a single training video. It must be noted that if conventional Fisher Vectors were used, the dimension of the super vector would have been twice as large. Using the same PCA matrix and GMM parameters that were obtained from training data, we obtain four FV encodings with reduced dimension (each representing one descriptor), for each test video also. Finally, we concatenate these four fisher encodings (in a similar fashion). So, now we have one super vector representing any single video (both test and train). The dimension of each Super Vector is $(48 + 54 + 48 + 48) * 64 = 12672$. It should be noted that, for the new super vector of every single video, the first 3072 ($48 * 64$) values represent the HOG descriptor, the next 3456 ($54 * 64$) values represent HOF, the next 3072 ($48 * 64$) values represent MBHx and the next 3072 ($48 * 64$) represent MBHy descriptor.

3.3 Action Classification Using Deep Belief Network

The obtained testing and training super vectors are then fed to the classifiers. The ensemble has been performed by Deep Belief Network. A deep belief network (DBN) is a generative graphical model, consisting several multiple layers of hidden units. Connections are present between such layers, but not between the hidden units. Such models learn to extract an ordered representation of the training data. The detailed information about LSTM-RNN has been presented in Sect. 4.2. More details about evaluation process are provided in Sect. 4.3.

4 Experimental Results

4.1 Datasets

The **KTH** dataset was introduced by Schuldt et al. [16] in 2004. The dataset has been commonly used to evaluate models where handcrafted feature extraction is required. Every video contains exactly one of the 6 contains : walking, jogging, running, boxing, hand-waving and hand-clapping, performed by 25 subjects in 4 different scenarios. Here, each person performs the same action 3 or 4 times in the same video, with some empty frames between every action sequence in the video. The dataset contains 599 videos. 383 videos (performed by 9 subjects: 2, 3, 5, 6, 7, 8, 9, 10, and 22)

are used for training and remaining 216 videos (performed by rest of the subjects) are used for testing.

The **Weizmann** dataset includes 10 actions running, walking, skipping, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping sideways, waving-two-hands, waving one-hand and bending performed by nine subjects [3]. It contains 93 videos in all, each of which contains exactly one action sequence.

4.2 Classifier Training

In our work, we train a DBN classifier with two hidden layers for classification of the testing super vectors. Each value in the training super vector acts as an input, which is fed to the DBN for training. The input dimension of the DBN is varied depending on the feature (or combination of features) under consideration. We choose to use 70 hidden units in each of the hidden layers.

As mentioned earlier, the KTH training set consists of 383 videos; each of them is represented by an FV of length 12672. We trained the DBN for several iterations over these 383 training vectors and tested it on 216 testing vectors. In case of Weizmann dataset, we apply Leave-one-out validation approach. Here, a single subject is used for testing, while all other subjects are used to train the DBN. Each of the subjects in Weizmann dataset is used for testing once. The configuration of the DBN is kept the same for both datasets.

As a comparison, we have also tested our ensemble using the LSTM-RNN classifier [7]. Recurrent Neural Networks (RNN) are commonly used for analysis of sequential data, because of the usage of time-steps. A time step is a hidden state of RNN, whose value is a function of the previous hidden state and the current input vector. Long Short Term Memory (LSTM) is a special variant of RNN, where along with the hidden states, special cell states are also defined for every time step. The value of the current cell state is a function of the input vector and the previous cell state. The output value of the hidden state is a function of the current cell state. LSTM-RNNs provide additive interactions over conventional RNNs, which help to tackle the vanishing gradient problem, while backpropagating. We use an LSTM-RNN architecture with one hidden layer of LSTM cells. Here also, the input dimension of the RNN is varied depending on the feature (or combination of features) under consideration. There exists a full connection between LSTM cells and input layer. Also the LSTM cells have recurrent connections with all the LSTM cells. The softmax output layer is connected to LSTM outputs at each time step. We have experimented by varying the number of hidden LSTM, in the network. A configuration of 70 LSTM was found to be optimal for classification. Backpropagation algorithm was used to train the LSTM-RNN model. The configuration of LSTM-RNN used for both datasets is the same.

Table 1 Comparison of action-wise recognition accuracy for KTH dataset obtained by using LSTM-RNN and DBN classifiers and by using different features or combination of features. Best results are obtained when a combination of HOG, HOF and MBH is used

ACTION	HOG		HOF		MBH		HOG+HOF		HOG+MBH		HOF+MBH		HOG+HOF+MBH	
_	RNN	DBN	RNN	DBN	RNN	DBN	RNN	DBN	RNN	DBN	RNN	DBN	RNN	DBN
Boxing	97.69	95.37	98.15	99.53	98.61	100.00	97.11	99.53	98.15	100.00	98.61	100.00	97.69	100.00
Walking	97.22	99.07	99.54	100.00	99.54	100.00	99.30	100.00	98.61	100.00	98.15	100.00	98.84	100.00
Running	94.91	96.30	94.44	97.22	96.30	98.15	96.30	96.30	96.30	97.69	95.37	97.22	97.22	97.22
Jogging	94.44	97.22	96.30	96.76	97.22	99.53	96.52	100.00	96.76	100.00	97.22	98.61	96.76	100.00
Handclapping	89.81	94.90	94.44	98.61	94.44	98.15	94.91	97.22	94.44	98.61	95.14	98.61	95.14	98.61
Handwaving	93.52	93.06	94.91	96.76	95.37	97.22	96.76	96.76	97.22	96.76	96.53	98.15	96.06	97.69
Accuracy	94.60	95.99	96.29	98.15	96.92	98.84	96.82	98.30	96.91	98.84	96.84	98.76	**96.95**	**98.92**

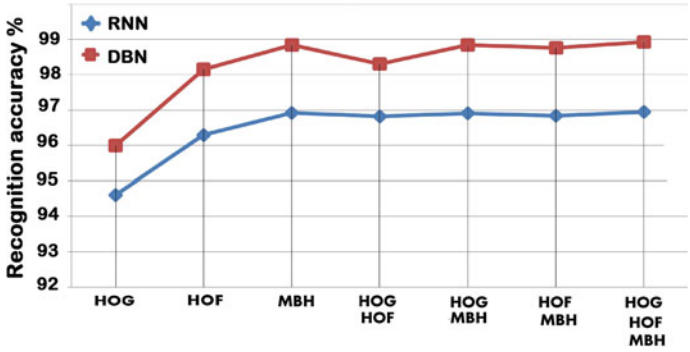


Fig. 3 Performance evaluation of proposed framework using different features (or combination of features), using different classifiers on KTH dataset

4.3 Quantitative Evaluation

We evaluate the results obtained by using different feature (or different combinations of features) for both KTH and Weizmann datasets. We report the accuracy obtained corresponding to the average across 10 trials. We use the values of the training super vectors obtained as inputs to be fed to the Deep Belief Network (DBN) classifier. We also test the ensemble with every possible combination of feature descriptors, so as to detect the combination which corresponds to the best recognition accuracy. Followed by this, the entire setup is again tested using the LSTM-RNN classifier and the results obtained are compared those obtained when using the DBN classifier.

KTH The recognition results for KTH dataset are shown in Table 1. Its observed that, for both LSTM-RNN and DBN classifiers, maximum accuracy is obtained when a combination of all the dense trajectory features are used. In such a case, we obtain a recognition accuracy 98.92 % using DBN classifier and 96.95 % using LSTM-RNN classifier. It is observed that since running and jogging are very similar actions, the ensemble wrongly classifies about 3 % of the testing data for running, as jogging. Also it can be inferred from the graph in Fig. 3 that, for both classifiers, the relative order of recognition accuracy obtained for a particular feature or a combination of features is the same.

Weizmann In case of Weizmann dataset, no variation in recognition accuracy is observed obtained for different features (or their different combination), when using the DBN classifier. For all combination of features, we obtain a perfect accuracy of 100 %. While using LSTM-RNN classifier also, negligible variance of results is observed amd an average accuracy of 98.96 % is achieved.

Hence, we conclude that the best recognition accuracy is obtained when a combination of HOG, HOF and MBH is used, while using a Deep Belief Network classifier. For such a setup, we report an average accuracy of **98.92** % for KTH and **100** % for

Table 2 Comparison of recognition accuracy of (a) KTH and (b) Weizmann dataset obtained by our pipeline with those obtained in previous work

METHOD	Accuracy (%)
(a)	
Our method	**98.92**
Baccouche et al. [1]	94.39
Kovashka et al. [9]	94.53
Gao et al. [6]	95.04
Liu et al. [11]	93.80
Bregonzio et al. [4]	93.17
Sun et al. [19]	94.0
Schindler and Gool [15]	92.70
Liu and Shah [12]	94.20
(b)	
Our method	**100**
Bregonzio et al. [4]	96.6
Sun et al. [18]	100.00
Ikizler et al. [8]	100
Weinland et al. [23]	93.6
Fathi et al. [5]	100
Sadek et al. [14]	97.8
Lin et al. [10]	100
Wang et al. [22]	100

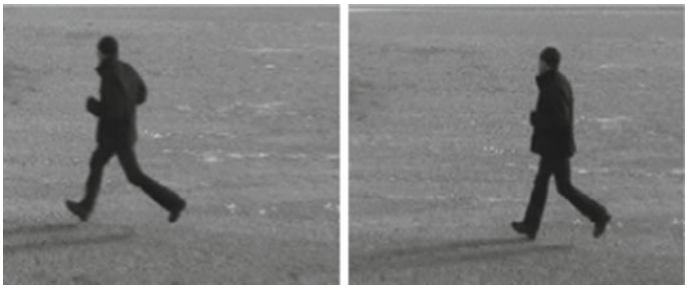


Fig. 4 Running and Jogging actions from KTH dataset. Since both the actions have similar trajectories, a considerable amount of test data for running is classified incorrectly as jogging

Weizmann dataset. Moving on with the best average accuracy obtained, we compare our results with some of the previous work done on the same datasets (using the same no. of samples) in Table 2a, b (Fig. 4).

5 Conclusion and Discussion

In this paper, we have proposed a framework, where for each stage, the best known approach has been used. Such a fusion of the best possible approaches has proved to be highly successful in the task of action classification. Experimental results show that the proposed pipeline gives competitive results, both on KTH (98.92 %) and Weizmann (100 %). As future work, we will examine a similar setup, where for feature encoding, only second order statistics would be encoded in the respective Fisher Vector of each descriptor. This would help us to evaluate the relative importance of Gaussian mean and co-variance, computed by Gaussian Mixture Model. Also, recent works are shifting towards other challenging video datasets, which contain in-the-wild videos. Therefore, we aim to confirm the generality of our approach by evaluating it on recent datasets, e.g. UCF sports, UCF-101, J-HMDB etc. Also, in the near future, we aim to investigate deep learning methods to classify actions in videos, in order to automate the process of feature learning.

References

1. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: Human Behavior Understanding, pp. 29–39. Springer (2011)
2. Bengio, Y., LeCun, Y., et al.: Scaling learning algorithms towards ai. Large-scale kernel machines 34(5) (2007)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: The Tenth IEEE International Conference on Computer Vision (ICCV'05). pp. 1395–1402 (2005)
4. Bregonzio, M., Gong, S., Xiang, T.: Recognising action as clouds of space-time interest points. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. pp. 1948–1955. IEEE (2009)
5. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp. 1–8. IEEE (2008)
6. Gao, Z., Chen, M.Y., Hauptmann, A.G., Cai, A.: Comparing evaluation protocols on the kth dataset. In: Human Behavior Understanding, pp. 88–100. Springer (2010)
7. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with lstm recurrent networks. The Journal of Machine Learning Research 3, 115–143 (2003)
8. Ikizler, N., Duygulu, P.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image and Vision Computing 27(10), 1515–1526 (2009)
9. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2046–2053. IEEE (2010)
10. Lin, Z., Jiang, Z., Davis, L.S.: Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th International Conference on Computer Vision,. pp. 444–451. IEEE (2009)
11. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. pp. 1996–2003. IEEE (2009)
12. Liu, J., Shah, M.: Learning human actions via information maximization. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp. 1–8. IEEE (2008)

13. Oneata, D., Verbeek, J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: 2013 IEEE International Conference on Computer Vision (ICCV). pp. 1817–1824. IEEE (2013)
14. Sadek, S., Al-Hamadi, A., Michaelis, B., Sayed, U.: An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. EURASIP Journal on Advances in Signal Processing 2011(1), 540375 (2011)
15. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp. 1–8. IEEE (2008)
16. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. vol. 3, pp. 32–36. IEEE (2004)
17. Sun, C., Nevatia, R.: Large-scale web video event classification by use of fisher vectors. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV). pp. 15–22. IEEE (2013)
18. Sun, C., Junejo, I., Foroosh, H.: Action recognition using rank-1 approximation of joint self-similarity volume. In: 2011 IEEE International Conference on Computer Vision (ICCV). pp. 1007–1012. IEEE (2011)
19. Sun, X., Chen, M., Hauptmann, A.: Action recognition via local descriptors and holistic features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. pp. 58–65. IEEE (2009)
20. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),. pp. 3169–3176. IEEE (2011)
21. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. In: BMVC 2009-British Machine Vision Conference. pp. 124–1. BMVA Press (2009)
22. Wang, Y., Mori, G.: Human action recognition by semilatent topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(10), 1762–1774 (2009)
23. Weinland, D., Boyer, E.: Action recognition using exemplar-based embedding. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. pp. 1–7. IEEE (2008)

Detection Algorithm for Copy-Move Forgery Based on Circle Block

Choudhary Shyam Prakash and Sushila Maheshkar

Abstract Today lots of software tools are available which are used to manipulate the images easily to change their originality. The technique which is usually used these days for tampering an image without leaving any microscopic evidence is copy-move forgery. There are many existing techniques to detect image tampering but their computational complexity is high. Here we present a robust and effective technique to find the tampered region. Initially the given image is divided into fixed size blocks and DCT is applied on each block for feature extraction. Circle is used to represent each transformed block with two feature vectors. In this way we reduce the dimension of the blocks to extract the feature vectors. Then lexicographical sort is applied to sort the extracted feature vectors. Matching algorithm is applied to detect the tampered regions. Results show that our algorithm is robust and has less computational complexity than the existing one.

Keywords Image forensics · Copy-Move forgery · Dimension reduction · Circle block · Region duplication detection

1 Introduction

These days there are many software (e.g. Photoshop) and applications are available which are used to edit a picture comfortably and modify it without any noticeable evidence. Thats why it is difficult to discern that a given image is original or forged. It cause many problem generally in insurance claims, courtroom witness and scientific scams. One of the famous example is shown in Fig. 1 One of the cosmonauts, Grigoriy Nelyubov from the Russian team which completed an orbit of the earth for

C.S. Prakash (✉) · S. Maheshkar
Department of Computer Science and Engineering,
Indian School of Mines, Dhanbad, India
e-mail: shyamprakash2008@yahoo.com
URL: http://www.ismdhanbad.ac.in/

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_32

Fig. 1 1961-cosmonauts in which left one is original and right one is tampered

the first time in 1961 led by Yuri Gagarin, was removed from a photo of the team taken after their journey. Further he was removed from the team after finding him guilty of misbehaving [1].

There are many types of image tampering method in which copy-move forgery is broadly used where any image is covered by any object or natural image. Previously many methods have been developed for detection of image forgery detection based on square block matching. DCT based features are used by Fridrich [5] for forgery detection, which is sensitive to variation in duplicated region when there is any additive noise in the image. Later on Huang et al. [6] reduce the dimension to improve the performance but their method did not succeeded in detection of multiple copy-move forgery. Popescu et al. [10] proposed a new method in which PCA based feature is used which is capable of detecting the forgery when there is additive noise but the accuracy of detection is not adequate. Luo et al. [7] proposed proposed a method in which color feature and block intensity ratio is used to show the robustness of their method. Bayram et al. [2] used Fourier-Mellin transform (FMT) to extract the feature for each block. They used FMT to project the feature vector in one dimension. Luo et al. [7] and Mahdian et al. [8] used blur moment invariants to locate the forgery region. Pan et al. [9] used SIFT features to detect the duplicated regions which is much robust. The methods discussed above have the higher computational complexity as they used quantized square blocks for matching. As the dimension of feature vectors are higher, the efficiency of detection is affected specially when the image resolution and size is high.

Here we come up with an efficient and robust detection method based on enhanced DCT. After comparing with the existing methods the prime features of proposed method are as follows:

- Feature vectors are reduced in dimension.
- Robustness against various attacks (Gaussian blurring, noise contamination, multiple copy-move).
- Computational complexity is low.

The rest of this paper is organized as follows. In Sect. 2 the proposed method is described in details. Results and discussion are presented in Sect. 3 whereas the proposed technique is concluded in Sect. 4.

2 Proposed Method

Normally it is known that in copy-move forged image there must be two identical regions are present. Exceptionally if two large regions are present in image such as blue sky, then it won't be considered. The task of forgery detection method is to find the input image have any duplicated region. Since the shape and size of duplicated region is undetermined and it is hardly possible to check each possible pairs of region with distinct shape and size. So it will be sufficient to divide the input image in fixed-sized overlapping blocks and apply the further process of matching to get the duplicated region. In this process of matching, first we represent the blocks by its features. To make our algorithm robust and effective we need a good feature extraction method. Once all the blocks are perfectly represented by some features then matching of block is done. The features of the matching block will be same and it is sorted lexicographically to make our matching process more effective. In this way computational complexity of the proposed detection algorithm is reduced as compared to the existing methods.

2.1 Algorithm Flow

The framework of detection algorithm is shown in Fig. 2.

- Dividing the image in to fixed size sub-blocks.
- To generate the quantized coefficients, DCT is applied on each sub-block.
- Each quantized sub-block is represented by a circle block and extracts appropriate features from each circle block.
- Explore the similar block pairs.
- Extract correct block and represent the output.

2.2 Implementation Details

Step 1: Let the input image I is gray scale of size $M \times N$, if the image is color then convert it by using the standard formula $I = 0.299R + 0.587G + 0.114B$. Firstly the input image is split into overlapping fixed size sub-block of $b \times b$ pixels in which each block have one different row and column. Every sub-block is denoted as b_{ij}, where i and j indicate the starting point of the block's row and column respectively.

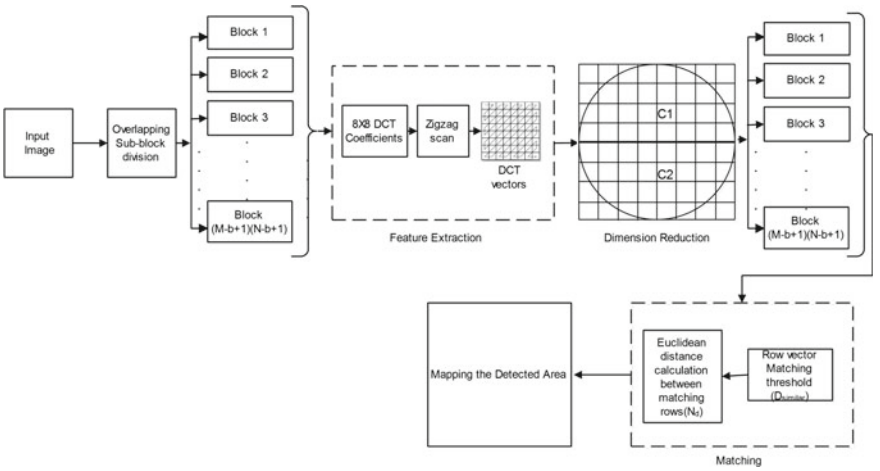


Fig. 2 Algorithm flow of detection method

$$b_{ij} = f(x+j, y+i) \; where \begin{cases} x, y \in (0, 1, 2, ...b-1) \\ i \in (1, 2, ...M-b+1) \\ j \in (1, 2, ...N-b+1) \end{cases} \tag{1}$$

In this way we obtain N_{blocks} of overlapping sub-blocks from input image.

$$N_{blocks} = (M-b+1)(N-b+1) \tag{2}$$

Step 2: Apply DCT on each block. After that we get a matrix of the same size, which represent the corresponding blocks.

Step 3: Assume that the block b_i where $i = 1, 2, ...N_{blocks}$ is of 8×8 size, hence the size of coefficient matrix is also 8×8 and there are 64 elements in the matrix. According to the nature of DCT, the energy focuses on the low frequency coefficients. Therefore we pruned the high frequency coefficients. Here the low frequency coefficients are extracted in a zigzag order which occupy $1/4$th DCT coefficients. For this reason, we take circle block is considered to represent the coefficient matrix.Hence we divide the circle block in two semicircles along with horizontal and vertical direction as shown in Fig. 3.

If r is radius of the circle, the ratio between area of circle and area of the block is given by

$$ratio = \frac{Area\ of\ Circle}{Area\ of\ Block} = \frac{\pi r^2}{4r^2} \approx 0.7853 \tag{3}$$

From above calculation it signifies that circle block can be used to represent the block as it consider most of the coefficients of the block and leave only few of them. So

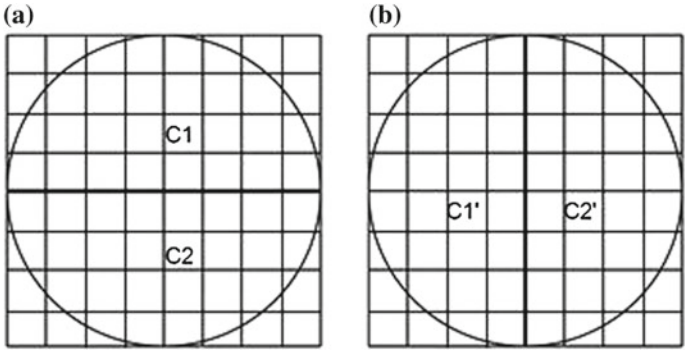


Fig. 3 Circle divided along with **a** horizontal and **b** vertical direction

it can be concluded that computational complexity can be reduced by using a circle block.
For block matching the circle is divided along with horizontal (Case 1) and vertical (Case 2) direction. These two cases are discussed below in detail.

Case 1: The circle is divided into two semicircle C1 and C2 along with horizontal direction as shown in Fig. 3a. Features of C1 and C2 are denoted by v_1 and v_2 respectively. It is calculated as given in equation (6).

$$v_i = \frac{\sum f(x,y)}{Area\ of\ Circle}\ where\ f(x,y) \in Area\ of\ semicircle_i\ ; i\ = 1,2. \tag{4}$$

Here v_i is the mean of the coefficient value analogous to each C_i. In this way two feature vectors are obtained which can be collectively represented as a feature vector with size 1×2 as

$$V = [v_1, v_2] \tag{5}$$

Thus the dimension reduction is more as compared to other methods [5, 6, 10] value reduction is by 1×64, 1×16, 1×32 features vectors respectively as shown in Table 2.

Case 2: Similarly, the circle is divided into two semicircles C1' and C2' along with vertical direction as shown in Fig. 3b. Features of C1' and C2' can be denoted as v_1' and v_2' and it can also be obtained as shown in equation (6) and the feature vectors are represented as

$$V' = [v_1', v_2'] \tag{6}$$

To check whether the feature vector is robust, post processing operations such as additive white noise (AWGN) and Gaussian blurring are applied and analysed. Gaussian blurring only affect some high frequency components and there is a little change in low frequency components. For justification of the robustness of feature

Table 1 The robustness of feature vectors

Feature vector	$V_{original}$	$V_{Post-Processing}$			
		AWGN (SNR = 25 db)	AWGN (SNR = 50 db)	Gaussian blurring (w = 3, = 1)	Gaussian blurring (w = 5, = 0.5)
v_1	0.7969	−0.7724	−0.7459	−0.1707	−0.4740
v_2	−9.8881	−0.1673	−0.1488	−0.0706	−0.1186
v'_1	−14.3910	0.3473	0.4204	0.0690	0.2480
v'_2	−3.0982	0.0489	0.0122	0.0547	0.0436
Correlation coefficient		1.0	1.0	1.0	1.0

vectors, we take a standard image (e.g. Baboon) and randomly select a 8×8 block and perform some post processing operation such as AWGN and Gaussian blurring with different parameters as shown in Table 1.

The correlation between post processed data is calculated. From Table 1, it is observed that the correlation is 1.0 which indicates that the feature vectors are robust. It also indicates that the reduction of dimension is successful.

Step 4: The extracted feature vectors are arranged in a matrix P which have dimension of $(M - b + 1)(N - b + 1) \times 2$. Now the matrix P is sorted in lexicographical order. As each element is a vector, the sorted set is defined as $\hat{P}$. The Euclidean distance m_match $(\hat{P}_i, \hat{P}_{i+j})$ between adjacent pair of $\hat{P}$. This distance is compared with the present threshold $D_{similar}$(Explained in detail in Sect. 3.1. If the distance m_match is smaller than the threshold $D_{similar}$, then the tampered region is detected. mathematically it can be represented as follows

$$m\_match(\hat{P}_i, \hat{P}_{i+j}) = \sqrt{\sum_{k=1}^{2} (v_i^k, v_{i+j}^k)} < D_{similar} \tag{7}$$

where $\hat{P}_i = (\hat{v}_i^1, \hat{v}_i^2)$ and $\hat{P}_{i+j} = (\hat{v}_{i+j}^1, \hat{v}_{i+j}^2)$
It is also possible that the neighbouring blocks may have the similar feature vectors. Hence between two similar blocks the actual distance is calculated as:

$$m\_distance(V_i, V_{i+j}) = \sqrt{(x_i - x_{i+j})^2 + (y_i - y_{i+j})^2} > N_d \tag{8}$$

where (x, y) is the centre of corresponding block and N_d is threshold.

Step 5: Apply the morphologic operations on black map image to remove the isolated regions to fill the holes in the marked regions, to get the final output.

Table 2 Comparison of Dimension Reduction

Literatures	Extraction method	Feature dimension
Fredrich et al. [5]	DCT	64
Popescu [10]	PCA	32
Huang et al. [6]	Improved DCT	16
Cao et.al. [3]	Block representing	4
Proposed	Block representing	2

3 Experimental Results and Analysis

The experiments are performed on the Matlab R2013a. All the images taken in this experiment are of size 256×256 pixel image in JPG format [4] (Table 2).

3.1 Threshold Setting

Let D_1 is the duplicated region, D_2 is the detected duplicated region. T_1 and T_2 are the altered region and detected altered region respectively. The Detection Accuracy Rate (DAR) and False Positive Rate (FPR) is calculated by the following equation:

$$DAR = \frac{|D_1 \cap D_2| + |T_1 \cap T_2|}{|D_1| + |T_1|}, \; FPR = \frac{|D_2 - D_1| + |T_2 - T_1|}{|D_2| + |T_2|} \tag{9}$$

Here overlapping of sub-block and circle representation method is used for extracting the features. To set the threshold parameters, the radius of the circle is selected very carefully. For this, we take different images with duplicated regions and set the radius of circle varying from 2 to 6, with unit increment. Then a set of value for b, $D_{similar}$ and N_d. If the circle radius(r) is set to 4 for color images then, $b = 8, D_{similar} = 0.0015$ and $N_d = 120$ is obtained. The results obtained are shown in Fig. 4. This technique is tested on set of standard database [4]. All the images are of size 256×256 for detection of the tampered region.

The result obtained for DAR, FPR, FNR which has been tested on the database [4] and the average result is shown in Table 3. The graphical representation for this is shown in Fig. 5. It is observed that the DAR in average case is more than 75 %. It signifies that the proposed technique is robust and efficient. In few cases it detect only 50 % of the tampered region but it can be a positive clue for the verification of originality of the image. FPR in most of the cases is zero and hence we can conclude that the proposed technique is efficient to detect the forgeries. It is also observed that

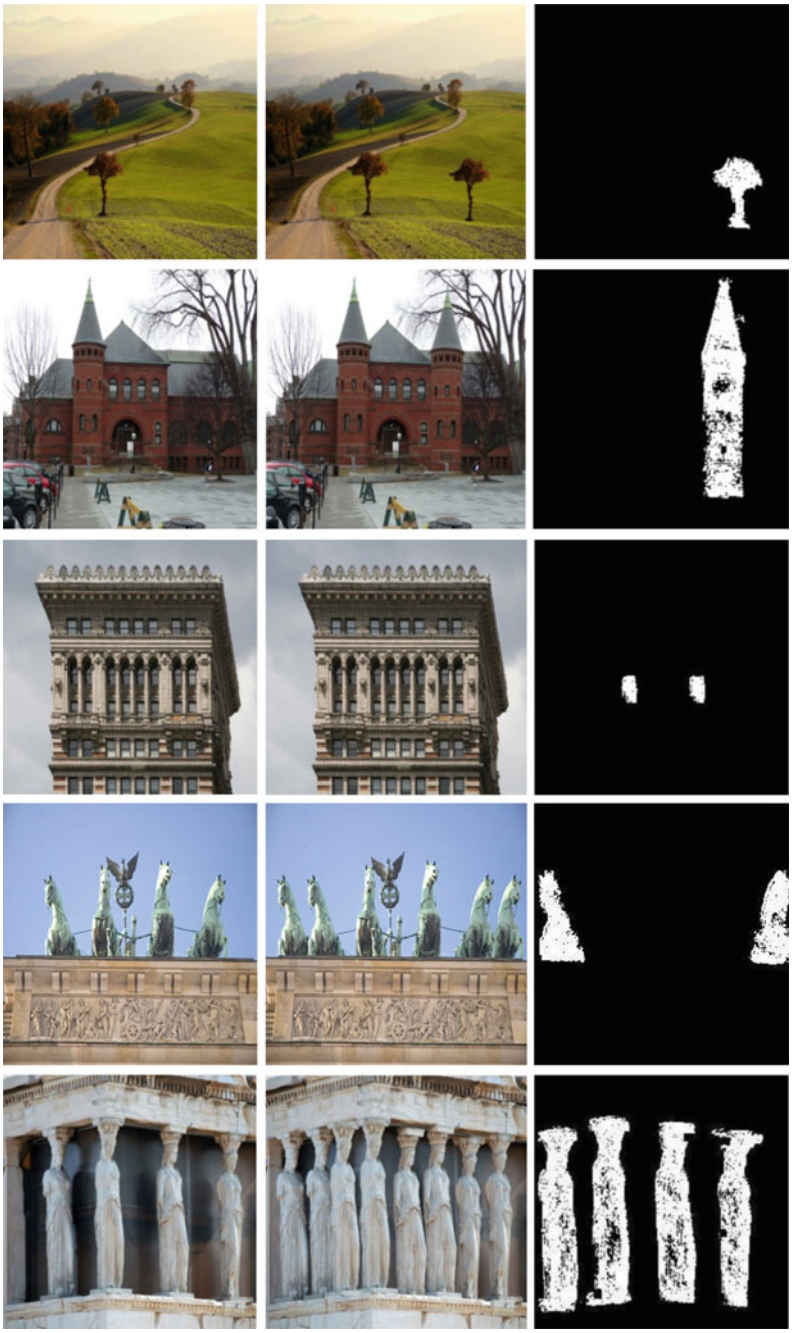


Fig. 4 The forgery detection results (original image, tampered image, detection image)

Table 3 Average valur of DAR, FNR, FPR

	DAR	FNR	FPR
Case 1	75.55	24.45	2.88
Case 2	71.40	28.58	1.38

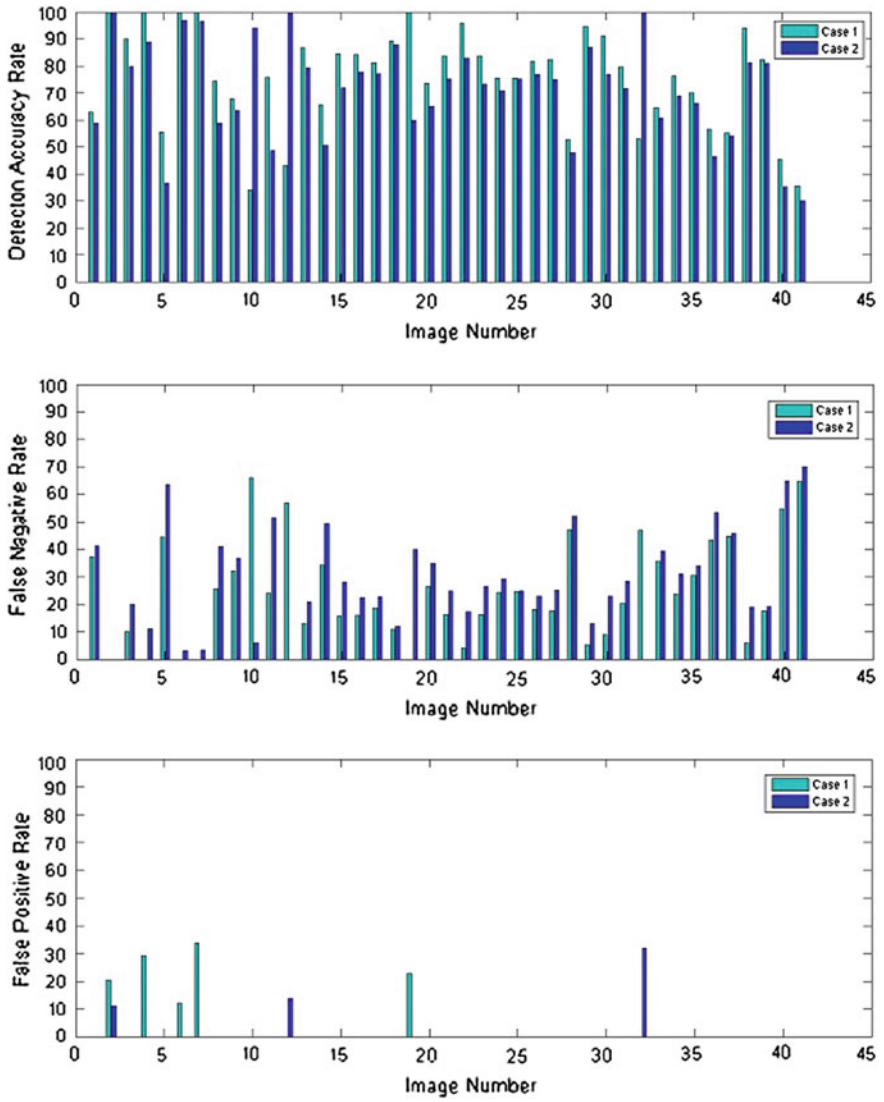


Fig. 5 DAR, FNR, FPR of different images respectively

in few cases the FPR is positive but the rate of detection is 100 % in such cases. FNR in most of the cases is less than 25 % which shows that the detection of forgeries are accurate and efficient.

4 Conclusion

In this paper a robust and effective algorithm for copy-move tamper detection is proposed. It is a passive method for tamper detection that means a priori knowledge about the tested is not required. The feature reduction is achieved up to the mark as compared to the existing methods [5, 6, 10] as shown in Table 2. It is also observed that the DAR is more than 75 % in average case indicating the efficiency of the proposed algorithm. The robustness of feature vectors is tested on AWGN for various SNR levels and Gaussian blurring and it is observed that the correlation coefficient is 1. This indicates the robustness of the proposed technique. hence, we believe that our method is efficient and robust enough to detect the image forgery.

References

1. http://www.fourandsix.com/photo-tampering-history/tag/science
2. Bayram, S., Sencar, H.T., Memon, N.: An efficient and robust method for detecting copy-move forgery. In: Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on. pp. 1053–1056. IEEE (2009)
3. Cao, Y., Gao, T., Fan, L., Yang, Q.: A robust detection algorithm for copy-move forgery in digital images. Forensic science international 214(1), 33–43 (2012)
4. Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E.: An evaluation of popular copy-move forgery detection approaches. Information Forensics and Security, IEEE Transactions on 7(6), 1841–1854 (2012)
5. Fridrich, A.J., Soukal, B.D., Lukáš, A.J.: Detection of copy-move forgery in digital images. In: in Proceedings of Digital Forensic Research Workshop. Citeseer (2003)
6. Huang, Y., Lu, W., Sun, W., Long, D.: Improved dct-based detection of copy-move forgery in images. Forensic science international 206(1), 178–184 (2011)
7. Luo, W., Huang, J., Qiu, G.: Robust detection of region-duplication forgery in digital image. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. vol. 4, pp. 746–749. IEEE (2006)
8. Mahdian, B., Saic, S.: Detection of copy–move forgery using a method based on blur moment invariants. Forensic science international 171(2), 180–189 (2007)
9. Pan, X., Lyu, S.: Detecting image region duplication using sift features. In: Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. pp. 1706–1709. IEEE (2010)
10. Popescu, A., Farid, H.: Exposing digital forgeries by detecting duplicated image region [technical report]. 2004-515. Hanover, Department of Computer Science, Dartmouth College. USA (2004)

FPGA Implementation of GMM Algorithm for Background Subtractions in Video Sequences

S. Arivazhagan and K. Kiruthika

Abstract Moving object detection is an important feature for video surveillance based applications. Many background subtraction methods are available for object detection. Gaussian mixture modeling (GMM) is one of the best methods used for background subtraction which is the first and foremost step for video processing. The main objective is to implement the Gaussian mixture modeling (GMM) algorithm in Field-Programmable Gate Array (FPGA). In this proposed GMM algorithm, three Gaussian parameters are taken and the three parameters with learning rate over the neighborhood parameters were updated. From the updated parameters, the background pixels are classified. The background subtraction has been performed for consecutive frames by the updated parameters. The hardware architecture for Gaussian mixture modeling has been designed. The algorithm has been performed in offline from the collected data set. It can able to process up to frame size of 240 × 240.

Keywords Background subtraction · Moving object detection · Gaussian mixture modeling (GMM) · Hardware architecture · Field programmable gate array (FPGA)

1 Introduction

Video processing is a method used to analyze video streams electronically. It helps end-users to detect events of interest and to recognize a range of behaviours in real time. Sequences of image, are actually called as videos, each of the image are called as a frame. The frames are visible with high frequency rate in which human eyes cannot percept the individual frames of its content. It is conspicuous that all image

S. Arivazhagan (✉) · K. Kiruthika
Department of ECE, Mepco Schlenk Engineering College, Sivakasi, India
e-mail: sarivu@mepcoeng.ac.in

K. Kiruthika
e-mail: k.keerthi.kiruthika@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_33

processing algorithm can be applied to each individual frame. Besides, the contents of each consecutive frames are visually closely related. Visual perception of a human can be modeled as a hierarchy of abstractions. At the first level are the raw pixels with RGB or brightness information. Further processing yields features such as lines, curves, colours, edges and corners regions. A next abstraction layer may interpret and combine these features as objects and their facets. At the highest level of abstraction, the human level concepts involving one or more objects and relationships among them. Object detection in a sequence of frames involves verifying the presence of an object in each frame and locating it precisely for recognition. There are two methods that realize object detection. One is changing at pixel level and another based on feature comparison. The first method is based on a few visual features, such as a specific colours are used to represent the object. It is fairly easy to identify all pixels with the same colour as the object. The first method is easy and better because of very fast detection of any kind of changes in the video sequences. In the later approach, is very difficult to be accurately detected and identified because of perceptual details of a specific person, such as different poses and illumination.

On the other hand, detecting a moving object has an important significance in video object detection and tracking. Compared with object detection, moving object detection complicates the detection problem by adding temporal change requirements. Several methods of moving object detection can be classified into groups approximately. The first method is thresholding technique over interface difference. These approaches based on the detection of temporal changes either at pixel level or block level. Using a predefined threshold value, the difference map is binarized to obtain the motion detection. The second method based on statistical tests constraints to pixelwise independent decisions. In this approach detection masks and filter have been considered but the masks cannot able to provide invariant change detection with respect to size and illuminations. And finally the third method is based on global energy frameworks, where the detection methods are performed using stochastic or deterministic random algorithm such as mean field or simulated annealing. For high speed parallel data processing such as video processing FPGA's are widely used. The design of VLSI techniques in video processing can be performed in FPGA. The new FPGA are having a features of VGA port and HDMI port for High definition video processing. The camera and high definition video source interfaces also available (Fig. 1).

The video processing in real time is time consuming. The implementation of video processing algorithms in hardware language offers parallelism, and thus significantly reduces the processing time. The video sequences are processed as each frame.

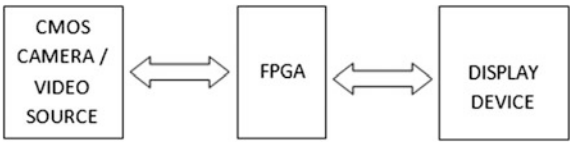


Fig. 1 FPGA Interfacing for video processing

The paper is organized as follows. Section 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. Future work has been discussed in Sect. 6.

2 Related Work

There are many research works have been done on video processing and moving object detection. Zhang Yunchu et al. described about the object detection techniques which was performed for the images captured at the night time, which having low contrast and SNR, poor distinction between the object and the background, that pose more challenges to moving object detection. A moving object detection algorithm for night surveillance based on dual-scale Approximate Median Filter background models was proposed. The moving object detection was robustly at night under adverse illumination conditions, with high spatial resolution resistant to noise and with low computational complexity [1]. Bo-Hao Chen, et al. proposed a moving object detection algorithm for intelligent transport system. The principal component was radial basis function which was utilized for the detection of moving objects [2]. An algorithm has developed for both high bit rate and low bit rate video streams.

Eun-Young Kang, et al. proposed a bimodal Gaussian approximation method using colour features and motion for moving object detection [3]. They introduce a compound moving object detection by combining motion and colour features. The accuracy of detected objects was increased by statistical optimization in high-resolution video sequences with motion camera and camera vibrations. Motion analysis involves information about the motion vector obtained H.264 decoder and a moving-edge map. A dedicated integrated circuit is used in real time moving object detection in video analysis. The designed circuits have been shown in the works of [4, 5].

Processing of the information determines whether the information contains specific object and exact location. This work is computationally profound and several attempts have been made to design hardware-based object detection algorithms. The majority of the proposed works target field-programmable gate-array (FPGA) implementations; additionally, targets on ASIC application Specific Integrated Circuits or operate on images of relatively small sizes in order to achieve real-time response in [6]. Bowmans, et al. proposed a statistical background modeling for moving object detection [7]. They assume the background is static in the video. This assumption implies that video is captured by a stationary camera like a common surveillance camera.

Implementation of OpenCV version of the Gaussian mixture algorithm has been shown in [8]. To reduce the circuit complexity they had utilized compressed ROM and binary multipliers. The enhancement of their work was done in [9]. The background and foreground pixels were identified based on the intensity of the

pixel. Xiaoyin et al. proposed fixed point object detection methodology using the histogram of oriented gradients (HOG) [10]. HOG has delivered just 1FPS (frames per second) on a high-end CPU but achieves high accuracy. The fixed point detection reduces circuit complexity.

Hongbo Zhu et al. proposed using row-parallel and pixel-parallel architectures for motion features from moving images in real. These architectures are based on the digital pixel sensor technology. To minimize the chip area the directional edge filtering of the input image was carried out in a row-parallel processing. As a result, self adaptive motion feature extraction has been established [11]. Most of these algorithms were assessed with software implementation on a general purpose processor (GPP) [12]. It is usually suficient for verifying the function of the algorithm. However, the difficulty occurs in real time at high data throughput video analysis. Hongtu Jiang et al. describe about the embedded automated video surveillance for a video segmentation unit [13]. The segmentation algorithm is explored with potential increase of segmentation results and hardware efficiency.

A. Yilmaz et al. presented an extensive survey on object tracking method and also reviewed the methods for object detection and tracking [14]. Describes the context of use, degree of applicability, evaluation criteria, and qualitative comparisons of the tracking algorithms.

3 System Model

In GMM algorithm, the moving objects are detected. The algorithm has been performed in hardware language. The input frames are stored in block RAM. For each frame the background subtraction has been performed from the updated parameters and the corresponding output frames are obtained as shown in Fig. 2.

3.1 System Architecture

For real time object detection, the input data capture from camera. The captured video sources are processed and Gaussian mixture algorithm has to be performed.

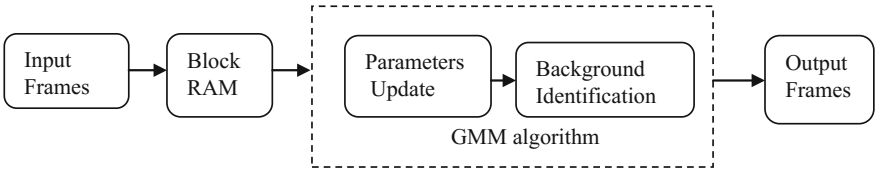


Fig. 2 Procedural block

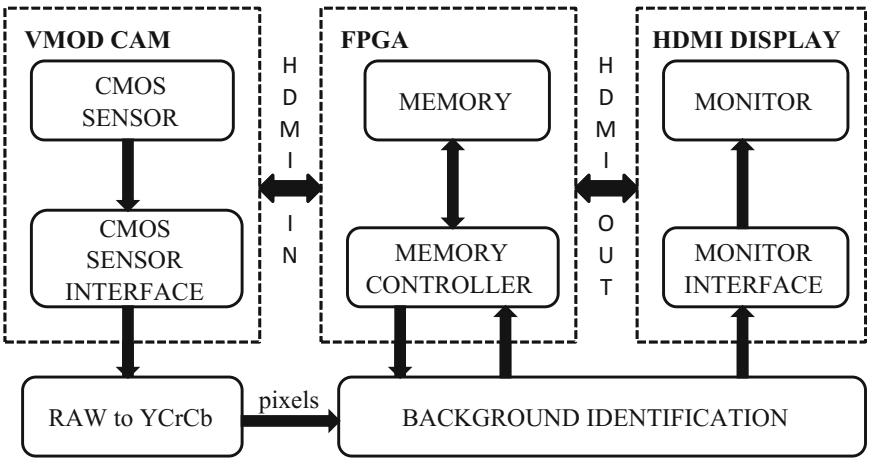


Fig. 3 System architecture

The motion objects are detected from the background subtraction which is to be displayed on display devices (Fig. 3).

3.2 GMM Algorithm

The GMM algorithm proposed by Stauffer and Grimson [15] has been modified, that deals with statistical background subtraction using a mixture of Gaussian distributions. The modified background subtraction requires only a minimum number of operations to perform. This reduces the design complexity in hardware language. A short description of GMM algorithm is given as follows.

Statistical model
The statistical model of video sequence of each pixel is composed by k Gaussian distributions with parameters mean, variance, weight and matchsum. For the each Gaussian of each pixel the Gaussian parameters differs and change every frame of the video sequence.

Parameters Update
When the frame is acquired, for each pixel, the K Gaussian distributions are sorted in decreasing order of a parameter, named Fitness $F_{k,t}$

$$F_{k,t} = w_{k,t}/\sigma_{k,t} \tag{1}$$

with K number of Gaussian distributions a match condition is checked that model the pixels. The match condition $Mk = 1$.

$$if \left|pixel - \mu_{k,t}\right| < \lambda^* \sigma_{k,t} \tag{2}$$

where λ represents the threshold value equal to 2.5 as in the OpenCV library. Equation (2) establishes if the pixel can be considered part of the background. A pixel can verify (1) for more than one Gaussian. The Gaussian that matches with the pixel ($Mk = 1$) is considered to be the "matched condition" with the highest fitness $F_{k,t}$ value and its parameters are updated as follows

$$\mu_{k,t+1} = \mu_{k,t} + \alpha_{k,t}(pixel - \mu_{k,t}) \tag{3}$$

$$\sigma^2_{k,t+1} = \sigma^2_{k,t} + \alpha_{k,t}\left[\left(pixel - \mu_{k,t}\right)^2 - \sigma^2_{k,t}\right. \tag{4}$$

$$w_{k,t+1} = w_{k,t} - \alpha_w w_{k,t+1} + \alpha_w \tag{5}$$

$$matchsum_{k,t+1} = matchsum_{k,t} + 1 \tag{6}$$

The parameter α_w is the learning rate for the weight from α_w the learning rate for mean and variance are calculated which is denoted as $\alpha_{k,t}$ and the equation as follows.

$$\alpha_{k,t} = \alpha_w / w_{k,t} \tag{7}$$

Equation (7) is the characteristic equation of the GMM algorithm proposed in OpenCV, where $\alpha_{k,t}$ is calculated as follows.

$$\alpha_{k,t} = \alpha_w . \eta(\text{pixel}, \mu_{k,t}, \sigma_{k,t}) \tag{8}$$

where η is the Gaussian probability density function.

For the matched Gaussian distributions, variances and means are unchanged while the weights are updated as

$$w_{k,t+1} = w_{k,t} - \alpha_w w_{k,t+1} \tag{9}$$

When the pixel does not match with any of the Gaussians function, a specific "No match" procedure is executed and the Gaussian distribution is updated with the smallest fitness value $F_{k,t}$.

$$\mu_{k,t+1} = pixel \tag{10}$$

$$matchsum_{k,t+1} = 1 \tag{11}$$

$$\sigma^2_{k,t+1} = vinit \tag{12}$$

$$w_{k,t+1} = 1/msumtot \tag{13}$$

where the fixed initialization value is represented by *vinit* and *msumtot* is the sum of the values of the *matchsum* of $k - 1$ Gaussians with highest fitness. The weight of the k-1 Gaussians with $F_{k,t}$ are reduced as in Eq. (12) while their variances and means are unchanged.

Background Identification

The background identification is performed by using the following equation. The algorithm for background subtraction in [6] is modified as follows.

$$B = \begin{cases} 0, & if \left| pixel - \mu_{k,t+1} \right| \leq \max(\mathrm{Th}, \lambda, \sigma_{k,t+1}) \\ 1, & if \left| pixel - \mu_{k,t+1} \right| \geq \max(\mathrm{Th}, \lambda, \sigma_{k,t+1}) \end{cases} \tag{14}$$

where $\sigma_{k,t+1}$ is the standard deviation calculated from variance. The set of the Gaussian distributions that verify the equation represents the background; if 0, then the pixel that matches one of these Gaussians will be classified as a background pixel. If the algorithm evokes 1, then "No match" condition occurs, the pixel will be classified as foreground.

3.3 Background Subtraction Architecture

From the video sequences, the variance, mean, weight and pixels of each frames are calculated. Those values are fed as input to the VLSI circuit and then for the next frame the parameter values are updated. The background subtraction is performed after processing consecutive frames.

The fitness function is sorted in decreasing order. The higher order fitness is used to update the parameters with match condition. The lower order fitness is used to update the parameters for unmatched condition. The number of fitness calculation depends upon the GMM parameter. The output variables are updated depending upon the selected line. Figure 4 shows the parameter updating block where the three parameters mean, variance and weight are updated with calculated fitness value and learning rate.

Match Flow

With k number of Gaussian distributions a match check condition is checked. If $Mk = 1$ *the* parameters are matched and updated. If $Mk = 0$ the parameters are unmatched, the mean and variance remains the same, the weight is updated (Fig. 5).

If match conditions are satisfied (i.e.,) $Mk = 1$ the mean are updated as shown in Fig. 6a, similarly for variance and weight the parameters are updated as in Fig. 6b, c.

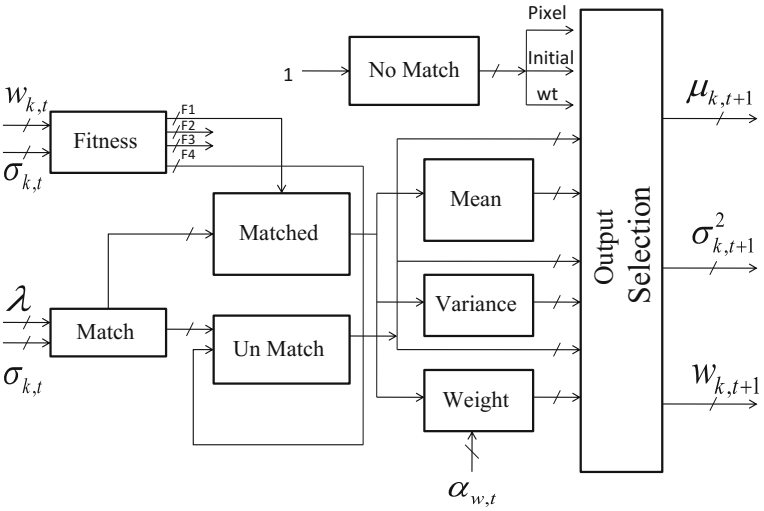


Fig. 4 Parameters update block

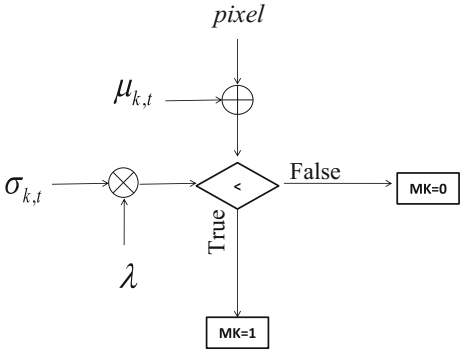


Fig. 5 Match flow chart

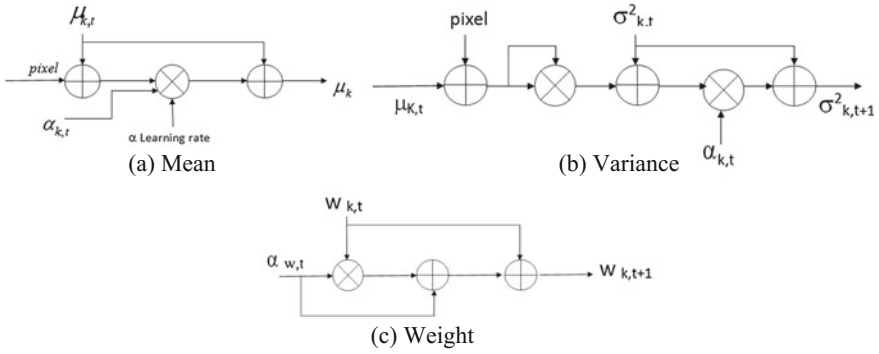


Fig. 6 Matched update circuit

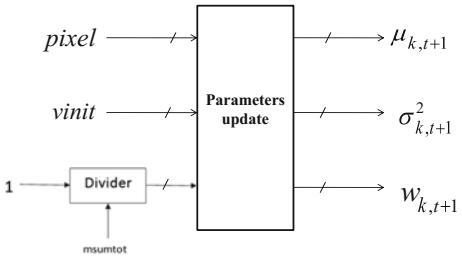


Fig. 7 No match block

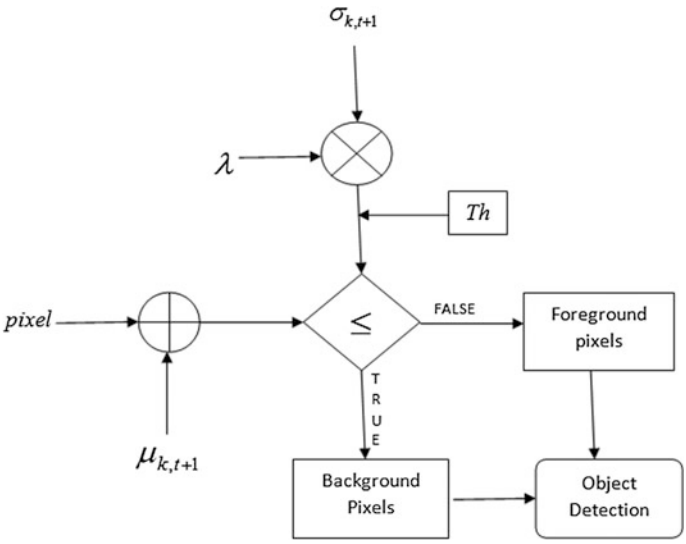


Fig. 8 Background subtraction block

For unmatched (i.e.,) $Mk = 0$ the mean, weight and variance are updated as the same values. For No match block the mean and variances are updated as shown in Fig. 7.

The background identification is performed by the Eq. (14). The pixels are classified as foreground and background and the moving objects are detected. The flow for background identification is shown in Fig. 8.

Fig. 9 Sample frames from Data set

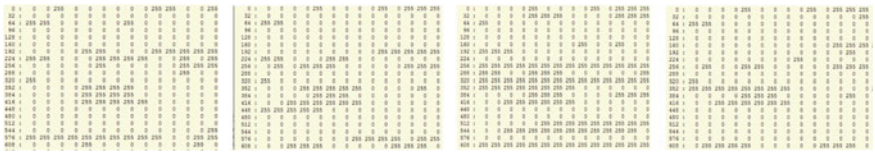


Fig. 10 Simulation results of background identification

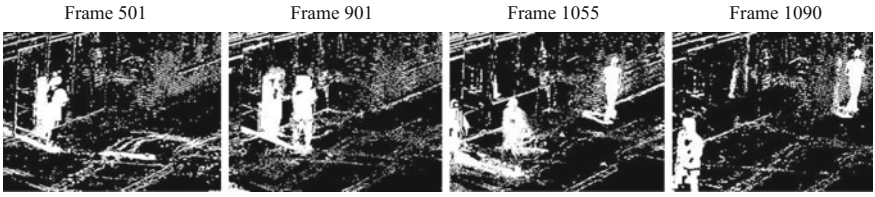


Fig. 11 Background subtracted frames

4 Results and Discussions

The algorithm has been performed in offline by storing the collected video sequences in FPGA memory. The collected video is consisting of 1250 frames, each frame of size 360 × 240 and of 24 bit depth having a horizontal and vertical resolution of 69 dpi. The samples are shown in Fig. 9.

Figure 10 shows the simulation results obtained from background identification block. The sample pixels for corresponding background subtracted frames are shown. The pixel format is 8 bits and each byte represent the intensity of the pixel. Typically 0 represents black and 255 is taken to be white. After the parameters are updated, the resultant values are stored as pixels in RAM.

The detected objects are verified in Matlab by processing the results obtained from the text file. The verified object detection frames have been shown in the Fig. 11.

5 Conclusion

In this proposed method, a moving object detection algorithm has been developed and the hardware circuit for Gaussian Mixture Modeling has been designed. The algorithm is modeled for three parameter values such as mean, variance, and weight. After updating the parameters for consecutive frames, the background subtraction has been performed. The fixed point representation has been performed to reduce the hardware complexity. The algorithm has been able to perform efficiently with vertical and horizontal resolution of 96 dpi.

6 Future Work

The implementation of GMM algorithm can be done in real time in the future. It has been planned to implement by interfacing CMOS camera with FPGA for real time moving object detection.

Acknowledgments 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. Zhang Yunchu, Li Yibin, and Zhang Jianbin, "Moving object detection in the low illumination night scene," *IET International Conference on Information Science and Control Engineering 2012 (ICISCE 2012),* Dec. 2012, pp. 1–4.
2. Bo-Hao Chen, Shih-Chia Huang, "An Advanced Moving Object Detection Algorithm for Automatic Traffic Monitoring in Real-World Limited Bandwidth Networks," *IEEE Transactions on Multimedia*, vol. 16. no. 3, pp. 837–847, April 2014.
3. V. Mejia, Eun-Young Kang, "Automatic moving object detection using motion and color features and bi-modal Gaussian approximation," *IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2011, pp. 2922–2927.
4. Hongtu Jiang, Hakan Ardo, and Viktor Owall, "Hardware Accelerator Design for Video Segmentation with Multimodal Background Modelling," *IEEE International Symposium on Circuits and Systems*, vol 2. May 2005, pp. 1142–1145.
5. Tomasz Kryjak, Mateusz Komorkiewicz, and Marek Gorgon, "Real-time Moving Object Detection For Video Surveillance System In FPGA," *Conference on Design and Architecture for signal and Image Processings*, pp. 1–8, Nov 2011.
6. Mariangela Genovese and Ettore Napoli, "ASIC AND FPGA Implementation Of The Gaussian Mixture Model Algorithm For Real-time Segmentation Of High Definition Video," *IEEE Transactions on very large scale integration (VLSI) systems*, vol. 22, no. 3, March 2014, pp. 537–547.
7. T. Bouwmans, F. El Baf and B. Vachon, "Statistical Background Modelling for Foreground Detection: A Survey," *in Handbook of Pattern Recognition and Computer Vision*, World Scientific Publishing, 2010, pp. 181–199.

8. Mariangela Genovese and Ettore Napoli, "An Fpga - based Real-time Background Identification Circuit For 1080p Video," *8th International Conference on Signal Image Technology and Internet Based Systems*, Nov 2012, pp. 330–335.
9. Ge Guo, Mary E. Kaye, and Yun Zhang, "Enhancement of Gaussian Background Modelling Algorithm for Moving Object Detection & Its Implementation on FPGA," *Proceeding of the IEEE 28th Canadian Conference on Electrical and Computer Engineering Halifax, Canada*, May 3–6, 2015, pp. 118–122.
10. Xiaoyin Ma, Walid A. Najjar, and Amit K. Roy-Chowdhury, "Evaluation And Acceleration Of High-throughput Fixed-point Object Detection On FPGA'S," *IEEE Transactions On Circuits And Systems For Video Technology*, Vol. 25, No. 6, June 2015, pp. 1051–1062.
11. H. Jiang, V. O wall and H. Ardo, "Real-Time Video Segmentation with VGA Resolution and Memory Bandwidth Reduction," *IEEE International Conference on Video and Signal Based Surveillance*, 2006. AVSS'06., pp. 104–109, Nov. 2006.
12. M. Genovese, E. Napoli and N. Petra,"OpenCV compatible real time processor for background, foreground identification," *Microelectronics (ICM), 2010 International Conference,* pp. 487–470, Dec. 2010.
13. H. Jiang, H. Ardö, and V. Öwall, "A hardware architecture for real-time video segmentation utilizing memory reduction techniques,"*IEEE Trans. Circuit Syst. Video Technol.,* vol. 19, no. 2, pp. 226–236,Feb. 2009.
14. A. Yilmaz, O. Javed and M. Shah,"Object tracking: A survey," *ACM Comput. Surv.,* vol. 38, no. 4, Dec. 2006.
15. http://dparks.wikidot.com/background-subtraction.

Site Suitability Evaluation for Urban Development Using Remote Sensing, GIS and Analytic Hierarchy Process (AHP)

Anugya, Virendra Kumar and Kamal Jain

Abstract An accurate and authentic data is prerequisite for proper planning and management. If one looks for proper identification and mapping of urban development site for any city, then accurate and authentic data on geomorphology, transport network, land use/land cover and ground water become paramount. In order to achieve such data in time satellite remote sensing and geographic information system techniques has proved its potentiality. The importance of this technique coupled with Analytic Hierarchy Process (AHP) in site suitability analysis for urban development site selection is established and accepted worldwide too and to know the present actual status of environmental impact in surrounding of urban development site. Remote Sensing, GIS, GPS and AHP method is a vital tool for identification, comparison and multi criterion decision making analysis of urban development site's proper planning and management. Now keeping in view the availability of high resolution data of IKONOS satellite, cartosat and IRS 1C/1D LISS—III data has been used for preparation of various thematic layers in Lucknow city and its environs. The study describes the detailed information on the site suitability analysis for urban development site selection. The final maps of the study area prepared using GIS software and AHP method, can widely applied to compile and analyze the data on site selection for proper planning and management. It is necessary to generate digital data on site suitability for urban development sites for local bodies/development authorities in GIS & AHP environment, in which data are reliable and comparable.

Keywords Remote sensing and GIS · Site suitability · Urban development · Multi criterion layers · Pairwise comparison and AHP

Anugya (✉)
IIT Roorkee, Roorkee, Uttarakhand, India
e-mail: anugya.shukla92@gmail.com

V. Kumar
Remote Sensing Application Centre, Lucknow, U.P, India

K. Jain
IIT Roorkee, Roorkee, Uttarakhand, India

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_34

1 Introduction

The urban areas in developing countries have witnessed tremendous changes in terms of population growth and urbanization [1]. In the absence of proper urban management practice, uncontrolled and rapid increase in population pose enormous challenges to governments in providing adequate shelter to the millions homeless and poor in urban areas. This has also posed great concern among urban planners. Urban growth due to immigration has led to increase in population density. There is an increase in slum and squatter settlements in cities and urban area [5]. This has led to shortage of facilities and increasing demand of urban land for residential purposes.

The migration of rural people to urban areas hoping for better job opportunities, better standard of living and higher level of education will not stop. One of the reports says that India is one of the most rapidly urbanizing countries in the world. Therefore, there is an urgent need to regulate the urbanization process in a systematic and scientific way for future development. Urbanization is a dynamic phenomenon, which keeps on changing with time. Therefore, accurate and timely data is required for proper urban planning. Urban planners need to use variety of data and methods to solve the problems of urban areas [4]. With the launching of high resolution satellites and availability of remotely sensed data, gives synoptic view and crucial means of the planning areas.

The decision makers/urban planners need authenticated and accurate data and sophisticated computer tools for making dynamic decisions. Remote sensing and GIS are such tools or aids, which help the planners to accurately create and manage data. GIS is used as analysis tool as a means of specifying logical and mathematical relationships among map layers to get new derivative map layers. Any new data can be added to existing GIS database easily. Thus remote sensing data provides reliable, timely, accurate and periodic spatial data while GIS provides various integrating tools for handling spatial and non-spatial data to arrive at solution for decision making tool for pairwise comparison of layer [10, 13]. This study is an endeavor for evaluation of suitable site for urban development.

2 Related Work Done and Literature Review

During the recent years, the rate of urbanization was so fast, as there were very little planned expansion, thereby leading to poor management practices, which have created many urban problems like slum, traffic congestion, sewage and transportation network etc. Consequently, it has led to immense deleterious impact on existing land and environment, eventually affecting the man himself. A high population growth rate combined with unplanned urban area has resulted environs pressure on our invaluable land and water resource and posing severe threat on fertile land. This practice or haphazard growth of urban area is particularly seen in

developing countries like India. India is close to 7933 municipalities as per census of India 2011 and population of Lucknow city is around 29.00 Lakhs.

The applicability of GIS in suitability analyses is vast, and new methods of spatial analysis facilitate suitability assessment and land use allocation [3, 9]. However, quick growth and sprawl bring new needs and demands for planners and designers, including the consideration of new growth and new alternatives for land use [2]. According to, GIS can also be used to evaluate plan for Smart Growth Developments (SGDs) and Transit Oriented Developments (TODs). Many cities in India are suffering this condition; these undefined areas are increasing rapidly and occupied by large number of people [6, 12]. There is a rapid increase in pre-urban boundary which in return results in increase in urban area and decrease in agriculture land [12]. Due to such haphazard growth of city land is also wasted. The increase in urban growth is also responsible for land transformation.

2.1 Data Used and Methodology

Study Area

The Lucknow city and its environs is selected for study. An additional area measuring 10–12 kms radius from the present boundary have been taken into consideration for study which extends from 26° ‘41’ “11.12” N to 26° ‘56’ “59.05” N latitude and 80° ‘54’ “55.55” E to 80° 48’ 0.57” E longitude. The geographical area of Lucknow district is about 3,244 sq. km. and city area is approximate 800 sq. km with population around 36,81,416 lakhs by 1456 per sq. km.

Data Sources

To meet the set objectives of the study, Survey of India topographical (SOI) Map sheet no. 63B/13 and 63B/14 on 1:50,000 scale and satellite imageries of IRS-1C/1D LISS-III 23.5 m resolution data on 1:50,000 scale acquired in 2001–02 and IKONOS satellite’s 1 m resolution data of 2014 and LISS III data has been used for preparation of multi criterion layers i.e. landuse/landcover, geomorphology, road/transport-network and ground water table data was collected from state/Central Ground Water Department.

Method

Multi-criteria decision making is a process of determining complex decision problems accurately. It involves breaking of complex decision problems into more simpler and smaller parts and by determining each part of problem in a logical and systematic manner in order to produce relevant result [11].

The expert choice software was used for finding the relative weights. There were four important steps applied in this process

1. determining the suitable factors for site selection procedure
2. then assign the weights to all the parameters
3. generating various land suitability thematic maps for urban development
4. Finally determining the most suitable area for urban development.

AHP model has been used on for these criteria pertaining to geomorphology (Fig. 1), landuse/landcover (Fig. 2), accessibility of road/transport network (Fig. 3) and groundwater.

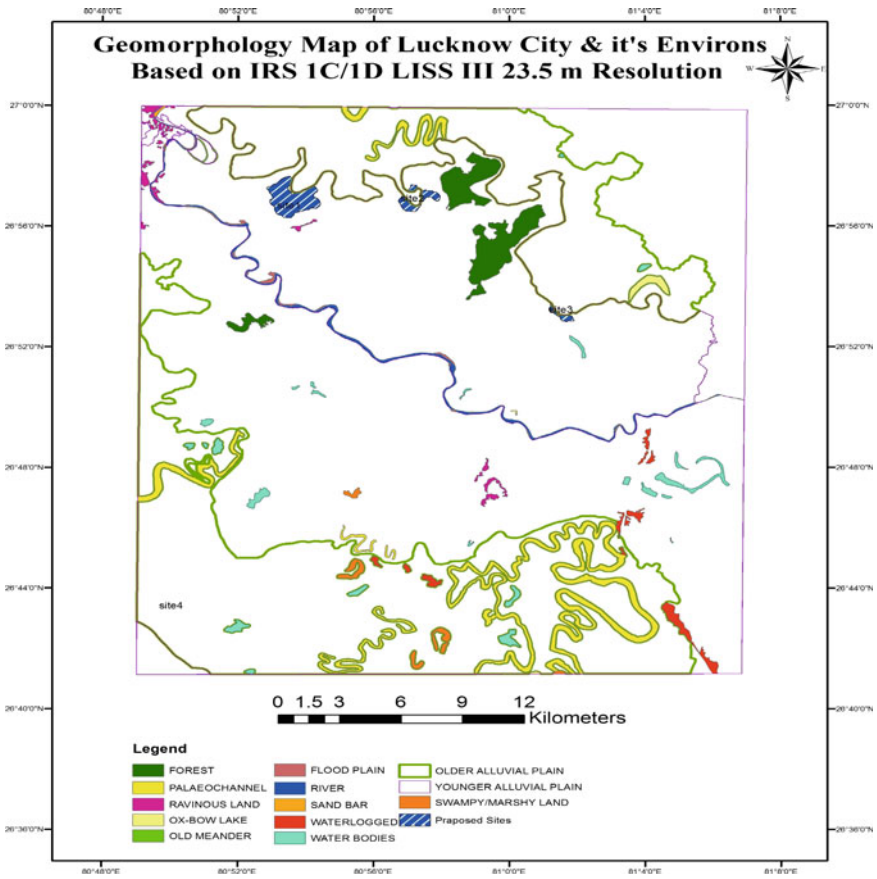


Fig. 1 Map showing different Geomorphology

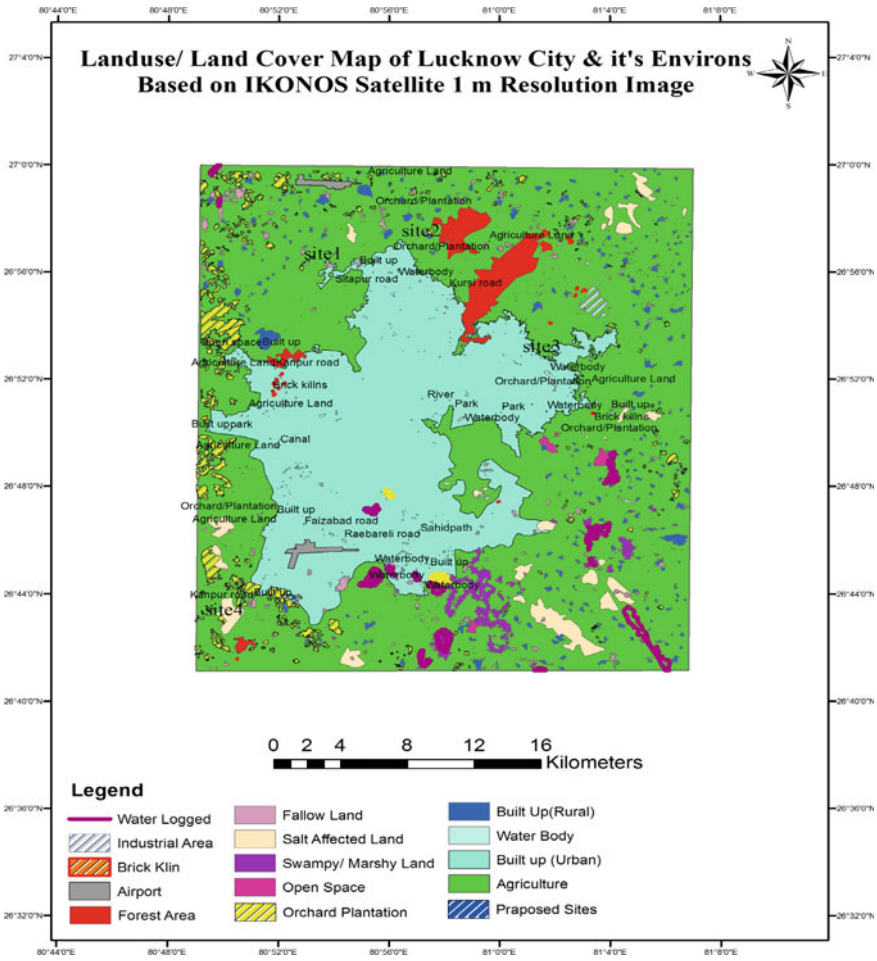


Fig. 2 Landuse/Land cover map of the study area

The steps involved in this process are described below:

A. Determining Suitable Factors

The suitable factors parameters are determined for urban development. In this study we have find four important parameters for urban development which are geomorphology, transport/road network, ground water table and land use/land cover.

Transport network is one of the important factors for determining urban development since accessibility to road for market, school, hospital etc. is necessary. Ground water table is also used as another parameter to determine the water table for each site, this performed with ground survey of each site. Type of land is

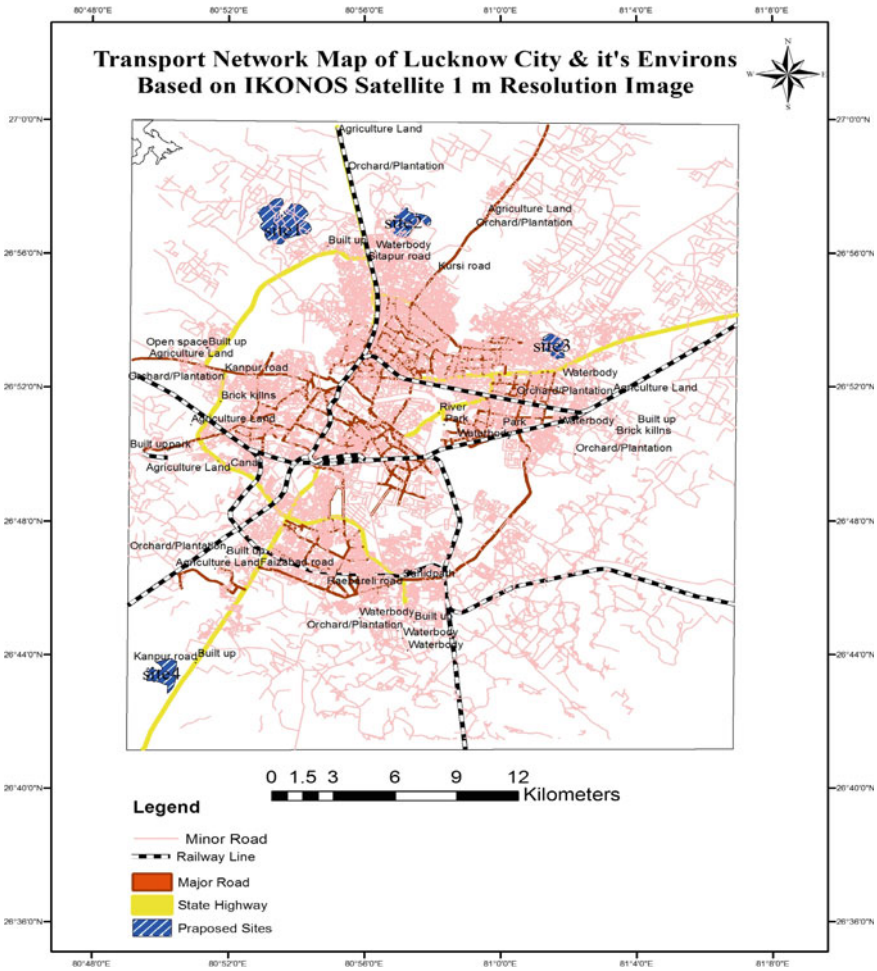


Fig. 3 Map showing Transport Network with help of IKONOS Image

determined by land used and land cover layer to protect the prime agriculture land from expansion to urban area, salt affected/waste lands are preferable for urban expansion [7].

The numbers of comparisons are calculated by using Eq. 1 as follows:

$$\frac{n(n-1)}{2} \tag{1}$$

where, n = number of things to compare (Table 1).

Table 1 This table shows the suitable factors

Selection crietria		Site-1	Site-2	Site-3	Site-4
Geomorphology		Younger alluvial plain	Younger alluvial plain	Older alluvial plain	Older Alluvial plain
Land-Use/Land-Cover		Open scrub	Open scrub	Open scrub	Salt affected land
Transport network	**RL**	2448 m (>500 m)	1390 m (>500 m)	2722 m (>500 m)	3677 m (>500 m)
	NH	1710 m (>1000 m)	1377 m (>1000 m)	486 m (<1000 m)	324 m (<1000 m)
	CMjR	5622 m (>1000 m)	1967 m (>1000 m)	845 m (<1000 m)	4123 m (>1000 m)
	CMnR	516 m (>30 m)	123 m (>30 m)	135 m (>30 m)	817 m (> 0 m)
Ground water table		130–150 Feet	80–100 Feet	50 Feet	85 Feet

B. Determining Relative Weight for each Criterion

After calculating number of comparisons the pair-wise comparison is performed, the selection criteria in one level relative to their significance to the goal of the study are made with help of square matrix. The diagonal elements of all square matrix equal to one or unit matrix. The principal Eigen value and the corresponding eigenvector (normalized Principal Eigen vector is also called **priority vector**) of the comparison matrix were calculated which gave the relative importance weights (RIW) of the criteria being compared [8].

After calculating weights consistency index (CI) and consistency ratio (CR) were calculated. In resulting, if the value of Consistency Ratio is greater than 0.1 the inconsistency is acceptable otherwise validity is again checked.

Index (CI) as deviation or degree of consistency using the Eq. 2 below:

$$CI = \frac{(\lambda_{\max} - n)}{(n-1)} \tag{2}$$

where, CI = Consistency Index, $\lambda_{\max}$ = Eigen vector, n = size of matrix.

Extracted the standard random consistency index (RI) value according to the size of the matrix from the Satty's random consistency index.

$$CR = \frac{CI}{RI} \tag{3}$$

where,

CR = Consistency Ratio, CI = Consistency Index, and RI = Random Consistency Index.

Urban Development Suitability Assessment

In this process overall composite weights calculated and then, the land suitability maps for urban development have been generated, based on the linear combination of each used factor's suitability score [14, 16] as shown in Eq. (4). AHP method applied to determine the importance of each parameter.

$$SI = \sum W_i X_i. \quad (4)$$

where, *S.I.* = Suitability Index of each alternative
W_i = RIW of particular selection criteria
W_i = RIWs of alternatives with respect to each criterion

3 Results and Discussions

Urban site selection model involves three steps to identify the most suitable alternatives for urban development which comprises as preliminary analysis, MCDM evaluation and identification of most suitable site. Preliminary analysis involves creation of various criterion maps into input the raster data layers.

3.1 Geomorphological Map

The geomorphological map of the study area has been obtained from available report and maps of Land-Use and Urban Survey Division, UP-RSAC. On the basis of regional geomorphology classification as shown in previous study. Two classes namely older alluvial plain and younger alluvial plain for the study area have been grouped and mapped as shown in Fig. 1. It is observed that two sites out of four sites namely Site-1 and Site-2 are located in identical class (younger alluvial plain), while Site-3 and Site-4 are located in older alluvial plain. Site-3 and Site-4 are considered as suitable to siting an urban development (Table 2).

3.2 Landuse/Landcover Map

Land-use/Land-cover map has been prepared using IKONOS data. Fourteen different land use classes have been identified in the study area namely agriculture land, built-up land, industrial area, brick-kiln, fallow land, open scrub, orchard and plantation, forest, built up rural, waste land/sodic land, river/drain, water bodies, water-logged area and other as shown in Fig. 2. Waste lands are degraded lands include sodic lands (salt affected land), scrub lands, which seem to be more suitable

Table 2 This table shows the relative weights for geomorphology

	Site-1	Site-2	Site-3	Site-4	RW
Site-1	*1*	1	1/4	1/4	**0.1**
Site-2	1	*1*	1/4	1/4	**0.1**
Site-3	4	4	*1*	1	**0.4**
Site-4	4	4	1	*1*	**0.4**

Table 3 This table shows the Relative Weights for Land use/Land Cover

	Site-1	Site-2	Site-3	Site-4	RW
Site-1	*1*	1	1	1/4	**0.136**
Site-2	1	*1*	1/4	1/5	**0.093**
Site-3	1	4	*1*	1/4	**0.203**
Site-4	4	5	4	*1*	**0.567**

for urban site because these land lack appropriate soil moisture, minerals etc. qualities of fertile land. However scrub lands are often appear like fallow land and look like crop land which could be discriminated but in essence these are also categorize as waste land and considered as moderately suitable for urban development. Consequently, it is noticed that Site-3 and Site-4 are more suitable area for urban development, while Site-1 and Site-2 are seen to be moderately suitable (Table 3).

3.3 Transport Network Map

The transportation network map has been digitized from the 1:25,000/1:50,000 scale topographical maps and IKONOS Satellite are grouped into five classes including railway line, national highway (NH), city major roads (CMjR) and city minor roads (CMnR) as shown in Fig. 3. According to [15], distance less than 500 km from NH, SH and CMjR should be avoided. The urban site should locate near to existing road networks to avoid the high cost of construction. Using Fig. 3, distances of different sites from existing transportation network have been measured and (Table 3) for proximity of transport network with respect to each site has been prepared. It is seen that Site-3 and Site-4 are suitable because these are near to NH, CMjR, and CMnR, whereas, Site-1 and Site-2 are unsuitable sites having distances beyond the required distance from NH, CMjR and CMnR (Table 4).

3.4 Water Table

The ground that water data was collected with help of field survey of various Sites from which it was found that Site-1, Site-2, Site-3 is having water at 130–150 ft, 80–100, 50 respectively and Site-4 at 85 ft (Tables 5 and 6).

Table 4 This table shows the relative weights for transport network

	Site-1	Site-2	Site-3	Site-4	RW
Site-1	**1**	1/3	1/3	1/3	**0.092**
Site-2	3	**1**	1/4	1/3	**0.152**
Site-3	3	4	**1**	1/2	**0.332**
Site 4	3	3	2	**1**	**0.424**

Table 5 This table shows the water table of various sites

Sl. no.	Site	Water table (in ft.)
2	Site-1	130 to 150
3	Site-2	80 to 100
4	Site-3	50
6	Site-4	85

Table 6 This table shows the relative weights for ground water

	Site-1	Site-2	Site-3	Site-4	RW
Site-1	*1*	1/3	1/4	1/2	**0.078**
Site-2	3	*1*	1/4	1/3	**0.147**
Site-3	4	4	*1*	2	**0.406**
Site-4	2	3	1/2	*1*	**0.370**

4 Results and Discussion

In this study all criteria were analyzed through literature review. Four different parameters were identified to select a suitable site for urban development. All four parameters thematic map were prepared as input map layers. These criteria include geomorphology, land-use/land-cover, transport network and ground water table. After the preparation of output criterion maps, four possible sites have been identified for urban development in the study area according to the data availability include Mubarakpur (Site-1), Rasoolpur Kayastha (Site-2), Chinhat (Site-3) and Aorava (Site-4). These sites are identified on the basis of flexibility of different criterion features. These four possible sites for urban development have been confirmed after the field investigation according to the urban development site suitability prerequisite (Figure 4 and Table 7).

The pair-wise comparison for determination of weights is more suitable than direct assignment of the weights, because one can check the consistency of the weights by calculating the consistency ratio in pair-wise comparison; however, in direct assignment of weights, the weights are depending on the preference of

Fig. 4 Overall best land suitability graph

Table 7 This table shows overall suitability index

W_i W_i					X_i	$SI = \sum W_i X_i$
	GEOM	LU/LC	ROAD	GWT		
Site-1	0.078	0.136	0.092	0.078	0.082	**0.094**
Site-2	0.147	0.093	0.152	0.147	0.198	**0.136**
Site-3	0.406	0.203	0.332	0.406	0.346	**0.348**
Site-4	0.37	0.567	0.424	0.37	0.374	**0.422**

decision maker [11, 17]. All the selected sites in the study area are ranked in terms of relative weight according to their suitability for urban development. The suitability index has the value 0.442 for Site-4 (Aorava) and 0.346 for Site-3 (Near Chinhat). Thus, Site-4 (Aorava) and Site-3 (Near Chinhat) having higher suitability index are identified as potential urban development sites in the study area. Site-1 and site-2 are found to have lowest suitability.

References

1. A G-O YEH (1999), Urban planning and GIS, Geographical information system, Second edition, Volume2 Management issues and applications, 62, pp 877–888.
2. Brueckner J K (2000) Urban Sprawl: Diagnosis and Remedies, International Regional Science Review 23(2): 160–171.
3. Brueckner, Jan K. "Strategic interaction among governments: An overview of empirical studies." *International regional science review* 26.2 (2003): 175–188.
4. Dutta, Venkatesh. "War on the Dream–How Land use Dynamics and Peri-urban Growth Characteristics of a Sprawling City Devour the Master Plan and Urban Suitability?." *13th Annual Global Development Conference, Budapest, Hungary*. 2012.
5. FAO, 1976. A framework for land evaluation. Food and Agriculture Organization of the United Nations, Soils Bulletin No. 32, FAO: Rome.
6. Kayser B (1990) La Renaissance rurale. Sociologie des campagnes du monde occidental, Paris: Armand Colin.
7. Malczewski, J., 1997. Propogation of errors in multicriteria location analysis: a case study, In: fandel, G., Gal, T. (eds.) Multiple Criteria Decision Making, Springer-Verlag, Berlin, 154–155.
8. Malczewski, J., 1999. GIS and Multicriteria Decision Analysis, John Wiley & Sons, Canada, 392 p.
9. Merugu Suresh, Arun Kumar Rai, Kamal Jain, (2015), Subpixel level arrangement of spatial dependences to improve classification accuracy, IV International Conference on Advances in Computing, Communications and Informatics (ICACCl), 978–1-4799-8792-4/15/$31.00 ©2015 IEEE, pp. 779–876.
10. Merugu Suresh, Kamal Jain, 2015, "Semantic Driven Automated Image Processing using the Concept of Colorimetry", Second International Symposium on Computer Vision and the Internet (VisionNet'15), Procedia Computer Science 58 (2015) 453–460, Elsevier.
11. Merugu Suresh, Kamal Jain, 2014, "A Review of Some Information Extraction Methods, Techniques and their Limitations for Hyperspectral Dataset" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 3, March 2014, ISSN: 2278–1323, pp 2394–2400.

12. McGregor D, Simon D and Thompson D (2005) The Peri-Urban Interface: Approaches to Sustainable Natural and Human Resource Use (eds.), Royal Holloway, University of London, UK, 272.
13. Myers, Adrian. "Camp Delta, Google Earth and the ethics of remote sensing in archaeology." *World Archaeology* 42.3 (2010): 455–467.
14. Mieszkowski P and E S Mills (1993) The causes of metropolitan suburbanization, Journal of Economic Perspectives 7: 135–47.
15. Mu, Yao. "Developing a suitability index for residential land use: A case study in Dianchi Drainage Area." (2006).
16. Saaty, T.L., 1990, How to make a decision: The Analytic Hierarchy Process. European Journal of Operational Research, 48, 9–26.
17. Saaty, T.L. 1994. Highlights and Critical Points in the Theory and Application of the Analytic Hierarchy Process, European Journal of Operational Research, 74: 426–447.

A Hierarchical Shot Boundary Detection Algorithm Using Global and Local Features

Manisha Verma and Balasubramanian Raman

Abstract A video is considered as high dimensional data which is tedious to process. Shot detection and key frame selection are activities to reduce redundant data from a video and make it presentable in few images. Researchers have worked in this area diligently. Basic shot detection schemes provide shot boundaries in a video sequence and key frames are selected based on each shot. Usually in video clips, shots repeat after one another, in that case the basic shot detection scheme gives redundant key frames from same video. In this work, we have proposed a hierarchical shot detection and key frames selection scheme which reduce a considerable amount of redundant key frames. For temporal analysis and abrupt transformation detection, color histogram has been used. After shot detection, spatial analysis has been done using local features. Local binary patterns have been utilized for local feature extraction. The proposed scheme is applied to three video sequences of news video, movie clip and tv-advertisement video.

Keywords Shot boundary · Keyframe · Local binary pattern · Temporal analysis · Spatial analysis

1 Introduction

A video is a collection of images with some temporal relation in between sequential images. A video scene is made of some shots and shots contains similar images. Keyframe is a frame which is assumed to be contained most of the information of a shot. Keyframe may be one or more according to the requirement of the system. Shot detection and key frame selection are the initial stages of a video retrieval model

M. Verma (✉)
Mathematics Department, IIT Roorkee, Roorkee, India
e-mail: manisha.verma.in@ieee.org

B. Raman
Computer Science and Engineering Department, IIT Roorkee, Roorkee, India
e-mail: balarfma@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_35

system. It is near impossible to process a video for retrieval or analysis task, without key frame detection. Key frame detection appears to reduce a large amount of data from video that makes it easy for further process.

A video shot transition happens in two ways, i.e., abrupt and gradual transition. The abrupt transition happens because of short cuts and gradual transition includes shot dissolve and fades. Many algorithms have been proposed to detect abrupt and gradual shot transition in video sequence [2]. A hierarchical shot detection algorithm was proposed using abrupt transitions and gradual transitions in different stages [3]. Wolf and Yu presented a method for hierarchical shot detection based on different shot transition analysis and used multi-resolution analysis. They used a hierarchical approach to detect different shot transitions, e.g., cut, dissolve, wipe-in, wipe-out, etc. [12]. Local and global feature descriptors have been used for feature extraction in shot boundary detection. Apostolidis et al. used local Surf features and global HSV color histograms for gradual and abrupt transitions for a shot segmentation [1].

In image analysis, only spatial information is required to extract. However, for a video study, temporal information should be recognized with spatial information. Temporal information defines the activity and transition of a frame to another frame. Rui et. al. proposed a keyframe detection algorithm using color histogram and activity measure. Spatial information was analyzed using color histogram and activity measure is used for temporal information detection. Similar shots are grouped later for better segmentation [9]. A two stage video segmentation technique was proposed using a sliding window. A segment of frame is used to detect shot boundary in first stage, and in second stage, the 2-D segments are propagated across the window of frames in both spatial and temporal direction [8]. Tippaya et al. proposed a shot detection algorithm using RGB histogram and edge change ratio, and three different dissimilarity measures have been used to extract difference between frame feature vectors [10].

Event detection and video content analysis have been done based on shot detection and keyframe selection algorithms [4]. Similar scene detection has been done using clustering approach. Story line has been made from a long video [11]. Event detection in sports video has been analyzed using long, medium and close-up shots, and play breaks are extracted for summarization of a video [5]. A shot detection technique has been implemented based on visual and audio content in video. Wavelet transformation domain has been utilized for feature extraction [6].

1.1 Main Contributions

In a video, many shots may be similar in visualization. In a conversation video or in general, shots usually repeat after one or more shots. Usually, shot boundary detection algorithms classify all shots in different clusters irrespective of redundant shots. In the proposed method, the authors have developed a hierarchical shot detection algorithm in two stages. First stage extracts temporal information of video and detect the initial shot boundary and extract the keyframes based on each shot. In the second

stage, spatial information of extracted key frames from first stage are analyzed, and redundant keyframes are excluded.

The paper has been organized in following way: In Sect. 1, a brief introduction of video shots and keyframes have been given with literature survey. Section 2 describes technique of the proposed method. In Sect. 3, framework of proposed algorithm is described. Section 4 represents the experimental results. Finally, the work has been concluded in Sect. 5.

2 Hierarchical Clustering for Shot Detection and Key Frame Selection

Shot detection problem is very common in video processing. Processing a full video at a time and extracting shot boundary may give results of similar shots. Frames of a video are shown in Fig. 1. Ten different shots are there in the video, in which 3, 5 and 7, and 4, 6 and 8 are of similar kind. Hence, keyframes extracted from these shots would be similar, and redundant information will be extracted from the video. It is a small example and it can happen in large video. To resolve this problem, a hierarchical scheme has been adopted for keyframe extraction from a video.

For abrupt shot boundary detection, we have used RGB color histogram. RGB color histogram provides global distribution of three color bands in RGB space. A quantized histogram of 8 bins for each color channel has been created. Initially, each color channel has been quantized in 8 intensities and histogram has been generated using the following equation.

$$Hist_C(L) = \sum_{a=1}^{m} \sum_{b=1}^{n} F(I(a,b,C),L) \quad (1)$$

where $C = 1, 2, 3$ for R, G, B color bands

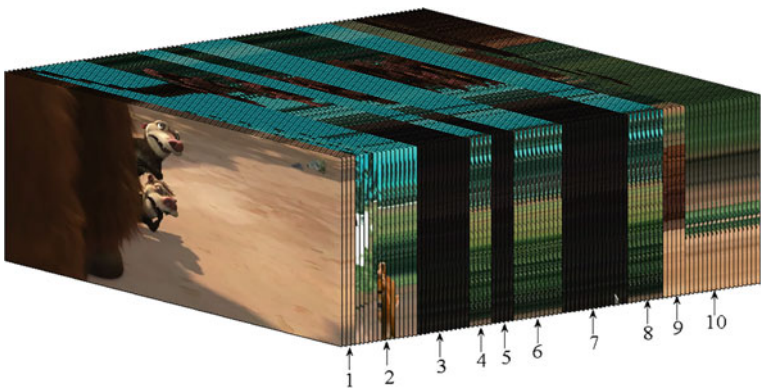


Fig. 1 Consecutive frames and shot boundary of a video

$$F(a,b) = \begin{cases} 1 \text{ if}, a = b \\ 0 \quad \text{else}. \end{cases} \tag{2}$$

where size of the image is $m \times n$ and L is total bins. $I(a, b, C)$ is the intensity of color channel C at position of (a, b).

For temporal information in a video sequence, each frame of video has been extracted and RGB color histogram has been generated. Difference of each frame to the next frame is extracted using the following distance measure.

$$Dis(DB^n, Q) = \sum_{s=1}^{L} \left| F_{db}^n(s) - F_q(s) \right| \tag{3}$$

If the measured distance between two frames is greater than a fixed threshold value, then those frames are separated in different clusters. This process is applied for each consecutive pair of frames in video sequence. In this process, we get different clusters of similar frames. After getting clusters, we extract one key frame from each cluster. For keyframe extraction, entropy has been calculated for each frame in one cluster using Eq. 4, and the maximum entropy frame has been chosen as a keyframe for that cluster.

$$Ent(I) = -\sum_{i} (p_i \times log_2(p_i)) \tag{4}$$

where p is the histogram for the intensity image I.

During this process, consecutive keyframes will not hold the similar information. However, except consecutive positions, two or more non-consecutive clusters may contain similar types of frames as a video sequence may hold similar shots at non-consecutive positions. Due to this, many keyframes may hold redundant information. To overcome this issue, hierarchical process is adopted in this work. Local binary pattern (LBP) is a well-known texture feature descriptor [7]. It computes relation of each pixel with neighboring pixels in the following manner:

$$LBP_{p,r} = \sum_{l=0}^{p-1} 2^l \times S_1(I_l - I_c) \tag{5}$$

$$S_1(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{else} \end{cases}$$

$$Hist(L) \mid_{LBP} = \sum_{a=1}^{m} \sum_{b=1}^{n} F(LBP(a,b), L); \tag{6}$$

$$L \in [0, (2^p - 1)]$$

where p and r represent number of neighboring pixels and radius respectively. After calculating the LBP pattern using Eq. 5, histogram of LBP pattern is created using

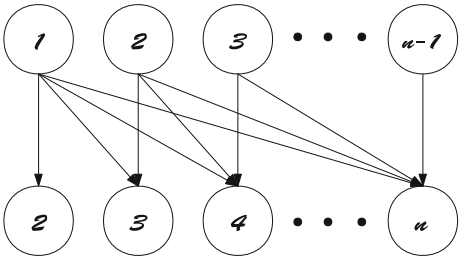


Fig. 2 Distance measure calculation in 2nd phase

Eq. 6. LBP is extracted from each of the keyframe obtained from the above process. Now, the distance between each frame mutually is calculated using Eq. 3 as shown in Fig. 2. Distance of frame 1 has been calculated with frame 2, 3 up to n. Distance of frame 2 has been calculated with frame 3, 4 up to n. In a similar process, distance of frame $n-1$ has been calculate with frame n. A tri-diagonal matrix has been created for all distance measures. Now, if the distance between two or more frames is less than a fixed threshold, then all those frames will be grouped into one cluster. In this process, even non-consecutive similar keyframes have been clustered, and completely non-redundant data in different clusters have been obtained. Again, the entropy of each of the frames in different cluster is calculated and maximum entropy frame is obtained as final keyframe. Finally, we get a reduced number of final key frames without any redundant information.

3 Proposed System Framework

3.1 Algorithm

Phase 1:

Input : Video clip
Output: Initial key frames

Upload video and extract all frames.
for i=1: n_1
Calculate the RGB histogram of frame i and i+1.
Calculate Dist(i,i+1).
If(Dist(i,i+1)> Th_1)
put i and i+1 in different clusters.
end
end
Calculate the entropy of each frame in different clusters.
Select maximum entropy frame from each cluster as a keyframe.

n_1=total number of frames in video

Phase 2:
Input : Initial key frames
Output: Selected final key frames

Load all keyframes extracted from Phase 1 and calculate LBP histogram for all.
Compute distance as explained in Fig. 2 and make a distance matrix 'D' for all frames.
Initialize a zero vector key_array of size n_2.
for i=1: n_2
  If(key_array(i))=0)
    assign key_array(i)=1
    Initialize a Stack 'S' of size n_2 *and push first element i.*
    while(S is not empty)
      t_1 *= pop an element from S*
      Check if the distances between t_1 *and other frames are less than* Th_2 *then put them in one cluster* t_2.
      Push all the elements of t_2 *in the Stack 'S'.*
    end
    Delete redundant frames from cluster if there is any.
  else
    continue
  end
end
Calculate the entropy of each frame in different clusters.
Select maximum entropy frame from each cluster as a keyframe.

n_2= number of keyframes extracted from Phase 1.

4 Experimental Results

For experimental purpose, we have used three different kinds of videos of news, advertisement and movie clip. General details about all three videos are given in Table 1. All three videos are of different size with respect to time and frame size.

Table 1 Video details

Video	Time (min.)	Frame size	Frame/sec
News video	02:55	1280×720	30
Movie clip	00:30	720×384	29
Advertisement	01:00	1920×1080	25

Fig. 3 Video 1: **a** Initial stage keyframes **b** final stage keyframes

In news video, anchor and guest are present at first. The camera is moving from anchor to guest and guest to anchor many times in the video. Hence, in shot detection, many shots are having similar kind of frames (either anchor or guest). Further, other events are shown in the video repeatedly one after another shot. All these redundant shots are separated initially and key frames are selected. In the second phase of algorithm, redundant key frames are clustered and keyframes of maximum entropy are extracted as final key frames. Initially, 63 key frames are extracted and after applying hierarchical process only 12 key frames are extracted at the end. This hierarchical process has removed a significant amount of redundant key frames for further processing.

Second video clip for the experiment is a small movie clip of a animation movie called 'Ice age'. The same hierarchical process is applied to the video clip. Initially, 11 key frames are extracted, and then by using LBP for spatial information, 6 final non-redundant key frames are extracted. In Fig. 3 keyframes of initial and final phase have been demonstrated.

The third video is taken for experiment is of a Tata-sky advertisement. The proposed method is applied to the video and two phase keyframes are collected. The keyframes of phase one and two are shown in Fig. 4. It is clearly visible that using hierarchical method the number of key frames has been reduced significantly and redundant key frames have been removed. Information regarding extracted key

Fig. 4 Video 1: **a** Initial stage keyframes **b** final stage keyframes

Table 2 Number of keyframes extracted in both phases

Video	Keyframes in phase 1	Keyframes in phase 2
News video	63	12
Movie clip	11	6
Advertisement	34	11

frames in phase one and two are given in Table 2. Summary of reduced keyframes explains that the proposed algorithm has removed repeated frames from keyframe detected from the color histogram method. Further, in phase two using LBP we have obtained optimum amount of key frames which summarize the video significantly.

Extracted keyframes can be saved as a database and used for video retrieval task. Keyframe extraction is an offline process and retrieving video using keyframes is a realtime online process. The proposed method of keyframe extraction is a two step process hence it is time consuming. However, it is an offline process and can be used to collect the keyframe database in one time. Number of keyframes has been reduced in the proposed method and those can be used in realtime video retrieval process. Since there are less number of keyframes hence the video retrieval will be less time consuming.

5 Conclusions

In the proposed work, shot boundary detection problem has been discussed and further, key frames have been obtained. A hierarchical approach is adopted for final keyframes selection. This approach helped in reducing similar keyframes in non-consecutive shots. Initially, a color histogram technique is used for temporal analysis and abrupt transition is obtained. Based on abrupt transition, shots are separated and keyframes are selected. Spatial analysis has been done in obtained keyframes using local binary pattern and finally redundant keyframes are removed. In this process, a significant amount of redundant keyframes are removed. The proposed method is applied on three videos of news reading, movie clip and tv advertisement for experiment. Experiments show that the proposed algorithm helped in removing redundant keyframes.

References

1. Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. pp. 6583–6587. IEEE (2014)
2. Brunelli, R., Mich, O., Modena, C.M.: A survey on the automatic indexing of video data. Journal of visual communication and image representation 10(2), 78–112 (1999)
3. Camara-Chavez, G., Precioso, F., Cord, M., Phillip-Foliguet, S., de A Araujo, A.: Shot boundary detection by a hierarchical supervised approach. In: Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on. pp. 197–200. IEEE (2007)
4. Cotsaces, C., Nikolaidis, N., Pitas, I.: Video shot detection and condensed representation. a review. Signal Processing Magazine, IEEE 23(2), 28–37 (2006)
5. Ekin, A., et al.: Generic play-break event detection for summarization and hierarchical sports video analysis. In: Multimedia and Expo (ICME), International Conference on. vol. 1, pp. I–169–172. IEEE (2003)
6. Nam, J., Tewfik, A.H., et al.: Speaker identification and video analysis for hierarchical video shot classification. In: Image Processing, International Conference on. vol. 2, pp. 550–553. IEEE (1997)
7. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24(7), 971–987 (2002)
8. Piramanayagam, S., Saber, E., Cahill, N.D., Messinger, D.: Shot boundary detection and label propagation for spatio-temporal video segmentation. In: IS&T/SPIE Electronic Imaging. vol. 9405, 94050D. International Society for Optics and Photonics (2015)
9. Rui, Y., Huang, T.S., Mehrotra, S.: Exploring video structure beyond the shots. In: Multimedia Computing and Systems, IEEE International Conference on. pp. 237–240. IEEE (1998)
10. Tippaya, S., Sitjongsataporn, S., Tan, T., Chamnongthai, K.: Abrupt shot boundary detection based on averaged two-dependence estimators learning. In: Communications and Information Technologies (ISCIT), 14th International Symposium on. pp. 522–526. IEEE (2014)
11. Yeung, M., Yeo, B.L., Liu, B.: Extracting story units from long programs for video browsing and navigation. In: Multimedia Computing and Systems, 3rd IEEE International Conference on. pp. 296–305. IEEE (1996)
12. Yu, H.H., Wolf, W.: A hierarchical multiresolution video shot transition detection scheme. Computer Vision and Image Understanding 75(1), 196–213 (1999)

Analysis of Comparators for Binary Watermarks

Himanshu Agarwal, Balasubramanian Raman, Pradeep K. Atrey and Mohan Kankanhalli

Abstract Comparator is one of key components of watermarking system that determines its performance. However, analysis and development of comparator is an undermined objective in the field of watermarking. In this paper, the core contribution is that five comparators for binary watermarks are analysed by theory and experiments. In the analysis, it is explored that negative pair of binary watermarks provide same information. Receiver operating characteristic curve is used for experimental analysis. It is observed that comparators based on similarity measure functions of symmetric normalized Hamming similarity (SNHS) and absolute mean subtracted normalized correlation coefficient (AMSNCC) have outstanding performance. Further, a range of threshold of SNHS based comparator that maximizes decision accuracy of a watermarking system is found by theoretical analysis. This range is verified by experiments.

Keywords Comparator · Threshold · Watermarking · Binary watermarks · Receiver operating characteristic curve

H. Agarwal (✉)
Department of Mathematics, Jaypee Institute of Information Technology, Sector 62, Noida, Uttar Pradesh, India
e-mail: himanshu203@gmail.com

B. Raman
Indian Institute of Technology Roorkee, Roorkee, India
e-mail: balarfma@iitr.ac.in

P.K. Atrey
State University of New York, Albany, NY, USA
e-mail: patrey@albany.edu

M. Kankanhalli
National University of Singapore, Singapore, Singapore
e-mail: mohan@comp.nus.edu.sg

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_36

1 Introduction

Digital watermarking is a technique that hides a watermark into multimedia to provide digital rights solutions [1–6, 8, 10–12, 15, 16]. Watermark refers to some important information, which can be a logo or a signature (binary/gray scale image). Multimedia refers to any image, video or audio. When watermark inserted media (watermarked media) is available in public domain, digital right(s) can be claimed using a watermark. Digital right claim is solved by comparing the watermark which is being presented and by extracting/detecting watermark from the watermarked media. Accurate decision rate of comparison of watermarks determines success of a watermarking system, which should be maximum.

Without loss of generality, a watermarking system consists of a watermarking scheme M which has two parts: a watermark embedding algorithm M_{emb} and a watermark extraction algorithm M_{ext}, set of host images H, set of watermarks X, comparator C and possible attack t on watermarked media.

This paper examines the effect of different comparators on watermarking system under the assumption that negative of a binary watermark must be treated same as itself. The origin of this assumption is fundamental of information theory, according to which, images of negative pair such as binary images of negative pair provide same information. Further, a formula is derived with theoretical analysis to compute a parameter of a comparator for maximum decision accuracy of watermarking system. Experiments are done on several watermarking systems to examine the different comparators and to verify the formula.

Remaining paper is described as follows. In Sect. 2, different comparators are defined. Receiver operating characteristic curve for performance evaluation of watermarking system is discussed in Sect. 3. In Sect. 4, theoretical analysis of a comparator is provided. Discussion on experiments, results and analysis is given in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Comparator

A comparator has two elements: a function, that measures level of similarity between two watermarks and a threshold. Two watermarks are said to be matched if the level of similarity is more than the threshold. Otherwise, watermarks are not matched. The mathematical formulation of a general comparator is as follows:

$$C_{(\tau,\ \mathrm{sim})}(x_1, x_2) = \begin{cases} \text{match} & \text{if } \mathrm{sim}(x_1, x_2) \geq \tau \\ \text{no match} & \text{otherwise} \end{cases}, \tag{1}$$

where, x_1 and x_2 are two watermarks, $\mathrm{sim}(x_1, x_2)$ is an arbitrary function that measures similarity between two watermarks and τ is a threshold value.

In watermarking, commonly used similarity measure functions are the normalized Hamming similarity (NHS) [1, 6, 11], the normalized correlation coefficient (NCC) [2, 4] and the mean subtracted NCC (MSNCC) [10]. In this paper, we have discussed two more functions: symmetric NHS (SNHS) and absolute MSNCC (AMSNCC). SNHS is derived from NHS and AMSNCC is derived from MSNCC. SNHS and AMSNCC are derived so that they treat negative of binary watermark same as itself. For two binary watermarks x_1 and x_2 of equal length of N, mathematical formula of these five functions are as follows:

$$\mathrm{NHS}(x_1, x_2) = 1 - \frac{1}{N}\sum_{i=1}^{N} x_1(i) \oplus x_2(i), \tag{2}$$

$\oplus$ is a bit-wise exclusive OR (XOR) operation;

$$\mathrm{SNHS}(x_1, x_2) = 0.5 + |\mathrm{NHS}(x_1, x_2) - 0.5|; \tag{3}$$

$$\mathrm{NCC}(x_1, x_2) = \frac{\sum_{i=1}^{N} x_1(i).x_2(i)}{\sqrt{\sum_{i=1}^{N} x_1(i)^2}\sqrt{\sum_{i=1}^{N} x_2(i)^2}}; \tag{4}$$

$$\mathrm{MSNCC}(x_1, x_2) = \frac{\sum_{i=1}^{N} (x_1(i) - \bar{x_1}).(x_2(i) - \bar{x_2})}{\sqrt{\sum_{i=1}^{N} (x_1(i) - \bar{x_1})^2}\sqrt{\sum_{i=1}^{N} (x_k(i) - \bar{x_1})^2}}, \tag{5}$$

$\bar{x_1}$ and $\bar{x_2}$ are mean (average) value of watermarks x_1 and x_2 respectively;

$$\mathrm{AMSNCC}(x_1, x_2) = \left|\mathrm{MSNCC}(x_1, x_2)\right|. \tag{6}$$

Respective range of the functions is [0, 1], [0.5, 1], [0, 1], [−1, 1] and [0, 1]. For same watermarks, value of each function is 1. For negative pair of watermarks, respective value of the functions is 0, 1, 0, −1 and 1. Moreover, two fundamental properties of NHS are as follows:

$$\mathrm{NHS}(x_1, x_2) + \mathrm{NHS}(x_2, x_3) \leq 1 + \mathrm{NHS}(x_1, x_3), \tag{7}$$

$$\mathrm{NHS}(x_1, x_2) + \mathrm{NHS}(x_2, x_3) \geq 1 - \mathrm{NHS}(x_1, x_3). \tag{8}$$

Note that NCC is failed if either of watermark is a pure black image and, MSNCC and AMSNCC are failed if either of watermark is an uniform intensity image.

3 Receiver Operating Characteristic Curve

We have evaluated the decision accuracy of a watermarking system by using receiver operating characteristic (ROC) curve [13, 15]. Decision is taken on the basis of matching result of an extracted watermark with a reference watermark. ROC curve is defined as two dimensional plot of false positive rate FPR (defined as ratio of total number of wrong matched case to the total number of matched case) versus false negative rate FNR (defined as ratio of total number of wrong non-matched case to the total number of non-matched case). FPR and FNR depend on each component of a watermarking system. The range of FPR and FNR is [0, 1]. (FPR, FNR) = (0, 0) is the *ideal point* on a ROC curve. Decision accuracy is maximum at the ideal point, that is, (i) *each extracted watermark is matched with correct watermark*, (ii) *each extracted watermark is uniquely matched*. We must tune the parameters of a watermarking system to obtain the ideal point. However, in practice, obtaining the ideal point is always not possible. In that case, we tune the parameters of the watermarking system to obtain an optimal point (a point on ROC curve nearest to the ideal point) on the ROC curve.

4 Theoretical Analysis of Comparator

Theoretical analysis of watermarking system is an important research issue [7, 9, 14, 17]. In this section, two conditions are discussed for obtaining ideal point of a watermarking system of SNHS based comparator. These conditions are used to develop a formula which helps to find range of threshold corresponding to ideal point. The analytic proof for the formula is also provided.

4.1 First Condition

The first condition for obtaining the ideal point is that set of watermarks (X) must not consist any pair of same or negative watermarks. This condition gives that maximum symmetric normalized Hamming similarity (MSNHS) of X must be strictly less than one. Mathematically,

$$\text{MSNHS}(X) = \max \left\{ \begin{array}{l} \text{SNHS}(x_i, x_j) : \\ x_i, x_j \in X, i \neq j \end{array} \right\} < 1. \tag{9}$$

By the definition of MSNHS(X), it is clear that for all two different watermarks x_i and x_j of X,

$$1 - \text{MSNHS(X)} \leq \text{NHS}\left(x_i, x_j\right) \leq \text{MSNHS(X)}. \tag{10}$$

4.2 Second Condition

The first condition is *necessary* for existence of second condition and second is *sufficient* for existence of the ideal point. The second condition is

$$\text{MSNHS} < 2P - 1. \tag{11}$$

where, the term P represents the minimum similarity of any embedded watermark with respect to the corresponding extracted watermark for a given watermarking system. Range of the P is [0.5, 1]. A method to compute P is provided in algorithm 1.

Algorithm 1: An algorithm to compute P
Input: $X, H, t, M = (M_{emb}, M_{ext})$ (for details, refer Sect. 1)
Output P

1. Create a matrix of size $|X| \times |H|$ with each element of 0. Store this matrix as SNHS=zeros($|X|, |H|$).
2. **for** $i = 1 : 1 : |X|$
3. **for** $j = 1 : 1 : |H|$
4. Select $x_i \in X, h_j \in H$
5. Embed watermark x_i in h_j by using the watermark embedding algorithm M_{emb} to obtain watermarked image $\hat{h}_{ij}$.
6. Apply noise/attack t on the $\hat{h}_{ij}$ to obtain noisy/attacked watermarked image $\hat{\hat{h}}_{ij}$ (for details of attack t, refer Tables 2 and 3).
7. Extract watermark $\hat{x}_{ij}$ from $\hat{\hat{h}}_{ij}$ by using the watermark extraction algorithm M_{ext}.
8. Compute $\text{SNHS}(i, j) = 0.5 + |\text{NHS}(\hat{x}_{ij}, x_i) - 0.5|$.
9. **end**
10. **end**
11. Find the minimum value of the matrix SNHS to obtain P.

Output P

4.3 Ideal Point Threshold Range and Its Proof

The ideal point threshold range is obtained from the (11) as

$$\frac{1 + \text{MSNHS}}{2} < \tau \leq P. \tag{12}$$

The formula (12) provides sufficient threshold range for ideal point, that is if threshold (τ) is in the range provided by the formula (12), then watermarking system is tuned at ideal point. However, converse need not to be true, that is if watermarking system is tuned at ideal point then threshold may or may not be in the range given by (12). The proof for "(12) *is sufficient*" is discussed.

Proof for correct match:

Let x be an extracted watermark corresponding to an original embedded watermark $x_i \in X$. According to the formula (12) and by definition of P (Sect. 4.2), we have,

$\text{SNHS}(x, x_i) \geq P$ *implies* $\text{SNHS}(x, x_i) \geq \tau$ *implies* x is matched with x_i.

Proof for unique match:

If possible, let us assume that an extracted watermark x is matched with two watermarks, x_i and x_j in X. Therefore, according to the definition of SNHS based comparator, we have

$$\text{SNHS}(x, x_i) \geq \tau \text{ and } \text{SNHS}(x, x_j) \geq \tau;$$

where τ satisfies (12). Thus, we have the following four possible cases.

case (i) $\text{NHS}(x, x_i) \geq \tau$ and $\text{NHS}(x, x_j) \geq \tau$.
case (ii) $\text{NHS}(x, x_i) \leq 1 - \tau$ and $\text{NHS}(x, x_j) \leq 1 - \tau$.
case (iii) $\text{NHS}(x, x_i) \geq \tau$ and $\text{NHS}(x, x_j) \leq 1 - \tau$.
case (iv) $\text{NHS}(x, x_i) \leq 1 - \tau$ and $\text{NHS}(x, x_j) \geq \tau$. Now, we will discuss one by one.

case (i):

$$\text{NHS}(x, x_i) \geq \tau \text{ and } \text{NHS}(x, x_j) \geq \tau \Rightarrow$$

$$\text{(by (12))}, \ \text{NHS}(x, x_i) \geq \tau > \tfrac{1+\text{MSNHS}}{2}$$
$$\text{and}$$
$$\text{(by (12))}, \ \text{NHS}(x, x_j) \geq \tau > \tfrac{1+\text{MSNHS}}{2} \Rightarrow$$

$$\text{NHS}(x, x_i) + \text{NHS}(x, x_j) > 1 + \text{MSNHS} \Rightarrow$$

$$\text{(by (10))}, \ \text{NHS}(x, x_i) + \text{NHS}(x, x_j) > 1 + \text{NHS}(x_i, x_j) \Rightarrow$$

a contradiction by a fundamental property of NHS (7).

case (ii):

$$\text{NHS}(x, x_i) \leq 1 - \tau \text{ and } \text{NHS}(x, x_j) \leq 1 - \tau \Rightarrow$$

$$\text{(by (12))}, \ \text{NHS}(x, x_i) \leq 1 - \tau < \tfrac{1-\text{MSNHS}}{2}$$
$$\text{and}$$
$$\text{(by (12))}, \ \text{NHS}(x, x_j) \leq 1 - \tau < \tfrac{1-\text{MSNHS}}{2} \Rightarrow$$

$$\text{NHS}(x, x_i) + \text{NHS}(x, x_j) < 1 - \text{MSNHS} \Rightarrow$$

$$\text{(by (10))}, \ \text{NHS}(x, x_i) + \text{NHS}(x, x_j) < 1 - \text{NHS}(x_i, x_j) \Rightarrow$$

Table 1 Explanation of data-sets used in experiments

Data set	Set of original images	Number of original images	Size of original images (pixels)	Set of watermarks	Number of watermarks	Size of watermarks (pixels)
D_1	H_1	6	256×256	X_1	10	128×128
D_2	H_2	10	256×256	X_2	10	128×128
D_3	$H_3 = H_2$	10	256×256	X_3	20	128×128
D_4	H_4	20	256×256	X_4	10	128×128
D_5	H_5	30	256×256	X_5	5	128×128
D'_1	H'_1	6	512×512	X'_1	10	64×64
D'_2	H'_2	10	512×512	X'_2	10	64×64
D'_3	$H'_3 = H'_2$	10	512×512	X'_3	20	64×64
D'_4	H'_4	20	512×512	X'_4	10	64×64
D'_5	H'_5	30	512×512	X'_5	5	64×64

a contradiction by a fundamental property of NHS (8).

case (iii):

$$\mathrm{NHS}(x, x_i) \geq \tau \text{ and } \mathrm{NHS}(x, x_j) \leq 1 - \tau \Rightarrow$$

$$\text{(by (12))}, \ \mathrm{NHS}(x, x_i) \geq \tau > \frac{1+\mathrm{MSNHS}}{2}$$
$$\text{and}$$
$$\text{(by (12))}, \ \mathrm{NHS}(x, x_j) \leq 1 - \tau < \frac{1-\mathrm{MSNHS}}{2} \Rightarrow$$

$$\mathrm{NHS}(x, x_i) - \mathrm{NHS}(x, x_j) > \mathrm{MSNHS} \Rightarrow$$

$$\text{(by (10))}, \ \mathrm{NHS}(x, x_i) - \mathrm{NHS}(x, x_j) > 1 - \mathrm{NHS}(x_i, x_j) \Rightarrow$$

$$\mathrm{NHS}(x, x_i) + \mathrm{NHS}(x_i, x_j) > 1 + \mathrm{NHS}(x, x_j) \Rightarrow$$

a contradiction by a fundamental property of NHS (7).

case (iv): This proof is similar to the proof for case (iii).

5 Experiments, Results and Analysis

Two main aims of the experiments are

- Compare the performance of different comparators in watermarking systems using ROC curve;

Table 2 Ideal point threshold range for various watermarking systems using ROC curve. Comparators are based on SNHS and AMSNCC. NGF: Negative + Gaussian Filter, NGN: Negative + Gaussian Noise, NR: Negative + Rotation, σ: variance, Q: Quality Factor, r: rotation in degree, counter clockwise, DNE: Does Not Exist

Watermarking scheme	Data-set	Attack (t)	Threshold range for	
			AMSNCC	SNHS
Wong and Memon [16]	D_1	No attack	[0.39, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_1	Negative attack	[0.39, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_2	No attack	[0.23, 1.00]	[0.68, 1.00]
Wong and Memon [16]	D_2	Negative attack	[0.23, 1.00]	[0.68, 1.00]
Wong and Memon [16]	D_3	No attack	[0.39, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_3	Negative attack	[0.39, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_4	No attack	[0.90, 1.00]	[0.96, 1.00]
Wong and Memon [16]	D_4	Negative attack	[0.90, 1.00]	[0.96, 1.00]
Wong and Memon [16]	D_5	No attack	[0.38, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_5	Negative attack	[0.38, 1.00]	[0.74, 1.00]
Wong and Memon [16]	D_1	NGF, $\sigma = 0.3$	[0.39, 0.97]	[0.74, 0.99]
Wong and Memon [16]	D_1	NGF, $\sigma = 0.7$	DNE	DNE
Wong and Memon [16]	D_1	NGN, $\sigma = 10^{-5}$	[0.28, 0.62]	[0.68, 0.84]
Wong and Memon [16]	D_1	JPEG, Q = 95	[0.15, 0.28]	[0.60, 0.66]
Wong and Memon [16]	D_1	NR, r = 0.2	[0.39, 0.99]	[0.74, 1.00]
Bhatnagar and Raman [4]	D'_1	No attack	[0.43, 1.00]	[0.73, 0.99]
Bhatnagar and Raman [4]	D'_2	No attack	[0.27, 0.97]	[0.69, 0.98]
Bhatnagar and Raman [4]	D'_3	No attack	[0.43, 0.97]	[0.73, 0.98]
Bhatnagar and Raman [4]	D'_4	No attack	[0.91, 0.93]	{0.96}
Bhatnagar and Raman [4]	D'_5	No attack	[0.43, 0.82]	[0.65, 0.91]
Bhatnagar and Raman [4]	D'_1	NGF, $\sigma = 0.3$	[0.42, 0.99]	[0.73, 0.99]
Bhatnagar and Raman [4]	D'_1	NGN, $\sigma = 10^{-4}$	[0.43, 0.99]	[0.73, 0.99]
Bhatnagar and Raman [4]	D'_1	JPEG, Q = 95	[0.43, 0.99]	[0.73, 0.99]
Bhatnagar and Raman [4]	D'_1	NR, r = 0.2	DNE	DNE

- Verify the analytically proved formula (12) with ROC curve for different watermarking systems.

Ten data-set and two watermarking schemes have been used in the experiments. Each data-set consists of a set of host images (original images) and a set of watermarks. Host images are eight bit gray scale images and watermarks are binary images. Further description of each data-set is provided in the Table 1. The data-set $D_1, D_2, \ldots, D_5$, are compatible with watermarking scheme of [16] and the data-set $D'_1, D'_2, \ldots, D'_5$ are compatible with watermarking scheme of [4]. Various attacks, such as Gaussian filter, Gaussian noise, JPEG compression, rotation, and negative operation have been applied on the watermarked images.

Table 3 Verification of formula (12) for various watermarking systems using ROC curve. Comparator is based on SNHS. NGF: Negative + Gaussian Filter, NGN: Negative + Gaussian Noise, NR: Negative + Rotation, σ: variance, Q: Quality Factor, r: rotation in degree, counter clockwise, DNE: Does Not Exist

Watermarking scheme	Data-set	Attack (t)	P	MSNHS	Ideal point threshold range		Verified
					ROC curve	Formula (12)	
Wong and Memon [16]	D_1	No attack	1	0.7347	[0.74, 1.00]	(0.86, 1.00]	Yes
Wong and Memon [16]	D_2	No attack	1	0.6799	[0.68, 1.00]	(0.83, 1.00]	Yes
Wong and Memon [16]	D_3	No attack	1	0.7347	[0.74, 1.00]	(0.86, 1.00]	Yes
Wong and Memon [16]	D_4	No attack	1	0.9501	[0.96, 1.00]	(0.97, 1.00]	Yes
Wong and Memon [16]	D_5	No attack	1	0.6371	[0.64, 1.00]	(0.81, 1.00]	Yes
Wong and Memon [16]	D_1	NGF, $\sigma = 0.3$	0.9904	0.7347	[0.74, 0.99]	(0.86, 0.99]	Yes
Wong and Memon [16]	D_1	NGN, $\sigma = 10^{-3}$	0.5	0.7347	DNE	DNE	Yes
Wong and Memon [16]	D_1	JPEG, Q = 95	0.6664	0.7347	[0.60, 0.66]	DNE	Yes
Wong and Memon [16]	D_1	NR, r = 0.2	1	0.7347	[0.74, 1.00]	(0.86, 1.00]	Yes
Bhatnagar and Raman [4]	D'_1	No attack	0.9963	0.7212	[0.73, 0.99]	(0.86, 0.99]	Yes
Bhatnagar and Raman [4]	D'_2	No attack	0.9880	0.6819	[0.69, 0.98]	(0.84, 0.98]	Yes
Bhatnagar and Raman [4]	D'_3	No attack	0.9880	0.7212	[0.73, 0.98]	(0.86, 0.98]	Yes
Bhatnagar and Raman [4]	D'_4	No attack	0.9673	0.9526	{0.96}	DNE	Yes
Bhatnagar and Raman [4]	D'_5	No attack	0.9133	0.6399	[0.65, 0.91]	(0.81, 0.91]	Yes
Bhatnagar and Raman [4]	D'_1	NGF, $\sigma = 0.3$	0.9963	0.7212	[0.73, 0.99]	(0.86, 0.99]	Yes
Bhatnagar and Raman [4]	D'_1	NGN, $\sigma = 10^{-4}$	0.9963	0.7212	[0.73, 0.99]	(0.86, 0.99]	Yes
Bhatnagar and Raman [4]	D'_1	JPEG, Q = 95	0.9954	0.7212	[0.73, 0.99]	(0.86, 0.99]	Yes
Bhatnagar and Raman [4]	D'_1	NR, r = 0.2	0.5039	0.7212	DNE	DNE	Yes

5.1 Comparison of Comparators

Performance of NHS, SNHS, NCC, MSNCC and AMSNCC based comparators are examined using ROC curve. One main observation is that if extracted watermark is not the negative of embedded watermark then all the comparators have same performance. However, if extracted watermark is a negative of embedded watermark then SNHS and AMSNCC based comparators have outstanding performance. Using the ROC curves, we have obtained the ideal point threshold range for various watermarking systems of SNHS and AMSNCC based comparators. Highlights of results are shown in Table 2. The important observations from the experiment are as follows:

- For Wong and Memon [16] scheme based watermarking systems, AMSNCC and SNHS based comparators are better than NHS, NCC and MSNCC based comparators against negative attack on watermarked images.
- For Bhatnagar and Raman [4] scheme based watermarking systems, all the comparators have same performance.
- The performance of AMSNCC and SNHS based comparators is very close.
- The length of ideal point threshold range decrease with increase in attack level and finally the range vanishes.
- The performance of the watermarking systems degrades with attack level.

5.2 Verification of Formula (12)

Analytic formula (12) for ideal point threshold range consists of two parameters, P and MSNHS. In the experiment, P and MSNHS are computed for several watermarking systems to find ideal point threshold range using the formula (12). This threshold range is compared with the ideal point threshold range obtained by ROC curve of corresponding watermarking system. Highlights of results are given in Table 3. The important observations from the experiments are as follows:

- The interval of ideal point threshold range obtained by the formula (12) is subinterval of threshold range obtained by ROC curve. This verifies the formula (12).
- If P is greater than $\frac{1+\text{MSNHS}}{2}$, then the ideal point threshold range obtained by ROC curve is [MSNHS, P].
- If P is less than $\frac{1+\text{MSNHS}}{2}$ then the upper bound of interval of ideal point threshold range obtained by ROC curve is P.

6 Conclusions

Choice of comparator affects performance of watermarking system. Experiments are done for watermarking schemes of Wong and Memon [16] and Bhatnagar and Raman [4]. Performance of watermarking systems of binary watermarks against negative

attack are the best when SNHS and AMSNCC based comparators are used. However, performance of Bhatnagar and Raman [4] scheme is independent of choice of comparator. Ideal point threshold range is found by using ROC curve for NHS, SNHS, NCC, MSNCC and AMSNCC based comparators. Further, the formula (12) is theoretically proved that finds ideal point threshold range for SNHS based comparator. This formula is verified for several watermarking systems and is computationally efficient than ROC curve to find ideal point threshold range. The formula (12) is sufficient, therefore, ideal point threshold range found by using (12) is sub-interval of the range found by ROC curve.

Acknowledgements The author, Himanshu Agarwal, acknowledges the grants of the University Grant Commission (UGC) of New Delhi, India under the JRF scheme and Canadian Bureau for International Education under the Canadian Commonwealth Scholarship Program. He also acknowledges research support of the Maharaja Agrasen Technical Education Society of India and Jaypee Institute of Information Technology of India.

References

1. Agarwal, H., Atrey, P. K. and Raman, B. Image watermarking in real oriented wavelet transform domain. *Multimedia Tools and Applications*, **74**(23):10883–10921, 2015.
2. Agarwal, H., Raman, B. and Venkat, I. Blind reliable invisible watermarking method in wavelet domain for face image watermark. *Multimedia Tools and Applications*, **74**(17):6897–6935, 2015.
3. Bender, W., Butera, W., Gruhl, D., et al. Applications for data hiding. *IBM Systems Journal*, **39**(3.4):547–568, 2000.
4. Bhatnagar, G. and Raman, B. A new robust reference watermarking scheme based on DWT-SVD. *Computer Standards & Interfaces*, **31**(5):1002–1013, 2009.
5. Cox, I. J., Kilian, J., Leighton, F. T. and Shamoon, T. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, **6**(12):1673–1687, 1997.
6. Kundur, D. and Hatzinakos, D. Digital watermarking for telltale tamper proofing and authentication. *Proceedings of the IEEE*, **87**(7):1167–1180, 1999.
7. Linnartz, J. P., Kalker, T. and Depovere, G. Modelling the false alarm and missed detection rate for electronic watermarks. In *Information Hiding*, pages 329–343, 1998.
8. Memon, N. and Wong, P. W. Protecting digital media content. *Communications of the ACM*, **41**(7):35–43, 1998.
9. Miller, M. L. and Bloom, J. A. Computing the probability of false watermark detection. In *Information Hiding*, pages 146–158, 2000.
10. Pandey, P., Kumar, S. and Singh, S. K. Rightful ownership through image adaptive DWT-SVD watermarking algorithm and perceptual tweaking. *Multimedia Tools and Applications*, **72**(1):723–748, 2014.
11. Rani, A., Raman, B., Kumar, S. A robust watermarking scheme exploiting balanced neural tree for rightful ownership protection. *Multimedia Tools and Applications*, **72**(3):2225–2248, 2014.
12. Rawat, S. and Raman, B. A blind watermarking algorithm based on fractional Fourier transform and visual cryptography. *Signal Processing*, **92**(6):1480–1491, 2012.
13. Tefas, A., Nikolaidis, A., Nikolaidis, N., et al. Statistical analysis of markov chaotic sequences for watermarking applications. In *IEEE International Symposium on Circuits and Systems*, number 2, pages 57–60, Sydney, NSW, 2001.

14. Tian, J., Bloom, J. A. and Baum, P. G. False positive analysis of correlation ratio watermark detection measure. In *IEEE International Conference on Multimedia and Expo*, pages 619–622, Beijing, China, 2007.
15. Vatsa, M., Singh, R. and Noore, A. Feature based RDWT watermarking for multimodal biometric system. *Image and Vision Computing*, **27**(3):293–304, 2009.
16. Wong, P. W. and Memon, N. Secret and public key image watermarking schemes for image authentication and ownership verification. *IEEE Transactions on Image Processing*, **10**(10):1593–1601, 2001.
17. Xiao, J. and Wang, Y. False negative and positive models of dither modulation watermarking. In *IEEE Fourth International Conference on Image and Graphics*, pages 318–323, Sichuan, 2007.

On Sphering the High Resolution Satellite Image Using Fixed Point Based ICA Approach

Pankaj Pratap Singh and R.D. Garg

Abstract On sphering the satellite data, classified images are achieved by many authors that had tried to reduce the mixing effect in image classes with the help of different Independent component analysis (ICA) based approaches. In these cases multispectral images are limited with small spectral variation in heterogeneous classes. For better classification, high spectral variance among different classes and low spectral variance within a particular class should exhibit. In the consideration of this issue, a Fixed point (FP) based Independent Component Analysis (ICA) method is utilized to get better classification accuracy in the existing mixed classes that consist similar spectral behavior. This FP-ICA method identifies the objects from mixed classes having similar spectral characteristics, on sphering high resolution satellite images (HRSI). It also helps to reduce the effect of similar spectral behavior between different image classes. The estimation of independent component related to non-gaussian distribution data (image) with optimizing the performance of this approach with the help of nonlinearity, which utilize the low variance between similar spectral classes. It is quite robust, effortless in computation and high convergence rate, even though the spectral distributions of satellite images are rigid to classify. Hence, this FP-ICA approach plays a key role in image classification such as buildings, grassland area, road, and vegetation.

Keywords Fixed point · Independent component analysis · Image classification · Mixed classes · Non-gaussianity · Negentropy

P.P. Singh (✉)
Department of Computer Science & Engineering, Central Institute of Technology Kokrajhar, BTAD, Kokrajhar, Assam, India
e-mail: pankajp.singh@cit.ac.in

R.D. Garg
Geomatics Engineering Group, Department of Civil Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
e-mail: garg_fce@iitr.ernet.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_37

1 Introduction

In the current scenario, rapid changes are found in environment, but the better classification is still a challenging task to provide the distinct image information in the respective of image classes in automated manner. One of the causes can be little spectral variation among the different land classes. In this regard, to achieve this challenge, the existing classification approaches stress to identify the accurate class information from existing mixed classes in high resolution satellite image (HRSI). While, this intricate problem is being resolved in different manner up to some level only, due to the occurrence of mixed classes deceive less accurate classification. In such scenario, the existence of edge information is reasonable useful to segregate the image objects.

Since, the conservation of image information and diminution of noise are not two interdependent parts in image processing. Hence, a single technique is not quite valuable for other applications. On the basis of spectral resolution of HRSI and the field of restoration, many premises were recommended for different noise reduction methods. The quasi-newton methods and related higher order cumulants are quite approachable to find the saddle points in these theories for segregating the spectral similar classes. Such approaches have played an important role to reduce mixed classes problem in the classification. The state-space based approach were utilized to resolve the difficulty occurs in blind source separation and thereafter, the deconvolution method was included to improve the existing classes [1–4].

Mohammadzadeh et al. [5] utilized particle swarm optimization (PSO) to optimize a proposed fuzzy based mean estimation method. It helps to evaluate better mean value for road detection in specific spectral band, due to improvement in fuzzy cost function by PSO. Singh and Garg [6] calculated threshold value in adaptive manner for extracting road network on considering the illumination factor. In addition, artificial intelligence (AI) related methods are used in image-processing techniques to classify the buildings in automatic way [7, 8].

Li et al. [9] designed a snake model for automatic building extraction by utilizing the region growing techniques and also mutual information. Singh and Garg [10] developed a hybrid technique for classification of HRSI which is comprised of a nonlinear derivative method and improved watershed transforms to extract an impervious surface (buildings and roads) from different urban areas.

The problem of mixed classes arises due to zero mean and unit variance among them that occurs availability of the gaussian random variable (spectral behavior of pixels). However, such limitation is an another issue to strength such kind of approximations, in particular, due to the non-lustiness meet with negentropy. To decrease the mutual information among (image classes) variables, negentropy is increased. This reflects the better autonomy among mixed image classes that shows the measurement of the non-gaussianity of independent components by negentropy. The better value of non-gaussianity illustrates the separated classes in images [11, 12]. Hyyarinen [13] provided a new approximation to resolve the issues related to

preceding approximations of negentropy. The principle based on maximum-entropy is used as key criteria of such approximations.

2 A Designed Framework for Image Classification Using FP-ICA Approach

The proposed framework explains fixed point based independent component analysis (FP-ICA) technique for image classification by using HRSI. FP-ICA is an ordered blind extraction algorithm that helps to extract independent components in non-Gaussian distributed data sources sequentially. At first this method whitens the monitored image pixels (variables), which act as a preprocessing scheme in ICA. This preprocessing stage shows that the initial transformation of the monitored image vector x_{si} is in linear form, consequently, the acquired new image vector $\tilde{x}_{si}$ is already whitened in nature, i.e. it consist the various components which are already in uncorrelated manner and their variances show similar agreement. The whitening transformation is also an important approach for image classification which uses eigenvalue decomposition (EVD) of image data related covariance matrix,

$$E\left\{x_{si}x_{si}^{T}\right\} = E_{oe}D_{de}E_{oe}^{T} \tag{1}$$

where, E denotes the orthogonal matrix of eigenvectors of $E\left\{x_{si}x_{si}^{T}\right\}$ and D shows the diagonal matrix having its eigenvalues, D = diag(d_1, …, d_n). In this case, the available sample x(1), …, x(T) is utilized to approximate the $E\left\{x_{si}x_{si}^{T}\right\}$ in a standard way. This whitening can be processed in the following way,

$$\tilde{x}_{si} = E_{oe}D_{de}^{-1/2}E_{oe}^{T}x_{si} \tag{2}$$

where, the diagonal matrix $D_{de}^{-1/2}$ is evaluated by using a easy component based operation as

$$D_{de}^{1/2} = diag(D_1^{-1/2}, \ldots, D_n^{-1/2}) \tag{3}$$

It is straightforward to make certain that $E\left\{\tilde{x}_{si}\tilde{x}_{si}^{T}\right\} = 1$ is achieved, due to the whiten nature of this obtained new image. The mixing matrix is transformed by whitening into a new matrix that is form of orthogonal. Now, it can be examine that the whitening process plays a crucial role to use limited parameters, as an alternative of utilizing related parameters of image matrix. Therefore, the procedure of whitening has become an essential component of the ICA. It is also an approachable technique to decrease the computational complexity in this way. Hence, this proposed approach is adjustable and utilized for identifying the different objects in satellite images.

2.1 Sphering the Satellite Images Using FP-ICA Approach

In this proposed approach, FP-ICA method is used to sphered (whiten) data by deriving a definite level maxima of the estimation of negentropy for satellite images. It identifies less correlated pixels that increase the redundancy level in adjacent pixels and segregate them in different classes. The non-gaussianity is calculated by using the value of negentropy in easy way. FP-ICA method gives a linear depiction of non-gaussian image data to maintain the significance of individual image pixel.

The FP-ICA is based on a fixed-point iteration method to find the non-gaussianity of $w_{si}^T x_{si}$ [11, 13]. To evaluate it, an approximate Newton iteration is assisted. According to Kuhn-Tucker conditions [14], Eq. (4) helps to find the optima of $\tilde{x}_{si}$.

$$E\left\{x_{si} g\left(w_{si}^T x_{si}\right)\right\} - \beta w_{si} = 0 \tag{4}$$

At this instant, the Eq. (5) after applying Newton method on each independent components of Eq. (4) to deliver s a Jacobian matrix $JF(w_{si})$,

$$JF(w_{si}) = E\left\{x_{si} x_{si}^T g'\left(w_{si}^T x_{si}\right)\right\} - \beta I \tag{5}$$

The first term of Eq. (5) estimate to make easy the inversion of this matrix.

$$E\left\{x_{si} x_{si}^T g'\left(w_{si}^T x_{si}\right)\right\} \approx E\left\{x_{si} x_{si}^T\right\} E\left\{g'\left(w_{si}^T x_{si}\right)\right\} \tag{6}$$

$$= E\left\{g'\left(w_{si}^T x_{si}\right)\right\} I \tag{7}$$

This Newton iteration is followed with the estimation of β,

$$w_{si}^+ = w_{si} - \left[E\left\{x_{si} g\left(w_{si}^T x_{si}\right)\right\} - \beta w_{si}\right] / \left[E\left\{g'\left(w_{si}^T x_{si}\right)\right\} - \beta\right] \tag{8}$$

After processing the normalization of Eq. (8), the subsequent algorithm is estimated for FP-ICA.

$$w_{si}^+ = E\left\{x_{si} g\left(w_{si}^T x_{si}\right)\right\} - E\left\{g'\left(w_{si}^T x_{si}\right)\right\} w_{si} \tag{9}$$

and

$$w_{si}^* = w_{si}^+ / \left\| w_{si}^+ \right\| \tag{10}$$

where, w_{si}^* describes a new value of w_{si} for sphere the satellite image data. FP-ICA algorithms exploited many independent components with a number of units (e.g. neurons) and their corresponding weight vectors $w_1, \ldots, w_n$ for the entire satellite

image. In the continuation of iteration, the decorrelation step for image outputs $w_1^T x_{si}, \ldots, w_n^T x_{si}$ that assists to avoid distinct vectors from converging to the similar maxima (single image class). An easy way for finding decorrelation is to approximate the independent components 'c' ('c' vector) in a sequence and $w_1, \ldots, w_c$ are approximated. To find w_{c+1}, a one-unit fixed-point algorithm is performed and the next weight vector is subtracted from w_{c+1} in next iteration that is the real projections $w_{c+1}^T w_j w_j$. Thereafter w_{c+1} is renormalized, where j = 1, …, c are the earlier estimated c vectors.

The proposed approach utilizes FP-ICA method to classify the image objects. It also determines and suppresses the mixed class problem that helps to maintain the principle of mutual exclusion with different classes. Initially, HRSI is taken input in matrix form 'x_{si}', which dimensions are x (sources) and s (samples). Different spectral values of pixels and image pixels are columns and rows of image matrix respectively. The other possible input argument is the number of independent components (c) that is an essential factor to extract image classes (default c = 1) and concurrently weighting coefficients (w_{si}) to identify distinct objects from mixed classes.

The proposed structure of FP-ICA approach is shown in Fig. 1 for classification of satellite images. Firstly a preprocessing step is taken for whitening or sphering process on the HRSI. Thereafter, non-gaussianity processing and negentropy evaluation on sphered data is done before utilizing FP-ICA method for classification. Non-gaussianity provides the less similarity between image classes and its maximum values reveals the separations between the mixed classes. These classes are generally less non-gausssian in comparison to existing classes which are

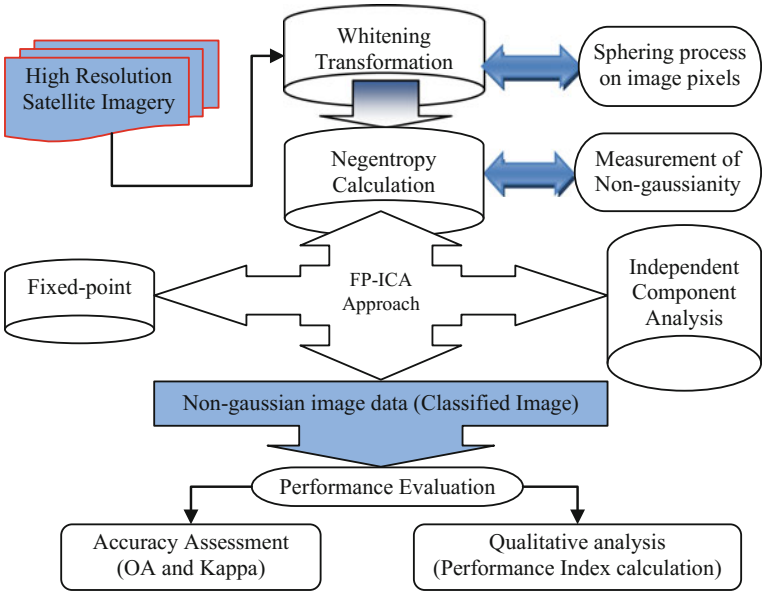


Fig. 1 A detailed framework for classification of mixed classes in HRSI

different spectral values. The non-gaussianity in FP-ICA approach is measured by the negentropy. The maximization of the negentropy is needed to reduce the mutual information among the different classes of HRSI. Finally, a classified image shows the good level of separation among in different classes.

3 Results and Discussion

Figure 2a and b show the different input satellite images SI1 and SI2 respectively. The classification results for input satellite images SI1 and SI2 are shown in Fig. 2c and d respectively. The lower values of performance index (PI) illustrate that the

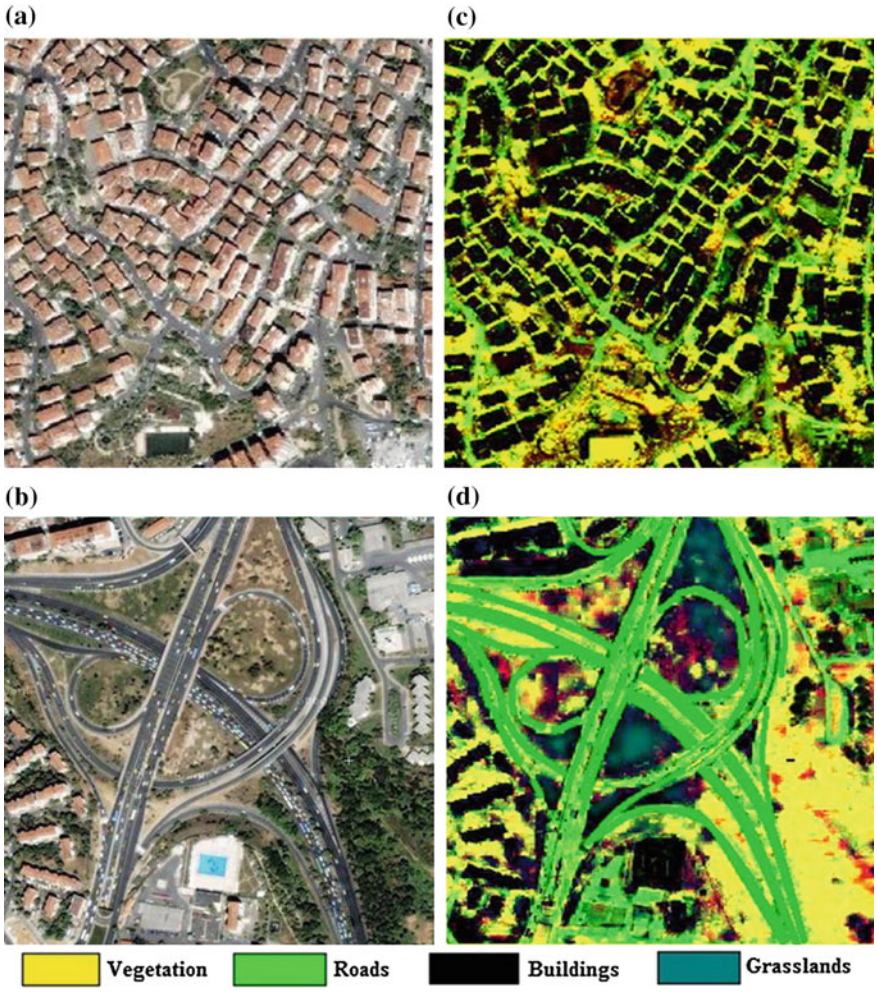


Fig. 2 FP-ICA based classified results: **a**, **b** input image; **c**, **d** result of the proposed approach

separation of classes is achieved efficiently. Figure 2d shows the roads, vegetation, buildings and grasslands area in green, yellow, black and grey colors respectively. In Fig. 2d buildings, grasslands looks in black color and dark grey respectively. soil background is seen in bluish color.

3.1 Performance Index and Accuracy Assessment

The PI values are used to calculate the qualitative analysis of the proposed FP-ICA algorithm. The value of PI is described as follows [15]:

$$PI = \frac{1}{n(n-1)} \sum_{k=1}^{n} \left\{ \left(\sum_{m=1}^{n} \frac{|gs_{km}|}{\max_l |gs_{kl}|} - 1 \right) + \left(\sum_{m=1}^{n} \frac{|gs_{mk}|}{\max_l |gs_{lk}|} - 1 \right) \right\} \tag{11}$$

where, gs_{kl} is indicated as the (k, l) component in the matrix of the entire image,

$$GS = WH \tag{12}$$

where, the maximum value of the components in the *i*th row vector of G express as $\max_l gs_{kl}$ and the maximum value of the components in the *i*th column vector of GS express as $\max_l gs_{kl}$. W and H stand for the separating matrix and mixing matrix respectively. If the PI value is almost zero or smaller, it signifies that the perfect separation has been acquired at the *k*th extraction of processing unit.

Table 1 shows the performance measurement of proposed approach for image classification. The given number of iteration is needed to converge the method in the mentioned elapsed time. Figure 2c and d demonstrate the convergence of the FP-ICA method for ES1 and ES2 images, which took 2.03 and 2.14 s time respectively to classify the satellite image in different classes. Further, PI values 0.0651 and 0.0745 respectively are calculated for ES1 and ES2 satellite images, which illustrate the better separation in different classes of different image. The suggested iterations are required to converge this method in the desired time. The comparison of the calculated Kappa coefficient (κ) and overall accuracy (OA) for image classification with the help of proposed FP-ICA method is shown in Table 2.

Tables 3 and 4 illustrates the producer's accuracy (PA) and user's accuracy (UA) respectively for classification of HRSI.

Table 1 Performance measurement of proposed FP-ICA algorithm for classification of HRSI

HRSI	Iteration	Time elapsed (s)	Performance index (PI)
SI1	1–14	2.03	0.0651
SI2	1–16	2.14	0.0745

Table 2 Kappa coefficient and overall accuracy of classification of HRSI

HRSI	Proposed approach	
	Overall accuracy (OA) (%)	Kappa coefficient (κ)
SI1	96.65	0.8675
SI2	95.78	0.8546

Table 3 Producer's accuracy of proposed FP-ICA algorithm for classification of HRSI

HRSI	Producer's accuracy (PA) (%)			
	Vegetation	Buildings	Roads	Grassland
SI1	95.65	96.81	97.96	–
SI2	93.89	93.74	96.74	96.58

Table 4 User's accuracy of proposed FP-ICA algorithm for classification of HRSI

HRSI	User's accuracy (UA) (%)			
	Vegetation	Buildings	Roads	Grassland
SI1	91.41	94.09	95.81	–
SI2	93.54	91.43	94.79	94.18

4 Conclusion

In this paper, FP-ICA based approach achieves good level of accuracy for satellite image classification in different classes. Therefore, the classified image facilitates to recognize the existing class information distinctly. The classified image results demonstrate the existing objects such as buildings, grasslands, roads, and vegetation in HRSI of emerging urban area. The uncorrelated image classes are identified as dissimilar objects by using this FP-ICA approach. This proposed approach is reasonably effective to decrease the issue of mixed classes in satellite images by incorporation of whitening process on suppressing the effect of spectral similarities among different classes. Thus, this proposed approach provides major role in classification of satellite images among four major classes. The experimental outcomes of the satellite image classification evidently specify that the proposed approach has been achieved a good level of accuracy and convergence speed of classification results. It is also suitable to resolve the classification of satellite image problem in the existence of mixed classes. Moreover, the expansion in ICA with other techniques can also be tested.

References

1. Amari, S.: Natural gradient work efficiently in learning. Neural Comput. **10**(2), 251–276 (1998).
2. Cichocki, A., Unbehauen, R., Rummert, E.: Robust learning algorithm for blind separation of signals. Electronics Letters **30**(17), 1386–1387 (1994).

3. Zhang, L., Amari, S., Cichocki, A.: Natural Gradient Approach to Blind Separation of Over- and Under-complete Mixtures. In Proceedings of ICA'99, Aussois, France, January 1999, pp. 455–460 (1999).
4. Zhang, L., Amari, S., Cichocki, A.: Equi-convergence Algorithm for blind separation of sources with arbitrary distributions. In Bio-Inspired Applications of Connectionism IWANN 2001. Lecture notes in computer science, vol. 2085, pp. 826–833 (2001).
5. Mohammadzadeh, A., ValadanZoej, M.J., Tavakoli, A.: Automatic main road extraction from high resolution satellite imageries by means of particle swarm optimization applied to a fuzzy based mean calculation approach. Journal of Indian society of Remote Sensing **37**(2), 173–184 (2009).
6. Singh, P.P., Garg, R.D.: Automatic Road Extraction from High Resolution Satellite Image using Adaptive Global Thresholding and Morphological Operations. J. Indian Soc. of Remote Sens. **41**(3), 631–640 (2013).
7. Benediktsson, J.A., Pesaresi, M., and Arnason, K.: Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. IEEE Transactions on Geoscience and Remote Sensing **41**, 1940–1949 (2003).
8. Segl, K., Kaufmann, H.: Detection of small objects from high-resolution panchromatic satellite imagery based on supervised image segmentation. IEEE Transactions on Geoscience and Remote Sensing **39**, 2080–2083 (2001).
9. Li, G., Wan,Y., Chen, C.: Automatic building extraction based on region growing, mutual information match and snake model. Information Computing and Applications, Part II, CCIS, vol. 106, pp. 476–483 (2010).
10. Singh, P.P., Garg, R.D.: A Hybrid approach for Information Extraction from High Resolution Satellite Imagery. International Journal of Image and Graphics **13**(2), 1340007(1–16) (2013).
11. Hyvarinen, A., Oja, E.: A Fast Fixed-Point Algorithm for Independent Component Analysis. Neural Computation **9**(7), 1483–1492 (1997).
12. Hyvarinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks **13**(4–5), 411-430 (2000).
13. Hyvarinen, A.: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks **10**(3), 626–634 (1999).
14. Luenberger, D.: Optimization by Vector Space Methods, Wiley (1969).
15. Singh, P.P., Garg, R.D.: Fixed Point ICA Based Approach for Maximizing the Non-gaussianity in Remote Sensing Image Classification. Journal of Indian Society of Remote Sensing **43**(4), 851–858 (2015).

A Novel Fuzzy Based Satellite Image Enhancement

Nitin Sharma and Om Prakash Verma

Abstract A new approach is presented for the enhancement of color satellite images using the fuzzy logic technique. The hue, saturation, and gray level intensity (HSV) color space is applied for the purpose of color satellite image enhancement. The hue and saturation component of color satellite image are kept intact to preserve the original color information of an image. A modified sigmoid and modified Gaussian membership functions are used for the enhancement of the gray level intensity of underexposed and overexposed satellite images. Performance measures like luminance, entropy, average contrast and contrast enhancement function are evaluated for the proposed approach and compare with histogram equalization, discrete cosine transform (DCT) method. On comparison, this approach is found to be better than the recent used approaches.

Keywords Satellite image enhancement · Singular value decomposition · Contrast assessment function

1 Introduction

Image enhancement is an essential step used in medical imaging, defence operations, metrological department, and satellites image etc. applications [1]. We modify the pixel values in order to get the enhanced image. Satellite images have poor resolution and contrast as the images are captured from long distances. The generally methods used for enhancement are histogram equalization, contrast stretching that may improve the resolution only or contrast only or both [2, 3]. Sometime pixels values may get saturated results the artifact which distorts the

N. Sharma (✉)
Maharaja Agrasen Institute of Technology, Rohini, Delhi, India
e-mail: sharmaisnitin@gmail.com

O.P. Verma
Delhi Technological University, Delhi, India
e-mail: opverma@dce.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_38

details of an image. SVE, SVD, DCT and DWT are other approaches which are also applied to enhance the satellite images [4]. The researchers are applied corrections either on whole image or in any one component of the image. These pixel operations applied on whole image do not preserve the edge information [5–7]. Sometimes they also enhanced the noisy pixels. The edge information is preserved if we apply the approaches in the lower region of image like applying gamma correction on LL component of discrete wavelet transformed satellite image [8]. Due to the presence of different region in the same satellite image the dynamic range of other components are not increased. The variant fuzzy algorithms are studied in literature to improve the uncertainty in the image enhancement [9–14]. The modified fuzzy based approach is proposed to increases the contrast as well as the dynamic range of the satellite image. The proposed approach removes the uncertainty in the different spectral region of satellite images and enhances the quality of the image.

The organization of the paper is given as follows: Sect. 2 introduces the concept of SVD in contrast enhancement. This decomposition increases the illumination of the image. Section 3 presents the fuzzy based contrast enhancement. In Sect. 4, we define the proposed approach use to increase the contrast of the satellite image. The analysis based on the performance measures are given in Sect. 5. The results and conclusions are drawn in Sects. 6 and 7 respectively.

2 Singular Value Decomposition (SVD) Based Enhancement

SVD technique is used in the image enhancement for improving the illumination of the given image. It is well-known that the SVD has optimal decorrelation and sub rank approximation properties. This technique match the each block on the basis of singular vectors. Use of SVD in image processing applications is described by [5]. The equation for singular value decomposition of X is as follows

$$X = UDV^T \tag{1}$$

where X is the closest rank approximation matrix. The U, V and D is the left, right and diagonal matrix. Among U, V and D, we kept our concentration on the diagonal Eigen values only. These values are employed in the correction step which gives the rank approximation. For example, the 4 × 4 block input to the SVD block will have a total of 4 – k eigenvalues, where k denotes the number of different diagonal elements. The significant eigenvalues are considered after arranging the Eigen values in the descending order along the diagonal of the 'D' matrix for applying correction factors in the input sample image.

3 Fuzzy Based Contrast Enhancement

In this paper HSV color model (Hue, Saturation and Gray level value) is considered for satellite image enhancement. As if we perform the image enhancement directly to Red, Green and Blue components of satellite image then it results the color artifact. The color image enhancement must preserve the original color information of the satellite image and increases the pixel intensity in such a way that it cannot exceed the maximum value of an image. The color image enhancement must be free from the gamut problem. Fuzzy set theory is one of the useful tools used to reduce the uncertainty and ambiguity. In reference to the satellite image enhancement the one important step have to involve is the creation of "IF…, Then…, Else" fuzzy rules for image enhancement. A set of neighbourhood pixels results the prior and subsequent clauses that behave as the fuzzy rule for the pixel to be enhanced. These rules give decision similar to the reasoning of human beings. The fuzzy approach mentioned above is employ on the luminance part of the color satellite image. The reason for the same is that doing enhancement operation the luminance value of pixels get overenhance or underenhance due to the fact that they will not sync with the range of histogram. The fuzzy satellite image processing involves image fuzzification, modification of membership values and defuzzification. The selection of membership function is based on the application.

The grayscale image of size M × N having the intensity levels I_{mn} in the range [0, L − 1] is described as

$$I = \bigcup \{\mu(I_{mn})\} = \left\{ \frac{\mu_{mn}}{I_{mn}} \right\}, \mathrm{m} = 1, 2, \ldots, \mathrm{M} \text{ and } \mathrm{n} = 1, 2, \ldots, \mathrm{N} \tag{2}$$

where $\mu(I_{mn})$ represents the membership with I_{mn} being the intensity at the (m^{th}, n^{th}) pixel. An image can be classified into the low, mid and high intensity region according to the defined three fuzzy rules (Table 1).

The above said rules are applied on different multispectral regions of the color satellite image. The image splits into three different intensity regions. These divided regions centred on the mean value for each region. The modified sigmoidal membership function used to fuzzify the low and high intensity region of satellite image is as follows,

$$\mu_X(x) = \frac{1}{1 + e^{-(x)}} \tag{3}$$

where x represents the gray level of low and high intensity region.

Table 1 List of rules

If the intensity is low then, region belongs to underexposed
If the intensity is mid then, region is moderate
If the intensity is high then, region belongs to overexposed

A modified Gaussian membership function is used to fuzzify the mid intensity region of the image mathematically as follows:

$$\mu_{X_g}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-x_{avg})^2}{2\sigma^2}} \tag{4}$$

where, x indicates the gray level intensity value of mid intensity region, x_{avg} is the average or mean gray level value in the given image and σ represents the deviation from the mean or average gray level intensity value. These membership functions transform the image intensity from spatial domain to the fuzzy domain. The values of the same are in the range of 0–1. The sigmoid membership function modifies the underexposed and overexposed regions of given satellite image. The mid intensity region are modifies using the Gaussian membership function. Finally, r_o is the output membership function are defuzzify by using constant membership function given by,

$$r_o = \sum_{i=1}^{3} \frac{\mu_i r_i}{\mu_i} \tag{5}$$

where, μ_i is the crossover points, r_i is the input pixel intensities.

4 Proposed Methodology

Satellite image contains multiple frequency components. These components are high, low and mixed frequency components. We have been increased the dynamic range in all these frequency range. Generally, these images contain edges which contains high frequency component. The fuzzy based satellite color image enhancement algorithm is described in Table 2.

Table 2 Fuzzy based satellite color image enhancement algorithm

Fuzzy based satellite color image enhancement algorithm
1. Read a color satellite image and obtain three color channels i.e. X = {R, G, B}
2. Transform the color channels X into HSV color space
3. Obtain the normalized value of each channel ($\bar{H}$, $\bar{S}$, $\bar{I}$}
4. The intensity channel $\bar{I}$ is fuzzified in three membership functions defined in Eqs. (2)–(4)
5. Obtain the modified membership values for underexposed and overexposed regions
6. Defuzzification is done using Eq. (5)
7. Finally, apply the correction factor to the degraded channel after performing SVD using Eq. (1)

5 Performance Measures

The contrast assessment function (CAF) is used to measure the quality of the proposed methodology. This function calculates by computing the brightness, entropy and contrast. These measures evaluate the quality of the enhanced image.

$$CAF = IE^{\alpha} + \bar{C}^{\beta} \tag{6}$$

where, the IE and $\bar{C}$ represents the average entropy value of the given image and the average color contrast in the image [15].

6 Results and Discussions

The proposed approach has been successfully implemented on Intel core i3 at 2.40 GHz using MATLAB version 2014. The low intensity satellite images used for experimentation are obtained from the NASA's earth observatory Lab. To show the effectiveness of the proposed approach we are chosen four test images shown in Fig. 1 a1–a4. These images are enhanced with proposed Fuzzy based approach and are compared with state of art approach like histogram equalization and discrete cosine transform. The images obtained with conventional approaches are shown from Fig. 1 c1–c4 with histogram equalization but these images are further improved by using DCT method are shown from Fig. 1 d1–d4. The images obtained with proposed method are enhanced the lower, middle and higher intensity region. There is an incremental change in the pixel intensity. This will increase the contrast in each of these regions. The images obtained with proposed methodology increase the contrast with details in each intensity level. The same results depict with the color assessment function which shows that the proposed approach improves the quality of image and measured the value of entropy, brightness and contrast. The CAF value achieved 22.9368 with the proposed approach which is higher than the CAF value obtained with the other methods like histogram based is 18.4498 and with DCT based approach is 19.4610 for Fig. 1 a1. Similarly, we achieved higher CAF value than other methods for other three satellite images is shown in Table 3. This value is higher than the other methods. These measures indicate the quality of the image.

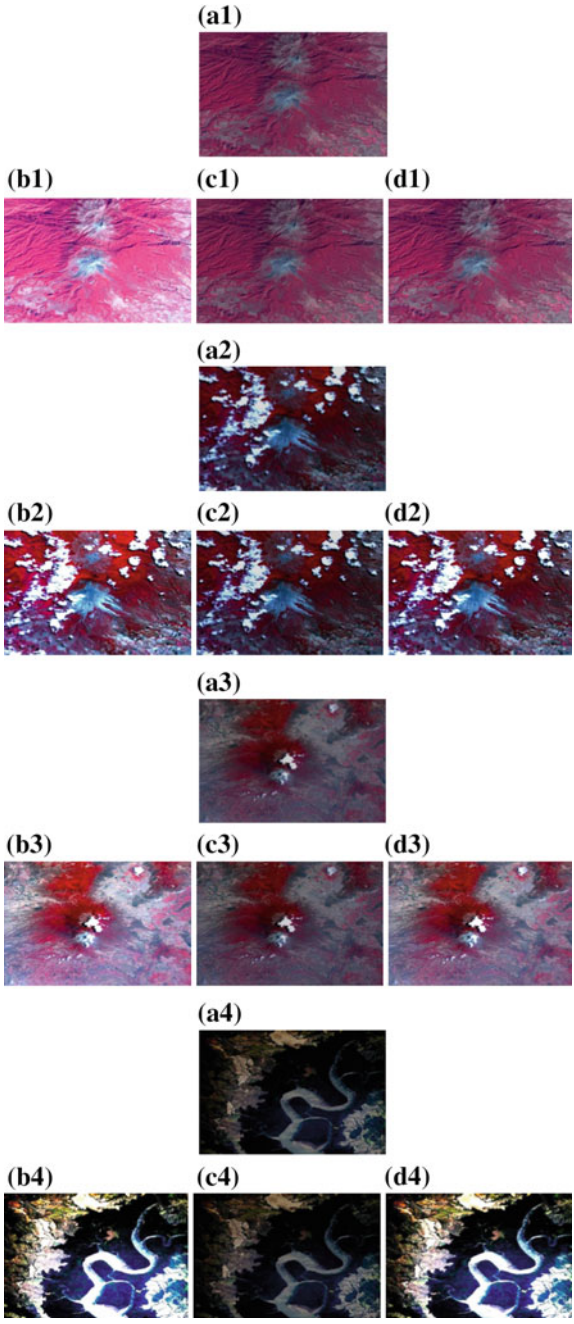


Fig. 1 **a1–a4** show the test image, **b1–b4** show the fuzzy based proposed approach image, **c1–c4** show the Histogram equalized image, **d1–d4** show the DCT based image

Table 3 Comparison of luminance, entropy, average contrast, CAF values using proposed approach, histogram equalization and DCT based methods over different images

Images		Luminance $\bar{L}$	Average Entropy IE	Average contrast $\bar{C}$	CAF
1	Test image	79.9986	6.4551	65.6653	18.3753
	Proposed approach	141.8108	7.0530	111.8474	22.9368
	HE based method	79.9941	6.4722	66.0319	18.4498
	DCT based method	89.4299	6.6564	73.0664	19.4610
2	Test image	65.4602	7.0769	99.3299	22.3417
	Proposed approach	102.2152	7.4782	148.6010	26.1097
	HE based method	65.5456	7.0912	96.0337	22.1985
	DCT based method	85.2260	7.3169	127.5119	24.5875
3	Test image	71.6810	6.6087	54.6698	17.9702
	Proposed approach	126.4009	7.5229	103.4835	23.9940
	HE based method	67.5417	6.6926	67.5481	19.1867
	DCT based method	101.3466	7.1918	81.3107	21.5960
4	Test image	39.8276	5.8954	56.0258	16.1292
	Proposed approach	91.5350	6.7875	149.6766	23.7411
	HE based method	39.9452	5.9362	55.8990	16.2316
	DCT based method	85.5918	6.7959	144.7823	23.5737

7 Conclusions

Fuzzy based satellite image enhancement has been implemented by fuzzifying the color intensity property of an image. An image may be categorized into overexposed and underexposed regions. A suitable modified membership functions are employed for the fuzzification of different regions. The results of the proposed fuzzy based contrast enhancement approach have been compared with the recent used approaches. The proposed approach successfully depicts the better result in terms of luminance, entropy, average contrast and CAF value.

References

1. Gonzalez, C. Rafael, and E. Richard. “Woods, digital image processing.” ed: Prentice Hall Press, ISBN 0-201-18075-8, 2002.
2. Gillespie, R. Alan, B. Anne, Kahle, and E. Richard Walker. “Color enhancement of highly correlated images. I. Decorrelation and HSI contrast stretches.” Remote Sensing of Environment 20, vol. 3, pp. 209–235, 1986.
3. P. Dong-Liang and X. An-Ke, “Degraded image enhancement with applications in robot vision,” in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 2005, vol. 2, pp. 1837–1842.
4. H. Ibrahim and N. S. P. Kong, “Brightness preserving dynamic histogram equalization for image contrast enhancement,” IEEE Trans. Consum. Electron., vol. 53, no. 4, pp. 1752–1758, Nov. 2007.

5. H. Demirel, G. Anbarjafari, and M. N. S. Jahromi, "Image equalization based on singular value decomposition," in Proc. 23rd IEEE Int. Symp. Comput. Inf. Sci., Istanbul, Turkey, Oct. 2008, pp. 1–5.
6. Bhandari, A. K., A. Kumar, and P. K. Padhy. "Enhancement of low contrast satellite images using discrete cosine transform and singular value decomposition." World Academy of Science, Engineering and Technology 79 (2011): 35–41.
7. Demirel, Hasan, Cagri Ozcinar, and Gholamreza Anbarjafari. "Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition." Geoscience and Remote Sensing Letters, IEEE 7, no. 2 (2010): 333–337.
8. Sharma, N., & Verma, O. P. (2014, May). Gamma correction based satellite image enhancement using singular value decomposition and discrete wavelet transform. In Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference on (pp. 1286–1289). IEEE.
9. S. K. Naik and C. A. Murthy, "Hue-preserving color image enhancement without gamut problem," IEEE Trans. Image Process., vol. 12, no. 12, pp. 1591–1598, Dec. 2003.
10. F. Russo, "Recent advances in fuzzy techniques for image enhancement," IEEE Trans. Image Process., vol. 47, no. 6, pp. 1428–1434, Dec. 1998.
11. Verma, O. P., Kumar, P., Hanmandlu, M., & Chhabra, S. (2012). High dynamic range optimal fuzzy color image enhancement using artificial ant colony system. Applied Soft Computing, 12(1), 394–404.
12. M. Hanmandlu and D. Jha, "An optimal fuzzy system for color image enhancement," IEEE Trans. Image Process., vol. 15, no. 10, pp. 2956–2966, Oct. 2006.
13. M. Hanmandlu, S. N. Tandon, and A. H. Mir, "A new fuzzy logic based image enhancement," Biomed. Sci. Instrum., vol. 34, pp. 590–595, 1997.
14. Hanmandlu, Madasu, Om Prakash Verma, Nukala Krishna Kumar, and Muralidhar Kulkarni. "A novel optimal fuzzy system for color image enhancement using bacterial foraging." Instrumentation and Measurement, IEEE Transactions on 58, no. 8 (2009): 2867–2879.
15. Xie, Zheng-Xiang, and Zhi-Fang Wang. "Color image quality assessment based on image quality parameters perceived by human vision system." Multimedia Technology (ICMT), 2010 International Conference on. IEEE, 2010.

Differentiating Photographic and PRCG Images Using Tampering Localization Features

Roshan Sai Ayyalasomayajula and Vinod Pankajakshan

Abstract A large number of sophisticated, yet easily accessible computer graphics softwares (STUDIO MAX, 3D MAYA, etc.) have been developed in the recent past. The images generated with these softwares appear to be realistic and cannot be distinguished from natural images visually. As a result, distinguishing between photographic images (PIM) and Photo-realistic computer generated (PRCG) images of real world objects has become an active area of research. In this paper, we propose that "a computer generated image" would have the features corresponding to a "completely tampered image", whereas a camera generated picture would not. So, the differentiation is done on the basis of tampering localization features viz., block measure factors based on JPEG compression and re-sampling. It has been observed experimentally, that these measure factors vary for a PIM from a PRCG image. The experimental results show that the proposed simple and robust classifier is able to differentiate between PIM and PRCG images with an accuracy of 96 %.

Keywords Image forensics · Photographic images · Photorealistic computer generated images · Tampering localization · Steganalysis

1 Introduction

The ease with which digital images can be modified to alter their content and meaning of what is represented in them has been increasing with the advancing technology. The context in which these images are involved could be a tabloid, an advertising poster, and also a court of law where the image could be legal evidence. Many algorithms are now developing using the advantage of making a machine

R.S. Ayyalasomayajula (✉) · V. Pankajakshan
Electronics and Communication Engineering Department,
Indian Institute of Technology Roorkee, Roorkee, India
e-mail: saiaruce@iitr.ac.in

V. Pankajakshan
e-mail: vinodfec@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_39

learn to classify datasets into various classes, this classification is basically done by identifying a set of features that are different for PIM and PRCG images. Later these features are used to train a set of sample images and a threshold is set by the so trained classifier, based on which testing is done.

There remains controversy in both forensics and human vision fields as there are no agreed standards for measuring the realism of computer generated graphics. Three varieties exists in computer graphics which mainly differ in the level of visual coding.

1. Physical realism: same visual stimulation as that of the scene.
2. Photo realism: same visual response as the scene.
3. Functional realism: same visual information as the scene.

Among the three, the first kind of realism is hard to achieve in real applications. Computer images which have the last kind of realism usually presented as cartoons and sketches and hence can be easily classified as Computer Generated images. The last kind of realism forms the crucial part of computer generated images as they look as real as a photograph. Farid [1] has categorized the various tools used for image forgery detection into five different categories of techniques used:

1. Detection of statistical anomalies at pixel level.
2. Detection of statistical correlation introduced by lossy compression.
3. Detection of the artefacts introduced by the camera lens, sensor or post-processing operations.
4. Detection of anomalies in object interaction with light and sensor in 3D domain.
5. Detection of real world measurements.

Ng and Chang [2], developed a classifier using 33D power spectrum features, 24D local patch features, and 72D higher order wavelet statistic features, which come under category 1 (Cat-1) defined in [1]. Lyu and Farid [3] proposed a 214D feature vector based on first four order statistics of wavelet sub-band coefficients and the error between the original and predicted sub-band coefficients. More recently, Wang et al. [4] have proposed a 70D feature vector based on statistical features extracted from the co-occurrence matrices of differential matrices based on contourlet transform of the image and homomorphically filtered image and texture similarity of these two images and their contourlet difference matrices (Cat-1). Based on the physics of an imaging and image processing of a real world natural image, Ng et al. [5] have proposed a 192D feature vector utilizing the physical differences in PIM and PRCG images, viz., local patch statistics (Cat-1), local fractal dimension, surface gradient, quadratic geometry, and Beltrami flow (Cat-4&5). Every PIM is obtained using a specific camera, Dehnie et al. [6] differentiated PIM and PRCG image based on camera response or sensor noise, which is specific to a given digital camera and that this sensor noise would be absent in a PRCG image. Gallagher and Chen [7] later used features extracted based on Bayer's color filter array (CFA) (Cat-3). More recently, methods that can discriminate computer generated and natural human faces (Cat-1, 4&5) have been developed [8].

Among all the proposed methods, an accuracy of 98.4 % reported by Gallagher and Chen [7] on Columbia Image dataset [9] was the highest, and 98 % by Wang et al. [4] tested on the same image dataset and an average of 94.5 % while training and testing over different datasets.

All the existing methods, used features which come under, Cat-1, 3, 4 and 5 mentioned in [1], to distinguish between PIM and PRCG. But due to the basic imaging process and on-chip post-processing done over an image, being different for a PIM and a PRCG image, there are significant traces left behind that can be traced back to differentiate between them. On-chip post-processing performed over a PIM and PRCG image, leaves different signatures i.e., the lossy compression statistical correlation artefacts are different for a PIM to a PRCG image. These artefacts are widely used to detect double compression in images, which is a footprint of a tampered image.

So far the need for differentiating PIM and PRCG images and existing methods were discussed. The rest of the paper is organized as follows. In Sect. 2, we discuss the proposed method and the reasoning behind the ability of classification of the chosen block artefact features. We show the experimental results and comparison with the existing algorithms in Sect. 3, and conclude in Sect. 4.

2 Proposed Method

After conducting a series of experiments based on various other tampering localization features based on double compression, it is found that the JPEG compression artefacts realized by Zuo et al. [10], tend to be more efficient in differentiating PIM and PRCG images. The explanation for how these tampering localization features identify the PRCG from PIM is as follows.

In [10], two different features; block measure factor based JPEG compression artefacts (β) and block measure factor based Re-sampling artefacts (α). It is shown that these two features identify the tampered portion of an image. In Fig. 1a the procedure for the extraction of feature 'β' has been detailed. In Fig. 1b we see the feature extraction procedure mentioned by Gallagher and Chen [7], which is based on CFA interpolation. We see in both the methods, we first calculate the difference matrix of the image (or pass it through a high-pass filter).

Let $I(i, j)$ be the given image, first the difference image $D(i, j)$ is computed as:

$$D(i,j) = I(i,j) - I(i+1,j) - I(i,j+1) + I(i+1,j+1) \tag{1}$$

This difference map is divided into 8 × 8 blocks, $K_1, K_2, \ldots, K_n$, n is the number of blocks. As the difference at the borders would be higher than that inside the block for each block this operation gives the information of JPEG blocking artefacts in an image. This operation also reveals the interpolation information in an image assuming a Bayer array and a linear interpolation [7]. Then the mean of (i, j)th pixel in each corresponding block is found. This mean, Blocking Artefacts Matrix (BAM),

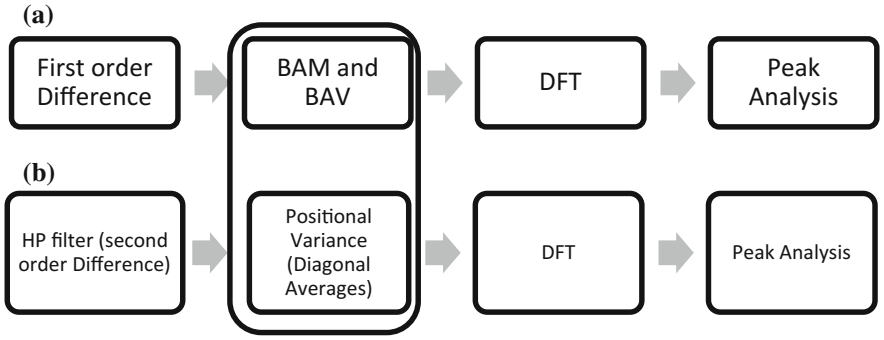


Fig. 1 **a** Calculation of β. **b** Calculation for features in [7]. We see that except for the second level rest of the analysis to calculate the features is more or less the same

which is an 8 × 8 matrix would have high values at the boundaries for a tampered image.

$$BAM(i,j) = \frac{1}{n}\sum_{c=1}^{n} |K_c(i,j)| \quad 1 < i,j < 8 \tag{2}$$

BAM is then converted to Blocking Artefacts Vector (BAV), a vector of size 64. Let the magnitude spectrum of BAV be P(w),

$$\mathrm{P(w)} = |\mathrm{DFT(BAV)}| \tag{3}$$

Now because there are sharp boundaries (transitions) at the edges of the BAM, the DFT of BAV would result in clearly distinguishable peaks at (normalized frequency) w = m/8, m = 1, 2, …, 7 as mentioned in [10]. But in experiments, it is observed that these peaks are more pronounced in the case of a PIM (Fig. 2a) than in the case of a PRCG (Fig. 2b). A PIM is generated with the help of CFA pattern, because of which interpolation occurs and thus the peaks P(w) are more pronounced for a PIM, than for a PRCG image. Since a PRCG image is created using a software, it would not show the artefacts related to CFA interpolation.

In Fig. 2a, b distinct peaks can be observed at w = m/8, m = 1, 2, …, 7 for both PIM and PRCG, but they are more stronger in the case of a PIM than a PRCG image. Thus, it is possible to differentiate a PIM and PRCG based on β, computed as:

$$\beta = \log\left(P\left(\frac{1}{8}\right) * P\left(\frac{2}{8}\right) * P\left(\frac{3}{8}\right) + \varepsilon\right) - \log(P(0) + \varepsilon) \tag{4}$$

Gallagher and Chen [7] have also calculated a feature to identify the CFA interpolation artefacts using a similar algorithm.

Similarly, another feature block measure factor, ‘α’ is calculated [10] based on JPEG resampling, which is determined using the row (column) average of the DFT of second order difference image [10].

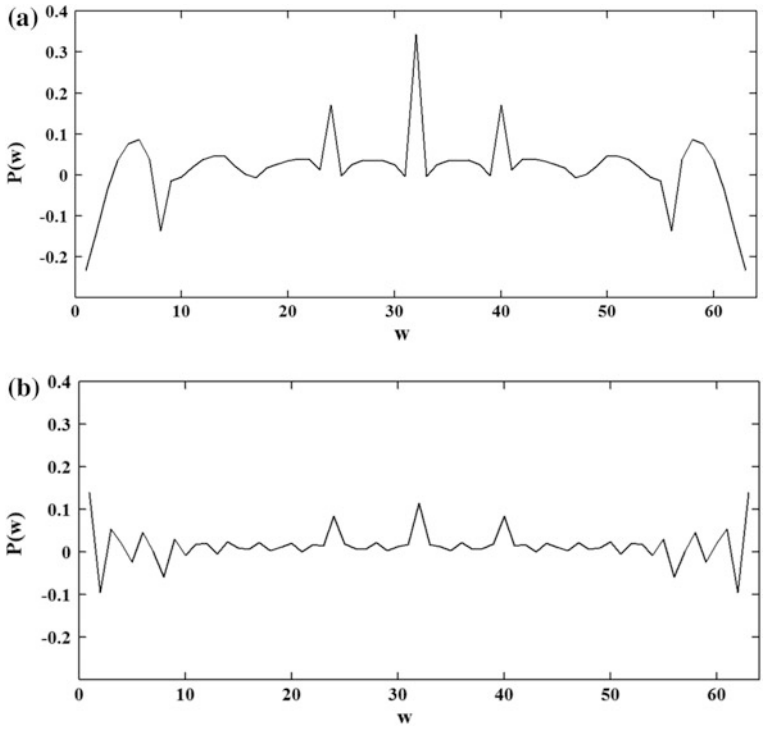


Fig. 2 P(w) versus w for **a** PIM, **b** PRCG

$$\mathrm{E}(m,n) = 2(i,\mathrm{j}) + (i,\mathrm{j}+1) - (i+1,\mathrm{j})\ 1 \le \mathrm{m} \le \mathrm{M}, 1 \le \mathrm{n} \le \mathrm{N}-2 \tag{5}$$

$$E_{DA}(w) = \frac{1}{M}\sum_{m=1}^{M} |DFT(E(m,n))| \tag{6}$$

$$\alpha = \left| \log\left(\frac{\sum (E_{DA}(w) + \varepsilon)}{\sum \log(E_{DA}(w) + \varepsilon)} \right) \right| \tag{7}$$

where ε is an arbitrary constant.

Both of these features to form a 2-D vector to classify the images as PIM or PRCG. For the classification of PIM and PRCG using these two features, SVM classifier with radial basis function from LIBSVM [11] is used.

3 Experiments and Results

Two different steps of testing were performed in order to evaluate the effectiveness of the proposed approach in differentiating PIM and PRCG images. The first step aims at measuring the accuracy of the proposed classifier and to compare it with the state-of-the-art, in particular with [4, 7]. In the second step, robustness of the proposed classifier is tested and its benefits over [4, 7] are detailed.

3.1 Training and Testing

To train and test the proposed 2-D feature vector classifier we have trained an SVM classifier using the pre-built libraries of LIBSVM integrated with MATLAB. Using images of different formats and conditions viz., 350 uncompressed TIFF images from Dresden Image Database [12], 1900 JPEG images taken from 23 different digital cameras from Dresden Image Database and 350 uncompressed TIFF images from UCID database [13], and 2500 computer generated images from ESPL Database [14, 15]. Also 800 JPEG compressed PIM images and 800 JPEG compressed PRCG images from Columbia Dataset [9] are considered for comparison with algorithms proposed in [4, 7].

The various values of β shown in Fig. 3 demonstrate clusters, different for different cameras and a cluster for PRCG images. This shows that the tampering based features are specific to a given digital camera. The magnitudes of β are significantly larger for a PRCG image, because of very small AC component values (Fig. 2b). This shows that there is little variations in pixel intensities from pixel to pixel in PRCG images than that in a PIM images, as mentioned in Sect. 2. From the similarities mentioned in Sect. 3 and the experimental results, it can be observed

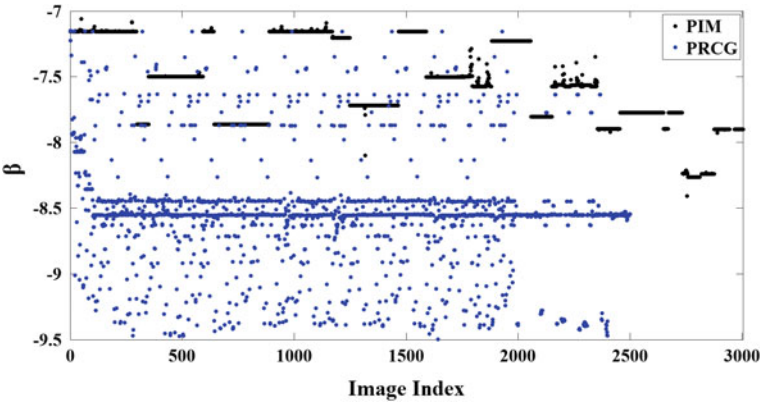


Fig. 3 Clustering of β for various cameras and PRCG images

Table 1 Classification accuracies for 10 different random combinations of images

Experiment no.	Rate of detection (%)		
	Total detection rate	True positives (PIM)	True positives (PRCG)
1	92.66	93.83	91.49
2	92.94	93.17	92.71
3	93.01	94.19	91.83
4	93.50	95.27	91.74
5	93.53	94.74	92.32
6	93.30	94.55	92.05
7	92.98	93.85	92.10
8	93.60	94.15	93.06
9	93.19	94.56	91.82
10	93.39	93.94	92.42

Table 2 Comparison with state-of-the-art and related features based classifiers

Method	Highest classification accuracy (%)
Wang et al. [4]	98
Gallagher and Chen [7]	98
Proposed method	96

that the source specific feature that has been estimated using these tampering based features is the CFA interpolation pattern.

These 5000 images is divided randomly into two sets of 2500 images each, one for training and the other for testing. This process is repeated for ten different random selections, the results of these are shown in Table 1, and average of these results is that there is a 7.8 % false positive rate for a 94.2 % true detection rate of camera generated images.

From Table 1 the average accuracy of detection rate can be observed to be 93.2 %. To validate the proposed approach the obtained results are compared with those of [4, 7] reported by Ng et al. [16]. The proposed classifier was tested on Columbia Image dataset with 800 PRCG and 800 PIM images using the same testing procedure. For PIM, 98.67 % true detection rate and 9.8 % false positive rate, with an average accuracy of about 95.82 % was observed. The best accuracy results of [4, 7] are shown in Table 2.

3.2 Robustness

The robustness of a classifier can be justified by Receiver Operating Characteristic (ROC) curves and Area Under ROC Curve (AUC). Experiments on Dresden Image Database and ESPL database give an ROC as shown in Fig. 4 with an AUC of 0.98 (Fig. 5).

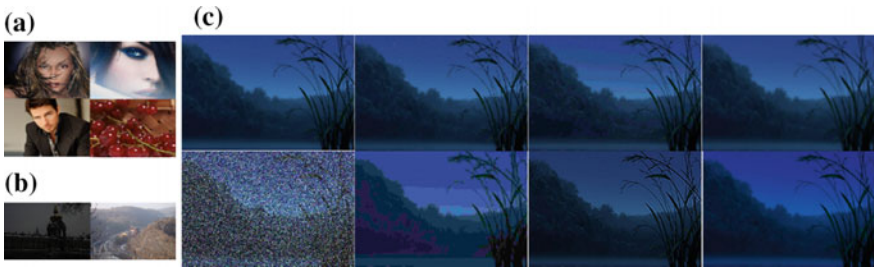


Fig. 4 **a** PRCG Images from various websites. **b** PIM from Dresden Database. **c** (*left* to *right* and *top* to *bottom*) Original example from ESPL Database, aliased, with banding artefacts, blurred, Gaussian noise added, JPEG compressed with q = 5, image with ringing artefacts, and color saturated

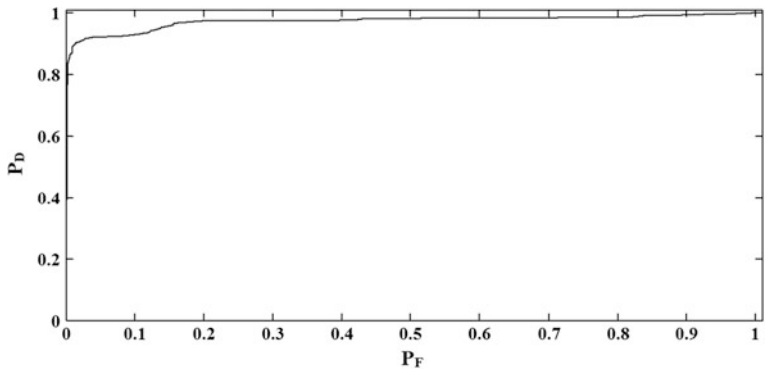


Fig. 5 ROC for Dresden [12] and ESPL [14, 15] database

Recent developments in Anti-image forensics tend to alter the image statistical and other features to hide the forgery traces, thus robustness test of a classifier is denoted by the minimal variations in detection accuracy under incidental change in image like JPEG compression, aliasing, banding, noise addition, blurring, etc. ESPL database on which the proposed classifier has been trained and tested, contains images with many of these manipulations. But, since the proposed features are derived utilizing JPEG compression based tampering artefacts, it should be verified that our classifier is resilient to various quality factors. To test this, the uncompressed UCID images are considered and the accuracy of the proposed method in differentiating images with varying quality factors is shown in Table 3.

In [4] the accuracy decreases to 81.5 % during manipulations. The average accuracy of the proposed method decrease only to 89 %, even for very low quality factor JPEG compressed PRCG images. From Table 3, it can be observed that for JPEG compression with as low quality factor as 20 and 5, the proposed method still differentiates PIM and PRCG images.

Table 3 Robustness analysis for various qualities in JPEG compression

Quality factor	Average detection accuracy rate (%)
90	96.31
70	96.93
50	96.93
20	96.62
5	94.94

This can also be justified as—"a PRCG image is a completely tampered image" and thus any other tampering or content hiding manipulations done to the image are nullified.

4 Conclusion

Block Measure Factors based on JPEG Compression and Re-sampling proposed by Gallagher and Chen [7], were used to differentiate PIM from a PRCG image. It was observed that the JPEG compression artefacts used for tampering localization can be used to differentiate PIM and PRCG images. And thus a 2D feature set was used to differentiate PIM and PRCG images with an average accuracy of 96 %. The existing state-of-the-art classifier to our knowledge [4], uses a 70D feature vector for this differentiation. Though [7] uses only one single feature to achieve a high accuracy rate, it is limited by the drastic decrease in accuracy with various image manipulations. Thus the proposed method overcomes the shortcomings of [4, 7] in terms of computation complexity and robustness respectively. Though the algorithm works well without degradation in accuracy under various other manipulations like aliasing, ringing, banding, Gaussian noise addition, speckle noise addition, histogram equalization and JPEG images compressed with low quality factors, the performance starts degrading when an image is compressed multiple number of times.

References

1. Farid, Hany. "Image forgery detection." Signal Processing Magazine, IEEE 26.2 (2009): 16–25.
2. Ng, Tian-Tsong, and Shih-Fu Chang. "Classifying photographic and photorealistic computer graphic images using natural image statistics," ADVENT Technical report, 2004.
3. Lyu, Siwei, and Hany Farid. "How realistic is photorealistic?" Signal Processing, IEEE Transactions on 53.2 (2005): 845–850.
4. Wang, Xiaofeng, et al. "A statistical feature based approach to distinguish PRCG from photographs," Computer Vision and Image Understanding 128 (2014): 84–93.

5. T. Ng, S. Chang, J. Hsu, L. Xie, MP. Tsui, "Physics motivated features for distinguishing photographic images and computer graphics," in: Proceedings of MULTIMEDIA 05, New York, NY, USA, ACM2005, pp. 239–248.
6. S. Dehnie, H.T. Sencar, N.D. Memon, "Digital image forensics for identifying computer generated and digital camera images", ICIP, IEEE, Atlanta, USA (2006), pp. 2313–2316.
7. Andrew C. Gallagher, Tsuhan Chen, "Image authentication by detecting traces of demosaicing," in: Proceedings of the CVPR WVU Workshop, Anchorage, AK, USA, June 2008, pp. 1–8.
8. H. Farid, M.J. Bravo, "Perceptual discrimination of computer generated and photographic faces," Digital Invest. 8 (3–4) (2012) 226 (10).
9. Tian-Tsong Ng, Shih-Fu Chang, Jessie Hsu, Martin Pepeljugoski, "Columbia Photographic Images and Photorealistic Computer Graphics Dataset," ADVENT Technical Report #205-2004-5, Columbia University, Feb 2005.
10. Zuo, J., Pan, S., Liu, B., & Liao, X., "Tampering detection for composite images based on re-sampling and JPEG compression," In Pattern Recognition (ACPR), 2011 First Asian Conference on (pp. 169–173). IEEE.
11. Chih-Chung Chang, Chih-Jen Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), v.2 n.3, pp. 1–27, April 2011.
12. Gloe, T., & Böhme, R., "The 'Dresden Image Database' for benchmarking digital image forensics," In Proceedings of the 25th Symposium on Applied Computing (ACM SAC 2010) (Vol. 2, pp. 1585–1591).
13. G. Schaefer and M. Stich (2003) "UCID - An Uncompressed Colour Image Database", Technical Report, School of Computing and Mathematics, Nottingham Trent University, U.K.
14. D. Kundu and B. L. Evans, "Full-reference visual quality assessment for synthetic images: A subjective study," in Proc. IEEE Int. Conf. on Image Processing., Sep. 2015, accepted, September 2015.
15. Kundu, D.; Evans, B.L., "Spatial domain synthetic scene statistics," Signals, Systems and Computers, 2014 48th Asilomar Conference on, vol., no., pp. 948–954, 2–5 Nov. 2014.
16. Tian-Tsong Ng, Shih-Fu Chang, "Discrimination of computer synthesized or recaptured images from real images," Digital Image Forensics, 2013, p. 275– 309.

A Novel Chaos Based Robust Watermarking Framework

Satendra Pal Singh and Gaurav Bhatnagar

Abstract In this paper, a novel logo watermarking framework is proposed using non-linear chaotic map. The essence of proposed technique is to use chaotic map to generate keys to be used in the embedding process. Therefore, a method for generating keys is first proposed followed by the embedding process. A robust extraction process is then proposed to verify the presence of watermark from the possibly attacked watermarked image. Experimental results and attack analysis reveal the efficiency and robustness of the proposed framework.

Keywords Digital watermarking · Chaotic map · PN sequence · security

1 Introduction

Recently, there is a rapid growth in the area of internet and communication technology. As a result, one can easily transmit, store and modify digital data, such as image, audio and video, which essentially leads to the issues of illegal distribution, copy, editing, etc. of digital data. Therefore, there is a need of some security measure to prevent these issues. The technology of watermarking has recently identified as one of the possible solutions [1, 2]. The basic idea of watermarking is to embed a mark or information related to digital data into the digital data. This mark could be a copyright information, time-stamp or any other useful information, which can be extracted later for different purposes.

In the recent years, a range of watermarking schemes have been proposed for diverse purposes. These techniques can be broadly classified into frequency and spatial domain schemes [3–6]. For the frequency domain schemes, the information is

S.P. Singh · G. Bhatnagar (✉)
Department of Mathematics, Indian Institute of Technology Jodhpur, Jodhpur, India
e-mail: goravb@iitj.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_40

hided in the coefficients obtained from various transforms such as Discrete Fourier Transform (DFT) [7], Discrete Cosine Transform (DCT) [8], Wavelet Transform (DWT) and others [9, 10]. In contrast, the basic idea of the spatial domain watermarking is to directly modify the values of digital data. That brings some advantages such as simple structure, lower computational complexity, easy implementation, etc., but also result in its lower security and weaker robustness to signal processing attacks.

In this paper, a novel spatial domain image watermarking scheme is proposed based on the pseudorandom noise (PN) sequence and non-linear chaotic map. The basic idea is to first generate a chaotic sequence considering the length of watermark and an initial seed. This sequence is then used to construct a feature vector using modular arithmetic. Now, this feature vector is used to generate PN sequence followed by the generation of a feature image by stacking PN sequence into an array. For each pixel in watermark, a feature image is constructed using the circular shift on the obtained feature image. Obtained feature images are then used for embedding and extraction of the watermark. The reverse process is applied for validation of the watermark. Finally, the attack analysis demonstrate the efficiency of proposed scheme against a number of intentional/un-intentional attacks.

The remaining paper is organized as follows. In Sect. 2, non-linear chaotic map and PN sequence are briefly described. Next, the detailed description of the proposed watermarking scheme is introduced in Sect. 3. The detailed experimental results and discussions are given in Sect. 4 followed by the conclusions in Sect. 5.

2 Preliminaries

In this section, the basics concepts used in the development of the proposed watermarking technique are discussed. These are as follows.

2.1 Non-linear Chaotic Map

In this work, piecewise non-linear chaotic map is used in order to generate a feature vector. A piece-wise non-linear map (PWNLCM) $\mathcal{G} : [0, 1] \rightarrow [0, 1]$ can be defined as [11]

$$\mathcal{G}(y_{k+1}) = \begin{cases} \left(\frac{1}{J_{l+1}-J_l} + b_l\right)(y_k - b_l) - \frac{a_l}{J_{l+1}-J_l}(y_k - b_l)^2, & \text{if } y_k \in [J_l, J_{l+1}) \\ 0, & \text{if } y_k = 0.5 \\ \mathcal{G}(y_k - 0.05), & \text{if } y_k \in (0.5, 1] \end{cases} \tag{1}$$

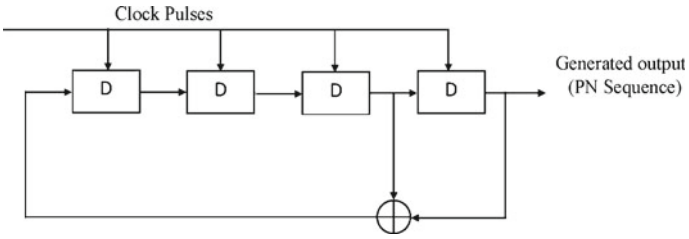


Fig. 1 A 4 bit LFSR structure

where $y_k \in [0, 1], 0 = J_0 < J_1 < \cdots < J_{m+1} = 0.5$ and $b_l \in (-1, 0) \cup (0, 1)$ is the tuning parameter for the lth interval sequence satisfying

$$\sum_{l=0}^{n-1} (J_{l+1} - J_l) b_l = 0 \tag{2}$$

2.2 Pseudo-Noise(PN)-Sequences

The PN-sequence is uniformly distributed combination of 0s and 1s which is defined over the binary field [12]. The sequence seems to be random means that the sequence is random within a period of length n which repeat again and again in purely deterministic manner and hence it preserve the periodic phenomena. The easiest way to generate PN sequence is to use a linear feedback shift register (LFSR) circuit. A LFSR with n shift register, the tap is placed between output and one of the shift register, where taping defines the XOR-operation of output and pre-defined shift registers. The output bit of the XOR operation is feedback into left most shift register and output is shifted at every clock pulse. The output of LFSR produce periodic PN sequence with finite range set. Figure 1 gives a 4-bit LFSR structure giving PN sequence using 4-shift registers.

3 Proposed Watermarking Framework

In this section, a robust logo watermarking framework, which explores the characteristics of the piece-wise non-linear map and PN-sequence, is proposed. Let us consider H be the gray-scale host and W be the binary watermark images of size $M \times N$ and $m \times n$ respectively. The proposed framework can be formalized as follows.

3.1 Watermark Embedding Process

1. Generate a chaotic sequence $K = \{k_j \in \{0,1\} | \ j = 1 \ldots L\}$ of length $L = m \times n$ by iterating piece-wise non-linear map with A_{key} as the initial seed.
2. Obtain a feature vector K_f from K considering the binary modulo operation as follows
$$K_f = \lfloor (K * 2^{16}) \bmod 2 \rfloor \tag{3}$$
3. Considering the feature vector K_f, generate a PN-sequence S as described in Sect. 2.2.
4. Construct an array (F) of size $m \times n$ by stacking S into it.
5. Feature images $\{F_i | i = 1, 2, \ldots, L\}$ are then obtained by applying circular shift on F.
6. Obtain a transformed feature image using the watermark (W) as follows.
$$F' = \sum_{i=1}^{L} W(x,y) * F_i = \begin{cases} \sum\limits_i F_i, & \text{if } W(x,y) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$
7. Embed the transformed image into the host image to get watermarked image H_W as follows.
$$H_W = H + \beta * F' \tag{5}$$
where β gives the strength to watermark embedding.

3.2 Watermark Extraction Process

The main objective of this sub-section is to extract an approximate watermark from possibly attacked/distorted watermarked image for the verification/authentication purposes. The complete extraction process can be formalized as follows.

1. Adopting steps 1–5 of embedding process, generate feature matrix F_i considering A_{key} as the initial seed for the map.
2. Obtain image H_W^f by applying the high pass filter on watermarked image H_W.
3. Obtain a vector C comprising the correlation coefficient between the feature images and the high pass filtered watermarked image H_W^f.
4. Partition the vector C into two different segments C_1 and C_2 of positive and negative correlation values.
5. Construct a threshold value T as follows.
$$T = \frac{1}{256} \frac{w_1 * \sum\limits_{i=1}^{l_1} C_1(i) + w_2 * \sum\limits_{j=1}^{l_2} C_2(j)}{w_1 + w_2} \tag{6}$$

where w_1 and w_2 are the weights for the negative and positive correlation coefficients whereas l_1 and l_2 be the lengths of segments C_1 and C_2 respectively.

6. Construct a binary sequence as follows.

$$b_{ext}(i) = \begin{cases} 1, & \text{if } C(i) \geq T \\ 0, & otherwise \end{cases} \tag{7}$$

7. Extracted watermark W_{ext} is finally formed by stacking binary sequence (b_{ext}) into an array.

4 Results and Discussions

The efficiency of the proposed watermarking scheme is evaluated on the basis of extensive experiments on different images under the platform of MATLAB. Four different gray-scale images of size 512×512 namely Pirate, Mandrill, Peppers and Elaine are considered as the host image. In contrast, four binary logos of size 32×32 having structure of Focus, Plus, Pyramid and alphabet T are used as watermarks. The logos having Focus, Plus, Pyramid and alphabet T structure are embedded into Pirate, Mandrill, Pepper and Elaine images respectively. All the host images and the corresponding watermark images are depicted in Fig. 2. The imperceptibility of the proposed framework is measured by peak signal-to-noise ratio (PSNR) and it was found to be 32.2598, 31.9947, 32.3687 and 32.7498 respectively for the watermarked Pirate, Mandrill, Peppers and Elaine images. These values essentially validates the fact that the proposed scheme has no perceptual degradation as per the human visual system. Therefore, the proposed algorithm is perceptually robust. All the results are obtained by considering the key as 0.75.

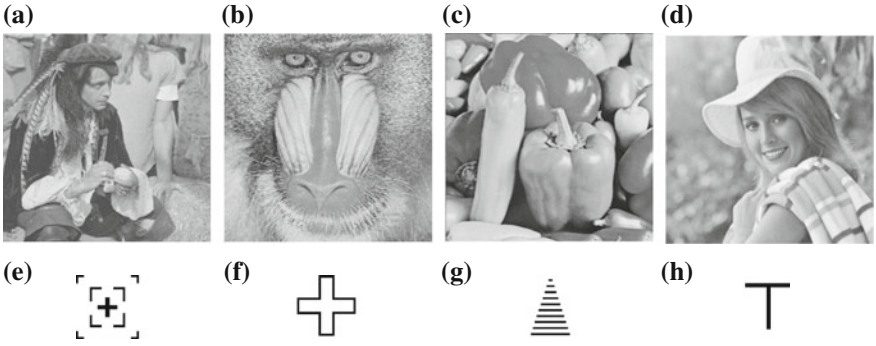


Fig. 2 Experimental images: **a–d** host, **e–h** original watermark images

The robustness of the proposed scheme is evaluated against a number of general attacks (Gaussian blurring, salt and pepper noise, JPEG-compression, histogram-equalization, sharpening, contrast-adjustment and gamma-correction) and geometric attacks (resizing and cropping). The watermark is extracted from the attacked watermarked image and is checked for the similarity with the initial one. The watermarked and corresponding extracted watermark images can be seen in Fig. 3 for visual verification. The presence of the watermark can be quantify by the linear relationship between the original and extracted watermark and is defined by correlation coefficients (ρ), which is given by the following equation.

$$r(\omega, \varpi) = \frac{\sum_{r,s}(\omega_{r,s} - \mu_\omega)(\varpi_{r,s} - \mu_\varpi)}{\sqrt{\sum_{r,s}(\omega_{r,s} - \mu_\omega)^2}\sqrt{\sum_{r,s}(\varpi_{r,s} - \mu_\varpi)^2}} \tag{8}$$

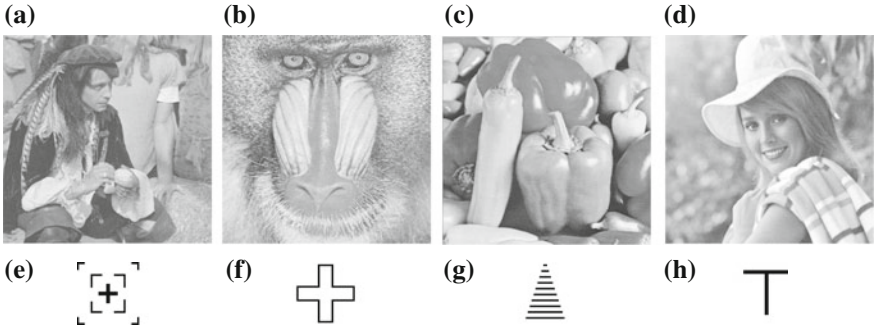


Fig. 3 **a–d** Watermarked host, **e–h** extracted watermark images

Table 1 Correlation coefficients of extracted watermarks after attack analysis

Attacks/distortions	Pirate	Mandrill	Elaine	Pepper
Without attack	1	1	1	1
Gaussian blurring (3 × 3)	0.9161	0.7990	0.9588	0.9219
Salt and pepper noise (10 %)	0.9170	0.8200	0.8350	0.8505
JPEG compression (70 %)	0.8945	0.7673	0.8833	0.8932
Resizing (512 → 256 → 512)	0.6759	0.3998	0.8497	0.8508
Cropping (50 % area)	0.9797	0.9903	0.9365	0.9938
Histogram equalization	1	1	1	1
Sharpening (increased by 50 %)	0.9949	0.9805	0.9812	0.9875
Contrast adjustment (decreased by 100 %)	1	1	1	1
Gamma correction ($\gamma = 5$)	1	1	1	1

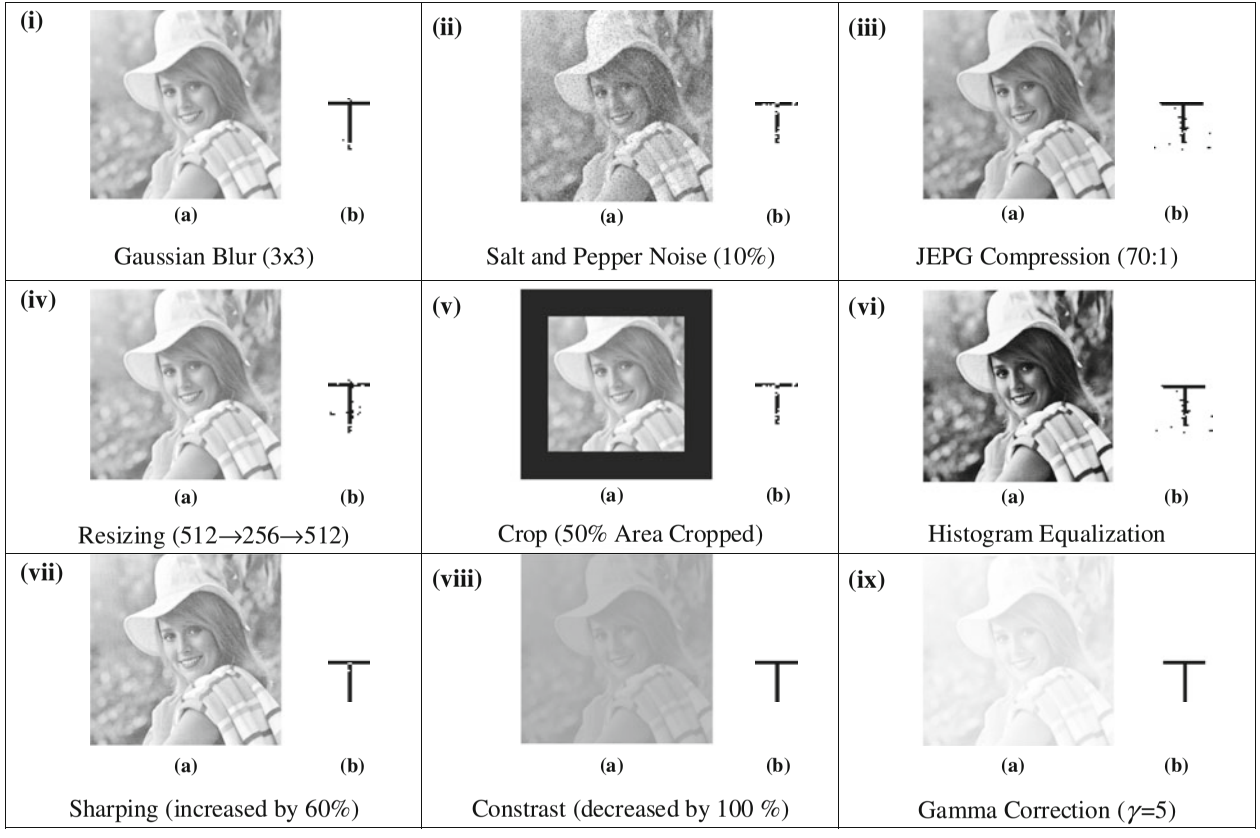


Fig. 4 **a** Attacked watermarked, **b** extracted watermark images after various attacks

where ω and ϖ denotes the original and extracted watermark images while μ_{ω} and μ_{ϖ} are their respective mean. The principle range for ρ is from -1 to 1. The value of ρ close to 1 indicates the better similarity between the images ω and ϖ. Here, the results are shown visually only for Elaine image because it attains the maximum PSNR. In contrast, correlation coefficients values are given for all images and are tabulated in Table 1.

The transmission of digital image over the insecure network introduce degradation in image quality and increase the possibility of different kind of attacks. These attacks generally distort the statistical properties of the image. As a result, these distortions effects the quality of the extracted watermark. Therefore, every watermarking scheme should robust against these distortions. In general, these also depend upon robustness of the scheme. Hence robustness is the key factor which represent the efficiency of the technique.

The most common distortions in digital imaging is blurring and noise additions. For blurring, Gaussian blur with filter size 3×3 is considered whereas 10 % slat and pepper noise is added in the watermarked image before the watermark extraction. JPEG compression, a lossy encoding schemes to store the data efficiently, is most another common distortion in real life applications. Therefore, the proposed scheme is further investigated for JPEG Compression having compression ratio 70:1. Performance of the proposed scheme is also evaluated for most common geometric distortions such as cropping and resizing. There are several ways to crop an image. Essentially, it is done by a function which maps a region of the image to zero. The 50 % information of the watermarked images is mapped to zero for cropping. In contrast, the effect of resizing is created by reducing the size of watermarked image to 256×256 and again up-sizing it to the original size. The proposed scheme is further examined for general image processing attacks like histogram-equalization, sharpening, contrast-adjustment and gamma-correction. For sharpening and contrast-adjustment, the sharpness and contrast of the watermarked image are increased/decreased by 60 %/100 % respectively where as for gamma-correction the watermarked image is corrected using a gamma of 5. The visual results for all these distortions are depicted in Fig. 4.

5 Conclusion

In this paper, a novel spatial domain watermarking scheme is proposed where a visually meaningful binary image is embedded in the host image. The core idea is to first obtain a feature vector based on non-linear map and then use it to generate a PN sequence. This sequence is used to obtain feature images using the circular shift followed by the embedding of watermark using feature images. In the proposed framework, the feature images are the integral part of security since the watermarked cannot be extracted without their exact knowledge. Finally, a detailed investigation is performed to validate the efficacy of the proposed scheme.

Acknowledgements This research was supported by the Science and Engineering Research Board, DST, India.

References

1. Katzenbeisser S. and Petitcolas, F.A.P., Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston (2002).
2. Cox I. J., Miller M. and Bloom J., Digital Watermarking, Morgan Kaufmann (2002).
3. Mandhani N. and Kak S., Watermarking Using Decimal Sequences, Cryptologia 29(2005) 50–58.
4. Langelaar G., Setyawan I., and Lagendijk R.L., Watermarking Digital Image and Video Data, IEEE Signal Processing Magazine, 17 (2009) 20–43.
5. Wenyin Z. and Shih F.Y., Semi-fragile spatial watermarking based on local binary pattern operators, Opt. Commun., 284 (2011) 3904–3912.
6. Botta M., Cavagnino D. and Pomponiu V., A modular framework for color image watermarking, Signal Processing, 119 (2016) 102–114.
7. Ruanaidh O. and Pun T., Rotation, scale and translation invariant spread spectrumdigital image watermarking, Signal Processing, 66 (1998) 303–317.
8. Aslantas V., Ozer S. and Ozturk S., Improving the performance of DCT-based fragilewatermarking using intelligent optimization algorithms, Opt. Commun., 282 (2009) 2806–2817.
9. Yu D., Sattar F. and Binkat B., Multiresolution fragile watermarking using com-plex chirp signals for content authentication, Pattern Recognition, 39 (2006) 935–952.
10. Bhatnagar G., Wu Q.M.J. and Atrey P.K., Robust Logo Watermarking using Biometrics Inspired key Generation, Expert Systems with Applications, 41 (2014) 4563–4578.
11. Tao S., Ruli W. and Yixun Y., Generating Binary Bernoulli Sequences Based on a Class of Even-Symmetric Chaotic Maps, IEEE Trans. on Communications, 49 (2001) 620–623.
12. Khojasteh M.J., Shoreh M.H. and Salehi J.A., Circulant Matrix Representation of PN-sequences with Ideal Autocorrelation Property, Iran Workshop on Communication and Information Theory (IWCIT), (2015) 1–5.

Deep Gesture: Static Hand Gesture Recognition Using CNN

Aparna Mohanty, Sai Saketh Rambhatla and Rajiv Ranjan Sahay

Abstract Hand gestures are an integral part of communication. In several scenarios hand gestures play a vital role by virtue of them being the only means of communication. For example hand signals by a traffic policeman, news reader on TV gesturing news for the deaf, signalling in airport for navigating aircrafts, playing games etc. So, there is a need for robust hand pose recognition (HPR) which can find utility in such applications. The existing state-of-the-art methods are challenged due to clutter in the background. We propose a deep learning framework to recognise hand gestures robustly. Specifically we propose a convolutional neural network (CNN) to identify hand postures despite variation in hand sizes, spatial location in the image and clutter in the background. The advantage of our method is that there is no need for feature extraction. Without explicitly segmenting foreground the proposed CNN learns to recognise the hand pose even in presence of complex, varying background or illumination. We provide experimental results demonstrating superior performance of the proposed algorithm on state-of-the-art datasets.

Keywords Hand gesture recognition · Deep learning · Convolutional neural network

A. Mohanty (✉) · S.S. Rambhatla · R.R. Sahay
Department of Electrical Engineering,
Indian Institute of Technology Kharagpur, Kharagpur, India
e-mail: aparnamhnty@gmail.com

S.S. Rambhatla
e-mail: sakethu.rambhatla@gmail.com

R.R. Sahay
e-mail: rajivsahay@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_41

1 Introduction

The human visual system (HVS) is very effective in rapidly recognising a large number of diverse objects in cluttered background effortlessly. Computer vision researchers aim to emulate this aspect of HVS by estimating saliency [1, 2] of different parts of the visual stimuli in conjunction with machine learning algorithms. Hand gesture recognition is an active area of research in the vision community because of its wide range of applications in areas like sign language recognition, human computer interaction (HCI), virtual reality, and human robot interaction etc.

Expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body [3] are called *gestures*. Hand gestures are either static or dynamic. In this work we focus on the recognition of static hand gestures. There are relatively few hand posture databases available in the literature, and so we have shown results on three databases [4–6].

2 Prior Work

In the literature basically two approaches are proposed for recognition of hand gestures. The first approach [7] makes use of external hardware such as gloves, magnetic sensors, acoustic trackers, or inertial trackers. Theses external devices are cumbersome and may make the person using them uncomfortable. Vision based approaches do not need any physical apparatus for capturing information about human hand gestures and have gained popularity in recent times. However, these methods are challenged when the image background has clutter or is complex.

Skin detection is the initial step to detect and recognise human hand postures [5]. However it is clear that different ethnicities have different skin features, which make skin-based hand detection difficult. In addition, skin models are sensitive to illumination changes. Lastly, if some skin like regions exist in the background then it also leads to erroneous results. Recently, Pisharady et al. [5] proposed a skin based method to detect hand areas within images. Furthermore in [5] an SVM classifier was used to identify the hand gestures.

There are various methods proposed to detect and recognise hand postures which do not use skin detection. Based on object recognition method proposed by Viola and Jones [8], Kolsch and Turk [9] recognise hand postures for the recognition of six hand postures. Bretzner et al. [10] use a hierarchical model, which consists of palm and the fingers and a multi-scale colour feature to represent hand shape. Ong and Bowden [11] proposed a boosted classifier tree-based method for hand detection and recognition. View-independent recognition of hand postures is demonstrated in [12]. Kim and Fellner [13] detect fingertip locations over dark environment. For gesture recognition Chang et al. [14] recognise hand postures under a simple dark environment. Triesch and Von der Malsburg [15] address the complex background problem in hand posture recognition using elastic graph matching. A self-organizing neural network is used by Flores et al. [16] in which the network topology determines

hand postures. In this work, we overcome the limitations of methods using hand crafted features.

Existing methods can recognize hand gestures in uniform/complex backgrounds with clutter only after elaborate pre-processing [4–6, 17]. In this work we propose a deep learning based method using convolutional neural networks (CNN) to identify static hand gestures both in uniform and cluttered backgrounds. We show the superiority of our CNN-based framework over approaches using handcrafted features [5, 17]. We do not need to use elaborate, complex feature extraction algorithms for hand gesture images. We show that the CNN-based framework is much simpler and more accurate on the challenging NUS-II dataset [5] containing clutter in the background. In contrast to the proposed method, earlier works on hand gesture recognition e.g. [18, 19] using CNN have have not reported results on the challenging datasets such as Marcel dataset [4] and NUS-II clutter dataset [5].

3 Hand Gesture Datasets

There are very few publicly available hand gestures datasets with clutter in the background. Therefore, we choose to present the results of our CNN-based framework on [4, 6] and NUS-II [5], which we describe below. The hand gesture database proposed by Jochen Triesch [6] has 10 different classes in uniform light, dark and complex background. In this paper we focus our work on the uniform dark background dataset to show the effectiveness of proposed approach on both uniform as well as complex background. A snapshot of the dataset [6] is shown in Fig. 1. The dataset has 24 images per class and a total of 10 classes.

The hand gesture database proposed by Marcel [20] has 6 different classes separated into training and testing data.The testing data also has separate data for the uniform and complex case. There are 4872 training images in total and 277 testing images for the complex background. A snapshot of the Marcel dataset is shown in Fig. 2

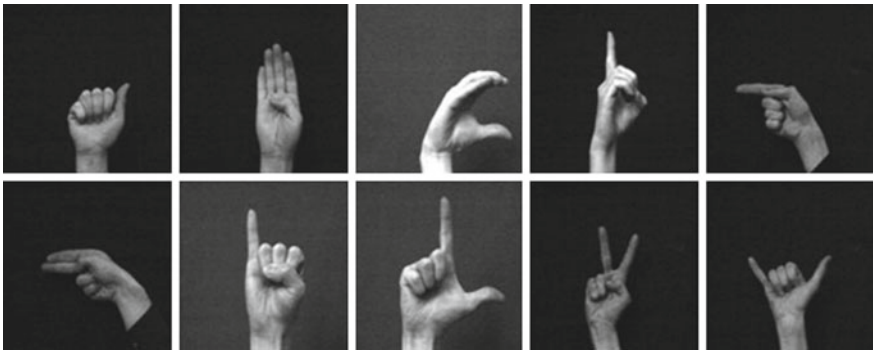


Fig. 1 Images from Triesch dataset [6] dataset with uniform dark background

Fig. 2 Marcel dataset [4]

Fig. 3 Sample images from NUS-II [5] dataset with clutter

Recently, a challenging static hand gesture dataset has been proposed in [5]. The NUS-II [5] dataset has 10 hand postures in complex natural backgrounds with varying hand shapes, sizes and ethnicities.The postures were collected using 40 subjects comprising of both male and female members in age group of 22 to 56 years from various ethnicities. The subjects showed each pose 5 times. The NUS-II [5] dataset has 3 subsets. The first subset is a dataset of hand gestures in presence of clutter consisting of 2000 hand posture color images (40 subjects, 10 classes, 5 images per class per subject). Each image is of size 160×120 pixels with complex backgrounds. Some sample images from the NUS-II dataset with clutter are shown in Fig. 3. In this paper we focus our work on the subset of NUS-II dataset with only non-human clutter in the background.

4 Deep Learning Framework: Convolutional Neural Network

Convolutional neural nets originally proposed by LeCun [21] have been shown to be accurate and versatile for several challenging real-world machine learning problems [21, 22]. According to LeCun [21, 23], CNNs can be effectively trained to recognize

objects directly from their images with robustness to scale, shape, angle, noise etc. This motivates us to use CNNs in our problem since in real-world scenarios hand gesture data will be affected by such variations.

4.1 Architecture

The general architecture of the proposed CNN is shown in Fig. 4. Apart from the input and the output layers, it consists of two convolution and two pooling layers. The input is a 32×32 pixels image of hand gesture.

As shown in Fig. 4, the input image of 32×32 pixels is convolved with 10 filter maps of size 5×5 to produce 10 output maps of 28×28 in layer 1. These output maps may be operated upon with a linear or non-linear activation function followed by an optional dropout layer. The output convolutional maps are downsampled with max-pooling of 2×2 regions to yield 10 output maps of 14×14 in layer 2. The 10 output maps of layer 2 are convolved with each of the 20 kernels of size 5×5 to obtain 20 maps of size 10×10. These maps are further downsampled by a factor of 2 by max-pooling to produce 20 output maps of size 5×5 of layer 4.

The output maps from layer 4 are concatenated to form a single vector during training and fed to the next layer. The quantity of neurons in the final output layer depends upon the number of classes in the image database. The output neurons are fully connected by weights with the previous layer. Akin to the neurons in the convolutional layer, the responses of the output layer neurons are also modulated by a non-linear sigmoid function to produce the resultant score for each class. [22].

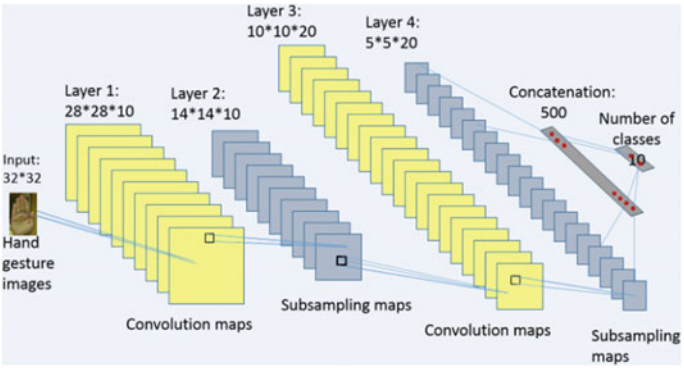


Fig. 4 Architecture of the proposed CNN model used for hand gesture classification

4.2 Data Augmentation

It has been shown in [22] that data augmentation boosts the performance of CNNs. The dataset of Triesch et al. [6] has 240 images each captured in three different conditions such as uniform light, dark or complex background. Therefore we augmented these images by cropping decrementing the size by 1 pixel successively along horizontal and vertical direction 5 times and then resizing to original size to obtain 1200 images in each category. For Marcel [4] dataset there are 4872 training images and 277 testing images with complex background. We combined them to have a total of 5149 images. We did not augment this database since the number of images was substantial. The NUS-II dataset has images with two different backgrounds i.e. with inanimate objects cluttering the background or with humans in the backdrop. The NUS-II [5] dataset with inanimate objects comprising the cluttered background has 10 hand postures with 200 images per class. We separated 200 images per class into a set of 120 training images and a set of 80 test images. Then we augmented the training and test data by 5 times to obtain 6000 training images and 4000 test images.

4.3 Activation Function

There is a layer of linear or non-linear activation function following the convolutional layer in a CNN depicted in Fig. 4. It has been shown in literature [22] that ReLu (rectified linear unit) non-linear activation function aids in making the network learn faster. Hence, we tested the impact of ReLu activation function on the performance of the system as shown in Table 2. The impact of using a non-linear activation function along with dropout is shown in Fig. 8. As can be seen in Fig. 8b the training error falls and stabilises at 800 epochs for the NUS-II dataset with clutter in the background.

4.4 Dropout

Deep networks have large number of parameters which makes the network train slower as well as difficult to tune. Hence, dropout is used to make the network train faster and avoid over-fitting by randomly dropping some nodes and their connections [24]. The impact of overfitting on the performance of the network for the Triesch dataset [6] with dark background and NUS-II dataset [5] with clutter in the background is shown in Table 3. We used a dropout of 50 % in the two dropout layers that follows the ReLu layer present after the convolutional layer in Fig. 4.

5 Experimental Results

We have trained a CNN with images of hand gestures from three standard datasets [4–6] considering the cases of both plain/uniform and complex backgrounds. The architecture of the CNN shown in Fig. 4 remains invariant for all our experiments. The weights of the proposed CNN are trained by the conventional back-propagation method using the package in [25]. The total number of learnable parameters in the proposed CNN architecture is 6282.

5.1 Training Phase

Firstly we trained the proposed CNN model on the original datasets of [4–6] without any data augmentation. The details of the training procedure such as values of the learning rate α, batch size, number of epochs are provided in Table 1. The variation of the mean square error with epochs while training the CNN is plotted in Fig. 5a,

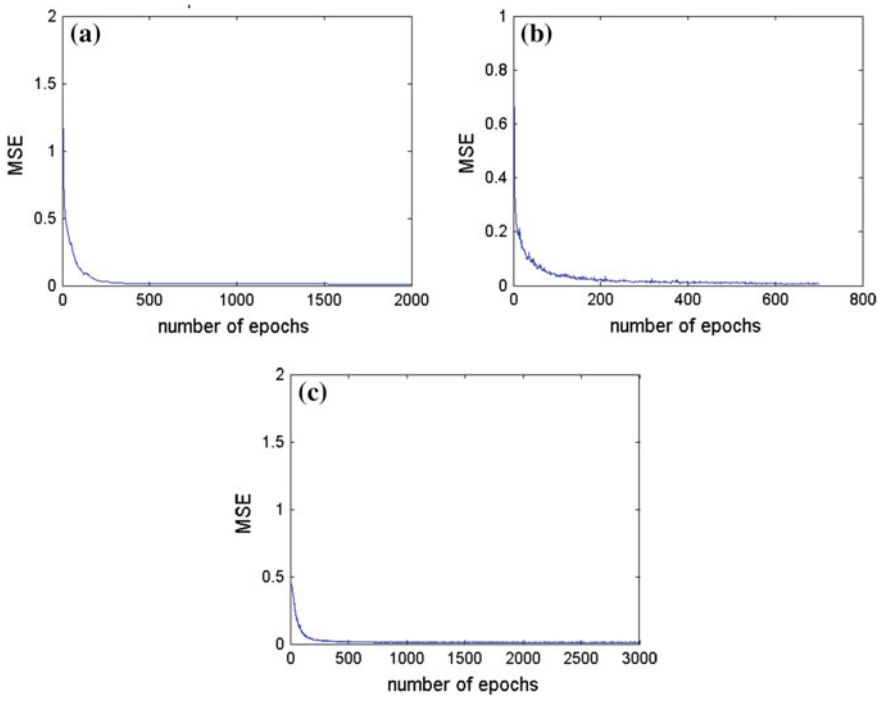


Fig. 5 The confusion matrix of un-augmented NUS-II dataset with inanimate clutter in the background with ReLu and dropout. Since we consider 80 test images per class so the sum of all elements of a row is 80

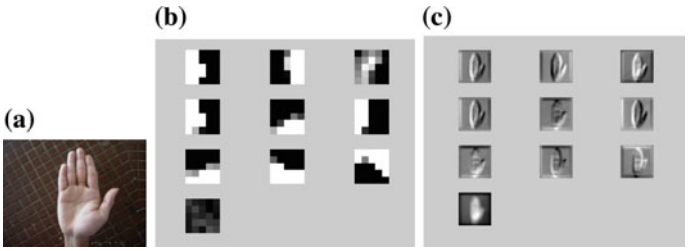


Fig. 6 Training phase: Variation of Mean squared error (MSE) for **a** the CNN trained on the Triesch dark [6] dataset. **b** the CNN trained on the Marcel [4] dataset. **c** the NUS-II [5] dataset containing clutter without augmentation

b, c for the three cases of Triesch dataset with dark background [6], Marcel dataset with complex background [4], NUS-II [5] with clutter. In Fig. 6, we show image visualizations of the 10 filter kernels obtained after training the proposed CNN model on a sample input image from NUS-II [5] (with inanimate clutter) dataset. From Fig. 6c we observe that the trained filter kernels automatically extract appropriate features from the input images emphasizing edges and discontinuities. Note that we tune the values of the parameters for optimal performance in all the experiments conducted on all datasets [4–6].

Table 1 Performance of the proposed CNN model on the Triesch and Von der Malsburg [6], Marcel [4] database and the NUS [5] datasets with sigmoid activation function without dropout and augmentation

Data	No. of classes	Training set	Testing set	α	Batch size	Epochs	MSE result	Without ReLu and dropout (%)
Triesch et al. [6] (dark background)	10	200	40	0.5	10	2000	0.02	77.50
Marcel [4] (complex case)	6	3608	1541	0.2	5	500	0.014	85.98
NUS-II [5] (with clutter)	10	1200	800	0.3	5	3000	0.0217	84.75

5.2 Testing Phase

As shown in Table 1, for the un-augmented dataset of Triesch et al. [6] containing hand gesture images in the dark background, using parameters $\alpha = 0.5$, batch size = 10 even after 2000 epochs the accuracy of the proposed CNN model was only 77.50 %.

The efficacy of CNN for images of hand gestures with complex backgrounds on a large dataset is demonstrated by the accuracy obtained on the test data in [4]. We obtained an accuracy of 85.59 % with 222 out of 1541 images being misclassified on this challenging dataset. The accuracy obtained is higher than state-of-the art result of 76.10 % reported on the Marcel dataset [4]. Similarly good accuracy was seen for the NUS-II dataset for the cluttered background with values as high as 84.75 % as shown in Table 1. The performance further improved by using ReLu as the non-linear activation function and using dropout to avoid overfitting. A dropout of 50 % was used in two layers following the ReLu layer that follows the convolutional layer in the CNN architecture depicted in Fig. 4. By using ReLu and dropout on the un-augmented datasets of Triesch [6] and NUS-II with cluttered background [5] an improved accuracy of 82.5 and 89.1 % respectively was achieved as shown in Table 2. We observed that the accuracy obtained using ReLu and dropout was higher than the accuracy obtained using a sigmoidal activation function, without dropout as reported in Table 1. The confusion matrix for the un-augmented NUS-II dataset with clutter in the background [5] having 80 images per class for testing is shown in Fig. 7. For the Triesch dataset [6] the probability that the predicted class of a test image is among the top 2 predictions made for the image is 93.6 %. The decrease in training and validation error for the un-augmented NUS-II dataset with clutter in the background, obtained using MatConvNet [26] is shown in Fig. 8. The testing accuracy further improved to 88.5 % by augmenting the datasets of Triesch [6] by augmenting it five times by successively cropping by one pixel and resizing it back to the original size. The performance on the augmented data is reported in Table 3.

63	2	2	7	0	0	1	0	3	2
2	68	7	1	0	0	1	0	0	1
0	3	67	0	1	2	0	4	3	0
3	1	2	65	2	1	1	1	2	2
0	0	0	0	75	0	1	1	2	1
0	0	0	1	1	73	1	0	4	0
0	1	0	1	4	2	71	0	0	1
2	0	4	1	2	1	0	68	0	2
0	0	1	0	1	4	0	0	73	1
0	0	0	0	1	1	0	0	1	77

Fig. 7 **a** Sample input image from NUS-II [5] dataset with clutter. **b** Filter kernels of the first layer of the CNN model. **c** Output of the first layer of the CNN model

Table 2 Performance of the proposed CNN model on the Triesch et al. [6] and the NUS [5] datasets with ReLu activation function and dropout without augmentation

Data	No. of classes	Training set	Testing set	α	Batch size	Epochs	MSE result	ReLu and dropout (%)
Triesch et al. (dark background) [6]	10	200	40	9×10^{-7}	10	500	0.005	82.5
NUS-II [5] with clutter	10	1200	800	5×10^{-6}(upto500) 1×10^{-6}(upto2000)	10	2000	0.015	89.1

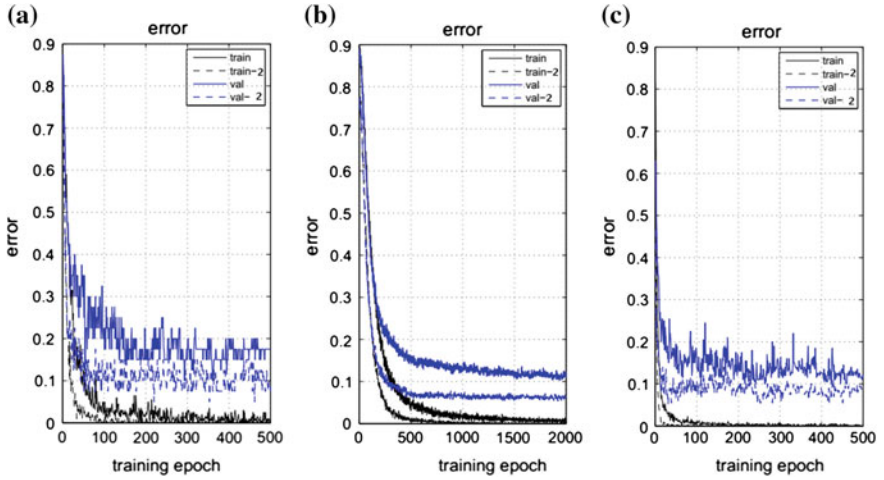


Fig. 8 The variation of training and validation error **a** for the un-augmented Triesch dataset [6]. **b** for un-augmented NUS-II [5] dataset with clutter in the background. **c** for the augmented Triesch dataset [6]

Table 3 Performance of the proposed CNN model on the Triesch et al. [6] dataset with data augmentation and using dropout and ReLu as the activation function

Data	Classes	Training set	Testing set	α	Batch size	Epochs	MSE	State of the art result	Proposed approach
Triesch and Von der Malsburg [6] (dark background)	10	1000	200	5×10^{-6}	10	500	0.001	95.83 %	88.5 %

6 Conclusions

The proposed CNN model has been demonstrated to be able to recognize hand gestures to a high degree of accuracy on the challenging popular datasets [4, 5]. Unlike other state-of-the-art methods we do not need to either segment the hand in the input image or extract features explicitly. In fact, with a simple architecture of the proposed CNN model, we are able to obtain superior accuracy on the dataset in [4] (complex background) and obtain comparable performance on the NUS-II dataset [5]. We believe that the proposed method can find application in areas such as sign language recognition and HCI. In future we plan to extend this work to handle dynamic hand gestures also.

References

1. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.
2. A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.
3. S. Mitra and T. Acharya, "Gesture recognition: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 3, pp. 311–324, 2007.
4. S. Marcel, "Hand posture recognition in a body-face centered space," in *CHI '99 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '99. New York, NY, USA: ACM, 1999, pp. 302–303. [Online]. Available: http://doi.acm.org/10.1145/632716.632901.
5. P. K. Pisharady, P. Vadakkepat, and A. P. Loh, "Attention based detection and recognition of hand postures against complex backgrounds," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 403–419, 2013.
6. J. Triesch and C. Von Der Malsburg, "Robust classification of hand postures against complex backgrounds," in *fg*. IEEE, 1996, p. 170.
7. T. S. Huang, Y. Wu, and J. Lin, "3d model-based visual hand tracking," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 905–908.
8. P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, pp. 51–52, 2001.
9. M. Kölsch and M. Turk, "Robust hand detection." in *FGR*, 2004, pp. 614–619.
10. L. Bretzner, I. Laptev, and T. Lindeberg, "Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 423–428.
11. E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 889–894.
12. Y. Wu and T. S. Huang, "View-independent recognition of hand postures," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2. IEEE, 2000, pp. 88–94.
13. H. Kim and D. W. Fellner, "Interaction with hand gesture for a back-projection wall," in *Computer Graphics International, 2004. Proceedings*. IEEE, 2004, pp. 395–402.
14. C.-C. Chang, C.-Y. Liu, and W.-K. Tai, "Feature alignment approach for hand posture recognition based on curvature scale space," *Neurocomputing*, vol. 71, no. 10, pp. 1947–1953, 2008.
15. J. Triesch and C. Von Der Malsburg, "A system for person-independent hand posture recognition against complex backgrounds," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 1449–1453, 2001.
16. F. Flórez, J. M. García, J. García, and A. Hernández, "Hand gesture recognition following the dynamics of a topology-preserving network," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*. IEEE, 2002, pp. 318–323.
17. P. P. Kumar, P. Vadakkepat, and A. P. Loh, "Hand posture and face recognition using a fuzzy-rough approach," *International Journal of Humanoid Robotics*, vol. 7, no. 03, pp. 331–356, 2010.
18. P. Barros, S. Magg, C. Weber, and S. Wermter, "A multichannel convolutional neural network for hand posture recognition," in *Artificial Neural Networks and Machine Learning–ICANN 2014*. Springer, 2014, pp. 403–410.
19. J. Nagi, F. Ducatelle, G. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, L. M. Gambardella *et al.*, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 342–347.
20. S. Marcel and O. Bernier, "Hand posture recognition in a body-face centered space," in *Gesture-Based Communication in Human-Computer Interaction*. Springer, 1999, pp. 97–100.

21. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.
22. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
23. Y. Lecun, F. J. Huang, and L. Bottou, "Learning methods for generic object recognition with invariance to pose and lighting," in *CVPR*. IEEE Press, 2004.
24. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
25. R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, 2012. [Online]. Available: https://github.com/rasmusbergpalm/DeepLearnToolbox.
26. A. Vedaldi and K. Lenc, "Matconvnet-convolutional neural networks for matlab," *arXiv preprint* arXiv:1412.4564, 2014.

A Redefined Codebook Model for Dynamic Backgrounds

Vishakha Sharma, Neeta Nain and Tapas Badal

Abstract Dynamic background updation is one of the major challenging situation in moving object detection, where we do not have a fix reference background model. The background model maintained needs to be updated as and when moving objects add and leave the background. This paper proposes a redefined codebook model which aims at eliminating the ghost regions left behind when a non-permanent background object starts to move. The background codewords which were routinely deleted from the set of codewords in codebook model are retained in this method while deleting the foreground codewords leading to ghost elimination. This method also reduces memory requirements significantly without effecting object detection, as only the foreground codewords are deleted and not background. The method has been tested for robust detection on various videos with multiple and different kinds of moving backgrounds. Compared to existing multimode modeling techniques our algorithm eliminates the ghost regions left behind when non permanent background objects starts to move. For performance evaluation, we have used similarity measure on video sequences having dynamic backgrounds and compared with three widely used background subtraction algorithms.

Keywords Motion analysis · Background subtraction · Object detection · Video surveillance

V. Sharma (✉) · N. Nain · T. Badal
Computer Science and Engineering Department,
Malaviya National Institute of Technology, Jaipur, India
e-mail: 2014pcp5168@mnit.ac.in

N. Nain
e-mail: nnain.cse@mnit.ac.in

T. Badal
e-mail: tapasbadal@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_42

1 Introduction

Moving object detection is one of the key step to video surveillance. Objects in motion are detected from the video and is then used for tracking and activity analysis of the scene. It has found use in areas such as security, safety, entertainment and efficiency improvement applications. Typically motion detection is the primary processing step in Activity recognition which is used in traffic analysis, restricted vehicle movements, vehicle parking slots, multi-object interaction, etc.

Background subtraction methods are widely used techniques for moving object detection. It is based on the difference of the current frame from the background model. A background model is maintained which is actually a representation of the background during the training frames. The difference of each incoming frame from this reference model gives the objects in motion in the video sequence.

In static background situations, where we have a fix background, it is relatively easy to segment moving objects in the scene. Many methods have been developed which are able to successfully extract foreground objects in video sequences. However, many techniques fail in major challenging situations such as non-static backgrounds, waving trees, sleeping and walking person, etc. as illustrated in [1].

In this paper, we propose an improved multi-layered codebook method to deal with dynamic backgrounds. It removes the false positives which were detected as ghost regions due to the area left behind when moving objects dissolved into background and starts moving again. We also place a memory limit on the pixel codewords length to avoid infinite unnecessary codewords in the model without affecting the detection rate, thus improving the speed.

Section 2 discuss widely followed methods for background subtraction available in literature. Section 3 gives an introduction to the basic codebook model used in our algorithm. Section 4 elaborates the problem of ghost region encountered in dynamic backgrounds. The proposed approach is enumerated in Sect. 5 and detailed algorithmic description is presented in Sect. 6. Section 7 evaluates the four object detection algorithms using similarity measure. Finally, Sect. 8 concludes the paper.

2 Related Work

Many methods are proposed in literature [2–6] where in the training time we model the background, perform background subtraction and then update the background model each time. Wren et al. [2] proposed an approach where the background is represented as a Gaussian with single modality. For dynamic backgrounds, the approach was extended to include not just a single Gaussian, but to represent background as a set of multiple gaussians by Grimson [3]. It also assigned weights to each gaussian, and adaptively learnt them. Many researchers [4, 5] has also followed the above

approach. But it has many limitations as it could not handle the shadows completely. Also it needs to estimate the parameters for the Gaussian in uncontrolled environments having varying illumination conditions. Apart from parametric representation, non-parametric approach has also been used by many researchers. The background is modelled using kernel density estimation [6], where probability density function (*pdf*) is estimated for the background for detecting background and foreground regions, but is not be appropriate for all dynamic environments. A codebook model was proposed by Kim et al. [7] for modelling the background. The Codeword contains the complete information of the background including the color and intensity information of the pixels.

The method proposed in this paper deals in removing the ghost regions produced due to dynamic backgrounds when non-permanent background suddenly starts to move. It also improves the memory requirement as we place a memory limit on each pixel codeword model length without affecting the detection results.

3 Codebook Model

The Codebook Model as illustrated in [7] uses a codebook for each pixel, which is a set of codewords. A codeword is a quantized representation of background values. A pixel is classified as foreground or background based on color distance metric and brightness bounds. A codebook for a pixel is represented as:

$$C = \{c_1, c_2 \ldots c_L\}$$

where, L is the total number of codewords.

Each codeword C_i consist of a *RGB* vector and 6-tuples as follows:

$$v_i = \bar{R}_i, \bar{G}_i, \bar{B}_i$$

$$aux_i = \langle \check{I}_i, \hat{I}_i, f_i, \lambda_i, p_i, q_i \rangle$$

where, $\check{I}_i, \hat{I}_i$ are the minimum and maximum brightness of all pixels for the codeword in the training period, f_i is the frequency of occurrence of codeword, λ_i(**MNRL**-Maximum Negative Run Length) is the maximum interval of non-occurrence of codeword, p_i and q_i are the first and the last access time of codewords respectively. The color distance of a pixel $x_t = (R, G, B)$ to a codeword $v_i = (R_i, G_i, B_i)$ is calculated as:

$$color\_dist(x_t, v_i) = \sqrt{(R^2 + G^2 + B^2) + \frac{(R_i^2 R + G_i^2 G + B_i^2 B)}{R_i^2 + G_i^2 + B_i^2}} \quad (1)$$

and brightness $\mathcal{B}$ is defined as:

$$\mathcal{B}(I, \langle \check{I}_i, \hat{I}_i \rangle) = \begin{cases} true \text{ if}(\alpha \hat{I}_i \leq \|x_t\| \leq \min(\beta \hat{I}_i, \frac{\check{I}_i}{\alpha})), \\ false \text{ otherwise.} \end{cases} \quad (2)$$

where, $\|x_t\| = \sqrt{(R^2 + G^2 + B^2)}$, and, $\alpha < 1$ and $\beta > 1$. Now, a pixel is classified as background if it matches to any codeword of the background model based on following two conditions:

- $color\_dist(x_t, c_m) \leq \varepsilon$
- $\mathcal{B}(I, \langle \check{I}_i, \hat{I}_i \rangle) = \text{true}$

where ε is the detection threshold. Also, for layered modelling an addition to the codebook model is made which includes a cache codebook $\mathscr{H}$. It contains the set of foreground codewords that might become background thus supporting background changes.

4 Ghost Region in Dynamic Backgrounds

Major challenging issues for motion detection are dynamic backgrounds. Many background subtraction algorithms as illustrated in Sect. 2 performs well for static backgrounds, but leads to false detections for changing backgrounds. Dynamic backgrounds may occur due to moving objects which were not part of the background model, to come to halt and become part of the background itself.

In motion detection algorithms, such objects should then be treated as background and be included in the background model.

For such dynamic backgrounds, Codebook model has proposed an approach of layered modelling and detection in [7]–[9]. Apart from the original set of codebook $\mathscr{M}$ for each pixel, it introduces another cache codebook $\mathscr{H}$. $\mathscr{H}$ holds the codewords for a pixel which may become part of the background. When a codeword is not matched to the background model $\mathscr{M}$, it is added to the cache codebook $\mathscr{H}$. Now, on successive incoming pixels for next frames, the codewords are updated as illustrated in [7]. Codewords which stays for a long time in $\mathscr{H}$ are added to the codebook $\mathscr{M}$ as it is the moving object which has come to stop and is now a part of the background.

According to the algorithm in [7], when an object such as a car is parked into an area, it is dissolved into background. Later when the car leaves the parking area, a ghost region is left behind for some time frames till the model learns the background. This is because according to the Codebook model in [7], the background codewords

not accessed for a long time are deleted routinely from the set of codewords of a pixel in $\mathscr{M}$. This also deletes the codewords belonging to the actual background which were not accessed for the time when the car was parked.

For example, initially after the training time, a pixel contains on average 3 codewords as follows:

$$\mathscr{M} = \{c_1, c_2, c_3\}$$

When a car is parked, it becomes part of the background and codewords for car are added to the codebook $\mathscr{M}$ for that pixel.

$$\mathscr{M} = \{c_1, c_2, c_3, c_{i_1}, c_{i_2}\}$$

When the background values not accessed for a long time are deleted from the background model $\mathscr{M}$, we have:

$$\mathscr{M} = \{c_{i_1}, c_{i_2}\}$$

Just when the car leaves, codewords belonging to the actual background will not be present in the background model. It will be classified as foreground. These codewords would be part of the cache $\mathscr{H}$ till some learning time frame. For this duration a ghost region will be present in the background. This is illustrated in the Fig. 1.

5 Proposed Approach

The approach proposed in this paper focuses on removing the ghost region left behind due to non-static background. It eliminates the background learning time where false positives are detected by retaining the background codewords in codebook $\mathscr{M}$ of each pixel. As a result, as and when the object leaves the area, immediately the left behind pixels are classified as background.

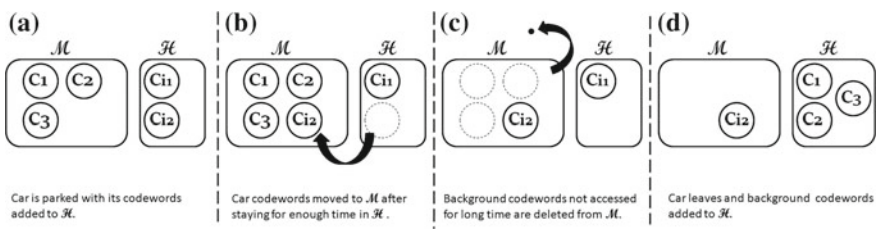


Fig. 1 Updation of codebooks $\mathscr{M}$ and $\mathscr{H}$ in multi-layered codebook model

We redefine the codeword tuple ***MNRL*** (Maximum Negative Run Length) which was the maximum amount of time a codeword has not recurred, to the ***LNRL*** (Latest Negative Run Length), which is the latest amount of time a codeword is not accessed. *LNRL* is obtained by the difference in the time since a codeword was last accessed to the current time.

Removing *MNRL* does not effect the temporal filtering step [7] and layered modelling step [7], since we can use the frequency component of codewords alone for filtering without the need of *MNRL*.

We replace the *MNRL* used in the temporal filtering [7] to remove the set of codewords belonging to the foreground from the background model in the training period, defined as *MNRL* greater than some threshold, which is usually taken as *N/2*, where ***N*** is the number of training samples.

Instead, we use the frequency component of a pixel for codeword C_m to perform the same thresholding by using:

$$f_{c_m} < \frac{N}{2} \tag{3}$$

This removes the codewords of foreground objects and only keeps the codewords staying for atleast half the training time frames.

Additionally, ***MNRL*** gives the maximum of all times the duration of non-occurrence, it alone gives no information about the recent/latest access of the codeword, which is provided by ***LNRL*** and hence it is used instead of *MNRL*.

For example, two different codewords say C_1 and C_2 representing different objects, can have same frequencies as: $f_{c_1} = f_{c_2}$, and same *MNRL*: $\Psi_{c_1} = \Psi_{c_2}$, irrespective of the fact that codeword C_2 is accessed recently for a considerable duration of time. Thus, simple codeword information of the two objects could not tell which among them was recently accessed.

However, ***LNRL*** defined in this paper records the latest information of active and inactive codeword in the set of codebook.

For multi-layered approach, we need to set a memory limit on the length of codebook model $\mathscr{M}$. It is been observed that the number of codewords for a pixel is on an average $[3 \cdots 5]$ [7]. Thus, we may set the maximum limit $\mathcal{M}_{max} = [5 \cdots 7]$ that is enough to support multiple layers of background, where the range also supports an average number of non permanent background codewords. The complete approach is illustrated in Algorithm 1.

Algorithm 1 Multi-layered Codebook Ghost Elimination Approach

I. **For** each pixel p of the incoming frame, **do** the following steps.

II. $\forall\ C_m \in \mathscr{M}$ where $m = 1$ to L, find c_m which matches to x.

(i) Update the codeword c_m and set:

- $\Psi_{c_m} \longleftarrow 0$

(ii) For all codewords C_m in $\mathscr{M}$ that do not match to x, and all codewords in $\mathscr{H}$, increase the ***LNRL*** as:

- $\Psi_{c_m} \longleftarrow \{\ \Psi_{c_m} + 1 \mid \forall C_k \in \mathscr{M}$ that does not match to x, $\wedge\quad \forall C_k \in \mathscr{H}\ \}$

III. $\mathbf{x} = \left\{ \begin{array}{ll} \textbf{background} & \text{if match found} \\ \textbf{foreground} & \text{otherwise} \end{array} \right\}$

IV. If no match is found in $\mathscr{M}$, find a matching codeword in $\mathscr{H}$.

(i) Update the matching codeword C_m and set :

- $\Psi_{c_m} \longleftarrow 0$

(ii) For all codewords in $\mathscr{M}$, and the non-matching codewords in $\mathscr{H}$, set:

- $\Psi_{c_m} \longleftarrow \{\ \Psi_{c_m} + 1 \mid \forall C_k \in \mathscr{H}$ that does not match to x, $\wedge\quad \forall C_k \in \mathscr{M}\ \}$

V. If no match is found in $\mathscr{H}$, create a new codeword with Ψ=0, and add it to $\mathscr{H}$. Also set:

- $\Psi_{c_m} \longleftarrow \{\ \Psi_{c_m} + 1 \mid \forall C_k \in (\mathscr{H}, \mathscr{M})\ \}$

VI. Delete the codeword belonging to foreground from $\mathscr{H}$, whose LNRL is greater than some threshold t_h :

$\mathscr{H} \longleftarrow \mathscr{H} \mid \{\ \forall\ C_k \in \mathscr{H}$, where $\Psi_{c_k} > t_h\ \}$

VII. Add the codeword C_m which stays for a long time t_m in $\mathscr{H}$ to $\mathscr{M}$ if it not exceeds the memory limit. When memory limit $\mathcal{M}_{max}$ is reached, then find the codeword with largest Ψ in $\mathscr{M}$ and replace it with C_m.

(i) $\forall\ C_m \in \mathscr{H}$, where $f_{c_m} > t_m$, **do**

(ii) if length($\mathscr{M}$) $< \mathcal{M}_{max}$

- $\mathscr{M} \longleftarrow \mathscr{M} + C_m$

(iii) else, replace C_k with C_m such as :

- $\Psi_{c_k} = \max\{\ \Psi_{c_t} \mid \forall\ C_t \in \mathscr{M}\ \}$

6 Algorithm Description

After the training as illustrated in [7] using N sample frames, ***LNRL***(Ψ) is calculated as:

$$\Psi_{c_m} = N - q_{c_m} \tag{4}$$

where, N is the total number of training frames and q_{c_m} is the last access time of codeword C_m.

For background codewords in $\mathscr{M}$, ***LNRL***(Ψ) is close to 0.

Considering the scenario of dynamic background, initially the codebook model after training $\mathscr{M}$ for pixel x is:

$$\mathscr{M}_x = \{c_1, c_2, c_3\}$$

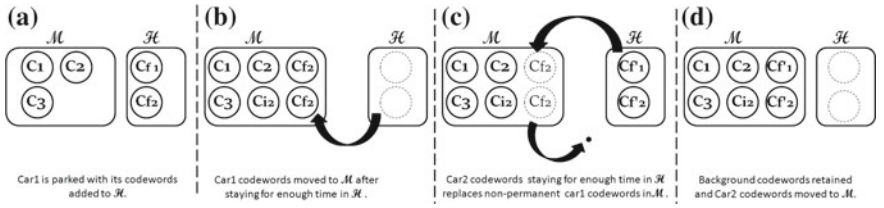


Fig. 2 Updation of codebooks $\mathscr{M}$ and $\mathscr{H}$ in improved multi-layered codebook model for ghost elimination

Now, when a car gets parked, its codewords c_{f_1} and c_{f_2} are added to the background model:

$$\mathscr{M}_x = \{c_1, c_2, c_3, c_{f_1}, c_{f_2}\}$$

When the car leaves the parking area, the set of codewords for background $\{c_1, c_2, c_3\}$ are still part of the background model $\mathscr{M}$ and the area left behind is recognised as background. For multi-layer scenarios, when a car leaves the parking area and another car occupies it to become background, we need to let the first car codeword be removed from the model and second car codewords ($c_{f_1'}, c_{f_2'}$ in our case) to be included in the model. This is done by filtering the codebook $\mathscr{M}$ using **LNRL(Ψ)**. After the first car leaves and before the second car arrives, the actual background is visible for some time frame. This sets the **LNRL(Ψ)** of the background codewords to 0.

As mentioned in step II of the algorithm 1, the codeword with maximum **LNRL(Ψ)** is deleted from the model. Since the background codewords now have **LNRL(Ψ)** close to 0, the first car codeword is deleted from the model as it is now having the maximum **LNRL(Ψ)**. Thus we have:

$$\mathscr{M}_x = \{c_1, c_2, c_3, c_{f_1'}, c_{f_2'}\}$$

This also removes the objects which are non-permanent and currently not part of the background. The complete scenario is illustrated in Fig. 2.

7 Results

The performance of the method is evaluated both qualitatively and quantitatively. Comparisons of the proposed method with some existing methods like basic Codebook model, Gaussian Mixture Model and ViBe [10] is done. The test sequences are downloaded from standard datasets [11]. For testing, the first 180 frames of the video sequence are used for training.

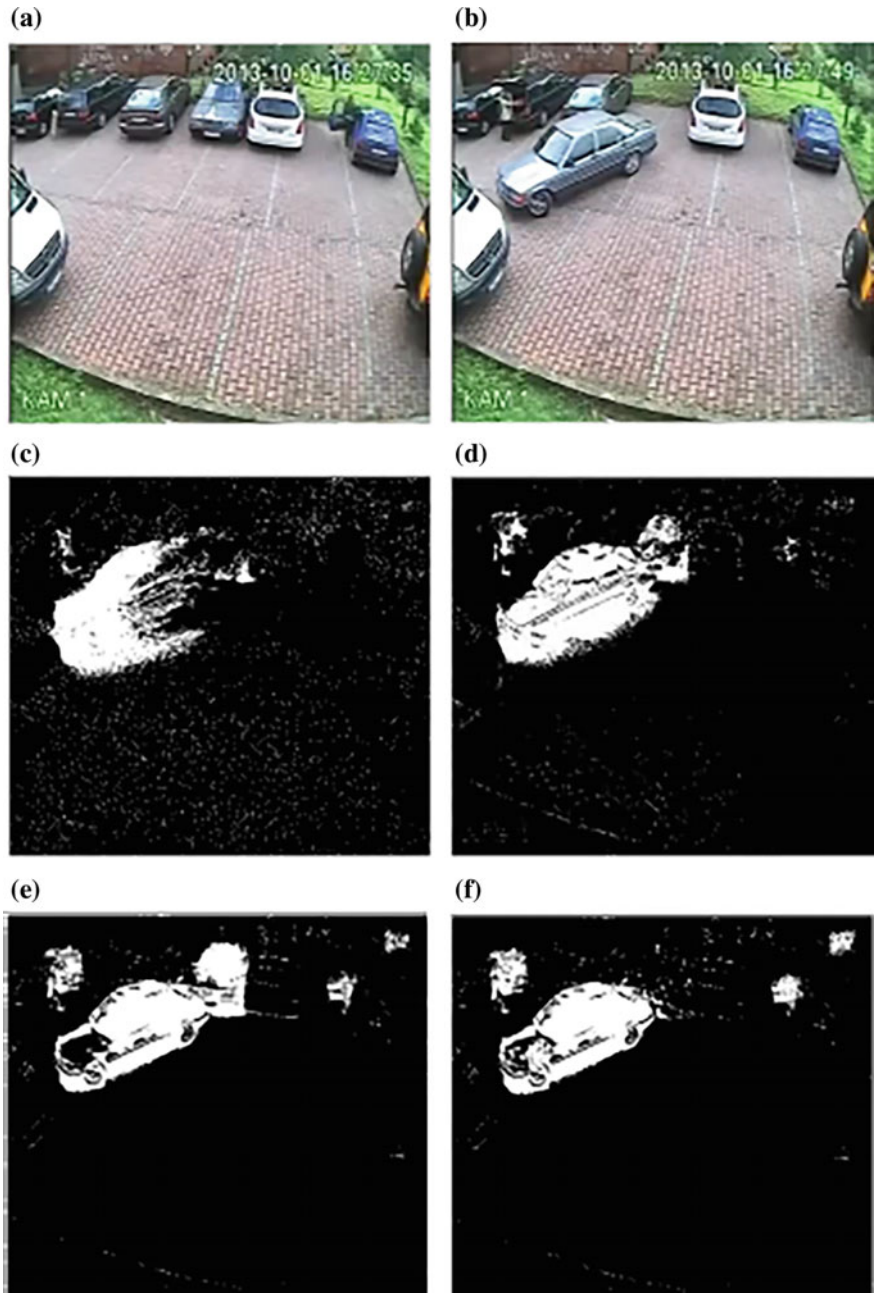


Fig. 3 Results of object detection. **a** Original image without moving object. **b** Image with a moving car. **c** Result of GMM. **d** ViBe. **e** Codebook. **f** Proposed approach

Table 1 Similarity measures S(A, B) of the compared methods

	GMM	ViBe	Codebook	Proposed
AbandonedBox (4500)	0.599	0.671	0.627	0.734
Parking (2500)	0.409	0.597	0.514	0.651
Sofa (2750)	0.368	0.471	0.413	0.529

Qualitative Evaluation:

Figure 3 shows results using the frame sequence where car moves out of parking area and creates empty background region. Figure 3 (c(ViBe), d(GMM), e(CodeBook)) shows that state-of-art methods creates ghost region where the car has left, while the proposed approach as shown in Fig. 3f adapts to changes in the background and successfully detects uncovered background region providing better result in comparison to other methods.

Quantitative Evaluation:

We used the similarity measure proposed in [12] for evaluating the results of foreground segmentation. Let A be a detected region and B be the corresponding "ground truth", then the similarity measure between the regions A and B is defined as:

$$S(A, B) = \frac{A \cap B}{A \cup B} \tag{5}$$

where, $S(A, B)$ has a maximum value 1.0 if A and B are the same, and 0 when the two regions are least similar.

Table 1 shows the similarity measures of the above three methods against the proposed method on three video sets of dynamic background from CDnet dataset [13] which includes videos of parking, abandonedBox and sofa along with their testing number of frames. By comparing the similarity values in Table 1, it is shown that the proposed method is successful in recognising background region more accurately than the other three methods for dynamic backgrounds as its similarity measure is higher than the other three with values varying from 0.529 to 0.734.

8 Conclusion

This paper proposed an approach that deals with regions detected falsely as foreground due to dynamic background when a moving object dissolved into background starts to move again. The ghost region left due to false detection in basic Codebook model is eliminated by retaining the actual permanent background codewords. The results are also compared to other motion detection approaches where our approach removes the false positives and shows improved results. Quantitative evaluation and

comparison with 3 existing methods have shown that the proposed method has provided an improved performance in detecting background regions in dynamic environments with the similarity measure ranging from 0.529 to 0.734. It also reduces the memory requirement significantly by keeping a memory limit of 5 to 7 which deletes the non-permanent background codewords without effecting the results.

References

1. K. Toyama, J. Krumma, B. Brumitt, and B. Meyers. "Wallflower: principles and practice of background maintenance" In ICCV99, 1999.
2. Wren, Christopher Richard, et al. "Pfinder: Real-time tracking of the human body." IEEE Transactions on Pattern Analysis and Machine Intelligence, pp: 780–785, 1997.
3. Stauffer, Chris, and W. Eric L. Grimson. "Adaptive background mixture models for real-time tracking." IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999.
4. Zivkovic, Zoran. "Improved adaptive Gaussian mixture model for background subtraction." Proceedings of the 17th International Conference on Pattern Recognition, ICPR Vol. 2. IEEE, 2004.
5. KaewTraKulPong, Pakorn, and Richard Bowden. "An improved adaptive background mixture model for real-time tracking with shadow detection." Video-based surveillance systems. Springer US, 135–144, 2002.
6. Mittal, Anurag, and Nikos Paragios. "Motion-based background subtraction using adaptive kernel density estimation." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 2. IEEE, 2004.
7. Kim, Kyungnam, et al. "Real-time foregroundbackground segmentation using codebook model." Real-time imaging 11.3, pp: 172–185, 2005.
8. Sigari, Mohamad Hoseyn, and Mahmood Fathy. "Real-time background modeling/subtraction using two-layer codebook model." Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 1. 2008.
9. Guo, Jing-Ming, et al. "Hierarchical method for foreground detection using codebook model." IEEE Transactions on Circuits and Systems for Video Technology, pp: 804–815, 2011.
10. Barnich, Olivier, and Marc Van Droogenbroeck. "ViBe: A universal background subtraction algorithm for video sequences." IEEE Transactions on Image Processing, pp:1709–1724, 2011.
11. Goyette, N.; Jodoin, P.; Porikli, F.; Konrad, J.; Ishwar, P., "Changedetection. net: A new change detection benchmark dataset," Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pp: 16–21, June 2012.
12. Li, L., Huang, W., Gu, I.Y.H., Tian, Q., "Statistical modeling of complex backgrounds for foreground object detection" IEEE Transactions on Image Processing, pp: 1459–1472, 2004.
13. Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, CDnet 2014: "An Expanded Change Detection Benchmark Dataset", in Proc. IEEE Workshop on Change Detection (CDW-2014) at CVPR-2014, pp. 387–394. 2014

Reassigned Time Frequency Distribution Based Face Recognition

B.H. Shekar and D.S. Rajesh

Abstract In this work, we have designed a local descriptor based on the reassigned Stankovic time frequency distribution. The Stankovic distribution is one of the improved extensions of the well known Wigner Wille distribution. The reassignment of the Stankovic distribution allows us to obtain a more resolute distribution and hence is used to describe the region of interest in a better manner. The suitability of Stankovic distribution to describe the regions of interest is studied by considering face recognition problem. For a given face image, we have obtained key points using box filter response scale space and scale dependent regions around these key points are represented using the reassigned Stankovic time frequency distribution. Our experiments on the ORL, UMIST and YALE-B face image datasets have shown the suitability of the proposed descriptor for face recognition problem.

Keywords Stankovic distribution · Time frequency reassignment · Feature detection · Local descriptor

1 Introduction

Face recognition plays a vital role in surveillance/security applications. Active research works in this area has contributed efficient algorithms to address some of the inherent problems present in face recognition like illumination, pose, age and scale variations. But the performance of these algorithms is found to be satisfactory in the case of face images that are acquired under controlled environment. Hence, research on new descriptors which can show superior performance on the more challenging datasets that are acquired under uncontrolled environment is still underway.

B.H. Shekar (✉) · D.S. Rajesh
Department of Computer Science, Mangalore University,
Mangalagangothri 574199, Karnataka, India
e-mail: bhshekar@gmail.com

D.S. Rajesh
e-mail: rajeshds1972@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_43

Recently, Zheng et al. [4] used the Discrete Wavelet Transform for face recognition. They fused the transform coefficients of the three detail sub bands and combined with the approximation sub-band coefficients to develop their descriptor. Ramasubramanian et al. [8] used the Discrete Cosine transform for face recognition. In their work, they proposed to retain only those coefficients of the transform which could contribute significantly for face recognition. Using principal component analysis on these retained coefficients, they computed the "cosine faces" similar to the eigenfaces to form their descriptor.

On the other hand, we have seen an excellent growth in signal processing domain where researchers came out with improved versions of the basic transforms. Recently, Auger and Flandrin [2] extended the concept of time frequency reassignment. Using the reassignment method, they proved that it is possible to obtain improvement of the time frequency distributions from their less resolute time frequency distributions. They derived expressions for time frequency reassignment of the (smoothed/pseudo) Wigner-Wille distribution, Margeneau-Hill distribution, etc. These reassigned time frequency distributions possess improvement over their original distribution. Flandrin et al. [7] came up with the reassigned Gabor spectrogram. On the similar line, Stankovic and Djurovic [3] derived a reassigned version of the Stankovic time frequency distribution. These improved distributions and their significant characteristics to capture the more accurate frequency distribution motivated us to investigate their suitability for face recognition problem.

2 Time Frequency Reassignment

2.1 The Time Frequency Reassignment Principle

The principle of time frequency reassignment may be explained using the relation between the well known Wigner-Ville time frequency distribution (WVD)and the spectrogram. The WVD of a signal **s** at a time **t** and frequency **f** is defined as [2] (Fig. 1):

$$WVD_s(t,f) = \int_{-\infty}^{+\infty} s(t+\frac{\tau}{2})\, s^*(t-\frac{\tau}{2}) e^{-i2\pi f\tau} d\tau \tag{1}$$

Here $\mathbf{s}^*$ stands for conjugate of **s** and **i** stands for imaginary part of a complex number. The spectrogram of a signal **s** at a time **t** and frequency **f** is defined as:

$$SP_s(t,f) = |STFT_s(t,f)|^2 \tag{2}$$

where

$$STFT_s(t,f) = \int_{-\infty}^{+\infty} w(\tau - t) s(\tau) e^{-i2\pi f\tau} d\tau \tag{3}$$

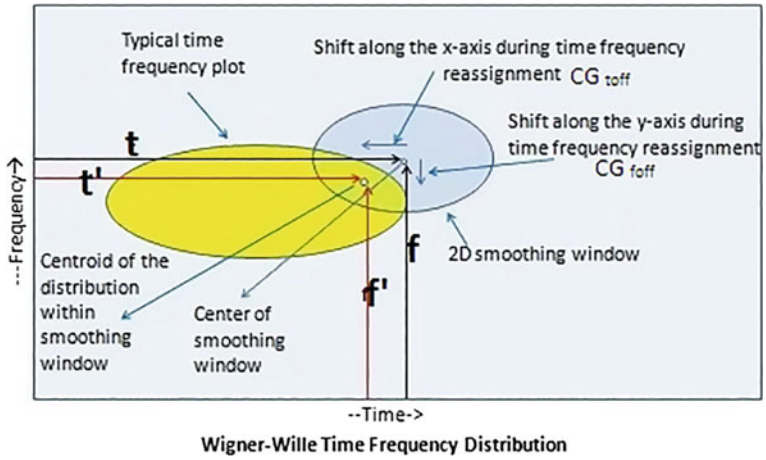


Fig. 1 The time frequency reassignment procedure being executed on the Wigner-Wille distribution of a typical signal

is the Short Term Fourier Transform (STFT) of the signal **s**(**t**) with $w(t)$ being the window function. Given two signals **s**(**t**) and $w(t)$, according to the unitarity property of WVD [14], they are related by

$$|\int_{-\infty}^{+\infty} s(t)w(t)dt|^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} WVD_s(t,f)WVD_w(t,f)dtdf \tag{4}$$

where WVD_s and WVD_w are the WVD functions of **s(t)** and $w(t)$ respectively. Suppose the function $w(t)$ is shifted in time by τ and in frequency by ν as follows:

$$w(\tau - t)e^{-i2\pi\nu\tau} \tag{5}$$

then, it can be shown using Eqs. 2, 3, 4 and 5 that

$$SP_s(t,f) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} WVD_w(\tau - t, \nu - f)WVD_s(\tau,\nu)d\tau d\nu \tag{6}$$

since the LHS of Eq. 4 becomes

$$|\int_{-\infty}^{+\infty} s(t)w(\tau - t)e^{-i2\pi\nu t}dt|^2 = |STFT_s(t,f)|^2 = SP_s(t,f) \tag{7}$$

The transformation of Eq. 4 to Eq. 6 may be proved using the time and frequency covariance properties of the WVD [11]. The Eq. 6 is the well known **Cohen's class** representation of spectrogram of a signal **s**. The spectrogram of a signal **s(t)** i.e. Eq. 6

may now be viewed as the smoothing of the WVD of the signal (WVD_s), by the WVD of the spectrogram smoothing function w(t) (i.e. WVD_w).

The method of improving the resolution of the spectrogram, by time frequency reassignment may be found by analyzing Eq. 6. The WVD_s localizes the frequency components of the given signal very well, but its time frequency distribution contains a lot of cross interference (between the frequency components present in the signal, as seen in Fig. 2d) terms. Though the smoothing operation in Eq. 6 reduces these interference terms, the sharp localization property of the WVD_s is lost (the frequency components plotted in the WVD_s get blurred or spread out) and hence the spectrogram is of poor resolution. The reason for this may be explained as follows. It shall be observed from Eq. 6 that the window function WVD_w runs all over the time frequency distribution WVD_s during convolution. Convolution at a particular location (t,f) can be seen in Fig. 1. The rectangular region is the WVD_s of the signal **s**. The yellow coloured region shows a region of the WVD_s distribution, with significant values (supposedly indicating the strong presence of signals corresponding to the frequencies and times of that region). Other such regions are not shown for simplicity of understanding. The other oval shaped region shown is the window function WVD_w in action at (t,f). The weighted sum of the distribution values WVD_s (weighted by the WVD_w values) within the arbitrary boundary of the window function is assigned to the location (t,f) which here is the center of the window. The WVD_s, which had no value at (t,f) (implying the absence of a signal corresponding to (t,f)) before convolution, gets a value (wrongly implying the presence of a signal corresponding to (t,f)) which is an indication of incorrect assignment of distribution values. All such incorrect assignments leads to a spread out low resolution spectrogram. One possible way of correction is by a more sensible method called time frequency reassignment. That is, during reassignment, the result of above convolution is assigned to a newly located position (t',f'), which is the center of gravity

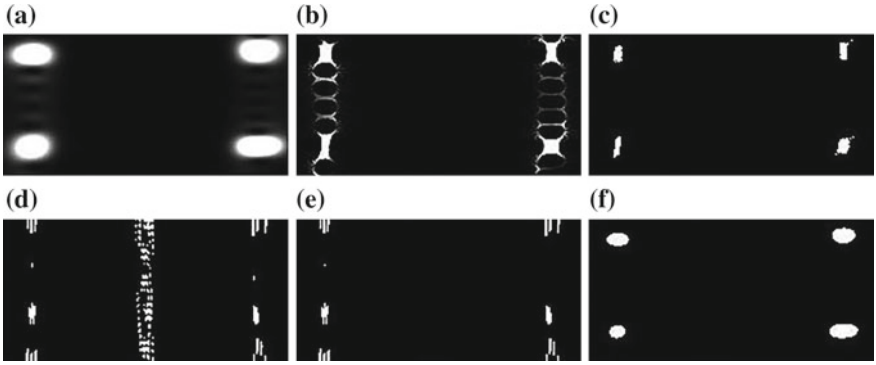


Fig. 2 TFD of a synthetic signal made of 4 freq components, shown as 4 bright spots in **a**. A low resolution Gabor time frequency distribution. **b**. Reassigned Gabor distribution. **c**. The reassigned Stankovic time frequency distribution. **d**. The WVD. **e**. The pseudo WVD **f**. The Stankovic time frequency distribution

of the region of the distribution within the window function. Reassigning the convolution value to the center of gravity gives us a more correct representation of the actual time frequency distribution. Thus the reassigned spectrogram will be more accurate than its original. The equation that gives the offset in the position of the centre of gravity (t',f'), from (t,f) is hence given by [2]

$$CG_{toff}(t,f) = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tau WVD_w(\tau - t, \nu - f)WVD_s(\tau,\nu)d\tau d\nu}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} WVD_w(\tau - t, \nu - f)WVD_s(\tau,\nu)d\tau d\nu} \tag{8}$$

and

$$CG_{foff}(t,f) = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \nu WVD_w(\tau - t, \nu - f)WVD_s(\tau,\nu)d\tau d\nu}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} WVD_w(\tau - t, \nu - f)WVD_s(\tau,\nu)d\tau d\nu} \tag{9}$$

where CG_{toff} and CG_{foff} are the offsets along the t-axis and the f-axis of the time frequency distribution, as shown in Fig. 1. The Eq. 8 may also be expressed in terms of the Rihaczek distribution [2] as

$$CG_{toff}(t,f) = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tau Ri_w^*(\tau - t, \nu - f)Ri_s(\tau,\nu)d\tau d\nu}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Ri_w^*(\tau - t, \nu - f)Ri_s(\tau,\nu)d\tau d\nu} \tag{10}$$

where

$$Ri_s(\tau,\nu) = s(\tau)S(\nu)e^{-i2\pi\nu\tau}, Ri_w(\tau,\nu) = w(\tau)W(\nu)e^{-i2\pi\nu\tau} \tag{11}$$

and $\mathbf{S}(\nu)$, $\mathbf{W}(\nu)$ are the Fourier transforms of signal and window functions respectively.

Expressing Eq. 8 in terms of the Rihaczek distribution, expanding it using the above Eq. 11 and rearranging the integrals of the resulting equation, gives the following equation for CG_{toff}

$$CG_{toff}(t,f) = \mathbf{Real}\left(\frac{STFT_s^{\tau w}(t,f)STFT_s^*(t,f)}{|STFT_s^*(t,f)|^2}\right) \tag{12}$$

where the superscript τw implies the computation of STFT with a window τw, which is the product of the traditional window function $w(t)$ (used in STFT) with τ. Similarly, the equation for $\mathbf{CG_{foff}}$ may be derived in terms of STFTs as

$$CG_{foff}(t,f) = \mathbf{Imag}\left(-\frac{STFT_s^{dw}(t,f)STFT_s^*(t,f)}{|STFT_s^*(t,f)|^2}\right) \tag{13}$$

where the superscript dw implies the computation of STFT with a window dw, which is the first derivative of the traditional window function used in STFT. Hence the reassigned location of the convolution in Eq. 6 is given by $\mathbf{t'} = \mathbf{t} - \mathbf{CG_{toff}}$ and $\mathbf{f'} = \mathbf{f} + \mathbf{CG_{foff}}$. Thus the result of convolution within the smoothing window is

assigned to $(\mathbf{t}', \mathbf{f}')$ instead of $(\mathbf{t}, \mathbf{f})$. This computation will happen at all pixels $(\mathbf{t}, \mathbf{f})$ of the time frequency distribution WVD_s in Fig. 1, to get a more resolute and true, **reassigned time frequency distribution**.

2.2 The Reassigned Stankovic Time Frequency Distribution

It is better to arrive at the equation for the Stankovic distribution starting with the pseudo Wigner-Wille distribution [12], which is given by

$$pWVD(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} w(\frac{\tau}{2})w(\frac{-\tau}{2})s(t+\frac{\tau}{2})s(t-\frac{\tau}{2})e^{-i\omega\tau}d\tau \tag{14}$$

We can see in the above equation that a narrow window function $w(t)$ limits the auto-correlation of signal s and hence does a smoothing of *WVD* in the frequency domain (we can see that the white interference strip present in Fig. 2d (WVD), is missing in Fig. 2e (pseudoWVD) (i.e. it has improved over WVD, due to this smoothing) in the distribution. The Eq. 14 may also be expressed as

$$pWVD(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} STFT_s(t,\omega+\frac{\nu}{2})STFT_s(t,\omega-\frac{\nu}{2})d\nu. \tag{15}$$

The Cohen's class representation of pseudo Wigner Wille distribution is given by [14]

$$pWVD(t,\omega) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W(\nu-f)WVD_w(t,f)dtdf \tag{16}$$

where W is the Fourier transform of the window function w. By inserting a narrow window B into the above Eq. 15 we get the Stankovic time frequency distribution [12] (the S-method)

$$S(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} B(\nu)STFT_s(t,\omega+\frac{\nu}{2})STFT_s(t,\omega-\frac{\nu}{2})d\nu \tag{17}$$

where B does a smoothing operation in the time domain (we can see that the white interference strip present in Fig. 2e (pseudoWVD), is missing in Fig. 2f (Stankovic distribution i.e. it has improved over pseudoWVD), due to this smoothing. The Cohen's class representation of Stankovic distribution shown above may be written using Eqs. 16 and 17 as [2, 14].

$$S(t,\omega) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} b(t)WVD_w(\nu-f)WVD_s(t,f)dtdf \tag{18}$$

where b is the Fourier inverse of the window function B. We can view this equation also (like Eq. 6) as a convolution operation of WVD of a signal $\mathbf{s}$ (WVD_s) by WVD of a window function w (WVD_w). Now the reassigned Stankovic TFD is [3]

$$CG^{S}_{toff}(t,f) = \mathbf{Real}\left(\frac{S^{\tau w}_{s}(t,f)S_s(t,f)}{|S_s(t,f)|^2}\right) \tag{19}$$

where superscript τw implies the computation of Stankovic distribution with a window τw and

$$CG^{S}_{foff}(t,f) = \mathbf{Imag}\left(\frac{S^{dw}_{s}(t,f)S_s(t,f)}{|S_s(t,f)|^2}\right). \tag{20}$$

where the superscript dw implies the computation of Stankovic distribution with a window dw which is the first derivative of the window used in Stankovic distribution. The reassigned Stankovic distribution shown in Fig. 2c has greater resolution than the traditional Stankovic distribution in Fig. 2f due to reassignment done using Eqs. 19 and 20.

2.3 The Reassigned TFD Based Descriptors

The reassigned Stankovic TFD based descriptor is computed as follows. Using the difference of box filter scale space [13], scale normalized interest points in face images are detected and scale dependent square regions around the interest points are represented using the reassigned Stankovic distribution as follows. Each region is scaled down to 24×24 size, further divided into 16, 6×6 subregions. The Stankovic distribution of each of these subregions is computed using Eq. 18, reassigned using Eqs. 19 and 20 (size of this distribution will be 36×36 and due to symmetry, we may neglect the distribution corresponding to the negative frequencies). The reassigned distribution of each subregion is converted into a row vector. Hence for each square region around an interest point, we get 16 row vectors, which are stacked one above the other to form a 2D matrix. Applying PCA, we reduce the dimension of the 2D matrix to 16×160, to develop the descriptor of each 24×24 size square region around the interest point.

2.4 Classification

In our experiments over face datasets that have pose variation (ORL and UMIST), we have formed interest point based descriptors as explained in Sect. 2.3. The gallery image descriptors are put together to form a descriptor pool. Each test face descriptor is made to vote a gallery subject, whose descriptor matches with the test face

descriptor, the most. When once all the test face descriptors have completed their voting, the gallery subject which earned the most number of votes is decided as the subject, to which the test face image belongs. We have used the Chi-square distance metric for descriptor comparison.

For experiments over face datasets without pose variation (YALE), we have formed fixed position block based descriptors. The size of these blocks is also 24×24 and the descriptor for this block is formed as explained in Sect. 2.3. For classification we have used fixed position block matching and Chi-square distance for descriptor matching.

3 Experimental Results

We have conducted reassigned Stankovic time frequency distribution descriptor based face recognition experiments using the ORL, UMIST and YALE face datasets (Figs. 3, 4 and 5).

The ORL dataset has 400 images belonging to 40 subjects, with 10 images of each subject, with variations in pose, lighting and expression. In our experiments, we have varied the train : test samples ratio as follows: 3:7, 4:6, and 5:5. The performance is shown in Table 1 and the CMC curves are shown in Fig. 5. Also, we have plotted the ROC curves by varying the client to imposter ratios from 20:20, 25:15 and 30:10 (see Fig. 5). For each of the cases we have varied the training to testing samples as 3:7, 4:6, and 5:5.

The UMIST dataset consists of 565 images of 20 subjects with pose and illumination variation. The number of images in this dataset varies from 18 to 45 images per subject. We have plotted the CMC curves with 3 images/subject, 4 images/subject and 5 images/subject for training (mentioned in 3rd column of Table 1 and in color

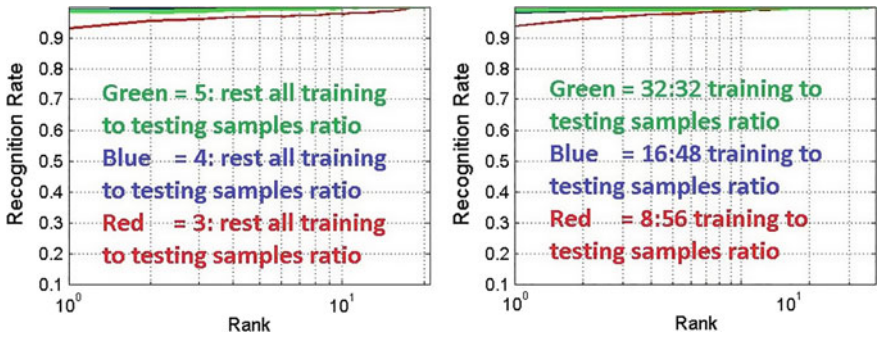


Fig. 3 The CMC curves of our method. Plot on the *left* shows the results on the UMIST dataset (with *red*, *blue*, *green curves* standing for 3, 4, and 5 images per subject for training) and on the *right* shows results on the YALE dataset (with *red*, *blue*, *green curves* standing for 8:56, 16:48, and 32:32 training to testing images ratio

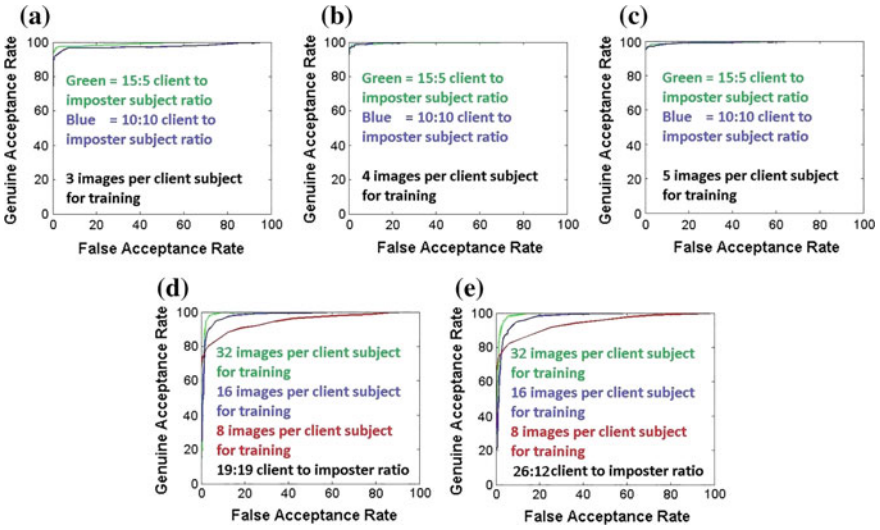


Fig. 4 The ROC curves (**a**), (**b**), (**c**) of our method on the UMIST dataset and **d**, **e** on the YALE dataset. (Each of the plot a, b, c contain 2 curves and ROC curves d, e contain 3 curves according to the ratios in *color text*)

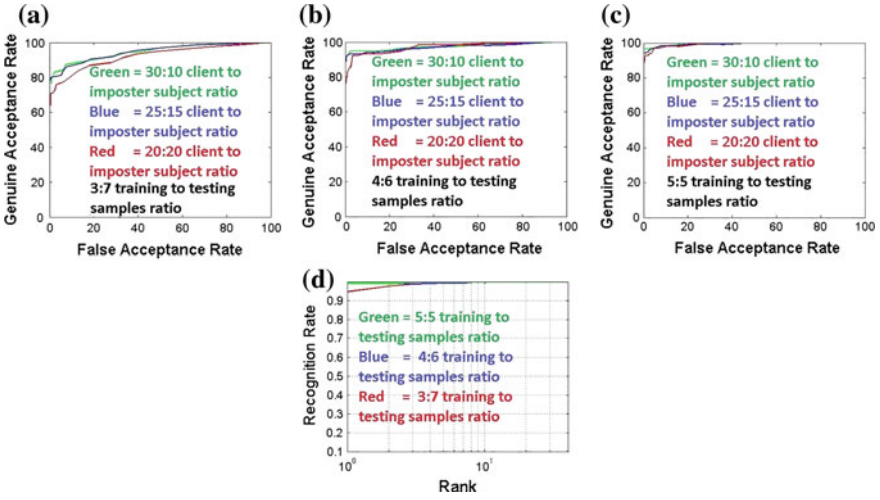


Fig. 5 The ROC curves (**a**), (**b**), (**c**) and CMC curve (**d**) of our method on the ORL dataset. (Each of the plot a, b, c contain 3 curves and d contains 3 curves according to the ratios in *color text*)

text/curves in Fig. 3). The results are also shown in Table 1 and Fig. 3. Also we have plotted the ROC curves taking client to imposter subject ratio of 10:10 and 15:5 (as mentioned in color text/curves in Fig. 4). In each of these cases, we have varied the training samples per subject as 3, 4 and 5 (see Fig. 4).

Table 1 Recognition rate (RR) of our method on various datasets (In the table T:T stands for train:test ratio)

ORL dataset		UMIST dataset		YALE dataset	
T:T	RR (%)	T:T	RR (%)	T:T	RR (%)
3:7	96.8	3:	94.25	8:56	93.74
4:6	98.75	4:	98.76	16:48	98.67
5:5	99	5:	99.6	32:32	99.5

Table 2 Comparative analysis of our method with state of the art methods (T:T stands for train:test ratio)

ORL T:T=3:7		UMIST T:T=4:rest all		YALE T:T=32:32	
Method	Recognition rate (%)	Method	Recognition rate (%)	Method	Recognition rate
ThBPSO [5]	95	ThBPSO [5]	95.1	GRRC [10]	99.1
SLGS [1]	78				
KLDA [9]	92.92				
2D-NNRW [6]	91.90				
Our method	**96.8**		**98.76**		**99.5 %**

The YALE dataset consists of 2432 images of 38 subjects, with 64 images/subject. We have plotted the CMC curves with 32 images/subject, 16 images/subject and 8 images/subject for training(mentioned in column 5 of Table 1 and in color text/curves in Fig. 3). Also we have plotted the ROC curves taking client to imposter subject ratio of 19:19 and 26:12 (see Fig. 4). In each of these cases we have varied the training samples per subject as 32, 16 and 8 (in color text/curves in Fig. 4). Table 2 indicates the comparative study of our method with other state of the art methods. Comparative study with some of the state of the art methods can be seen in Table 2.

4 Conclusion

We have proposed a local descriptor based on reassigned Stankovic distribution. We have explored its use for face recognition and found that our descriptor exhibits good discriminative capability. Our experiments on ORL, YALE and UMIST face datasets exhibit better accuracy over other methods. We found that our method provides good recognition rate with just few training samples.

References

1. M. F. A. Abdullah, M. S. Sayeed, K. S. Muthu, H. K. Bashier, A. Azman, and S. Z. Ibrahim. (2014) 'Face recognition with symmetric local graph structure (slgs)'. Expert Systems with Applications, Vol.41, No.14, pp.6131–6137.
2. F. Auger and P. Flandrin. (1995) 'Improving the readability of time frequency and time-scale representations by the reassignment method'. IEEE Transactions on Signal Processing, Vol.43, No.5 pp.1068– 1089.
3. I. Djurovic and L. Stankovic. (1999) 'The reassigned s-method'. In Telecommunications in Modern Satellite, Cable and Broadcasting Services, 1999. 4th International Conference on, Vol. 2, pp.464 467.
4. Z. Huang, W. Li, J. Wang, and T. Zhang. (2015) 'Face recognition based on pixel-level and feature-level fusion of the toplevels wavelet sub-bands'. Information Fusion, vol.22, No. pp.95–104.
5. N. A. Krisshna, V. K. Deepak, K. Manikantan, and S. Ramachandran. (2014) 'Face recognition using transform domain feature extraction and pso-based feature selection'. Applied Soft Computing, Vol.22, pp.141–161.
6. J. Lu, J. Zhao, and F. Cao. (2014) 'Extended feed forward neural networks with random weights for face recognition'. Neurocomputing, Vol.136, pp.96–102.
7. F. A. P. Flandrin and E. Chassande-Mottin. (2002) 'Time-frequency reassignment from principles to algorithms'. Applications in Time Frequency Signal Processing, Vol.12 No.1, pp.234 778.
8. D. Ramasubramanian and Y. Venkatesh. (2001) 'Encoding and recognition of faces based on the human visual model and fDCTg'. Pattern Recognition, Vol.34, No.12, pp.2447–2458.
9. Z. Sun, J. Li, and C. Sun. (2014) 'Kernel inverse fisher discriminant analysis for face recognition'. Neurocomputing, Vol.134, pp.4652.
10. M. Yang, L. Zhang, S. C. Shiu, and D. Zhang. (2013) 'Gabor feature based robust representation and classification for face recognition with gabor occlusion dictionary'. Pattern Recognition, Vol.46, No.7, pp.1865–1878.
11. Rene Carmona, Wen-Liang Hwang, Bruno Torresani. 'Practical Time-Frequency Analysis, Volume 9: Gabor and Wavelet Transforms, with an Implementation in S (Wavelet Analysis and Its Applications)' 1st Edition
12. LjubiSa StankoviC, (1994) 'A Method for Time-Frequency Analysis'. IEEE Transactions on Signal Processing, pp.225–229.
13. M. Agrawal, K. Konolige, and M. Blas, (2004) 'CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching'. Springer, pp.102 115.
14. Franois Auger, Patrick Flandrin, Paulo Gonalvs, Olivier Lemoine, (1996) 'Time-Frequency Toolbox', CNRS (France), Rice University (USA).
15. F. S. Samaria and F. S. Samaria *t and A.C. Harter and Old Addenbrooke's Site, 1994 'Parameterisation of a Stochastic Model for Human Face Identification'.

Image Registration of Medical Images Using Ripplet Transform

Smita Pradhan, Dipti Patra and Ajay Singh

Abstract For image fusion of geometrically distorted images, registration is the prerequisite step. Intensity-based image registration methods are preferred due to higher accuracy than that of feature-based methods. But, perfect registered image using intensity based method leads towards improvements in computational complexity. Conventional transform like wavelet transform based image registration reduces the computational complexity, but suffers from discontinuities such as curved edges in the medical images. In this paper, a new registration algorithm is proposed that uses the approximate-level coefficients of the ripplet transform, which allows arbitrary support and degree as compared to curvelet transform. The entropy-based objective function is developed for registration using ripplet coefficients of the images. The computations are carried out with 6 sets of CT and MRI brain images to validate the performance of the proposed registration technique. The quantitative approach such as standard deviation, mutual information, peak signal to noise ratio and root mean square error are used as performance measure.

Keywords Image registration · Ripplet transform · Standard deviation · Mutual information · Root mean square error · Peak signal noise ratio

S. Pradhan (✉) · D. Patra · A. Singh
IPCV Lab, Department of Electrical Engineering,
National Institute of Technology, Rourkela, India
e-mail: ssmita.pradhan@gmail.com

D. Patra
e-mail: dpatra@nitrkl.ac.in

A. Singh
e-mail: ajayniya.singh@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_44

1 Introduction

Now a days, in medical imaging applications, high spatial and spectral information from a single image is required to monitor and diagnose during treatment process. These informations can be achieved by multimodal image registration. Different modalities of imaging techniques gives several information about the tissues and organ of human body. According to their application range, the imaging techniques mostly used are CT, MRI, FMRI, SPECT, and PET. A Computed tomography (CT) image detects the bone injuries, whereas MRI defines the soft tissues of an organ such as brain and lungs. CT and MRI provide high resolution image with biological information. The functional imaging technique such as PET, SPECT, and fMRI gives low spatial resolution with basic information. To get the complete and detailed information from single modality is a challenging task, which necessitates the registration task to combine multimodal images [1]. The registered image is more suitable for radiologist for further image analysis task.

Image registration has several applications such as remote sensing and machine vision etc. Several researchers have been discussed and proposed different registration techniques in literature [2]. Image registration technique can be divided into intensity based and feature based technique [3]. Intensity based techniques are associated with pixel values whereas feature based techniques considers the different features such as line, point and textures etc.

The steps of the registration technique are as follows:-

- Feature detection: Here, the salient or distinctive objects such as edges, contours, corners are detected automatically. For further processing, these features can be represented by their point representatives i.e. centers of gravity, line endings, distinctive points, which are called control points (CPs).
- Feature matching: In this step, the correspondence between the features detected in the floating image and those detected in the reference image is established. Different similarity measures along with spatial relationships among the features are used for matching.
- Transform model estimation: The mapping function parameters of the floating image with respect to the reference image are estimated. These parameters are computed by means of the established feature correspondence.
- Image resampling and transformation: Finally, the floating image is transformed by means of the mapping functions. The non-integer coordinates of the images are computed by interpolation technique.

Wavelet transform (WT) based low level features, provide a unique representation of the image, and are highly suitable for characterizing textures of the image [4]. As WT is inherently non-supportive to directionality and anisotropy, the limitations were overcome by a new theory Multi-scale Geometric Analysis (MGA). Different MGA tools were proposed by the researchers such as Ridgelet, Curvelet, Bandlet and Contourlet transform for high-dimensional signals [5–10]. The principles and methods of fusion are described in [11]. Manu et.al. proposed a new statistical fusion

rule based on Weighted Average Merging Method (WAMM) in the Non Subsampled Contourlet Transform (NSCT) domain and compared with Wavelet domain [12]. Alam et.al. proposed entropy-based image registration method using the curvelet transform [13].

In this paper, we proposed a new registration technique for multimodal medical image with the help of ripplet transform. The detailed Ripplet transform (RT) is depicted in Sect. 2. In Sect. 3, proposed method is described. Performance evaluation is given in Sect. 4. Experimental Results are discussed in Sect. 5 with a conclusion in Sect. 6.

2 Ripplet Transform

Fourier transforms, wavelet transform suffers from discontinuation such as boundary and curve in image. Hence, to solve this problem, Xu et al. proposed another transform called Ripplet transform (RT) [14]. RT is high dimensional generalization of curvelet transform (CVT), appropriate for representation of image or two dimensional signals in different scale and direction [15, 16]. Now a days RT gives a new compact frame work with representation of image.

2.1 Continuous Ripplet Transforms (CRT)

For a 2D integrable function $f(\vec{x})$, the continuous ripplet transform is defined as the inner product of $f(\vec{x})$ and ripplets

$$R(a, \vec{b}, \theta) = \langle f \ , \ \rho_{\vec{ab\theta}} \rangle = \int f(\vec{x}) \overline{\rho_{\vec{ab\theta}}(\vec{x})} d\vec{x} \tag{1}$$

where $R(a, \vec{b}, \theta)$ are the ripplet coefficients. When ripplet function intersects with curves in images, the corresponding coefficients will have large magnitude, and the coefficients decay rapidly along the direction of singularity as $a \to 0$.

2.2 Discrete Ripplet Transforms (DRT)

For digital image processing, discrete transform is mostly used than continuous transform. Hence, discretization of Ripplet transform is defined. The discrete ripplet transform of an M x N image f (n1, n2) will be in the form of

$$R_{\vec{j\,kl}} = \sum_{n_1}^{M-1} \sum_{n_2}^{N-1} f(n_1, n_2) \overline{\rho_{\vec{j\,kl}}(n_1, n_2)} \tag{2}$$

where $R_{(j,\vec{k},l)}$ are the ripplet coefficients. Again the image can be reformed by inverse discrete ripplet transform.

$$\vec{f}(n_1, n_2) = \sum_{j} \sum \sum_{l} R_{\vec{jkl}} \rho_{\vec{jkl}}(n_1, n_2) \tag{3}$$

3 Proposed Methodology

During acquisition of brain images, changes are found in brain shape and position with the skull over short periods between scan. These images require corrections for a small amount of subject motion during imaging procedure. Some anatomical structures appear with more contrast in one image than other. These structures in various modalities are convenient to have more information about them. Brain MR images are more sensitive to contrast changes. The local neighboring coefficients of wavelet-like transforms of an image results better as compared to the processing of the coefficients of the entire sub-band. Alam et.al. proposed local neighboring coefficients of the transform in the approximate sub-band [13]. In this paper, the cost function is computed based on the probability distribution function (PDF) of ripplet transform. The distorted parameters of the mapping function are obtained by minimizing the cost function. The cost function is derived by the conditional entropy between the neighboring ripplet transform of reference image and floating image. The joint PDF of the local neighboring ripplet coefficients in the approximate band of the reference and floating images are considered as the bivariate Gaussian PDF. The conditional entropies can be calculated by differencing the joint and marginal entropies. At minimum conditional entropy, the floating image geometrically aligned to the reference. The cost function of the for minimization of the conditional entropies can be expressed as

$$C_{\Phi} = H(\xi_r, \xi_f) - \alpha H(\xi_r) - (1 - \alpha) H(\xi_f) \tag{4}$$

where ξ_r and ξ_f are the random variable with conditional dependencies, and α is the weight parameter and $0 < \alpha < 1$.

4 Performance Evaluation

For statistical analysis of the proposed scheme, several performance measures have been carried out in the simulation.

4.1 Peak Signal to Noise Ratio (PSNR)

Considering two images, the PSNR can be computed as

$$PSNR = 10 * \log_{10}\left(\frac{MAX_I^2}{RMSE}\right) \tag{5}$$

where Root Mean Square Error (RMSE) between the registered image and the reference image represents the error of mean intensity of both images. It can be stated as

$$RMSE = \sqrt{\frac{1}{(m \times n)}\left[\sum\sum\left(I_{ref}(x,y) - I_{reg}(x,y)\right)\right]^2} \tag{6}$$

4.2 Standard Deviation (STD)

Standard deviation defines the variation of the registered image. Image with high dissimilarity results high STD value. It can be defined as

$$STD = \frac{1}{mn}\sum\left(I_{reg}(x,y) - mean\right)^2 \tag{7}$$

4.3 Mutual Information (MI)

Mutual information defines the amount of dependency of the two images. Higher the value of MI means better registered.

$$MI = I(R) + I(F) - I(R,F) \tag{8}$$

5 Experimental Results

As CT and MRI provide high resolution images with biological information, this combination of brain images is considered for evaluation of proposed technique. For parametric analysis, several set of images have been taken. Among them 6 set of CT and MRI are shown in this paper. The existing and proposed techniques were evaluated with those images. The proposed technique is implemented using MATLAB version 13.

Four set of image with the corresponding registered images are presented in Fig. 1. The performance value of the registration process using ripplet transform and curvlet transform are tabulated in Table 1. The obtained performance values such as STD,

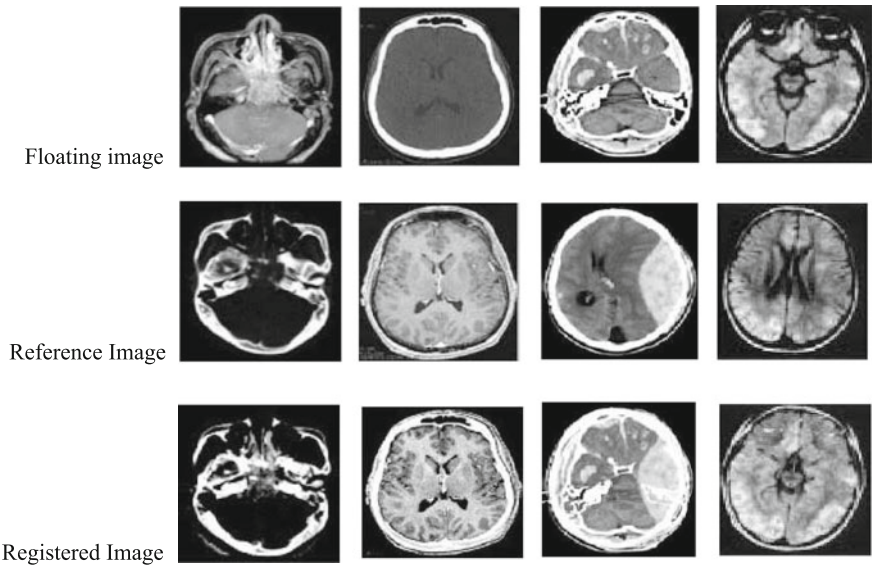


Fig. 1 1st row: floating images, 2nd row: reference images, 3rd row: registered image using ripplet transform

Table 1 Comparison of performance measures among proposed and existing method

Image set	Transform	STD	MI	RMSE	PSNR
1	Ripplet	47.84	4.28	133.24	15.77
	Curvlet	45.84	2.18	–	–
2	Ripplet	47.61	4.48	265.76	18.33
	Curvlet	45.61	2.17	–	–
3	Ripplet	88.23	6.08	246.45	13.22
	Curvlet	85.63	3.98	–	–
4	Ripplet	39.42	3.86	254.48	12.68
	Curvlet	37.31	1.78	–	–
5	Ripplet	69.0	5.16	178.23	11.11
	Curvlet	65.28	3.27	–	–
6	Ripplet	40.21	3.85	165.34	20.95
	Curvlet	36.78	1.77	–	–

MI, RMSE, and PSNR of the ripplet based registration technique are compared with curvlet based registration technique. The graph plot for the same performance measure are plotted for the 6 set of images in Fig. 2. From the graph it is analysed that the proposed scheme outperforms the existing method.

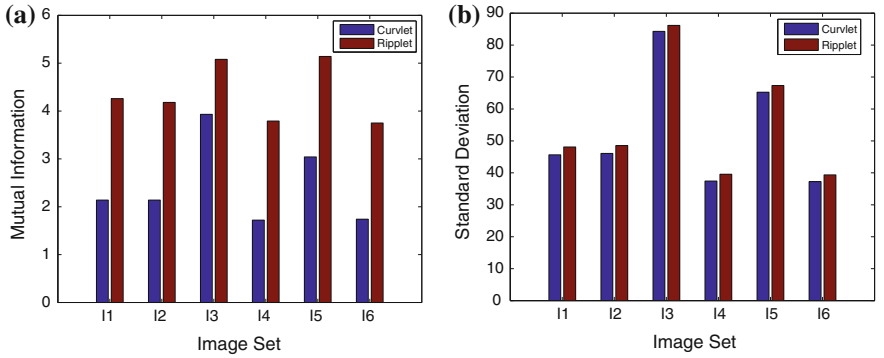


Fig. 2 Performance plot for **a** mutual information, **b** standard deviation of all sets of images

6 Conclusion

Registration of multimodal medical images plays an important role in many clinical applications. Also, transform based technique performs faster than intensity based methods. The sub-bands of CT and MRI images are extracted through ripplet transform and the entropy based objective function is computed for registration. Image registration using ripplet transform supports more widespread and accurate information than curvlet transform. The proposed technique shows improved performance measures such as standard deviation, mutual information, root mean square error, and peak signal to noise ratio as compared to the existing technique. In this paper, we have only focused mainly on brain images, we will analyze for other human body part such as abdomen, chest etc. in future.

References

1. Pradhan, S., Patra, D. RMI based nonrigid image registration using BF-QPSO optimization and P-spline, AEU-International Journal of Electronics and Communications, 69 (3), 609–621 (2015).
2. Mani, V.R.S and Rivazhagan, S. Survey of Medical Image Registration, Journal of Biomedical Engineering and Technology, 1 (2), 8–25 (2013).
3. Oliveira, F. P., Tavares, J.M.R. Medical image registration: a review, Computer methods in biomechanics and biomedical engineering, 17 (2), 73–93 (2014).
4. Acharyya, M., Kundu, M.K., An adaptive approach to unsupervised texture segmentation using M-band wavelet tranform, Signal Processing, 81(7), 1337–1356, (2001).
5. Starck, J.L., Candes, E.J., Donoho, D.L., The curvelet transform for image denoising, IEEE Transactions on Image Processing 11, 670–684 (2002).
6. Candes, E.J., Donoho, D., Continuous curvelet transform: II. Discretization and frames, Applied and Computational Harmonic Analysis 19, 198–222 (2005).
7. Candes, E.J., Donoho, D., Ridgelets: a key to higher-dimensional intermittency, Philosophical Transactions: Mathematical, Physical and Engineering Sciences 357 (1760) 2495–2509 (1999).

8. Do, M.N., Vetterli, M., The finite Ridgelet transform for image representation, IEEE Transactions on Image Processing 12 (1), 16–28 (2003).
9. Do, M.N., Vetterli, M., The contourlet transform: an efficient directional multiresolution image representation, IEEE Transactions on Image Processing 14 (12), 2091–2106 (2005).
10. Pennec, E. Le, Mallat, S.: Sparse geometric image representations with bandelets, IEEE Transactions on Image Processing 14 (4), 423–438 (2005).
11. Flusser, J., Sroubek, F., Zitova, B., Image Fusion:Principles, Methods, Lecture Notes Tutorial EUSIPCO (2007).
12. Manu, V. T., Simon P., A novel statistical fusion rule for image fusion and its comparison in non-subsampled contourlet transform domain and wavelet domain, The International Journal of Multimedia and Its Applications, (IJMA), 4 (2), 69–87 (2012).
13. Alam, Md., Howlader, T., Rahman S.M.M., Entropy-based image registration method using the curvelet transform, SIViP, 8, 491505, (2014).
14. Xu, J., Yang, L., Wu, D., A new transform for image processing, J. Vis. Commun. Image Representation, 21, 627–639 (2010).
15. Chowdhury, M., Das, S., Kundu, M. K., CBIR System Based on Ripplet Transform Using Interactive Neuro-Fuzzy Technique, Electronic Letters on Computer Vision and Image Analysis 11(1), 1–13, (2012).
16. Das, S., Kundu, M. K., Medical image fusion using ripplet transform type-1, Progress in electromagnetic research B, 30, 355–370, (2011).

3D Local Transform Patterns: A New Feature Descriptor for Image Retrieval

Anil Balaji Gonde, Subrahmanyam Murala, Santosh Kumar Vipparthi, Rudraprakash Maheshwari and R. Balasubramanian

Abstract In this paper, authors proposed a novel approach for image retrieval in transform domain using 3D local transform pattern (3D-LTraP). The various existing spatial domain techniques such as local binary pattern (LBP), Local ternary pattern (LTP), Local derivative pattern (LDP) and Local tetra pattern (LTrP) are encoding the spatial relationship between the neighbors with their center pixel in image plane. The first attempt has been made in 3D using spherical symmetric three dimensional local ternary pattern (SS-3D-LTP). But, the performance of SS-3D-LTP is depend on the proper selection of threshold value for ternary pattern calculation. Also, multiscale and color information are missing in SS-3D-LTP method. In the proposed method i.e. 3D-LTraP, the first problem is overcome by using binary approach Similarly, the other lacunas are avoided by using wavelet transform which provide directional as well as multiscale information and color features are embedded in feature generation process itself. Two different databases which included natural and biomedical database (Coral 10 K and OASIS databases) are used for experimental purpose. The experimental results demonstrate a

A.B. Gonde (✉)
Department of Electronics and Telecommunication Engineering,
SGGSIE and T, Nanded, Maharashtra, India
e-mail: abgonde@sggs.ac.in

S. Murala
Department of Electrical Engineering, IIT, Ropar, India
e-mail: subbumurala@iitrpr.ac.in

S.K. Vipparthi
Department of Computer Science Engineering, MNIT, Jaipur, Rajasthan, India
e-mail: skvipparthi@mnit.ac.in

R. Maheshwari
Department of Electrical Engineering, IIT, Roorkee, India
e-mail: rudrafee@iitr.ac.in

R. Balasubramanian
Department of Computer Science and Engineering, IIT, Roorkee, India
e-mail: balarfma@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_45

noteworthy improvement in precision and recall as compared to SS-3D-LTP and recent methods.

Keywords Local binary pattern (LBP) • Texture • Image retrieval • Wavelet transform • 3D local transform patterns

1 Introduction

1.1 Introduction

As technology grow rapidly and percolates into hearts of society, new means of image acquisition (cameras, mobiles) enter into our day to day activities and consequently digital images become an integral part of our lives. This exponential growth in the amount of visual information available coupled with our inherent tendency to organize things resulted in the use of text based image retrieval. This method has some limitations such as image annotation and human perception. Later a more potent alternative, content based image retrieval (CBIR) came into picture to address the aforementioned problems. All CBIR systems are primarily a two-step process. The first is feature extraction and second one query matching. Feature extraction is a process where a feature vector is constructed to represent the abstract from of the image. A feature vector represent the characteristics of the image which utilizes global image features like color or local descriptors like shape and texture. The second process in CBIR is query matching based on similarity measurement which uses the distance of the query from each image in database to find the closest image. A thorough and lucid summary of the literature survey on existing CBIR techniques is presented in [1–6].

As texture based CBIR has matured over the years, texture extraction centered about local patterns have emerged as frontrunners because of their relatively simplistic yet potent texture descriptor and its relative invariance to intensity changes. The underlining philosophy of all these descriptors is depicting the relationship between a pixel and its neighbors. The trailblazer of this method was Ojala et al. who proposed local binary pattern (LBP) [7]. Zhang et al. [8] modified the LBP method by introducing the concept of derivative and suggested local derivative patterns (LDP) method for face recognition. The problems related to variation in illumination for face recognition in LBP and LDP is sorted out in local ternary pattern (LTP) [9].

Several local patterns have been proposed by Subrahmanyam et al. including: local maximum edge patterns (LMEBP) [10], local tetra patterns (LTrP) [11] and directional local extrema patterns (DLEP) [12] for natural/texture image retrieval and directional binary wavelet patterns (DBWP) [13], local mesh patterns (LMeP) [14] and local ternary co-occurrence patterns(LTCoP) [15] for natural, texture and biomedical image retrieval applications.

Volume LBP (VLBP) is proposed by Zhao and Pietikainen [16] for dynamic texture (DT) recognition. They have collected the joint distribution of the gray levels for three consecutive frames of video for calculating the feature. Also, they used 3D-LBP features for lip-reading recognition [17]. Subrahmanyam and Wu [18] proposed the spherical symmetric 3-D LTP for retrieval purpose.

1.2 Main Contributions

The SS-3D-LTP [18] inspired us to propose the 3D-LTraP method. The main contributions of the 3D-LTraP method are as follows: (i) it is very clear that the performance of LTP is mostly depend on the proper selection of threshold value for ternary pattern calculation. An attempt has been made to resolve this problem. (ii) 3D-LTraP used five different planes at different scale with directional information whereas in SS-3D-LTP, both the information are missing. (iii) Color information is incorporated in the proposed method which is absent in the SS-3D-LTP.

This paper is organized as follows: Sect. 1gives overview of content based image retrieval for different applications. Section 2 gives the information about 2D LBP, SS-3D-LTP and 3D-LTraP. Section 3 explains the proposed algorithm and different performance measures. Sections 4 and 5 show experimental outcomes for natural and biomedical image database. Section 6, is dedicated to conclusion.

2 Local Patterns

2.1 Local Binary Patterns

Ojala et al. [7] devised the LBP method for face recognition application. For a given gray value of center pixel, LBP value is calculated using Eqs. 1 and 2 as presented in Fig. 1.

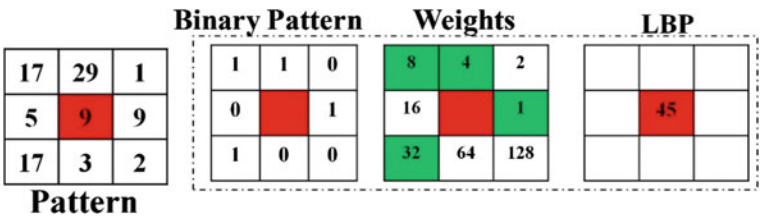


Fig. 1 LBP computation for subsample of given image

$$LBP_{P,R} = \sum_{i=1}^{P} 2^{(i-1)} \times f_1(g_i - g_c) \tag{1}$$

$$f_1(x) = \begin{cases} 1 & x \geq 0 \\ 0 & else \end{cases} \tag{2}$$

where g_c indicates middle pixel value, g_i represents pixel values of neighbors, P indicates total neighbors and R represents radius of the surrounding pixels.

2.2 Spherical Symmetric 3D-LTP (SS-3D-LTP)

Subrahmanyam and Wu [18] used the VLBP concept to define the spherical symmetric 3-D LTP for image retrieval application. From a given image, three multiresolution images are generated using 2D Gaussian filter bank. With these three multiresolution images, a 3D grid is constructed with five spherical symmetric directions. Next, neighbors are collected for each direction and 3D-LTP features are obtained. Detailed information of SS-3D-LTP is available in [18].

2.3 3D Local Transform Patterns (3D-LTraP)

Motivation behind the proposed 3D-LTraP method is the shortcomings of SS-3D-LTP method. The first hurdle in the performance of SS-3D-LTP is the proper selection of threshold value for ternary pattern calculation. Also, multiscale and color information is not utilized in SS-3D-LTP method. In the proposed method i.e. 3D-LTraP, the first problem is overcome by using binary approach. Similarly, the other lacuna is avoided by using wavelet transform. Also, the color feature is embedded in feature generation process.

Multiscale images are obtained by performing the wavelet transform on each color plane plane of input color image.

Figure 2 illustrates the directions of 3D LTraP method which formed using transform domain images R_T, G_T and B_T. These five spherical symmetric directions (α) are used for 3D-LTraP patterns generation. For the given center pixel $I_c(G_T)$; the patterns are collected based on direction α as follows:

$$V|_{P=8} = \begin{cases} v(I_0(G_T), I_1(G_T), I_2(G_T), I_3(G_T), I_4(G_T), I_5(G_T), I_6(G_T), I_7(G_T)); & \alpha = 1 \\ v(I_2(R_T), I_c(R_T), I_6(R_T), I_6(G_T), I_6(B_T), I_c(B_T), I_2(B_T), I_2(G_T)); & \alpha = 2 \\ v(I_5(R_T), I_c(R_T), I_1(R_T), I_1(G_T), I_1(B_T), I_c(B_T), I_5(B_T), I_5(G_T)); & \alpha = 3 \\ v(I_4(R_T), I_c(R_T), I_0(R_T), I_0(G_T), I_0(B_T), I_c(B_T), I_4(B_T), I_4(G_T)); & \alpha = 4 \\ v(I_3(R_T), I_c(R_T), I_7(R_T), I_7(G_T), I_7(B_T), I_c(B_T), I_3(B_T), I_3(G_T)); & \alpha = 5 \end{cases} \tag{3}$$

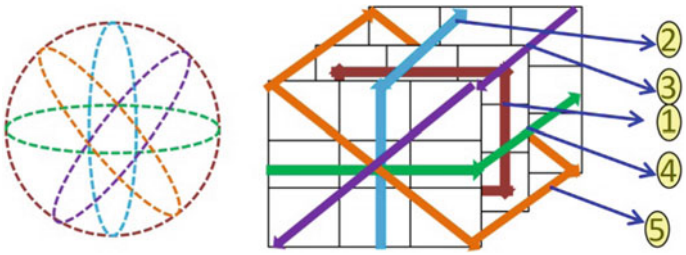


Fig. 2 Selected directions of 3D LTraP method

With specific direction α, 3D-LTraP values are calculated using Eqs. 4 and 5, considering relationship among middle pixel $I_c(G_T)$ and its surrounding pixels as given below:

$$V_\alpha|_{P=8} = \begin{cases} f(I_0(G_T)-I_c(G_T)), f(I_1(G_T)-I_c(G_T)), f(I_2(G_T)-I_c(G_T)), f(I_3(G_T)-I_c(G_T)), f(I_4(G_T)-I_c(G_T)), f(I_5(G_T)-I_c(G_T)), f(I_6(G_T)-I_c(G_T)), f(I_7(G_T)-I_c(G_T)); \alpha=1 \\ f(I_2(R_T)-I_c(G_T)), f(I_c(R_T)-I_c(G_T)), f(I_6(R_T)-I_c(G_T)), f(I_6(G_T)-I_c(G_T)), f(I_6(B_T)-I_c(G_T)), f(I_c(B_T)-I_c(G_T)), f(I_2(B_T)-I_c(G_T)), f(I_2(G_T)-I_c(G_T)); \alpha=2 \\ f(I_5(R_T)-I_c(G_T)), f(I_c(R_T)-I_c(G_T)), f(I_1(R_T)-I_c(G_T)), f(I_1(G_T)-I_c(G_T)), f(I_1(B_T)-I_c(G_T)), f(I_c(B_T)-I_c(G_T)), f(I_5(B_T)-I_c(G_T)), f(I_5(G_T)-I_c(G_T)); \alpha=3 \\ f(I_4(R_T)-I_c(G_T)), f(I_c(R_T)-I_c(G_T)), f(I_0(R_T)-I_c(G_T)), f(I_0(G_T)-I_c(G_T)), f(I_0(B_T)-I_c(G_T)), f(I_c(B_T)-I_c(G_T)), f(I_4(B_T)-I_c(G_T)), f(I_4(G_T)-I_c(G_T)); \alpha=4 \\ f(I_3(R_T)-I_c(G_T)), f(I_c(R_T)-I_c(G_T)), f(I_7(R_T)-I_c(G_T)), f(I_7(G_T)-I_c(G_T)), f(I_7(B_T)-I_c(G_T)), f(I_c(B_T)-I_c(G_T)), f(I_3(B_T)-I_c(G_T)), f(I_3(G_T)-I_c(G_T)); \alpha=5 \end{cases} \tag{4}$$

3D-LTraP values for a particular selected block is calculated using Eq. 5.

$$3D-LTrap_{\alpha,P,R} = \sum_{i=0}^{P-1} 2^i \, V_\alpha(i) \tag{5}$$

In LBP, the feature vector length is 2^P [7]. Further, feature vector length is reduced to $P(P-1)+2$ using uniform pattern suggested by Guo et al. [19].

$$H_S(l) = \frac{1}{N_1 \times N_2} \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} f_1(3D-LTraP(j,k), l); \;\; l \in [0, P(P-1)+2] \tag{6}$$

$$f_1(x,y) = \begin{cases} 1, & x=y \\ 0, & x \neq y \end{cases} \tag{7}$$

where, $N_1 \times N_2$ is the image size.

Figure 3 shows the calculation of 3D-LTraP feature of proposed method. Let select $\alpha = 2$, the respective patterns are collected as per the Eq. 3 are {35, 20, 15, 55, 72, 17, 90, 32}. The corresponding binary pattern for given center pixel (gray value "25") is {1, 0, 0, 1, 1, 0, 1, 1}. Finally, this binary pattern is converted into unique 3D-LTraP value of 195 as per Eq. 1.

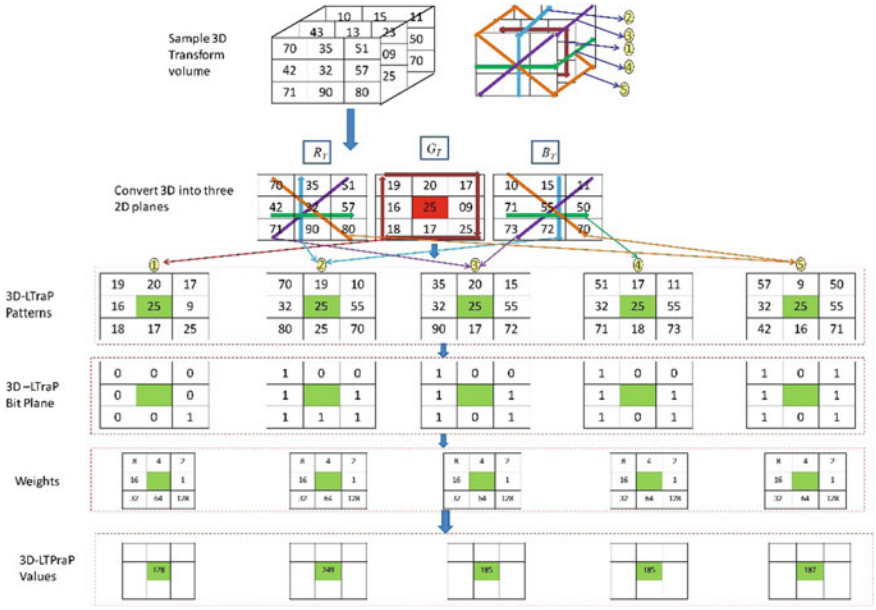


Fig. 3 Calculation of 3D-LTraP feature in five directions

3 Proposed Algorithm

3.1 Feature Extraction

The block schematic of proposed method is presented in Fig. 4 and its algorithm is given below:

Algorithm:

$H_S(l)$ 3D LTraP histogram
α spherical symmetric directions, $\alpha \in \{1, 2, \ldots N\}$
J Total decomposition levels in wavelet transform, $J \in \{1, 2, \ldots M\}$

Input: Color image; Output: 3D-LTraP *feature vector*

1. Read the input color image.
2. Apply wavelet transform to each *R, G* and *B* plane for *J* number of decomposition levels.
3. For $J = 1$ to M.

   - Create the volumetric representation (as shown in Fig. 3) using R_T, G_T and B_T transform images.
   - Generate 3D LTraP patterns in α spherical symmetric directions.
   - Generate histogram for each 3D LTraP patterns using Eq. 7.
   - Concatenated the histograms

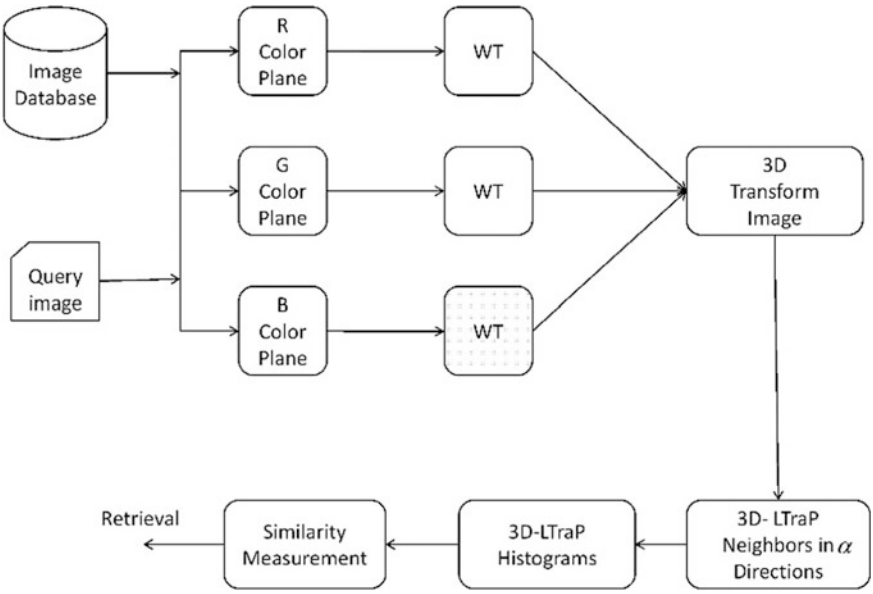


Fig. 4 Block schematic of proposed method

$$H_J = [H_{S_{\alpha=1}}, H_{S_{\alpha=2}}, \ldots, H_{S_{\alpha=N}}] \tag{8}$$

4. End of J.
5. Finally, 3D LTraP feature vectors are constructed by concatenating the histograms obtained at step 3 for each decomposition level.

$$H_{Final} = [H_{J=1}, H_{J=2}, \ldots, H_{J=M}] \tag{9}$$

3.2 Similarity Measurement

Similarity measurement is used to select similar images that look like the query image. 3D- LTraP features are collected from the given database as per the procedure discussed in Sect. 3.1. Query image feature vector is compared with the feature vector of images in the test database using the distance D (Eq. 10).

The distance is calculated as:

$$D(Q, I) = \sum_{i=1}^{N} \left| \frac{f_{I,i} - f_{Q,i}}{1 + f_{I,i} + f_{Q,i}} \right| \tag{10}$$

where

Q Query image;
N Feature vector length;
I Images in database;
$f_{I,i}$ ith feature of I^{th} image in the database;
$f_{Q,i}$ ith feature of query image Q

3.3 Evaluation Measures

Performance of proposed method is tested on the following evaluation measures:

- Average retrieval precision (ARP)
- Average retrieval rate (ARR)

Precision of query image Q is given as:

$$\mathrm{P(Q)} = \frac{\text{Number of relevant images retrieved}}{\text{Total number of images retrieved}} \tag{11}$$

Similarly, the recall is given as:

$$R(Q) = \frac{\text{Number of relevant images retrieved}}{\text{Total number of relevant images in the database}} \tag{12}$$

The ARP and ARR are defined as:

$$ARP = \frac{1}{N_q} \sum_{k=1}^{N_q} P(Q_k) \tag{13}$$

$$ARR = \frac{1}{N_q} \sum_{k=1}^{Nq} R(Q_k) \tag{14}$$

where

N_q Number of queries
Q_k kth image in the database

4 Experiments on Natural Image Database

Corel 10 K database [20], is used as natural image database which is a collection of 10000 images with 100 different categories. Each category contains 100 images of different groups i.e. Africa, Horses, Flowers, Food etc. with a sizes either 126 × 187 or 187 × 126. Precision, recall, ARP and ARR are the evaluation measures used to verify the performance of 3D-LTraP method.

Figure 5 shows the experimental outcomes on natural database. Figure 5a, b show performance of hundred groups of Corel 10-K database in terms of precision

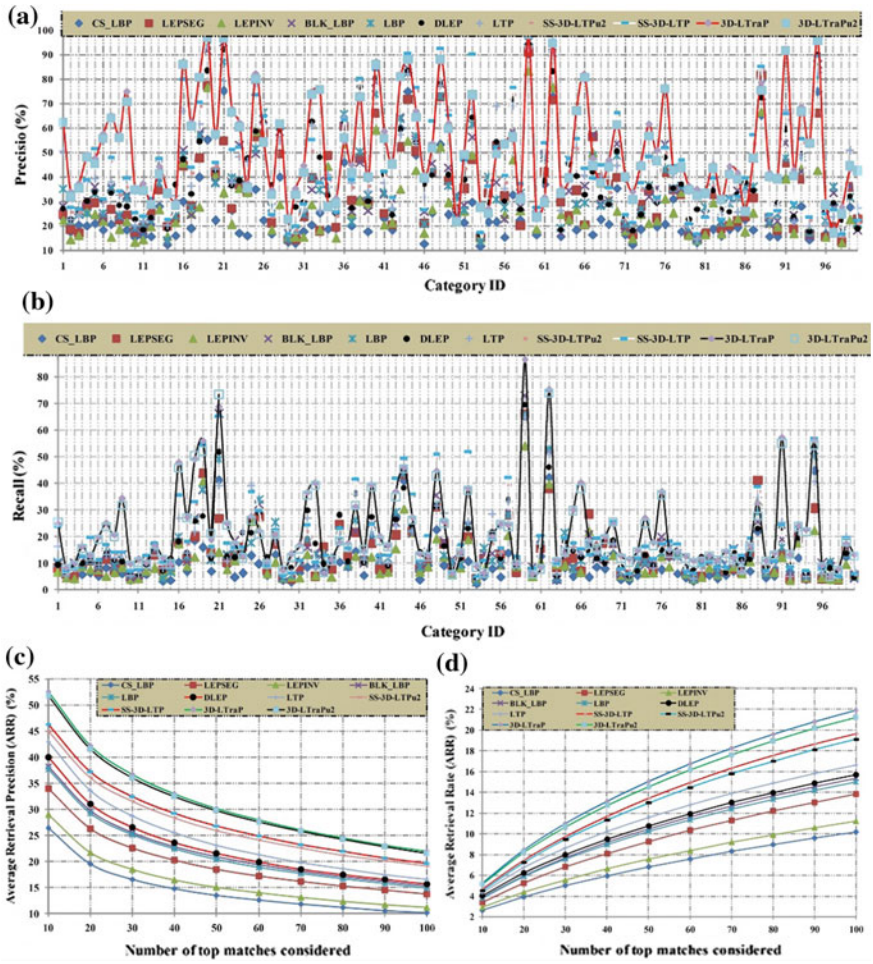


Fig. 5 **a** Average retrieval precision versus category. **b** Average retrieval rate versus category **c** Average retrieval precision versus number of top matches. **d** Average retrieval rate versus number of top matches

Table 1 Comparative results of average retrieval precision and average retrieval rate on natural database

Database	Performance	Method										
		CS_LBP	LEPSEG	LEPINV	BLK_LBP	LBP	DLEP	LTP	SS-3D-LTP	SS-3D-LTPu2	3D-LTraP	3D-LTrapu2
Corel–10 K	Precision (%)	26.4	34	28.9	38.1	37.6	40	42.95	46.25	44.97	**52.46**	**51.71**
	Recall (%)	10.1	13.8	11.2	15.3	14.9	15.7	16.62	19.64	19.09	**21.90**	**21.21**

and recall respectively. Similarly, Fig. 5c, d demonstrate the performance of 3D-LTraP method in terms of ARP and ARR. Table 1 and Fig. 5 demonstrate that 3D-LTraP and 3D-LTraPu2 methods outperform the other existing methods.

5 Experiments on Biomedical Image Database

Experiment is performed on the Open Access Series of Imaging Studies (OASIS) database [21]. OASIS database contained magnetic resonance imaging (MRI) of 421 subjects those aged in-between 18 and 96 years. For experimental work, 421 images are divided on the basis of shape of ventricular into four groups which contained 124, 102, 89 and 106 images in each group.

Figure 6 depicts that the performance of proposed method i.e. 3D-LTraP, 3D-LTraPu2 is outperforming the existing local pattern methods on the basis of average retrieval precision versus number of top matches.

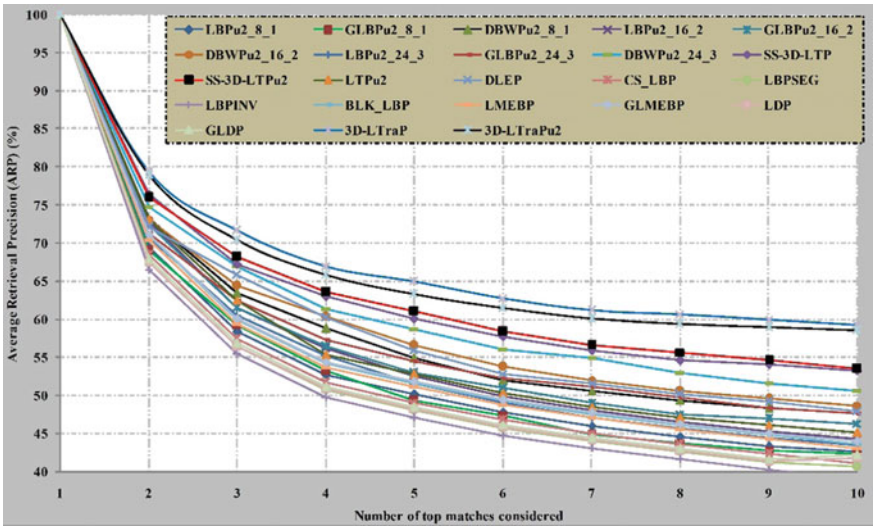


Fig. 6 Average retrieval precision versus number of top matches on biomedical image database

6 Conclusions

Proposed method resolved the problem of proper selection of threshold value in SS-3D-LTP method and add directional information using transform domain approach.

Performance of 3D-Ltrap is confirmed by performing the experimental work on natural and biomedical database and detailed outcomes are as below:

- ARP (%)/ARR (%) of 3D-LTraP show significant improvement from 26.4/10.1, 34/13.8, 28.9/11.2, 38.1/14.9, 37.6/14.9, 40/15.7, 42.95/16.62, 46.25/19.64 to **52.46/21.90** as compared to CS_LBP, LEPSEG, LEPINV, BLK_LBP, LBP, DLEP, LTP and SS-3D-LTP respectively on Core-10 K database.
- ARP (%) of 3D-LTraP has been improved from 47.05 %, 45.17 %, 53.32 % to **59.24** % as compared to DBWPu2-8-1, LTPu2 and SS-3D-LTP respectively on OASIS database.

References

1. M. L. Kherfi, D. Ziou and A. Bernardi. Image Retrieval from the World Wide Web: Issues, Techniques and Systems. ACM Computing Surveys, 36 35–67, 2004.
2. Ke Lu and Jidong Zhao, Neighborhood preserving regression for image retrieval. Neurocomputing 74 1467–1473, 2011.
3. Tong Zhaoa, Lilian H. Tang, Horace H.S. Ip, Feihu Qi, On relevance feedback and similarity measure for image retrieval with synergetic neural nets. Neurocomputing 51 105−124, 2003.
4. Kazuhiro Kuroda, Masafumi Hagiwara, An image retrieval system by impression words and specific object names-IRIS. Neurocomputing 43 259–276, 2002.
5. Jing Li, Nigel M. Allinson, A comprehensive review of current local features for computer vision. Neurocomputing 71 1771–1787, 2008.
6. Akgül C. B., Rubin D. L., Napel S., Beaulieu C. F., Greenspan H. and Acar B. Content-Based Image Retrieval in Radiology: Current Status and Future Directions. Digital Imaging, 24, 2 208–222, 2011.
7. T. Ojala, M. Pietikainen, D. Harwood. A comparative study of texture measures with classification based on feature distributions. Pattern Recognition, 29 51–59, 1996.
8. B. Zhang, Y. Gao, S. Zhao, J. Liu. Local derivative pattern versus local binary pattern: Face recognition with higher-order local pattern descriptor. IEEE Trans. Image Process., 19, 2 533–544, 2010.
9. X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process., 19, 6 1635–1650, 2010.
10. Subrahmanyam Murala, R. P. Maheshwari, R. Balasubramanian. Local Maximum Edge Binary Patterns: A New Descriptor for Image Retrieval and Object Tracking. Signal Processing, 92 1467–1479, 2012.
11. Subrahmanyam Murala, Maheshwari RP, Balasubramanian R. Local tetra patterns: a new feature descriptor for content based image retrieval. IEEE Trans. Image Process 2012; 21(5): 2874–86.
12. Subrahmanyam Murala, R. P. Maheshwari, R. Balasubramanian. Directional local extrema patterns: a new descriptor for content based image retrieval. Int. J. Multimedia Information Retrieval, 1, 3 191–203, 2012.

13. Subrahmanyam Murala, R. P. Maheshwari, R. Balasubramanian. Directional Binary Wavelet Patterns for Biomedical Image Indexing and Retrieval. Journal of Medical Systems, 36, 5 2865–2879, 2012.
14. Subrahmanyam Murala, Jonathan Wu QM. Local mesh patterns versus local binary patterns: biomedical image indexing and retrieval. IEEE J Biomed Health In format 2013, http://dx.doi.org/10.1109/JBHI.2013.2288522.
15. Subrahmanyam Murala, Jonathan Wu QM. Local ternary co-occurrence pat-terns: a new feature descriptor for MRI and CT image retrieval. Neurocomputing, 119 (7):399–412, 2013.
16. G. Zhao, M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell., 29, 6 915–928, 2007.
17. Guoying Zhao, Mark Barnard, and Matti Pietikäinen. Lipreading with Local Spatiotemporal Descriptors. IEEE Trans. Multimedia, 11, 7 1254−1265, 2009.
18. Subrahmanyam Murala and Q. M. Jonathan Wu, "Spherical Symmetric 3D Local Ternary Patterns for Natural, Texture and Biomedical Image Indexing and Retrieval," *Neurocomput*ing, 149 (C) 1502–1514, 2015.
19. Z. Guo, L. Zhang and D. Zhang. Rotation invariant texture classification using LBP variance with global matchning. Pattern recognition, 43 706–716, 2010.
20. Corel-10 K image database. [Online]. Available: http://www.ci.gxnu.edu.cn/cbir/Dataset.aspx.
21. D. S.Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C.Morris, and R. L. Buckner. Open access series of imaging studies (OASIS): Crosssectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci., 19, 9 1498–1507, 2007.

Quaternion Circularly Semi-orthogonal Moments for Invariant Image Recognition

P. Ananth Raj

Abstract We propose a new Quaternion Circularly Semi-Orthogonal Moments for color images that are invariant to rotation, translation and scale changes. In order to derive these moments we employ the recently proposed Circularly Semi-Orthogonal Moment's expression. Invariant properties are verified with simulation results and found that they are matching with theoretical proof.

Keywords Circularly semi orthogonal moments · Exponent fourier moments

1 Introduction

Processing and analysis of color images represented by a quaternion algebra provides better results than the traditional methods like processing of three (R, G, B) images separately because in the quaternion representation, a color image pixel is treated as a single unit. One of the first persons who employed the quaternion algebra for color images is Sangwine [1]. Since then may techniques like Fourier transform [2, 3], Winer filter [4], Zernike moments [5, 6], Disc-Harmonic moments [7–11], Legendre-Fourier moments [12], FourierMellin moments [13] and Bessel Fourier moments [14–16] are extended to color images using the Quaternion algebra. Recently, Karakasis et al. [17] published a unified methodology for computing accurate quaternion color moments. Another recent paper by Chen et al. [18] suggested a general formula for quaternion representation of complex type moments.

All these moments provided mixed results for image reconstruction, object recognition and water marking problems. Recently, Hu et al. [19] proposed an Exponent Fourier moments for gray level images. These moments are similar to Polar Complex Exponential Transforms and Radial Harmonic Fourier Moments.

P. Ananth Raj (✉)
Department of ECE, College of Engineering, Osmania University,
Hyderabad 500 007, Telangana, India
e-mail: ananthraj0123@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_46

Exponent Fourier moments are computationally inexpensive as compared with other moments like Zernike Bessel Fourier moments [17]. Xia et al. [20] pointed out some errors in the above paper and proposed a better radial function. Unfortunately, this expression turns out to be incorrect. Hence, Hu et al. [21] recently suggested an improved version of it. Most of these moments suffer from numerical and geometric errors. In order to minimize these errors, Wang et al. [22] proposed a Circularly Semi-Orthogonal (CSO) moments for both binary and multilevel images only. Hence, in this paper, we extend the CSO moments proposed for gray level images to color images using the algebra of quaternion and propose a new quaternion circularly semi orthogonal (QCSO) moments for color images. Further, we also derive invariants properties of QCSO moments and verified them with simulation results. In this study we have chosen CSO moments because of the following advantages: (a) Higher order moments are numerically stable than the lower order moments. (b) Zeroth order approximation is more robust to numerical errors compared with other approximations and (c) no factorial terms in the radial function definition

This paper is organized into eight sections. In Sect. 2 we present the quaternion number system. In Sect. 3 circularly semi orthogonal moments are discussed in detail. In Sects. 4 and 5, expressions are derived for Quaternion circularly semi orthogonal forward and inverse moments. Invariant properties of QCSO moments are derived in Sect. 6. Finally, simulation results and conclusions are presented in Sects. 7 and 8 respectively.

2 Quaternion Number System

These numbers are extensions of complex numbers, that consists of one real part and three imaginary parts. A quaternion number with zero real part is called pure quaternion. Quaternion number system was introduced by the mathematician Hamilton [23] in 1843. Then Sang wine [1, 23] applied them for color image represntation. A quaternion number q is written as

$$q = a + bi + cj + dk. \tag{1}$$

Where a, b, c and d are real numbers, i, j and k are orthogonal unit axis vectors satisfies the following rules

$$i^2 = j^2 = k^2 = -1,\ ij = -ji = k \tag{2}$$

$$jk = -kj = i, ki = -ik = j$$

From these equations one can say that quaternion multiplication is not commutative. Both conjugate and modulus of a quaternion number q is

$$\bar{q} = a - bi - cj - dk$$
$$|q| = \sqrt{a^2 + b^2 + c^2 + d^2}$$

For any two quaternion numbers say p and q we have $\overline{p.q} = \bar{p}.\bar{q}$. Quaternion representation of a pixel in a color image is

$$f(x, y) = f_R(x, y)i + f_G(x, y)j + f_B(x, y)k \tag{3}$$

It is assumed that real part is zero. In the above expression $f_R(x, y), f_G(x, y)\, and f_B(x, y)$ represents the red, green and blue components of a color pixel, similarly, polar representation of an image using the quaternion representation is

$$f(r, \theta) = f_R(r, \theta)i + f_G(r, \theta)j + f_B(r, \theta)k \tag{4}$$

In this expression $f_R(r, \theta), f_G(r, \theta)\, and f_B(r, \theta)$ denote red, green and blue components of polar representation of image.

3 Circularly Semi-orthogonal Moments

Let $f(r, \theta)$ be the polar representation of a gray level image of size N × M, then the general expression for circularly orthogonal moments (E_{nm}) of order n and repetition m of a polar image $f(r, \theta)$ is

$$E_{nm} = \frac{1}{Z} \int_{r=0}^{1} \int_{\theta=0}^{2\pi} f(r, \theta) T_n(\mathrm{r}) \exp(-jm\theta) r dr d\theta \tag{5}$$

where n = ±0, ±1, ±2 ..... and $m = \pm 0, \pm 1, \pm 2 \ldots\ldots$ are the moment order and repetition of a radial function $T_n(r)$ and $T_n^*(r)$ is the conjugate of radial function $T_n(r)$. It is noted from the available literature that, most of the circularly orthogonal moments differ only in Radial functions and normalization constants. Hence, Table 1 shown below lists some of the radial functions proposed recently for defining moments.

Xia et al. [20] pointed out some errors in the radial function of Exponent Fourier moment-I. In order to correct them, a new expression was suggested for Exponent Fourier moments-I, which is given by

$$E_{nm} = \frac{1}{2\pi a_n} \int_{r=\pi}^{\pi+1} \int_{\theta=0}^{2\pi} f(r, \theta) T_n(\mathrm{r}) \exp(-jm\theta) r d\theta dr$$

Table 1 Radial functions and normalization constants of various moments

Moments	Radial function $T_n(r)$	Normalization constant $\frac{1}{Z}$
Exponent fourier—I	$\sqrt{\frac{1}{r}}e^{-j2\pi nr}$	$\frac{2}{\pi(M^2+N^2)}$
Circularly semi orthogonal moments	$(15)^{-\frac{r}{4}}\sin(n+1)\pi r$	$\frac{2}{\pi(M^2+N^2)}$
Polar harmonic	$e^{-j2\pi nr^2}$	$\frac{4}{\pi(M^2+N^2)}$
Exponent fourier—II	$\sqrt{\frac{2}{r}}e^{-j2\pi nr}$	$\frac{1}{\pi(M^2+N^2)}$

where $a_n = \exp(-j4n\pi^2)$ is normalization constant. Modified radial function $T_r(r)$ is given by

$$T_n(r) = \sqrt{\frac{1}{\pi + r}}e^{-j2\pi n(r+\pi)}$$

Recently, Hu et al. [21] suggested an improved version of the above expression, because incorrect orthogonality condition was employed to find the normalization constant and when the limits for r are applied, no value will be within the 0 to 2 π range. Hence, an Improved Exponent Fourier moments (IEF) moment expression is suggested and it is given by.

$$E_{nm} = \frac{1}{2\pi}\int_{r=k}^{k+1}\int_{\theta=0}^{2\pi} f(r-k,\theta)T_n(r)\exp(-jm\theta)rd\theta dr$$

where k is a non negative integer and $f(r-k,\theta)$ is the translated version of original image $f(r,\theta)$. Another circularly semi orthogonal moments whose radial bases functions are same for forward and inverse transforms was proposed by Wang et al. [22]. Like the other radial bases functions this radial bases function also do not use factorial terms. Hence, it is computationally not expensive. Figure 1 display the graphs of real parts of $T_n(r)$ for orders n = 0, 1, 2, 3, 4 and 5. These graphs are numerically stable and avoids the large value in the above expression when r = 0. Given a finite number of moments (N_{max} and M_{max}) image reconstruction can be obtained using the expression given below

$$f(r-k,\,\theta) = \sum_{n=1}^{N_{max}}\sum_{m=1}^{M_{max}} E_{nm}T_n^*(r)\exp(jm\theta)$$

According to the above equation, pixels of reconstructed image must be shifted by a distance k along the opposite direction. If we substitute $T_n(r) = (15)^{-\frac{r}{4}}\sin(n+1)\pi r$ and Z = 2π in Eq. (5), we obtain an expression for Circularly semi orthogonal moments that is given below

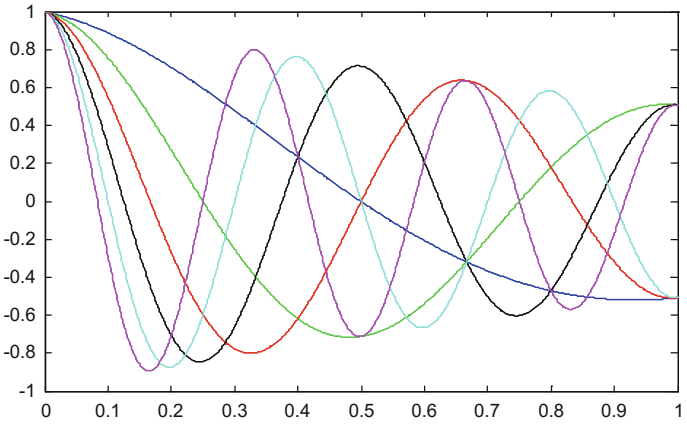


Fig. 1 Real part of radial function $T_n(r)$ for n = 0, 1, 2, 3, 4, 5 values. x axis denotes 'r' values (1 to 2, in steps of 0.001) and y real part of radial function. *Colors* n = 0, *blue*, n = 1, *green*, n = 2, *red*, n = 3 *black*, n = 4, *cyan*, n = 5 *magenta*

$$E_{nm} = \frac{1}{Z} \int_{r=0}^{1} \int_{\theta=0}^{2\pi} f(r,\theta)(15)^{-\frac{r}{4}} \sin(n+1)\pi r \exp(-jm\theta) r dr d\theta \tag{6}$$

In order to convert the above equation suitable for 2D images of size M × N, we need a polar representation of the image and replace the integrals by summation. In our work we employed the 'Outer Unit Disk Mapping' employed by Hu et al. [19] which fixes the image pixels inside the unit circle. Expression for coordinate mapping is

$$x_i = \frac{2i+1-N}{\sqrt{M^2+N^2}} \quad x_j = \frac{2j+1-M}{\sqrt{M^2+N^2}} \tag{6a}$$

where i = 0, 1…M, j = 0, 1, …N. Final expression when zeroth order approximation of Eq. (6) is used

$$E_{nm} = \frac{1}{Z} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} f(x_i, y_j) T_n(r_{ij}) e^{-jm\theta_{ij}} \tag{7}$$

where $\frac{1}{Z} = \frac{2}{\pi(M^2+N^2)}$, $r_{ij} = \sqrt{x_i^2 + y_j^2}$ and $\theta_{ij} = \tan^{-1}\frac{y_j}{x_i}$

More details can be seen in paper [19]. Next section presents Quaternion circularly semi orthogonal moments.

4 Quaternion Circularly Semi-orthogonal Moments

According to the definition of circularly semi-orthogonal moments (CSOM) (Eq. 6) for gray levels and quaternion algebra, the general formula for the right side CSOM of a color image $f(r,\theta)$ of order n with repetition m is

$$E_{nm}^{R} = \frac{1}{2\pi} \int_{r=0}^{1} \int_{\theta=0}^{2\pi} f(r,\theta) T_n(r) \exp(-\mu m\theta) r d\theta dr$$

where μ is a unit pure quaternion, generally it is a linear combination of i, j and k such that its magnitude is unity. In this work μ is taken as $\mu = \frac{i+j+k}{\sqrt{3}}$. Quaternion is not commutative. Hence, we obtain left side expression, which is given by

$$E_{nm}^{L} = \frac{1}{2\pi} \int_{r=0}^{1} \int_{\theta=0}^{2\pi} \exp(-\mu m\theta) f(r,\theta) T_n(r) r d\theta dr \tag{8}$$

In this work we consider only right side expression and drop out the symbol R. Relationship between these two expressions is $E_{nm}^{L} = -\overline{E_{n,-m}^{R}}$, it can be derived using the conjugate property. Next, we derive an expression for implementation of E_{nm},. substituting Eq. (4) into Eq. (8), we get

$$E_{nm} = \frac{1}{2\pi} \int_{r=0}^{1} \int_{\theta=0}^{2\pi} T_n(r) [f_R(r,\theta) i + f_G(r,\theta) + f_B(r,\theta) k] \exp(-\mu m\theta) r d\theta dr \tag{9}$$

Let

$$A_{nm} = \frac{1}{2\pi} \left[\int_{r=0}^{1} \int_{r=0}^{2\pi} f_R(r,\theta) T_n(r) \exp(-\mu m\theta) r d\theta dr \right]$$

$$B_{nm} = \frac{1}{2\pi} \left[\int_{r=0}^{1} \int_{\theta=0}^{2\pi} f_G(r,\theta) T_n(r) \exp(-\mu m\theta) r d\theta dr \right]$$

$$C_{nm} = \frac{1}{2\pi} \left[\int_{r=0}^{1} \int_{\theta=0}^{2\pi} f_B(r,\theta) T_n(r) \exp(-\mu j m\theta) r d\theta dr \right]$$

Equation (9) can be written as

$$E_{nm} = iA_{nm} + jB_{nm} + kC_{nm}.$$

A_{nm}, B_{nm} and C_{nm} are complex values, hence, the above equation can be expressed as

$$E_{nm} = i\left(A_{nm}^{R} + \mu A_{nm}^{I}\right) + j\left(B_{nm}^{R} + \mu B_{nm}^{I}\right) + k\left(C_{nm}^{R} + \mu C_{nm}^{I}\right)$$

Substituting for $\mu = \frac{i+j+k}{\sqrt{3}}$ and simplifying the above expression using Eq. 2 we get

$$E_{nm} = i\left(A_{nm}^{R} + \frac{(i+j+k)}{\sqrt{3}}A_{nm}^{I}\right) + j\left(B_{nm}^{R} + \frac{(i+j+k)}{\sqrt{3}}B_{nm}^{I}\right) + k\left(C_{nm}^{R} + \frac{(i+j+k)}{\sqrt{3}}C_{nm}^{I}\right)$$

$$E_{nm} = -\frac{1}{\sqrt{3}}\left[A_{nm}^{I} + B_{nm}^{I} + C_{nm}^{I}\right] + i\left[A_{nm}^{R} + \frac{1}{\sqrt{3}}(B_{nm}^{I} - C_{nm}^{I})\right] + j\left[B_{nm}^{R} + \frac{1}{\sqrt{3}}(C_{nm}^{I} - A_{nm}^{I})\right] + k\left[C_{nm}^{R} + \frac{1}{\sqrt{3}}(A_{nm}^{I} - B_{nm}^{I})\right].$$

In order to express the above equation in a better way we let $S1 = -\frac{1}{\sqrt{3}}\left[A_{nm}^{I} + B_{nm}^{I} + C_{nm}^{I}\right]$, $S2 = \left[A_{nm}^{R} + \frac{1}{\sqrt{3}}\left(B_{nm}^{I} - C_{nm}^{I}\right)\right]$, $S3 = \left[B_{nm}^{R} + \frac{1}{\sqrt{3}}(C_{nm}^{I} - A_{nm}^{I})\right]$ and $S4 = \left[C_{nm}^{R} + \frac{1}{\sqrt{3}}(A_{nm}^{I} - B_{nm}^{I})\right]$, now the above equation can be expressed as

$E_{nm} = S1 + iS2 + jS3 + kS4$. In next we derive the expression for quaternion inverse circularly semi-orthogonal moments.

5 Quaternion Inverse Circularly Semi-orthogonal Moments

Given a finite number up to a given order L of Quaternion Circularly Semi orthogonal moments, we find the approximated image $f(r, \theta)$ using the equation given below

$$f(r, \theta) = \sum_{n=0}^{L} \sum_{m=-L}^{L} E_{nm} T_n(r) e^{\mu m \theta} \tag{10}$$

Substituting E_{nm} from the above equation we get

$$f(r, \theta) = \sum_{n=0}^{L} \sum_{m=-L}^{L} (S1 + iS2 + jS3 + kS4) T_n(r) e^{\mu m \theta}$$

Substituting, $\mu = \left(\frac{i+j+k}{\sqrt{3}}\right)$ in the above expression and after simplification we get expression for inverse QCSO moments as

$$f(r,\theta) = \overline{f_{s1}}(r,\theta) + i\overline{f_{s2}}(r,\theta) + j\overline{f_{s3}}(r,\theta) + k\overline{f_{s4}}(r,\theta)$$

where each term is equal to

$$\overline{f_{s1}}(r,\theta) = real(\overline{s_1}) - \frac{1}{\sqrt{3}}[imag(\overline{s_2}) + imag(\overline{s_3}) + imag(\overline{s_4})]$$
$$\overline{f_{s2}}(r,\theta) = real(\overline{s_2}) + \frac{1}{\sqrt{3}}[imag(\overline{s_1}) + imag(\overline{s_3}) - imag(\overline{s_4})]$$
$$\overline{f_{s3}}(r,\theta) = real(\overline{s_3}) + \frac{1}{\sqrt{3}}[imag(\overline{s_1}) - imag(\overline{s_2}) + imag(\overline{s_4})]$$
$$\overline{f_{s4}}(r,\theta) = real(\overline{s_4}) + \frac{1}{\sqrt{3}}[imag(\overline{s_1}) + imag(\overline{s_2}) - imag(\overline{s_3})]$$

In this expression real (.) and imag (.) terms denote real and imaginary part of the value within the bracket. Each term represents the reconstruction matrix of s1, s2, s3 and s4 respectively and they are determined using

$$\overline{s_1} = \sum_{n=0}^{L}\sum_{m=-L}^{L} S_1 T_n(r) e^{jm\theta}, \overline{s_2} = \sum_{n=0}^{L}\sum_{m=-L}^{L} S_2 T_n(r) e^{jm\theta}$$
$$\overline{s_3} = \sum_{n=0}^{L}\sum_{m=-L}^{L} S_3 T_n(r) e^{jm\theta}, \overline{s_4} = \sum_{n=0}^{L}\sum_{m=-L}^{L} S_4 T_n(r) e^{jm\theta}$$

These expressions can be easily implemented using equation Eq. 7.

6 Invariant Properties of QCSO Moments

Let (r,θ) and $f(r,\theta-\varphi)$ be the un rotated and rotated (by an angle φ) images expressed in polar form, then QCSO moments of a rotated image is

$$E_{nm} = \frac{1}{2\pi}\int_{r=0}^{1}\int_{\theta=0}^{2\pi} f(r,\theta-\varphi)T_n(r)\exp(-\mu m\theta) r d\theta dr$$

Let $\bar{\theta} = \theta - \varphi$ then d $d\theta = d\bar{\theta}$, substituting it in the above equation we obtain a rotated QCSO moments as

$$E^r_{nm} = E_{nm}\exp(-\mu m\varphi)$$

Applying modulus operation on both sides of the above equation we get

$$\left\|E_{nm}^{r}(f)\right\| = \|E_{nm}(f)\|.1$$

Rotation of an image by an angle φ does not change the magnitude, but the phase changes from $-\mu m\theta$ *to* $-(\mu m\theta + \mu m\varphi)$. Hence, we say that rotation does not change magnitude, therefore it is invariant to rotation. Translation invariant is achieved by using the common centroid (x_c, y_c) obtained using R, G, B images. This procedure was suggested by Fluser [24] and employed by number of researchers like Chen et al. [18], Nisrine Das et al. [7]. Procedure consists of fixing the origin of the coordinates at the color image centroid obtained using

$$x_c = \frac{(m_{1,0}^{R} + m_{1,0}^{G} + m_{1,0}^{B})}{m_{00}}, y_c = \frac{(m_{0,1}^{R} + m_{0,1}^{G} + m_{0,1}^{B})}{m_{00}}, m_{0,0} = m_{00}^{R} + m_{00}^{G} + m_{00}^{B},$$

where m_{00}^{R}, m_{10}^{R}, m_{01}^{R} are the geometric moments of R image, whereas G and B superscripts denote green and blue images, using the above coordinates, QCSO moments invariants to translation is given by

$$E_{nm}\frac{1}{2\pi}\int_{r=0}^{1}\int_{\theta=0}^{2\pi} f(\bar{r},\bar{\theta})T_n(\bar{r})\exp(-\mu m\bar{\theta})\bar{r}d\bar{r}d\bar{\theta} \tag{11}$$

where $\bar{r} = \sqrt{(x-x_c)^2 + (y-y_c)^2}$ $\bar{\theta} = \tan^{-1}\left(\frac{y-y_c}{x-x_c}\right)$.

Moments calculated using the above expression are invariant to translation. In most of the applications like image retrieval, images are scaled moderately, then the scale invariant property is fulfilled automatically, because, QCSO moments are defined on the unit circle using Eq. 6a [10].

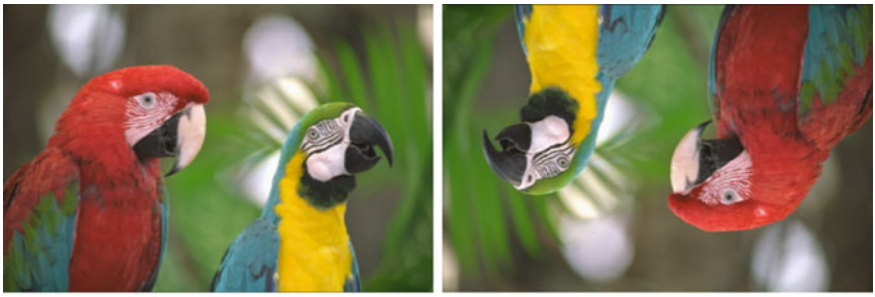
Another useful property is flipping an image either vertical or horizontal. Let $f(r,\theta)$ be the original image, $f(r,-\theta)$ and $f(r,\pi-\theta)$ be the vertical and horizontal flipped images. One of the color images flipped vertically and horizontally are shown in Fig. 2. We derive its QCSO moments. QCSO moments of the flip vertical image is

$$E_{nm}^{V} = \frac{1}{2\pi}\int_{\theta=0}^{2\pi}\int_{r=0}^{\infty} f(r,-\theta)T_n(r)e^{-\mu m\theta}rdrd\theta$$

Substituting $\varnothing = -\theta$, $d\varnothing = -d\theta$, and Simplifying the above equation, we obtain the moments as

$$E_{nm}^{v} = -E_{nm}^{\cdot}$$

where $E_{nm}^{\cdot}$ is the complex conjugate of the QCSO moments. QCSO moments of flip horizontal image is given by



(a) Vertical flipped image (b) Horizontal flipped image

(c) Rotated image 2 deg clockwise (d) Translated image x=2, y=1 units

Fig. 2 Flipped, rotated and translated images

$$E_{nm}^{h} = \frac{1}{2\pi} \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} f(r, \pi - \theta) T_n(r) e^{-\mu m \theta} r dr d\theta$$

Substituting $\varnothing = \pi - \theta$, $d\varnothing = -d\theta$, we obtain the moments as

$$E_{nm}^{h} = -E_{nm}^{\cdot} e^{\frac{-(i+j+k)}{\sqrt{3}} m\pi}$$

Hence, one can compute the flipped image moments using the above equation. Some of the properties like invariance to contrast changes can be verified by normalizing moments by E_{00}.

7 Simulation Results

In order to verify the proposed Quaternion circularly semi orthogonal moments for both reconstruction capability and invariant for rotation, translation, and flipping we have selected four color images namely, Lion image, Vegetable image, Parrot

(c) Reconstructed Lion image

(d) Reconstructed Parrot Image

(e) Original Image

f) Reconstructed Image using QCSO

(g) Original Image

(h) Reconstructed image using QCSO

Fig. 3 Original and reconstructed Images using QCSO Moments

Table 2 Rotation invariants of QCSO moments

Moments	Rotated vegetable image	Rotated parrot image	Rotated Lion image	Monalisa image	Rotated image
$\|E_{0,0}\|$	0.5188	0.398	0.1961	0.382	0.3792
$\|E_{2,0}\|$	0.0331	0.0343	0.0716	0.0396	0.0393
$\|E_{3,1}\|$	0.0413	0.033	0.0320	0.1055	0.1034
$\|E_{4,2}\|$	0.0186	0.0131	0.0012	0.0114	0.0109
$\|E_{5,1}\|$	0.0283	0.0277	0.0123	0.0594	0.05801

Table 3 Translation invariants of QCSO moments

x = 2, dy = 1	Vegetable image	Translated image	Lion image	Translated image
$\|E_{0,0}\|$	0.5274	0.4944	0.1987	0.1803
$\|E_{2,0}\|$	0.0338	0.029	0.0723	0.0608
$\|E_{3,1}\|$	0.0415	0.0397	0.0325	0.037
$\|E_{4,2}\|$	0.0196	0.0182	0.0012	0.0021
$\|E_{5,1}\|$	0.0276	0.0203	0.0123	0.0102

Table 4 Original and flipped image QCSO moments

Vertical flipped	Parrot image	Flipped image	Lion image	Flipped image
$\|E_{0,0}\|$	0.399	0.4052	0.1987	0.1987
$\|E_{2,0}\|$	0.0338	0.0328	0.0723	0.0726
$\|E_{3,1}\|$	0.0330	0.0289	0.0325	0.0324
$\|E_{4,2}\|$	0.0134	0.0137	0.0012	0.0012
$\|E_{5,1}\|$	0.0275	0.0267	0.0123	0.0124

image and painted Mona Lisa images and computed their QCSO moments and these color images are shown in Fig. 2, down loaded from Amsterdam Library of objects, reconstructed using only moments of order 40 (L = 40 in Eq. 10). Obtained results are shown in Fig. 3. High frequency information like edges is well preserved. These images (Lion image is shown in Fig. 2) are rotated by 2 degrees in clock wise using IMROTATE function available in MATLAB 2010 software and magnitude of only few QCSO moments are computed and results are reported in Table 2. From these results we can note that before and after rotation QCSO moments are almost equal. We have also verified translated property by translating Lion image and Vegetable images by 2 units in x direction (dx = 2) and 1 unit (dy = 1) in y direction. Their results (magnitude of E_{nm}) computed using Eq 11 are shown in Table 3. Difference, between the moments before and after translation is very small. Finally, moments (magnitude of E_{nm}) calculated for flipped vertical images are reported in Table 4.

8 Conclusions

In this paper we proposed a new Quaternion circularly semi orthogonal moments for color images. Further, we have showed that these moments are invariant to rotation, scale, translation and we also derived moments for flipped color images. Presently, we are applying these moments both for color and monochrome super resolution problems.

Acknowledgements Author would like to thank Dr. B. Rajendra Naik, Present ECE Dept Head and Dr. P. Chandra Sekhar, former ECE Dept Head for allowing me to use the Department facilities even after my retirement from the University.

References

1. S.J. Sangwine, "Fourier Transforms of color images using quaternion or Hypercomplex, numbers", Electronics Letters vol 32, Oct, 1996.
2. T.A. Ell and S.J. Sangwine, "Hyper-complex Fourier Transforms of color images", IEEE Trans Image Processing, vol 16, no1 Jan 2007.
3. Heng Li Zhiwen Liu, yali Huang Yonggang Shi, " Quaternion generic Fourier descriptor for color object recognition, " Pattern recognition, vol 48, no 12, Dec 2015.
4. Matteo Pedone, Eduardo, Bayro-Corrochano, Jan Flusser, Janne Hakkila, "Quaternion wiener deconvolution for noise robust color image registration", IEEE Signal Processing Letters, vol 22, no 9, Sept, 2015.
5. C.W. Chong, P. Raveendran and R. Mukundan, " The scale invariants of Pseudo Zernike moments", Patter Anal Applic, vol 6 no pp 176–184, 2003.
6. B.J. Chen, H.Z. Shu, H. Zhang, G. Chen, C. Toumoulin, J.L. Dillenseger and L.M. Luo, "Quaternion Zernike moments and their invariants for color image analysis and object recognition", Signal processing, vol 92, 2012.
7. Nisrine Dad, Noureddine Ennahnahi, Said EL Alaouatik, Mohamed Oumsis, " 2D color shapes description by quaternion Disk harmonic moments", IEEE/ACSint Conf on AICCSA, 2014.
8. Wang Xiang-yang, Niu Pan Pan, Yang Hong ying, Wang Chun Peng, Wang Ai long, " A new Robust color image watermarking using local quaternion exponent moments", Information sciences, vol 277, 2014.
9. Wang Xiang, Li wei-yi, Yang Hong-ying, Niu Pan Pan, Li Yong-wei, "Invariant quaternion Radial Harmonic fourier moments for color image retrieval", Optics and laser Technology, vol 66, 2015.
10. Wang Xiang-yang, Wei-yi Li, Yang Hong-ying, Yong-wei Li " Quaternion polar complex exponential transform for invariant color image description", Applied mathematical and computation, vol 256, April, 2015.
11. Pew-Thian Yap, Xudong Jiang Alex Chi Chung Kot, " Two Dimensional polar harmonic transforms for inavariant image representation, IEEE Trans on PAMI (reprint).
12. C. Camacho. Bello, J.J. Baez-Rojas, C. Toxqui-Quitl, A. Padilla-vivanco, "Color image reconstruction using quaternion Legendre-Fourier moments in polar pixels". Int Cof on Mechatronics, Electronics and Automotive Engg, 2014.
13. J. Mennesson, C. Saint-Jean, L. Mascarilla, " Color Fourier-Mellin descriptors for Image recognition", Pattern Recognit, Letters vol, 40, 2014 pp.

14. Zhuhong Sha, Huazhong Shu, Jiasong wu, Beijing Chen, Jean Louis, Coatrieux, “Quaternion Bessel Fourier moments and their invariant descriptors for object reconstruction and recognition”, Pattern Recognit, no 5, vol 47, 2013.
15. Hongqing Zho, Yan Yang, Zhiguo Gui, Yu Zhu and Zhihua Chen, “Image Analysis by Generalized Chebyshev Fourier and Generalized Pseudo Jacobi-fourier Moments, Pattern recognition (accepted for publication).
16. Bin Xiao, J. Feng Ma, X. Wang, “ Image analysis by Bessel Fourier moments”, Pattern Recognition, vol 43, 2010.
17. E. Karakasis, G. Papakostas, D. Koulouriotis and V. Tourassis, “A Unified methodology for computing accurate quaternion color moments and moment invariants”, IEEE Trans. on Image Processing vol 23 no 1, 2014.
18. B. Chen, H. Shu, G. Coatrieux, G. Chen X. Sun, and J.L. Coatrieux, “Color image analysis by quaternion type moments”, journal of Mathematical Imaging and Vision, vol 51, no 1, 2015.
19. Hai-tao-Hu, Ya-dong Zhang Chao Shao, Quan Ju, “Orthogonal moments based on Exponent Fourier moments”, Pattern Recognition, vol 47, no 8, pp. 2596–2606, 2014.
20. Bin Xiao, Wei-sheng Li, Guo-yin Wang, “ Errata and comments on orthogonal moments based on exponent functions: Exponent-Fourier moments”, Recognition vol 48, no 4, pp 1571–1573, 2015.
21. Hai-tao Hu, Quan Ju, Chao Shao, “ Errata and comments on Errata and comments on orthonal moments based on exponent functions: Exponent Fourier moments, Pattern Recognition (accepted for publication).
22. Xuan Wang, Tengfei Yang Fragxia Guo, “Image analysis by Circularly semi orthogonal moments”, Pattern Recognition, vol 49, Jan 2016.
23. W.R. Hamilton, Elements of Quaternions, London, U.K. Longman 1866.
24. T. Suk and J. Flusser, “Affine Moment Invariants of Color Images”, Proc CAIP 2009, LNCS5702, 2009.

Study of Zone-Based Feature for Online Handwritten Signature Recognition and Verification in Devanagari Script

Rajib Ghosh and Partha Pratim Roy

Abstract This paper presents one zone-based feature extraction approach for online handwritten signature recognition and verification of one of the major Indic scripts–Devanagari. To the best of our knowledge no work is available for signature recognition and verification in Indic scripts. Here, the entire online image is divided into a number of local zones. In this approach, named Zone wise Slopes of Dominant Points (ZSDP), the dominant points are detected first from each stroke and next the slope angles between consecutive dominant points are calculated and features are extracted in these local zones. Next, these features are supplied to two different classifiers; Hidden Markov Model (HMM) and Support Vector Machine (SVM) for recognition and verification of signatures. An exhaustive experiment in a large dataset is performed using this zone-based feature on original and forged signatures in Devanagari script and encouraging results are found.

Keywords Online handwriting · Signature recognition · Signature verification · Zone-wise feature · Dominant points · SVM and HMM

1 Introduction

An online signature is a method of personal authentication biometrically to execute automated banking transactions, online voting system or physical entry to protected areas. Signatures are used for identifying different persons because each signature

R. Ghosh (✉)
Department of Computer Science & Engineering, National Institute of Technology, Patna, India
e-mail: rajib.ghosh@nitp.ac.in

P.P. Roy
Department of Computer Science & Engineering, Indian Institute of Technology, Roorkee, India
e-mail: proy.fcs@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_47

possess both static as well as some dynamic features [1]. Dynamic features include elevation and pressure signals which make each person's signature unique. Even if skilled forgers are able to produce the same shape of the original signature, it is unlikely that they will also be able to produce the dynamic properties of the original one. In this paper a zone based feature [2] extraction approach has been used for recognition and verification of online handwritten Devanagari signatures. Zone based features [2] have been used and shown efficient results in online character recognition purpose.

Several studies are available [1–12] for online handwritten signature recognition and verification in non-Indic scripts, but to the best of our knowledge no work is available for signature recognition and verification in Indic scripts. In our system, we perform preprocessing such as interpolating missing points, smoothing, size normalization and resampling on each stroke of the signature. Then each online stroke information of a signature is divided into a number of local zones by dividing each stroke into a number of equal cells. Next, using the present approach, named ZSDP, dominant points are detected for each stroke and next the slope angles between consecutive dominant points are calculated separately for the portion of the stroke lying in each of the zones. These features are next fed to classifiers for recognition and verification of signatures. We have compared SVM and HMM based results in this paper.

The rest of the paper is organized as follows. Section 2 details the related works. In Sect. 3 we discuss about the data collection. Section 4 details the preprocessing techniques used and the proposed approaches of feature extraction methods. Section 5 details the experimental results. Finally, conclusion of the paper is given in Sect. 6.

2 Literature Survey

To the best of our knowledge, no study is available for online handwritten signature recognition and verification in Indic scripts. Some of the related studies available in non-Indic scripts are discussed below.

Plamondon et al. [3] reported an online handwritten signature verification scheme where signature features related to temporal and spatial aspects of the signature, are extracted. Several methods have been proposed for using local features in signature verification [4]. The most popular method uses elastic matching concept by Dynamic Warping (DW) [5, 6]. In the literature, several hundreds of parameters have been proposed for signature recognition and verification. Among these, the parameters like position, displacement, speed, acceleration [7, 8], number of pen ups and pen downs [8], pen down time ratio [7], Wavelet transform [9], Fourier transform [10] have been extensively used. Dimauro et al. [11] proposed a function-based approach where online signatures are analysed using local properties

based on time sequences. In this approach, a signature is characterized by a time function [11]. In general, better performances are obtained from function-based approaches than the parameter-based approach but time-consuming matching/comparison procedures are involved in function-based approach. However, another study [12] shows that both parametric and function-based approaches are equally effective. During matching, the authenticity of test signatures are validated by comparing the features of test signatures against the model created from the training set. The matching techniques based on Dynamic time warping (DTW), Hidden Markov Model (HMM), Support vector machine (SVM) and Neural Networks (NN) are commonly used.

3 Devanagari Script and Online Data Collection

Devanagari (or simply Nagari), the script used to write languages such as Sanskrit, Hindi, Nepali, Marathi, Konkani and many others. Generally, in Devanagari script, words or signatures are written from left to right and the concept of upper-lower case alphabet is absent in this script. Most of the words or signatures of Devanagari script have a horizontal line (*shirorekha*) at the upper part. Figure 1 shows two different online handwritten signatures in Devanagari script where *shirorekha* is drawn in the upper part of both the signatures.

Online data acquisition captures the trajectory and strokes of signatures. In online data collection, the rate of sampling of each stroke remains fixed for all signature samples. As a consequence, the number of points in the series of co-ordinates for a particular sample does not remain fixed and depends on the time taken to write the sample on the pad.

For our data collection, a total of 100 native Hindi writers belonging to different age groups contributed handwritten signature samples. Each writer was prompted to provide five genuine samples of each signature in Devanagari script. So, a total of 500 samples have been collected for each genuine signature in Devanagari script. The training and testing data for genuine signatures are in 4:1 ratio. Each writer was also prompted to provide five forged signatures of five other people. So, a total of 500 samples have been collected of forged signatures.

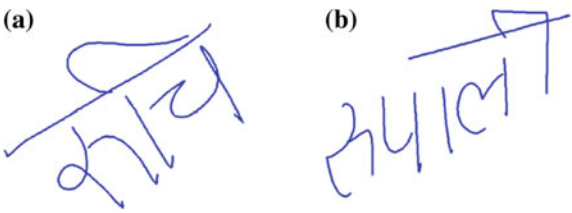


Fig. 1 Two different online handwritten signatures in Devanagari script

4 Feature Extraction

Before extracting the features from strokes, a set of preprocessing tasks is performed on the raw data collected for each signature sample. Preprocessing includes several steps like *interpolation*, *smoothing*, *resampling* and *size normalization* [13]. Figure 2 shows the images of one online handwritten signature in Devanagari script before and after smoothing. The detailed discussion about these preprocessing steps may be found in [13].

During feature extraction phase, the features that will be able to distinguish one signature from another, are extracted. The feature extractions are done on the entire signature image, irrespective of the number of strokes it contains. We discuss below the proposed zone-based feature extraction approach for signature recognition and verification.

Zone wise Slopes of Dominant Points (ZSDP): The whole signature image is divided into a number of local zones of r rows $\times$ c columns. But, instead of directly local feature extraction, we divide the portion of the strokes lying in each zone into dominant points. These dominant points are those points where the online stroke changes its slope drastically. In order to extract this feature, at first, *slopes* are calculated between two consecutive points for the portion of the trajectory lying in each local zone. The *slope* angles are quantized uniformly into 8 levels. Let, the resultant quantized slope vector for any particular zone is $Q = \{q_1, q_2, \ldots, q_n\}$. Formally, a point, p_i is said to be a dominant point if the following condition is satisfied:

$$|q_{i+1} - q_i| \,\%\, k \geq CT$$

where, CT is *Curvature Threshold* and % is *modulo* operator. Here, $k = 8$ is used for modulus operations because each element q_i of the quantized slope vector can take any value from 0, …, 7. By default, the first and last point of a stroke are considered as dominant points. Figure 3 illustrates this concept. It is noted that, when CT = 0, it contains all points as dominant points. The number of dominant points keeps decreasing with increasing CT. When CT is more than 3, very few dominant points remain.

Next, the slope angles between consecutive dominant points are calculated in each zone. The slope angles are quantized uniformly into 8 levels. If the resultant angular values of *slope* lie between 0° and 45° then the corresponding dominant point is placed in bin1, if the values lie between 46° and 90° it is placed in bin2, and

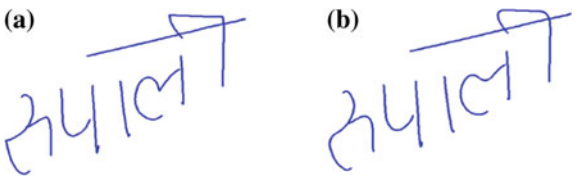


Fig. 2 Example of one online handwritten signature in devanagari script **a** After interpolation but before smoothing. **b** After smoothing the trajectories of strokes

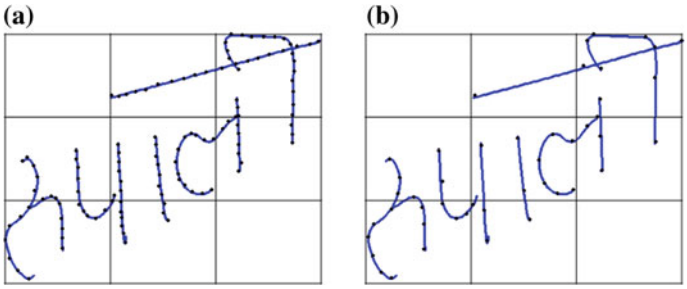


Fig. 3 Dominant points of different strokes of an online handwritten signature in devanagari script for different threshold values a) CT = 0, b) CT = 1

so on. We have tested with other bin divisions, but the one using pi/4 gives the best accuracy. The histograms of feature values are normalized and we get 8 dimensional feature vector for each zone. So, the total dimension for 9 zones is $8 \times 9 = 72$.

5 Experimental Results and Discussion

Support Vector Machine (SVM) and Hidden Markov Model (HMM) classifiers are used for our online signature recognition and verification system. Support Vector Machine (SVM) has been applied successfully for pattern recognition and regression tasks [14, 15].

We apply Hidden Markov Model (HMM) based stochastic sequential classifier for recognizing and verifying online signatures. The HMM is used because of its capability to model sequential dependencies [16]. HMM classifier has been implemented through HTK toolkit [17].

The experimental testing of the proposed approach was carried out using online handwritten Devanagari genuine and forged signatures. The feature vectors of genuine signatures are used to build the training set which is used to build a model for validating the authenticity of test signatures. Two separate testing sets are created—one for genuine signatures and another for forged signatures.

5.1 Signature Recognition Result

Results using SVM: Using the current approach, the system has been tested using different kernels of SVM and by dividing the entire signature image into different zones. Using this approach, best accuracy is obtained using the combination of 16 zone division, CT = 2 and linear kernel of SVM. The detailed result analysis, using ZSDP approach, is shown in Table 1.

Table 1 Signature recognition results using SVM with different kernels for ZSDP approach

ZSDP				
Zones	CT	RBF kernel (%)	Linear kernel (%)	Polynomial kernel (%)
9 (3 × 3)	2	87.03	92.57	88.23
9 (3 × 3)	3	84.89	89.89	85.17
9 (3 × 3)	4	81.85	86.85	82.11
16 (4 × 4)	2	**93.42**	**98.36**	**94.73**
16 (4 × 4)	3	90.23	95.23	91.47
16 (4 × 4)	4	87.76	92.76	88.20

Table 2 Signature recognition results using HMM for ZSDP approach

HMM states	Gaussian mixture	*ZSDP (%)*
3	16	84.58
4	16	70.21
3	**32**	**88.23**
4	32	74.59
3	64	88.23
4	64	74.59

Results using HMM: The testing datasets for HMM based experimentation are same as used in SVM based experimentation. Table 2 shows the recognition accuracies using *ZSDP* approach. In our experiment, we have tried different Gaussian mixtures and state number combinations. We noted that with 32 Gaussian mixtures and 3 states, HMM provided the maximum accuracies. Figure 4 shows the signature recognition results using ZSDP approach based on different top choices for both SVM and HMM.

5.2 Signature Verification Result

To validate the authenticity of each genuine signature, forged signatures are used as test samples. For signature verification, generally two different measurement techniques are employed to measure the performance of the verification system—False Acceptance Rate (FAR) and False Rejection Rate (FRR). The first one indicates the rate of accepting forgeries as genuine signatures and the second one indicates the rate of rejecting forged signatures. For a good signature verification system, the value of FAR should be very low and FRR should be high. Table 3 shows the signature verification results through FAR and FRR.

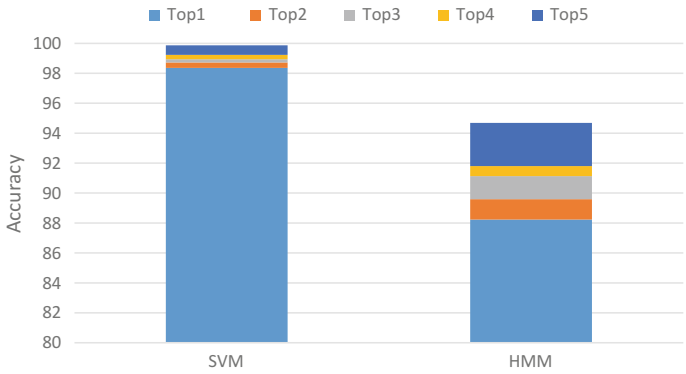


Fig. 4 Signature recognition results for ZSDP approach based on different Top choices using SVM and HMM

Table 3 Signature verification results using ZSDP approach

Classifier	FAR (%)	FRR (%)	Accuracy (%)
SVM	2.89	97.11	97.11
HMM	6.70	93.30	93.30

5.3 Comparative Analysis

Among the existing studies in the literature, to the best of our knowledge, no work exists on online handwritten signature verification system in Devanagari script. So, the present work cannot be compared with any of the existing studies.

6 Conclusion

In this paper, we have described one approach of feature extraction for online handwritten signature recognition and verification in Devanagari script. In our dataset we considered five samples each for signatures of 100 different persons in Devanagari script. The experimental evaluation of the proposed approach yields encouraging results. This work will be helpful for the research towards online recognition and verification of handwritten signatures of other Indian scripts as well as for Devanagari.

References

1. W. Nelson, E. Kishon, "Use of Dynamic Features for Signature Verification", Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1991, Charlottesville, USA, pp. 201–205.
2. R. Ghosh, P.P. Roy, "Study of two zone based features for online Bengali and Devanagari character recognition", Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, Nancy, France, pp. 401–405.
3. R. Plamondon, G. Lorette, "Automatic signature verification and writer identification- the state of the art", Pattern Recognition, 1989, Vol. 22, Issue 2, pp. 107–131.
4. M. Parezeau, R. Plamendon, "A Comparative Analysis of Regional Correlation, Dynamic Time Warping and Skeleton Matching for Signature Verification", IEEE Transaction on Pattern Recognition and Machine Intelligence, 1990, Vol. 12, Issue 7, pp. 710–717.
5. Y. Sato, K. Kogure, "Online signature verification based on shape, motion and writing", Proceedings of the 6th International Conference on Pattern Recognition, 1982, Munich, Germany, pp. 823–826.
6. P. Zhao, A. Higashi, Y. Sato, "On-line signature verification by adaptively weighted DP matching", IEICE Transaction on Information System, 1996, Vol. E79-D, Issue 5, pp. 535–541.
7. W.T. Nelson, W. Turin, T. Hastie, "Statistical Methods for On-Line Signature Verification", International Journal of Pattern Recognition and Artificial Intelligence, 1994, Vol. 8, Issue 3, pp. 749–770.
8. L.L. Lee, T. Berger, E. Aviczer, "Reliable On-Line Signature Verification Systems", IEEE Transaction on Pattern Analysis and Machine Intelligence, 1996, Vol. 18, Issue 6, pp. 643–649.
9. D. Letjman, S. Geoge, "On-Line Handwritten Signature Verification Using Wavelet and Back Propagation Neural Networks", Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR), 2001, Seattle, USA, pp. 596–598.
10. Q.Z. Wu, S.Y. Lee, I.C. Jou, "On-Line Signature Verification Based on Logarithmic Spectrum", Pattern Recognition, 1998, Vol. 31, Issue 12, pp. 1865–1871.
11. G. Dimauro, G. Impedevo, G. Pirlo, "Component Oriented Algorithms for Signature Verification", International Journal of Pattern Recognition and Artificial Intelligence, 1994, Vol. 8, Issue 3, pp. 771–794.
12. J.F. Aguilar, S. Krawczyk, J.O. Garcia, A.K. Jain, "Fusion of Local and Regional Approaches for On-Line Signature Verification", Proceedings of the International Workshop on Biometric Recognition System, 2005, Beijing, China, pp. 188–196.
13. S. Jaeger, S. Manke, J. Reichert, A. Waibel, "Online handwriting recognition: The NPen++ recognizer," International Journal on Document Analysis and Recognition, 2001, Volume 3, Issue 3, pp. 169–180.
14. C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol.2, pp. 1–43.
15. U. Pal, P. P. Roy, N. Tripathy, J. Lladós, "Multi-Oriented Bangla and Devanagari Text Recognition", Pattern Recognition, vol. 43, 2010, pp. 4124–4136.
16. P.P. Roy, P. Dey, S. Roy, U. Pal, F. Kimura, "A Novel Approach of Bangla Handwritten Text Recognition using HMM", Proceedings of the 14th International Conference on Frontiers in Handwriting Recognition, 2014, Heraklion, Greece, pp. 661–666.
17. S. Young. The HTK Book, Version 3.4. Cambridge Univ. Eng. Dept., 2006.

Leaf Identification Using Shape and Texture Features

Thallapally Pradeep Kumar, M. Veera Prasad Reddy and Prabin Kumar Bora

Abstract Identifying plant species based on a leaf image is a challenging task. This paper presents a leaf recognition system using orthogonal moments as shape descriptors and Histogram of oriented gradients (HOG) and Gabor features as texture descriptors. Th e shape descriptors captures the global shape of leaf image. The internal vein structure is captured by the texture features. The binarized leaf image is pre-processed to make it scale, rotation and translation-invariant. The Krawtchouk moments are computed from the scale and rotation normalized shape image. The HOG feature is computed on rotation normalized gray image. The combined shape and texture features are classified with a support vector machine classifier (SVM).

Keywords Plant identification · Krawtchouk moments · Histogram of oriented gradients · Gabor filter

1 Introduction

Identifying a plant is a difficult task even for experienced botanists, due to huge number of plant species existing in the world. This task is important for a large number of applications such as agriculture, botanical medicine (ayurvedic treatment), cosmetics and also for biodiversity conservation [1]. Plant identification is generally done based on the observation of the morphological characteristics of the plant such as structure of stems, roots and leaves and flowers followed by the consultation of a guide or a known database. Flowers and fruits are present only for a few weeks whereas leaves are present for several months and they also contain taxonomic identity of a plant. This is why many plant identification methods work on leaf image databases [2–4]. A leaf image can be characterized by its color, texture, and shape. Since the color of a leaf varies with the seasons and climatic conditions and also

T. Pradeep Kumar (✉) · M. Veera Prasad Reddy · P.K. Bora
Department of Electronics and Electrical Engineering,
Indian Institute of Technology Guwahati, Guwahati 781039, India
e-mail: pradeepiitg29@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_48

most plants have similar leaf color, this feature may not be useful as discriminating feature for the species recognition. Hence only shape and texture features are used as discriminating features for plant species recognition. In this paper, we use shape and texture features to identify the plant species. They are used to calculate the shape features in many pattern recognition tasks, e.g. object detection. Shape descriptors can be broadly categorized into contour-based descriptors and region based descriptors. By using the shape feature, we get the global shape of the leaf. To get the complete description about the leaf interior structure is important. The interior structure extracted using the histogram of oriented gradients (HOG) and Gabor filters. The combined shape and texture features is given to a Support Vector Machine (SVM) for classification.

Leaf identification is a major research area in the field of computer vision and pattern recognition because of its application in many areas. Many methods have been proposed in the literature for leaf identification. Yahiaoui et al. [3] used Directional Fragment Histogram (DFH) [5] as a shape descriptor for plant species identification. They applied this method on plant leaves database containing scanned images which gave an accuracy of 72.65 %. Wu et al. [6] proposed a leaf recognition using probabilistic neural network (PNN) with image and data processing techniques. They also created the Flavia database [6], which consists of 32 classes of leaf species and total of 1907 images. Kumar et al. [4] used Histograms of Curvature over Scale (HoCS) features for identifying leaves. They obtained an accuracy of 96.8 % when performed on dataset containing 184 tree species of Northeastern United States. Bhardwaj et al. [7] used moment invariant and texture as features for plant identification. This method obtained an accuracy of 91.5 % when performed on a database containing 320 leaves of 14 plant species. Recently, Tsolakidis et al. [8] used Zernike moments and histogram of oriented gradients as features for leaf image, for which they obtained an accuracy of 97.18 % on the Flavia database.

The Zernike moments, being continuous moments suffer from the discretization error and hence discrete orthogonal moments, like Krawtchouk moments (KM) [9] are proved to be better alternatives. The Gabor filters [10] are widely used as to extract the texture features because of their superior performance. This paper proposes to investigate the following:

1. To investigate the superiority of the Krawtchouk moments over Zernike moments as leaf-shape descriptors.
2. To propose Gabor features as alternative leaf texture features and study their performance with and without integration with the shape features.

The rest of the paper is organized as follows. Section 2 gives a brief introduction about the orthogonal moments, then Sect. 3 presents HOG and Gabor filters. Our proposed KMs and Gabor filter methods are presented in Sect. 4, followed by Simulation results in Sect. 5 and finally the paper is concluded in Sect. 6.

2 Orthogonal Moments

Moments are used as shape descriptors in many computer vision applications. Teague [11] introduced moments with orthogonal basis function. Orthogonal moments has minimum information redundancy, and they are generated using continuous and discrete orthogonal polynomials.

2.1 Continuous Orthogonal Moments

Continuous are generated using continuous orthogonal polynomials [12]. Zernike and Legendre moments are used widely in analyzing images.

2.2 Discrete Orthogonal Moments

Discrete orthogonal moments (DMOs) use discrete orthogonal polynomials as basis set. The discrete orthogonal polynomials used are Tchebichef polynomials, Meixner polynomials, Krawtchouk polynomials [13] and Charlier polynomials [12]. In this paper, we study mainly about Krawtchouck moments which are used in our method for leaf identification. We particularly consider the 2D Krawtchouk moments. These moments have been first used in image processing by Yap et al. [9], and Priyal and Bora [14] apply them for hand gesture recognition.

Krawtchouk Moments

The *n*th order krawtchouk polynomial is defined as

$$K_n(x;p,N) = \sum_{k=0}^{N} a_{k,n,p^{x^k}} =_2 F_1(-n,-x;N;\frac{1}{p}) \tag{1}$$

where $x, n = 1, 2, 3, \ldots., N, N > 0, p \in (0, 1)$,
n is order of krawtchouk polynomial,
N is number of sample points,
and p is parameter of binomial distribution.
${}_2F_1$ is the hypergeometric function defined as

$${}_2F_1(a,b;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{z^k}{k!} \tag{2}$$

and $(a)_k$ is the Pochhammer symbol given by

$$(a)_k = a(a+)(a+2)\ldots\ldots(a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)} \tag{3}$$

The weight is given by

$$w(x;p,N) = \binom{N}{x} p^x (1-p)^{N-x} \tag{4}$$

and satisfies the orthogonality condition

$$\sum_{x=o}^{N} w(x;p,N) K_n(x;p,N) K_m(x;p,N) = \rho_n(x;p,N)\delta_{nm} \tag{5}$$

The normalized Krawtchouk (weighted Krawtchouk) polynomials is given as

$$\bar{K}_n(x;p,N) = K_n(x;p,N)\sqrt{\frac{w_n(x;p,N)}{\rho_n(x;p,N)}} \tag{6}$$

The computation complexity can be reduced using following recurrence relation [9]

$$\begin{aligned} \rho(n-N)\bar{K}_{n+1}(x;p,N) &= A(Np - 2np + n - x) \\ &\times K_n(\bar{x;}p,N) - Bn(1-p)\bar{K}_{n-1}(x;p,N) \end{aligned} \tag{7}$$

where

$$A = \sqrt{\frac{(1-p)(n+1)}{p(N-n)}},$$

and

$$B = \sqrt{\frac{(1-p)^2(n+1)n}{p^2(N-n)(N-n+1)}}$$

With first two weighted krawtchouk polynomial given by

$$= \sqrt{w(x;p,N)},$$

$$K_1(\bar{x;}p,N) = (1 - \frac{x}{pN})\sqrt{w(x;p,N)}$$

Similarly the weight function can also be calculated recursively using function

$$w(x;p,N) = \left(\frac{N-x}{x+1}\right)\frac{p}{1-p}w(x;p,N) \tag{8}$$

with

$$w(0;p,N) = (1-p)^N = e^{Nln(1-p)}$$

The 2D KM of order $n+m$ for an image with intensity function $f(x,y)$ is defined as

$$Q_{nm} = \sum_{x=0}^{N}\sum_{y=0}^{M}\hat{K}_n(x;p1,N-1)\hat{K}_m(y;p2,M-1)f(x,y) \tag{9}$$

where $0 < p1, p2 < 1$ are constraints. They are used in the region-of-interest feature extraction of KM.

3 Texture Features

3.1 Histogram of Oriented Gradients

Consider the gray image $f(x,y)$ of size 512×512. The components $f(x,y)$ with horizontal kernel $D_X = \begin{bmatrix}-1 & 0 & 1\end{bmatrix}$ and vertical kernel $D_Y = \begin{bmatrix}-1 & 0 & 1\end{bmatrix}^T$. Thus

$$\nabla f_X(x,y) = f(x,y) * D_X \tag{10}$$

$$\nabla f_Y(x,y) = f(x,y) * D_Y \tag{11}$$

where $*$ denotes the convolution operator, $\nabla f_X(x,y)$ and $\nabla f_Y(x,y)$ are horizontal and vertical gradients.
The magnitude and orientation of the gradients are given by

$$G = \sqrt{\nabla f_X(x,y)^2 + \nabla f_Y(x,y)^2} \tag{12}$$

$$\theta = arctan\left(\frac{\nabla f_Y(x,y)}{\nabla f_X(x,y)}\right) \tag{13}$$

respectively.

The input image is divided into 128×128 pixels. Thus we get total of 16 patches of the image, of size 128×128 pixels each. Each pixel within a patch casts weighted vote for an orientation θ based on the magnitude of gradient G over 9 bins evenly

spaced over 0–180°. Thus by taking 50 % overlapping of image patches we get output vector of size $7 \times 7 \times 9 = 441$. The histogram is normalized by using the L_2 norm. i.e.

$$f = \frac{v}{\sqrt{\|v\|^2 + e}} \tag{14}$$

where v is vector and e is small constant.

3.2 Gabor Filter Bank

2-D Gabor filter is modelled according to simple cells in mamal visual cortex [10, 15, 16]. The filter is given by

$$\psi(x, y; f_0, \theta) = \frac{f_0^2}{\pi\gamma\eta} e^{-\left(\left(\frac{f_0}{\gamma}\right)^2 x'^2 + \left(\frac{f_0}{\eta}\right)^2 y'^2\right)} e^{j2\pi f_0 x'} \tag{15}$$

where

$$x' = x\cos(\theta) + y\sin(\theta)$$

$$y' = -x\sin(\theta) + y\cos(\theta)$$

f_0 is central frequency of the filter, θ is rotation angle of major axis of the ellipse, γ is sharpness along the major axis and η is sharpness along the minor axis.

4 Proposed Methods

4.1 Leaf Identification Using Krawtchouk Moments and HOG

In this section, we explain our proposed method of for leaf identification with KMs and HOG. The block diagram is shown in Fig. 1. The steps of proposed method are as follow

Pre-processing Stage

The given RGB image is converted to gray-scale image and further gray-scale image is converted to binary image.

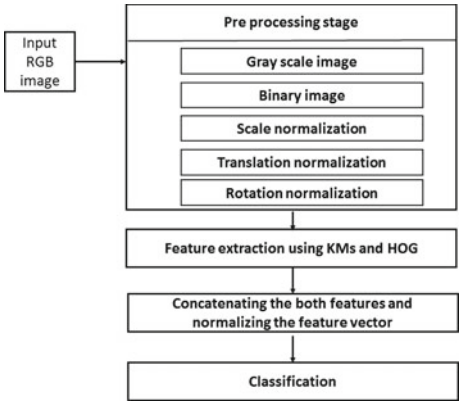


Fig. 1 Block diagram of leaf identification using KMs and HOG

(a) Original Image. (b) Gray Image. (c) Binary Image. (d) Scale Invariant Image. (e) Shift-Invariant Image. (f) Rotationa Invariant Image.

Fig. 2 Pre-processing stage

Scale Normalization

Scale invariance is achieved by enlarging or reducing each shape such that object area (i.e., zeroth order moment m_{00}) is set to a predetermined value. The scale normalized image shown in Fig. 2d.

Translation Normalization

Translation invariance is achieved by transforming the object such that centroid is moved to the origin. The translated image shown in Fig. 2e.

Rotation Normalization

The image should be rotated such that the major axis is vertical or horizontal as shown in Fig. 2f.

Feature Extraction and Classification

The KMs upto order 12 are computed on pre-processed image and HOG are computed on rotation normalized image. Both features are concatenated and normalized. The normalized feature vectors are given to the SVM for classification.

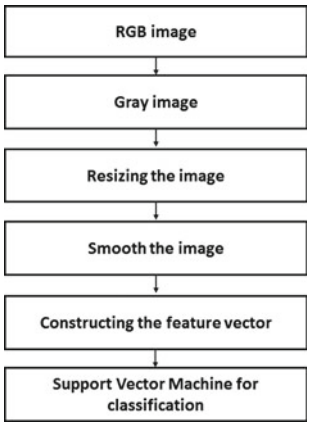


Fig. 3 Block diagram of leaf identification using Gabor filtering

4.2 Leaf Identification Using Gabor Filters

The steps of our proposed Gabor features based system is shown in Fig. 3. In this method we used 5 scales (frequencies) and 8 orientations. Thus, a total of 40 Gabor filters are used to get the texture features of leaves. The RGB image is converted to gray image and then resized to 512 × 512 pixels. Perform low pass filtering to remove dc content in image. Then, the smoothened image is convolved with Gabor filter bank. The feature vector obtained by dividing the filtered image into 64 blocks and take the local maximum from each image portion. So from each filtered image we get 64 dimensional vector. Hence from all the 40 Gabor filters we get 2560 dimensional vector.

4.3 Classification

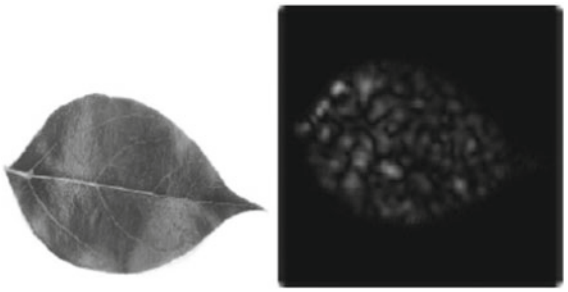
The features obtained above are given to a SVM classifier for classification. We have used the One-Against-All SVM [17, 18] technique for multiclass classification.

5 Simulation Results

The combined feature of KMs+HOG is tested on the Flavia database which consists of 32 classes and each class having 65 leaf images. We obtained an accuracy of 97.12 % for 80 % images for training and 20 % images for testing. We also implemented the method [8] ZMs+HOG features, and obtained an accuracy of 96.86 % on Flavia database (Fig. 4).

Since, the Krawtchouk moments overcome the discretization error of Zernike moments, the combined feature KMs+HOG gave an accuracy of 97.12 % compared

Fig. 4 Gabor filtering



(a) Gray image. (b) one of the filtered images.

to that of 96.86 % in the case of ZMs+HOG features. The Gabor filter bank alone has given an accuracy of 97.64 %. We have also tested the combined features of ZMs+Gabor and KMs+Gabor but accuracy did not improve.

The simulation results obtained by our proposed method are listed in Table 1. We have also compared our results with different existing methods in Table 2.

Table 1 Performance of the different features

Feature	Dataset	Training and testing (%)	Accuracy (%)
KMs+HOG	FLAVIA	80 and 20	97.12
Gabor	FLAVIA	80 and 20	97.64
Gabor+KMs	FLAVIA	80 and 20	97.64

Table 2 Comparision of diffrent algorithems with our on FLAVIA dataset

Author	Features	Classifier	Training and testing (%)	Accuracy (%)
Wu et al. [6]	Geometrical features	PNN	80 and 20	90.31
Kulkarni et al. [19]	Zernike moments, colour vein features	RBPNN	80 and 20	93.82
Tsolakidis et al. [8]	ZM+HOG	SVM	80 and 20	97.18
Proposed	KMs+HOG features	SVM	80 and 20	97.12
Proposed	Gabor features	SVM	80 and 20	97.64
Proposed	KMs+Gabor features	SVM	80 and 20	97.64

6 Conclusions

This paper investigated the leaf species classification problem. The shape features are computed using Krawtchouk moments. The texture features are computed using histogram of oriented gradients and Gabor features. The Gabor filter bank outputs give better texture description over HOG, because of multiple scales and orientations. For extracting the complete description of the leaf, we have combined shape feature with texture feature. The simulation results show that our proposed method outperform the existing methods.

References

1. Arun Priya, C. T. Balasaravanan, and Antony Selvadoss Thanamani, "An efficient leaf recognition algorithm for plant classification using support vector machine," 2012 International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), IEEE, 2012.
2. Z. Wang, Z. Chi, and D. Feng, "Shape based leaf image retrieval," Vision, Image and Signal Processing, IEE Proceedings-. Vol. 150. No. 1. IET, 2003.
3. Itheri Yahiaoui, Olfa Mzoughi, and Nozha Boujemaa, "Leaf shape descriptor for tree species identification," 2012 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2012.
4. Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida Lopez and Joao V. B. Soares, "Leafsnap: A computer vision system for automatic plant species identification," Computer Vision ECCV 2012, Springer Berlin Heidelberg, 2012, 502–516.
5. Itheri Yahiaoui, Nicolas Herv, and Nozha Boujemaa, "Shape-based image retrieval in botanical collections," Advances in Multimedia Information Processing-PCM 2006. Springer Berlin Heidelberg, 2006. 357–364.
6. Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang and Qiao-Liang Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network," 2007 IEEE International Symposium on Signal Processing and Information Technology, IEEE, 2007.
7. Anant Bhardwaj, Manpreet Kaur, and Anupam Kumar, "Recognition of plants by Leaf Image using Moment Invariant and Texture Analysis," International Journal of Innovation and Applied Studies 3.1 (2013): 237–248.
8. Dimitris G. Tsolakidis, Dimitrios I. Kosmopoulos, and George Papadourakis, "Plant Leaf Recognition Using Zernike Moments and Histogram of Oriented Gradients," Artificial Intelligence: Methods and Applications, Springer International Publishing, 2014, 406–417.
9. Pew-Thian Yap, Raveendran Paramesran, and Seng-Huat Ong, "Image analysis by Krawtchouk moments," Image Processing, IEEE Transactions on 12.11 (2003): 1367–1377.
10. Tai Sing Lee, "Image representation using 2D Gabor wavelets," Pattern Analysis and Machine Intelligence, IEEE Transactions on 18.10 (1996): 959–971.
11. Michael Reed Teague, "Image analysis via the general theory of moments," JOSA 70.8 (1980): 920–930.
12. R. Mukundan, S. H. Ong, and P. A. Lee, "Discrete vs. continuous orthogonal moments for image analysis," (2001).
13. M. Krawtchouk, "On interpolation by means of orthogonal polynomials," Memoirs Agricultural Inst. Kyiv 4 (1929): 21–28.
14. S. Padam Priyal, and Prabin Kumar Bora, "A robust static hand gesture recognition system using geometry based normalizations and Krawtchouk moments," Pattern Recognition 46.8 (2013): 2202–2219.

15. Joni-Kristian Kamarainen, Ville Kyrki, and Heikki Klviinen "Invariance properties of Gabor filter-based features-overview and applications," Image Processing, IEEE Transactions on 15.5 (2006): 1088–1099.
16. Mohammad Haghighat, Saman Zonouz, and Mohamed Abdel-Mottaleb "CloudID: Trustworthy cloud-based and cross-enterprise biometric identification," Expert Systems with Applications 42.21 (2015): 7905–7916.
17. Fereshteh Falah Chamasemani, and Yashwant Prasad Singh, "Multi-class Support Vector Machine (SVM) Classifiers-An Application in Hypothyroid Detection and Classification," 2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA). IEEE, 2011.
18. Chih-Wei Hsu, and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," Neural Networks, IEEE Transactions on 13.2 (2002): 415–425.
19. Kulkarni, A. H. Rai, H. M. Jahagirdar, and K. A. Upparamani, "A leaf recognition technique for plant classification using RBPNN and Zernike moments," International Journal of Advanced Research in Computer and Communication Engineering, 2(1), 984–988.
20. Alireza Khotanzad, and Yaw Hua Hong, "Invariant image recognition by Zernike moments," Pattern Analysis and Machine Intelligence, IEEE Transactions on 12.5 (1990): 489–497.
21. Navneet Dalal, and Bill Triggs "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

Depth Image Super-Resolution: A Review and Wavelet Perspective

Chandra Shaker Balure and M. Ramesh Kini

Abstract We propose an algorithm which utilizes the Discrete Wavelet Transform (DWT) to super-resolve the low-resolution (LR) depth image to a high-resolution (HR) depth image. Commercially available depth cameras capture depth images at a very low-resolution as compared to that of the optical cameras. Having an high-resolution depth camera is expensive because of the manufacturing cost of the depth sensor element. In many applications like robot navigation, human-machine interaction (HMI), surveillance, 3D viewing, etc. where depth images are used, the LR images from the depth cameras will restrict these applications, thus there is a need of a method to produce HR depth images from the available LR depth images. This paper addresses this issue using DWT method. This paper also contributes to the compilation of the existing methods for depth image super-resolution with their advantages and disadvantages, along with a proposed method to super-resolve depth image using DWT. Haar basis for DWT has been used as it has an intrinsic relationship with super-resolution (SR) for retaining the edges. The proposed method has been tested on Middlebury and Tsukuba dataset and compared with the conventional interpolation methods using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) performance metrics.

Keywords Discrete wavelet transform · Depth image · Interpolation · Normalization

C.S. Balure (✉) · M. Ramesh Kini
National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, India
e-mail: balure1986a@gmail.com

M. Ramesh Kini
e-mail: rameshkinim@gmail.com

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_49

1 Introduction

With the increasing need for HR images, the size of image sensors have increased. To accommodate more pixels in the same sensor area, the pixel size started shrinking. As the pixel size decreases, it may not register the image accurately due to shot noise, thus there is a limit to the pixel size, beyond which it effects the imaging. The optical cameras generally available are of the order of 8 mega-pixels (MP), but the depth cameras (Mesa Swiss Ranger, CanestaVision) which are commercially available are of lower than 0.02 mega-pixels, which leads to LR depth images. Applications like robot navigation, human-machine interfaces (HMI), video surveillance, etc., HR images are desirable for decision making. As the power consumed by the depth sensor camera increases with increase in the sensors resolution, the power consumption can become a critical requirement in mobile applications like robot navigation. Thus its convenient to capture the depth images from LR depth cameras and super-resolve remotely using hardware/software solutions.

SR is a class of techniques which enhances the image quality, and the quality of an image is determined by the resolution of an image which is defined by, how closely the lines are separated. SR is an ill-posed method. It has many solutions for a given input image. Resolution can be classified into spatial, temporal, spectral, and radiometric, but in this paper we could be discussing about the spatial resolution enhancement of depth images unless otherwise specified.

The principle behind classical SR method is to fuse multiple LR images of a same static scene viewed by a single camera which are sub-pixel apart from each other. Each LR image contributes to the final reconstruction of the HR image. The super-resolved output must have the high-frequency details (e.g. edges, textures) and should be plausible to human eye. Varied methods have been developed based on this principle to enhance the intensity image, but are limited to magnification factor not beyond $\times 4$. Similar methods have been tried to super-resolve the depth image which give better results with higher magnification factor (e.g. $\times 4$, $\times 8$, or $\times 16$).

Depth image contains the information of the distance of an object from the camera position. Depth images are used in many applications like machine vision where inspection and logistic system uses for detecting and measuring the volumes, or in robotics where depth images are used for finding obstacle, in industrial applications for assembling and monitoring, in medical field endoscopic surgery and for patient monitoring, in surveillance to determine the presence, location and number of persons. Since many applications require depth image, and due to the lack of depth camera sensors resolution (which is typically low), there is a need of a solution which produces HR image without disturbing the image integrity for better visibility. The solution could be to increase the number of pixels on the same sensor size, but it becomes the dense set of pixels which introduces shot noise during imaging, or the other solution is by keeping enough distance between the pixels and increase the sensor size to hold more number of pixels, but it becomes costlier and heavier. So playing around with the no. of pixels on a given sensor plate is not in our hands. Thus, appropriate solution would be to super-resolve the captured LR images off-

line, which maintains the commercially available pricing of the cameras and reduces the computational cost of system.

Depth image super-resolution (DISR) methods are mainly concerned about preserving the discontinuity information (e.g. edges). As depth image doesn't have texture, it can be super resolved to a higher magnification factor (beyond ×4), as opposed to that of other images like natural images, medical images, or domain specific images (e.g. face images).

The available methods are classified into two groups: DISR without intensity image, and DISR with registered intensity image as a cue (also called RGB-D methods). We have reviewed the work of RGB-D methods, which is our first contribution, and we have proposed a method for DISR using DWT transform, which is our second contribution.

Use of DWT has been widely used in super-resolving the intensity images, but DWT has not been tried for depth image super-resolution. The proposed method uses interpolation of the DWT coefficients (LL, LH, HL, and HH) in super-resolving the depth image. The proposed method is an intermediate step to recover the high-frequency information by normalization the input image to the values of LL image, and then finally combining all the sub-bands (except the interpolated LL sub-band) with the normalized input image to generate the HR depth image using inverse discrete wavelet transform (IDWT).

In Sect. 2 we will be discussing about the depth imaging techniques. Section 3 will provide the description of the available methods for depth image super-resolution. We will mainly be concentrating on DISR methods with HR color image as a cue. The proposed DWT method for DISR and its comparative results will be discussed in detail in Sect. 4, followed by the conclusions drawn in Sect. 5.

2 Depth Imaging

Depth images can be estimated/captured in various ways. The common way of estimating the depth is the stereo image set-up. This is a passive technique which is suitable for static scenes. In stereo imaging set-up there will be two optical cameras facing the same scene with their optical axes making a finite angle. An angle of $\angle 0$ means, their optical axes are parallel, and $\angle 180$ means they are looking at each other, but to have a proper depth estimation the axes should make a finite angle (between $\angle 0$ to $\angle 180$), which makes their image planes non-coplanar. Less angle means less common scene between the two views, thus less search area for matching, and similarly, more angle means there will be more common are between the views for proper matching. These images need to be brought to the same plane (called rectification) before reconstructing the 3D scene.

These cameras are placed apart by a distance called baseline. Small baseline gives less precise distance estimation, and large baseline gives more precise distance estimation but at a cost of accuracy [1]. Since the scene is viewed from two different locations, it gives the sense of the depth information of the scene. The depth map

obtained by this method will be of same size as that of the intensity images captured by the stereo camera. The review of most of the point correspondence stereo methods can be seen at [2].

The other way of finding the depth image is by using the depth camera which use time-of-flight principle to find the distance of an object from the camera position. These are the active sensors which are suitable for dynamic scenes with higher frame rates. There are several depth cameras available in market e.g. Microsoft Kinect (640×480), Swiss Ranger SR4000 (176×144), Photonic Mixer Devices (PMD), CanestaVision, which use this principle. The IR light emitting diode (LED) emits IR light and the reflected IR light from the reflective object is received by the IR camera. The system calculates the distance of the object by measuring the phase of the returned IR light. The ToF method for calculating the distance is based on Eq. (1),

$$D_{max} = \frac{1}{2} c t_0, \tag{1}$$

where D_{max} is the distance in pixels from the camera position, c is the speed of light (2.997924583×10^8 m/s), t_0 is the pulse width in ns.

3 Depth Image Super Resolution (DISR)

In this section we would see how high-resolution RGB image would help achieve the task of super-resolving the LR depth image. The task here is to super-resolve the input LR depth image (using LR depth camera) to a size equal to that of the size of high-resolution RGB image (using HR optical camera) of the same scene. The assumption is that these two input images are co-aligned with each other, such that a point in LR depth image should coincide with the same point in the HR intensity image.

3.1 DISR with RGB as a Cue (RGB-D Methods)

Markov random field (MRF) framework [3] has been used in many tasks in image processing and computer vision. To mention few, they are, image segmentation, image registration, stereo matching, and super-resolution. The MRF method has been used for range image super-resolution [4]. MRF is used for graphical modeling the problem of data integration of the LR range image with the HR camera image. The mode of probability distribution defined by the MRF provides us the HR range image. It uses conjugate gradient fast optimization algorithm to MRF inference model for finding the mode. For example, x and z are the two types variables mapped to the measurements of range and image pixels respectively. The variable y represent the reconstructed range image whose density is same as the variable z.

Variables u and w are the image gradient and range discontinuity respectively. The MRF is defined through the following potentials:
The *depth measurement potential*,

$$\Psi = \sum_{i \in L} k(y_i - z_i)^2 \tag{2}$$

where L is the set of indexes for which depth measurement is available, k is the constant weight, y is estimated range in HR grid, and z is the measured range which is on LR grid.
The *depth smoothness prior*,

$$\Phi = \sum_{i} \sum_{j \in N(i)} w_{ij}(y_i - y_j)^2 \tag{3}$$

where $N(i)$ is the neighborhood of i, w_{ij} is the weighting factor between the center pixel i and its neighbor j. The resulting MRF is defined as the conditional probability over variable y which is defined through the constraints Ψ and Φ,

$$p(y|x, z) = \frac{1}{Z} exp\left(-\frac{1}{2}(\Phi + \Psi)\right) \tag{4}$$

where Z is normalizer (partition function).

With the advent of bilateral filtering (BF) [5], many researches have tried improving it to match the real scenario for depth images (cases like depth image with heavy noise). BF is a non-linear, non-iterative, local and simple to smooth image while preserving the edges. It is bilateral because it combines the domain filtering (based on closeness) and range filtering (based on similarity). The filtering is generally the weighted average of pixels in the neighborhood. The intuition for filter weights to fall off slowly over space as we move out from the center pixel is because the image vary slowly over the space, but this is not true at the edges which results in blurring the edges. Bilateral filtering for range image averages image values with weights which decay with dissimilarity.

The filters are applied to image $f(x)$ produces output image $h(x)$. Combining domain and range filtering enforces both the geometric and photometric locality, which is described as,

$$\mathbf{h}(\mathbf{x}) = k^{-1}(\mathbf{x}) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbf{f}(\xi).c(\xi, \mathbf{x}).s(\mathbf{f}(\xi), \mathbf{f}(\mathbf{x}))d\xi \tag{5}$$

where $c(\xi, \mathbf{x})$ and $s(\mathbf{f}(\xi), \mathbf{f}(\mathbf{x}))$ are the *geometric closeness* and *photometric similarity* respectively between neighborhood center x and nearby points ξ, $k(\mathbf{x})$ is the normalizing factor. In smooth regions, this bilateral filter acts as standard domain filter, but at the boundaries the similarity function adjust according the center pixel of calculation and maintains a good filtering keeping the edge information preserved.

Bilateral filtering was further extended for image upsampling [6]. They showed that this method is useful for applications like stereo depth, image colorization, tone mapping, and graph-cut based image composition, which need a global solution to be found. Recently, *joint bilateral filters* have been introduced in which the range filter is applied to a second guidance image, $\tilde{I}$. Thus,

$$J_p = \frac{1}{k_p} \sum_{q \in \Omega} I_q . f(\| p - q \|) . g(\| \tilde{I}_p - \tilde{I}_q \|) \tag{6}$$

Here the only difference is that the range filter uses $\tilde{I}$ instead of I. Using this joint bilateral filter, a method called joint bilateral upsampling (JBU) [6] has been proposed to upsample the LR solution S, with a given HR image $\tilde{I}$. The upsampled solution $\tilde{S}$ is obtained as,

$$\tilde{S}_p = \frac{1}{k_p} \sum_{q\downarrow \in \Omega} S_{q\downarrow} . f(\| p \downarrow - q \downarrow \|) . g(\| \tilde{I}_p - \tilde{I}_q \|) \tag{7}$$

where p and q denote the integer coordinates of pixels in $\tilde{I}$, and $p \downarrow$ and $q \downarrow$ are the corresponding coordinates in LR solution S.

The direct application of JBU on depth image results into artifacts which is attributed to the erroneous assumption of color data and depth data relation (i.e. when the neighboring pixel has same color, then they also will have same depth) [7]. The use of HR color camera image for guiding the LR depth data super-resolution has a risk of copying the texture from the color image into the smooth areas in the captured depth image. Where these assumptions does not hold true, we can see two artifacts, one may arise in *edge blurring* in depth map, and the *texture copying* from color image. So, they proposed multi-lateral noise aware filter for depth upsampling (NAFDU) by extending the original bilateral upsampling approach, which takes the following form,

$$\tilde{S}_p = \frac{1}{k_p} \sum_{q\downarrow \in \Omega} I_{q\downarrow} . f(p \downarrow - q \downarrow \|) . [\alpha(\Delta_\Omega) . g(\| \tilde{I}_p - \tilde{I}_q \|) + (1 - \alpha(\Delta_\Omega) . h(\| I_{p\downarrow} - I_{q\downarrow} \|)] \tag{8}$$

where $f(\cdot)$, $g(\cdot)$ and $h(\cdot)$ are the Gaussian functions, and $\alpha(\Delta_\Omega)$ is a blending function.

Another work was by [8] to increase the range image resolution of time-of-flight (ToF) camera using multi-exposure data acquisition technique and Projection onto Convex Sets (POCS) reconstruction. Their method can be used for other image modalities which are capable of capturing range image (e.g. Light Detection and Ranging LADAR). The problem with the assumption of utilizing the HR color image to super resolve the LR range image is that it does not always hold good at situations where the shading and illumination variation can produce false depth discontinuity and it sometimes introduce the texture copy artifacts due to noise in the region. The novel idea is to utilize the *alternating integration time* (exposure) during range image acquisition for depth super-resolution. Low integration time is suitable for capturing

the near field but boosts noise in far field, while high integration time captures objects in far field but cause the depth saturation in near field. So, the idea of multi-exposure will merge useful depth information from different levels and eliminate saturation and noise. Modeling the depth map is a function of various parameters, thus image formation becomes,

$$\mathbf{D}_i = f(\alpha_i \mathbf{H}_i \mathbf{q} + \eta_i), \tag{9}$$

where $\mathbf{D}_i$ is ith low resolution depth image, and $\mathbf{q}$ is the high resolution ground truth surface, $\mathbf{H}_i$ is linear mapping (consists of motion, blurring and downsampling), α_i is time dependent exposure duration, and $f(\cdot)$ is the opto-electronic conversion function that convert the phase correlation at each spatial location to depth value.

Many a times, the estimated HR depth images does not lead back to the initial LR depth image on subsampling which were used as inputs. This leads the HR solution to be slightly away from the ground truth. Thus, [9] combines the advantages of *guided image filtering* [10] and *reconstruction constraints*. The range image F is a linear function of camera image I,

$$F_i \approx a_k^T I_i + b_k, \quad \forall i \in w_k \tag{10}$$

where I_i is a color vector, w_k is a small image window, a_k is 3×1 coefficient vector, and F_i and b_k are scalars. The linear coefficients a_k and b_k can be determined by minimizing the below cost function in the window w_k,

$$E(a_k, b_k) = \sum_{i \in w_k} \left((a_k^T I_i + b_k - P_i)^2 + \epsilon \parallel a_k \parallel_2^2 \right) \tag{11}$$

where ϵ is the regularization parameter preventing a_k to be too large. This solution to this cost function results in an initial estimation of HR range image F_c. Since this might not satisfy the reconstruction constraint exactly. So the solution is estimated by projecting f_c onto the solution space of $DHf = g_r$,

$$f^* = \underset{f}{arg\ min} (\parallel DHf - g_r \parallel_2^2 + \alpha \parallel f - f_c \parallel_2^2) \tag{12}$$

The solution of this optimization problem can be computed using steepest descent method.

A fusion of weighted median filters and bilateral filters for range image upsampling was proposed by [11] and named it as *bilateral weighted median filter*. The weighted median filter finds the value minimizing the sum of weighted absolute error of given data,

$$\underset{b}{arg\ min} \sum_{\mathbf{y} \in N(\mathbf{x})} W(\mathbf{x}, \mathbf{y}) |b - I_{\mathbf{y}}|, \tag{13}$$

where $W(\mathbf{x}, \mathbf{y})$ is weight assigned to pixel $\mathbf{y}$ inside local region centered at pixel $\mathbf{x}$. The bilateral weighted median filter corresponds to the following minimization problem,

$$\underset{b}{arg\ min} \sum_{\mathbf{y} \in N(\mathbf{x})} f_S(\mathbf{x}, \mathbf{y}) \cdot f_R(I_\mathbf{x}, I_\mathbf{y})\ |b - I_\mathbf{y}| \tag{14}$$

where $f_S(\cdot)$ and $f_R(\cdot)$ are the spatial and range filter kernel.
This method alleviates texture transfer problem, more robust to misalignment, and also removes depth bleeding artifacts.

Lu and Forsyth [12] proposed a method which fully utilizes the relationship between the RGB image segmentation boundaries and depth boundaries. Since each segment will have its depth field, which will be constructed independently using novel smoothing method. They have shown results for super-resolution from ×4 to ×100. They have even contributed a novel dataset. The best part of this work is that they could obtain the high resolution depth image output which is almost near ground truth from aggressive subsampling.

4 Proposed DWT Method for DISR

DWT has been widely used in image processing tasks. It gives an insight into the images spectral and frequency characteristics. One DWT operation decomposes the input image into four coefficients, i.e. one approximation coefficients (low-low (LL) sub-band), and three detail coefficients (low-high (LH), high-low (HL) and high-high (HH) sub-bands).

The haar wavelet basis has shown the intrinsic relationship with the super-resolution task, thus haar wavelet basis has been used throughout the paper unless otherwise stated.

The use of DWT for super-resolution have been used for super-resolving the intensity images. It proves useful for applications like satellite imaging where the cameras mounted on satellites are of low-resolution. For intensity image super-resolution, [13–15] have proposed methods for multi-frame super-resolution problem, where [14] uses Daubechies db4 filter to super-resolve to a factor of ×4, and [13] uses batch algorithm for image alignment and reconstruction using wavelet based iterative algorithm, and [15] combines Fourier and wavelet for deconvolution and de-noising for multi-frame super-resolution.

Our work is in line with [16, 17], where DWT method has been used to proposed the intermediate steps for obtaining the high-frequency information from the available LR images and its corresponding interpolated coefficients (LL, LH, HL, and HH). In available methods, the LL sub-band of the DWT transform of input LR depth image were untouched and the LR input image was directly used with the estimated intermediate high-frequency sub-bands for super-resolution. Our method is different from the existing methods for two reasons: first, we have used the DWT method for depth image super-resolution which others have not tried, and secondly, we modify

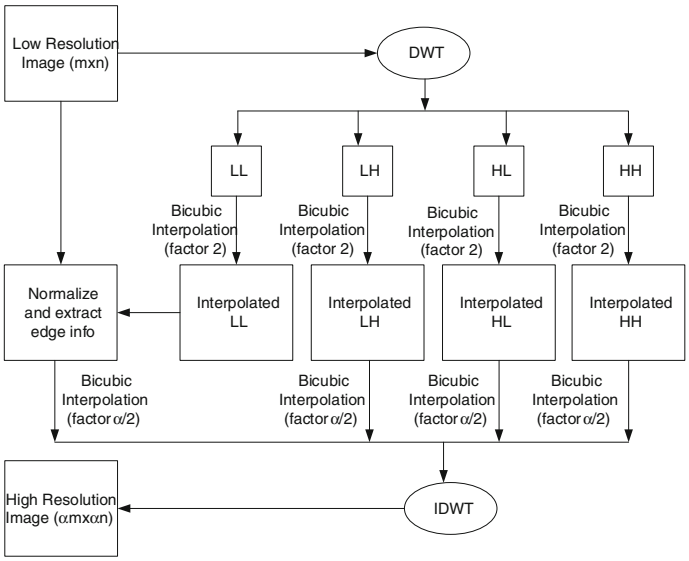


Fig. 1 Block diagram of proposed algorithm for depth image super-resolution using DWT transform

the input depth image w.r.t. the interpolated LL sub-band and then combined it with the interpolated high-frequency sub-bands (LH, HL and HH).

The proposed method uses haar wavelet type. Haar wavelet is squared-shaped rescaled sequence which together forms a wavelet family or wavelet basis. The wavelet basis haar was chosen because it has some advantageous property of analyzing sudden transition, which can be incorporated in super-resolution method for edge preserving.

The Fig. 1 shows the block diagram of the proposed method for depth image super-resolution using DWT transform. It takes the input image which is taken from the LR depth camera and then decomposes it into the approximation sub-band and detail sub-bands. All these sub-bands were interpolated by a factor of $\times 2$ using bicubic interpolation because bicubic is the sophisticated method compared to other interpolation methods. As we are trying to super-resolve the depth image by a factor of $\times 2$, so is the reason that the input image is decomposed only once (level-1 decomposition). Once the interpolation of sub-band images are done, then the original LR depth input image is used with the interpolated LL sub-band image to extract the high-frequency detail by first normalizing to values [0 1], and then using it to normalized to the values between minimum and the maximum of the interpolated LL sub-band. The output of this process is then combined with the high-frequency interpolated detail sub-band images (LH, HL, and HH) and apply IDWT to get the high-resolution image.

(a) Sawtooth (b) Cone (c) Teddy (d) Art (e) Aloe (f) Tsukuba

Fig. 2 Test images from stereo dataset of Middlebury and Tsukuba with their original resolution. **a** 380×434, **b** 375×450, Fig. 3b 375×450, Fig. 3c 370×463, **e** 370×427, **f** 480×640, (**all images are left image of the stereo pair**)

In order to perform the super resolution by a higher magnification factor, say $\times\alpha$ (>2), then the estimated high-frequency sub-bands and the normalized input image needs to be interpolated by $\times(\alpha/2)$, and then combining all of them by applying IDWT to get the resolution enhanced output image. The pseudo code of the overall proposed algorithm is shown in Algorithm 1.

Algorithm 1 depth image super resolution using DWT

1: INPUT: Read the input LR depth image I_d of size $m \times n$.
2: Apply the DWT on I_d to produce the subbands LL, LH, HL and HH of size $m/2 \times n/2$, $(I_{LL}, I_{LH}, I_{HL}, I_{HH})$.
3: Apply bilinear interpolation on all the subband image to the subbands of size $m \times n$ $(I_{iLL}, I_{iLH}, I_{iHL}, I_{iHH})$.
4: Normalize the I_d between [0 1] to produce I_{dNorm}.
5: Normalize the I_{dNorm} obtained from the step-4 between $[\min(I_{iLL}), \max(I_{iLL})]$.
6: Finally apply the IDWT on I_{dNorm} and $I_{iLH}, I_{iHL}, I_{iHH}$ to produce I_{SR}.
7: For higher magnification factor ($\times\alpha$), the inputs of IDWT need to be interpolated by a factor ($\times\alpha/2$) before using.
8: OUTPUT: I_{SR} is the super resolved image for the magnification factor $\times\alpha$ of size $\alpha m \times \alpha n$.

Existing super-resolution methods using DWT has been applied to intensity images only. We have gone one step ahead and tried with depth images. The image used for testing and comparison are taken from the stereo dataset of Middlebury [2, 18] and Tsukuba [19]. These images are shown in Fig. 2 in its original resolution.

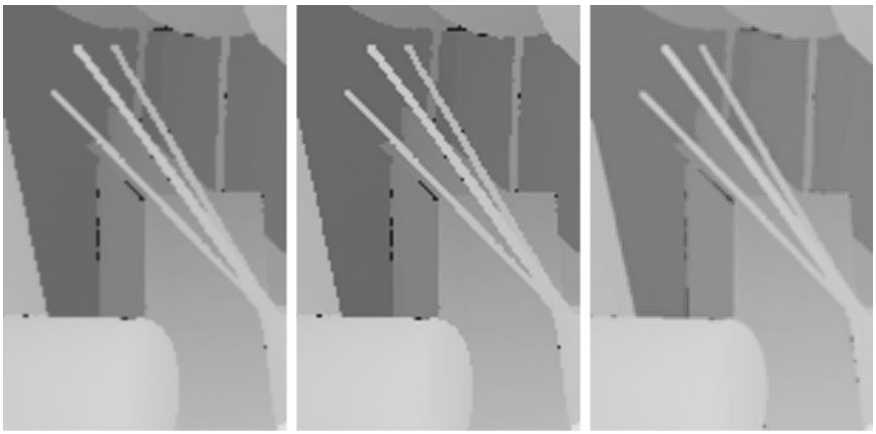
For quantitative evaluation, PSNR and SSIM performance metrics for the proposed DWT based super-resolution method for $\times 2$ magnification factor is shown in Table 1. It is seen that the results are better than the bilinear and bicubic interpolation methods. The cropped region of Teddy and Art images are shown in Figs. 3 and 4 respectively. As one can notice that the cropped region of the of the Teddy/Art in the proposed method shows less blocking artifacts compared to that of the conventional methods of interpolation.

Table 1 PSNR and SSIM performance metrics of proposed DWT method for depth image super resoution on Middlebury and Tsukuba dataset for magnification factor ×2

Images	PSNR (SSIM)		
	Bilinear	Bicubic	Proposed
Sawtooth	38.72 (0.98)	38.51 (0.98)	39.64 (0.98)
Cone	29.06 (0.94)	28.78 (0.94)	30.47 (0.94)
Teddy	28.49 (0.95)	28.22 (0.95)	29.55 (0.95)
Art	31.21 (0.96)	30.95 (0.96)	32.10 (0.98)
Aloe	31.59 (0.97)	31.29 (0.97)	32.36 (0.99)
Tsukuba	42.00 (0.99)	41.70 (0.99)	43.31 (0.99)

(a) Bilinear (b) Bicubic (c) Proposed

Fig. 3 Cropped region 150 × 150 of *Teddy* from the proposed DWT method



(a) Bilinear (b) Bicubic (c) Proposed

Fig. 4 Cropped region 300 × 200 of *Art* from the proposed DWT method

5 Discussion and Conclusion

After looking at the existing methods to find the solution of an ill-posed problem of depth image super-resolution, we realized that there is still a scope of super resolving the depth images captured from the low-resolution depth sensor cameras using the high-resolution RGB image as a cue which is available from the optical cameras of higher mega-pixels and at a low cost.

We have seen methods like joint bilateral filter which upsamples the LR depth image using the RGB image of the same scene, but the problem of texture transfer occurs in the presence of heavy noise. To overcome the problem of texture copy, a noise-aware filter NAFDU for depth upsampling was proposed which eliminates the problem of texture copy. The NLM has also been used for upsampling, but it results in the depth bleeding the fine edges. A weighted median filter and bilateral filter solves the problem of depth bleeding. With recent development of sparse methods, we have seen that the aggressive subsampled depth image can be recovered to almost near ground truth. When looking at these methods and their results, it seems like a lot of the algorithms can be combined to produce the better results.

The intermediate step of obtaining the high frequency details has been proposed which normalize the input image with respect to the LL sub-band of the DWT transform of the input LR depth image. The proposed method has been tested on widely used datasets of Middlebury and Tsukuba. The proposed method performs better than the conventional interpolation methods. The proposed method shows an improvement of 1.33 dB (on an average) in the PSNR value over the selected test images when compared with the bicubic interpolation technique for a $\times 2$ magnification factor.

References

1. Okutomi, M., Kanade, T.: A multiple-baseline stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(4), 353–363 (1993)
2. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47(1–3), 7–42 (2002)
3. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 721–741 (1984)
4. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: NIPS. vol. 5, pp. 291–298 (2005)
5. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Sixth International Conference on Computer Vision, 1998. pp. 839–846. IEEE (1998)
6. Kopf, J., Cohen, M.F., Lischinski, D., Uyttendaele, M.: Joint bilateral upsampling. In: ACM Transactions on Graphics (TOG). vol. 26, p. 96. ACM (2007)
7. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A noise-aware filter for real-time depth upsampling. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008 (2008)
8. Gevrekci, M., Pakin, K.: Depth map super resolution. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3449–3452. IEEE (2011)

9. Yang, Y., Wang, Z.: Range image super-resolution via guided image filter. In: Proceedings of the 4th International Conference on Internet Multimedia Computing and Service. pp. 200–203. ACM (2012)
10. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(6), 1397–1409 (2013)
11. Yang, Q., Ahuja, N., Yang, R., Tan, K.H., Davis, J., Culbertson, B., Apostolopoulos, J., Wang, G.: Fusion of median and bilateral filtering for range image upsampling. IEEE Transactions on Image Processing 22(12), 4841–4852 (2013)
12. Lu, J., Forsyth, D.: Sparse depth super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2245–2253 (2015)
13. Ji, H., Fermuller, C.: Robust wavelet-based super-resolution reconstruction: theory and algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 649–660 (2009)
14. Nguyen, N., Milanfar, P.: A wavelet-based interpolation-restoration method for superresolution (wavelet superresolution). Circuits, Systems and Signal Processing 19(4), 321–338 (2000)
15. Robinson, M.D., Toth, C., Lo, J.Y., Farsiu, S., et al.: Efficient fourier-wavelet super-resolution. IEEE Transactions on Image Processing 19(10), 2669–2681 (2010)
16. Demirel, H., Anbarjafari, G.: Discrete wavelet transform-based satellite image resolution enhancement. IEEE Transactions on Geoscience and Remote Sensing 49(6), 1997–2004 (2011)
17. Demirel, H., Anbarjafari, G.: Image resolution enhancement by using discrete and stationary wavelet decomposition. IEEE Transactions on Image Processing 20(5), 1458–1460 (2011)
18. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–195. IEEE (2003)
19. Peris, M., Maki, A., Martull, S., Ohkawa, Y., Fukui, K.: Towards a simulation driven stereo vision system. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1038–1042. IEEE (2012)

On-line Gesture Based User Authentication System Robust to Shoulder Surfing

Suman Bhoi, Debi Prosad Dogra and Partha Pratim Roy

Abstract People often prefer to preserve a lot of confidential information in different electronic devices such as laptops, desktops, tablets, etc. Access to these personalized devices are managed through well known and robust user authentication techniques. Therefore, designing authentication methodologies using various input modalities received much attention of the researchers of this domain. Since we access such personalized devices everywhere including crowded places such as offices, public places, meeting halls, etc., the risk of an imposter gaining one's identification information becomes highly feasible. One of the oldest but effective form of identity theft by observation is known as shoulder surfing. Patterns drawn by the authentic user on tablet surfaces or keys typed through keyboard can easily be recognized through shoulder surfing. Contact-less user interface devices such as Leap Motion controller can be used to mitigate some of the limitations of existing contact-based input methodologies. In this paper, we propose a robust user authentication technique that has been designed to counter the chances of getting one's identity stolen by shoulder surfing. Our results reveal that, the proposed methodology can be quite effective to design robust user authentication systems, especially for personalized electronic devices.

Keywords Contact-less authentication · Shoulder surfing · Personalized device authentication · Gesture recognition

S. Bhoi (✉) · D.P. Dogra
School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar, India
e-mail: sb31@iitbbs.ac.in

D.P. Dogra
e-mail: dpdogra@iitbbs.ac.in

P.P. Roy
Department of Computer Science and Engineering,
Indian Institute of Technology, Roorkee, India
e-mail: proy.fcs@iitr.ac.in

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7_50

1 Introduction

The process by which a system recognizes a user or verifies the identity of a user trying to access it, is known as authentication. Installing a robust authentication technique that prevents impersonation, is of utmost importance for any personalized system since it plays a major role to defend against unauthorized access of the system. The procedure for establishing the identity of a user can be broadly branched into three categories [1]:

1. Proof by Knowledge—A user's identity can be authenticated with the help of information which is known only to the actual user. (e.g. Password)
2. Proof by Possession—Here the authentication is done with the help of an object specific to and in possession of the real user. (e.g. Smart Card)
3. Proof by Property—The user's identity is validated by measuring certain properties (e.g. Biometrics) and comparing these against the claimed user's original properties (e.g. Fingerprint)

Majority of the research in this domain mainly focuses on the proof by knowledge domain. Here, the validation is done with the use of password, PIN or pattern based techniques. These authentication schemes are mainly victim of shoulder surfing as shown in Fig. 1. It is a form of spying to gain knowledge of one's password or identity information. Here, the forger or imposter may observe or glance at the password, PIN or pattern being entered during authentication and may use it to impersonate a valid user. Extensive research is going on in this field to aid various applications such as authentication to prevent e-financial incidents [7], etc. Most of these applications use keystroke patterns [8] or biometrics [9, 11] or password entry [10] for authentication. The visual feedback provided by above mentioned techniques make them vulnerable to theft of identity. A possible solution to this may be to exploit the fact that the field of view of the valid user will be different as compared to the impersonator while shoulder surfing. Combining minimization of visual feedback with the above possible solution is likely to create a robust system resistant to user impersonation.

This paper proposes a novel authentication technique to avoid identity theft mainly caused by shoulder surfing. Here, we have used pattern based authentication technique without visual feedback (unlike pattern based authentication used in touch enabled devices) where Leap Motion device serves as the sensor to capture input signal. The device's interface has been used to create patterns with the help of on-air gestures. Leap Motion sensor[1] is a recent release by Leap Motion Inc. It can capture real-time movement of fingers and it can track precise movement of hand and fingers in three-dimensional space. It has a tracking accuracy of 0.01 millimetre. The device is currently being used for various gesture based applications like serious gaming [13], human computer interface, augmented reality, physical rehabilitation [12], etc. It is a low-cost device that is small in size. It supports a number of frameworks and is fairly accurate. These features of the device makes it a good choice as compared

[1] https://www.leapmotion.com/.

Fig. 1 An instance portraying authentication by a legitimate user while an imposter is applying shoulder surfing

to other similar devices such as Microsoft's Kinect or Intel's RealSense. For proper tracking of hand or fingers, a user should place his/her hand in the field of view of the device. Its range is about 150° with the distance constrained to less than a meter. The device comprises of a pair of infra-red cameras and three LEDs providing a frame rate varying from 20 to 200 fps. Information regarding the position of fingers, palm, or frame time-stamp can be obtained from each frame.

We have developed a methodology to use this for authentication on personalized devices. We start with partitioning the 2D screen or display into non-overlapping rectangular blocks and map it with the 3D field of view of the device. Assuming each of these blocks represent one character or symbol of the alphabet, users are asked to draw patterns on air. However, during the process, we do not provide any visual feedback to the user. Therefore, no cursor movement is seen on the screen. Then the task of recognising these patterns can be done by classifiers such as Hidden Markov Model (HMM) [5], Support Vector Machine (SVM), Conditional Random Field (CRF), etc. Here we have used HMM as the classifier due to its ability to model sequential dependencies and its robustness to intra-user variations. We train independent HMM for each distinct pattern in the training set and then a given sequence is verified against all trained model. The model having maximum likelihood is assumed to be the best choice.

Rest of the paper is organized as follows. In Sect. 2, proposed methodology is presented. Results obtained using large set of samples collected in laboratory involving several volunteers, are presented in Sect. 3. We conclude in Sect. 4 by highlighting some of the possible future extensions of the present work.

2 Proposed Methodology of Authentication

This section describes about the signal acquisition, field of view mapping, training and testing the authentication methodology.

2.1 Device Mapping and Feature Extraction

First, we divide the whole screen or display into non-overlapping rectangular boxes and label each of those boxes. As an example, the screen can be arranged as a matrix of size 4×4 labelled "A" to "P" as depicted in Fig. 2. Using the finger and hand tracking utility of the Leap Motion device, we track movement of a user's index finger while performing the gesture during authentication. Initially, we provide a visual feedback to the user in the form of a visible cursor that helps the user to get an idea of the initial position of his/her finger with respect to the screen. Therefore, before drawing the authentication pattern, the user first executes a predefined gesture (e.g. circle gesture) that is used as a marker to start the authentication pattern and thereafter we hide the cursor. Therefore, the visual feedback is removed and the user draws the pattern through sense and anticipation. We have tested various patterns such as swipe, screen-tap, key-tap or circle to understand the best choice for the start marker. Circle gesture was found to be the most suitable and comfortable by the volunteers. Based on their feedback and the fact that the execution of the gesture should facilitate the knowledge of the finger position on screen before the cursor is hidden, circle gesture was used.

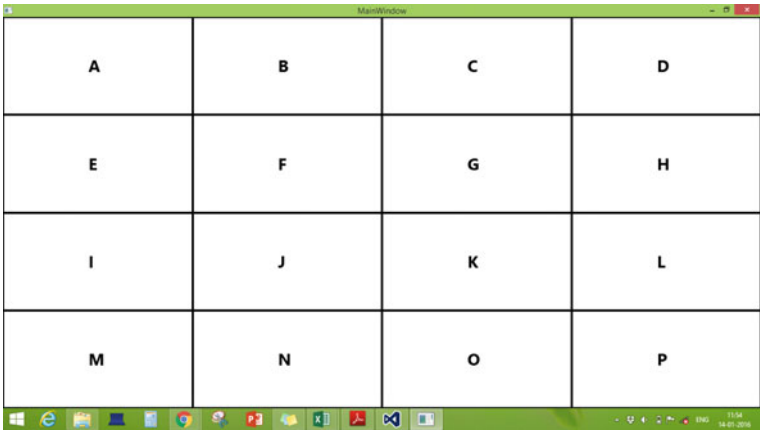


Fig. 2 Partitioning of the display into non-overlapping blocks and assignment of symbols or alphabets

Next, we present the method to map the field of view of the Leap Motion device to the display screen (e.g. computer screen). Since the display screen is rectangular in shape, therefore, instead of mapping the entire inverted pyramid interaction-space of the device (3D) to the 2D screen, we create an interaction box within the field of view to ease the movement and mapping of fingers. The height of interaction box can be set according to the user's preference of interaction height. Respective coordinate systems of display screen and Leap Motion are shown in Fig. 3. From the figure, it is evident that we need to flip the Y-axis of Leap Motion to map the coordinates correctly to the display screen. We normalize the real world position of the finger so that the coordinates lie between 0 and 1 and then translate the coordinates to the screen position as described in (1) and (2). This helps us to localize the position of the finger-tip on the display screen segment towards which the finger is pointing. We have not included Z-axis of the real-world position (with respect to the device) of the finger since we want to portray the movement on the 2-D screen.

$$X_s = (X_n)W_s \tag{1}$$

$$Y_s = (1 - Y_n)H_s \tag{2}$$

where,

X_s, Y_s represent X and Y coordinate of the finger position mapped on to the screen, respectively. X_n and Y_n represent the normalized X and Y coordinate of the finger-tip within the field of view of the device. W_s and H_s represent width and height of the screen.

Next, acquisition of authentication patterns with respect to the above mentioned mapping is described. Suppose, a user wants to draw a pattern "AEIJKL" as depicted in Fig. 4. The user needs to move his/her finger over the device's field of view to traverse the labelled boxes in the following order of sequence, A, E, I, J, K, L. To accom-

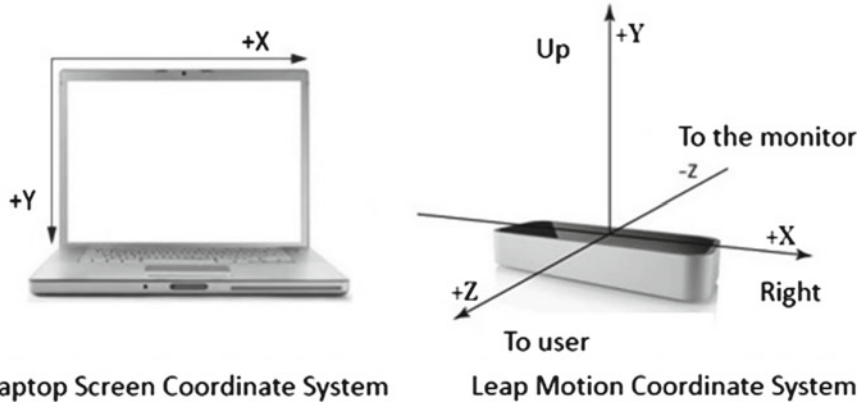


Fig. 3 Respective coordinate systems and possible mapping

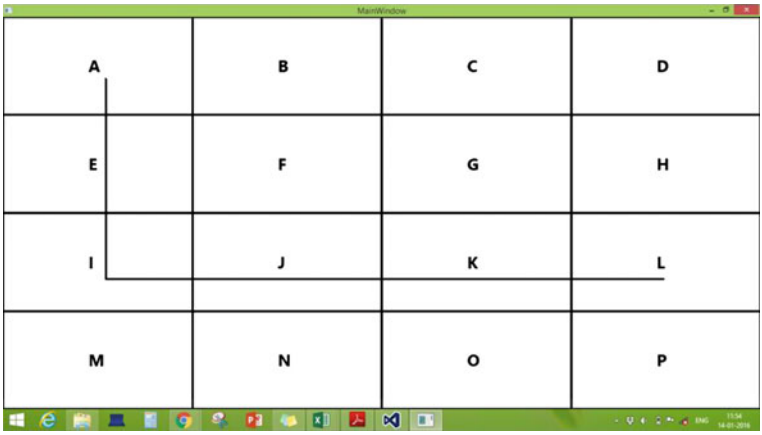


Fig. 4 A sample pattern "AEIJKL" drawn over the field of view of the device and its corresponding 2D mapping onto the screen

plish this, the user must bring his/her finger within the device's field of view and try to point box "A". After making a small circle gesture on box "A" (as described earlier), the user needs to traverse the other boxes in the above mentioned order. Although there is no visual feedback, position of the finger-tip in each frame is recorded. This information is used for generating the pattern. A pattern of such movement can be represented as follows,

$$p = (x_1, y_1), \ldots \ldots .., (x_k, y_k) \tag{3}$$

where, p represents the pattern under consideration, (x_k, y_k) represents the coordinate of the finger-tip with respect to the screen-space in the kth frame. Figure 5 depicts some of the patterns engaged in this experiment.

2.2 Training of Hidden Markov Model and Recognition

In this section, we present a methodology to implement the authentication protocol. We have applied Hidden Markov Model (HMM) based stochastic sequential classifier to train our system and classify test patterns. In our authentication scheme, users were asked to register their favourite patterns or secret sequence of symbols.

HMM is a preferred choice for such pattern classification tasks because of its ability to model sequential dependencies. An HMM can be defined by initial state probabilities π, state transition matrix $A = [a_{ij}]$, $i, j = 1, 2, \ldots, N$, where a_{ij} denotes the transition probability from state i to state j, and output probability $b_j(O)$ is modelled with discrete output probability distribution with S number of states. After several experiments, we find that, $S = 5$ provides optimum results. Vector quantization with 16 clusters has been used to discretize the input patterns or sequences. We perform

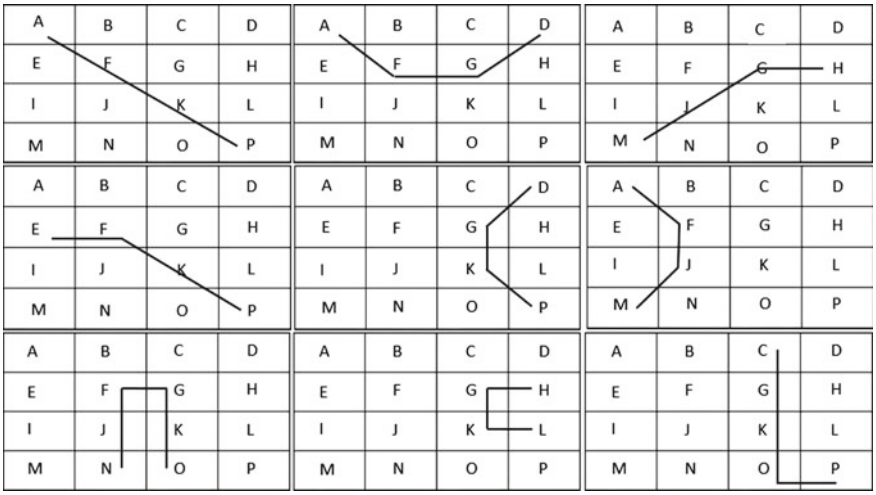


Fig. 5 Different test patterns involved in the study

the task of recognition using the Viterbi decoding algorithm [2–4]. We assume that the observation variable depends only on the present state. Therefore, a first order left to right Markov model has been presumed to be correct in the present context. The estimation of maximum likelihood parameters is carried out using Baum-Welch training algorithm. It uses EM technique for maximization of the likelihood where $\theta = (A, b_j, \pi)$ describes the Hidden Markov chain. The algorithm finds a local maximum of θ for a given set of observations (Y) as depicted in (4), where Y represents the observation sequence. More on the method can be found in Rabiner's pioneering documentation on HMM [5].

$$\theta^* = \max_{\theta} P(Y|\theta) \tag{4}$$

The parameter θ that maximizes the probability of the observation can be used to predict the state sequence for a given vector [6]. We compute the probability of observing a particular pattern ($p_j \in S$) using (5), where θ_i represents the parameters of the ith HMM that are learned through training and X denotes the hidden state sequence. Finally, given a test pattern we can classify it into one of the classes using (6) assuming there are C such distinct patterns in the dataset.

$$P(p_j, \theta_i) = \sum_{X} P(p_j|X, \theta_i)P(X, \theta_i) \tag{5}$$

$$\arg\max_{\theta_i} P(p_j, \theta_i) \;\; i = 1, 2, \ldots, 10 \tag{6}$$

Using the normalized coordinate vector representing all samples including training and testing patterns, the approach seems fairly robust to intra-user variations. In addition to that, since HMMs are scale-invariant in nature, the recognition process works fairly well regardless of the size of the coordinate vector. The procedure is summarized in Algorithm 1.

Algorithm 1 Recognition of 2D patterns using HMM

Input: $p \in P_{test}$ = Set of test sequences, P_{train} = Set of training sequences, C = Number of classes or distinct patterns, S = Number of states, O = Number of observation symbols.

Output: c_j(Class of p)where $c_j \in C$.

1: Create codebook using all training data (P_{train}).
2: **Training:** Vector quantize all training samples of a user with chosen feature set using the codebook.
3: Initialize $\pi, A, b_j \in B$, where B is the observation matrix.
4: Train and fix the model for the present pattern or a user's sequence.
5: Repeat steps 2 to 4 for all C patterns.
6: **Recognition:** Pass a test pattern (p) through all trained models and find the model with θ^*.
7: Repeat step 6 for all the test samples $p \in P_{test}$.
8: Return c_j for each test pattern.

3 Results

This section presents the results obtained during experiment involving 10 unbiased volunteers. To test the robustness of the proposed system, we have selected 10 varying authentication patterns (simple as well as complex patterns). Users were asked to mimic these patterns. Each volunteer was involved for the data acquisition phase where they were given a short demonstration to make them familiar with the Leap Motion device. A total of 1000 patterns were collected and 80 % of this data was used for training and remaining 20 % was used for testing.

A total of 10 models were created, one for each of the 10 unique patterns (essentially represent 10 distinct users). These models were trained (HMM) following the procedure described in Algorithm 1. Out of 1000 samples, 800 patterns were used for training and 200 patterns were used for testing. Confusion matrix of the classification is presented in Table 1. It is evident from the results that, accuracy is quite high for majority of these patterns except a few patterns. For example, it may be noted that, single instance of two of the patterns, namely 7 and 9, often getting confused with 9 and 6, respectively. 9 ("HGKL") is being recognized as 6 ("DGKP"). This is due to the fact that, while the user was trying to draw "HGKL" pattern, he/she might have traversed the path representing "DGKP" as depicted in Fig. 6. Therefore,

Table 1 Confusion matrix depicting accuracy of test pattern recognition (authentication)

Pattern	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	20	0	0	0	0	0	0	0	0	0
P_2	0	20	0	0	0	0	0	0	0	0
P_3	0	0	20	0	0	0	0	0	0	0
P_4	0	0	0	20	0	0	0	0	0	0
P_5	0	0	0	0	20	0	0	0	0	0
P_6	0	0	0	0	0	20	0	0	0	0
P_7	0	0	0	0	0	0	19	0	1	0
P_8	0	0	0	0	0	0	0	20	0	0
P_9	0	0	0	0	0	1	0	0	19	0
P_{10}	0	0	0	0	0	0	0	0	0	20

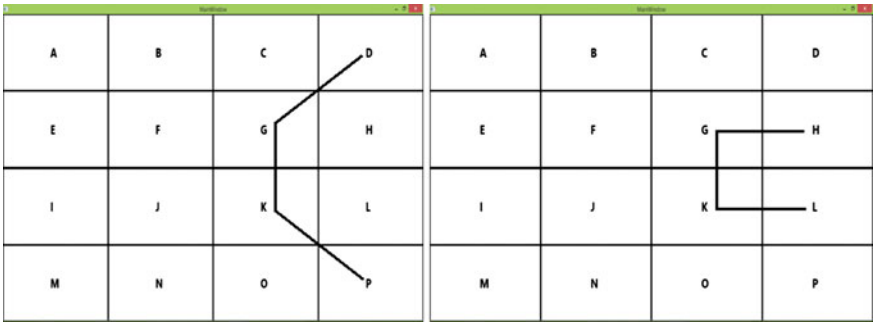


Fig. 6 Illustration of closely matching patterns "DGKP" and "HGKL"

unintentionally visiting nearby blocks during the gesture may cause failure in the logging-in procedure. However, our experiments reveal that only in two cases, we got a mismatch. Remaining cases were detected correctly with an overall accuracy of 99 %.

4 Conclusion

The paper proposes a novel technique for personalized device authentication via patterns without visual feedback. Here, we can conclude that, if the visual feedback is eliminated during authentication, the process becomes robust. However, existing touch-less or touch-based systems rely on visual feedbacks. On the contrary, the proposed Leap Motion based interface is robust against shoulder surfing attacks. This happens due to the difference in the field of views of the authentic user and the imposter.

The proposed system can be used for designing robust authentication schemes for personalized electronic devices. This will mitigate some of the limitations of existing contact-based or visual feedback based authentication mechanisms. However, the proposed system needs to be tested against real-imposter attacks and experiments need to be carried out to test its protection potential against such attacks.

References

1. Jansen, W.: Authenticating users on handheld devices. In: Proceedings of the Canadian Information Technology Security Symposium, pp. 1–12 (2003)
2. Iwai, Y., Shimizu, H., Yachida, M.: Real-time context-based gesture recognition using HMM and automaton. In: International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 127–134 (1999)
3. Rashid, O., Al-Hamadi, A., Michaelis, B.: A framework for the integration of gesture and posture recognition using HMM and SVM. In: IEEE International Conference on Intelligent Computing and Intelligent Systems, vol. 4, pp. 572–577 (2009)
4. Shrivastava, R.: A hidden Markov model based dynamic hand gesture recognition system using OpenCV. In: 3rd IEEE International Conference on Advance Computing, pp. 947–950 (2013)
5. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. In: Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286 (1989)
6. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden Markov model. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 379–385 (1992)
7. Seo, H., Kang Kim, H.: User Input Pattern-Based Authentication Method to Prevent Mobile E-Financial Incidents. In: Ninth IEEE International Symposium on Parallel and Distributed Processing with Applications Workshops (ISPAW), pp. 382–387 (2011)
8. Sheng, Y., Phoha, V. V., Rovnyak, S. M.: A parallel decision tree-based method for user authentication based on keystroke patterns. In: IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 35, no. 4, pp. 826–833 (2005)
9. Mengyu, Q., Suiyuan, Z., Sung, A. H., Qingzhong, L.: A Novel Touchscreen-Based Authentication Scheme Using Static and Dynamic Hand Biometrics. In: 39th Annual IEEE conference on Computer Software and Applications, vol. 2, pp. 494–503 (2015)
10. Syed, Z., Banerjee, S., Qi, C., Cukic, B.: Effects of User Habituation in Keystroke Dynamics on Password Security Policy. In: IEEE 13th International Symposium on High-Assurance Systems Engineering (HASE), pp. 352–359 (2011)
11. Frank, M., Biedert, R., Ma, E., Martinovic, I.; Song, D.: Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. In: IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, pp. 136–148 (2013)
12. Vamsikrishna, K., Dogra, D. P., Desarkar, M. S.: Computer Vision Assisted Palm Rehabilitation With Supervised Learning. In: IEEE Transactions on Biomedical Engineering, DOI:10.1109/TBME.2015.2480881 (2015)
13. Rahman, M., Ahmed, M., Qamar, A., Hossain, D., Basalamah, S.: Modeling therapy rehabilitation sessions using non-invasive serious games. In: Proceedings of the IEEE International Symposium on Medical Measurements and Applications, pp. 1–4 (2014)

Author Index

B. Raman et al. (eds.), *Proceedings of International Conference on Computer Vision and Image Processing*, Advances in Intelligent Systems and Computing 460,
DOI 10.1007/978-981-10-2107-7