

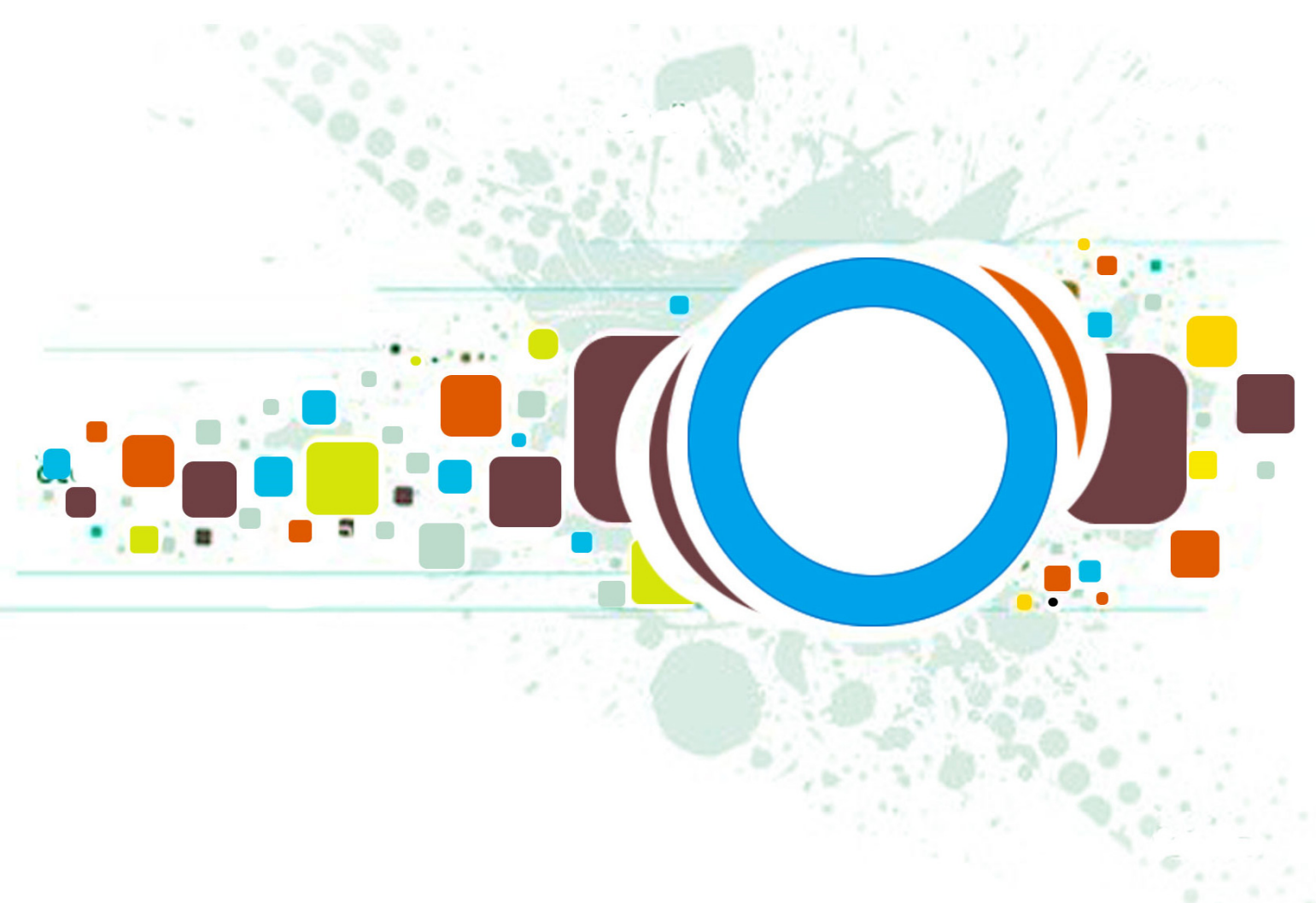
Volume 7 • Issue 4 • September 2013

Editor-in-Chief
Professor Hu, Yu-Chen

INTERNATIONAL JOURNAL OF
IMAGE PROCESSING (IJIP)

ISSN : 1985-2304

Publication Frequency: 6 Issues Per Year



CSC PUBLISHERS
<http://www.cscjournals.org>

INTERNATIONAL JOURNAL OF IMAGE PROCESSING (IJIP)

VOLUME 7, ISSUE 4, 2013

**EDITED BY
DR. NABEEL TAHIR**

ISSN (Online): 1985-2304

International Journal of Image Processing (IJIP) is published both in traditional paper form and in Internet. This journal is published at the website <http://www.cscjournals.org>, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJIP Journal is a part of CSC Publishers

Computer Science Journals

<http://www.cscjournals.org>

INTERNATIONAL JOURNAL OF IMAGE PROCESSING (IJIP)

Book: Volume 7, Issue 4, September 2013

Publishing Date: 15-09-2013

ISSN (Online): 1985-2304

This work is subjected to copyright. All rights are reserved whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication of parts thereof is permitted only under the provision of the copyright law 1965, in its current version, and permission of use must always be obtained from CSC Publishers.

IJIP Journal is a part of CSC Publishers

<http://www.cscjournals.org>

© IJIP Journal

Published in Malaysia

Typesetting: Camera-ready by author, data conversion by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2013

EDITORIAL PREFACE

The International Journal of Image Processing (IJIP) is an effective medium for interchange of high quality theoretical and applied research in the Image Processing domain from theoretical research to application development. This is the Fourth Issue of Volume Seven of IJIP. The Journal is published bi-monthly, with papers being peer reviewed to high international standards. IJIP emphasizes on efficient and effective image technologies, and provides a central for a deeper understanding in the discipline by encouraging the quantitative comparison and performance evaluation of the emerging components of image processing. IJIP comprehensively cover the system, processing and application aspects of image processing. Some of the important topics are architecture of imaging and vision systems, chemical and spectral sensitization, coding and transmission, generation and display, image processing: coding analysis and recognition, photopolymers, visual inspection etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Starting with volume 7, 2013, IJIP appears in more focused issues. Besides normal publications, IJIP intends to organize special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

IJIP gives an opportunity to scientists, researchers, engineers and vendors from different disciplines of image processing to share the ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development. This journal is helpful for the researchers and R&D engineers, scientists all those persons who are involve in image processing in any shape.

Highly professional scholars give their efforts, valuable time, expertise and motivation to IJIP as Editorial board members. All submissions are evaluated by the International Editorial Board. The International Editorial Board ensures that significant developments in image processing from around the world are reflected in the IJIP publications.

IJIP editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJIP. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJIP provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members

International Journal of Image Processing (IJIP)

EDITORIAL BOARD

EDITOR-in-CHIEF (EiC)

Professor Hu, Yu-Chen
Providence University (Taiwan)

ASSOCIATE EDITORS (AEiCs)

Professor. Khan M. Iftekharruddin
University of Memphis
United States of America

Assistant Professor M. Emre Celebi
Louisiana State University in Shreveport
United States of America

Assistant Professor Yufang Tracy Bao
Fayetteville State University
United States of America

Professor. Ryszard S. Choras
University of Technology & Life Sciences
Poland

Professor Yen-Wei Chen
Ritsumeikan University
Japan

Associate Professor Tao Gao
Tianjin University
China

Dr Choi, Hyung Il
Soongsil University
South Korea

EDITORIAL BOARD MEMBERS (EBMs)

Dr C. Saravanan
National Institute of Technology, Durgapur West Benga
India

Dr Ghassan Adnan Hamid Al-Kindi
Sohar University
Oman

Dr Cho Siu Yeung David

Nanyang Technological University
Singapore

Dr. E. Sreenivasa Reddy
Vasireddy Venkatadri Institute of Technology
India

Dr Khalid Mohamed Hosny
Zagazig University
Egypt

Dr Chin-Feng Lee
Chaoyang University of Technology
Taiwan

Professor Santhosh.P.Mathew
Mahatma Gandhi University
India

Dr Hong (Vicky) Zhao
Univ. of Alberta
Canada

Professor Yongping Zhang
Ningbo University of Technology
China

Assistant Professor Humaira Nisar
University Tunku Abdul Rahman
Malaysia

Dr M.Munir Ahamed Rabbani
Qassim University
India

Dr Yanhui Guo
University of Michigan
United States of America

Associate Professor András Hajdu
University of Debrecen
Hungary

Assistant Professor Ahmed Ayoub
Shaqra University
Egypt

Dr Irwan Prasetya Gunawan
Bakrie University
Indonesia

Assistant Professor Concetto Spampinato
University of Catania
Italy

Associate Professor João M.F. Rodrigues

University of the Algarve
Portugal

Dr Anthony Amankwah

University of Witswatersrand
South Africa

Dr Chuan Qin

University of Shanghai for Science and Technology
China

Associate Professor Vania Vieira Estrela

Fluminense Federal University (Universidade Federal Fluminense-UFF)
Brazil

Dr Zayde Alcicek

firat university
Turkey

Dr Irwan Prasetya Gunawan

Bakrie University
Indonesia

TABLE OF CONTENTS

Volume 7, Issue 4, September 2013

Pages

- 314 - 329 Preferred Skin Color Enhancement of Digital Photographic Images
Huanzhao Zeng, Ronnier Luo
- 330 - 338 Fast Motion Estimation for Quad-Tree Based Video Coder Using Normalized Cross-Correlation Measure
Eskinder Anteneh Ayele, Ravindra Eknath Chaudhari, S. B. Dhok
- 339 - 352 Unsupervised Categorization of Objects into Artificial and Natural Superordinate Classes Using Features from Low-Level Vision
Zahra Sadeghi, Majid Nili Ahmadabadi, Babak Nadjar Araabi
- 353 - 357 Cut-Out Animation Using Magnet Motion
Srinivas Anumasa, Avinash Singh, Rishi Yadav
- 358 - 371 Faster Training Algorithms in Neural Network Based Approach For Handwritten Text Recognition
Haradhan Chel, Aurpan Majumder, Debashis Nandi
- 372 - 384 Performance Analysis of Daubechies Wavelet and Differential Pulse Code Modulation Based Multiple Neural Networks Approach for Accurate Compression of Images
Siripurapu Sridhar, P.Rajesh Kumar, K.V.Ramanaiah
- 385 - 394 Skin Color Detection Using Region-Based Approach
Rudra PK Poudel, Jian J Zhang, David Liu, Hammadi Nait-Chairf

- 395 - 401 Recognition of Offline Handwritten Hindi Text Using SVM
Naresh Kumar Garg, Dr. Lakhwinder Kaur, Dr. Manish Jindal
- 402 - 417 Elliptic Fourier Descriptors in the Study of Cyclone Cloud Intensity Patterns
Ishita Dutta, S. Banerjee
- 418 - 429 3D Position Tracking System for Flexible Cystoscopy
Munehiro Nakamura, Yusuke Kajiwara, Tatsuhito Hasegawa, Haruhiko Kimura

Preferred Skin Color Enhancement of Digital Photographic Images

Huanzhao Zeng

*Color Imaging Scientist, Qualcomm Research Center
Qualcomm Inc.
San Diego, CA 92121, USA*

hzeng@qti.qualcomm.com

Ronnier Luo

*Professor, Colour, Imaging and Design Research Centre
University of Leeds
Leeds LS2 9JT, UK*

M.R.Luo@leeds.ac.uk

Abstract

Reproducing skin colors pleasingly is essential for photographic color reproduction. Moving skin colors toward their preferred skin color center improves the skin color preference. Two main factors to successfully enhance skin colors are: a method to detect skin colors effectively and a method to morph skin colors toward a preferred skin color region properly. This paper starts with introducing a method to enhance skin colors using a static skin color detection model. It significantly improves the color preference for skin colors that are not far off from regular skin tones. To enhance a greater range of skin tones effectively, another method that automatically adapts the skin color detection model to the skin tone of each individual image is proposed. It not only enhances skin colors effectively, but also adjusts the overall image colors to produce more accurate white balance on the image.

Keywords: Skin Color Enhancement, Image Enhancement, Skin Tone, Preferred Color, Memory Color, Skin Color Model.

1. INTRODUCTION

Preferred color rendering is essential for enhancing the perceived image quality of photographic images. Previous research efforts for preferred color reproduction may be traced back to more than half a century ago. It has been reported that people prefer to see an image in which the color appearance of a familiar object agrees with its memory color rather than with the actual colorimetric content of the original scene, and certain memory colors such as skin, grass, and sky are preferred to be produced with slightly different hues and with greater purity [1]-[3]. Since the color sensations evoked by a reproduction are compared with a mental recollection of the color sensations previously experienced when looking at objects similar to the ones being appraised, observers are able to rate the quality of an image without the original object presented [4]-[6].

Reproducing skin tones pleasingly is critical in preferred color reproduction. Moving skin colors toward their preferred skin color center improves the color preference. Braun [7] proposed a method for preferred skin color enhancement by squeezing the hue angles of skin colors toward a preferred point over a limited chroma range while keeping chroma unchanged. The overall color preference is improved even if skin colors of non-skin objects are modified. Because the chroma is not adjusted, it does not enhance skin tones that are within the preferred hue range but are too pale or too chromatic. Kim et al. [8] applied adaptive affine transform in Yu'v' to enhance skin color reproduction of video streams. Skin colors are defined within an ellipse, and preferred skin colors are set in a smaller ellipse within the skin color ellipse. An input skin color is converted into the preferred skin color by a linear transformation. Post-processing is required to fix contouring in the skin color boundary. Park et al. [9] developed a method to optimize memory colors in YCbCr

color space for the color reproduction on digital TV. The skin color distribution is modeled with a bivariate Gaussian probability density function. The Mahalanobis distance is used to determine the skin boundary. A smaller region within the center of the skin color ellipse is determined as the preferred skin color region. Skin colors outside the small central color region are moved toward this color region. Because the preferred skin color center is usually different from the skin color distribution center, the skin color enhancement is not optimal. Xu and Pan [10] and Pan and Daly [11] applied a sigma filter to decompose an image into a primary image and a residue image. The primary image generated from a sigma filter contains limited details but maintains sharp edges, and the residue image contains details/noises/artifacts but relatively few sharp edges. To avoid amplifying noise and artifacts, the residue image that contains noise and artifacts are not modified. Skin colors in the primary image are adjusted and mapped from sRGB to an LCD display color space.

In all of these approaches, skin color models to detect skin pixels or face pixels are generated from statistical analysis of a larger number of images, and skin detection models are not adapted to different images in which skin tones may be shifted far away from the statistical skin center. If the skin color distribution of an image is far off from the general statistical skin color distribution (e.g. skin colors are very pale or over-saturated, or skin colors are shifted to very pinkish or very yellowish direction), skin colors may not be detected or probabilities of skin colors may be very low. Thus skin colors may not be adjusted or the adjustment may be insufficient. Increasing the skin color range of a skin color model enables the model to capture more skin tones, but the false detection rate is increased as well. Face feature detections may be applied to resolve the problem [12].

The skin color modeling and skin color preference were comprehensively studied by the authors for preferred skin color reproduction [13]. A skin color model that is trained with a large number of images is applied to detect skin colors. The detected skin colors are morphed toward a preferred skin color center. Although the method improves the skin color preference, it is not effective in enhancing low quality images in which skin tones are very different from normal skin tones. To enhance skin colors more effectively, an image-dependent skin color model is developed to detect skin colors. The method results in more effective and more accurate skin color detection and leads to more effective skin color enhancement.

The rest of the paper is organized as below: the skin color modeling is briefly introduced in the following section; preferred skin color regions for skin color enhancement is briefly described in Section 3; a skin color enhancement method based on a static skin color model is presented in Section 4; skin color enhancement using image-dependent skin color modeling is presented in Section 5; a new skin color morphing method is presented in Section 6; Section 7 briefly describes experimental result; and the last section is the conclusion.

2. SKIN COLOR MODELING

The cluster of skin colors may be approximated with an elliptical shape [14]-[15]. An elliptical model is applied for baseline skin color detection. Let X_1, \dots, X_n be the distinctive colors of a training data set and $f(X_i)$ ($i=1, \dots, n$) be the occurrence counts of a color, X_i . An elliptical boundary model is defined as

$$\Phi(X) = [X - \Psi]^T \Lambda^{-1} [X - \Psi] \quad (1)$$

where Ψ is the elliptical center, and Λ is the covariance matrix.

Given a threshold ρ and an input color X of a pixel, X is classified as a skin color if $\Phi(X) < \rho$ and as a non-skin color otherwise. The threshold ρ trades off correct detections by false detections. As ρ increases, the correct detection rate increases, but the false detection rate increases as well. $\Phi(X) = \rho$ defines an elliptical boundary between skin and non-skin colors. The elliptical center is given by ψ and the principal axes are determined by Λ .

A 3-D modeling of skin colors using an ellipsoid function can be written as:

$$\begin{aligned} \Phi(x, y, z) = & \\ & u_0(x - x_0)^2 + u_1(x - x_0)(y - y_0) + u_2(y - y_0)^2 + \\ & u_3(x - x_0)(z - z_0) + u_4(y - y_0)(z - z_0) + u_5(z - z_0)^2 \end{aligned} \quad (2)$$

A 2-D modeling of skin colors in chrominance space (x, y) ignoring the lightness axis is expressed with an ellipse function:

$$\Phi(x, y) = u_0(x - x_0)^2 + u_1(x - x_0)(y - y_0) + u_2(y - y_0)^2 \quad (3)$$

To improve the accuracy of 2-D modeling, the ellipse model may be trained on a serial of lightness buckets to formulate a lightness-dependent ellipse model.

3. PREFERRED SKIN COLOR REGION

It is generally agreed that preferred skin colors are different from actual skin colors and preferred skin colors are within a smaller region of skin colors. While some reports conclude that preferred skin colors are not significantly different among observers with different culture backgrounds, others indicate that they are statistically different for preferred color reproduction [16]-[21]. In order to have a better understanding of skin color preference of digital photographic images, skin color preferences of different ethnic skin tones judged by mixed ethnic groups and by each of unique culture backgrounds were studied by the authors [22]-[23]. The psychophysical experimental results of the skin tone preference by ethnicity (the preferred African skin tone judged by Africans, the preferred Caucasian skin tone judged by Caucasians, and the preferred Oriental skin tone judged by Orientals) in CIELAB a^* - b^* coordinates with the D50 adapted white are shown in Figure 1. The orange, black, and green ellipses are the preferred skin color regions of Caucasian, African, and Oriental culture backgrounds, respectively. A large dot at each ellipse center is the preferred skin color center of the corresponding ethnical group. The blue ellipse is the overall preferred skin color region. The preferred hue angle in CIELAB adapted to the D50 white point is about 49° in all three groups.

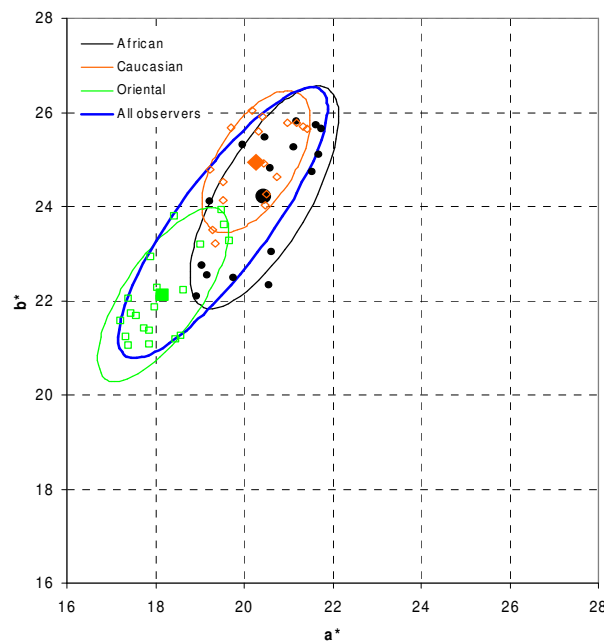


FIGURE 1: Preferred skin colors of African, Caucasian, and Oriental judged by African, Caucasian, and Oriental observers, respectively.

Statistical analysis of skin color preference among African, Caucasian and Oriental culture backgrounds reveals that all three preferred skin color centers are significantly different from each other in 5% significance level; Orientals prefer slightly less chromatic skin colors than Africans and Caucasians; the inter-observer variation of the skin color preference of Africans is larger than that of Caucasians and that of Orientals; Caucasians may prefer slightly more yellowish skin tones than Africans. In cross-culture preference, Orientals prefer slight less chromatic skin colors than Caucasians and Africans, and Africans prefer more chromatic Caucasian and Oriental skin colors than Caucasians and Orientals. The result of preferred skin colors are to be used for skin color enhancement presented in following sections.

4. PREFERRED SKIN COLOR ENHANCEMENT ALGORITHM

The ellipsoid skin color model is chosen for implementing skin color enhancement for the tradeoff of its efficiency in computation and its accuracy in skin color detection [15]. Equation (2) is used to calculate Mahalanobis distance of a point (x, y, z) to the ellipsoid center (x0, y0, z0). $\Phi(x, y, z) < \rho$ defines the region of skin tones.

Figure 2 is a sketch diagram for skin color adjustment (drawn in 2-D space for simplification). The large ellipse represents the skin color region, A is the center of the region (the statistical skin color center), and B is the preferred skin color center (PSCC). Skin colors (in the large ellipse) are morphed toward the preferred skin color region (in the smaller ellipse) for skin color enhancement.

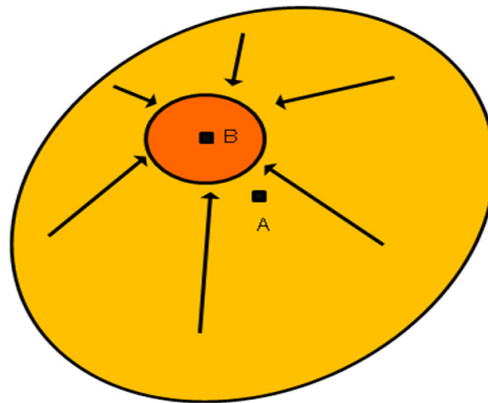


FIGURE 2: A Sketch Diagram for Skin Color Enhancement.

Figure 3 is a flowchart for skin colors adjustment. P is an input color and P' is its corresponding output color. The source color of each pixel is converted to CIELAB (or another luminance-chrominance color space, such as YCbCr). A skin color detection model is applied to compute Mahalanobis distance, $\Phi(L, a, b)$. If $\Phi > \rho$, the color is not a skin color and no color adjustment is performed for the pixel. Otherwise, the color is a skin color, and a weight, w, is computed to adjust $L^*a^*b^*$.

At the ellipsoid center, $\Phi(L, a, b) = 0$ and w is maximized; on the ellipsoid boundary, $\Phi(L, a, b) = \rho$ and $w = 0$. A weight for color adjustment may be calculated by

$$w = 1 - \Phi(L, a, b) / \rho. \tag{4}$$

Other linear or nonlinear formulae may be applied to calculate the weight. A basic concept is that the smaller Φ , the smaller w. Based on the desired strength for color enhancement, w may be further adjusted, i.e.,

$$w = w_0(1 - \Phi(L, a, b) / \rho), \tag{5}$$

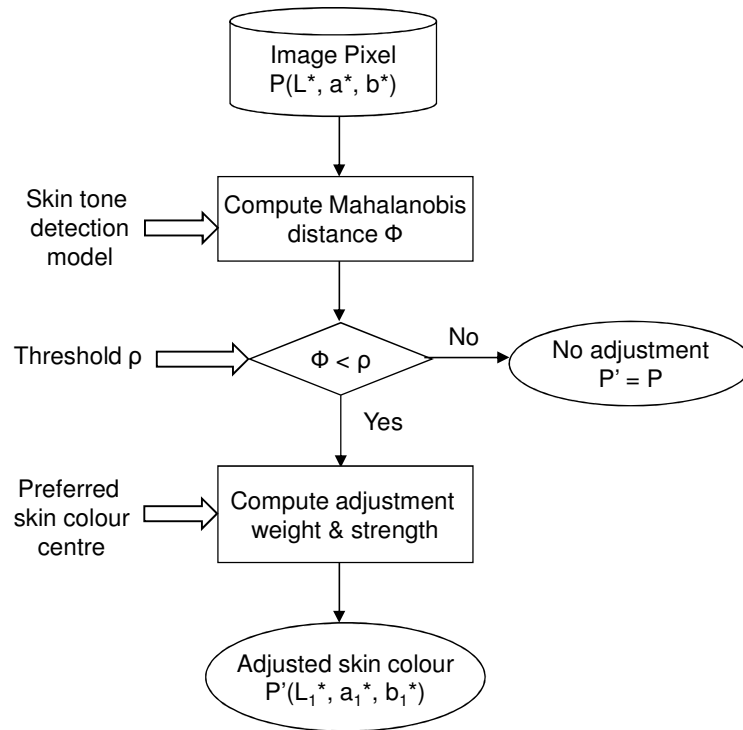


FIGURE 3: A Flow Chart for Skin Color Adjustment.

where w_0 is a strength factor for color adjustment. Without adjusting L^* , a^* , and b^* are adjusted by equations:

$$\begin{aligned} a_{new} &= a + w \cdot (a_{center} - a) \\ b_{new} &= b + w \cdot (b_{center} - b) \end{aligned} \quad (6)$$

where (a_{center}, b_{center}) is PSCC; (a, b) is a^*b^* of an original skin color, and (a_{new}, b_{new}) is a^*b^* of the corresponding enhanced skin color. The PSCC is an ellipse center shown in Figure 1. For preferred color reproduction of mixed culture backgrounds, a PSCC corresponding to mixed culture backgrounds is chosen.

The skin color enhancement presented by Park et al. [8] takes the statistical skin color center (the center of the skin color model) as the target color center for skin color adjustment. Because the objective is to morph skin colors toward a PSCC, which is different from the skin color center of the skin color model, (a_{center}, b_{center}) in Eq. (6) should be a PSCC.

Lightness dependency and highlight contrast are optimized with following implementations.

4.1. Lightness-Dependent Skin Color Enhancement

PSCCs are slightly different among different lightness. To enable different PSCCs for different lightness, three PSCCs (light, medium, and dark PSCCs) are determined. A PSCC at a given lightness is interpolated from these three PSCCs. In Figure 4, the large ellipsoid illustrates the skin color boundary; each of the three PSCCs is at the center of a preferred skin color region on a constant-lightness ellipse plane; and the red dash curve represents PSCCs at different lightness. A PSCC, (a_{center}, b_{center}) , becomes a function of lightness. In each lightness level, skin colors are morphed toward its PSCC.

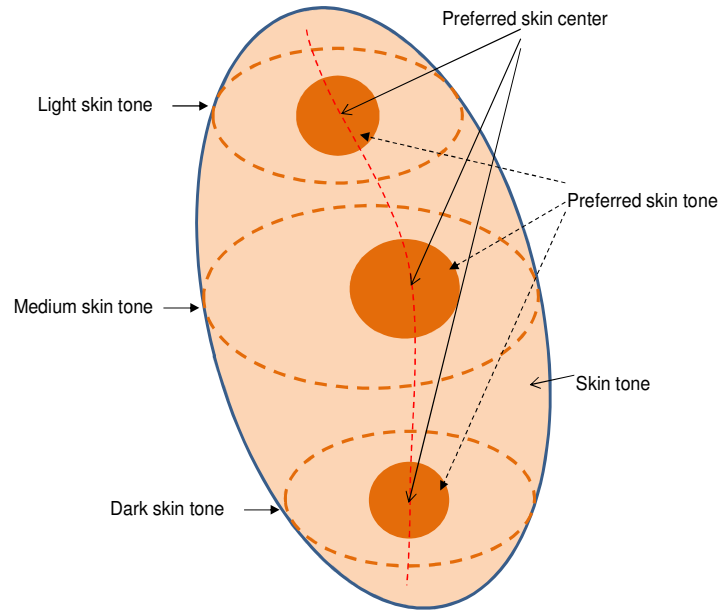


FIGURE 4: A Sketch Diagram for Skin Color Adjustment Using Three PSCCs.

4.2. Highlight Skin Color Enhancement

The perceptual contrast in the highlight skin color region may decrease after the skin color adjustment. This is due to the fact that low chroma skin colors in the highlight region may be moved toward more chromatic colors. If the adjusted colors are out of the device gamut or the encoding gamut, they are clipped to the gamut boundary. These effects result in a slight reduction in visual contrast. One approach to fix the problem is to reduce the amount of adjustment for highlight colors. The adjustment strength, w , in Eq. (6) is modulated with a factor, w_L , for lightness-dependent adjustment as shown in Eq. (7).

$$\begin{aligned} a_{new} &= a + w \cdot w_L \cdot (a_{center} - a) \\ b_{new} &= b + w \cdot w_L \cdot (b_{center} - b) \end{aligned} \quad (7)$$

w_L is 1 when L^* is smaller than or equal to a threshold (it is set to 65 in our experiment). It gradually decreases as L^* increases, and it becomes 0 at $L^*=100$.

Another approach to fix the highlight problem is to preserve chroma and to only adjust hue of highlight skin colors. After applying Eq. (7) to calculate an adjusted skin color, its hue angle and chroma (h_1 and C_1) are computed. Chroma of the original color, C_0 , is computed as well. In order to have a smooth transition from medium to highlight skin tones, chroma adjustment is slowly adapted from mid-tone to highlight by Eq. (8).

$$C = C_0 + w_L \cdot (C_1 - C_0) \quad (8)$$

Since the aim is to adjust the original hue toward a preferred hue and not to change chroma for highlight skin colors, chroma, C , and hue angle, h_1 , are used to compute a final skin color (a_{new} , b_{new}).

5. APPLYING FACE DETECTION INFORMATION FOR PREFERRED SKIN ENHANCEMENT

There are a few limitations in the skin color enhancement algorithm described in the earlier section. Because a static elliptical skin color model is the statistical modeling of a large set of images, it is suitable for detecting skin colors that are not distributed too far away from a

statistically distributed skin color region. If skin color distribution of an image is too much biased toward a certain direction, skin colors may be out of the edge of the skin color boundary or the probability of the skin color may be very low. Thus skin colors will not be adjusted or will be adjusted insufficiently.

Nachlieli et al [14] proposed face-dependent skin color enhancement method to resolve the problem. Since faces that are not detected are not enhanced, the method is unreliable for enhancing images that contain multiple faces. Improper face detection (e.g. faces partially detected) may result in color artifacts after color enhancement. A new image-dependent skin color enhancement method proposed below is to resolve the problem and the problem using static skin color models. Face detection is applied to assist analyzing skin tones in the image and to adapt the skin color model to each scene. A scene-dependent (aka image-dependent) skin color model is then applied to detect skin colors and to compute skin color weighting factors for skin color adjustment.

First, a face detection method is applied to detect faces in an image [24]-[26]. Because of potential false face detections, a static skin color model is applied to verify each detected face. The order of these two operations may be exchanged or merged into a single step, depending on how the face detection algorithm is implemented. The skin color detection model at this step is relatively high tolerant for false skin color detection rates. Since an accurate skin detection model is not required in this step, the ellipse skin color model, which is more efficient in computation but less accurate than the ellipsoid model and the lightness-dependent ellipse model, is applied [29]. The threshold, ρ , can be larger than that used in a static skin color model. A subsequent step using information of face boxes is applied to modify the skin color model for accurate skin color detection on the image. The procedure for skin color enhancement is illustrated in Figure 5 and described below.

Step 0: Initialize a skin color detection model. In this study, the ellipse model in CIELAB a^*b^* space (adapted to D50) is selected for skin color modeling.

Step 1: Apply a face detection method to detect faces.

Step 2: Remove false detected face boxes. The skin color model formulated in Step 0 is applied to classify each pixel in a face box as a skin pixel or a non-skin pixel. If the ratio of skin pixels over all pixels within a face box is lower than a pre-determined threshold, this face is classified as a false detected face and is removed from the face list. A mean skin color is then computed from detected skin colors in each remaining face box. The histogram of skin colors in each face box is generated, and skin colors in the top 25% occurrence rates are applied to compute mean skin color, a^*b^* . If the color difference between the mean skin color and the statistical skin color center is larger than a tolerance threshold, this face is considered a false detected face and is removed from the face list.

Step 3: Compute a global mean skin color from the remaining face boxes. A global mean skin color is averaged from the mean skin colors of all remaining face boxes weighted with their skin pixel counts. The mean a^*b^* are compared with a preferred skin color. If the difference is less than a pre-determined threshold, the overall skin tone of the image is considered to be within the preferred color region, therefore no skin color adjustment is performed for the image and the skin color adjustment ends here. Otherwise, skin color adjustment on the whole image is performed by remaining steps below.

Step 4: Construct an image-dependent skin color ellipse model for skin color adjustment. It starts from the static skin color ellipse in Step 0. The center of the skin color ellipse is then replaced with the mean skin color of the image computed in Step 3. Because the skin color ellipse is to represent the skin color distribution of this image instead of the skin color distribution of a training image set, the size of the ellipse can be reduced. This skin color ellipse is applied to detect skin colors of the whole image.

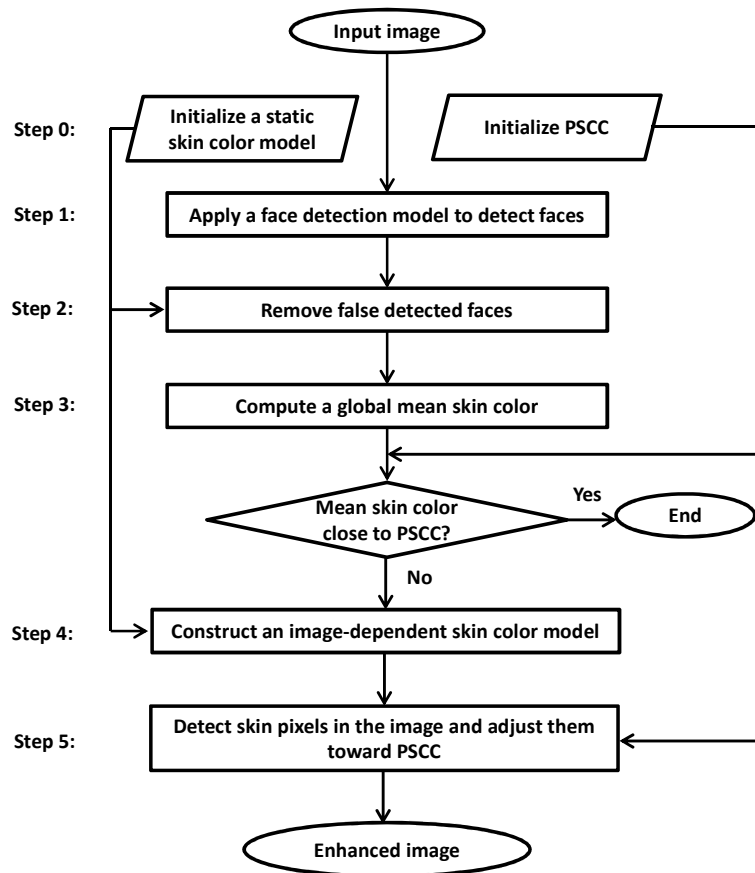


FIGURE 5: A Block Diagram for Skin Color Enhancement.

Step 5: Apply the image-dependent skin color model to adjust skin colors of the image. Instead of limiting to process skin colors within face boxes, all skin colors (skin color of any objects) in the whole image are adjusted. If the Mahalanobis distance of a pixel is smaller than a skin color threshold, the pixel is a skin color, and the Mahalanobis distance is applied to calculate a weight, w . Eq. (6) is applied to compute a preferred skin color, (a_{new}, b_{new}) , corresponding to an original skin color, (a, b) .

Large color adjustment on skin colors may have an effect that skin colors are not harmonized with the overall color balance of the image. If the mean skin color of an image is very different from the target PSCC, the adjustment is reduced. To further optimize the preferred skin color reproduction, the preferred skin color center, (a_{center}, b_{center}) , may be optimized as a function of lightness using Eq. (7).

Figure 6 shows a few test images and the enhancement result. All square boxes (black and yellow boxes) are the faces detected at Step 1. Yellow boxes are the faces removed in Step 2. Although an eye and an ear in the upper-right image are detected as faces, a mean skin color computed from these two boxes still represents the face tone of the image. Therefore, the algorithm still works well in adjusting skin colors. This demonstrates that the image-dependent skin color enhancement is more robust than face-dependent skin color enhancement. Since the skin color enhancement adjusts pixels that are skin colors regardless of their object types, it modifies the overall color balance, i.e., it adjusts the white balance of the image. Therefore it produces more accurate color balance and fixes inaccurate illuminant detection on the image. This can be seen in the bottom-left image where the wall in the back is adjusted as well.

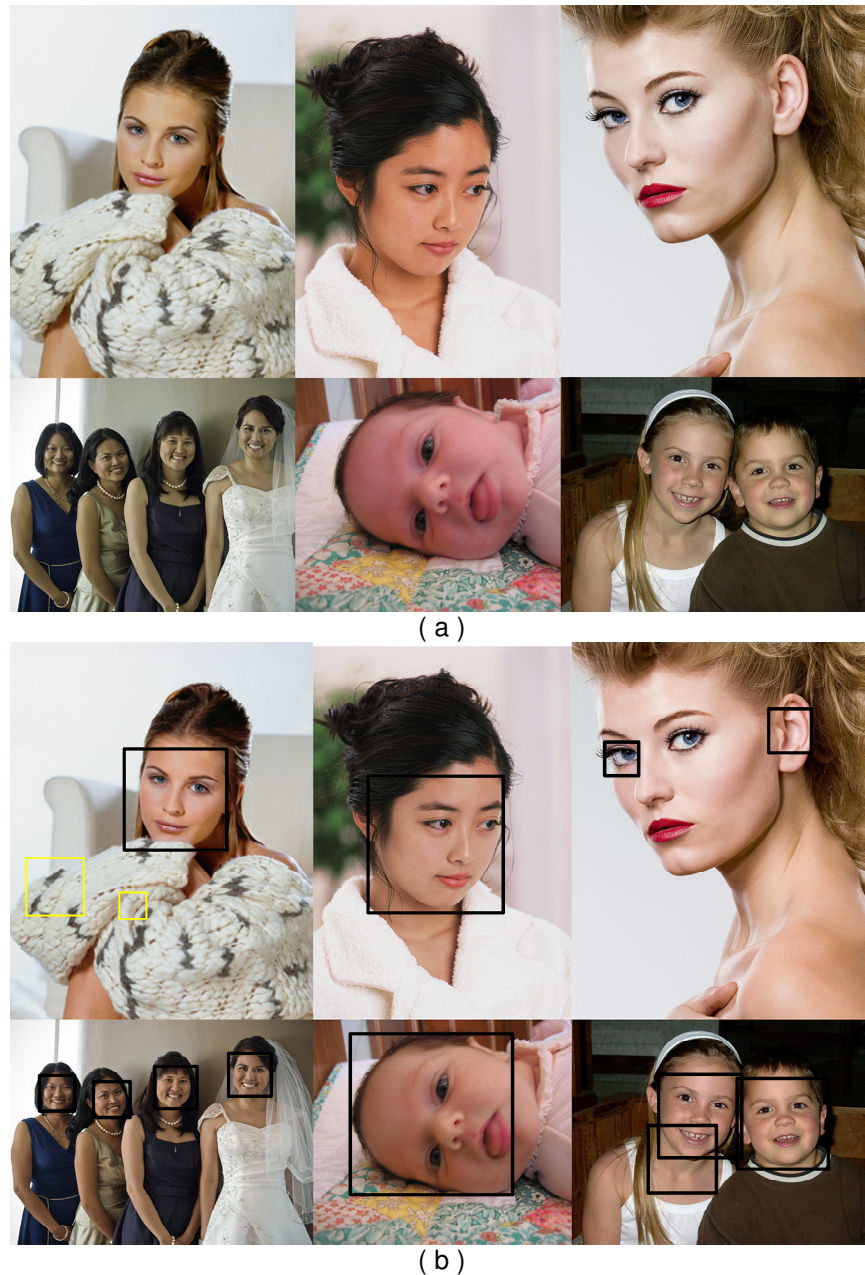


FIGURE 6: A Block Diagram for Skin Color Enhancement.

Because the skin color detection model in Step 0 is set to detect a larger region of skin colors than that for the skin color enhancement using a static skin color model, the new method adjusts a wider range of skin colors that deviate more seriously from normal skin colors. Since the ellipse used in Step 3 and Step 4 is adapted to a specific image, this method adjusts skin colors more effectively than those approaches using a static skin color model.

This algorithm works well in a uniform luminance-chrominance color space, such as CIELAB or CIECAM02 even if a single PSCC is applied to all lightness levels. Replacing the working color space with YCbCr improves the computation efficiency for processing RGB images and videos for display color enhancement. However, chroma coordinates (CbCr) of preferred skin colors are changed significantly among different luma values, Y. CbCr of preferred skin colors and skin color ellipses should be adapted to different luma for skin color enhancement.

A few cases explaining the improvement of the approach over approaches utilizing static skin color models are discussed in Appendix.

6. SKIN COLOR MORPHING BY TRIANGULAR SUB-DIVISION

Once a weight is computed for a skin color and a target preferred color is determined, Eq. (6) is applied to adjust the skin color. To avoid contouring, the weight, w in Eqs. (4-5), should be less than 1 [29]. In this section, an alternative color adjustment method is proposed. Figure 7 shows a skin color ellipse that is divided into a set of triangles [30]. Each triangle connects two neighbor points on the skin color ellipse boundary and the central point A. The more triangles are used, the more accurate the ellipse is represented. As the statistical skin color center, A, is moved toward a PSCC, B, all colors in each triangle (e.g. A-C0-C1) are morphed to its corresponding triangle (e.g. B-C0-C1). Since gamut boundary colors are not changed, it guarantees smooth color adjustment.

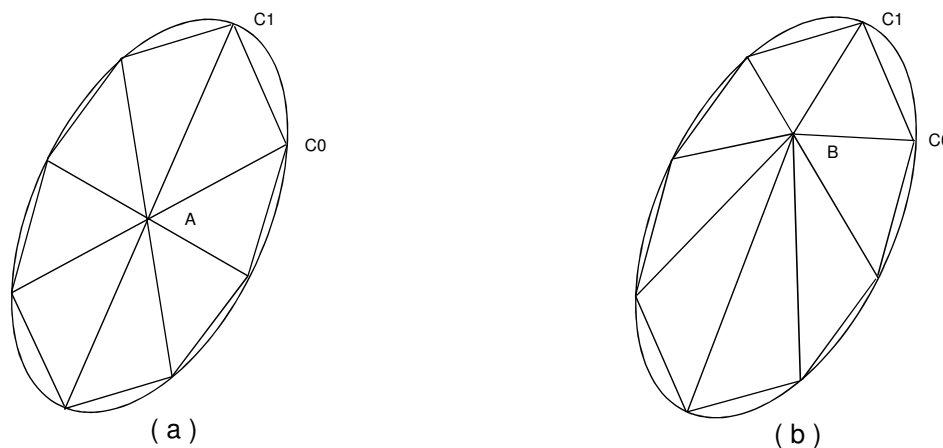


FIGURE 7: A skin color ellipse is approximated with a set of triangles for color adjustment. The statistical skin center (A) in (a) is mapped to a PSCC (B) in (b).

To adjust a skin color, a triangle that contains the color is found and a triangle-based interpolation is performed. Let's assume a skin color, P, is within the triangle A-C0-C1, as shown in Figure 8. As the mean skin color, A, is morphed to a PSCC, B, while C0 and C1 stay unchanged, P is mapped to P'. A method to preserve area ratio of three sub-triangles may be used to compute the adjusted color P'. Areas of triangles A-P-C0, A-P-C1, and P-C0-C1 are noted S_{A-P-C0} , S_{A-P-C1} , and $S_{P-C0-C1}$, respectively. They are computed using three known colors, A, C0, and C1, and the original skin color P. P' is computed by the following equation:

$$P = \frac{(C0 \cdot S_{A-P-C1} + C1 \cdot S_{A-P-C0} + B \cdot S_{P-C0-C1})}{(S_{A-P-C1} + S_{A-P-C0} + S_{P-C0-C1})} \quad (9)$$

An advantage of this method over the method described in earlier sections is that the statistical color center of the original skin colors A is exactly mapped to a target preferred skin color B and all other colors are morphed smoothly.

If lightness is modified together with chrominance, the morphing method must be expanded to a 3-D space. The original central color, A, in each sampled lightness level is mapped to a corresponding PSCC, B, in which lightness may be different. A set of sampling points to represent the boundary of the skin color gamut defined with a skin color model are generated. The boundary sampling points and the lightness-dependent points, A and B, are used to construct a set of tetrahedrons. Tetrahedral interpolation is applied to map each original skin color to a corresponding preferred skin color.

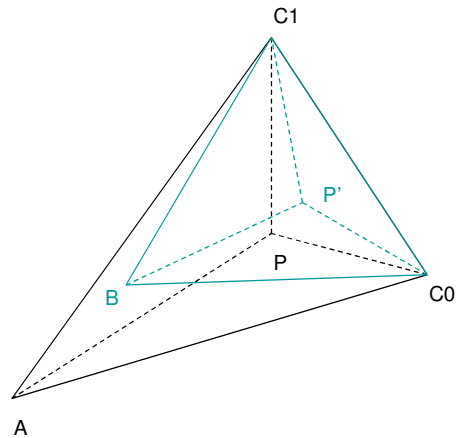


FIGURE 8: An original skin color, P, on a triangle A-C0-C1 is mapped to P' on another triangle B-C0-C1.

7. EXPERIMENTAL RESULT AND DISCUSSION

A psychophysical experiment was conducted to validate the effectiveness of the skin color enhancement algorithm. A calibrated monitor was used to display samples. 50 Caucasian images and 50 Oriental images were selected for testing. Most of images were captured with consumer digital cameras. 25 Caucasian observers and 25 Asian observers, who had normal visions, judged Caucasian images and Oriental images, respectively.

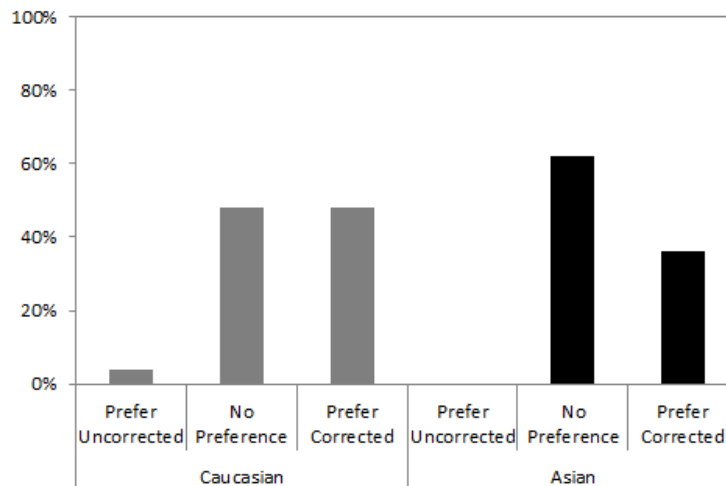


FIGURE 9: Psychophysical Experimental Result.

The result is plotted in Figure 9. For Caucasian images, 48% of images are significant improved while 4% have negative impact and the rest have no visual impact. For Asian images, 36% are significantly improved and the rest are visually not changed. The original images with skin tones close to optimal are mostly not changed visually. It verifies that the skin color enhancement method effectively improves skin color preference and does not damage good skin tones.

We also compared our method with skin color enhance using static skin color models. Our method is able to enhance images captured under more extended range of conditions. Skin color enhancement on each detected face, although able to process skin colors captured under a white range of conditions, is much less reliable than our method because of higher requirement on the accuracy of face detection and less number of pixels for statistical analysis of face tones.

8. CONCLUSION

A new skin color enhancement method is introduced for preferred skin color reproduction. If face detection is not enabled, it applies a statistical skin color model to detect skin colors and morphs skin colors toward a preferred skin color center. Preferred skin colors are adapted to different lightness to improve skin color preference. The highlight region is processed differently to maintain the contrast.

If skin colors are far away from the regular skin color distribution, applying a static skin color model for skin color enhancement is not able to detect skin colors effectively and therefore the method will not effectively enhance skin colors. With face detection, a skin color model adapted to the scene is constructed to resolve the limitation of the algorithm using statistical skin color models. By applying the scene-dependent skin color model, skin colors that are far away from normal skin colors are detected more effectively and the result of skin color enhancement is greatly improved. In addition, the result of the skin color analysis is applied to adjust overall color balance of the image, and therefore the method produces more accurate white balance on the image.

The current method is not able to enhance face tones of different culture backgrounds in a single image differently. This is a direction for further research. Another area for further research is to study the effectiveness of the method working on different color spaces, especially in YCbCr space to improve the computation efficiency.

9. APPENDIX: CASE STUDY OF SKIN COLOR ENHANCEMENT UTILIZING SCENE-DEPENDENT SKIN COLOR MODEL

A few cases illustrating the benefit of utilizing image-dependent skin color model for skin color enhancement are discussed below.

Case 1: Skin tones are highly chromatic. In Figure A1, the thick-black ellipse is the skin color region of an elliptical skin color model and the dash ellipse is derived from the skin colors of an image. If the static skin color model is applied to detect skin colors, skin colors that are out of the thick-black elliptical region will not be adjusted, and the adjustment strength of each skin color is proportional to the Mahalanobis distance calculated using the static skin color model. A color at location A has the highest adjustment strength and colors around B that are close to the boundary of the thick-black ellipse are adjusted very weakly. With the image-dependent skin color modeling, the ellipse skin color model is adjusted to center at the mean skin color of the face boxes of image. The gray ellipse becomes the elliptical model for skin color detection. The strength for skin color adjustment is proportional to the Mahalanobis distance calculated using the gray ellipse. A color at location B has the highest adjustment strength, and a color on the boundary of the gray ellipse is not adjusted. Because the image-dependent skin color ellipse (the gray ellipse) is adapted to colors of the skin pixels of the image, the size of the ellipse may be reduced. It is clear that the image-dependent skin color modeling is more effective for skin color adjustment.

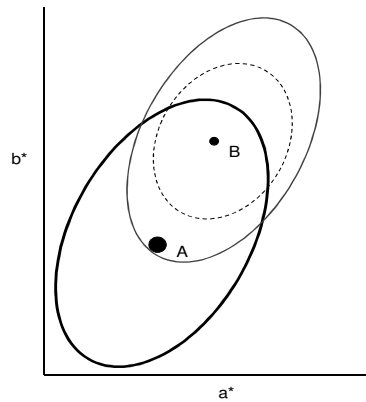


FIGURE A1: A sketch diagram illustrating effective adjustment of highly chromatic skin colors.

Case 2: Skin hues are very different from regular skin hues. In Figure A2, the large solid-black ellipse illustrates an elliptical skin color model and the dashed ellipse illustrates the skin tones of an image. Skin colors are very yellowish. Again, the strength of the skin color adjustment is proportional to the Mahalanobis distance computed using the skin color model. With the static skin color model, a color at location A has the highest adjustment strength and a color on the boundary of the black elliptical region or out of the ellipse has zero strength (no adjustment). Thus, the yellowish skin colors that are out of the black skin color ellipse are not adjusted. With the image-dependent skin color model, the ellipse skin color model is adjusted to center at the mean skin color of the face boxes of the image. Thus the skin color detection ellipse is shifted to the gray ellipse. Skin colors to be adjusted are detected using this gray ellipse and the strength of adjustment is proportional to the Mahalanobis distance computed using the gray ellipse. A color at location B has the highest adjustment strength, and a color on the boundary of the gray ellipse is not adjusted. Because the image-dependent skin color ellipse (the gray ellipse) is adapted to skin colors of the image, the size of the ellipse may be reduced. In this case, it is very likely that the illuminant detection on the image is incorrect and the entire image is shifted to the yellowish direction. Because the method of skin color enhancement adjusts all colors within the gray ellipse, all yellowish colors in the image are adjusted. Therefore, an additional benefit of the method is that it removes the yellow cast. The relative location of B and A can actually be used to assist illuminant detection and white balance.

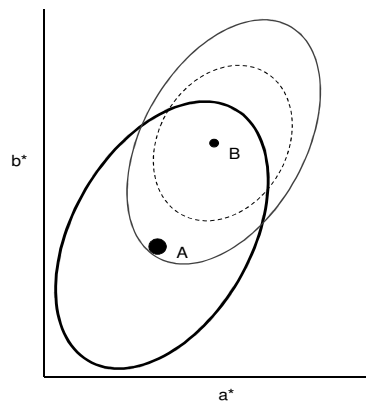


FIGURE A2: A sketch diagram illustrating the effective adjustment of yellowish skin tones.

Case 3: Bright and dark skin colors. Figure A3 shows an elliptical skin color model in a large solid-black ellipse and the skin color region of an image in a small dash ellipse. Very bright and very dark skin colors in digital photographic images tend to have lower chroma. If a skin color model is not adapted to very bright (or overexposed) and very dark (or underexposed) images, the skin elliptical model tend to have a mean chroma higher than the mean skin chroma of the image. With a static skin color model, high chroma skin colors close to the center, A, of the skin

color model is adjusted the most. And very low chroma skin colors that are close to the boundary of the black ellipse or out of the ellipse are adjusted very weakly or are not adjusted. With the image-dependent skin color modeling, an ellipse skin color model is adjusted to have the ellipse centered at the mean skin color of the face boxes. The skin color detection ellipse is moved from the solid-black ellipse to the gray ellipse. Skin colors to be adjusted are detected using this gray ellipse and the strength of the adjustment is computed based on the Mahalanobis distance to the center B instead to the point A. It is clear that the image-dependent method adjusts skin colors more properly and more effectively. In this case, the skin color ellipse may be reduced because color gamut in highlight or shadow is smaller than that in medium tones.

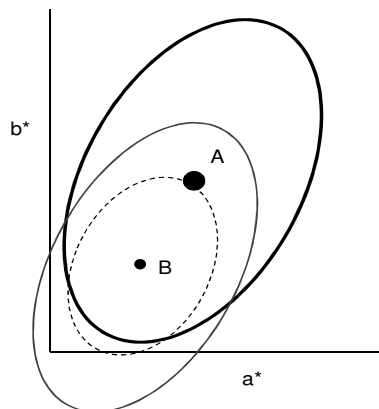


FIGURE A3: A sketch diagram illustrating effective and proper adjustment of pale skin tones.

In summary, a skin color model adapted to face tones of an image can be constructed by utilizing the face detection information. It models skin tones of the image more accurately than a static skin color model, and results in more effective skin color detection and skin color enhancement.

10. Acknowledgment

The authors thank Prof. Jon Y. Hardeberg from the Norwegian Color Research Laboratory at Gjøvik University College for his comments and suggestions.

11. REFERENCES

- [1] C.J. Bartleson, "Some observations on the reproduction of flesh colors", *Photographic Science and Engineering* 3, No. 3. 114-117 (1959).
- [2] C.J. Bartleson, "Memory colors of familiar objects", *J. Opt. Soc. Am.*, vol. 50, no. 1, pp. 73-77, 1960.
- [3] R.W.G. Hunt, I.T. Pitt, and L.M. Winter, "The preferred reproduction of blue sky, green grass and Caucasian skin in color photography", *J. Photographic Science* 22, 144-149 (1974).
- [4] T.J.W.M. Janssen and F.J.J. Blommaert, "Predicting the usefulness and naturalness of color reproductions", *J. Imaging Sci. Technol.* 44, No. 2, 93-104 (2000).
- [5] S.N. Yendrikhovskij, F.J.J. Blommaert, and H. Ridder, "Color reproduction and the naturalness constraint", *Color Res. Appl.* 24, no. 1, 54-67 (1999).
- [6] S. R. Fernandez and M. D. Fairchild, "Preferred Color Reproduction of Images with Unknown Colorimetry", *Proc. IS&T/SID 9th Color Imaging Conference*, 274-279 (2001).
- [7] K.M. Braun, "Memory color enhancement algorithm", *International Congress of Imaging Science*, 227-229 (2006).

- [8] D. H. Kim, H. C. Do, S. Il Chien, "Preferred skin color reproduction based on adaptive affine transform", IEEE Trans. Consumer Electron. 51, No. 1, 191-197 (2005).
- [9] D.S. Park, Y. Kwak, H. Ok, and C.Y. Kim, "Preferred skin color reproduction on the display", J Electronic Imaging 15, 041203, (2006).
- [10] X. Xu and H. Pan, "Skin and sky color detection and enhancement system", US Patent Application, US 2010/0322513.
- [11] H. Pan and J. Daly, "Skin color cognizant GMA with luminance equalization", US Patent Application, US 2010/0194773.
- [12] Hila Nachlieli, Gitit Ruckenstein, Darryl Greig, Doron Shaked, Ruth Bergman, Carl Staelin, Shlomo Harush, Mani Fischer, "Face and skin sensitive image enhancement", US Patent, 8,031,961, 2011.
- [13] H. Zeng, "Preferred skin colour reproduction", PhD dissertation, University of Leeds, Leeds, UK, 2011.
- [14] J. Y. Lee and S. I. Yoo, "An elliptical boundary model for skin color detection", Proc. International Conference on Imaging Science, Systems, and Technology, 579-584 (2002).
- [15] H. Zeng and R. Luo, "Skin color modeling of digital photographic images", J. Imaging Sci. Technol. 55, No. 3, 030201 (2011).
- [16] C.P. Bartleson, and C.P. Bray, "On the preferred reproduction of flesh, blue-sky, and green-grass colors", Photographic Science and Engineering 6, No. 1, 19-25 (1962).
- [17] C.L. Sanders, "Color preference for natural objects", Illumination Engineering, 452-456 (1959).
- [18] T. Yano, K. Hashimoto, "Preference for Japanese complexion color under illumination", Color Res. Appl. 22, No. 4, 269-274 (1997).
- [19] J. Kuang, X. Jiang, S. Quan, and A. Chiu, "A psychophysical study on the influence factors of color preference in photographic color reproduction", Proc. SPIE 5668, 12-19 (2005).
- [20] S.R. Fernandez, M.D. Fairchild, and K. Braun, "Analysis of observer and cultural variability while generating preferred color reproductions of pictorial images", J. Imaging Sci. Technol. 49, No. 10, 96-104 (2005).
- [21] P. Bodrogi, T. Tarczali, "Colour Memory for Various Sky, Skin, and Plant Colours: Effect of the Image Context", Color Res. Appl. 26, No. 4, 278-289 (2001).
- [22] H. Zeng, and R. Luo, "Color and tolerance of preferred skin colors", Color Res. and Appl., accepted for publication.
- [23] H. Zeng and R. Luo, "Preferred skin colours of African, Caucasian, and Oriental", Proc. IS&T/SID 19th Color Imaging Conference, 211-216 (2010).
- [24] P. Viola, and M. Jones, "Robust real-time face detection," International J. Computer Vision 57, No. 2, 137-154 (2004).
- [25] S.A. Inalou and S. Kasaei, "AdaBoost-based face detection in color images with low false alarm", Second International Conference on Computer Modeling and Simulation, 107-111 (2010).

[26] Z. Zakaria and S. A. Suandi, "Face detection using combination of neural network and Adaboost", IEEE Region 10 Conference, 335-338 (2011).

[27] K. J. Karande and S. N. Talbar, "Independent Component Analysis of Edge Information for Face Recognition", International Journal of Image Processing, Vol. 3(3) 120-130 (2009).

[28] L. Pitchai and L. Ganesan, "Face Recognition Using Neural Networks", Signal Processing: An International Journal, Vol. 3(5) 153-160 (2009).

[29] H. Zeng and R. Luo, "A preferred skin color enhancement method for photographic color reproduction", Proc. SPIE 7866, 786613 1-9 (2011).

[30] S. Hung, X. Jiang, and H. Li, "Image processing method and systems of skin color enhancement", World Intellectual Property Organization, WO 2010/071738.

Fast Motion Estimation for Quad-Tree Based Video Coder Using Normalized Cross-Correlation Measure

Eskinder Anteneh Ayele

*Research Scholar/ Department of Electronics Engineering
Visvesvaraya National Institute of Technology
Nagpur,440022, India*

eskinderanteneh@yahoo.co.uk

R. E. Chaudhari

*Asst. Professor/Dept. of ECE
St. Francis Institute of Technology
Mumbai,400103, India*

rec77@rediffmail.com

S. B. Dhok

*Asso. Professor/Department of Electronics Engineering
Visvesvaraya National Institute of Technology
Nagpur,440022, India*

sbdhok@vnit.ece.ac.in

Abstract

Motion estimation is the most challenging and time consuming stage in block based video codec. To reduce the computation time, many fast motion estimation algorithms were proposed and implemented. This paper proposes a quad-tree based Normalized Cross Correlation (NCC) measure for obtaining estimates of inter-frame motion. The measure operates in frequency domain using FFT algorithm as the similarity measure with an exhaustive full search in region of interest. NCC is a more suitable similarity measure than Sum of Absolute Difference (SAD) for reducing the temporal redundancy in video compression since we can attain flatter residual after motion compensation. The degrees of homogeneous and stationery regions are determined by selecting suitable initial fixed threshold for block partitioning. An experimental result of the proposed method shows that actual numbers of motion vectors are significantly less compared to existing methods with marginal effect on the quality of reconstructed frame. It also gives higher speed up ratio for both fixed block and quad-tree based motion estimation methods.

Keywords: FFT, Motion Estimation, Normalized Cross Correlation, Quad-tree, Video Compression.

1. INTRODUCTION

Motion estimation and compensation are the two crucial processes in block based video coding standards. A major technique known as motion estimation is used to compress the videos by removing the redundant information from successive frames. Inter-prediction explores temporal redundancy between frames to save coding bits [1]. By using motion compensated prediction, the best matching position of current block is found within the reference picture so that only prediction difference needs to be coded. Each prediction unit coded using inter-prediction, has a set of motion parameters, which consists of a motion vector, a reference picture index, and a reference list flag. Motion estimation is widely used in various applications related to computer vision and image processing, such as object tracking, object detection, pattern recognition and video compression, etc. Especially, block-based motion estimation is very vital for motion-compensated video compression, since it reduces the data redundancy between frames to achieve high compression ratio. Because of the high redundancy that exists between the consecutive frames of a video image sequence, a current frame can be reconstructed from a previous reference

frame and the difference between the current and previous frames by using the motion information.

The idea behind block matching is to divide the current frame into a matrix of macro blocks that are then compared with corresponding block and its adjacent neighbors in the previous frame to create a vector that stipulates the movement of a macro block from one location to another in the previous frame. This movement calculated for all the macro blocks comprising a frame, constitutes the motion estimated in the current frame. International standards for video communications such as MPEG-1/2/4 and H.261/3/4 employ motion compensation prediction which is based on regular (fixed- or near-fixed-size) block-based partitions of incoming frames. While such partitions require a minimal amount of overhead information they provide little or no adaptation to picture content. A notable departure from this practice has been the recently emerged H.264 standard which allows a degree of flexibility in the choice of block size. Motion estimation based on quad-tree partitions achieves a good balance between a degree of adaptation to picture content on one hand and low-complexity, low-overhead implementation on the other [2].

Block matching algorithms used for motion estimation in video compression differ in the matching criteria (e.g. Mean Square Error (MSE), SAD, cross-correlation), the search strategy (e.g. Full Search, Three Step Search, Four Step Search etc.), and the determination of block size (e.g. hierarchical, adaptive) [3]. In this paper, we adopt the normalized cross-correlation methodology and employ it in the framework of a quad-tree motion estimation scheme that provides a level of adaptation to picture contents without incurring substantial overheads. Our approach lies in the combination of these two concepts, namely quad-tree decomposition and cross correlation applied in the frequency domain using Fast Fourier Transform (FFT) algorithm, yielding partitions for which a monotonic decrease of the motion compensated prediction error.

Upon this introductory section, the rest of the paper is organized as follows: In Section 2, we briefly review the block partitioning principles underlying quad-tree partitioning of a frame for motion estimation. In Section 3, we formulate our quad-tree FFT-based normalized cross-correlation algorithm approach. In Section 4, we present the experimental/simulation results with a brief observational discussion. At the end, Section 5 concludes the paper.

2. QUAD TREE PARTITIONING

The proposed motion estimation scheme involves the quad-tree partitioning of a frame which provides a better level of adaptation to scene contents compared to fixed block size approaches. Quad-tree decompositions are achieved by using the motion compensated prediction error to control the partition of a parent block to four children quadrants [1] [4] [5] [6].

Figure 1a demonstrates a quadtree structure implemented in this paper. Here, we employed four level of quad-tree partitioning with block sizes, 32x32, 16x16, 8x8, and 4x4 pixels in which the macro block (32x32) can be partitioned into four 16x16. A sub-macro block (16x16) can be further partitioned to four 8x8 blocks and finally an 8x8 block can be also portioned into four 4x4 blocks as quadtree structure depending on threshold. The algorithm is implemented along the recursive raster scanning path [7], which has been traditionally used in the quadtree decomposition as can be seen from Figure 1b. To simplify the search for a good scanning path, we require that child blocks, which belong to the same parent block, are scanned in sequence. Therefore a high correlation between successive blocks will result in an efficient encoding.

A motion vector is generated for each block after a search is conducted to best match the movement of each block from a previous reference frame. However, large block-sizes generally produce poor motion estimation, thereby producing a large motion-compensated frame difference (error signal). Conversely, small block-sizes generally produce excellent motion estimation at the cost of increased computational complexity and the overhead of transmitting the increased number of motion vectors to a receiver. Thus, the balance between high motion vector overhead

and good motion estimation is the focus of a quadtree-based variable block-size motion estimation method. The scheme is based on normalized cross correlation and uses key features of the cross correlation to control the partition of a parent block to four children quadrants.

2.1 Partition Criteria

In this paper, different sizes of block partition are used to minimize the number of motion vectors to be sent. This is based on the assumption that the higher the degree of homogeneous and stationary blocks, the larger is the block partition used. However, the thresholds to determine the degree of homogeneity are empirically selected, and the resulting criteria cannot provide very accurate block partitioning.

Initially the current frame is divided into non-overlapping blocks of size 32x32. The first threshold th_1 is decided based on the output video quality of homogeneous and stationary regions and tested at the same location in the previous reconstructed frame. If the error exceeds the threshold th_1 , the bigger size of macroblock is partitioned into four quadrants of size 16x16. For higher levels another threshold is applied at four different cases. Results show that, the scheme provides a better level of adaptation to scene contents and outperforms fixed block size scheme in terms of different number of motion vectors for the same level of motion compensated prediction error and Structure Similarity (SSIM). The partition criterion also guarantees a monotonic decrease of the motion compensated prediction error with an increasing number of iterations [8].

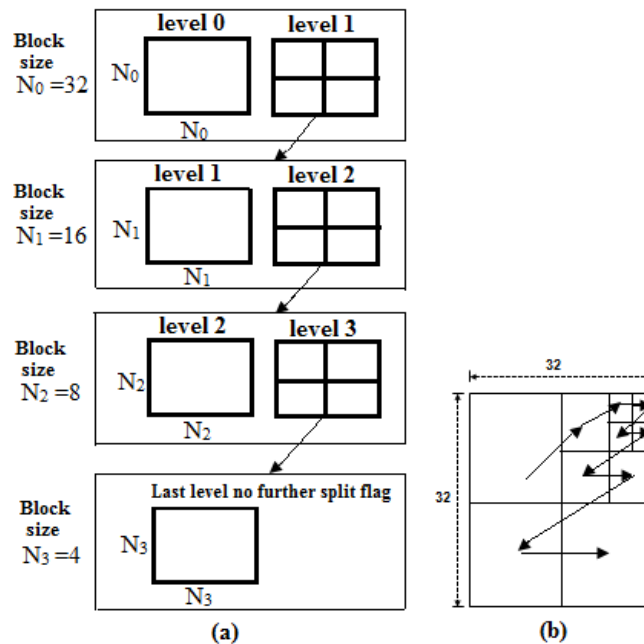


FIGURE 1: (a) A Quad tree Structure, (b) Recursive raster scan for Quad tree decomposition.

3. FFT-BASED NORMALIZED CROSS-CORRELATION

Many block-based motion estimation algorithms have been proposed and developed for finding the block with the smallest matching error including, phase-domain methods [9], time/space-domain methods [10], and spline-based methods [11]. Time-domain (1-D) or space-domain (2-D) methods have been widely and frequently used because of their high accuracy, precision, and resolution, and relative simplicity in implementation [12]. In terms of block distortion measure, the SAD is commonly used. In addition to SAD, the NCC is also a popular similarity measure. The NCC measure is more dynamic than SAD under uniform illumination changes. NCC can improve subjective visual quality as well as coding efficiency in video compression [2]. However, the NCC

is a more complex criterion compared to SAD. SAD is used to find the best match with the lowest matching error and NCC is to find the best macro block whose overall intensity variation is most similar to current macro block. Though the error of NCC for motion estimation is larger, but it is more uniformly distributed than SAD based. These flat error results in large DC term and smaller AC term DCT coefficients, which mean less information loss after quantization.

One of the main motivations for this paper has been the current interest in motion estimation techniques operating in the frequency domain. These are commonly based on the principle of cross correlation and offer well-documented advantages in terms of computational efficiency due to the employment of fast algorithms [1]. Correlation is widely used as an efficient similarity measure in matching tasks. However, traditional correlation based matching methods are limited to the short baseline case. NCC is the most robust correlation measure for determining similarity between points in two frames (images) and provides an accurate foundation for motion estimation. However, implementing directly in spatial domain is too computationally intense especially for rapidly managing several large frames [2]. A significantly faster method of calculating the NCC is presented using FFT method to speed up block matching for computationally efficient video encoding.

3.1 The Algorithm

The best match is defined in terms of NCC [13] by shifting a macro block pixel by pixel across the search window. Correlation planes provide information where the macro block best match the search window. The correlation coefficient conquers the difficulties in [14, 15] by normalizing the current and reference frame vectors to unit length, yielding a cosine-like correlation coefficient. It is defined in spatial domain as:

$$NCC(x, y) = \frac{c(x, y)}{\|T(x, y)\| \|I(x + i, y + j)\|} \quad (1)$$

The pixel location (x, y) corresponding to the maximum NCC value matches to best location (motion vector) of MB in the search window. At which, $x \in \{0, 1, \dots, N-M\}$ and $y \in \{0, 1, \dots, N-M\}$. For example if $N = 24$ and $M = 8$, the number of NCC coefficients are (17x17).

Where, $c(u, v)$ is the cross-correlation, T is the current macro block of size $M \times M$, and I is the search window of reference frame of size $N \times N$ ($N > M$). The norms of the current and the reference frame in (1) are defined, respectively, as follows:

$$\|T(x, y)\| = \sqrt{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} T(i, j)^2}$$

$$\|I(x + i, y + j)\| = \sqrt{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} I(x + i, y + j)^2}$$

A significantly more efficient way of calculating the NCC is by computing the numerator of (1) via FFT because the calculation of the numerator dominates the computational cost of the NCC. More specifically, cross-correlation in the spatial domain, which is the numerator in (1), is equivalent to multiplication in the frequency-domain:

$$\text{i.e } c(x, y) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} T(i, j) \cdot I(x + i, y + j)$$

$$\Rightarrow C(u, v) = T(u, v)I(u, v)$$

$$c(x, y) = \mathcal{F}^{-1}(C(u, v)) \quad (2)$$

Basically Equation (2) corresponds to computing a 2D FFT on the current and the search window of the frames followed by a complex-conjugate multiplication of the resulting Fourier coefficients. However, in order to avoid a complex-conjugate multiplication, we computed the current frame macro block via IFFT as shown in Figure 2. The final products are then inverse Fourier transformed to produce the actual coefficient cross correlation plane. The use of FFT in numerator calculations of (1) is used to reduce the number of NCC calculations.

Similarly the denominator calculations are performed by pre-computing the energy of the entire searching windows of a frame. Whenever the search window moves from block to block with respect to current macro block, we change the location of energy block which is stored as a look-up table. Thus it further reduces the computation complexity of the algorithm.

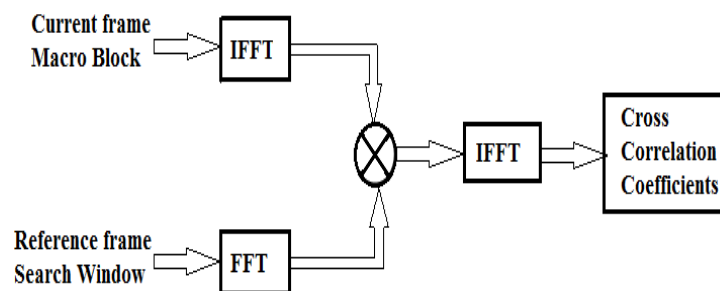


FIGURE 2: Implementation of the numerator of NCC by using FFT algorithm.

4. OBSERVATION AND RESULTS

As performance measure parameters, MSE and Peak Signal to Noise Ratio (PSNR) are used to evaluate the subjective quality of a reconstructed video sequence. Applying the NCC as the matching criterion to motion estimation leads to more uniform residuals. Hence, the NCC can improve subjective visual quality as well as coding efficiency in video compression. Recently, visual quality measures focusing on the human visual system (HVS) have been devised in place of PSNR. Among these measures, SSIM has become popular. The SSIM index is more consistent with human perception and is designed to measure structural information degradation, including the three comparison points of luminance, contrast, and structure.

Apart from the prediction error criterion, computational complexity is also a key criterion for the performance evaluation of fast block matching algorithms. The computational complexity can be directly compared by counting the number of searching points required. The number of searching points is a measure of search speed whereas the computation time is another speed measure that also takes into account the overhead of the algorithm. The overhead includes time spent on storing and fetching spatio-temporal predictors, making comparisons etc. Hence in general the computation time is a better measure of determining computational speedup.

The experimentations are performed on five standard video sequences with frame size of 288x352 through four different threshold cases for quadtree partitioning: case-I (10, 16), case-II (12, 18), case-III (15, 21) and case-IV (18, 24). For each case (th2, th3), the threshold th2, th3 are randomly chosen in the second and third levels of partitioning respectively to get minimum number of Motion Vectors (MVs). Based upon visual quality observation from the simulated result, we fixed the first level threshold =10 for the bigger size of macro block (32x32) as a

stationary (static) block. By keeping this threshold, we determined the PSNR and SSIM results of the reconstructed videos, which are similar to fixed block result with less number of bits. The corresponding average numbers of static blocks and bits per block is shown in table-1.

Video Sequences	Avg. No. of static blocks	Avg. no. bits per statics blocks
Foreman	51.20	2.50
Paris	3.58	3.00
Carphone	61.58	2.00
Tennis	66.18	2.54
News	94.30	1.00

TABLE 1: Average numbers of Static/Stationary blocks and bits per block of the test videos

Tables 2 shows simulations of the average PSNR, SSIM, and encoding times of different videos for fixed block size (8x8) using SAD and FFT-based NCC. All simulations were done on Matlab-7.9 using a Pentium 4 desktop with 3.0GHz CPU and 1.0Gmb of RAM. The experimental results show that efficient FFT based NCC full search algorithm can provide slightly higher PSNR and better SSIM in the reconstructed frame than the traditional SAD-based fixed block ME method. Further it reduces the encoding time by more than 50%.

Video sequences	PSNR (dB)	SSIM	Encoding Time(Sec)
Using SAD			
Foreman	34.51	0.8995	3.87
Paris	30.98	0.9514	3.89
Carphone	39.29	0.9279	3.86
Tennis	29.94	0.8013	3.84
News	39.48	0.9309	3.89
Using FFT-based NCC			
Foreman	34.75	0.9042	1.50
Paris	31.02	0.9536	1.53
Carphone	39.68	0.9342	1.50
Tennis	29.97	0.8028	1.52
News	39.54	0.9290	1.48

TABLE 2: Comparisons of average PSNR, SSIM, and Encoding time for fixed block size (8x8)

Table 3 shows the comparison of the performance parameter viz. average number of motion vectors, encoding time, PSNR and SSIM for the Quadtree FFT-based NCC method with four threshold cases as mentioned above. The results show in all cases that the Quadtree FFT-based NCC method is just as accurate as SAD method but the Quadtree FFT-based NCC method faster than the standard SAD method. Depending on the type of the video sequences and the threshold levels, the method is about 2 to 5 times faster than SAD-based search criteria.

Figure 3 summarizes the number of motion vectors for the luminance components of the first 12 inter-frames of the 'Tennis' video sequence. The graph shows that the proposed algorithm is substantially dependent on threshold levels and it shows the variation of MVs from frame to frame.

Video Sequences	PSNR (dB)	SSIM	Time (Sec)	No. of Motion Vectors			
				32x32	16x16	8x8	4x4
Case-I							
Foreman	34.07	0.8852	1.28	51.36	151.95	131.73	98.55
Paris	30.83	0.9436	2.16	3.55	262.27	452.18	104.00
Carphone	38.82	0.9075	1.15	60.82	125.45	83.73	101.45
Tennis	29.64	0.7979	1.55	65.64	45.18	320.82	129.09
News	39.42	0.8967	0.77	93.36	15.45	21.00	29.45
Case-II							
Foreman	34.01	0.8851	1.14	51.45	165.45	82.91	64.00
Paris	30.78	0.9432	1.91	3.55	297.45	319.00	73.82
Carphone	38.80	0.9076	1.07	60.64	134.64	56.64	74.55
Tennis	29.92	0.7977	1.17	65.73	96.00	125.54	91.27
News	39.44	0.8967	0.70	93.36	17.09	16.45	21.45
Case-III							
Foreman	33.98	0.8849	1.03	51.45	177.73	40.91	35.64
Paris	30.78	0.9423	1.71	3.55	322.54	224.09	52.00
Carphone	38.89	0.9075	0.98	60.55	142.18	34.45	48.36
Tennis	29.87	0.7968	0.99	65.64	115.45	59.36	50.55
News	39.46	0.8967	0.67	93.36	18.91	11.73	11.27
Case-IV							
Foreman	33.97	0.8850	0.98	51.45	183.27	23.18	17.82
Paris	30.73	0.9419	1.60	3.55	339.64	156.45	49.09
Carphone	38.92	0.9074	0.94	60.55	146.91	20.09	30.18
Tennis	29.82	0.7961	0.92	65.18	126.54	25.73	36.73
News	39.47	0.8967	0.65	93.36	19.91	8.82	6.91

TABLE 3: Performance parameters for Quadtree FFT-based NCC method with four randomly selected threshold.

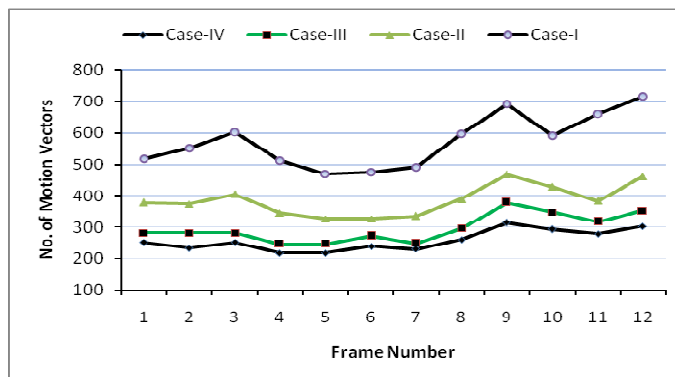


FIGURE 3: Number of MVs vs frame number for Tennis video at different thresholds.

The NCC can be often a better criterion than the SAD in terms of PSNR. In order to verify this, a full search based on SAD was compared to that based on NCC, where the search range and matching block size were fixed to ± 8 pixels and 8×8 pixels in terms of integer-pel accuracy. Figure 4 shows the experimental results for 12 sequences of Tennis video. Here, the PSNR values are computed from the motion compensated version of the second frame of each sequence in order to evaluate the performance of the motion estimation only. The results show that the NCC provides slightly better PSNR performance than the SAD. This means that the

motion compensated frames using NCC-based motion estimation are visually better than that those using SAD-based motion estimation in general.

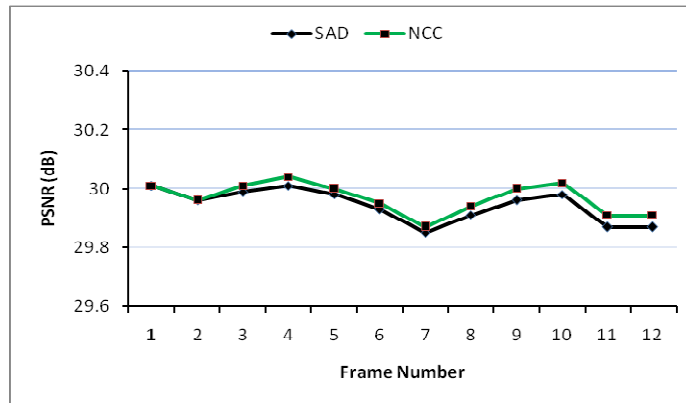


FIGURE 4: PSNR vs frame number using SAD and FFT-based NCC methods for Fixed Block Size.

Finally, the proposed algorithm (QT-NCC using FFT) was compared with Fixed Block (FB)-NCC and FB-SAD using same video frames. As seen from figure 5, the proposed algorithm can improve the speed-up ratio up to about 2.5 and 4.0 times in comparison with the FB-NCC and FB-SAD algorithms respectively but keeping SSIM and PSNR values almost the same for all algorithms.

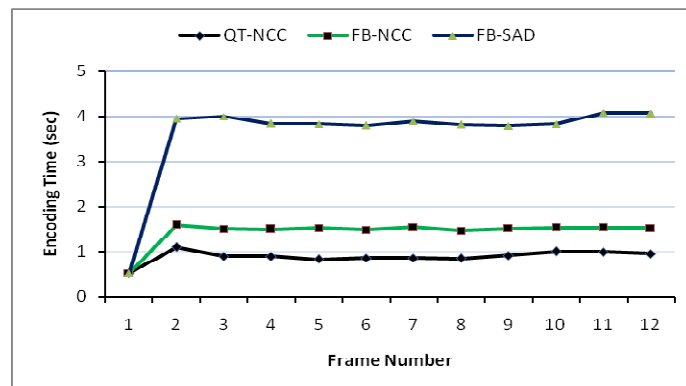


FIGURE 5: Encoding Time vs frame number for QT-NCC, FB-NCC, and FB-SAD algorithms.

5. CONCLUSIONS

This paper proposes a fast Quadtree FFT- based NCC, where re-using the energy part of search window is employed to skip unnecessary block-matching calculation and the cross correlation is determined in frequency domain based on FFT algorithm. Because of the quad-tree partitioning of a frame, it provides a better level of adaptation to scene contents compared to fixed block size approaches. Hence, the proposed algorithm considerably reduces the computational complexity and improves the speed-up ratio of about 4 times in comparisons with FB-NCC and FB-SAD algorithms. Moreover, for video sequences which contain more static data it requires less number of bits to encode without motion vector. For further quality improvement of reconstructed frames, one can use half or quarter pixel interpolation techniques. Correspondingly to enhance the speed ratio the algorithm can be modified and implemented using basis functions.

6. REFERENCES

- [1] V. Argyriou and T. Vlachos, "Motion estimation using quad-tree phase correlation", IEEE International Conference on Image Processing, 2005, vol. 1, pp. I-1081-I-1084.
- [2] B. C. Song, "A Fast Normalized Cross Correlation-Based Block Matching Algorithm Using Multilevel Cauchy-Schwartz Inequality", ETRI Journal, vol. 33, no.3, pp. 401-406, June 2011.
- [3] A. Barjatya, "Block Matching Algorithms for Motion Estimation", Digital Image Processing (DIP 6620) ,Final project paper, Utah State University, Spring 2004.
- [4] G. J. Sullivan and R. L. Baker, "Efficient Quadtree Coding of Images and Video", IEEE Transactions on Image Processing, vol. 3, issue 3, pp. 327-331, May 1994.
- [5] V. Seferidis and M. Ghanbari, "Generalised Block-Matching Motion Estimation using Quad-Tree Structured Spatial Decomposition", IEE Proceedings- Vision, Image and Signal Processing, vol. 141, issue 6, pp. 446-452, 1994.
- [6] J. Lee, "Optimal quadtree for variable block size motion estimation", IEEE International Conference on Image Processing, Oct. 1995, vol. 3, pp. 480-483.
- [7] G. M. Schuster and A. K. Katsaggelos, "An Optimal Quadtree Based Motion Estimation and Motion-Compensated Interpolation Scheme for Video Compression", IEEE Transactions on Image Processing, vol. 7, issue 11, pp. 1505-1523, Nov. 1998.
- [8] V. Argyriou and T. Vlachos, "Quad-Tree Motion Estimation in the Frequency Domain Using Gradient Correlation", IEEE Transactions on Multimedia, vol 9, issue 6, pp. 1147-1154, Oct. 2007.
- [9] C. Kasai, K. Namekawa, A. Koyana and R. Omoto "Real-Time Two-Dimensional Blood Flow Imaging Using an Autocorrelation Technique", IEEE Transaction on Sonics and Ultrasonics, vol. 32, issue 3, pp. 458-464, May 1985.
- [10] S. Langeland, J. D'hooge, H. Torp, B. Bijnens, and P. Suetens, "Comparison of Time-Domain Displacement Estimators for Two-Dimensional RF Tracking", Ultrasound in Medicine and Biology, vol. 29, no. 8, pp. 1177–1186, 2003.
- [11] F. Viola and W. F. Walker, "A spline-based algorithm for continuous time-delay estimation using sampled data", IEEE Transactions on Ultrasonics. Ferroelectrics. Frequency Control, vol. 52, no. 1, pp. 80–93, 2005.
- [12] J. Luo and E. E. Konofagou, "A Fast Normalized Cross-Correlation Calculation Method for Motion Estimation", IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, vol. 57, issue 6, pp. 1347-1357, June 2010.
- [13] A. J. H. Hii, C. E. Hann, J. G. Chase, and E. E. W. Van Houten, "Fast Normalized Cross Correlation for Motion Tracking Using Basis Functions", Journal of Computer Methods and Programs in Biomedicine, vol. 82, no. 2, pp. 144-156, 2006.
- [14] P. Anandan, "A Computational Framework and an Algorithm for the Measurement of Visual Motion", International Journal of Computer Vision, vol. 2, pp. 283-310, 1989.
- [15] S. D. Wei, W. H. Pan and S. H. Lai, "A novel motion estimation method based on normalized cross correlation for video compression", Proceedings-14th International Multimedia Modeling Conference, MMM 2008, Kyoto, Japan, Jan. 2008, pp. 338-347.

Unsupervised Categorization of Objects into Artificial and Natural Superordinate Classes Using Features from Low-Level Vision

Zahra Sadeghi

zahra.sadeghi@ut.ac.ir

*Cognitive Robotics Lab
School of Electrical and Computer Engineering
College of Engineering, University of Tehran, Iran
School of Cognitive Sciences
Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

Majid Nili Ahmadabadi

mnili@ut.ac.ir

*Cognitive Robotics Lab
School of Electrical and Computer Engineering
College of Engineering, University of Tehran, Iran
School of Cognitive Sciences
Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

Babak Nadjar Araabi

araabi@ut.ac.ir

*Cognitive Robotics Lab
School of Electrical and Computer Engineering
College of Engineering, University of Tehran, Iran
School of Cognitive Sciences
Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*

Abstract

Object recognition problem has mainly focused on classification of specific object classes and not much work is devoted to the problem of automatic recognition of general object classes. The aim of this paper is to distinguish between the highest levels of conceptual object classes (i.e. artificial vs. natural objects) by defining features extracted from energy of low level visual characteristics of color, orientation and frequency. We have examined two modes of global and local feature extraction. In local strategy, only features from a limited number of random small windows are extracted, while in global strategy, features are taken from the whole image.

Unlike many other object recognition approaches, we used unsupervised learning technique for distinguishing between two classes of artificial and natural objects based on experimental results which show that distinction of visual object super-classes is not based on long term memory. Therein, a clustering task is performed to divide the feature space into two parts without supervision. Comparison of clustering results using different sets of defined low level visual features show that frequency features obtained by applying Fourier transfer could provide the highest distinction between artificial and natural objects.

Keywords: Objects' Super-class Categorization, Low Level Visual Features, Categorization of Objects to Artificial and Natural, Local and Global Features, Color, Orientation, Frequency.

1. INTRODUCTION

Object recognition is a prominent problem in many fields of study such as computer vision, robotics, and cognitive sciences and has been studied for four decades [1]. The ultimate goal of this problem is to find a proper visual representation to identify each object effectively. However,

the emphasis has been mainly laid on classifying very specific groups of objects in which the members of each class have many similarities in shapes and textures. In contrast, not much work has been devoted to the problem of recognizing objects of more general classes due to vagueness in identifying the common properties for all the members of a high level conceptual group.

Moreover, object recognition problem has been mostly studied as a similarity measurement problem in a supervised environment and so it needs many labeled training examples from some predefined and known classes before making prediction about the label of unknown and unseen examples. This process is tedious and time-consuming and relies heavily on the goodness of the selected training examples. Also, as stated in [2] there is no comparison between the different methods of object classification and each paper took an ad hoc approach on a special dataset.

In this paper we intend to categorize members of classes in the highest generality level. We assume that artificial/natural objects are positioned in the highest level of abstraction and hence investigate the problem of finding an efficient representation for each object to show this difference. Although the problem of indoor vs. outdoor scene or city versus landscape image classification is studied by many authors [3],[4],[5],[6],[7] little attention has been given to the problem of artificial vs. natural object distinction which is the subject of this paper.

2. RELATED WORK

The topic of artificial/natural discrimination is studied in both computer vision and neuroscience papers. In [8] a method for identification of artificial objects in natural scenes based on applying Zipf's law is proposed. Their idea is based on the observation that man-made objects are much simpler in their texture and generally contain much more uniform and flatter regions compared to natural ones. The idea of uniformity is also used by Caron et. al [9]. They measured the amount of uniformity by computing gradient or derivative of images. Fractal patterns are another approach applied to measure the amount of uniformity in images [10],[11]. In a specific problem, artificial and natural fruits are tried to be classified using color and texture features [12].

In a separate study the energy of Gabor orientation filter maps for Natural/man-made object classification is used as feature representatives [13]. They showed that Gabor orientation energies of man-made objects have more variations than natural objects and their associated Gabor diagrams have sharper points. KNN is used to classify feature vectors obtained from computing orientation energy of each image. Line and edge properties are also used to categorize images into natural and artificial classes [14].

Yet in another attempt, a model based approach is used to detect artificial objects in which basic shapes are applied as templates to represent for artificial shapes [15].

In addition, there are many significant studies on the animate/inanimate recognition in primate's brain in the field of neuroscience. A detailed review on two different theories related to animate/inanimate object categorization is studied in [16]. It is shown that the representation of animate and inanimate objects in brain is separated both functionally and anatomically by recording the BOLD activity in visual cortex [17]. In [18] an experience is conducted to evaluate the effect of different conceptual levels of animate and inanimate objects based on the response time. Bell et al. studied the relation between three different hypotheses about the organization of IT cortex for object representation based on animacy, semantic category, and visual features using fMRI [19].

3. SUPERORDINATE OBJECT CLASSES

Our purpose is to make the most general distinction between object categories. While in some studies, animate/inanimate categories are assumed to be at the highest level of abstraction [20],[21], in our view, artificial/natural categories encompass more general visual characteristics and are located in the highest level of inclusiveness. In animate/inanimate categorization the

discerning characterization is life. However, in artificial/ natural classification, the question is whether the object is a man-made product or not. Our proposed object categorization diagram is depicted in Figure 1 in which objects are first divided into artificial and natural classes. The natural objects can further be subdivided into animate and inanimate objects. However, artificial group include only inanimate objects. For example, shells and rocks belong to inanimate natural objects.

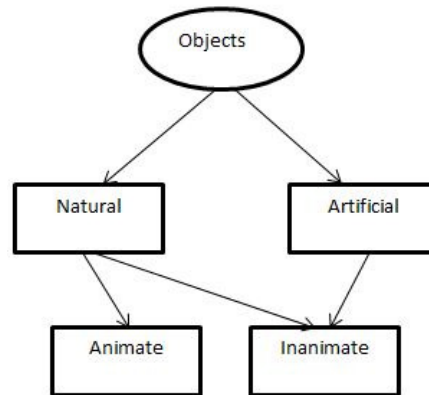


FIGURE 1: The Highest Levels of Object Categories.

3.1 Rationale Behind Unsupervised Learning of Superordinate Object Classes

Object recognition is mostly considered as a supervised problem in which the learning system needs to see many training samples to determine the label of test data with respect to the stored class representatives (eg. [22],[23],[24],[25].) One problem with supervised learning methods is that they are devoid of the power of generalization and they are dependent on the pre-computed features which are selected subjectively by their designers.

In this paper, we aim to perform the categorization of artificial/natural entities without the intervention of long term memory. In other words, instead of object classification, we try to group or cluster image data according to their similarities. We conjecture that the visual differences between these two categories can be found without any recall from memory, i.e., the difference between man-made and natural objects is visually observable and it doesn't depend on prior information. This hypothesis is supported with studies that have shown infants under 20-month year old can distinguish between superordinate level classes but cannot discriminate between basic level categories [26],[27],[28],[29]. Other studies have shown that children at the earliest age of living can categorize more inclusive classes much better than other levels [30],[31],[32],[33]. Recently, Rostad et al. have shown that the starting age at which children are capable of animate/inanimate categorization is found to be around 14 month year old [34]. These findings encourage the idea that the task of superordinate class discrimination is not relied on prior knowledge. In this direction, we first define feature vectors for each image and then perform a clustering task using k-means algorithm in order to divide feature space into separate regions.

4. DESCRIPTION OF PROPOSED FEATURES FROM LOW LEVEL VISION

In order to distinguish between artificial and natural classes we need to find general visual characteristic of objects. As Rosch and Lloyd pointed out superordinate categories are low in category resemblance [35], i.e., the superordinate level contains a very wide range of basic and subordinate object categories with different appearances and shapes, and so finding an effective representation with a high capability of discrimination is a very challenging task because it is not visually clear what is the common set of features which is shared among all the members of a superordinate category.

In the present study, we intend to show the power of several complex features derived from basic visual features, i.e., color, orientation, and frequency feature sets.

Note that we considered low level attributes of visual information i.e., orientation, frequency and color as basic features and the resulted features after particular computation on them as complex features. This is similar to Huble and Wisel Findings [36] that there exist simple and complex sets of neurons in visual system for processing visual information. In addition, according to psychologists, there are special visual receptors that respond to particular visual features [37]. We now explain how these features are computed.

For extracting orientation features we used Gabor wavelet functions:

$$g(x, y) = \frac{1}{\pi(\sigma s)^2} e^{-\frac{x^2+y^2}{2\sigma^2}} (\cos(x) + i \sin(x)) = G \cos(x) + G \sin(x) \quad (1)$$

$$x = (x_0 \cos(\alpha) + y_0 \sin(\alpha)) / s$$

$$y = (y_0 \cos(\alpha) - x_0 \sin(\alpha)) / s$$

$$-(2s + .5) \leq x_0 \leq 2s + .5,$$

$$-(2s + .5) \leq y_0 \leq 2s + .5$$

$$s \in \{.5, 1, 1.5, 2\}$$

$$\alpha \in \{0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6\}$$

We then removed the DC component from the cosine part.

To obtain complex features of orientation, we took the same approach as described in [13] by computing the sum of absolute values of pixel of images convolved with a bank of 24 Gabor or log Gabor filters.

Frequency features are computed using Fourier Transform functions of input images and then the sum of squared absolute values are computed on both phase and amplitude. We noticed that the shape of magnitude and phase of images in frequency domain is different for artificial and natural groups. The effect of using phase and amplitude spectrum of Fourier transform in man-made and natural scene images is discussed in [38].

The entropy of each input image is computed as well to measure the amount of randomness in object images. For this, we have computed the entropy of both RGB and gray input images.

We also defined four more complex features based on statistical analysis of histogram of edge and color attributes motivated by the fact that in general, artificial objects are much simpler in their texture properties. These features can be grouped into two main attributes, namely diversity and variability characteristics which represent two basic characteristics of artificial and natural images. Diversity feature demonstrates the number of distinct orientations and colors in each input image.

For computing the diversity of orientations, we convolved the input image with Gabor filters at 360 different orientations. We then applied a max pooling operator by selecting the max orientation that causes the highest output in convolution, i.e., in each pixel we look for an orientation corresponding to the maximum magnitude of the result of convolution. For Gabor filter computation we followed the same approach proposed by Riesenhuber and Poggio [39] and its parameter values are represented in equations (2-6) and Table 1 correspondingly.

$$g(x, y) = \exp\{-(x^2 + G^2 \cdot y^2)/(2\sigma^2)\} \cos(2\pi x/\lambda) \quad (2)$$

$$x = x_0 \cos(\theta) - y_0 \sin(\theta + \rho) \quad (3)$$

$$y = x_0 \sin(\theta) + y_0 \cos(\theta + \rho) \quad (4)$$

$$-2 \leq x_0 \leq 2, -2 \leq y_0 \leq 2 \quad (5)$$

$$\lambda = (RF_size)^2 / div \quad (6)$$

$$\sigma = \lambda * .8 \quad (7)$$

G	θ	ρ	RF_size	div
.3	0,1,2,...,359	0	5	4

TABLE 1: Gabor Filter Parameter's Value.

For diversity of color, RGB space is transformed into HSV color space and then a 360 bin histogram of hue values is computed. While RGB space is so sensitive to intensity variation, in HSV color space, intensity is separated from other components. It has also been found that this space is more likely to human visual system [40]. Having computed the histogram of edge and color, we finally count the number of different orientations which ranges from 0 to 359 and the number of hue colors which differs from 1 to 360.

In contrast to diversity, the variability attribute checks the number of peaks in each histogram which represents the number of dominant orientations and colors in each input image. This property shows the amount of change and variation and measures the coarseness of input image, i.e. how often we have a noticeable alteration in the orientations and colors. We implemented this feature by counting the number of peaks in the histogram of orientations and colors of input images. In a nutshell, having computed the histogram of orientation/color, the defined features can be explained as follows:

Diversity= the number of filled bins of histogram

Variability=the number of peaks of histogram

Generally, in comparison to artificial objects, natural objects group are characterized by higher values in Gabor energy, and much more amount of values in entropy, variability and diversity of pixels. In other words, large regions of artificial objects are of the same color and orientation, but in contrast, it is unlikely to find a region in natural objects with exactly identical color and orientation. However, sometimes, the opposite property can also be observed in some artificial and natural objects. For instance, artificial objects with complicated patterns in texture tend to have the attribute of naturalness due to its high variation in color and orientation.

5. LOCAL PROCESSING AND RANDOM EYE MOVEMENT

As was mentioned earlier, we have applied two different modes for feature extraction which are local and global strategies. In the local computation, complex features are extracted from random patches and are averaged and then the overall decision is made with respect to the average normalized feature values obtained from randomly selected regions from each object. However, in the global strategy, the defined features are extracted from the whole image. Thus based on the global or local method, we select the whole image or patches of image as the input data. The whole procedures of these strategies are described in Algorithm 1 and 2 respectively.

```
for im = 1 to size(ImageDataset) do
  InputImage = ImageDataset(im)
  BF = BasicFeatures(InputImage)
  CF(im; :) = ComplexFeatures(BF)
end for
Cluster(CF)
```

ALGORITHM 1: Global strategy

```
for im = 1 to size(ImageDataset) do
  image = ImageDataset(im)
  for i = 1 to exploringWindowsNum do
    InputImage = randomPatch(image)
    BF(i; :) = BasicFeatures(InputImage)
    CFi(i; :) = ComplexFeatures(BF; inputImage)
  end for
  CF(im; :) = Average(CFi)
end for
Cluster(CF)
```

ALGORITHM 2: Local strategy

Our local approach is based on random area selection, i.e. we explore different regions of objects randomly based on the assumption that the visual system has no prior knowledge for discerning between artificial and natural objects. In other words, we hypothesize that for a neutral subjective viewer with no previous knowledge, fovea attention wanders around the central middle point of an image which we call it gaze point.

Moreover, we didn't take the saliency detection approaches for selecting the local windows, because the salient regions (which are basically defined as regions with great amount of edges and colors) are not the only source of information to help us make a decision about the objects' type as being artificial or natural. Rather, the steady and monotonous regions are informative as well, and therefore, all areas of objects are important in order to decide the artificial/natural group of each object.

Computing the features locally have this advantage that the extracted features can represent small variation in features' differences much more accurately. In addition, since each image includes only one object, a considerable amount of each image contains plain background and it can reduce the accuracy of performance in the global strategy. Note that in local hue computation, the minimum and maximum of each block are computed for each local patch instead of the whole image. And in orientation and frequency computation, the Gabor filters and Fourier transform are applied only on a single patch. Figure 2 compares the local and global approach in hue computation for two sample images.

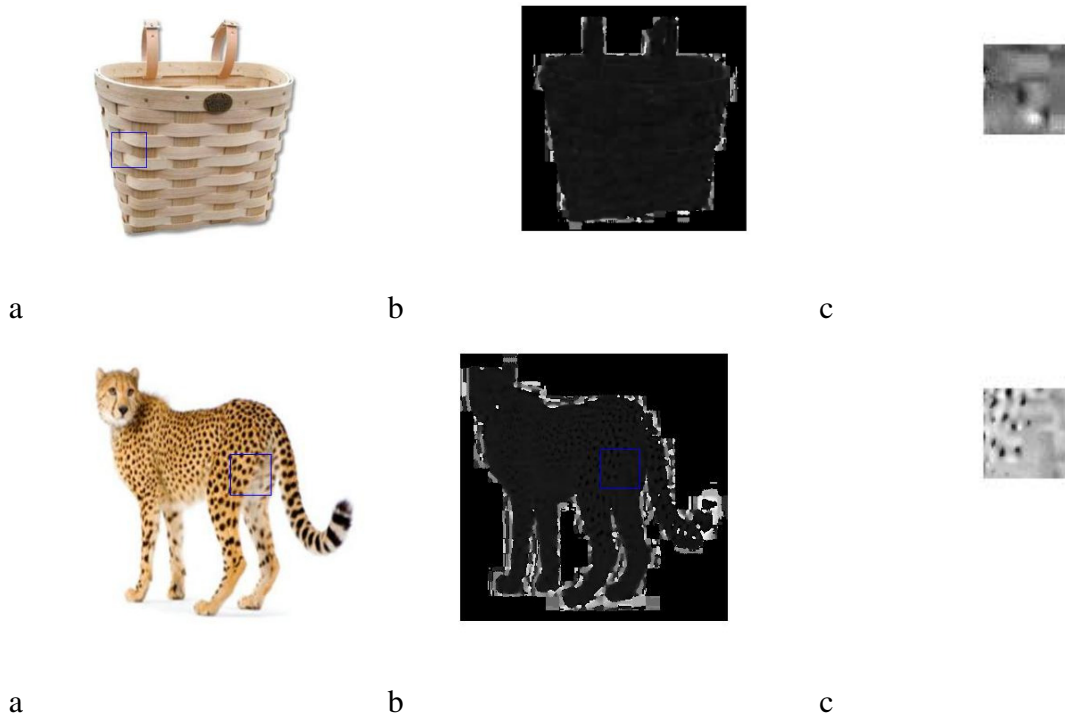


FIGURE 2: Global vs. Local Hue Computation
a: original image with a random selected patch,
b: Hue component computed globally for the whole image,
c: Hue component computed locally for the selected patch

6. EXPERIMENTAL RESULTS

One of the major issues in object recognition evaluation is the lack of proper dataset [41]. Most of the available datasets are composed of very limited range of classes. For example, UIUC database [42] contains only car images, CORE database [43] contains only vehicle and animals, and Pascal VOC challenges [44] are made of motorbikes, bicycles, cars, and people classes. Other datasets which cover a wider range of classes (ex. SUN dataset [45]) are specified to scene understanding purposes.

For the task of artificial/natural object discrimination we decided to use the images from two available datasets, i.e. Caltech-101 [46] and Coil-100 [47] object collections. Our dataset contains two main categories of artificial objects (selected from COIL-100) and natural objects (selected from Caltech-101 object libraries). The selected subclasses for artificial and natural groups are listed in Table 2.

In addition, we created another dataset by selecting 454 images from Hemera object collection database [48] and divided them into two groups of artificial and natural objects. In addition, as a preprocessing all images are converted to Jpeg format.

Natural Objects			
Ant	Dalmatian	Leopard	Rhino
Bass	dolphin	kangaroo	Rooster
Beaver	dragonfly	llama	scorpion
Bonsai	elephant	Lobster	sea-horse
brontosaurus	emu	Lotus	starfish
butterfly	flamingo	nautilus	stegosaurus
cougar-body	flamingo-head	Octopus	strawberry
cougar-face	Gerenuk	Okapi	sunflower
Crab	hawksbill	Panda	Tick
crayfish	hedgehog	Pigeon	water-lilly
crocodile	lbis	Pizza	wild-cat
crocodile-head	joshua-tree	platypus	
Artificial Objects			
Obj1,obj3			
Obj5 to obj62			
obj76 to obj82			
Obj64 to obj74			

TABLE 2: Artificial and Natural object classes selected from Caltech and Coil datasets.

As was mentioned before, we have defined complex feature vectors derived from different basic features of frequency, orientation and color.

The frequency features is a 3-dimensional feature vector obtained from:

$$FI = F(input_image) \quad (8)$$

$$magnitude(x, y) = \sqrt{\text{Re}(FI(x, y))^2 + \text{Im}(FI(x, y))^2} \quad (9)$$

$$phase(x, y) = \tan^{-1}(\text{Im}(FI(x, y)) / \text{Re}(FI(x, y))) \quad (10)$$

$$FreqFeat(1) = \sum_{x, y \in input_image} |magnitude(x, y)| \quad (11)$$

$$FreqFeat(2) = \sum_{x, y \in input_image} \log(1 + magnitude(x, y)) \quad (12)$$

$$FreqFeat(3) = \sum_{x, y \in input_image} |phase(x, y)| \quad (13)$$

Where FI is the result of Fourier Transform of input images.

For orientation feature, 24 dimensional feature vectors are obtained from sum of the absolute energy of convolution of images with Gabor filters which are computed for scale values of .5 to 2 with step sizes of .5 and orientations from 0 to 150 with step size of 30:

$$GI(x, y, s, \alpha) = G(input_image, s, \alpha) \quad (14)$$

$$gaborFeat(s, \alpha) = \sum_{x, y \in input_image} |GI(x, y, s, \alpha)| \quad (15)$$

$$s \in \{.5, 1, 1.5, 2\}$$

$$\alpha \in \{0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6\}$$

Where GI is obtained by convolving the input image with Gabor filters with a specific scale s and orientation α .

The entropy feature vector is a 2-dimensional feature vector including entropy of orientation and color which is obtained by computing the entropy of both RGB and gray input images using the following equation:

$$EntropyFeat(1) = entropy(Input_image_RGB) \quad (16)$$

$$EntropyFeat(2) = entropy(Input_image_gray) \quad (17)$$

$$entropy = \sum -H(I).log(H(I)) \quad (18)$$

Where $H(I)$ stands for 256-bin histogram counts of the input image I .

The histogram of orientation and color feature vectors are a two dimensional vector composed of diversity and variability attributes which were explained in section 4 and can be obtained by:

$$orientHFeat(1) = diversity(GrayInputImage) \quad (19)$$

$$orientHFeat(2) = variability(GrayInputImage) \quad (20)$$

$$colorHFeat(1) = diversity(RGBInputImage) \quad (21)$$

$$colorHFeat(2) = variability(RGBInputImage) \quad (22)$$

All the local strategy results are averaged for 10 independent runs of the whole procedure and the number of selected patches in each run is selected as 20 and 30 for the first and second datasets respectively. We used more patches for the second dataset (Hemera objects) because the size of images is larger than the images of the first dataset (Coil-Caltech). Note that, all the images of the first dataset are resized to 128*128 pixels, but the size of images of the second dataset is more than 200 pixels in each dimension.

As was mentioned before, for grouping the images, K-means clustering technique is used with value of $k=2$. To evaluate the quality of the generated clusters we used R-index, F-measure, precision, and recall of the obtained clusters [49] which are defined by:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (23)$$

$$P = \frac{TP}{TP + FP} \quad (24)$$

$$R = \frac{TP}{TP + FN} \quad (25)$$

$$F = \frac{(b^2 + 1).P.R}{b^2.P + R}, b = 2 \quad (26)$$

Where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative respectively.

The performance results of clustering with the local and global methods for both datasets are listed in Tables 3 to 6. All the results are rounded to two decimal points. Each Table shows the evaluation for the corresponding feature dimensions (explained in equations (8-22)). In bold are represented the best results obtained from each feature set.

It can be inferred from the results that frequency features showed dominance in making distinction between artificial and natural images. While local strategy is applied on a sequence of image patches instead of the whole image, it has generated superior or equal results in comparison to global strategy due to high amount of similarity and cohesion between pixels of each patch. Note that we only have considered the patches which are located mostly on foreground (i.e. the objects) and the patches that fall on background are automatically removed. It may be concluded that in artificial/natural distinction, texture and local information plays more important role than shape and global information.

Feature name	Feature dimension	RI	F	P	R
FreqFeat	1	0.89	0.93	0.88	0.94
FreqFeat	2	0.88	0.88	0.93	0.86
FreqFeat	3	0.51	0.60	0.57	0.61
FreqFeat	1:3	0.90	0.89	0.94	0.89
GaborFeat	1:24	0.82	0.90	0.79	0.93
EntropyFeat	1	0.61	0.62	0.75	0.60
EntropyFeat	2	0.55	0.58	0.65	0.58
EntropyFeat	1:2	0.58	0.60	0.73	0.58
colorHFeat	1	0.56	0.57	0.64	0.56
colorHFeat	2	0.54	0.55	0.62	0.53
colorHFeat	1:2	0.54	0.55	0.62	0.54
orientHFeat	1	0.59	0.60	0.67	0.59
orientHFeat	2	0.50	0.52	0.58	0.51
orientHFeat	1:2	0.56	0.59	0.63	0.58

TABLE 3: Performance of Local methods for Caltech-Coil Dataset.

Feature name	Feature dimension	RI	F	P	R
FreqFeat	1	0.79	0.87	0.78	0.89
FreqFeat	2	0.87	0.88	0.90	0.88
FreqFeat	3	0.87	0.86	0.92	0.84
FreqFeat	1:3	0.90	0.90	0.94	0.89
GaborFeat	1:24	0.69	0.82	0.69	0.86
EntropyFeat	1	0.56	0.57	0.65	0.55
EntropyFeat	2	0.64	0.65	0.72	0.63
EntropyFeat	1:2	0.61	0.62	0.69	0.60
colorHFeat	1	0.53	0.57	0.60	0.56
colorHFeat	2	0.50	0.52	0.58	0.51
colorHFeat	1:2	0.52	0.55	0.59	0.54
orientHFeat	1	0.58	0.60	0.66	0.58
orientHFeat	2	0.64	0.64	0.72	0.62
orientHFeat	1:2	0.65	0.64	0.73	0.63

TABLE 4: Performance of global methods for Caltech-Coil dataset.

Feature name	Feature dimension	RI	F	P	R
FreqFeat	1	0.70	0.72	0.69	0.73
FreqFeat	2	0.71	0.71	0.72	0.71
FreqFeat	3	0.64	0.66	0.64	0.67
FreqFeat	1:3	0.72	0.72	0.73	0.72
GaborFeat	1:24	0.54	0.70	0.53	0.76
EntropyFeat	1	0.72	0.72	0.73	0.72
EntropyFeat	2	0.73	0.73	0.73	0.73
EntropyFeat	1:2	0.73	0.73	0.73	0.73
colorHFeat	1	0.70	0.70	0.69	0.70
colorHFeat	2	0.73	0.73	0.73	0.73
colorHFeat	1:2	0.73	0.73	0.73	0.73
orientHFeat	1	0.69	0.69	0.69	0.69
orientHFeat	2	0.63	0.64	0.63	0.64
orientHFeat	1:2	0.68	0.68	0.68	0.68

TABLE 5: Performance of Local methods for Hemera dataset.

Feature name	Feature dimension	RI	F	P	R
FreqFeat	1	0.82	0.75	0.83	0.82
FreqFeat	2	0.80	0.74	0.80	0.80
FreqFeat	3	0.78	0.74	0.78	0.78
FreqFeat	1:3	0.84	0.76	0.85	0.84
GaborFeat	1:24	0.56	0.60	0.56	0.61
EntropyFeat	1	0.51	0.56	0.51	0.58
EntropyFeat	2	0.51	0.56	0.51	0.58
EntropyFeat	1:2	0.51	0.56	0.51	0.58
colorHFeat	1	0.50	0.50	0.51	0.50
colorHFeat	2	0.51	0.51	0.51	0.51
colorHFeat	1:2	0.51	0.51	0.51	0.51
orientHFeat	1	0.56	0.73	0.55	0.79
orientHFeat	2	0.59	0.72	0.57	0.78
orientHFeat	1:2	0.56	0.73	0.55	0.79

TABLE 6: Performance of Global methods for Hemera dataset.

7. CONCLUSION & DISCUSSION

One possible approach for solving object categorization problem is a top-down view. In this direction, different levels of categories need to be recognized subsequently in which the inclusiveness of recognition levels decreases in a descending order. In this paper, the first conceptual level of abstraction is associated with artificial and natural categories. Based on neuroscientific views, artificial and natural objects are represented differently in brain. However, the processing and encoding of visual features is under debate. Experimental studies on children can support the theory that human may distinguish between these two categories without referring to their long term memory and hence our feature definition mechanism is an unsupervised learning algorithm which doesn't use pre-learned parameter sets for dividing the feature space into two general categories. However, automatic unsupervised artificial/natural grouping is a complicated task. First, the artificial/natural category is located in the highest level of abstraction. Thus, finding appropriate generic properties is not an easy task. Second, in contrast to classification problems in which there exists a set of labeled data that helps the categorization problem, in clustering problem, there is no access to any prior information in advance. Taking into account objects' characteristics we derived different high level features from basic low level features which can make distinction between artificial and natural categories of objects. We compared the discriminating effect of different features obtained by using Fourier transform, Gabor filter, entropy, and histogram of color and orientation for artificial/natural object distinction. Feature extraction is applied by using two different strategies of local and global processing and then a clustering task is performed to group the similar features with regard to Euclidean distance between them. The obtained simulation results showed that frequency features derived from Fourier Transform achieved the first highest efficiency tier in distinguishing between artificial and natural objects. Also, local strategy which is based on random patch selection corresponding to random eye movement resulted in comparable performance in term of accuracy and with regard to the lower amount of information processing.

8. REFERENCES

- [1] D. Marr. (1982). Vision. A computational investigation into the human representation and processing of visual information. New York. W.H. Freeman.
- [2] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, W. Buntine. "Unsupervised Object Discovery: A Comparison."

- [3] M. Szummer, R. W. Picard. (1998). "Indoor-outdoor image classification, Proceeding of International Workshop on Content-Based Access of Image and Video Databases, Bombay".
- [4] E.C. Yiu, (1996). "Image classification using color cues and texture orientation."
- [5] M.J. Kane, A.E. Savakis. (2004). "Bayesian Network Structure Learning and Inference in Indoor vs. Outdoor Image Classification."
- [6] N. Serrano, A.E. Savakis, L. Luo. (2002). "A Computationally Efficient Approach to Indoor/Outdoor Scene Classification."
- [7] M. Szummer, R.W. Picard. (1998). "Indoor-Outdoor Image Classification."
- [8] Y. CARON, P. MAKRIS, N. VINCENT. (2002). "A method for detecting artificial objects in natural environments." pp.600-603.
- [9] Y. Caron, P. Makris, N. Vincent. (2002). "A method for detecting artificial objects in natural environments." IPCR, pp. 600 603.
- [10] D. Chenoweth, B. Cooper, J. Selvage. (1995). "Aerial Image Analysis Using Fractal Based Models." IEEE Aerospace Applications Conference Proceedings, pp. 277-285.
- [11] G. Cao, X. Yang, and Z. Mao. (2005). "A two-stage level set evolution scheme for man-made objects detection in aerial images." Proc. IEEE Conf. Comput. Vis. Pattern Recog., San Diego, CA, pp. 474-479.
- [12] B. S., Anami, V. Burkapalli, V. Handur, H.K.: Bhargav. "Discrimination of Artificial Objects from Natural Objects."
- [13] M., Kim, C., Park K. Koo. "Natural / Man-Made Object Classification Based on Gabor Characteristics, Image and Video Retrieval, LNCS 3568, pp. 550-559
- [14] J. Tajima, H. Kono. (2008). "Natural Object/Artifact Image Classification Based on Line Features." IEICE Transactions. 91-D(8), pp. 2207-2211.
- [15] Z. Wang, J. Ben Arie. (1999). "Generic Object Detection using Model Based Segmentation." IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
- [16] A. Caramazza, J. R. Shelton. (1998). "Domain-specific knowledge systems in the brain - the animate-inanimate distinction." Neuroscience, Vol. 10, pp. 1-34.
- [17] T. Naselaris, D.E. Stansbury, J. L. Gallant. "Cortical representation of animate and inanimate objects in complex natural scenes."
- [18] A.J., Wiggett, I.C., Pritchard P.E. Downing. (2009). "Animate and inanimate objects in human visual cortex: Evidence for task-independent category."
- [19] A.H., Bell, F. Hadj-Bouziane, J.B., Frihauf, R.B., Tootell, L.G. Ungerleider. (2009). Object representations in the temporal cortex of monkeys and humans as revealed by functional magnetic resonance imaging. J Neurophysiol 101, pp. 688-700.
- [20] D. Poulin-Dubois, S. Graham, L. Sippola. (1995). Early lexical development: The contribution of parental labeling and infants categorization abilities. Journal of Child Language: 22, pp. 325-343.

- [21] D. H. Rakison. Parts, motion, and the development of the animate inanimate distinction in infancy. In D. H. Rakison, L. M. Oakes (Eds.) (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*, pp.159-192.
- [22] P.F. Felzenszwalb, D.P. Huttenlocher. (2000). "Efficient matching of pictorial structures." *IEEE Conference on Computer Vision and Pattern Recognition*, pp.66-73.
- [23] B. Heisele. (2003). "Visual Object Recognition with Supervised Learning." *IEEE Intelligent Systems - AI's Second Century*, pp. 38-42.
- [24] D.J., Crandall, P.F., Felzenszwalb, D.P. Huttenlocher, (2005). "Spatial priors for part-based recognition using statistical models". In *IEEE Conference on Computer Vision and Pattern Recognition*, pp.10-17.
- [25] T. Deselaers, G. Heigold, H. Ney. "Object Classification by Fusing SVMs and Gaussian Mixtures."
- [26] J. M. Mandler, P. J. Bauer. (1988). "The cradle of categorization: Is the basic level basic?" *Cognitive Development*, 3(3), pp. 247-264.
- [27] J. M. Mandler, P. J. Bauer, L. McDonough. (1991). "Separating the sheep from the goats: Differentiating global categories." *Cognitive Psychology*, 23, pp.263-298.
- [28] M. H. Bornstein, M. E. Arterberry. (1991). "The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute and group versus individual analyses." *Developmental Psychology*: 46, pp. 350-365.
- [29] J. M. Mandler, L. McDonough. (1993). "Concept formation in infancy. *Cognitive Development*." 8, pp. 291-318.
- [30] B.A., Younger, D.D. (2000). "Fearing, A global-to-basic trend in early categorization: Evidence from a dual-category habituation task." *Infancy*. 1, pp. 47-58.
- [31] P.C. Quinn, M.H. Johnson, D. Mareschal, D.H. Rakison, B.A. Younger. (2000.). "Understanding early categorization: One process or two?." *Infancy*. 1, pp.111-122.
- [32] S. Pauen. (2002). "The global-to-basic level shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioral Development*." 26, pp. 492-499.
- [33] S. Pauen. (2002). "Evidence for knowledge-based category discrimination in infancy." *Child Development*. 73, pp. 1016-1033.
- [34] K. Rostad, J. Yott, D. Poulin-Dubois. (2012). "Development of categorization in infancy: Advancing forward to the animate/inanimate level." *Infant Behav Dev*.
- [35] E. Rosch, B.B. Lloyd. (1978). *Cognition and categorization*. Pp.27-48.
- [36] D.H. Hubel, T.N. Wiesel. (2005). "Brain and visual perception: the story of a 25-year collaboration." Oxford University.
- [37] S.M. Zeki. (1976). "The functional organization of projections from Striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harbor Symposia on Quantitative Biology*." 5, pp. 591-600.
- [38] A. Oliva, A. Torralba. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), pp. 145-175.

- [39] M. Riesenhuber, T. Poggio. (1999). "Hierarchical models of object recognition in cortex." *Nature Neuroscience*.
- [40] M. Khosrowpour. *Encyclopedia of Information Science and Technology*, 1-5
- [41] J. Ponce, T.L. Berg, M. Everingham, D.A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, , Russell B.C., A. Torralba, C.K.I, Williams, J. Zhang, A Zisserman. "Dataset Issues in Object Recognition."
- [42] Agarwal, S., Roth, D. (2002). "Learning a sparse representation for object detection. In: Proc. European Conf. Comp. Vision." LNCS 23-53., Copenhagen, Denmark. pp.113-127.
- [43] A. Farhadi, I. Endres, D. Hoiem. (2010). "Attribute-Centric Recognition for Cross-category Generalization."
- [44] R. Fergus, P. Perona, A. Zisserman. (2003.). Object class recognition by unsupervised scale-invariant learning. In: Proc. IEEE Conf. Comp. Vision Patt. Recog. 2, pp. 264-271.
- [45] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba. (2010). "SUN Database: Large-scale Scene Recognition from Abbey to Zoo." IEEE Conference on Computer Vision and Pattern Recognition.
- [46] L. Fei-Fei, R. Fergus, P. Perona. (2004). "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories." CVPR, Workshop on Generative-Model Based Vision.
- [47] S. A., Nene, S. K. Nayar, H. Murase. (1996). "Columbia Object Image Library (COIL-100)" Technical Report CUCS-006 96.
- [48] <http://www.hemera.com>
- [49] C.J., Van Rijsbergen. (1979). *Information retrieval*, Butterworths, London, second edition.

Cut-Out Animation Using Magnet Motion

Srinivas Anumasa

Computer Science

Shri Ramswaroop Memorial University

Lucknow, 225003,India

srinu.0922@gmail.com

Avinash Singh

Computer Science

Shri Ramswaroop Memorial University

Lucknow, 225003,India

singhavinash66@gmail.com

Rishi Yadav

Computer Science

Shri Ramswaroop Memorial University

Lucknow, 225003,India

rishiyadavjec@gmail.com

Abstract

In this paper a new video-based interface for creating cutout-style animation using magnets is presented. This idea makes room for users of all skill levels to animate. We created an interface which is generally a closed box with a camera placed at the bottom. Roof of the box acts like a multi touch interface. A cast of physical characters are designed by the animator using paper, markers and scissor. If animator wants to animate using this physical characters, he pastes them to magnets (we call them as PMagnets) and place them under the roof (puppets facing the camera). These characters are controlled by other magnets (CMagnets) on the top of the roof. Here we require, a simple foreground extraction algorithm to extract the characters and to render them onto a new background. Our system "Cut-Out Animation using Magnet Motion" runs in real time (i.e 30 Frames/Sec). Therefore animator and the audience can instantly see the animation.

Keywords: Cut-Out Style Animation, Background Replacement, Magnets.

1. INTRODUCTION

Animation is an illusion of movement created by rapidly displaying a sequence of images. Creating an animation sequence is a time consuming process. Hand drawn or stop motion animation is a laborious process in which every frame is manually drawn or crafted. For computer generated animation several tools like Maya, Flash, etc. are available in the market. Even though they provide a good interface for creating easy animation sequence one must be an expertise and requires training to gain control over the interface provided.

Puppetry animation is a real time story telling animation which is famous from past few centuries. Puppeteers use threads to control the motion of puppets and bring them to life. Puppeteers and their controls are often visible to the audience. Although the quality of Puppetry animation is not good as hand drawn animation, still it attracts kids because of its own way of storytelling.

Cut-out animation is one way of storytelling in which the characters are flat 2d characters designed using papers, scissors and colours. Traditional Cut-out animation is not a real time like Puppet shows, each frame is manipulated by adjusting the character cut-outs. It is a time consuming process.

In this paper we present a novel technique for cutout style animation which is mostly inspired by the work of Connelly Barnes, David E. Jacobs Video Puppetry: A Performative Interface for

Cutout Animation. Authors in paper [1] proposed a simple cut-out style animation in which the characters are controlled by hand of a puppeteer. These Puppets are tracked and rendered onto a new background in the computer. In our approach each character is attached to magnets (PMagnet(s)) and placed on one side of the screen which are controlled by magnets (CMagnet(s)) on the other side of the screen. Thus giving life to the Puppets by concealing the Puppeteer. Also, if the animator wants to render on to the new background no tracking and object detection algorithms are required, just a simple background subtraction algorithm works well.

The main contribution of our work is the introduction of a new interface for creating cut-out style animated stories by using magnets and simple background subtraction algorithms. Designing this interface is easy, cheap, neither required complicated algorithms nor specialized skills, also no formal training was required for animation. However the animation produced by this method cannot match the visual quality of a studio production, but they are useful in many scenarios, including kids' productions or the class of animations such as "South Park" or "JibJab" that tries to simulate a "cut-out" style animation.

2. RELATED WORK

2.1. Video Puppetry.

Our work is inspired by the work of Connelly Barnes, David E. Jacobs [1]. The system in [1] is divided into two modules Puppet builder and Puppet theatre. For the first module authors created an interface which allows users to add new puppets into the database. A user first draws a character on paper with markers, crayons, or other high-contrast media. Next the user cuts out the puppet and places it under the video camera. The puppet builder captures an image of the workspace and processes the image in several stages before adding the puppet to a puppet database. Scale - invariant feature transform SIFT [2] features of each character are detected and stored in a database which is used in the second module.

The second module is puppet theatre, generally it is an object tracking environment. The user moves the puppets for animation, an over head camera which is used to track the animated characters by removing the user's hands and renders the characters on a new background. The SIFT [2] features are used to identify all puppets every 7-10 frames. Between SIFT updates, optical flow on Kanade- Lucas- Tomasi KLT [3] features are used to track the movement of puppets in real time.

But, this system has some limitations in handling every puppet. For example every puppet must occupy some significant portion in every video frame. So that every character can be tracked robustly. This system can handle 5 to 6 characters at a time, because as the number of characters gets increased the system gets slow. It is difficult to handle occlusions which may interfere with the tracking part. Because in animation there are many situations where characters move over each other. Puppets cannot be moved quickly because the KLT algorithm assumes that the displacement of tracking objects in between frames is small.

2.2. Sketch-n-stretch.

Authors in [4] designed an interface which is specially designed to simulate cut-out style animation. This is a two (virtual) layered interface. The user first selects the background image on which the foreground needs to render, then he selects cut-out pen from the menu and draws the border of the cut-out and selects the pencil tool to draw the character or he selects an image of a character stored in computer memory and then he selects an empty cut-out. By rotating, scaling and translation the animator can animate the character. This is not a multi touch interface it is not possible to animate two or more characters at the same time, so a time line option is provided. Each character is individually animated and combined using the time line.

3. PRINCIPLE OF OUR APPROACH

Traditional cut-out animation is a time consuming process, because for each frame the animator has to adjust the position of each Puppet or a part of the Puppet. Authors in [1] came with an which are time expensive. These algorithms are required to remove hands of animator which are

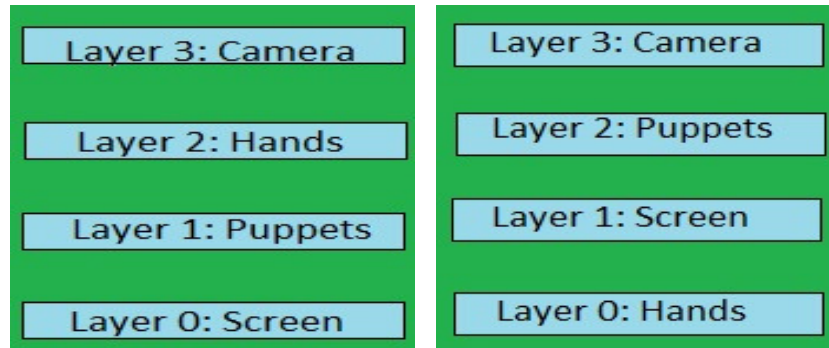


FIGURE 1: (a) (b)

in between the camera and Puppets. If we think this process as four layered animation, as shown in Figure 1(a) Layer 3 is camera, Layer 2 is hands (controller), Layer 1 Puppets, Layer 0 as screen. We tried to rearrange this layered system for simple cut-out style animation Figure 1(b) shows our layered system Layer 4 is camera, Layer 2 Puppets(With PMagnets), Layer 1 screen, Layer 0 controller (controlling the motion of Puppets (pasted with PMagnets)) with CMagnets.

4. PROPOSED WORK

Our system follows two steps. Puppet design and deciding number of P(C)Magnets required for his character is the first step in our system. Generating animation sequence is the second step. In this step animator control's the motion of characters using CMagnets. A background subtraction algorithm is required in this step.

4.1` Puppet Design

As shown in the Figure 2(a) animator can design the physical animation character for his animation sequence.



FIGURE 2: (a) (b)

When animator gets completed with their designing part, they have to decide what kind of motion does their character or part of the character will have. This motion decides number of P(C)Magnets required. For example, the character may have motion such that it does not include any rotation so it requires one P(C)Magnet. In Figure 2(a) the character will have some rotational motion. So, it requires two P (C) Magnets Figure 2(b) is the back side of the physical puppet attached with 2 PMagnets which are used to control the motion using other CMagnets. Between

PMagnets and CMagnets there will be a cardboard (roof of our interface) of small thickness, so that CMagnets can control PMagnets using magnetic field.

4.2 Animating

In this step animator generates the animation sequence using the Puppets. For this step we have

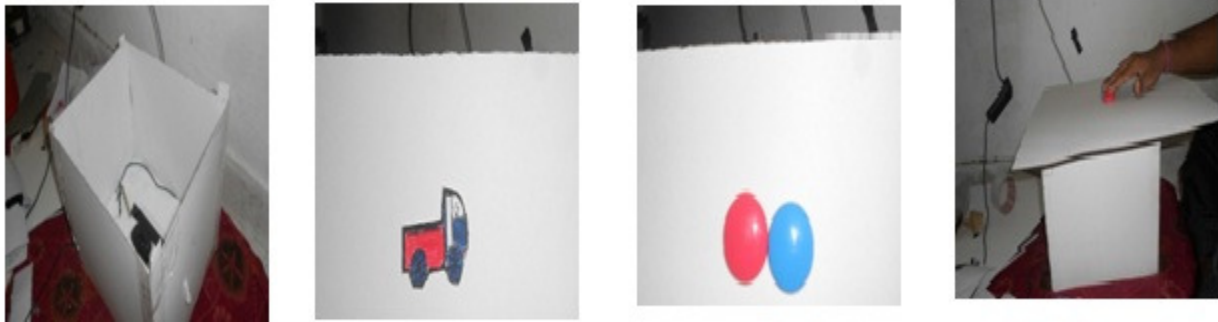


FIGURE 3: (a) (b) (c) (d)

designed a small box using a white cardboard sheet which is used as a puppet theatre. As shown in Figure 3 (a) this box creates a closed environment, and has its own light source providing uniform illumination so that the lighting cannot be affected by the outside environment. A camera is placed at bottom of the box, the roof which is a white cardboard sheet acts as an interface. The camera is projected on inside of the roof where it records the motion of physical characters. The other side of the roof acts as a medium for an animator for controlling PMagnets with CMagnets. Figure 3(b) a 4 wheeler puppet is placed on inside of the roof, Figure 3(c) is the other side of the roof with CMagnets (white board magnets). Figure 3(d), user controlling the motion of the puppet.

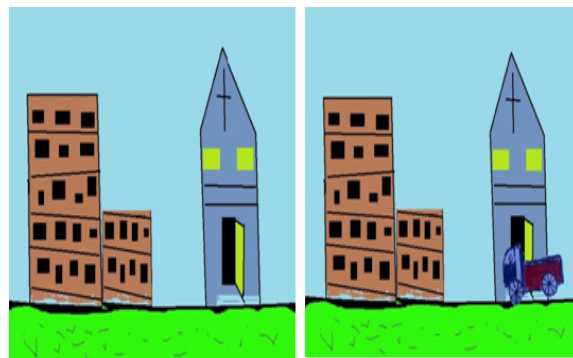


FIGURE 4: (a) (b)

If the animator wants to render onto a new background than the foreground must be extracted. There are methods proposed in [5],[6] for extracting foreground from a scene where background information is known and it is same for all frames. In our system the background is constant, so a simple background subtraction algorithm is sufficient. Figure 4(a) is the new background, Figure 4(b) is the snapshot of an animation sequence after rendering Figure. 2(a) onto Figure 4(a). Instead of rendering onto computer generated background, the animator can use his hand drawn background, so no background replacement is needed.

5. RESULTS

Figure 5 are the snapshots of an animation sequence using Figure 2(a) as puppet. We haven't done much animation, but we believe this interface provides a simple and cost effective way of producing simple animation sequences.



FIGURE 5

6. LIMITATIONS

Articulated animation is not possible with this simple approach Because there is no information about the relation between character parts which are controlled by the animator. We know only foreground and background. Another drawback is when two characters come closer or even overlap(when there is a situation like cross over) the magnets may attract or may ripple causing a distortion in the animation sequence

7. CONCLUSION AND FUTURE WORK

We presented a new interface which uses magnets (PMagnets and CMagnets) for controlling animation. Up to the best of our knowledge this is the first system which is designed by using the concept of Magnets. The advantage of using our system is that it can be designed in a cost effective manner. Our system have several advantages over existing system [1] like we are not using any object tracking or object detection algorithm which is computationally expensive. Computational cost never increases as the number of characters increases. There is no need to maintain a separate database. Partial visibility of a physical character is possible and simple. Even size of physical character (puppet) is not an issue in our approach. There is no limit on the speed at which a physical character is moved. In future we will try to work on the articulated animation of physical puppets. We will also try to add effects discussed in [1] in our system.

8. REFERENCES

- [1] Barnes, Connelly, David E. Jacobs, Jason Sanders, Dan B. Goldman, Szymon Rusinkiewicz, Adam Finkelstein, and Maneesh Agrawala. "Video puppetry: a performative interface for cut out animation." In ACM Transactions on Graphics (TOG), vol. 27, no. 5, p. 124. ACM, 2008.
- [2] Lowe, David G. "Object recognition from local scale-invariant features." In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, vol. 2, pp. 1150-1157. IEEE, 1999.
- [3] Sinha, Tomasi, Carlo, and Takeo Kanade. "Detection and tracking of point features". School of Computer Science, Carnegie Mellon Univ., 1991.
- [4] Sohn, Eisung, and Yoon-Chul Choy. "Sketch-n-Stretch: sketching animations using cutouts." Computer Graphics and Applications, IEEE 32.3 (2012): 59-69.
- [5] Qian, Richard J., and M. Ibrahim Sezan. "Video background replacement without a blue screen." In Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on, vol. 4, pp. 143-146. IEEE, 1999.
- [6] Horprasert, Thanarat, David Harwood, and Larry S. Davis. "A statistical approach for real-time robust background subtraction and shadow detection." In IEEE ICCV, vol. 99, pp. 1-19. 1999.

Faster Training Algorithms in Neural Network Based Approach For Handwritten Text Recognition

Haradhan Chel

*Dept. of Electronics and Communication
CIT Kokrajhar, Assam, India*

h.chel@cit.ac.in

Aurpan Majumder

*Dept. of Electronics and Communication
NIT Durgapur, West Bengal, India*

reach2am@gmail.com

Debashis Nandi

*Dept. of Information Technology,
NIT Durgapur, West Bengal, India*

debashisn2@gmail.com

Abstract

Handwritten text and character recognition is a challenging task compared to recognition of handwritten numeral and computer printed text due to its large variety in nature. As practical pattern recognition problems uses bulk data and there is a one step self sufficient deterministic theory to resolve recognition problems by calculating inverse of Hessian Matrix and multiplication the inverse matrix it with first order local gradient vector. But in practical cases when neural network is large the inversing operation of the Hessian Matrix is not manageable and another condition must be satisfied the Hessian Matrix must be positive definite which may not be satisfied. In these cases some repetitive recursive models are taken. In several research work in past decade it was experienced that Neural Network based approach provides most reliable performance in handwritten character and text recognition but recognition performance depends upon some important factors like no of training samples, reliable features and no of features per character, training time, variety of handwriting etc. Important features from different types of handwriting are collected and are fed to the neural network for training. It is true that more no of features increases test efficiency but it takes longer time to converge the error curve. To reduce this training time effectively proper train algorithm should be chosen so that the system provides best train and test efficiency in least possible time that is to provide the system fastest intelligence. We have used several second order conjugate gradient algorithms for training of neural network. We have found that *Scaled Conjugate Gradient Algorithm*, a second order training algorithm as the fastest for training of neural network for our application. Training using SCG takes minimum time with excellent test efficiency. A scanned handwritten text is taken as input and character level segmentation is done. Some important and reliable features from each character are extracted and used as input to a neural network for training. When the error level reaches into a satisfactory level (10^{-12}) weights are accepted for testing a test script. Finally a lexicon matching algorithm solves the minor misclassification problems.

Keywords: Transition Feature, Sliding Window Amplitude Feature, Contour Feature, Scaled Conjugate Gradient.

1. INTRODUCTION

As the computing technologies and high speed computing processors has been developed, pattern recognition field got a wider dimension. Recognition of handwritten text is a real challenging task and research on which started from early sixties. Along with the up gradation of computational techniques researchers always tried to realize human perception and to implement into mathematical logic and designed different perception to train the computer system.

Researchers took wide varieties approaches [7],[9], [20] to recognize handwritten text. In most of the research works some common logical steps were used for recognition of handwritten text such as Segmentation, Feature Extraction and pattern classification technique. Among different classification techniques popular approaches are Neural Classifier [6], [14], [23], Hidden Markov Model [9], Fuzzy Classifier, or hybridized technique like Neuro-fuzzy, Neuro-GA, or some other statistical methods [25]. Neural Classifier has high discriminative power [1],[11],[12] for different patterns. Handwritten text contains wide varieties of styles in nature. It is really difficult to get real character level segmentation. A lot of research works has been done on segmentation of hand written text in last two decades. Improper segmentation results inaccurate feature extraction and poor recognition performance. Similarly reliable and only discriminative features give better recognition performance. If number of features is increased recognition performance increases but it increases computational complexities and leads to much longer time of training of the neural network. The most important task is to choose a proper training algorithm which train faster the network with large no of features and provide best recognition performance. Many of these algorithms are based on Gradient Descent Algorithm such as Back Propagation Algorithm [15], [18]. But all the algorithms are practically inapplicable in large scale systems because of its slow nature and its performance also depends on some user dependent parameters like learning rate and momentum constant. Some second order training algorithm such as Fletcher Reeves [5], Polak Riebler[16],[19] or Powell-Beale Restarts algorithm[24] may be applied in appropriate applications. However *Scaled Conjugate Gradient* algorithm [22] results in much better recognition performance. By virtue of Scaled Conjugate Algorithm fastest training is obtained with almost 98 percent recognition efficiency. This paper not only shows the comparison of different second order training algorithm but also the reliable feature extraction and efficient lexicon matching technique. A small relevant description of Scaled Conjugate Algorithm is also given in later section. Different feature extraction schemes are also described in brief and finally the result of all experiments are shown.

2. IMAGE PREPARATION AND SEGMENTATION

A handwritten text written over A4 size paper is scanned through optical scanner and stored in bitmap image format. The image is then converted in to a binary image. One important point may be mentioned regarding image preparation is choosing the proper threshold of intensity level so that image does not get any discontinuity in any continuous part. Image is segmented [4],[8],[3] text level to word level and word level to character level segmentation all relevant information such as no of word, no of characters, no of lines are stored. Pixel coordinates of each words and characters in the bitmap image are stored very cautiously .The algorithm for segmentation used in image separation may be described below [20].

Algorithm 1:

Step1: The base lines such as upper, lower and middle base lines are calculated.

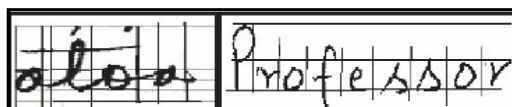
Step 2: Word is over segmented.

Step 3: The word segmentation points are modified using various rules.

Step 4: Undesired and incorrect segmentation points are removed.

Step 5: The correct segmentation points are used for final segmentation of words.

Step 6: Position of characters and words in the script are stored.



3. FEATURE EXTRACTION

It was mentioned in the introduction that reliable and meaningful features increase the recognition performance. Reliable feature are those features which produce almost same numeric value irrespective of slight positional variations in different sample images of same character. In this experiment four kinds of features extraction schemes are used and a total 100 nos. of features were collected per character. Before feature extraction all the segmented character is resized [17] in a fixed size (60 x40). One runtime separated unbounded character image, bounded image and resized image is shown in figure 1. Short description of all four kinds of features is described in the following subsections.



FIGURE 1: Unbounded Image, Bounded Image and Resized Image.

3.1 Transition Feature

Assume that the character image has a fixed dimension of X rows and Y columns. An M x N grid is superimposed over the image as shown in figure 2. The grid is prepared in such a way that all start and end values of each row and column got an integer value. The start values of row m and column n may be written as under.

$$x_1 = y_1 = 1 \quad (1)$$

$$x_m = \text{Int} \left[\frac{(m-1)X}{M-1} \right] \quad m = 2,3 \dots M \quad (2)$$

$$y_n = \text{Int} \left[\frac{(n-1)Y}{N-1} \right] \quad n = 2,3 \dots N \quad (3)$$

$\text{Int} [x]$ refers to nearest integer value to x. Assume the intensity of coordinate(x,y) of the X x Y image is A(x,y). The original image is scanned along every row and column and the gradient information along each row and columns are collected.

$$\Delta(x_m, y) = A(x_m, y+1) - A(x_m, y) \quad (4)$$

where $y = 1,2 \dots Y-1$

And

$$\Delta(x, y_n) = A(x+1, y_n) - A(x, y_n) \quad (5)$$

where $x = 1,2 \dots X-1$

As the image is a binary image both $\Delta(x_m, y)$ $\Delta(x, y_n)$ return one of the three values 0, 1 or -1. 0 indicate no transition 1 means white to black transition and -1 indicate transition from black to white. A 5x5 grid over character image is shown on figure 2.a and transitions are also shown in figure 2.b. In this case we would consider only -1 value and in each row and columns of the M x N grid total number of transitions in every row and columns are important features of the image. As it was mentioned earlier that the approach of the experiment is neural network based and the immediate operation after feature extraction is training, it is found experimentally that neural network works reliably and faster if the input feature values are within 0 to 1 limit. As in this case all values are more than 1, they are normalized between 0 and 1 by a nonlinear transfer function as mentioned below.

$$f(x) = \frac{1}{1 + e^{-kx}} \quad \dots \dots \dots 0 \leq x \leq 5 \quad (6)$$

k is a constant and values may be chosen between .3 to 1 for better result. The upper limit of x is taken as 5 because maximum possible white to black transition in handwritten characters found 5. In this way we found (M+N) nos. feature for each handwritten character. At the time of scanning the original image through all M rows and N columns using M x N grid and the coordinates of first and last transitions in each rows and columns are stored and a new kind of transition features were collected. Suppose in m^{th} row first and last transition occurs at (x_m, y_f) and (x_m, y_l) then two features can be collected from that row as mentioned below.

$$F_1 = \frac{y_f}{Y} \quad (7)$$

$$F_2 = \frac{Y-y_l}{Y} \quad (8)$$

Similarly if in n^{th} column of the M x N grid first and last transition occurs at (x_f, y_n) and (x_l, y_n) then two features can be collected from that column as mentioned below.

$$F_3 = \frac{x_f}{X} \quad (9)$$

$$F_4 = \frac{X-x_l}{X} \quad (10)$$



FIGURE 2.a: 5 x 5 Grid Over Original Image.

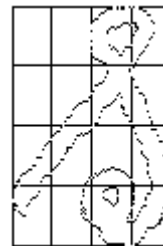


FIGURE 2.b: 5 x 5 Grid Over Grid Image.

3.2 Sliding Window Amplitude Feature

In this feature extraction scheme the image is subdivided into some overlapping windows such that half of the window breadth and width are overlapped with previous and next window along row and column wise except the windows at boundaries and corners. The windows at left boundary have no windows left at left side that it will overlap. Similarly top, bottom and right boundary side windows do not overlap with any windows beyond their boundary. The corner block has only two blocks to overlap for example the top left corner has only right and down windows to overlap. Black pixel density in each block or window is calculated. The original image of handwritten character has a fixed size of X rows and Y columns. As shown in figure (2.a) the image is superimposed with M x N grid. The four corners of the rectangular window R(m,n) are $[(X_m, Y_n) (X_m, Y_{n+2}) (X_{m+2}, Y_n) (X_{m+2}, Y_{n+2})]$. Values of x_m and y_n is defined in eqn. 1-3. Each block produces one feature and the feature is normalized between 0 and 1 and a total (M-2) (N-2) features are collected with normalized feature values [13]. In our experiment we have taken both M and N as 8 and a total 36 features were collected per character.

$$f = \frac{\text{no of black pixel in the block}}{\text{total no of pixel in the block}}$$

3.3 Contour Feature

The most important feature extraction scheme is contour feature [2][3]. Image is scanned and filtered. The filtered image is superimposed with an 11 X 1 grid as shown in figure (3.a) to find the external boundary intersection points of the character and total 22 such points are collected. All the 22 points are numbered in a circular fashion that from the top to bottom as 1 to 11 and from bottom to top as 12 to 22. All points are connected with the next point to draw an exterior boundary contour as shown in figure (3.b). For example the n^{th} line is bound between two coordinate (x_n, y_n) and (x_{n+1}, x_{n+1}) . The alignment of all the boundary lines is unique property of that image and it differs from character to character. The two unique parameters of a straight line that is the gradient and a constant value which solely depend on the coordinates of the two points are taken as feature normalizing by a nonlinear hyperbolic function. Equation of a straight line and its representation by two points may be mentioned by equation 12-14. Suppose a straight line in equation is bounded between two points (x_1, y_1) and (x_2, y_2) . The m is the gradient and c is a constant parameter of the straight line. Both m and c are two unique parameters which contain the boundary patterns of the characters.

$$y = mx + c \tag{12}$$

$$m = \frac{y_1 - y_2}{x_1 - x_2} \tag{13}$$

$$c = \frac{y_1 x_2 - y_2 x_1}{x_2 - x_1} \tag{14}$$

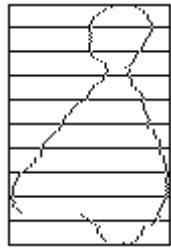


FIGURE 3.a: 11 X 1 grid superimposed over boundary image.



FIGURE 3.b: Contour of image represented with straight line.

Both the m and c may have any positive and negative values between negative infinity and positive infinity. As the immediate process after feature extraction is training of neural network with feature matrix all the feature values should be normalized between 0 and 1 as it was shown by experiment that neural network shows better response if the inputs are in range within 0 to 1 limit. For normalizing all the values of gradient (m) and constant (c) a hyperbolic nonlinear function is used as transfer function which is shown in equation 15 and a graphical representation is also shown in figure (4). In equation 15 any value of x maps the output between 0 and 1. k is a constant and value of it depends upon the no of features extracted. For 22 nos. of contour feature a value range from $k=.3$ to $.5$ is selected.

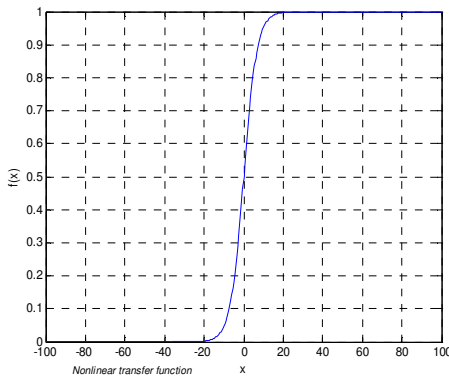


FIGURE 4: Nonlinear Hyperbolic Transfer Function.

$$f(x) = \frac{1}{1 + e^{-kx}} \tag{15}$$

If number of feature is increased its value should be decreased and similarly if number of features is decreased its value should be increased. In figure 4 plots for x versus $f(x)$ is shown for an x range from -100 to 100. In this way 22 nos. features are collected and finally put in to the feature matrix.

3.4 Shadow Feature

In this feature extraction scheme the resized image is segmented in to four segments as shown in figure (5). This feature is similar feature like transition feature as described in section 3.1. The term 'Shadow' is taken symbolically from the concept of formation of shadow when light falls over an object. When light is fallen from the upper side of the subsection it creates three shadows one is over the ground and other two are over the two inclined side. The phenomenon is shown in figure (6). When light is projected at a perpendicular direction to the horizontal line over the object shown in segment -1 shadow is formed over two sides AC and BC of the triangle and shadow also falls in the ground. The length of the shadows in side AC and BC are DC and EC respectively. The length of the shadow fallen over the ground is D'E' which is basically projection of point D and E over the ground. The features may be defined from these parameters in the following way.

$$\text{Shadow feature} = \frac{\text{Length of the shadow over the side}}{\text{length of the side}}$$

Using the above definition following three nos. of features can be extracted from each segment.

$$SF(1) = DC/AC \quad SF(2) = EC/BC \quad \text{and} \quad SF(3) = D'E'/AB$$

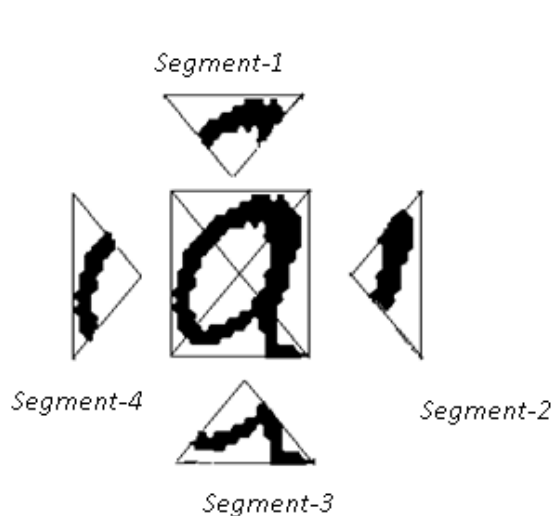


FIGURE 5: Four Subsections of the Resized Image

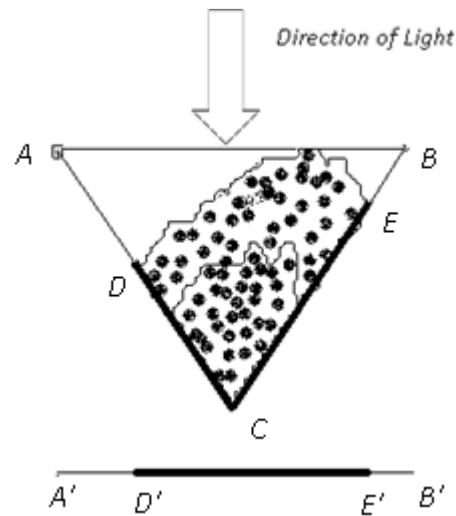


FIGURE 6: Shadow at Ground and Over Two Inclined Sides.

In the similar way features are extracted from all four segments and a total 12 nos. of features are added in the feature matrix for each character.

In our experiment four kinds of features extraction schemes are applied and we have taken 30 transition features, 36 sliding window amplitude features, 22 contour features and 12 shadow feature and a total 100 numbers of features are collected. Content of the training script are 10 sets of capital no(i.e from A-Z) and 10 sets of small letters (i.e from a-z) which are processed, filtered and resized as described in the previous sections. Finally the feature matrix having dimension 520x 100 is fed to the neural network for training. The description of the neural network and the algorithms used for training are described in the next section.

4. TRAINING OF NEURAL NETWORK

Training and design of neural network is the most important part in this experiment. Here we have used single and multilayer feed forward network with different first and second order back propagation training algorithm. It is not surprising that increase in no of feature improves

recognition efficiency. But in opposite side it takes a longer time to train the neural network. In our experiment we used four kinds of feature extraction scheme. Name of the feature extraction scheme and no of features of each type are mentioned in the previous section. But these large nos. of features takes very longer time to train the network if a proper training algorithm is not chosen. We have used different conjugate gradient methods for training and the speed of training improved significantly. All the algorithms showed better performance in respect to the speed of the training and recognition performance of the Neural Network are second order conjugate direction method such as Fletcher-Reeves [15], [21], Powell Beale [24], Polak Ribiere [16],[19] and Scaled Conjugate Gradient algorithm by Moller [22] .A comparative study also been given among the above algorithms and a comparison of converging performance of all the second order conjugate gradient methods is presented in result section. It has also been found experimentally that among the entire training algorithms Powell Beale method and Scaled Conjugate Gradient methods have shown better performance in respect of quick convergence of the error curve. Using *Scaled Conjugate Gradient* algorithm [22] the Hessian matrix of error equation is always positive over all iterations. But in all other algorithms mentioned above it is uncertain. This property of SCG algorithm increases learning speed reliably in successive iteration. Let us we define an error function Taylor Series expansion.

$$E(\tilde{w} + \Delta\tilde{w}) = E(\tilde{w}) + E'(\tilde{w})^T\Delta\tilde{w} + \Delta\tilde{w}^TE''(\tilde{w})\Delta\tilde{w} + \dots \quad (16)$$

Suppose the notations are used as $A = E'(\tilde{w})$ and $H = E''(\tilde{w})$. Weight vector in n^{th} iteration may be mentioned as \tilde{w}_n . H is the Hessian matrix and A is the local gradient vector. We will use a second order approximated error equation for further calculations and the same equation may be as follows

$$E(\tilde{w} + \Delta\tilde{w}) = E(\tilde{w}) + E'(\tilde{w})^T\Delta\tilde{w} + \Delta\tilde{w}^TE''(\tilde{w})\Delta\tilde{w} \quad (17)$$

The solution of the quadratic difference equation may be found as follows [15].

$$\Delta\tilde{w} = H^{-1}A \quad (18)$$

The above solution can be achieved subject to condition that H is positive definite matrix. The above equation states that how much amount of shifts is required for all the weights so that the error curve converges significantly. The above equation (18) is the essence of Newtown's Method [15] and from the equation it is found that the equation can be converged in one step if the inverse of the Hessian matrix is calculated. But in practical situation this phenomenon does not occur because of the some constraints as mentioned below.

- a) When number of Weights is more calculation of inverse of the Hessian Matrix is computationally expansive and highly time consuming matter.
- b) There is no guarantee that the inverse of the Hessian Matrix will be positive definite.
- c) The system converges in one step if and only if the error equation is perfectly quadratic in nature. But generally all error equation has some higher order terms.

To avoid the above limitations the only way was found to achieve the solution through an iterative process. Suppose a set of nonzero vectors $p_1, p_2, p_3, \dots, p_N$ are basis vectors in R^N and are H conjugate. If H is a positive definite matrix then the following conditioned must be satisfied [10].

$$\begin{aligned} \tilde{p}_k^T H \tilde{p}_i &= 0 \quad \text{for all } k \text{ and } i \text{ except } k = i \\ \tilde{p}_k^T H \tilde{p}_i &> 0 \quad \text{for } k = i \end{aligned} \quad (19)$$

Let $x_n = \tilde{p}_n^T H \tilde{p}_i$. When $k = i = n$ then x_n is nonzero quantity.

$$x_n = \tilde{p}_n^T H \tilde{p}_n \quad (20)$$

$$= \tilde{p}_n^T \tilde{s}_n$$

Here $\tilde{s}_n = H \tilde{p}_n = E''(\tilde{w}_n) \tilde{p}_n$. The idea to estimate the term s_n with a non symmetric approximation [22] it may be written as:

$$\tilde{s}_n = \frac{E'(\tilde{w}_n + \sigma_n \tilde{p}_n) - E'(\tilde{w}_n)}{\sigma_n} \quad \text{Where } 0 < \sigma_n \ll 1 \quad (21)$$

The above equation is the essence of second order error estimation used by Hestene [10] in conjugate gradient method. Scaled Conjugate Gradient method is a slightly different concept applied over it. In eqn.(20) the sign of x_n may be negative or positive but by definition if H is a positive definite matrix then x_n must be a greater than zero. Hestene [10] combines the concept of introducing a dumping scalar value and modifies the equation as in eqn. 22. The other steps are same to the conjugate direction method. The function of the scalar parameter λ is to compensate

$$s_n = \frac{E'(\tilde{w}_n + \sigma_n \tilde{p}_n) - E'(\tilde{w}_n)}{\sigma_n} + \lambda_n \tilde{p}_n \quad \text{Where } 0 < \sigma_n \ll 1 \quad (22)$$

the indefiniteness of value of $E''(w_k)$ when it is negative. In every iteration sign of x_n is checked i.e. whether $x_n > 0$ or $x_n < 0$. When x_n is less than zero the value of λ_n is increased and similarly when x_n is greater than zero the value of λ_n is decreased. All other steps are similar to the conjugate direction method.

Two up gradation equation governs the whole process. Firstly the weight up gradation equation (eqn. 23) and secondly the basis vector up gradation equation (eqn. 24). The target is to find such a solution set of weight vector that H become positive definite. As the iterations go forward the first order error gradient $E'(w_n)$ decreases and at last it reaches very near to zero. The equations are as follows

$$\tilde{w}_{n+1} = \tilde{w}_n + \alpha_n \tilde{p}_n \quad (23)$$

$$\tilde{p}_{n+1} = \tilde{r}_n + \beta_n \tilde{p}_n \quad (24)$$

Where $r_n = -E'(w_n)$, α_n and β_n may be calculated in each iteration by the following equations.

$$\alpha_n = \frac{\tilde{p}_n^T \tilde{r}_n}{\tilde{p}_n^T H \tilde{p}_n} = \frac{\tilde{p}_n^T \tilde{r}_n}{x_n} \quad (25)$$

And
$$\beta_n = \frac{\langle \tilde{r}_{n+1}, \tilde{r}_{n+1} \rangle - \langle \tilde{r}_{n+1}, \tilde{r}_n \rangle}{\tilde{p}_n^T H \tilde{p}_n} \quad (26)$$

The most important matter is to discuss that how the value of λ in eqn. 22 is chosen. As described earlier that, the function of λ_n is to check the sign of x_n in each iteration and to set a proper value of λ_n so that \tilde{x}_n get a positive value.

Let it is found that $x_n \leq 0$ and λ_n is raised by $\bar{\lambda}_n$ so that s_n get a new value

$$\overline{\tilde{s}}_n = \tilde{s}_n + (\bar{\lambda}_n - \lambda_n) \tilde{p}_n \quad (27)$$

$$\overline{\tilde{x}}_n = \tilde{p}_n^T \overline{\tilde{s}}_n \quad (28)$$

Putting the value of $\overline{\tilde{s}}_n$ from eqn. no 27 in eqn. 28

$$\overline{\tilde{x}}_n = x_n + (\bar{\lambda}_n - \lambda_n) |\tilde{p}_n|^2 > 0 \quad (29)$$

$$\Rightarrow \bar{\lambda}_n > \lambda_n - \frac{x_n}{|\tilde{p}_n|^2} \quad (30)$$

From eqn. 29 some guide line is found that the new λ_n should be greater than by $\frac{x_n}{|\tilde{p}_n|^2}$ but no such particular value can be obtained that so that the optimal solution can be obtained. However we have used

$$\bar{\lambda}_n = 3 \left(\lambda_n - \frac{x_n}{|\tilde{p}_n|^2} \right) \quad (31)$$

The above assumption is put into eqn. 29 we get

$$\tilde{x}_n = -2x_n + 3\lambda_n |\tilde{p}_n|^2 \quad (32)$$

Combining eqn. 32 and 25 we found the modified value of α_n

$$\alpha_n = \frac{\tilde{p}_n^T \tilde{r}_n}{-2x_n + 3\lambda_n |\tilde{p}_n|^2} \quad (33)$$

the above equation clears that with the increase of the value of λ_n decrease the height of step size and decrease in λ_n increases the height of step size which agrees to all the assumptions made earlier.

In the above steps it was realized that how the Hessian Matrix remain positive in all iterations using a scalar parameter λ . But the second order approximation of error equation which was used during entire calculation cannot assure for the best performance though we get positive definite Hessian matrix over all iterations. The above calculation steps assure that the error curve will be converging towards minima but in some situation choosing a proper value of λ is required otherwise the rate of convergence become slow. So a proper mechanism for scaling λ is adopted. A parameter ρ is defined which measure in what extent the second approximation of the error curve matches the original error curve. The following equation defines ρ as

$$\rho_n = \frac{E(\tilde{w}_n) - E(\tilde{w}_n + \alpha_n \tilde{p}_n)}{E(\tilde{w}_n) - E_{2q}(\alpha_n \tilde{p}_n)} \quad (34)$$

Here ρ_n measures how finely the second order approximated error equation $E_{2q}(\alpha_n \tilde{p}_n)$ matches to the $E(\tilde{w}_n + \alpha_n \tilde{p}_n)$. The above equation tells than the value of λ nearer to 1 means better the approximation. For faster convergence λ is scaled by the following equations.

$$\lambda_n = \frac{1}{3} \lambda_n \quad \text{if } \rho_n > .75$$

$$\lambda_n = \lambda_n + \frac{x_n(1 - \rho_n)}{|p_n|^2} \quad \text{if } \rho_n < .25 \quad (35)$$

In our experiments the initial values was chosen as $\lambda_1=10^{-3}$ and $\sigma = 10^{-5}$ was taken and in later steps $\sigma_n = \frac{\sigma}{|\tilde{p}_n|^2}$ was assumed. Initially weights are chosen as random nos. between 0 and 1. At starting of training in first iteration $p_1 = r_1 = -E'(\tilde{w}_1)$ is assumed. The algorithm may be summarized as below.

- a) Initialization of parameters like σ, p, r in first iteration.
- b) Calculation of second order parameters like s, x and σ
- c) Check whether Hessian matrix H is positive definite or not
- d) If false adjust the value of s by increasing λ and recalculate s again.
- e) Calculate ρ and readjust the value of λ
- f) Calculate step size
- g) If error > minimum error limit go to next iteration.
- h) Accept the weight vector for test.

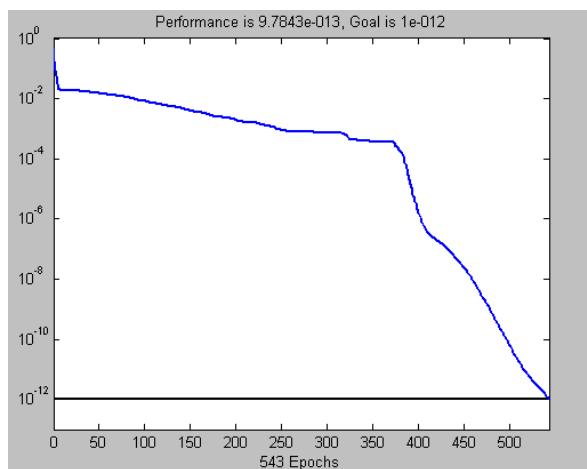


FIGURE 7: epoch versus error curve.

During the training operation the epoch versus error data is collected and the nature of the convergence is also noted. A run time error curve is plotted in figure 7 which shows that the nature of convergence is very fast and it takes just one or two min for completion of training and error limit reaches in 10^{-12} range within only 543 epoch.

5. RESULT AND DISCUSSION

This experiment is an extension of [26] where three types of feature extraction schemes were applied. Here a new kind of feature extraction scheme 'Shadow Feature' was implemented in section 3.4. It accelerated both the training performance and recognition performance. Though the

maximum recognition performance did not changed but most of the time it showed better performance. The results are shown in Table 1 and Table 2. The experiment for recognition of handwritten text was conducted in MATLAB environment where two types of images are used as input i.e. train image and test image. The purpose of the experiment was to recognize individual's handwriting by a neural network which is trained to identify the patterns of handwriting of the same person. Here both the train and test script was written by same person and the texts are casually written over the script. For this reason this experiment was not conducted and compared by any standard data base like CEDER or IAM data base of handwritten text. Two major features are highlighted in this experiment which differs it from all other researches [6],[13],[23] done earlier may be mentioned that is i) the superfast training speed ii) high recognition performance. Ten sets of handwritten capital and small letters of English alphabets are taken as training script and various handwritten texts written by same person are used as test script. A sample train and test scripts are shown in figure 8 and figure 9. All the characters are written in natural way over a sheet of A4 size. In both training and testing stage common character level segmentation and feature extraction is done over the train and test script. Training characters are English alphabets so there is no concept of forming words but test script is a text or may be called as group of words without any special symbol like ;, : and ?. In case of test script the segmentation is done from text level to word level and later from word level to character level. The start and end locations of each word in the text are stored. After recognition of all the characters in test script the computer printed words are regrouped by the NN output characters using the start and end location information of each words. Character level and Word level efficiency may be defined as follows.

$$\text{Character Level Efficiency} = (\text{No of characters recognized correctly}) / (\text{Total no of characters in the script})$$

$$\text{Word Level Efficiency} = (\text{No of words recognized correctly}) / (\text{Total no of words in the script})$$

One character misclassification in a word decreases the word level recognition performance. But it gets easily be improved if the same word is checked through a dictionary for validity of the word. If the word exists in the dictionary it returns the same word otherwise nearest word of same length is returned. For implementing this one dictionary is formed in MATLAB as an m-file which contains more than two thousand common words

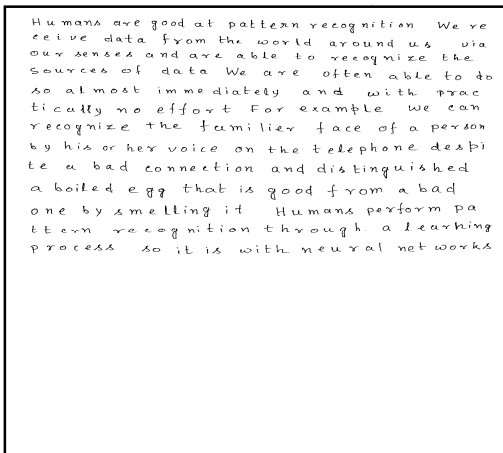


FIGURE 8: Train Script.

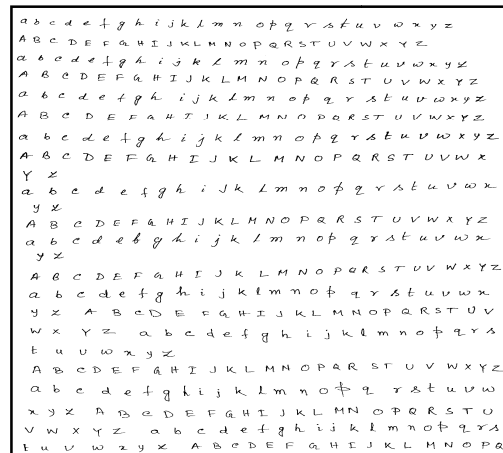


FIGURE 9: Test Script.

No of Features / character	NN structure	Training algorithm	No of epochs	Train efficiency	Error limit
100	100,120,52	SCG	543	100%	10 ⁻¹²
100	100,120,52	Powell-Beale	1412	100%	10 ⁻¹²
100	100,120,52	Fleture-Reeves	4322	100%	10 ⁻¹²

TABLE 1: Comparison Between Different Training Algorithms.

taken from a standard oxford dictionary. In this paper various algorithms were used for training such as a) Scaled Conjugate direction Method b) Powell Beale method and c) Fleture-Reeves method among the three algorithm Scaled conjugate algorithm shown the best training speed. A comparative statement is described in table.1.the above table shows that total 100 nos. of features were taken and NN structure indicates 52 nos. of

output neurons for 52 characters (a-z and A-Z) and the result was obtained for 120 nos. of hidden neurons. Both the three algorithms provide good convergence of the error curve but Scaled Conjugate Gradient algorithm provides fastest training speed and the training completes within only 543 epoch which takes less than one minute for reaching 100 percent train efficiency. The neural network trained with ten sets of handwritten alphabets that is total 520 nos. of characters when asked to make response to test feature matrix its response is really satisfactory only a few misclassification arises for very similar type of characters and for very bad handwritings. A character level recognition performance is shown in table 2 which shows that using double hidden layer with 150 and 100 hidden neurons we get maximum 83.60 percent case sensitive character level recognition and a maximum 92.32 percent case insensitive character level recognition performance is achieved. Though the degree of recognition is good but it may further be improved by lexicon matching technique which was discussed earlier. Some other researchers have reported like in [27] for 27 classes output recognition performance was 82 percent and 94 percent for lowercase and uppercase characters. In [12] a different type of feature extraction

scheme was applied and which gave 86 percent and 84 percent for lowercase and uppercase characters. In [13] a neural network based segmentation method has been presented and a global feature extraction scheme has been applied which shows character segmentation with a performance in the range of 56.11 percent for case sensitive and 58.50 percent which was further improved by a lexicon matching method and result was found as 85.71 percent. Compared to all the mentioned works this approach has given better result from three different aspects i) Very less number of features per characters was taken and it showed better result ii) It gives higher speed as scaled conjugate algorithm has been used for neural network training and iii) better recognition performance compared to the referred [12] [13] [27] works. Word label performance after lexicon matching is shown in table 3. This shows that the case insensitive performance shown in table 3 improves by maximum 99.02 percent and because some of the word which contain lesser misclassified characters get corrected by the lexicon matching technique which increases word level.

NN structure	MSE obtained	Train efficiency %	Test efficiency (Character Level) %	
			Case sensitive	Case insensitive
100,160,52	9.63×10^{-13}	100	81.43	89.42
100,200,52	9.75×10^{-13}	100	82.64	89.42
100,250,52	9.32×10^{-13}	100	79.25	88.70
100,120,80,52	9.44×10^{-13}	100	74.17	89.42
100,150,100,52	9.69×10^{-13}	100	83.60	92.32
100,200,100,52	9.98×10^{-13}	100	83.60	92.32

TABLE 2: Character Level Classification Performance.

NN Structure	Without Lexicon Matching		After Lexicon Matching	
	Character Level (%)	Word Level (%)	Character Level (%)	Word Level (%)
100,160,52	89.17	62.87	98.43	96.74
100,200,52	89.42	48.65	99.02	97.83
100,250,52	88.70	47.56	99.02	97.83
100,120,80,52	87.34	38.96	98.43	95.65
100,150,100,52	92.32	66.13	98.43	98.91
100,200,100,52	89.72	47.56	99.02	98.91

TABLE 3: Performance Before and After Lexicon Matching.

recognition and when word get corrected the misclassified characters are also changes to its correct version. As a whole both the word and character level performance increases. In this experiment maximum word level and character level recognition performance achieved as 98.91 and 99.02 percent respectively. Though the result of the experiment was very well and most of the words and characters were recognized correctly except a few misclassifications that were found during experiment which may be shown in table 4.

6. CONCLUSION

Reliable feature extraction methods are shown which is most important in Neural based approach for pattern recognition. When first order standard back propagation algorithms fails to produce result in a bulky neural network in a limited time frame, second order training algorithm work surprisingly. In this paper we focused both the training speed and recognition performance of handwritten alphabet based text. When no. of features are higher and no. of output classes are 52 all first order training algorithms basically fails or generates very poor result in training. Basic reason is the slower convergence of error curve and proper second order training algorithm become suitable replacement of those algorithms. Using this algorithm, Hessian matrix of error equation always remain positive definite and in every iteration the error curve converges in a faster way. In this paper comparison

Original	Recognized	Original	Recognized
H	M,n	n	m
r	v,Y	Y	r,y
e	c,l,C	y	r,Y
a	Q,u,l,G	c	C
f	t	C	c
s	a,b,	O	o
l	x	o	O,b,D
l	j	j	i
b	o	t	f,F,q

TABLE 4: General Misclassifications.

between several second order training algorithms has been shown and it was found experimentally that Scaled Conjugate Gradient algorithms works with fastest speed and recognition performance is also excellent. Training part was very fast but regarding the complexity of the test script, scripts characters are simple, easy to understand by human eye. Test script contains some inter-line horizontal space and some inter-word and inter-character space also. This pattern is not always available in natural handwriting. So it needs more experimental effort for faithful conversion from difficult handwriting to text conversion. However, this approach may be a true guideline for future research for giving computer an intelligence which a human being applies everyday and at every moment.

7. REFERENCES

- [1] S-B. Cho, "Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals", IEEE Trans. on Neural Networks, vol.8, 1997, pp. 43-53.
- [2] Verma, B. "A Contour Code Feature Based Segmentation For Handwriting Recognition", 7th IAPR International conference on Document Analysis and Recognition, ICDAR'03, 2003, pp. 1203-07.
- [3] N.W. Strathy, C.Y. Suen and A. Krzyzak, "Segmentation of Handwritten Digits using Contour Features", ICDAR '93, 1993, pp. 577-580.2003.
- [4] R G Casey and E Lecolinet "A Survey of Methods and Strategies in Character Segmentation," IEEE Trans. Pattern analysis and Machine Intelligence, vol. 18, 1996, pp. 690-706.
- [5] Fletcher, R., and C.M. Reeves "Function minimization by conjugate gradients" the computer journal, vol-7 149-153, 1964.
- [6] D. Gorgevik and D. Cakmakov, An Efficient Three-Stage Classifier for Handwritten Digit Recognition, ICPR, vol. 4, 2004, pp. 507-510.
- [7] Gernot A. Fink, Thomas Plotz, "On Appearance-Based feature Extraction Methods for Writer-Independent Handwritten Text Recognition" Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 1520-5263/05.
- [8] C. E. Dunn and P. S. P. Wang, "Character Segmentation Techniques for Handwritten Text A Survey Proceedings of the 11th International Conference on Pattern Recognition, The Hague, The Netherlands, 1992 pp 577-580.
- [9] Matthias Zimmermann and Horst Bunke "Optimizing the Integration of a Statistical Language Model in HMM based Offline Handwritten Text Recognition" .Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) 1051- 4651/04.
- [10]. Hestenes, M." conjugate Direction Methods In Optimization", Springer-verlag, New York, 1980.
- [11] S-W. Lee, "Off-Line Recognition of Totally Unconstrained Handwritten Numerals Using Multilayer Cluster Neural Network". IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 18, 1996, pp. 648-652.
- [12] P. D. Gader, M. Mohamed and J-H. Chiang, 'Handwritten Word Recognition with Character and Inter-Character Neural Networks", IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, vol .27, 1997, pp. 158-164.

- [13] Blumenstein and B. Verma. "Neural-based solutions for the segmentation and recognition of difficult handwritten words from a benchmark database". In Proc. 5th International Conference on Document Analysis and Recognition, pages 281–284 Bangalore, India, 1999.
- [14] J-H. Chiang, "A Hybrid Neural Model in Handwritten Word Recognition", Neural Networks, vol. 11, 1998, pp. 337-346.
- [15] Simon Haykin 'Neural Networks A comprehensive Foundation', second edition.
- [16] R.O.Duda ,P.E. Hart , D.G.stock 'Pattern classification' second edition.
- [17] William K.Pratt 'DIGITAL IMAGE PROCESSING' Third edition.
- [18]. S.Rajeshkaran and G.A. Vijayalakshmi Pai 'Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications.' Eastern Economy Edition.
- [19]. K.M.Khoda, Y.Liu and C. Storey "Generalized Polak –Ribiere Algorithm" journal of optimization theory and application: vol 75,No 2,November 1992.
- [20] Verma, B.; Hong Lee;' A Segmentation based Adaptive Approach for Cursive Hand written Text Recognition Neural Networks, 2007. IJCNN 2007. International Joint Conference on 12-17 Aug. 2007 Page(s):2212 – 2216 Digital Object Identifier 10.1109/IJCNN.2007.4371301.
- [21] Fletcher, R.(1975). "practical methods of optimization ". New York: John Wiley & Sons.
- [22] Martin Fodslette Moller. "A Scaled Conjugate Gradient Algorithm For Fast Supervised Learning." Neural Networks, vol 6:525-533, 1993.
- [23] M. Blumenstein and B. Verma "Neural-based Solutions for the Segmentation and Recognition of Difficult Handwritten Words from a Benchmark Database" In Proc. 5th International Conference on Document Analysis and recognition, pages 281–284, Bangalore, India, 1999.
- [24] Y. H. Dai and Y. Yuan, Convergence properties of the Beale-Powell restart algorithm, Sci.China Ser. A, 41 (1998), pp. 1142-1150.
- [25] U. Pal, N. Sharma, T. Wakabayashi, F. Kimura, "Off-line handwritten character recognition of Devanagari script", Proceedings of 9th international conference on document analysis and recognition , vol. 1, pp. 496-500, 2007.
- [26] Haradhan Chel, Aurpan Majumder, Debashis nandi, "Scaled Conjugate Gradient Algorithm in Neural NetworkBased Approach for Handwritten Text Recognition" D. Nagamalai, E. Renault, M. Dhanushkodi (Eds.): CCSEIT 2011, CCIS 204, pp. 196–210, 2011.
- [27] P. Gader, M. Whalen, M. Ganzberger, and D. Hepp. "Handprinted word recognition on a NIST dataset. Machine Vision and Applications, 8:31–41, 1995

Performance Analysis of Daubechies Wavelet and Differential Pulse Code Modulation Based Multiple Neural Networks Approach for Accurate Compression of Images

S.Sridhar

Faculty-ECE

LENDI Institute of Engg&Technology

Vlzanagaram, Andhra Pradesh, INDIA

sridhar.vskp@gmail.com

P.Rajesh Kumar

Associate Professor-ECE

Andhra University College of Engg

Visakhapatnam, Andhra Pradesh, INDIA

rajeshauce@gmail.com

K.V.Ramanaiah

Associate Professor-ECE

YSR Engg College of Yogi Vemana University

Proddatur, Andhra Pradesh, INDIA

ramanaiahkota@gmail.com

Abstract

Large Images in general contain huge quantity of data demanding the invention of highly efficient hybrid methods of image compression systems involving various hybrid techniques. We proposed and implemented a Daubechies wavelet transform and Differential Pulse Code Modulation (DPCM) based multiple neural network hybrid model for image encoding and decoding operations combining the advantages of wavelets, neural networks and DPCM because, wavelet transforms are set of mathematical functions that established their viability in the areas of image compression owing to the computational simplicity involved in their implementation, Artificial neural networks can generalize inputs even on untrained data owing to their massive parallel architectures and Differential Pulse Code Modulation reduces redundancy based on the predicted sample values. Initially the input image is subjected to two level decomposition using Daubechies family wavelet filters generating high-scale low frequency approximation coefficients A2 and high frequency detail coefficients H2, V2, D2, H1, V1 and, D1 of multiple resolutions resembling different frequency bands. Scalar quantization and Huffman encoding schemes are used for compressing different sub bands based on their statistical properties i.e the low frequency band approximation coefficients are compressed by the DPCM while the high frequency band coefficients are compressed with neural networks. Empirical analysis and objective fidelity metrics calculation is performed and tabulated for analysis.

Keywords: Backpropagation, Daubechies Wavelet, DPCM, PSNR, MSE, Neural Networks.

1. INTRODUCTION

The growing energy requirements of wireless data services, biomedical applications, computer graphics and many other web based applications disclosed an urge to innovate new techniques in the areas of signal and image processing to compress and decompress signals as well as still images and videos of various types and sizes to meet the everlasting storage space and channel bandwidth requirements. Wavelets perform better and provide good compression ratios for high resolution images relative to other competing technologies like JPEG objectively and subjectively as well. Unlike JPEG, wavelet does not show any blocking effects and allow degradation of the whole image quality while preserving the significant details of an image [1].The rapid development of high performance computing and communications opened up tremendous

opportunities in the development of different telecommunication applications, Image compression is the context where images of different sizes are compressed using different methodologies to meet demand for ever growing bandwidth requirements.

Since Images can be regarded as two dimensional signals, many digital Image compression techniques for one dimensional signal are extended to 2-D images to exploit the correlations between the neighboring pixels to eliminate the redundancies. Traditional techniques of compression aims at reducing the Coding, Interpixel and Psycho visual redundancies, [2] additionally new soft computing technologies like Neural Networks are developed for image compression owing to their features of Parallelism, Learning capabilities, Noise Suppression, Transform extraction and Optimized Approximations which encouraged researchers to use multiple combination techniques of wavelets and neural networks for image compression applications.

Image compression techniques are basically Lossy and Lossless. Lossless image compression techniques encode data exactly such that decoded image is almost identical to original image but they are limited in terms of compression ratio [3]. Few lossless image compression techniques are

- i) Run Length encoding
- ii) Huffman encoding
- iii) LZW coding
- iv) Area coding

Lossy image compression techniques encode an approximation of original image with good compression ratios and less distortion in the reconstructed image. Lossy compression techniques include transform coding, quantization and entropy encoding operations, In transform encoding input image is mathematically transformed by separating image information on gradual spatial variation of brightness from regions with faster variations in brightness at edges of the image [3][4] Few lossy compression techniques are:

- i) Transformation Coding techniques
- ii) Vector quantization
- iii) Fractal coding
- iv) Block Truncation coding
- v) Sub band coding

The proposed methodology of hybrid compression is a combination of both the lossy compression and lossless compression techniques.

This paper is organized as follows. Section 2, briefs the objective fidelity design metrics. Section 3, explains the Daubechies wavelet transform and Differential Pulse Code Modulation. In section 4, neural networks and backpropagation algorithm for training them are discussed. Section 5, discusses the proposed hybrid methodology of image compression and decompression system. Section 6, elaborates the Experimental results. Section 7 discusses the conclusion reached by analysis.

2. DESIGN METRICS

Digital image compression techniques are normally analyzed with objective fidelity measuring metrics like Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE), Compression Ratio (CR), Encoding time, Decoding time and Transforming time etc[2][5].

2.1 Mean Square Error (MSE)

MSE for monochrome images is given by

$$\frac{1}{N^2} \sum_i^N \sum_j^N [X(i, j) - Y(i, j)]^2 \quad (1)$$

MSE for color images is given by

$$\frac{1}{N^2} \sum_i^N \sum_j^N \{ [r(i, j) - r^*(i, j)]^2 + [g(i, j) - g^*(i, j)]^2 + [b(i, j) - b^*(i, j)]^2 \} \quad (2)$$

Where $r(i, j)$, $g(i, j)$ and $b(i, j)$ represents the color pixels at location (i, j) of the original image. $r^*(i, j)$, $g^*(i, j)$ and $b^*(i, j)$ represent the color pixel of the reconstructed image, while $N \times N$ denotes the size of the pixels of the color images [2]

2.2 Peak Signal to Noise Ratio (PSNR)

Peak signal to Noise Ratio is the ratio between signal variance and reconstruction error variance. PSNR is usually expressed in Decibel scale. The PSNR is a most common measure of the quality of reconstructed image in case of image compression.

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \quad (3)$$

Here 255 represent the maximum pixel value of the image, when the pixels are represented using 8 bits per sample. PSNR values range between infinity for identical images, to 0 for images that have no commonality. PSNR is inversely proportional to MSE and compression ratio i.e PSNR decreases as the compression ratio increases.

2.3 Compression Ratio (CR)

Compression ratio is defined as the ratio between the original image size and compressed image size.

$$Compression\ Ratio = \frac{OriginalImageSize}{CompressedImageSize} \quad (4)$$

3. COMPRESSION TECHNIQUES

3.1 Wavelet Transforms

Wavelet transforms allow good localization in frequency and space, Wavelet transforms represent image as a sum of wavelet functions with different locations and scales [18]. Wavelet transforms are continuous and discrete. Continuous wavelet transforms are time consuming for long signals, as the signal needs to be integrated at all times. Discrete wavelet transform (DWT) is implemented through sub band coding, it can localize signals in time and scale, the scaling operation is done by changing the resolution of signal through sampling [10].

Often signal processing in time domain require frequency related information, Mathematical transforms translate the information of signals into different forms. For example the Fourier transforms converts the signals in both time domain and frequency domain, but they failed to provide time specific frequency information however in Short Term Fourier Transform (STFT) window based technique, different parts of the signal can be viewed specifically [13]. But in accordance with the Heisenberg's Uncertainty Principle, resolution gets worse in frequency domain, if it is improved in time domain by zooming different sections. The power of wavelets

comes from the use of multiresolution i.e. different parts of the wave are viewed through different sized windows where high frequency parts in the signal use smaller windows to give good time resolution while the low frequency parts use big windows to extract frequency information [5].

In case of wavelet decomposition, wavelet function represent the high frequency detail parts clearly showing the Vertical, Horizontal and Diagonal details of the image while the scaling function represent the low frequencies or smooth parts of the image clearly corresponding to the approximation coefficients. If the number of high frequency coefficients are smaller than the threshold values they can be set to zero without significantly changing the image, If the number of zeros are greater, large compression can be achieved. If the threshold value is set to zero, then the energy or the amount of information retained is 100% and the compression is said to be lossless as the image can be reconstructed exactly. However, as more zeros are obtained more energy is lost; hence a balance is required [18].

3.1.1 Daubechies Wavelets

A major problem in the development of wavelets during the 1980's was the search for scaling functions that are compactly supported, orthogonal and continuous. These scaling functions were first constructed by Ingrid Daubechies, this construction amounts to finding the low pass filter h , or equivalently, the Fourier series. Ingrid Daubechies invented compactly supported orthonormal wavelets- thus making discrete wavelet analysis practicable [20].

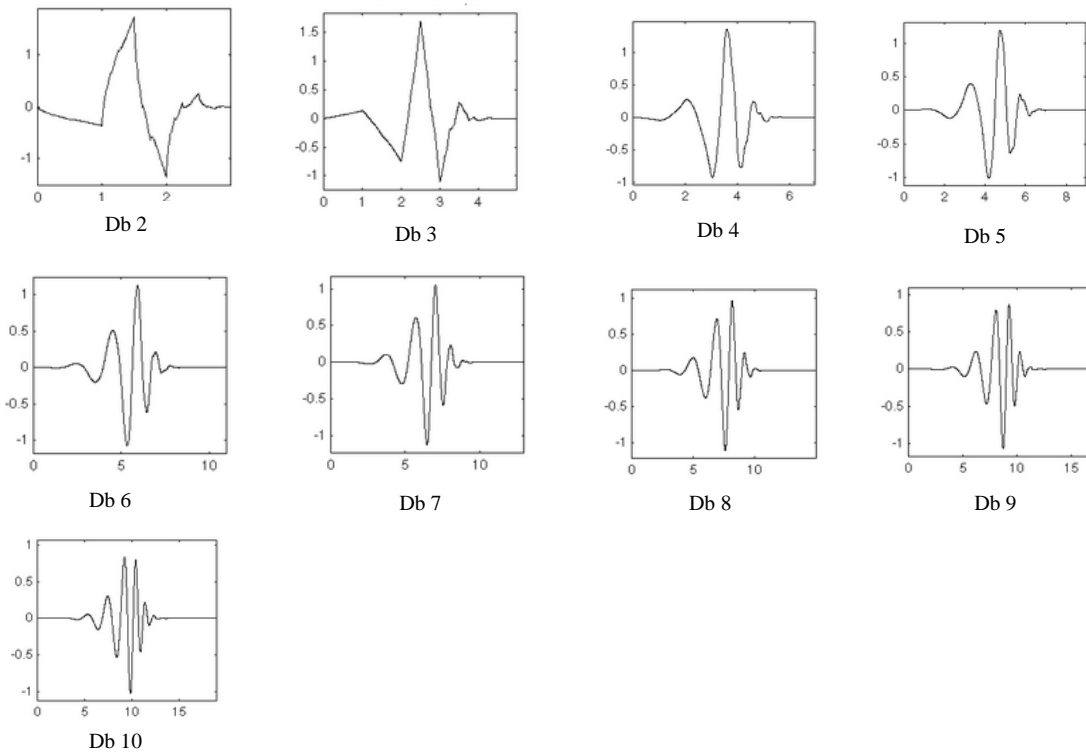


FIGURE 1: Wavelet Functions of Daubechies.

Daubechies wavelet transform signal is defined by the scaling and wavelet functions that are expressed in terms of α and β coefficients, respectively. Daubechies 1 represents same wavelet as Haar wavelet.

$$\alpha_1 = \frac{1 + \sqrt{3}}{\sqrt[4]{2}} \quad (5)$$

$$\alpha_2 = \frac{3 + \sqrt{3}}{\sqrt[4]{2}} \quad (6)$$

$$\alpha_3 = \frac{3 + \sqrt{3}}{\sqrt[4]{2}} \quad (7)$$

$$\alpha_4 = \frac{1 + \sqrt{3}}{\sqrt[4]{2}} \quad (8)$$

Daubechies wavelet transforms are defined similar to the Haar wavelet by obtaining running averages and differences through scalar products with scaling signals and wavelets. For high order Daubechies wavelets DbN, N denotes the order of wavelet and the number of vanishing moments, Daubechies wavelets have the highest number (A) of vanishing moments for given support width N=2A, The length of the wavelet transform is easy to put into practice using the fast wavelet transform, the approximation and detail coefficients are of length [16] [21].

$$\text{Floor} \left(\frac{n-1}{2} \right) + N \quad (9)$$

If n is the length of f (t), this wavelet has balanced frequency responses but non-linear phase responses. Wavelets with fewer vanishing moments give less smoothing effects and remove less details, but wavelets with more vanishing moments produce distortions. Daubechies wavelets are widely used to solve broad range of problems like for example, self-similarity Properties of a signal or fractal problems, signal discontinuities etc. The wavelet functions of Daubechies family are listed in fig.1, in which x-axis represents the time and y-axis represents the frequency.

3.2 Differential Pulse Code Modulation

Differential pulse code modulation (DPCM) [14] [15] is a signal encoder that uses the baseline of pulse code modulation (PCM) but adds some functionality based on the prediction of signal samples. Input to a DPCM is an analog or digital signal. If the input is a continuous time analog signal, it needs to be sampled first so that a discrete time signal is the input to the DPCM encoder. In DPCM, We transmit the difference e (n), between x (n) and its predicted value y (n) but not the present sample x (n). At the receiver, we generate y (n) from the past sample value to which the received x (n) is added to generate x (n). There is, however, one difficulty associated with this scheme. At the receiver, instead of the past samples x (n-1), x (n-2)... as well as e(n), we have their quantized version xs (n-1), xs (n-2),... This will increase the error in reconstruction. In such a case, a better strategy is to determine y (n), the estimate of xs (n) (instead of x (n), at the transmitter also from the quantized samples xs (n-1), xs (n-2),... The difference e (n)=x (n)-y (n) is now transmitted via PCM. At the receiver, we can generate y (n), and from the received e (n), we can reconstruct xs (n). [16]

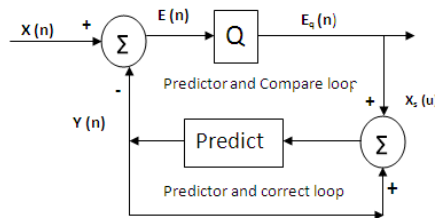


FIGURE 2: DPCM Encoder.

The difference of the original image data, x (n), and prediction image data, y(n) is called estimation residual, e(n). So

$$\mathbf{e(n) = x(n) - y(n)} \quad (10)$$

Is quantized to yield

$$e_Q(n) = x(n) + q(n) \tag{11}$$

Where $q(n)$ is the quantization error and $e_q(n)$ is quantized signal and

$$q(n) = e_q(n) - e(n) \tag{12}$$

$$q(n) = \frac{I_{max}}{2^b} = \frac{(simg)_{max}}{2^b} \tag{13}$$

Here b is number of bits. I_{max} (Simg) $_{max}$ is maximum value of an image signal. The prediction output $y(n)$ is fed back to its input so that the predictor input $x_s(n)$ is

$$x_s(n) = y(n) + e_q(n) \tag{14}$$

$$= x(n) - e(n) + e_q(n)$$

$$= x(n) + q(n)$$

This shows $x_s(n)$ is quantized version of $x(n)$. The prediction input is indeed $x_s(n)$, as assumed [19].

4. Artificial Neural Networks and LM Algorithm

Artificial neural networks pre-process the input patterns to produce patterns of sufficient compression rates preserving the information security [6]. An artificial neural network is a nonlinear system and powerful data modeling tool meant for solving optimization problems. Few advantages of neural networks are, they are self adaptive and adjust themselves to the data, they approximate any function with arbitrary accuracy, they are fault tolerant via redundant information coding, and can retain their capabilities despite major network damage with minimum degradation in the performance. Finally, neural networks model the real world complex relationships [7].

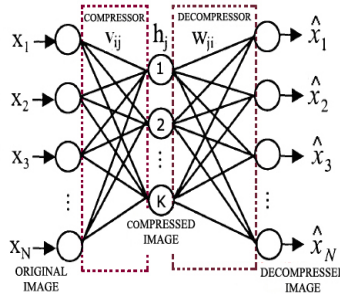


FIGURE 3: Basic Image Compression using ANN.

In case of a multilayered perceptron (MLP) type feed forward neural network architecture, number of connections between any two layers is the summation of number of bias neuron connections of the second layer (bias connections of a layer is equal to the number of layer neurons) and product of total number of neurons in the two layers. If there are N_i neurons in the input layer, N_h neurons in the hidden layer and N_o neurons in the output layer, total number of connections are given by the equation:

$$\text{Network Size : } (N_w) = [(N_i * N_h) + N_h] + [(N_h * N_o) + N_o] \tag{15}$$

Levenberg-Marquardt (Backpropagation) algorithm [4] is a common supervised training methods used for training the artificial neural networks which is based on error-correction learning rule. Here the error propagation through the network involves a forward pass and a backward pass. In

the forward pass the synaptic weights of the network are fixed, however, in the backward pass the synaptic weights are adjusted in accordance with an error-correction rule. The Network is trained by iterative updation of weights to minimize the mean square error. [8] The computed error signal is then propagated backward to the lower layers and the synaptic weights of the network are adjusted accordingly such that the error is decreased along the descent direction to move the actual response of the network closer to the desired response. In case of neural networks with more than one hidden layer, backpropagation algorithm converges slowly, as the output is saturated due to the activation function used, and the descent gradient takes a very small value, even if the output error is large, leading to a little progress in the adjustment of weights. Learning rate and momentum factor are two parameters used for weights adjustments in the direction of the descent to suspend oscillations [9].

5. IMAGE COMPRESSION/ DECOMPRESSION SYSTEM

The proposed architecture analyses the performance of Daubechies wavelet and Differential Pulse Code Modulation based hybrid model using multiple neural networks for accurate compression of images. Scalar quantization and Huffman encoding are also used as well to eliminate the psychovisual and coding redundancies. Initially, the selected standard input image is compressed by decomposing it twice using Daubechies (Db10) filter wavelet transforms to generate the low frequency band approximation coefficients and the high frequency band detail coefficients clearly showing the horizontal, vertical and diagonal details of the image after the two levels of decomposition. The low frequency approximation coefficients in the second level are now compressed using differential pulse code modulation encoder while the high frequency band coefficients after both levels of decomposition are compressed in a parallel arrangement of artificial neural networks of dimensions M-N-P where M, N, P represent the number of artificial neurons in the Input layer, Hidden layer and the Output layer. Further compressed hidden layer outputs of the five proposed neural networks are scalar quantized together and Huffman encoded in combination with the DPCM output, this operation generates the overall compressed image output. Decompression process involves the reverse operations of Huffman decoding, reverse quantization; decompression in neural networks between hidden and output layers of the respective neural networks, inverse DPCM operation or DPCM decoding and inverse Dabechies filter wavelet transform operations to retrieve the reconstructed image.

Bench mark images circuit, lifting body, rice, testpat1 and Lena of different sizes ranging from 256 x 256 pixels down to 32 x 32 pixels are considered for analysis.

5.1 Image Encoding Scheme

Initially the selected bench mark image of size 256 x 256 is decomposed first using Daubechies filter wavelet transform(Db2) to generate low frequency approximation coefficients A1 and three high frequency detail coefficients H1, V1, D1 of resolutions 128 X 128 each, after the first level of decomposition. The first level approximation coefficients so obtained are now decomposed at the second level generating approximation coefficients A2 and three detail coefficients H2, V2, D2, of resolutions 64 x 64 giving rise to a total of seven frequency bands after two level decomposition. The first band high-scale low frequency approximation coefficients A2 contain significant information while the low-scale, high frequency detail coefficients represent the second, third and fourth bands respectively. Band1 low frequency approximation coefficients A2 are now compressed using DPCM to reduce the inter pixel redundancy; DPCM predicts the value of neighboring pixel based on the previous pixel information, the difference between current pixel and predicted pixel is then given to an optimal quantizer which reduces the granular noise and slope over load noise. Finally the error output is obtained from DPCM.

The second level decomposed low-scale, high frequency detail coefficients H2, V2, D2 are encoded using three different multi layer Perceptron type feed forward neural networks of dimensions 16-12-16. Similarly the first level decomposed low-scale, high frequency detail coefficients H1, V1 are encoded using two different MLP type feed forward neural networks of dimensions 16-8-16. Compression normally takes place between the input layer and hidden layer of the selected neural network; the compressed hidden layers coefficients at the outputs of the

five different neural networks are scalar quantized, the quantized bits in combination with DPCM encoded data are further Huffman encoded to generate the compressed image, which can be stored for the purpose of transmission .

In the entire process of encoding and decoding operations the first level decomposed low-scale, high frequency detail coefficients D1 are discarded for the current analysis since they contain no useful data. Throughout the analysis all the artificial neural networks are trained with error backpropagation algorithm or Levenberg-Marquardt algorithm.

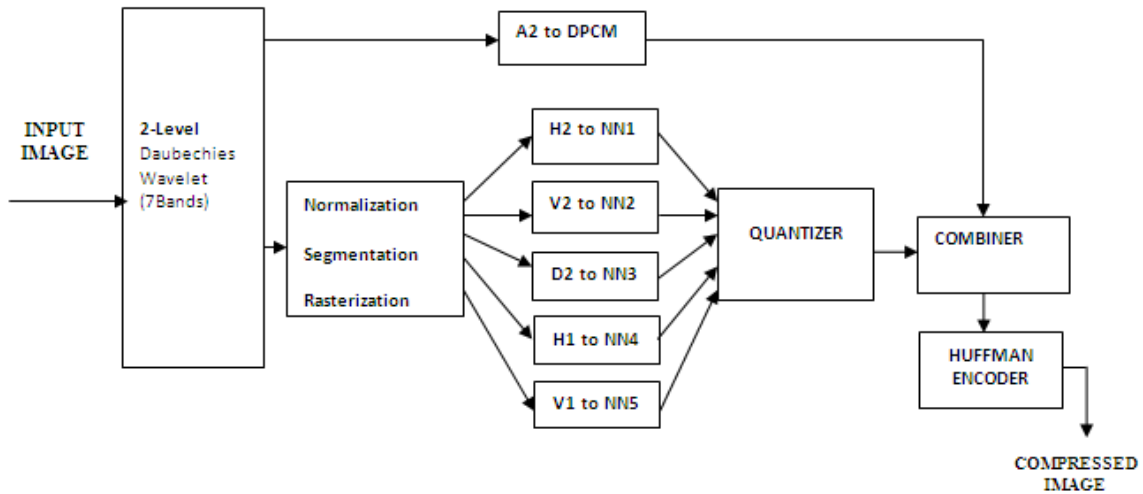


FIGURE 4: Proposed Image Compression System Architecture.

5.2 Image Decoding Scheme

In the decoding process as shown in Fig. 5, the compressed image coefficients are decoded in the Huffman decoder initially; the reconstructed bit streams are now split to separate the band1 high-scale low frequency approximation coefficients A2 and the remaining five bands of high frequency detail coefficients H2, V2, D2, H1 and V2. The compressed low frequency band-1 coefficients are now fed to the inverse DPCM unit for decoding operation while band 2 to band 6 high frequency detail coefficients are reverse quantized and fed to the output layers of respective neural networks for decoding purpose. Reconstructed sub band coefficients of inverse DPCM unit and neural networks are reconstructed with Inverse Daubechies filter Wavelet Transform (IDWT) operation to generate the desired reconstructed image.

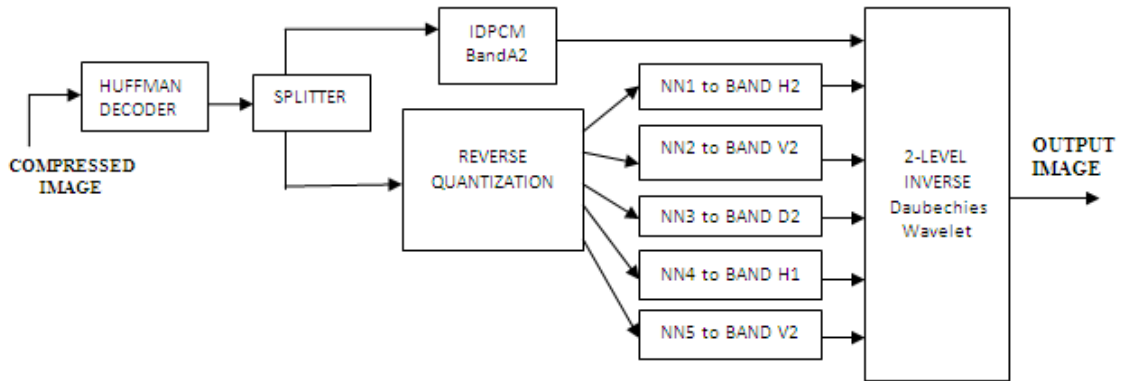


FIGURE 5: Proposed Image Decompression System Architecture.

6. EXPERIMENTAL RESULTS

Experiments are conducted on several standard bench mark images and the results of few of the images are presented here.

Figures 6-10, as shown below contain four different images in each figure. They are arranged in the order of top row and bottom row with two images in each row. They can be read as the original input image and 2-Level wavelet compressed image in the top row starting from the left, and the output image, error image in the bottom row from the left.

Measured objective fidelity metrics PSNR, MSE and CR for each image analysed after experimentation are tabulated for relative analysis purpose.

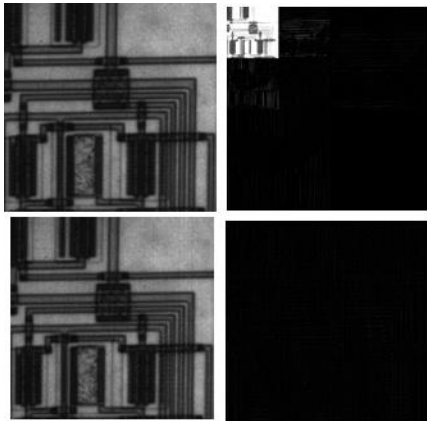


TABLE 1: Results of Cameraman Image.

Input Image	BAND	A2	H2	V2	D2	H1	V1
Circuit	NN Size		16-12-16	16-12-16	16-12-16	16-8-16	16-8-16
	Encoding Time	307.8511					
	PSNR	54.00	36.10	30.55	42.94	45.69	45.09
	MSE	0.25	15.92	57.23	3.29	1.75	2.01
	Overall PSNR	31.9580					
	Overall MSE	41.4267					

FIGURE 6: Circuit Image.

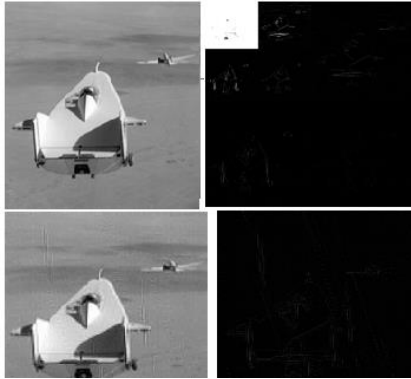


FIGURE 7: Lifting Body Image.

Table 2. Results of Lifting Body Image

Input Image	BAND	A2	H2	V2	D2	H1	V1
Lifting Body	NN Size		16-12-16	16-12-16	16-12-16	16-8-16	16-8-16
	Encoding Time	274.9183					
	PSNR	52.88	34.24	31.89	39.09	41.22	39.90
	MSE	0.33	24.44	42.04	8.01	4.90	6.64
	Overall PSNR	30.6541					
	Overall MSE	55.9337					

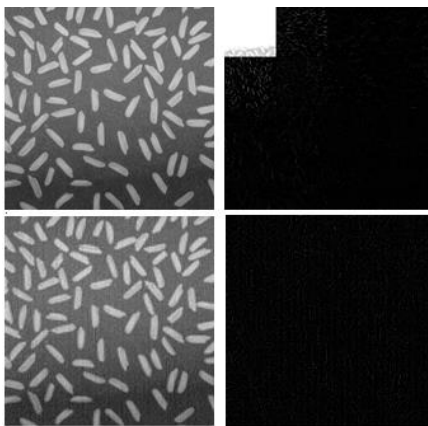


FIGURE 8: Rice Image.

TABLE 3: Results of Rice Image.

Input Image	BAND	A2	H2	V2	D2	H1	V1
Rice	NN Size		16-12-16	16-12-16	16-12-16	16-8-16	16-8-16
	Encoding Time	512.5771					
	PSNR	53.04	30.07	25.59	33.07	27.25	28.60
	MSE	0.32	63.96	179.24	32.06	122.26	89.73
	Overall PSNR	26.1428					
	Overall MSE	158.0532					

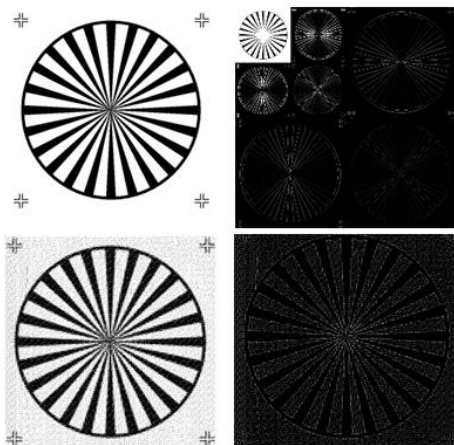


FIGURE 9: Testpat1 Image.

TABLE 4: Results of Testpat1 Image.

Input Image	BAND	A2	H2	V2	D2	H1	V1
TESTPAT1	NN Size		16-12-16	16-12-16	16-12-16	16-8-16	16-8-16
	Encoding Time	0.00214					
	PSNR	45.07	19.07	16.32	24.98	10.71	18.52
	MSE	2.022	803.75	0.005	206.38	0.005	913.7
	Overall PSNR	16.0276					
	Overall MSE	0.00162					



FIGURE 10: Lena Image.

TABLE 5: Results of Lena Image.

Input Image	BAND	A2	H2	V2	D2	H1	V1
LENA	NN Size		16-12-16	16-12-16	16-12-16	16-8-16	16-8-16
	Encoding Time	238.698					
	PSNR	51.27	26.96	29.98	35.43	39.61	35.98
	MSE	0.48	130.67	65.24	18.60	7.11	16.40
	Overall PSNR	23.9491					
	Overall MSE	261.921					

7. CONCLUSION

In proposed hybrid encoding and decoding scheme five bench mark input images Circuit, Lifting Body, Rice, Testpat1 and Lena of size 256 x 256 are tested and analysed for variations in objective fidelity metric measures PSNR, MSE, CR and Encoding time. It was observed that Circuit image produced better PSNR of order 31.958; Testpat1 image has the merit of being faster in performing the encoding operation and demerits of producing least PSNR and highest MSE values. When compared to neural networks based image compression techniques, Wavelet based image compression combined with DPCM and neural networks dramatically improve the quality of reconstructed images.

The proposed methodology can be explored to obtain better metrics with more number of hidden layers in the selected neural networks and varying the number of neurons in the hidden layers for training the network properly for early convergence. The proposed architecture can be tested with neural networks based on learning vector quantization and code book maintenance technique, arithmetic coding instead of Huffman encoding technique etc. This work can be further extended to explore the possibilities of applying hybrid combination techniques for effective data, image and video compression also.

There are many other existing and new wavelet functions, whose combination with other methodologies can always create wonderful statistics.

8. ACKNOWLEDGEMENTS

The authors express their deep sense of gratitude to the department of ECE, Lendi College of Engineering for provision of excellent facilities that made this work possible. The authors would also like to express their thanks to the passed out graduate engineers for their contribution.

9. REFERENCES

- [1] Aran Namphol, Steven H.Chin and Mohammed Arozullah, "Image Compression with a Hierarchical Neural Network", IEEE Transactions on Aerospace and Electronic Systems vol 32, no 1 January 1996.
- [2] Liu-Yue Wang and EARKKI Oja, "Image Compression by Neural Networks: A comparison study".

- [3] Sonal and Dinesh Kumar, "A study of various Image Compression Techniques", *Guru Jhmbheswar university of science and technology, Hisar*.
- [4] S.Anna Durai and E.Anna Saro, "Image Compression with Back-Propagation Neural Network using Cumulative Distribution Function", *World Academy of Science Engineering and Technology 17, 2006*.
- [5] Marta Mrak and Sonia Grgic, "Picture quality Measures in Image Compression Systems", *EUROCON 2003 Ljubljana, Slovenia*.
- [6] G.L.Sicuranzi, G.Ramponi and S.Marsi, "Artificial Neural Network for Image Compression", *Electronic Letters, vol26, no.7,pp. 477-479, March 29 1990*.
- [7] Hahn-Ming Lee, Tzong-Ching Huang and Chih-Ming Chen, "Learning Efficiency Improvement of Backpropagation Algorithm by Error Saturation Prevention Method, 0-7803-5529-6/992 @1999 IEEE.
- [8] Amjan Shaik and Dr.C.K.Reddy,"Empirical Analysis of Image Compression through wave transform and Neural Network", *International Journal of Computer Science and Information Technologies (IJCSIT), vol.2 (2), 2011, 924-931*.
- [9] K.Siva Nagi Reddy, Dr.B.R.Vikram,, B.Sudheer Reddy and L.Koteswararao, "Image Compression and Reconstruction using a new approach by Artificial Neural Network", *International Journal of Image Processing (IJIP), Volume (6): Issue (2):2012*.
- [10]B.Eswara Reddy and K.Venkata Narayana, "A lossless image compression using traditional and lifting based wavelets"
- [11]Yogendra Kumar Jain and Sanjeev Jain, "Performance Evaluation of Wavelets for Image Compression".
- [12]Faisal Zubir Quereshi, "Image Compression using Wavelet Transform".
- [13]Kareen Lees, "Image compression using wavelets".
- [14]Ranbeer Tyagi, " Image Compression using DPCM with LMS algorithm" *an international society of thesis publications*.
- [15]Petros T BouFounos, " Universal rate efficient scalar quantization" *IEEE transactions on information theory ,VOL 58, No 3, March 2012*
- [16]Jose Prades Nebot, Edward J.Delp," Generalized PCM coding of images" *IEEE transactions on image processing , VOL 21,N o 8, August 2012*
- [17]Christopher J.C.Burges, Ptrice Y.Simrad ," Improving Wavelet image compression with Neural Networks:
- [18]Chun-Lin, Liu, " A tutorial of the Wavelet Transform".
- [19]S.Sridhar, P.Rajesh Kumar and K.V.Ramanaiah, " An efficient hybrid image coding scheme combining neural networks, wavelets and DPCM for image compression" *International Journal of Computer Applications*.
- [20]Priyanka Singh, Priti Singh," JPEG Image Compression based on Biorthogonal, coiflets and Daubechies Wavelets".

- [21] Mohammed A. Salem, Nivin Ghamry, and Beate Meffert, "Daubechies versus Biorthogonal Wavelets for Moving Object Detection in Traffic Monitoring Systems".

Skin Color Detection Using Region-Based Approach

Rudra PK Poudel

*Media School, Bournemouth University
Poole, BH12 5BB, UK*

rpoudel@bournemouth.ac.uk

Jian J Zhang

*Media School, Bournemouth University
Poole, BH12 5BB, UK*

jzhang@bournemouth.ac.uk

David Liu

*Siemens Corporate Research
755 College Road East, Princeton, NJ 08540, USA*

david-Liu@siemens.com

Hammadi Nait-Charif

*Media School, Bournemouth University
Poole, BH12 5BB, UK*

hncharif@bournemouth.ac.uk

Abstract

Skin color provides a powerful cue for complex computer vision applications. Although skin color detection has been an active research area for decades, the mainstream technology is based on the individual pixels. This paper, which extended our previous work [1], presented a new region-based technique for skin color detection which outperformed the current state-of-the-art pixel-based skin color detection technique on the popular Compaq dataset [2]. Color and spatial distance based clustering technique is used to extract the regions from the images, also known as superpixels followed by a state-of-the-art non-parametric pixel-based skin color classifier called the basic skin color classifier. The pixel-based skin color evidence is then aggregated to classify the superpixels. Finally, the Conditional Random Field (CRF) is applied to further improve the results. As CRF operates over superpixels, the computational overhead is minimal. Our technique achieved 91.17% true positive rate with 13.12% false negative rate on the Compaq dataset tested over approximately 14,000 web images.

Keywords: Skin Color Detection, Bayes Classifier, Superpixels, MRF.

1. INTRODUCTION

Skin color provides a powerful cue in complex computer vision applications such as hand tracking, face tracking, and pornography detection. Skin color detection is computationally efficient yet invariant of rotation, scaling and occlusion. Which are the major reasons for its popularity. The main challenges of skin color detection are illumination, ethnicity background, make-up, hairstyle, eyeglasses, background color, shadows and motion [3]. Most of the skin color detection problems could be overcome by using infrared [4] and spectral imaging [5]. However, such systems are expensive as well as cumbersome to implement. Moreover, there are many situations such as image retrieval on the internet where such systems cannot be used.

Most of the skin color detection techniques are pixel-based, which treat each skin or non-skin pixel individually without considering its neighbors. However, it is natural to treat skin or non-skin as regions instead of individual pixels. Hence, this research focuses on the region-based skin color detection technique. Surprisingly, there are only few region-based skin detection techniques [6], [7], [8] and [9]. Kruppa [7], Yang and Ahuja [6] searched for elliptical skin color shape to find the face. Sebe [9] used fixed 3x3 pixel patches to train a Bayesian network, and Jedyank [8] smoothed the results using hidden Markov model. This paper proposes a new technique purely

based on the concept of regions, irrespective of the underlying geometrical shape. As such, this technique can be easily integrated into any skin detection based system.

Our technique uses a segmentation technique called superpixel [10] and [11] to group the similar color pixels together. Then each superpixel is classified as skin or non-skin by aggregating pixel-using the evidence obtained from a histogram based (also known as non-parametric) Bayesian classifier similar to [2]. However, any suitable pixel-based or superpixel-based skin color classification technique can be used. The result is further improved with Conditional Random Fields (CRF) which operates over superpixels instead of individual pixels. Even though the segmentation cost is an overhead over the pixel-based approach, it greatly reduces the processing cost further down the line, such as smoothing with CRF. Besides aggregation of pixels into regions helps to reduce local redundancy and the probability of merging unrelated pixels [12]. Since superpixels preserve the boundary of the objects, it helps to achieve accurate object segmentation [13].

The presented technique not only outperforms the current state-of-the-art pixel-based skin color detection techniques but also extracts larger skin regions while still keeping the false-positive rate lower (see Table 1 and Figure 2), providing semantically more meaningful skin regions. This could in turn benefit higher-level vision tasks, such as face, hand or human body detection. Related work is discussed in section 2; section 3 presents the proposed region-based skin color detection technique; experiments and results are discussed in section 4. Finally, we summarize our work in section 5.

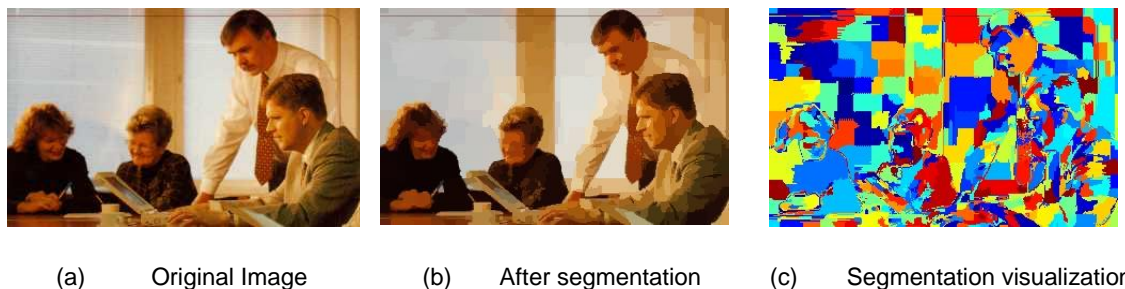


FIGURE 1: An example of superpixel segmentation. A five dimensional vector is used to extract the superpixels: three RGB color channels and two positional coordinates of the pixel in the image.

2. RELATED WORK

Skin color detection has two important parts: one is color space selection and another is color modeling. RGB: [14], [15], [16], [2], [9], HSV: [17], [18], [19], CIE-Lab: [20], [21], YCbCr: [22], [23], and normalized RGB: [16] are popular color spaces, with RGB and HSV being the most frequently used. CIE-Lab uniformly represents the color based on how two colors differ to the human observer. HSV shows better results under varying illumination [3]. Most systems choose RGB color space because the illumination variation can be eliminated by increasing sample size [2]. Due to this reason the RGB color space is chosen in our experiments.

Skin color modeling techniques fall into three categories: explicitly defined skin region [24], non-parametric and parametric techniques. Histogram based Bayes classifier is a popular non-parametric modeling approach. Jones and Rehg [2] used RGB color space and histograms based Bayes classifier and obtained 90% true positive rate with 14.5% false positive rate on unconstrained web images, a dataset made up of approximately 14,000 images. On parametric skin modeling technique, Gaussian mixture has been found to be producing the best result [25], [26]. However, Jones and Rehg [2] showed that given enough samples, the histogram based Bayes classifier technique is slightly better than Gaussian mixture. Neural Network [27], self-organizing map [16], Bayesian network [9] and a few other techniques have also been used for skin color modeling.

This paper presents a region-based skin color detection technique with no prior knowledge on the geometric shape of the skin regions. The works of Yang and Ahuja [6], Kruppa [7], Jedyank [8] and Sebe [9] are the closest to ours. However, Yang and Ahuja [6] used multi-scale segmentations to find elliptical regions for face detection. Hence, their model is biased toward the skin colored elliptical objects. Kruppa [7] also used a similar concept to find the elliptical regions using color and shape information for the face detection. Sebe [9] used 3x3 fixed size pixel patches. Our presented technique uses patches with varying sizes, which is purely based on the image evidence, i.e. skin color in this case. Also, Jedyank [8] used hidden Markov model at pixel level, while we use conditional random fields and operate on superpixel, as described in the section 3.4.

3. PROPOSED FRAMEWORK

We argue that skin is better presented as regions rather than individual pixels. The proposed region-based approach has four major components: basic skin classifier (section 3.1), extraction of regions called superpixels (section 3.2), superpixels classification (section 3.3), and a smoothing procedure with conditional random fields (CRF) (section 3.4). Each step is discussed in detail below.

3.1 Basic Skin Color Classifier

Any good skin color classification technique can be used as a basic skin color classifier. This paper uses the histogram based Bayesian classifier similar to that of Jones and Rehg [2], a state-of-the-art skin color detection technique.

Learning Skin and Non-Skin Histograms: Densities of skin and non-skin color histograms are learned from the Compaq dataset [2]. The Compaq skin color dataset has approximately 4,700 skin images and 9,000 non-skin images collected from free web crawling. It has images from all ethnic groups with uncontrolled illumination and background conditions. The number of manually labeled pixels is nearly 1 billion. Skin and non-skin histograms are obtained in RGB color space with 32 bins for each color channel, exactly the same to the settings as in Jones and Rehg [2]. Equal numbers of skin images are randomly selected for training and testing. Similarly, equal numbers of non-skin images are randomly selected for training and testing.

Bayesian Skin Classifier: Naive Bayes is used to build the skin and non-skin classifier. The probability of a color being skin s given a color c , $P(s|c)$, is given by

$$P(s|c) = \frac{P(c|s)P(s)}{P(c)} \quad (1)$$

where, $P(c|s)$ is the likelihood of a given color c being skin, $P(s)$ is skin color prior and $P(c)$ is color prior. Similarly, the probability of a color being non-skin ns given a color c is given by

$$P(ns|c) = \frac{P(c|ns)P(ns)}{P(c)} \quad (2)$$

where, $P(c|ns)$ is the likelihood of a given color c being non-skin and $P(ns)$ prior for non-skin color. Further $P(c)$ could be calculated as following

$$P(c) = P(c|s)P(s) + P(c|ns)P(ns) \quad (3)$$

$P(c|s)$ and $P(c|ns)$ are directly calculated from skin and non-skin histograms. Prior probabilities: $P(s)$ and $P(ns)$ can also estimate from the total number of skin and non-skin samples in the training dataset. However, for skin and non-skin classification, we can simply compare $P(s|c)$ to $P(ns|c)$. Using equations (1) and (2), the ratio of $P(s|c)$ to $P(ns|c)$ can be simplified to

$$\frac{P(s|c)}{P(ns|c)} = \frac{P(c|s)P(s)}{P(c|ns)P(ns)} \quad (4)$$

Equation (4) can be threshold to produce a skin and non-skin classification rule. Further, P(s) and P(ns) are also constant so this can be simplified as follows

$$\frac{P(c|s)}{P(c|ns)} > \theta \quad (5)$$

where, θ is a constant threshold value.

In the experiments, equation (5) is used to find the skin and non-skin probability for pixels. The values of P(c|s) and P(c|ns) are directly looked-up from normalized skin and non-skin histograms respectively.

3.2 Superpixels

A region or a collection of pixels is called a superpixel. A five dimensional vector is used to extract the superpixels, three RGB color channels and two positional coordinates of the pixel, using the quick shift [28] image segmentation algorithm. Superpixels generated from this approach vary in size and shape. Hence the number of superpixels in each image is highly dependent upon the complexity of the image. An image with low color variation will have a less number of superpixels than an image with high color variation, as there is no penalty for boundary violation. Generally, the concept of boundary is not used when extracting the superpixels, however different objects have different texture or color which will implicitly act as boundaries. Figure 1 shows an example of superpixels of an image. In our work we have used "the Superpixel extraction library" [29] for superpixel segmentation.

3.3 Superpixel Classification

First, the pixel based skin color classifier defined on section 3.1 is used to classify the pixels of the images. Then the probability of being skin for a given superpixel sp with N number of color pixels c is defined as follows

$$P(s|sp) = \frac{1}{N} \sum_{i=1}^N P(s|c_i) \quad (6)$$

Similarly, the probability of being non-skin for a given superpixel sp with N number of color pixels c is defined as follows

$$P(ns|sp) = \frac{1}{N} \sum_{i=1}^N P(ns|c_i) \quad (7)$$

3.4 Smoothing with CRF

Skin regions have varying size and shape, depending upon the camera angle, distance from the camera and human body factors. Hence, to obtain smooth skin regions but still preserve the skin and non-skin boundaries, it is necessary to introduce some constraints. Conditional Random Field (CRF) provides a natural way of combining pairwise constraints. Color difference and length of boundary between adjacent superpixels are used as pairwise constraints similar to Fulkerson [13]. Optimum skin and non-skin labeling L of all superpixels S of an image is defined as follows

$$-\log(P(L|S; \omega)) = - \sum_{s_i \in S} \Psi(l_i|s_i) + \omega \sum_{(s_i, s_j) \in E} \Phi(c_i, c_j | s_i, s_j) \quad (8)$$

where ω is the weight of pairwise constraint, E is the set of edges of superpixel, and i and j are index nodes in superpixel level graph of an image.

Color potential ($\Psi(l_i|s_i)$): The color potential Ψ captures the skin and non-skin probability of the superpixel s_i . We have used skin and non-skin probability for superpixel directly from superpixel classification defined in the section 3.3 for color potential Ψ as follows

$$\Psi(l_i|s_i) = \log (P(l_i|s_i)) \tag{9}$$

Edge and boundary potential ($\Phi(c_i, c_j|s_i, s_j)$): Pairwise edge and boundary potential Φ is defined similar to those of [13]

$$\Phi(c_i, c_j|s_i, s_j) = \left(\frac{L(s_i, s_j)}{1 + \|s_i - s_j\|} \right), [c_i \neq c_j] \tag{10}$$

Where, $L(s_i, s_j)$ is the shared boundary length and $\|s_i - s_j\|$ is Euclidean norm of the color difference between s_i and s_j superpixels.

Only one pairwise potential is used to make the system as simple as possible to show that treating skin color with regions is more effective than with pixels. To improve the effectiveness of our skin color detection technique, we could add more pairwise potentials similar to those in Shotton [30]. This implementation has only one weighting factor ω , which is optimized using cross validation. We use the multi-label graph optimization library of [31], [32] and [33] for the inference of skin and non-skin regions. CRF graph is built on the superpixel level hence CRF optimization is fast.

4. EXPERIMENTS AND RESULTS

Method	True Positive	False Positive
Jones and Rehg (2002)	90%	14.2%
Our (Superpixel only)	91.44%	13.73%
Our (superpixel and CRF)	91.17%	13.12%

TABLE 1: The Results of Pixel-based and Region-based Techniques.

Equal numbers of training and testing sets are randomly chosen from the Compaq dataset [2] and same training and testing sets are used for all experiments. The Compaq dataset has approximately 4,700 skin and 9,000 non-skin images, freely collected from the web. Basic pixel-based skin color classifier mentioned in section 3.1 achieves similar results to those in Jones and Rehg [2]. We have used RGB bin size = 32 for each channels, and threshold constant $\theta = 1$. It roughly detects 90% skin color with 14.2% false positive rate.

Superpixel extraction using quick shift is controlled by three parameters: (i) λ the tradeoff between spatial and color consistency, (ii) σ the deviation of density estimator, and (iii) τ maximum distance in the quick shift tree. We have used $\sigma = 2$, $\tau = 6$, and $\lambda = 0:9$ for our experiment. Which are chosen using grid search as there is no explicit mechanism to preserve the skin boundaries; with above selected parameters we have noticed that 97.43% skin pixels are correctly grouped into superpixels with 0.35% false positive rate. Average size of the superpixels increases with the larger value of τ and σ and vice versa. Lower value of λ gives importance to the spatial factor while higher value gives importance to the color value. Average size of superpixels are larger when λ is around 0.5. Skin color detection depends upon the values of the color channels, hence higher importance is given to the color consistency in superpixel extraction. Besides experiments show that the skin boundary is not well preserved with higher spatial importance. The average size of superpixel is 65 in our experiments. However, the size of superpixels is not fixed and fully depends on the complexity of the images.



FIGURE 2: Comparison between pixel-based [2] and region-based skin color classification techniques. Left column shows the original images. Middle column shows the result of pixel-based classification technique and right column shows the result of region-based classification technique with CRF.

Table 1 shows the results for both the presented region-based technique and the current state-of-the-art pixel-based skin color detection [2] on unconstrained illumination and background. The region based technique without CRF has 91.44% true positive rate with 13.73% false positive rate and with CRF has 91.17% of true positive rate with 13.12% false positive rate. Simply grouping the pixel-based evidence onto superpixels increased the true positive rate by 1.44% and decreased the false positive rate by 0.48%. This shows treating skin as a region yields better results than using pixels only. Both results from the region-based techniques are better than the pixel-based technique.

The results on figure 2 show the effectiveness of the region-based technique with CRF over pixel-based technique. Region-based technique first groups the skin and non-skin evidence from each pixel into superpixels level using basic skin color classifier, which helps to remove noise. This is the main reason why only grouping the pixel-based evidence into superpixels increases the true positive rate by 1.44% and reduces the false positive rate by 0.5% (see table 1). Moreover, CRF helps further extract larger smooth skin regions by exploiting neighboring color information and boundary sharing between superpixels.

However, there are also some cases where region-based technique performs worse than pixel-based technique when we apply the CRF. Figure 4 and Figure 3 are such examples. Skin-like looking pixels and high boundary sharing between skin and non-skin regions are the main reason of the failure. However, we also experimented using the color difference constraint only on CRF instead of both color difference and boundary sharing constraints and found that it performs better when skin regions are very small and narrow. But overall CRF with both neighbor color difference and length of boundary sharing constraints performed better. Figure 5 shows an example where CRF with both neighbors color difference and length of boundary sharing performs better than only with neighbors color difference.

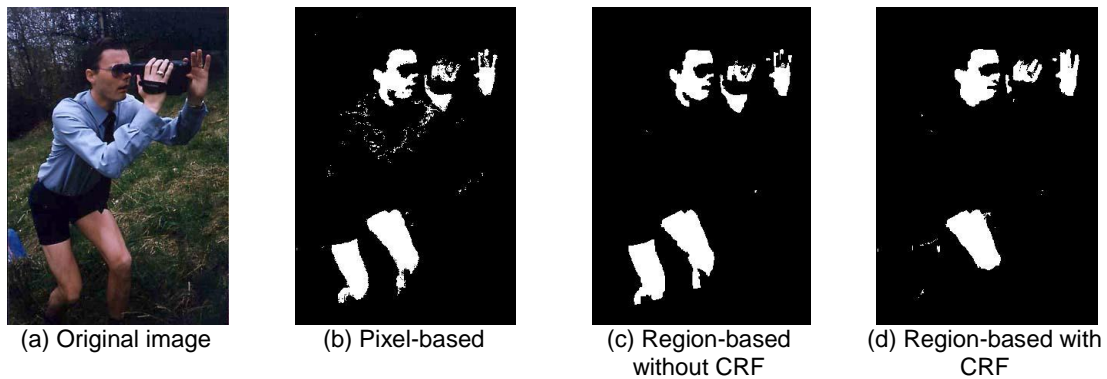


FIGURE 3: This example shows the advantages of the region-based approach even without CRF (see sub figures b and c). Sub figures c and d show the failure case when CRF is applied.

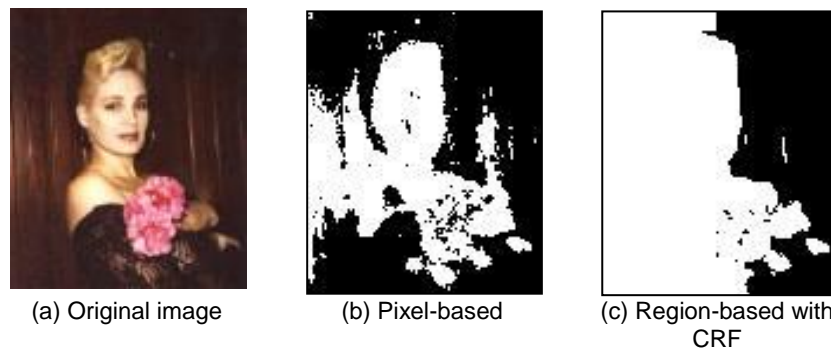


FIGURE 4: This example shows the failure of the region-based approach when border information is applied in CRF smoothing.

5. CONCLUSIONS AND FUTURE WORKS

This paper presented a region-based skin color detection technique, which outperforms the current state-of-the-art pixel-based skin color detection technique. Color and spatial distance based clustering technique is used to extract the regions from the images, also known as superpixels. In the first step, our technique uses the state-of-the-art non-parametric pixel-based skin color classifier [2] which we call the basic skin color classifier. The pixel-based skin color evidence is then aggregated to classify the superpixels. Finally, the Conditional Random Field (CRF) is applied to further improve the results. As CRF operates over superpixels, the computational overhead is minimal.

The proposed region-based technique achieved 91.44% true positive rate with 13.73% false positive rate without CRF optimization and 91.17% true positive rate with 13.12% false positive rate with CRF optimization. Grouping the pixel-based evidence into superpixels increased the true positive rate by 1.44% and reduced the false positive rate by 0.48%. Moreover, the region-based approach produced smoother results than the pixel-based techniques. Skin commonly appears as regions of similar pixels, so treating skin as a region is advantageous over treating it as an individual pixel. Due to the illumination, background reflection and other noise factors, pixel values vary greatly and grouping them into a region helps to remove noise by collecting evidence from neighboring pixels.

These results suggest that skin color detection should be region-based rather than pixel-based. Further, by adding more constraints on the CRF similar to [30], the detection rate can be

improved. Moreover, any better skin color classification technique can be used as our basic skin color classification module and can be easily combined with our region-based skin color detection framework defined in section 3 to improve the results.

Skin regions do not have the same color values; even the closest skin color pixels within superpixels have different color values. Also, other skin-look-like objects exist. Hence, results can be further improved using texture information. This is left for our future work.

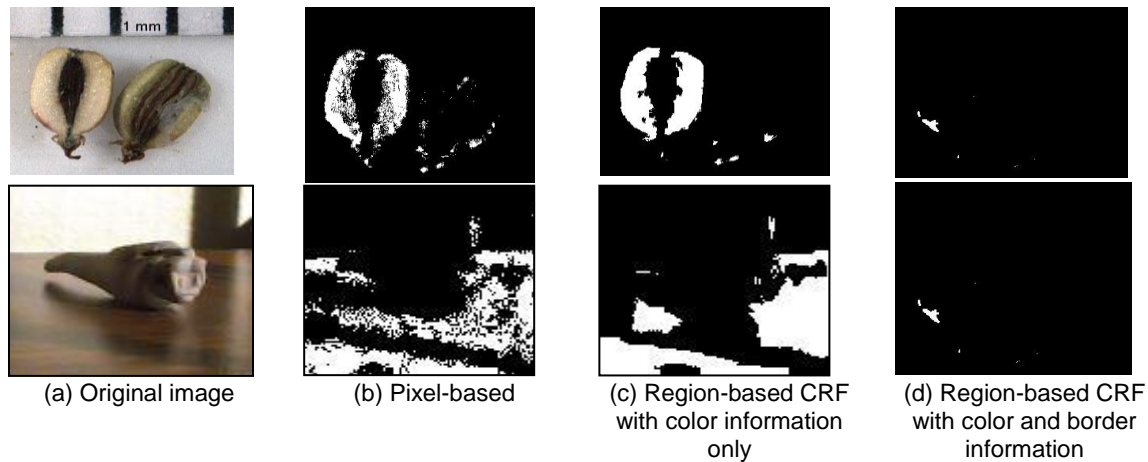


FIGURE 5: Example shows the failures of region-based approach when only a color difference constraint is used on CRF optimization.

6. REFERENCES

- [1] R. PK Poudel, H. Nait-Charif, J.J. Zhang, D. Liu. "Region-Based Skin Color Detection", in VISAPP, 2012.
- [2] M.J. Jones, J.M. Rehg. "Statistical color models with application to skin detection". *International Journal of Computer Vision* 46 (2002) 81–96.
- [3] P. Kakumanu, S. Makrogiannis, N. Bourbakis. "A survey of skin-color modelling and detection methods". *Pattern Recognition* 40 (2007) 1106–1122.
- [4] D.A. Socolinsky, A. Selinger, J.D. Neuheisel. "Face recognition with visible and thermal infrared imagery". *Computer Vision and Image Understanding* 91 (2003) 72–114.
- [5] Z. Pan, G. Healey, M. Prasad, B. Tromberg. "Face recognition in hyperspectral images". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (2003) 1552–1560.
- [6] M.H. Yang, N. Ahuja. "Detecting human faces in color images". In *International Conference on Image Processing, 1998. Volume 1 (1998)* 127–130.
- [7] H. Kruppa, M. Bauer, B. Schiele. "Skin patch detection in real-world images". In Van Gool, L., (Ed.) : *Pattern Recognition. Volume 2449 of Lecture Notes in Computer Science*. Springer Berlin/Heidelberg (2002) 109–116.
- [8] B. Jedynek, H. Zheng, M. Daoudi. "Maximum entropy models for skin detection". In *Energy Minimization Methods in Computer Vision and Pattern Recognition (2003)* 180–193.

- [9] N. Sebe, I. Cohen, T. Huang, T. Gevers. "Skin detection: A Bayesian network approach". In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK (2004) 903–906.
- [10] A.P. Moore, S. Prince, J. Warrell, U. Mohammed, G. Jones. "Super pixel lattices". In IEEE Conference on Computer Vision and Pattern Recognition. (2008).
- [11] X. Ren, J. Malik. "Learning a classification model for segmentation". In IEEE International Conference on Computer Vision. Volume 1 (2003).
- [12] S. Soatto. "Actionable information invasion". In Proceedings of the International Conference on Computer Vision. Volume 25 (2009).
- [13] B. Fulkerson, A. Vedaldi, S. Soatto. "Class segmentation and object localization with super pixel neighbourhoods". In Proceedings of International Conference on Computer Vision. Volume 5 (2009).
- [14] J.L. Crowley, F. Berard. "Multi-modal tracking of faces for video communications". In Computer Vision and Pattern Recognition, Published by the IEEE Computer Society (1997).
- [15] L.M. Bergasa, M. Mazo, A. Gardel, M.A. Sotelo, L. Boquete. "Unsupervised and adaptive gaussian skin-color model". Image and Vision Computing 18 (2000) 987–1003.
- [16] D. Brown, I. Craw, J. Lewthwaite. "A som based approach to skin detection with application in real time systems". In Proceedings of the British Machine Vision Conference. Volume 2. (2001) 491–500.
- [17] Q. Huynh-Thu, M. Meguro, M. Kaneko. "Skin-color extraction in images with complex background and varying illumination". In Proceedings of IEEE Workshop on Applications of Computer Vision, IEEE (2002) 280–285.
- [18] Y. Wang, B. Yuan. "A novel approach for human face detection from color images under complex background". Pattern Recognition 34 (2001) 1983–1992.
- [19] Q. Zhu, K.T. Cheng, C.T. Wu, Y.L. Wu. "Adaptive learning of an accurate skin-color model". In Sixth IEEE International Conference on Automatic Face and Gesture Recognition. (2004) 37–42.
- [20] J. Cai, A. Goshtasby. "Detecting human faces in color images. Image and Vision Computing 18 (1999) 63–75.
- [21] S. Kawato, J. Ohya. "Automatic skin-color distribution extraction for face detection and tracking". In International Conference on Signal Processing. Volume 2., IEEE (2002) 1415–1418.
- [22] R.L. Hsu, M. Abdel-Mottaleb, A.K. Jain. "Face detection in color images". IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 696–706.
- [23] K.W. Wong, K.M. Lam, W.C. Siu. "A robust scheme for live detection of human faces in color images". Signal Processing: Image Communication 18 (2003) 103–114.
- [24] P. Peer, J. Kovac, F. Solina. "Human skin colour clustering for face detection. In International Conference on Computer Tool. (2003).

- [25] M.H. Yang, N. Ahuja. "Gaussian mixture model for human skin color and its application in image and video databases. In Proceedings of SPIE: Storage and Retrieval for Image and Video Databases VII. Volume 3656., Citeseer (1999) 458–466.
- [26] J.C. Terrillon, H. Fukamachi, S. Akamatsu, M.N. Shirazi. "Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In Fourth IEEE International Conference on Automatic Face and Gesture Recognition (2000) 54.
- [27] S.L. Phung, D. Chai, A. Bouzerdoum. "A universal and robust human skin color model using neural networks". In Proceedings of International Joint Conference on Neural Networks Volume 4. (2002) 2844–2849.
- [28] A. Vedaldi, S. Soatto. "Quick shift and kernel methods for mode seeking". Proceedings of European Conference on Computer Vision (2008) 705–718.
- [29] A. Vedaldi, B. Fulkerson. "VLFeat: An open and portable library of computer vision algorithms". <http://www.vlfeat.org> (2008).
- [30] J. Shotton, J. Winn, C. Rother, A. Criminisi. "Text on boost: Joint appearance, shape and context modelling for multi-class object recognition and segmentation. Proceedings of European Conference on Computer Vision (2006) 1–15.
- [31] Y. Boykov, O. Veksler, R. Zabih. "Fast approximate energy minimization via graph cuts". IEEE Transactions on Pattern Analysis and Machine Intelligence (2001) 1222–1239.
- [32] Y. Boykov, V. Kolmogorov. "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision". IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 1124–1137.
- [33] V. Kolmogorov, R. Zabih, R. What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 147–159.

Recognition of Offline Handwritten Hindi Text Using SVM

Naresh Kumar Garg
GZSPTU Campus, CSE Department
Bathinda-151001, India

naresh2834@rediffmail.com

Dr. Lakhwinder Kaur
UCOE, Punjabi University,
Patiala, India

mahal2k8@yahoo.com

Dr. Manish Jindal
Punjab University Regional Centre,
Muktsar, India

manishphd@rediffmail.com

Abstract

Handwritten Hindi text recognition is emerging areas of research in the field of optical character recognition. In this paper, a segmentation based approach is used to recognize the text. The offline handwritten text is segmented into lines, lines into words and words into character for recognition. Shape features are extracted from the characters and fed into SVM classifier for recognition. The results obtained with the proposed feature set using SVM classifier is very challenging.

Keywords: Handwritten Hindi Text, Segmentation, Shape Based Features, Recognition Rate, SVM Classifier.

1. INTRODUCTION

Devanagari is the script for writing Hindi language. Hindi is the official language of India. Offline handwritten Hindi text recognition is need of the hour due to large number of application of Hindi OCR. Development of handwritten OCR is very difficult due to different writing styles of the individuals. The techniques developed for recognition of printed characters can not be directly applied on Handwritten text. Due to large number of characters and presence of half characters makes the recognition process even more complex.

There are mainly two approaches for recognition of text- Holistic approach and segmentation based approach. Due to different writing styles of writers and various shapes of characters it is very difficult to use the holistic approach. We have used the segmentation based approach to develop the recognition system for handwritten Hindi text.

Further the paper is divided into following sections- section 2 discussed the related work, section 3 explains the database taken for experimental work, section 4 is about proposed technique used for the recognition of handwritten Hindi text, section 5 discusses the results and last section is about future scope. References are given at the end of the paper.

2. PRELIMINARIES

A lot of work has been done in the past on recognition of printed Hindi text and Hindi numeral recognition. A few research reports are available in the field of handwritten text recognition. Most of the work done in handwritten Hindi text recognition is on recognition of isolated characters. To the best of author's knowledge, no commercial OCR for handwritten Hindi text is available, yet.

A good survey about OCR is given in [1]. The performance of any classifier depends upon the quality of features fed into it. A very good survey about recognition of Devanagari script is given in [2]. It is mentioned in this paper that a lot of research has been done in the past in the recognition of printed text and isolated characters of handwritten Devanagari text, but only few research reports are available on recognition of handwritten text. Work on recognition of printed devanagar text is explained by veena bansal in [3].

A good survey about feature extraction is given in [4]. Trier et al. [5] present an interesting survey of feature extraction method for off-line recognition of segmented characters. The authors describe important aspects that must be considered before selecting a specific feature extraction method.

To the best of author's knowledge, no commercial OCR for handwritten Hindi text is available, yet. The structural and statistical features are very useful for character recognition [6].

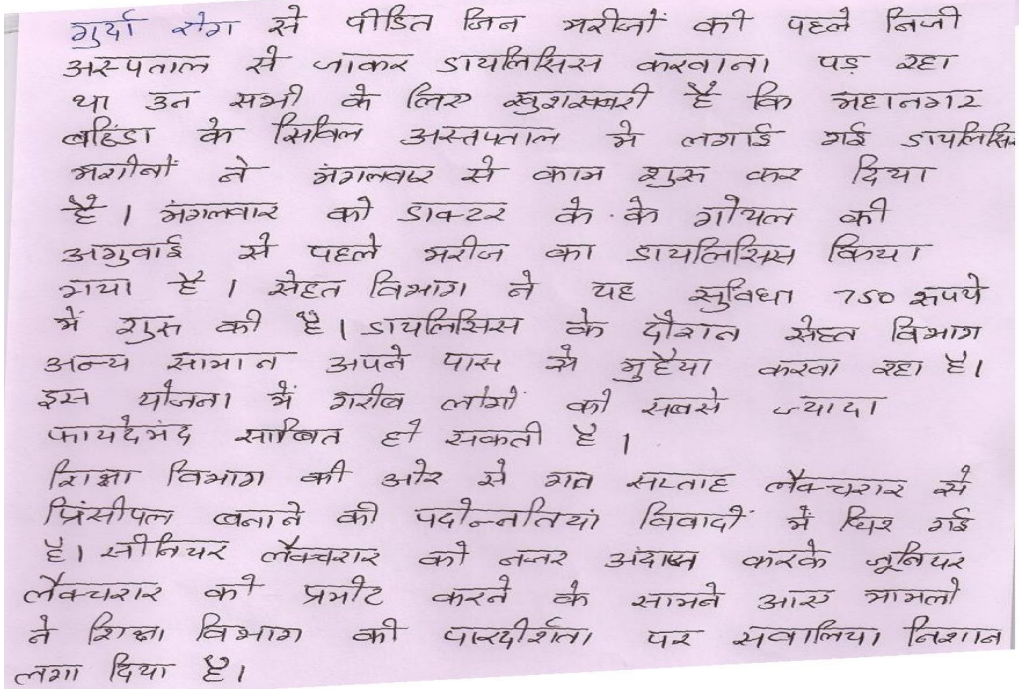
In [7], Hanmandlu et al. had used Fuzzy model based techniques for recognition of Handwritten Hindi Characters and the recognition rate of 90.65% was reported at character level.

In [8], Kumar and Singh had used Zernike moments for recognition of Devnagari handwritten characters and reported recognition rate of 80%. Shaw et al.[9] worked on recognition of handwritten devnagari words using segmentation approach.

The work on line segmentation, consonant segmentation, upper modifier segmentation and lower modifier segmentation in Handwritten Hindi text were explained by us in [10, 11]. The algorithm for segmentation of Half characters in handwritten Hindi text is explained in [12]. We have explained a method based on structural features for segmentation of half characters in handwritten Hindi text. Recognition of non compound handwritten Devanagari characters using MLP and minimum edit distance is explained in [13].

3. DATABASE

All experiments were conducted on database constructed by taking handwritten data from fifteen writers. Documents are scanned at 300 dpi. The handwritten documents were reduced in size in paint to 35% to increase the speed of execution. The percentage of stretching of the document in horizontal and vertical direction was same. The sample database is shown in figure 1.



गुर्या रोग से पीड़ित जिन मरीजों की पहले निजी अस्पताल से जाकर डायलिसिस करवाना पड़ रहा था उन सभी के लिए सुझावकारी है कि महानगर बहिडा के सिविल अस्पताल में लगई गई डायलिसिस मशीनों ने अंगल्वर से काम शुरू कर दिया है। अंगल्वर को डाक्टर के के गौयल की अंगुवाई से पहले मरीज का डायलिसिस किया गया है। सेहत विभाग ने यह सुविधा 750 रुपये में शुरू की है। डायलिसिस के दौरान सेहत विभाग अन्य सामान अपने पास में जुड़ेया करवा रहा है। इस योजना में मरीज लोगों की सबसे ज्यादा फायदेमंद साबित हो सकती है।

विज्ञान विभाग की ओर से गत सप्ताह लेक्चरर का प्रिंसीपल बनाने की पदोन्नतियों विवादी में फिर गई है। सीनियर लेक्चरर को जूनर अंदाज करके जूनियर लेक्चरर को प्रमोट करने के सामने आरंभ मामलों ने शिक्षा विभाग की पारदर्शिता पर सवालिया निशान लगा दिया है।

FIGURE 1: Sample Database.

4. PROPOSED TECHNIQUE

Handwritten Hindi text written by different persons was scanned and binarized in Matlab. Segmentation of the text was performed in the following sequence:-

1. Text was segmented into lines.
2. Lines were segmented into words.
3. Upper modifiers were segmented from words.
4. Lower modifiers were segmented from words.
5. Consonants, half characters, matras and joint characters were segmented from words.

The techniques used for segmentation was explained in [6][7]. The strip wise vertical projection method was used for line segmentation. Word segmentation was done using vertical projection method. For character segmentation after upper and lower modifier removal from the word, a header line was detected again for each word and then vertical projection along with other constraints for joining characters were used for segmentation. Segmentation of text is very tedious task. The segmentation error propagates to recognition and reduces the recognition rate. Holistic approach was not used due to heavy character set and large number of compound characters available in handwritten Hindi text.

After segmentation, feature extraction was another tedious task performed on each character. The recognition rate of characters mainly depends upon the correctness of the features used for recognition. The efforts were made on the correctness of the features. The shape based features were extracted by applying many heuristics depending upon the shape of the character for each feature. The programming was done to extract each feature by applying many heuristics to make the feature unique for each character.

Total 59 features are selected to make a unique feature set for recognition of handwritten Hindi text. After carefully analyzing the characters set of Hindi language, different features are selected. Feature set include bars (End bar, Middle bar), end points, loops, crossings, presence of

particular horizontal and vertical lines, groves, curves and projection profiles in front, back, bottom and top of the character.

Many heuristics are applied in the extraction of each feature. The heuristics are applied to differentiate similar shaped characters. Some of the similar shaped characters are given in table 2.

Total 41 characters are considered for character recognition. These are most commonly used characters of handwritten Hindi text shown in figure 2.

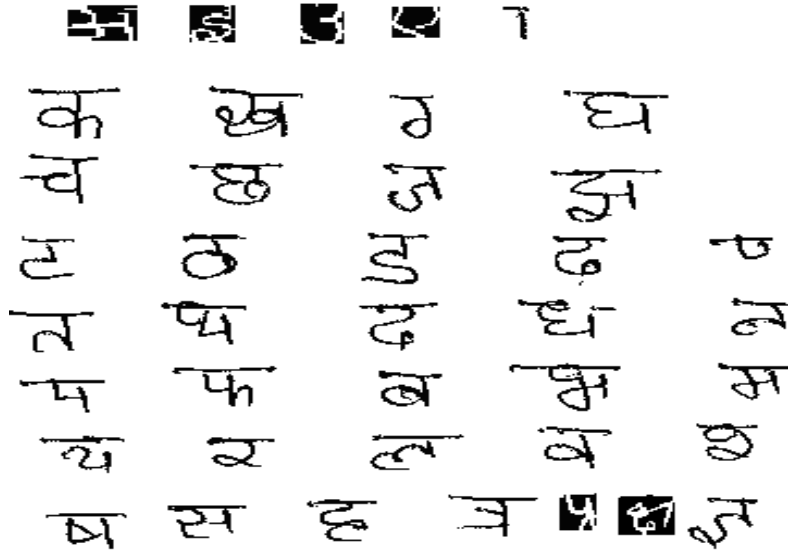


FIGURE 2: Most Commonly Used Characters.

5. RESULTS

The results obtained with shape based features and SVM classifiers are given in table 1.

Ten samples of each character are used for training the classifier. For 41 characters 410 samples are used for training purpose. All the other characters are used for testing purpose.



























TABLE 1: Recognition Accuracy of Characters.







No of characters	% of characters correctly recognized (including segmentation errors)	% Accuracy of Characters recognized from correctly segmented characters
2016	76.4	89.6



The errors in segmentation propagate to recognition. The overall recognition rate is less due to segmentation errors. The recognition rate obtained from correctly segmented characters is 89.6%, which is very promising. Some of the similar shaped characters which create confusion during recognition are given in table 2.

Till now most of the work is done on recognition of isolated characters. The feature set for isolated characters can not be directly applied on the handwritten text.

TABLE 2: Similar Shaped Characters.

S No.	Character	Confused with
1.	 p	 Jai
2	 k	 f
3	 l	 t
4	 r	 sh
5	 r	 tt
6	 a	 m
7	 adh	 e
8	 th	 dh
9	 s	 m
10	 j	 n
11	 ch	 b
12	 d	 b
13	 s	 kh

Similar characters like r , g  and sh  are very much confusing and difficult to recognize. They can be recognized with the help of complete word only. Also characters ch , jai  and p  are very much confusing due to different writing style used by the

different writer's. Characters  and  are very similar in shape. If the upper left loop of character 'bhh' is very small and merges with the character than it looks like character 'm'. Shapes of these characters are very similar and minor differences in shapes are difficult to detect even with human eye. These types of problems can be solved during post processing stage.

The obtained results can not be compared with the literature work because most of the work available in literature is on recognition of isolated characters. The results of recognition of handwritten text can not be compared with the results of recognition of isolated characters due to non availability of standard database for handwritten Hindi text. The results obtained in our work are still comparable with results of recognition of isolated handwritten Hindi characters.

6. DISCUSSION AND FUTURE SCOPE

From the results it is clear that shape based features and SVM classifier are very useful to develop an OCR for handwritten Hindi text. The segmentation errors affect the recognition rate. The similar shaped characters creates problem in recognition. The post processing can reduce the errors in recognition that occur due to similar shaped characters and improve the recognition rate. The efforts can be made in the future in the following direction:

- 1) Segmentation techniques can be improved to reduce the segmentation errors and recognition rate.
- 2) More features can be added in the feature set to differentiate similar shaped characters.
- 3) Other classifiers can be tried with shape based features.

7. REFERENCES

- [1] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR Research and development", Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058, 1992.
- [2] R. Jayadevan, S.R. Kohle, P.M. Patil and U. Pal, "Offline recognition of Devanagari script: A survey." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions*, Vol.41, No. 6, pp:782-796, 2011.
- [3] V. Bansal, "Integrating knowledge sources in Devanagari text recognition", Ph.D. thesis, IIT Kanpur, INDIA, 1999.
- [4] N. Arica and F. T. Y. Vural, "An overview of character recognition focused on offline handwriting", *IEEE Trans. On Systems, Man, and Cybernetics – Part C: Applications and Reviews*, Vol. 31, No. 2, pp. 216-233, 2001.
- [5] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition: A survey", *Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.
- [6] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier, and C. Olivier, "A structural/statistical feature based vector for handwritten character recognition", *Pattern Recognition Letters*, Vol. 19, No.7, pp:629-641, 1998.
- [7] M. Hanmandlu, O.V. Ramana Murthy and Vamsi Krishna Madasu, "Fuzzy Model based recognition of handwritten Hindi characters", *Digital Image Computing Techniques and Applications*, pp:454-461, 2007.

[8] S. Kumar and C. Singh, "A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition", International Conference on Cognition and Recognition, pp:514-520, 2005.

[9] B. Shaw, S. K. Parui, and M. Shridhar, "A segmentation based approach to offline handwritten Devanagari word recognition," Proceedings of IEEE International Conference on Information Technology, pp: 256–257, 2008.

[10] N. K. Garg, L. Kaur and M. K. Jindal, "Segmentation of Handwritten Hindi Text", *International Journal of Computer Applications (IJCA)*, Vol. 1, No. 4, pp.22-26, 2010.

[11] N. K. Garg, L. Kaur and M. K. Jindal, "A new method for line segmentation of Handwritten Hindi Text", Proceedings of the 7th International IEEE Conference on Information Technology: New Generations (ITNG), pp.392-397, 2010.

[12] N. K. Garg, L. Kaur and M. K. Jindal, "The Segmentation of Half Characters in Handwritten Hindi Text", Proceedings of the ICISIL 2011, Springer, pp.48-53, 2011.

[13] Sandhya Arora et al. "Recognition of Non-Compound Handwritten Devanagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue 1, pp. 107-120, 2010.

Elliptic Fourier Descriptors in the Study of Cyclone Cloud Intensity Patterns

Ishita Dutta

*Department of Computer Science and Engineering
West Bengal University of Technology
Kolkata 700064, India*

idutta_kalyani@yahoo.co.in

S. Banerjee

*Department of Natural Science
West Bengal University of Technology
Kolkata 700064, India*

sreeparnab@hotmail.com

Abstract

Cyclone cloud intensity analysis is conducted to study the evolution of a cyclone storm mainly using two approaches, namely: wind field analysis and pattern recognition. Of the pattern recognition based approaches, the Dvorak technique has been a pioneering effort which is widely used today. However, the Dvorak technique is subjective, as it relies on human judgment and is, therefore, error prone. Efforts have been described in the literature to automate the classification process. In this paper, we describe our efforts to perform a semi-automatic computer analysis of the cyclone cloud intensity evolution pattern which compares preprocessed visible (VIS) and enhanced infra-red (EIR) satellite images with the corresponding prototype Dvorak patterns using Elliptic Fourier Descriptors (EFD) and Principal Component Analysis (PCA) techniques. This novel approach is simple and intuitive and is robust to noise, and at the same time provides classification in cases where the cyclone exhibits fluctuations during its evolutionary cycle.

Keywords: Cyclone Cloud Intensity, Dvorak Technique, Elliptic Fourier Descriptors, PCA, Spiral Band.

1. INTRODUCTION

A Tropical Cyclone (also known as hurricane / typhoon) is an area of low atmospheric pressure characterized by rotating and converging winds and ascending air, with the central core being warmer than the surrounding atmosphere. The evolution of the cyclone is manifest as changing cloud intensity patterns, with the development of a central eye at an early stage of its evolution surrounded by spiraling cloud bands of varying intensity. The cyclone then intensifies, matures and finally dissipates. At each stage of its evolutionary sequence, the cyclone is characterized by a set of physical parameters related to wind intensity, as well as characteristic cloud intensity patterns obtained from Visible (VIS) and Enhanced Infra-Red (EIR) images obtained using satellite technology. Hence, there are two approaches to TC intensity determination: namely pattern matching and wind field analysis [1,2]. In this note we will describe our attempts in the study of cyclone cloud intensity patterns from VIS and EIR images. A pioneering effort in comprehensive pattern recognition based analysis for estimating tropical cyclone intensity from satellite imagery can be attributed to Dvorak [3] for visible and, later EIR [4] images. The latter set of images wears the additional advantage that intensity estimates can be made at night. Dvorak describes the evolution of the cyclone through the various stages of evolution, viz., formation from a disorganized cloud structure through intensification, maturity and finally dissipation using different cloud intensity templates, as depicted in Figure 1. Satellite images are matched by visual inspection with these templates to obtain a classification using a characteristic number referred to as T-number. There are eight different T-numbers for the different stages.

However, Dvorak's technique [5] is based on human judgement, requires expert training and is thus, subjective. Hence, there have been a number of attempts to automate this analysis procedure over the past three decades. This paper presents a semi-automated approach to classify cyclone images based on Dvorak's technique.

2. PREVIOUS WORK

The pattern recognition based approaches can be broadly classified into three categories. The earlier approaches focussed on the location of the cyclone eye. These include Griffin et. al. [6] who have defined the geometric centre of the TC eye wall. Wood [7] used an axi-symmetric hurricane vertex flow model to locate the ideal TC. Using Doppler velocity data, a TC is identified by locating areas with cyclonic shear, and its center is then located by the identification of extreme Doppler velocity value duplex. However, this method only works well for an ideal TC. Later approaches in this category include those of Tsang et. al. [8] and Pao and Yeh [9]. Tsang et al. [8] suggested morphology operations like erosion and region growing to automatically locate the TC center but their approach is limited to TCs whose main bodies are the most significant characteristics of circle regions. Using infrared and visible time series satellite images, Pao and Yeh [9] have attempted to locate the center of the typhoon and segment it from the background using morphology operations and statistical image classification methods. Some other distinctive features from slices of typhoon satellite cloud images, especially the rotation feature of wind movement vector were also found.

Another set of approaches involved the extraction of the contours of the dominant cyclone and compared with templates either generated from modeling based on the spiral helix equation or from prototype images. The works of Lee et. al. [10], Lee et. al. [11], Lui et. al. [12], Zhang et. al. [13] and Pineros et. al. [14], fall into the category of matching with prototype images. Lee et.al.[10] proposed a neural oscillatory elastic graph matching (NOEGM) model, for automatic TC pattern identification and track mining. The procedure is comprised of three steps, feature extraction, segmentation of cyclone contours using neural oscillators and elastic graph matching. This procedure could not develop a high level data mining and pattern prediction model by the generation of the time dependent relationship of the TC templates based on the past TC cases. Therefore they (Lee et. al.[11]) have proposed another elastic graph dynamic link model (EGDLM) based on the elastic contour matching to automate Dvorak technique. Lui et al. [12] proposed the use of angle features and time warping for TC forecast. The Gradient Vector Flow (GVF) snake model is applied to extract the contour points a tropical cyclone from the satellite image. Similarity among Dvorak templates and the candidate cyclone were retrieved using angle features found among the successive contour points. Zhang et. al. [13] extended the model of artificial ant colony (AAC) to continuous space by aid of multi-kernel Gaussian functions. The whirling shapes of real unclear typhoon eyes are simulated by snake contour boundaries. However, the stability of calculation, the selection of an effective initial Gaussian kernel and its deviation need to be improved. In addition, the performance of Gaussian parameter calculations may limit the number of selected kernels. Pineros et. al. [14] propose a technique using gradient vectors for obtaining features associated with shape and dynamics of cloud structures in cyclones. This method was not able to characterize intensity curves of some systems which exhibit extremely strong oscillations on time frames of 18–20 hours that overwhelm the intensity trend.

The third category, which is an extension of the second category, uses gradient vectors along the contours of the cyclone pattern. Contour matching based on the mathematical modeling of the spiral band of the cyclone is also performed. Wong et. al. [15] modelled the spiral rain-band of a TC by a polar equation given below, in which all vectors are tangents to the logarithmic spiral

$$R=ae^{\theta \cot(\alpha)} \quad (1)$$

where (R, θ) are the polar co-ordinates at any point R is radial coordinate and θ the angular coordinate, a determines the rate of growth of the spiral, and α (pitch angle) is the angle between the radial line and the tangent to the spiral at (R, θ). $\cot(\alpha)$ is the rate of change of R w.r.t. θ per

unit R. Templates generated by the estimated parameters are then used to match against radar images at plausible latitude– longitude positions. By using a genetic algorithm suggested by Yip and Wong [16] this method is automated by Wong et. al. [17]. Such a model may be unsuitable if images are sampled infrequently, or when TCs are rapidly moving. In another paper Wong et. al. [18] introduced a method of finding the centers of circulating and spiraling vector field patterns that can handle vector fields with multiple centers and is robust against noise. However, some vector field parameters must be defined beforehand, which limits the method's applicability. Wei and Jing [19] have also performed optimization of the spiral band model. Also, a novel Spiral Band Model (SBM) is designed to extract and describe the spiral pattern of a spiral band which spirals out from a TC's center.

Although Fourier descriptors have been used extensively for boundary description, matching and recognition, no work appears to have been done in using Fourier Descriptors or Elliptic Fourier Descriptors (EFD) for decomposing cyclone cloud intensity shapes. Abidi and Gonzalez [20] have decomposed time varying shapes associated with cells of tornadic thunderstorms using EFD. These time varying shapes evolve rapidly, in a matter of a few seconds. In this work we only use primary, EIR and VIS patterns. The next sections outline the Methodology, followed by Results and Discussion and finally some Concluding Remarks and Future Work.

3. METHODOLOGY

Satellite images of several cyclones /hurricanes/typhoons spanning over two decades have been used for the study. Of these 252 images, 227 are EIR images and 15 are visible images. As depicted in the flowchart of Figure 2, the input images are first pre-processed followed by shape analysis in the post-processing stage.

3.1 Pre-processing

In the pre-processing step, image enhancement and filtering is applied to obtain high quality images. A median filter is used to remove additive noises. In order to separate the target image (dominant cyclone) from the background, segmentation is performed. The segmentation, binarization, opening and hole filling operations are carried out subsequently, following Guo et. al.'s [21] methodology for galaxy image segmentation. This is done using Otsu's [22] thresholding which separates the predominant cyclone from the background. Otsu's algorithm for automated thresholding is a popular choice in 2D scenes because it is simple to implement, easy to use and gives satisfactory results in 2D when number of pixels in each class are close to each other. A small offset ranging from -0.2 to 0.3 was given in some cases to improve the visual quality of the results. After binarization, an opening operation was performed to remove neighboring cloud disturbances from the dominant cyclone (to be referred to as the region of interest or ROI). A hole filling operation was then performed for boundary extraction. The boundary can then be chain coded using the Freeman code [23] for segmentation purposes. This chain coding procedure is implemented as part of the SHAPE package to be described below.

DEVELOPMENTAL PATTERN TYPES	PRE STORM	TROPICAL STORM		HURRICANE PATTERN TYPES		
		(Minimal)	(Strong)	(Minimal)	(Strong)	(Super)
	T1.5 ± .5	T2.5	T3.5	T4.5	T5.5	T6.5 - T8
CURVED BAND PRIMARY PATTERN TYPE						
CURVED BAND EIR ONLY						
CDO PATTERN TYPE VIS ONLY						
SHEAR PATTERN TYPE						

FIGURE 1: Typical Dvorak templates for cyclone cloud intensity patterns (from [35]).

3.2 Post Processing

The second part of our procedure is the post-processing part which involves shape analysis. For this purpose we use the SHAPE software package developed by Iwata and Ukai [24]. SHAPE is an open-source software that was originally used for the analysis of biological shapes. This package extracts the contour shape from a full color bitmap image, delineates the contour shape with Elliptic Fourier Descriptors and finally performs the Principal Component Analysis (PCA) of the EFDs for summarizing the shape information. The package, SHAPE is open source and easy to use as a researcher can analyze 2 D shapes on a personal computer without special knowledge about procedures related to the method. However, in the following, details of the processes involved will be described below, for completeness. It may also be noted in passing that, SHAPE is characterized by the following features: (1) the packaged programs are easily operated with the aid of a graphical user interface (GUI); (2) No special computer devices for image processing are required; (3) A large number of samples (say 1,000) can be treated; (4) The scores of principal components are stored in tabbed text format files and can be easily exported for analysis by other software; and (5) The variations in shape accounted for by the principal components can be visualized and printed out.

SHAPE essentially performs the following operations on our images: After noise reduction, the closed contour is extracted by edge detection. The contours of the cyclone cloud shapes (candidate images) are then encoded in the Freeman chain code form and then approximated with the coefficients of Elliptic Fourier Descriptors. (EFD). The coefficients of the EFDs are normalized to be invariant with respect to size, rotation and starting point.

Principal Component Analysis (PCA) is then performed to reduce dimensionality, based on the variance-covariance matrix of scores. The scores of derived principal components are also calculated and stored in text format files which were used for quantitative analysis. SHAPE also visualizes shape variations accounted for by each principal component. Reconstructed contours can be printed. The classification is performed with the help of PCA. There are 6 classes of prototype images used for the classification of both for EIR and VIS images, corresponding to T-numbers T1.5, T2.5 etc., increasing in steps of unit T numbers. There are 252 candidate images of which 227 are EIR images which were classified. Since there are at most three to five spiral turns due to the banding effect of the cyclones, we have used five harmonics, although for our results, results converged for three harmonics.

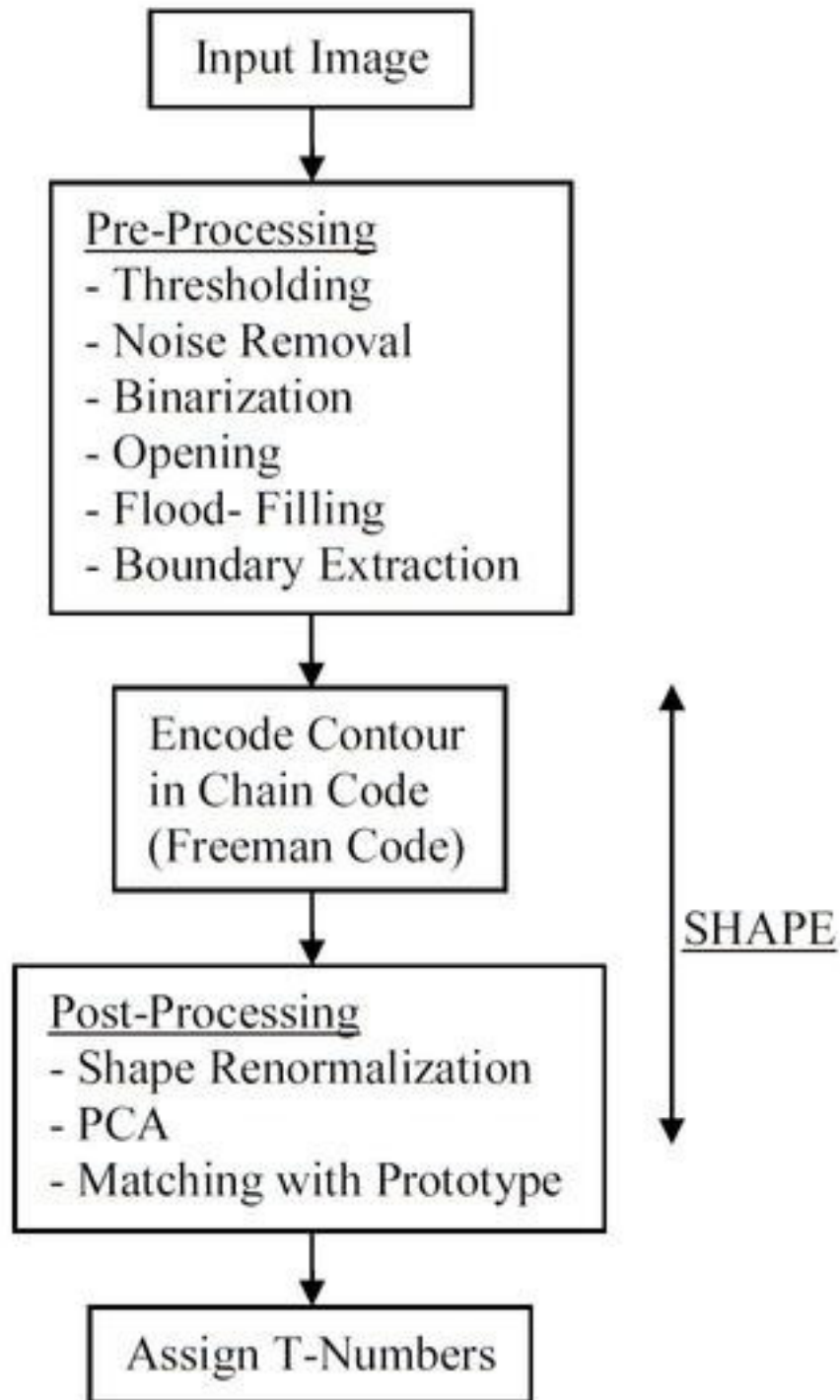


FIGURE 2: Flowchart for the Proposed Algorithm.

Cyclone Name	Number of Images	Number of Images Whose Computed T-numbers are in Agreement with Indian Meteorological Department (IMD)	Percentage (%) Agreement	Comments
Aila	29	28	96.55	Excellent
Bijli	50	35	70	Fairly Good
Phyan	24	19	79.2	Fairly Good
Rashmi	14	10	71.42	Fairly Good
Ward	4	4	100	Excellent

TABLE 1: Comparison of calculated table numbers with IMD T-numbers.

Cyclone Name	Number of Images	Number of Images Whose Computed T-numbers are in Agreement with Cyclone Evolution Trend	Percentage (%) Agreement	Comments
Emily	9	8	89	Very Good
Andrew	4	3	75	Fairly Good
Roxanne	10	7	70	Fairly Good
Nargis	7	6	86	Very Good
Irene	9	8	88	Very Good
Others (Elida, Jeanne, Darby, Flossie)	4	3	75	Fairly Good

TABLE 2: Comparison of Calculated T-numbers with Cyclone Evolution Trend.

A similar procedure is repeated for prototype images corresponding to the Dvorak [3,4] templates. The Euclidean distances between the score value of candidate images we have used and the prototype images are then computed. The best match corresponds to the minimum Euclidean distance between the candidate and prototype image, and the T-number obtained from the prototype image is taken to be the T-number of the candidate image.

After chain coding using the Freeman code [23], SHAPE then approximates the cyclone contours using Elliptic Fourier Descriptors (EFD) proposed by Kuhl and Giardina [25], who claim that there are several advantages of EFDs over standard Fourier descriptors. Firstly, integration and Fast Fourier transforms are not required. Moreover, bounds on the accuracy of image contour reconstruction are easy to specify. In addition, EFDs provide a convenient and intuitively pleasing procedure of normalizing a Fourier contour representation. The steps that have been followed for obtaining EFDs in SHAPE have been outlined below (Yoshioka et.al. [26]).

Each contour is represented as a sequence of x and y coordinates of ordered points that are measured counter-clockwise from an arbitrary starting point. Assume that the contour between the (i-1)-th and the i-th chain coded points is linearly interpolated, and that the length of the contour from the starting point to the p-th point and the perimeter of the contour are denoted by t_p and L, respectively. The quantity t_p is defined as:

$$t_p = \sum_{i=1}^p \Delta t_i \tag{2}$$

and $L = t_k$

Here Δt_i and K are the distances between the $(i - 1)$ -th and the i -th points and the total number of the chain-coded points on the contour, respectively. One point to note is that the K -th point is equivalent to the starting point. The x and y coordinates of the p -th point are

$$x_p = \sum_{i=1}^p \Delta x_i \tag{3}$$

And

$$y_p = \sum_{i=1}^p \Delta y_i \tag{4}$$

where Δx_i and Δy_i are the distances along the x and y axes between the $(i - 1)$ -th and the i -th point. Thus, the elliptic Fourier expansions of the coordinates on the contour are

and

$$y_p = C_0 + \sum_{n=1}^{\infty} \left(c_n \cos \frac{2n\pi t_p}{P} + d_n \sin \frac{2n\pi t_p}{P} \right) \tag{6}$$

with summation $n=1, \dots, \infty$, and a_n, b_n, c_n , and d_n being the Elliptic Fourier coefficients of the n -th harmonic and P being the period. As said earlier the coefficients of an elliptic Fourier descriptor [20],[25], are not invariant in size, rotation, shift and starting point of chain-coding about a contour, the Fourier coefficients are standardized (Yoshioka et al. 2008). Let the standardized coefficients of the n -th harmonic be $a_n^{**}, b_n^{**}, c_n^{**}$ and d_n^{**} . Then,

$$\begin{bmatrix} a_n^{**} & b_n^{**} \\ c_n^{**} & d_n^{**} \end{bmatrix} = \frac{1}{E^*} \begin{bmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{bmatrix} \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix} \tag{7}$$

Where $E^* = [(A_0 - x_q)^2 + (C_0 - y_q)^2]^{1/2}$,

$$\psi = \arctan \left[\frac{y_q - C_0}{x_q - A_0} \right] (0 \leq \psi < 2\pi)$$

and

$$\theta = \frac{2\pi t_q}{T} (0 \leq \theta < 2\pi)$$

In the above equations, E^* is the distance between the centre point (A_0, C_0) and a specific point (x_q, y_q) , and ψ is the spatial rotation angle. These two parameters are for the size invariance and the rotation invariance. θ is a parameter for chain-code starting point invariance. This standardization makes a_n^{**} , b_n^{**} , c_n^{**} and d_n^{**} independent of the size, rotation, shift and chain-code starting point of a contour. The coefficients of the EFDs are thus, subsequently normalized to be invariant with respect to the size, rotation, and starting point, with the procedure based on the ellipse of the first harmonic. The normalized coefficients of the EFDs can still not be used directly as shape characteristics because the number of coefficients is generally very large and the morphological meaning of each coefficient is difficult to interpret separately and so, Principal Component Analysis (PCA) is to be performed.

Principal component analysis is effective for summarizing the information of the variations contained in the coefficients. The scores of the derived principal components are also calculated and stored in text format files, which can be provided as input files for the various subsequent analysis. Then scores of the derived principal components are calculated. Then, score values for the Principal component for candidate images and model images are calculated. Then the Euclidean distance between a particular candidate image and all the model images are calculated and the best match is chosen.. This process was performed both for candidate and prototype images.

4. RESULTS

The images that we have used in our study include image sequences (i.e. images depicting successive stages of evolution of the cyclone) of cyclone Ward, cyclone Aila, cyclone Phyan, as well as images of hurricane Emily, hurricane Andrew, hurricane Roxanne, Nargis and individual images of hurricanes Elida, Jeanne, Darby and Flossie. Some images of the recent hurricane, Irene have also been included in our study.

Input images are first digitized and subsequently Otsu's method is applied with a small offset ranging from -0.2 to 0.3 to threshold the images. Then morphological opening operation is applied with the structuring element of disk shaped with a radius of 60 pixels. This is followed by a hole filling operation to enable boundary extraction. Figure 3 depicts images of cyclone Ward, hurricanes Emily, Andrew and Roxanne (first column) and the images after Otsu thresholding (column 2), morphological opening (column 3) and hole filling (column 4), respectively.

During the post-processing stage, the software package SHAPE extracts contours and assigns chain codes to the contours. Then EFDs are extracted followed by PCA. PCA results of candidate images are compared with PCA results of prototype images corresponding to the Dvorak templates. Best matches correspond to the minimum Euclidean distance between candidate and prototype images. Figure 4 gives principal components for Andrew (EIR T1.5) and Emily (VIS T1.5) with the template images with which they are matched.

Results obtained will be discussed under two categories. In the first category, our results will be compared with the T-numbers estimated by meteorological experts. In the second category, we correlate our results with reports of cyclone evolution trends. In the second set of experiments, T numbers of the cyclones were determined with the algorithm described above and compared with the description of the cyclone evolution given in the reports, since some cyclones have been described with the help of the Saffir Simpson Hurricane Scale (SSHS) a gradation scheme for hurricanes, based on wind field analysis. This set of experiments have been conducted on hurricanes Emily, Andrew, Roxanne, Irene and Nargis image sequences, as well as as individual images of Darbie, Elida, Flossie and Jeanne.

Cyclone Ward [27] formed and intensified to storm status and further intensified to cyclone status before finally weakening due to wind shear and eventually dissipating. Cyclone Nargis [28] formed and quickly intensified to severe storm status before weakening and dissipating. Table 1 gives the comparisons of the T-numbers obtained using our methodology with the data for the

cyclones obtained from the Indian Meteorological Department (IMD). Hurricane Emily [29] formed and intensified to cyclone status and again fluctuated from moderate to severe cyclone storm status before weakening and finally dissipating. Our analysis indicates an increase of T number from 1.5 to 4.5, followed by a subsequent decrease to 3.5, an increase to 6.5 and finally a fairly stable period with T number of 5.5, followed by another increase to 6.5 and finally, a gradual decrease. This fluctuation corroborates with the report [29]. Hurricane Andrew [30] also followed a similar pattern of fluctuations after formation and intensification to storm status and eventually dissipating. Hurricane Roxanne had a confusing formation [31] and fluctuated frequently between low intensity and severe cyclone storm status before finally dissipating. Such behaviour patterns of cyclone storms indicate irregularities and deviations from a model cyclone evolution pattern, as is true of any natural phenomena. The cyclones had been assigned categories based on the SSHS scheme at different stages. The T-numbers corresponding to each of these categories can be obtained from conversion tables and the T-numbers thus obtained give an indication of the cyclone evolution trend. We compared our results with the T-numbers obtained from these conversion values. In Table 2, we compare our results with the values obtained from this cyclone evolution trend (obtained from reports on the Internet [32]). Table 3 provides calculations for obtaining the Receiver Operating Characteristics (ROC) and Table 4 provides the Confusion Matrix for the T-numbers that we have classified versus the T-numbers manually assigned. This table gives an indication of the true and false classification rates. Figure 8 shows the ROC curve with TP indicating True Positive and FP indicating False Positive. The Area Under the Curve of the ROC curve (Figure 5) was 0.8278 indicating an 82.78% agreement with predicted values.

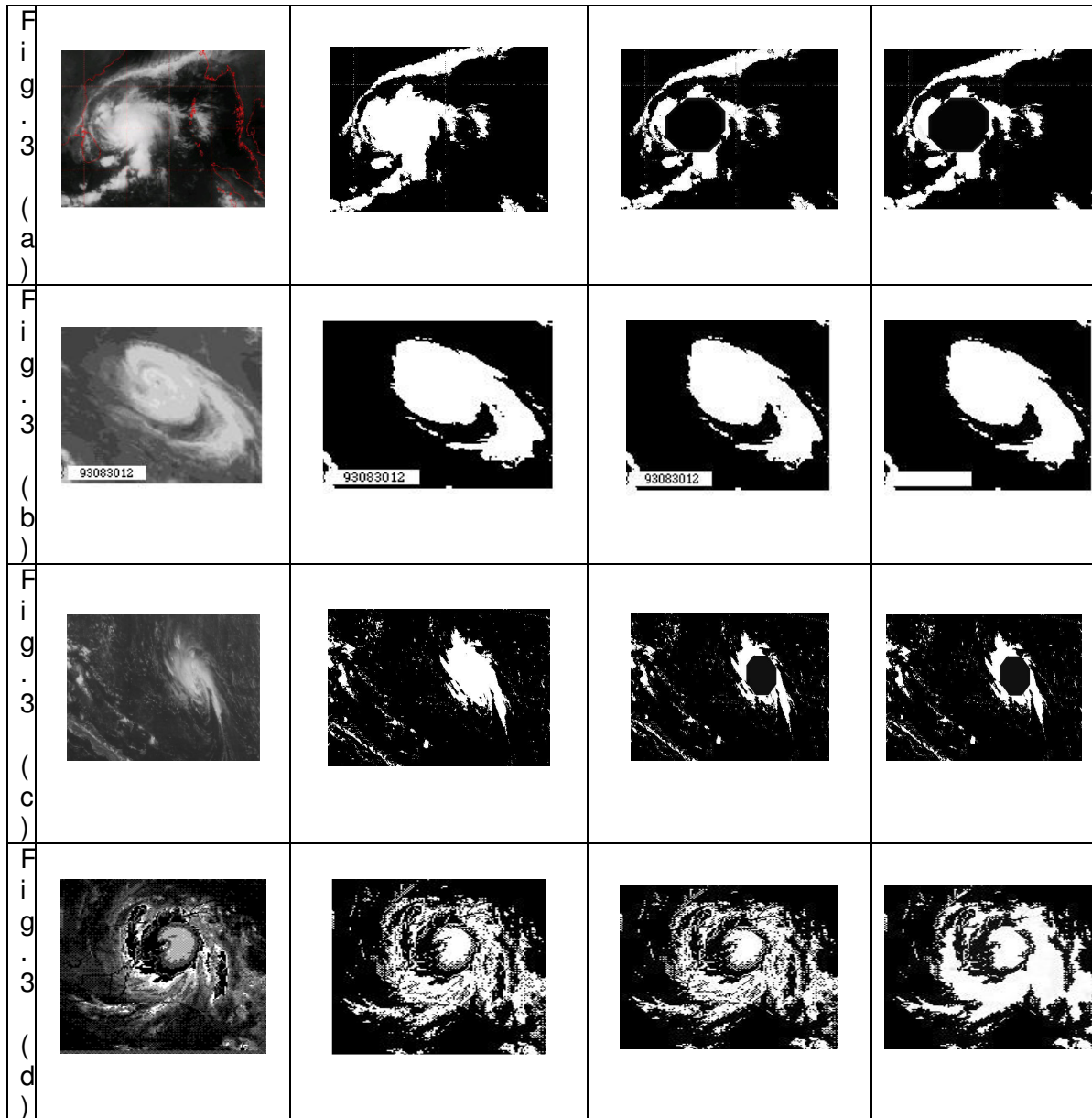


FIGURE 3: Images of Ward, Emily, Andrew and Roxanne after image preprocessing in rows 1-4, respectively.

Our proposed methodology falls under the second category. As with other approaches under this category, we preprocess the images using noise removal, Otsu thresholding, morphological operations like opening and filling, as appropriate. After preprocessing we use the SHAPE package for boundary extraction, chain coding, use of the EFD and PCA analysis, after which comparisons are made between the candidate and prototype (Dvorak template) images. Hence our matching procedure is based on shape descriptors. It may be mentioned here that, SHAPE can convert a color image to a binary image, remove noise and perform thresholding before boundary extraction and subsequent operations that we have used the package for.

T-number	Total number of images matched with the original T-number (TP)	Total number of images mismatched with the original T-number (FP)	Cumulative Rate	
			TP	FP
T1.5	11	1	0.0558	0.0256
T2.5	110	22	0.614	0.564
T3.5	35	9	0.792	0.821
T4.5	18	2	0.883	0.872
T5.5	13	4	0.949	0.974
T6.5	11	1	1.0	1.0
Total	197	39		

TABLE 3: Calculation of Receiver Operating Characteristic (ROC).

		Predicted Class					
		T1.5	T2.5	T3.5	T4.5	T5.5	T6.5
Actual Class	T1.5	11	1	0	0	0	0
	T2.5	12	107	5	3	0	0
	T3.5	0	3	35	5	1	0
	T4.5	0	0	1	15	1	0
	T5.5	0	0	2	2	13	1
	T6.5	0	1	0	0	0	9

TABLE 4: Confusion Matrix for the Algorithm.

5. DISCUSSIONS AND FUTURE WORK

This paper describes a novel approach to study the evolution of cyclones using both VIS and EIR satellite images of cyclone cloud intensity patterns. As discussed earlier, there are broadly three categories of pattern matching analysis of cyclones. The first approach focuses on the extraction of the eye. Most of the earlier approaches and also some later approaches fall under this category. The second and third categories involve a segmentation procedure after image pre-processing to extract cyclone cloud intensity contours. Subsequently, efforts in the second category attempt to match these images with Dvorak prototypes using soft computing techniques like neural networks, genetic algorithms, ant colony optimization, etc.,. Another set of approaches under this category involve generating gradient vectors or angle features at different points along the contour and matching these features with prototypes. The third category involves fitting the contours to curves generated from mathematical models (e.g. logarithmic helix, spiral band, etc.,).

5.1 Comparison with Other Work

The location of the cyclone “eye” is important as the very existence and metamorphosis of the cyclone is dependent on its presence, and hence, the detection of the eye does play an important role in cyclone evolution analysis. However, the typical cyclone contour is “comma” shaped and so the degree of spiralling of the curved band is indicative of the different stages of evolution of the cyclone and thus, the approaches of the second and third categories provide a clearer picture of the cyclone evolution. Mathematical models provide exact shapes, but cyclones being natural phenomena do not always correspond to regular curves. So an empirical image matching scheme with prototypes could give a good estimate of cyclone evolution trends. Our correct classification rate of 83% percent compares favorably with the other techniques in all these categories. Lee and Liu [11] have claimed an overall accuracy of 82% and found that EIR images give better results because of a better spiral pattern. Their earlier effort [10] yield an overall value 86% for track intensity mining but their efficiency depended on the inter-relationship of successive pictures. Pao and Yeh [9] claimed a correct classification rate of 82% while locating the center and contour of the typhoon. Pineros et. al. [14] achieved correlation rates ranging between 82-86% with the highest being for instances where the maximum hurricane strength was achieved. Liu et. al. [12] achieved a 10% (72.41%) improvement of human visual justification (62.86%).

The methodology that we have proposed has not been used previously, to the best of our knowledge. In an earlier attempt [33] we have used a different preprocessing technique involved a classical edge extraction template followed by erosion and dilation. However, edge detectors produce broken contours and so we have replaced [34] this earlier methodology with a filling algorithm to produce closed contours before performing boundary extraction followed by Freeman chain code implementation that is available in SHAPE. This has improved the results by 16%. The use of Elliptic Fourier Descriptors has several advantages over Fourier descriptors. Integration or use of fast Fourier techniques are not required and bounds on the accuracy of the image contour representation are easy to specify.

EFDs are convenient and Fourier contour representations can be conveniently normalized and is thus useful for the analysis of well defined 2D contours. Dvorak templates are widely used and we have also used them as they capture the essence of the intensity patterns effectively. As EFDs involve the use of many coefficients, PCA is used to reduce the dimensionality. In the earlier attempt [33] we had used 20 harmonics in the PCAbased classification, but because there are at most three to five spiral turns, we have used five harmonics, although our results converged after three harmonics. The degree of banding and thus the evolutionary stage of the cyclone can thus be described.





Particular of Candidate and Model image	PCA of Candidate image	PCA of model image
The principal component of the candidate image Andrew 2 and the model image EIR(T 1.5) with 5 harmonics.		
The principal component of the candidate image Emily and the model image VIS (T1.5) with 5 harmonics.		

FIGURE 4: Principal component of candidate images of Hurricanes Andrew and Emily and template images with which they are matched.

EFDs are convenient and Fourier contour representations can be conveniently normalized and is thus useful for the analysis of well defined 2D contours. Dvorak templates are widely used and we have also used them as they capture the essence of the intensity patterns effectively. As EFDs involve the use of many coefficients, PCA is used to reduce the dimensionality. In the earlier attempt [33] we had used 20 harmonics in the PCA based classification, but because there are at most three to five spiral turns, we have used five harmonics, although our results converged after three harmonics. The degree of banding and thus the evolutionary stage of the cyclone can thus be described.

5.2 Future Work

Future work will have to focus on a clear extraction of the eye and also try to integrate the preprocessing stage with boundary extraction, chain coding, etc. and to automate the whole process. This work will have to be extended to larger datasets in order to standardise the technique for cyclone intensity prediction. Also, a more accurate registration scheme between the SSHS scheme and Dvorak technique needs to be formulated and designed, in order to provide a more accurate prediction technique. The methodology will also have to be improved to cater to situations where the storm interacts with unfavourable environments such as land or wind shear. The database to be used should also include complicated situations where there are several cloud structures located within the disturbance, or when the shape of the cloud structures are elongated.

One of the most challenging problems is to predict the cyclone formation at an early stage of development, and hence, another long term goal is to incorporate cyclogenesis in this study.

6. ACKNOWLEDGMENTS

The authors are thankful to the University Grants Commission (UGC) Government of India (major grant no.37-534(09)) for their support. Images of cyclones in the Bay of Bengal and Indian Ocean were obtained from the Indian Meteorological department, New Delhi, India. Images of hurricanes and typhoons were obtained from the NOAA website.

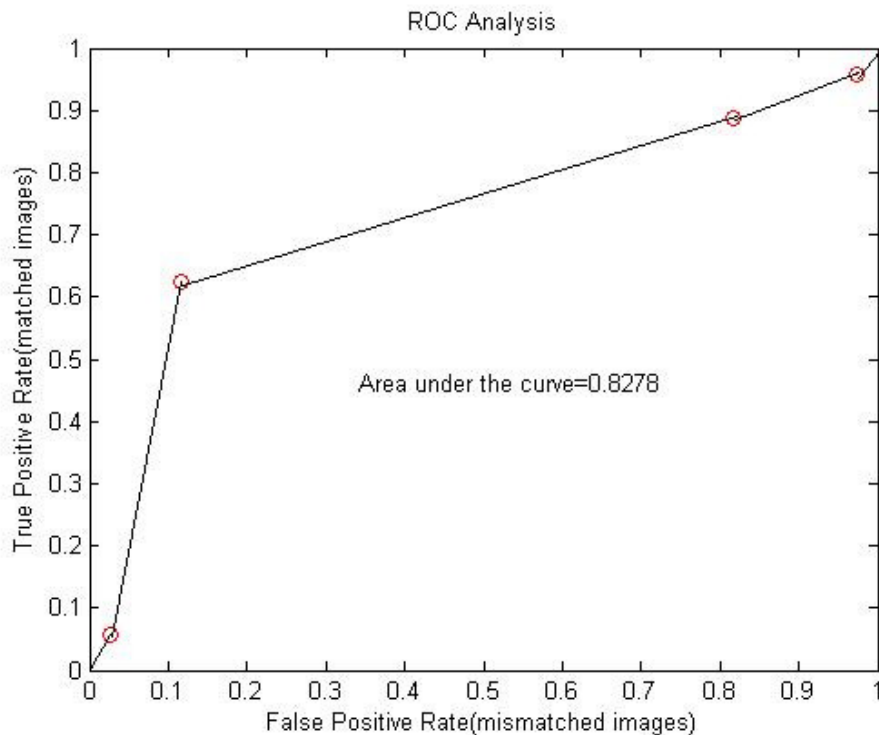


FIGURE 5: Data fitted with Receiver Operating Characteristic Curve.

7. REFERENCES

- [1] K. Emanuel "A Statistical Analysis of Tropical Cyclone Intensity", Monthly Weather Review, 128, pp.1139-1152 (2000).
- [2] C. C . Chao G. R. Liu, C. Liu , "Estimation of the upper-layer rotation and maximum wind speed of tropical cyclones via satellite imagery", Journal of Applied Meteorology and Climatology , 50(3), pp.750-766 (2011).
- [3] V. Dvorak. "Tropical cyclone intensity analysis and forecasting from satellite imagery", Monthly Weather Review, .103, pp.420-430 (1975).
- [4] V. Dvorak "Tropical cyclone Intensity Analysis Using Satellite Data", NOAA Technical Report NESDIS 11, (1984).
- [5] J.A. Knaff , D.P. Brown, J. Courtney , G. M. Gallina, J.L. Beven , "An Evaluation of Dvorak Technique-Based Tropical Cyclone Intensity Estimates", Weather and Forecasting, 25(5), 1362-1379, (2010).
- [6] J.S. Griffin, R.W. Burpee, F. D. Marks, and J. L. Franklin, " Real Time airborne analysis of aircraft data supporting operational hurricane forecasting", Weather and Forecasting, 7, pp.480-490, (1992).
- [7] V. T.Wood, " A technique for detecting a tropical cyclone centre using a Doppler radar", Journal of Atmospheric and Oceanic Technology, 11, pp. 1207-1216, (1994).

- [8] L. Tsang , LYeh , M. Liu, Y.Hsu, "Locating the Typhoon Center from the IR Satellite Cloud Images", IEEE International. Conference on System, Man and Cybernetics (SMC 2006) pp. 484-488, (2006).
- [9] T. L.Pao, and J.H. Yeh, " Typhoon locating and reconstruction from the infra-red satellite cloud image", Journal of Multimedia, 3, 2, pp. 45-51, (2008).
- [10] R. S. T. lee and j n k liu "Tropical cyclone identification and tracking system using integrated neural oscillatory elastic graph matching and hybrid RBF network track mining techniques," IEEE Trans Neural Networks, 11,pp. 680–689, (2000).
- [11] R.S. T. Lee , and J.N. K. Liu, " An Elastic Contour Matching Model for Tropical Cyclone Pattern Recognition", IEEE Transactions on Systems Man and Cybernetics-Part B: Cybernetics, 31, 3, 413-417, (2001).
- [12] J.N.K. liu, b. feng , m. wang. and w. luo "Tropical Cyclone forecast using Angle Features and Time Warping", International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, pp-4330-37, (2006).
- [13] Q. zhang , l. lai and h. wei " Continuous space optimized artificial ant colony for real-time typhoon eye tracking", IEEE International Conference on Systems, Man and Cybernetics (SMC) Washington DC: IEEE, pp.1470–1475 , (2007).
- [14] M. F. Pineros, E.A. Ritchie, and J.S. Tyo, " Objective Measures of tropical cyclone structure and Intensity change from remotely-sensed infra-red data", IEEE Transactions On Geosciences and Remote Sensing, 46, 11, pp. 3574-3579, (2008).
- [15] K. Y. WONG, C.L., YIP, P.W. LI, W. W. TSANG, "Automatic template matching method for tropical cyclone eye fix", Proceedings of the 17th International Conference on Pattern Recognition(ICPR-2004), Cambridge UK., 3, pp. 650-653, (2004).
- [16] C. L.YIP, K.Y. WONG, Efficient and effective tropical cyclone eye fix using genetic algorithms, Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES-2004), 3213 of Lecture Notes on Artificial Intelligence, Springer-Verlag, Wellington, pp. 654-660, (2004).
- [17] K. Y. Wong, C.L. Yip, and P.W. Li, "A tropical cyclone eye fix using genetic algorithm", Expert Systems with Applications 34, 643–656, (2008).
- [18] K.Y. Wong and C.L. Yip "Identifying centers of circulating and spiraling vector field patterns and its applications". Pattern Recognition, 42, pp. 1371–1387, (2009).
- [19]K. Wei , Z. L.Jing , "Spiral band model optimization by chaos immune evolutionary algorithm for locating tropical cyclones", Atmospheric Research, 97 (1-2), 266-277, (2010) .
- [20]M.A Abidi. and R.C. Gonzalez "Shape Decomposition Using Elliptic Fourier Descriptors", Proceedings. 18th. IEEE South-east Symposium on System Theory, Knoxville, TN, pp.53-61, (1986).
- [21] Q. I.Guo, F. Guo, J. Shao "Irregular Shape Symmetry Analysis: Theory and Application to Quantitative Galaxy Classification", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 32, no. 10, pp.1730-1743,(2010).
- [22] N. Otsu "A threshold selection method from gray-level histograms," IEEE Trans. Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62-66, (1979).

- [23] H.F. reeman "On encoding of arbitrary geometric configurations," IRE Transactions on Electronic computers EC 10, pp.260-268, (1961).
- [24]H. Iwata and Y. Ukai "SHAPE: A Computer Program Package for Quantitative Evaluation of Biological Shapes",Journal of Heredity, 93(5), pp.384-385, (2002).
- [25] F.P. Kuhl AND C.R. Giardina "Elliptic Fourier features of a closed contour", Computer Graphics and Image Processing, 18, 236–258, (1982).
- [26] Y. Yoshioka, H. Iwata, R. Ohsawa, S. Ninomiya "Analysis of Petal Shape Variation of *Primula sieboldii* by Elliptic Fourier Descriptors and Principal Component Analysis", Annals of Botany 94, pp. 657–664, (2004) .
- [27] INDIAN METEOROLOGICAL DEPARTMENT "Hurricane Ward: A Preliminary Report", Available at: <http://www.imd.gov.in/section/nhac/dynamic/cycrptward.pdf> [2009]
- [28] NASA" NASA Study Finds 'Pre-Existing Condition' Fueled Killer Cyclone", Available at: <http://www.nasa.gov/topics/earth/features/nargis-20090226.html>, (2009).
- [29] J. L.Franklin and D. P. Brown "Tropical Cyclone Report: Hurricane Emily", pp.1-18, (2006).
[30] E. N. RAPPAPORT "Hurricane Andrew", Weather (49), pp. 51-61, (1992).
- [31] L. AVILA L.Preliminary Report Hurricane Roxanne, Available at: <http://www.nhc.noaa.gov/1995roxanne.html>
- [32] INDIAN METEOROLOGICAL DEPARTMENT (2009) Hurricane Aila: A Preliminary Report, Available at: [http://www.imd.gov.in/section/nhac/dynamic/aila.pdf\(2009](http://www.imd.gov.in/section/nhac/dynamic/aila.pdf(2009)
INDIAN METEOROLOGICAL DEPARTMENT (2009) Hurricane Phyan: A Preliminary Report, Avail-able at: (<http://www.imd.gov.in/section/nhac/dynamic/cyclone2008>)
- [33] NATIONAL HURRICANE CENTRE (2001) Tropical Cyclone Report Hurricane Flossie Available at:<http://www.nhc.noaa.gov/2001flossie.html>
- [34]PASCH AND RICHARD Hurricane Jeanne Preliminary Report, National Hurricane Centre,(1999) STEWART S. R. Tropical Cyclone Report Hurricane Darby, pp. 1-12, (2010).
- [35] I. Dutta and S. Banerjee, "Shape Analysis of Satellite Images of Cyclone" ICCS 2013-Second International Conference on Computing and Systems", Burdwan, India, (2013).
- [36] I. Dutta, S. Banerjee and M.De, "An Algorithm for Pre-Processing of Satellite Images of Cyclone Clouds", in press, International Journal of Computer Applications (2013).
- [37] V. F. Dvorak, "A technique for the analysis and forecasting of tropical cyclone intensities from satellite pictures," unpublished, 1973.

3D Position Tracking System for Flexible Cystoscopy

Munehiro Nakamura

*Department of Natural Science and Engineering
Kanazawa University
Kanazawa, 9201192, Japan*

m-nakamura@blitz.ec.t.kanazawa-u.ac.jp

Yusuke Kajiwara

*Department of Information Science
Ritsumeikan University
Kusatsu, 525877, Japan*

kajiwara@de.is.ritsumei.ac.jp

Tatsuhito Hasegawa

*Department of Natural Science and Engineering
Kanazawa University
Kanazawa, 9201192, Japan*

t-hasegawa@blitz.ec.t.kanazawa-u.ac.jp

Haruhiko Kimura

*Department of Natural Science and Engineering
Kanazawa University
Kanazawa, 9201192, Japan*

kimura@blitz.ec.t.kanazawa-u.ac.jp

Abstract

Flexible cystoscopy is an examination that allows physicians to look inside the bladder. In flexible cystoscopy, beginner physicians tend to lose track of the observation due to complex handling patterns of a flexible cystoscope and poor characteristics of the bladder. In this paper, as a diagnostic support tool for beginner physicians in flexible cystoscopy, we propose a system for tracking the observation using cystoscopic images. Our system discriminates three handling patterns of a flexible cystoscope, namely bending, rotation, or insertion. To discriminate the handling patterns accurately, we propose to use the degree of bending, rotation, or insertion as features for the discrimination as well as ZNCC-based optical flows. These features are learned by a Random Forest classifier. The classifier discriminates sequential handling patterns of the cystoscope by a time-series analysis. Experimental results on ten videos obtained in flexible cystoscopy show that each of the three handling patterns were correctly discriminated over 90% in average. In addition, we reproduced the observation in a virtual bladder we propose.

Keywords: Flexible Cystoscopy, Position Tracking, Optical Flow, Zero-mean Normalized Cross-Correlation, Handling Pattern.

1. INTRODUCTION

With increase of aged people in the world, incidents of bladder disease are gradually increasing[1]. Bladder disease can be detected in cystoscopy or non-invasive examinations such as blood test, MRI, CT, PET, and ultrasonography. The non-invasive examinations are painless and less stressful. However, it is still difficult to detect tiny lesions in a non-invasive imaging examination[2]. Cystoscopy is conducted when a lesion found in a non-invasive imaging examination or severe symptoms are appeared in a patient. Cystoscopy enables physicians to look inside the bladder to confirm a patient's lesion directly. There are two types of cystoscopes, namely rigid and flexible. The examination with the latter one is less painful and is used widely. In this paper, we deal the examination with flexible cystoscope.

In flexible cystoscopy, images of the bladder are obtained from a camera embedded in the tip of flexible cystoscope and displayed in the monitor. However, there are three major difficulties for physicians to check the whole inner of the bladder completely. First, since the bladder has similar

shape and color, sometimes beginner physicians lose track of the observation. Second, cystoscopic images are sometimes unclear due to halation. Third, it requires some experiments to control flexible cystoscope properly due to complex handling patterns of the equipment.

As a diagnostic support tool for beginner physicians in flexible cystoscopy, this paper presents a system for tracking the observation. To achieve the objective, it is considered to attach acceleration sensors or location sensors to a flexible cystoscope. In that case, it is necessary to obtain an approval for usage of the cystoscope according to the pharmaceutical law, as well as to buy the cystoscope which would be at least 10,000\$. Another approach to track the observation in cystoscopy is mapping observed regions in a 3D space. As a representative 3D tracking system, Choi et al. proposed a robust segment-based object tracking system that uses the backside of a car image[3]. Their system measures depth information by calculating the enlargement factor of a target region. Ramisa et al. proposed to measure the distance between a single camera and a person[4]. However, since cystoscopic images are significantly unclear than the car image and human image, existing algorithms for 3D tracking[3, 4] could not be applied to cystoscopic images.

In this paper, we propose to discriminate the handling patterns of a flexible cystoscope. First, the proposed system extracts ZNCC-based optical flows[5] from cystoscopic images as features for estimating the handling patterns. Next, various features including the ZNCC-based optical flows are learned by a Random Forest classifier[6]. The classifier discriminates sequential handling patterns of the cystoscope by a time-series analysis. Finally, the observation in flexible cystoscopy is reproduced in a virtual 3D bladder we propose.

In section 2, we will introduce the process of flexible cystoscopy. In section 3, we will explain the proposed system. In section 4, we will examine the performance of the proposed system regarding the accuracy in estimating the handling patterns and tracking the observation in the virtual bladder.

2. FLEXIBLE CYSTOSCOPY

The human bladder is a hollow and balloon shaped organ that is broadly distinguished into seven regions; trigone, neck, left side wall, right side wall, posterior wall, dome, and anterior wall. Figure 1 shows images of the bladder. From the figure, we could perceive the images except neck are similar to each other. In addition, sometimes cystoscopic images are noisy due to halation which often occurs when the cystoscope is close to the bladder wall.

The flexible cystoscopy is conducted by a physician as below.

- (1) The physician inserts a flexible cystoscope into a patient's urethra.
- (2) The physician pushes the cystoscope slowly to the bladder.
- (3) By adjusting the position of the cystoscope, the physician observes the whole inner of the bladder.
- (4) The physician pulls the cystoscope after checking all the regions.

Figure 2 shows three handling patterns of the cystoscope to adjust the position of the cystoscope in step (3). The problem in this examination is that a beginner physician in the step (4) is sometimes unsure that all the regions were completely observed. Since oversight of severe legion would cause fatal case, diagnostic support tools for beginner physicians in flexible cystoscopy have been required.

3. PROPOSED SYSTEM

3.1 Overview

As mentioned in Sec. 2, the bladder can be distinguished into seven regions. However, the definitive region of each part is not defined. Considering a normal bladder, we construct a sphere bladder model which has seven regions. Figure 3 shows the virtual bladder we propose. And, Table 1 shows definitions of each region for the virtual bladder, which determined by an expert physician in flexible cystoscopy.

3.2 Preprocessing

Figure 4 shows the interface for a flexible cystoscopy. In the proposed system, first, ROI (Region of Interest) is set on the rectangle in the interface. The size of ROI is 300 pixel \times 300 pixel. Next, Figure 5 (b) shows the cystoscopic image applied 8-bit gray scale transformation to Figure 5 (a). And,

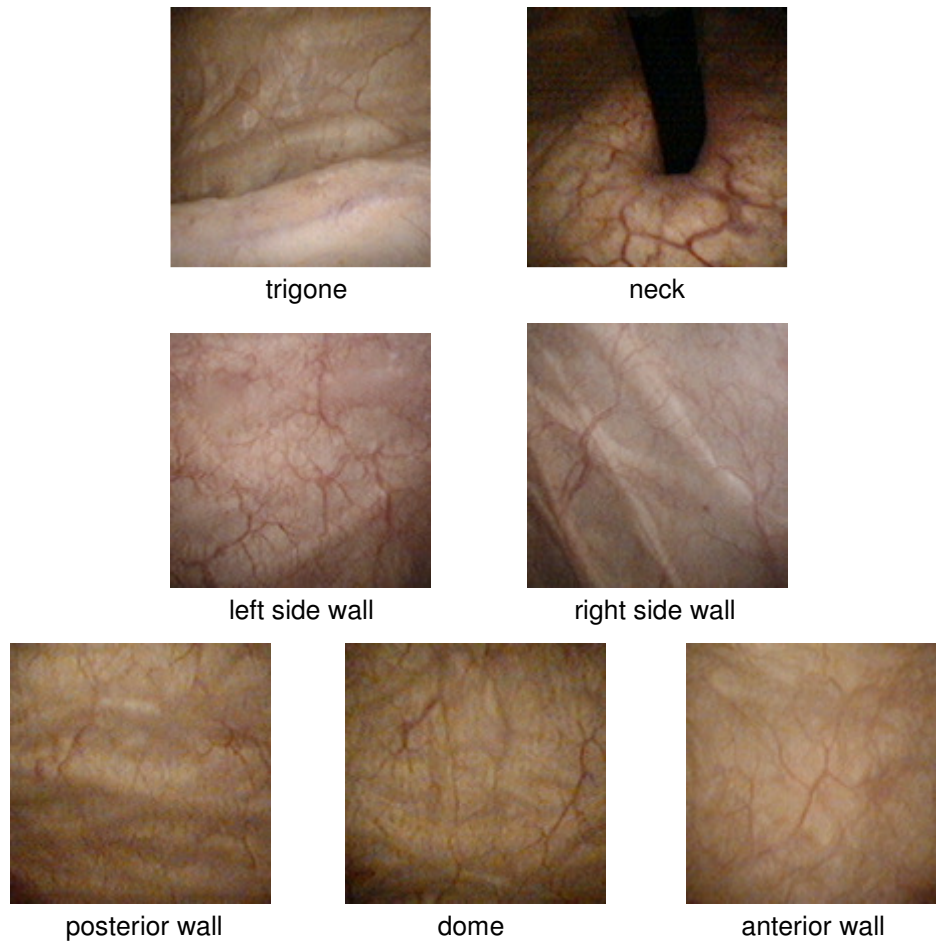


FIGURE 1: Example of Images Obtained from a Flexible Cystoscope.

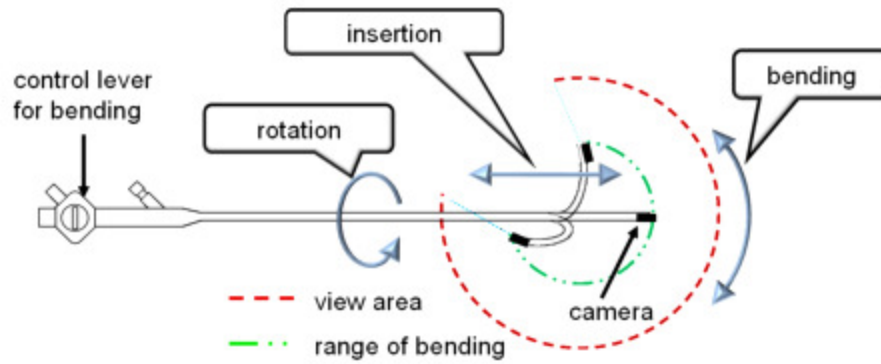


FIGURE 2: Three Handling Patterns of a Flexible Cystoscope.

Figure 5 (c) shows the image applied the histogram stretching to Figure 5 (b). Finally, Figure 5 (d) shows the image applied the selective local averaging to Figure 5 (c). Compared with Figure 5 (b), we could perceive that blood vessels are enhanced in the images of Figure 5 (d).

3.3 Extraction of Optical Flows

Approaches of well-known object tracking can be distinguished into gradient method and block matching method. Gradient method is effective on videos where target objects move slowly[7].

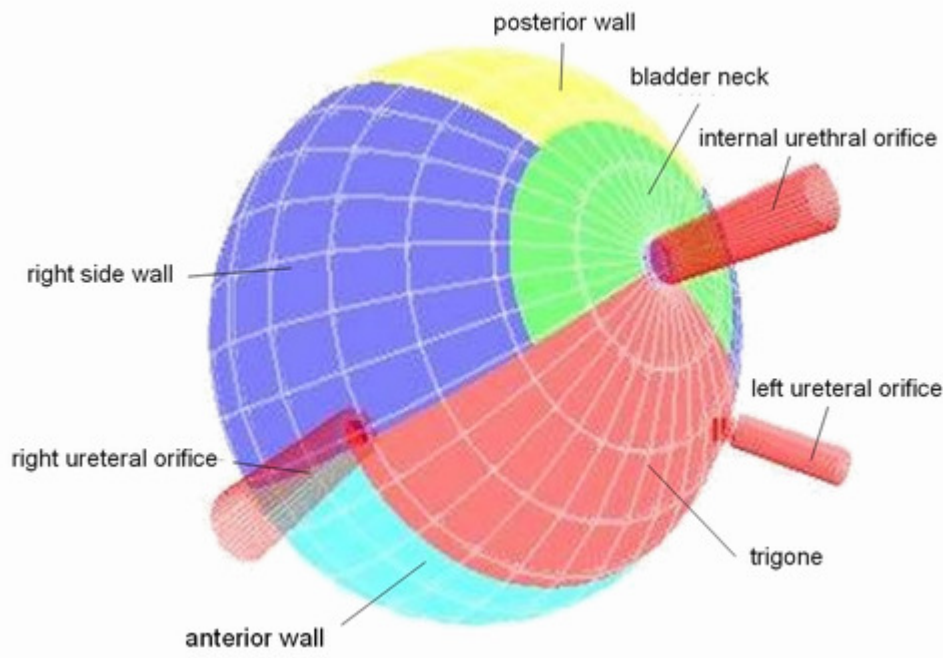


FIGURE 3: 3D Virtual Bladder We Propose.

Bladder region	Definition
Dome	More than lat. 70 degrees S, excluding the trigone defined below.
Bladder neck	More than lat. 70 degrees S, excluding the trigone defined below.
Trigone	Triangle part is surrounded with lat. 45 degrees S, a parallel of lines of longitude of long. 60 degrees E / long. 60 degrees W.
Posterior wall	The quadrangle surrounded with a parallel of lat. 60 degrees N/ lat. 70 degrees S, a line of longitude of long. 60 degrees E/ long. 60 degrees W.
Anterior wall	The quadrangle surrounded with a parallel of lat. 60 degrees N/ lat. 70 degrees S, a line of longitude of long. 120 degrees E/ long. 120 degrees W.
Right side wall	The reminded region in the east side.
Left side wall	The reminded region in the west side.

TABLE 1: Definition of the Virtual Bladder Map.

Meanwhile, block matching method is robust to quick movement of target objects, sudden illumination changes, and noises. Since cystoscopic images are often unclear due to noises such as halation and focus error, the proposed method applies the block matching[8]. As a robust method to measure the movements of a flexible cystoscope, the proposed method extracts ZNCC-based optical flows from consecutive cystoscopic images. In ZNCC (Zero-mean Normalized Cross-Correlation), an image is divided into $A \times B$ blocks and optical flows are extracted from each block by the following formula

$$R_{ZNCC} = \frac{\sum_{i=1}^N \sum_{j=1}^M (I(i, j) - \bar{I})(T(i, j) - \bar{T})}{\sqrt{\sum_{i=1}^N \sum_{j=1}^M (I(i, j) - \bar{I})^2 \times \sum_{i=1}^N \sum_{j=1}^M (T(i, j) - \bar{T})^2}} \quad (1)$$

where $I(i, j)$ is the pixel value at i th row and j th column in the template image, $T(i, j)$ is the pixel value at i th row and j th column in the target image, N is the search range towards x axis, and M is

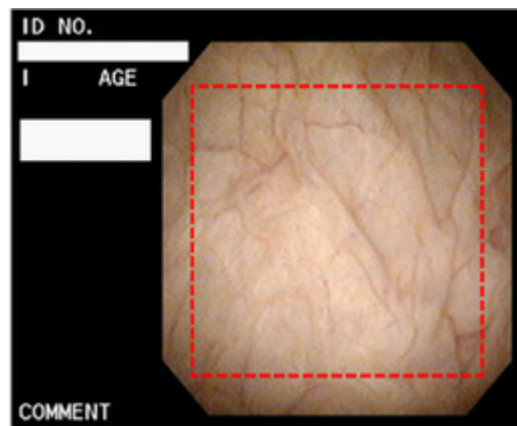


FIGURE 4: Interface for the Examination with a Flexible Cystoscope.

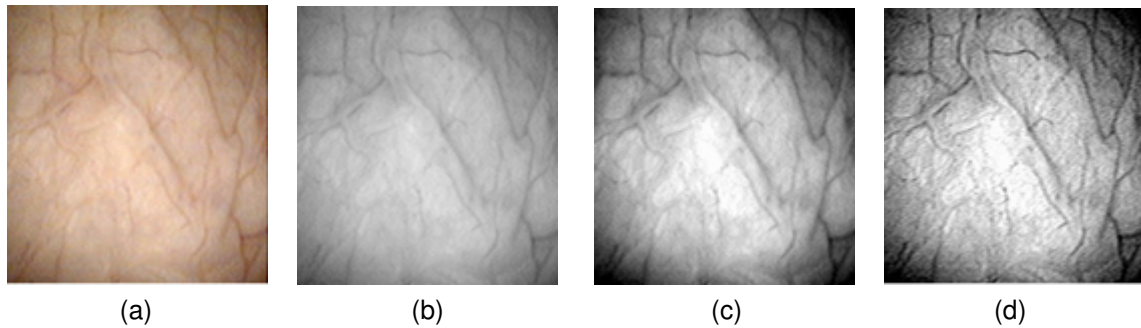


FIGURE 5: Preprocessing for the Enhancement of Blood Vessels.

Handling pattern			
rotation	left	right	neutral
bending	up	down	neutral
insertion	push	pull	neutral

TABLE 2: Handling Patterns of Flexible Cystoscope.

the search range towards y axis. As an algorithm of ZNCC, the proposed system uses the one proposed by Yoo and Han[5].

3.4 Discrimination of Handling Patterns for a Flexible Cystoscope

Table 2 shows handling techniques of the flexible cystoscope. From the table, we can find that the combination of the handling techniques is up to 27 patterns. And, figure 6 shows an example of optical flows obtained by rotation (a) and insertion (b) when the cystoscope is 90 degrees down or zero degrees or 90 degrees up. From figure 6, we can find that the ZNCC-based optical flows are depended on the degree of rotation and insertion.

We define three features that improve the discrimination of the handling patterns as below

$$Rot_t = \sum_{i=1}^t R_i \tag{2}$$

$$Bend_t = \sum_{i=1}^t B_i \tag{3}$$

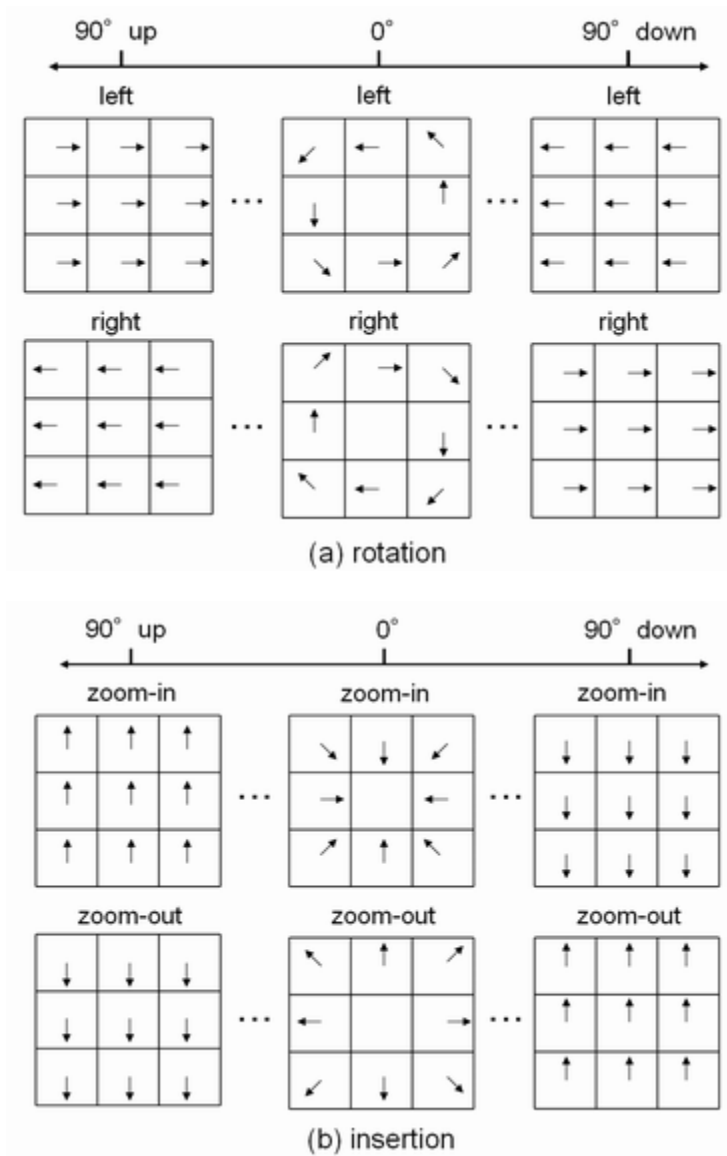


FIGURE 6: Example of Optical Flows Obtained by Rotation (a) and Insertion (b).

$$Ins_t = \sum_{i=1}^t I_i \tag{4}$$

where Rot_t is the degree of rotation at t th frame, R_t is the handling pattern of rotation at t th frame and returns 1 when $R_t=left$, -1 when $R_t=right$, and 0 when $R_t=neutral$. Similarly, $Bend_t$ is the degree of bending at t th frame, B_t is the handling pattern of bending at t th frame and returns 1 when $B_t=up$, -1 when $B_t=down$, and 0 when $B_t=neutral$, and Ins_t is the degree of insertion at t th frame, I_t is the handling pattern of insertion at t th frame and returns 1 when $I_t=push$, -1 when $I_t=pull$, and 0 when $I_t=neutral$.

Thus, the proposed system extracts features as $A \times B$ optical flows, Rot_t , $Bend_t$, and Ins_t from a cystoscopic image. By feeding all the features, a RF (random forest) classifier discriminates the 27 handling patterns frame by frame. RF is a noise-robust classification algorithm proposed by Breiman.

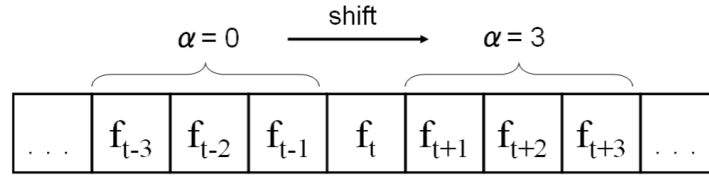


FIGURE 7: Selection of consecutive k frames whose prediction probabilities are similar to those of f_t ($k=5$).

3.5 Discrimination of Sequential Handling Patterns for a Flexible Cystoscope

In flexible cystoscopy, the same handling pattern would last for several frames. Hence, in a case that the discriminate handling for a frame f_t is different from the one for f_{t-1} and f_{t+1} , f_t is expected to be another handling pattern. Considering the case, we propose to correct the handling pattern for f_t . Figure 7 shows a process to discriminate sequential k frames, where $k = 5$ is configured in this case. The sequential k frames are determined by the following steps.

- (1) Select the sequential k frames from $f_{t-k+\alpha}$ to $f_{t+1+\alpha}$ (Initially, $\alpha = 0$).
- (2) Calculate a similarity of the class prediction probabilities for each frame.
- (3) $\alpha = \alpha + 1$ if $\alpha < k$ and go back to step (1).
- (4) Determine the k frames when the similarity in step (2) is maximum.

And then, the similarity Sim_t is calculated by the following formula

$$Sim_t[\alpha] = \sum_{j=1}^{27} |aprob_t[j] - prob_t[j]| \tag{5}$$

where $prob_t[j]$ is the class prediction probability of f_t for j th handling pattern, $aprob_t[j]$ is the average class prediction probability of selected k frames for j th handling pattern. Thus, the proposed method selects k frames whose average class prediction probability are similar to that of f_t . And then, the discriminated handling pattern for f_t is replaced by the j th handling pattern where $aprob_t[j]$ becomes maximal in k frames.

Next, regarding f_{num} as the number of replaced handling patterns ($f_{num}=1,2,3,\dots,k-1$), f_{num} is increased in 1 by 1 (f_{num} is 1 in the initial step above). Then, regarding k as k' , k' is set as $k-f_{num}+1$. Note that the replacement of k' frames is conducted when the following term is fulfilled.

$$P_B > P_A \tag{6}$$

where P_B is the average class prediction probability for a handling pattern B in k' frames and P_A is the average class prediction probability for a handling pattern A in k' frames. These terms can prevent a case that the handling pattern of the correct sequential k' frames is replaced to incorrect one. The method above can be applied to frames before f_t when the discrimination of the handling pattern for $t+k$ frames is finished.

4. PROPOSED SYSTEM

4.1 Experimental Environment

We applied the proposed system to ten videos of flexible cystoscopy at the Kanazawa University Hospital in Japan. All the examinations were conducted by an expert physician. The flexible cystoscope used in the examinations is OLYMPUS CYF TYPE VA2. Regarding the videos, the format is AVI (24-bit color), the frame rate is 29.97 fps, and the average length is 115 seconds. Each of the videos was cut manually as they start from the scene that the image of the bladder wall appears at the first time to the scene that the image of the urethra appears at the first time after the start scene. To evaluate the performance of the proposed system, by observing the

videos, the expert physician judged the handling patterns frame by frame. Then, a RF (Random Forest) classifier discriminated the handling patterns. The accuracy of estimating the handling patterns is evaluated by 10-fold cross validation[9] as below.

- (1) Choose the first 1000 frames in each of the ten videos.
- (2) Extract ZNCC-based optical flows and Rot_t , $Bend_t$, and Ins_t (refer to Sec. 3.4) from each of the 1000 frames.

Handling pattern	Number of frames	Base line	proposed method
neutral	877	70.6%	93.2%
left	1049	81.0%	94.5%
right	932	82.1%	93.4%
up	741	83.3%	95.0%
down	1214	82.0%	96.2%
push	801	76.0%	94.6%
right	932	82.1%	93.4%

TABLE 3: Average correct rate for each of the handling patterns. *base line* is the case that only ZNCC-based optical flows were learned by the Random Forest classifier.

No.	1	2	3	4	5
<i>Cr</i>	54.9%	81.9%	91.6%	91.2%	62.5%
<i>Fr</i>	6.4%	2.6%	1.2%	1.3%	5.4%
No.	6	7	8	9	10
<i>Cr</i>	83.5%	84.1%	82.9%	92.1%	90.4%
<i>Fr</i>	2.4%	2.3%	2.4%	1.1%	1.4%

TABLE 4: Parameter *Cr* and *Fr* obtained in the Experiment.

- (3) Choose the 1000 frames in one of the videos as test data and the other 9000 frames as training data.
- (4) RF learns the training data.
- (5) RF discriminates the handling patterns frame by frame in the test data.
- (6) Sequential handling patterns of the cystoscope are discriminated by the time series analysis described in Sec. 3.5.
- (7) Repeat the procedure from (3) until all the 1000 frames are selected as test data.

4.2 Results

First, we discriminated the handling patterns of the flexible cystoscope in each of the 1000 frames. Table 3 shows the average correct rate in the discrimination. In this experiment, parameters for the Random Forest classifier were optimized, namely the number of the trees was configured as 525. As the table shows, the proposed method outperforms *base line*.

Next, we reproduced each of the cystoscopic examinations in the virtual bladder defined in Sec. 3.1. In each examination, observed regions were painted on the virtual bladder. To evaluate the performance of reproducing the observation, we assume that the physician could not observe one of the whole regions. Such a region is called as target region in the rest of this paper. For example, suppose that the physician could not observe trigone, the region of trigone in the virtual

bladder is unpainted ideally and the other regions are painted. Here, we define Cr as the percentage of the unpainted area in the target region and Fr as the percentage of the painted area in the other region. Therefore, the ideal Cr is 100% and the ideal Fr is 0%. Table 4 shows the average Cr and Fr for each of the video. From the table, we can find that Cr has been more than 80% except the video No.1 and No.5.

4.3 Failure Cases

In Table 4, both of Cr and Fr for the video No.1 has been the worst among all the videos. Observing the video, we could find that bladder stones were floating around trigone. Figure 8 shows the stones marked in circle. In 114 frames, although the camera was being stopped, the block-matching method detected movement of the stones and the proposed system incorrectly judged the handling patterns in the 114 frames as left or insertion. The bladder stones are sometimes observed in cystoscopy. Therefore, we need to detect the white stones as noise. In Figure 8, average density of the stones in areas of each circle has been 223.4 (standard deviation is 15.9) while average density of all the 64072 frames in ten videos is 124.9. The densities above were



FIGURE 8: Bladder Stones Appeared Around Trigone in the Video No.1.



FIGURE 9: One of the Images where Halation Observed in the Video No.5.

measured from the original images obtained from the cystoscope. Hence, areas of the bladder stones in each image would be extracted using the density distributions of the whole image.

In Table 4, Cr and Fr for the video No.5 has been the second worst among all the videos. Observing the video No.5, we could find that halation lasted for 125 consecutive frames. Figure 9 shows an image of the halation. Since the halation covered overall area in each image, the block-matching method could not extract movement of the flexible cystoscope. Hence, in the case when handling patterns are judged incorrectly due to the halation, it is necessary to discriminate the handling patterns from handling patterns in other frames.

4.4 Future Works

From the experimental results shown in Sec. 4.2, this paper has indicated that our proposed system can be used as a trainer for beginner physicians. Although cystoscopic images are significantly unclear compared with images used in related works [3, 4, 10], it is seen that the proposed system works well on tracking the observation in a flexible cystoscopy. However, sometimes the tracking would fail due to the failure cases described in Sec. 4.3. In such case, it is necessary to estimate the actual position according to landmarks placed in the virtual bladder. One of the ways for generating landmarks is the use of HOG (Histogram of Oriented Gradients) which is robust against rotation and size difference. In addition, we would like to correct turbulent flows obtained from noisy images.

5. CONCLUSION

This paper has presented a system of tracking the observation in flexible cystoscopy. The proposed system discriminates three handling patterns of flexible cystoscope. To achieve this objective accurately, we proposed to extract ZNCC-based optical flows and three features that nd the represent the degree of the handling patterns from cystoscopic images. In addition, we proposed to discriminate sequential handling patterns of the flexible cystoscope. Experimental results using ten videos have shown the average correct ratio of the three handling patterns has been at least 90%. We also reproduced the observation in a flexible cystoscopy in a virtual 3D bladder we constructed.

Considering the failure cases for tracking the observation, we need to estimate the actual position in case of the tracking failure and correct turbulent flows obtained from cystoscopic images where bladder stones are floating and halation is happened. Besides, we would like to estimate the shape and size of the bladder using CT or MRI images.

6. REFERENCES

- [1] M. Froehner, M. A. Brausi, H. W. Herr, G. Muto, U E. Studer. "Complications following radical cystectomy for bladder cancer in the elderly" *European Urology*, vol.56, no.3, pp.443-454, 2009.
- [2] J. Key, D. Dhawan, D. K. Knapp, K. Kim, I. C. Kwon, K. Choi, J. F. Leary. "Method and apparatus for estimating the velocity vector of multiple vehicles on non-level and curved roads using a single camera" in *Proc. SPIE 8225, Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues X*, 82251F, 2012, pp.1-8.
- [3] H C. Choi, S. Y Oh. "Robust segment-based object tracking using generalized hyperplane approximation" *Pattern Recognition*, vol.45, no.8, pp.2980-2991, 2012.
- [4] A. Ramisa, G. Alenya, C. Torras. "Single image 3D human pose estimation from noisy observations" in *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp.2673-2680.
- [5] J. C. Yoo and T. Han. "Fast normalized cross-correlation" *Circ. Syst. Signal Process*, vol.28, no.6, pp.819-843, 2009.
- [6] L. Breiman. "Random Forests" *Machine Learning*, vol.45, no.1, pp.5-32, 2001.

- [7] S.L. Tanimoto. "Template Matching in Pyramids" *Computer Graphics and Image Processing*, vol.16, no.4, pp.356-369, 1981.
- [8] B.K.P Horn and B.G. Schunck. "Determining optical flow" *Artif. Intell.*, vol.17, pp.185-203, 1981.
- [9] M. Fosteller. "A k -sample slippage test for an extreme population" *Annals of Mathematical Statistics*, vol.19, no.1, pp.58-65, 1948.
- [10] L. Beyang, S. Gould, and D. Koller. "Single image depth estimation from predicted semantic labels" in *Proc. 2010 IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp.1253-1260.
- [11] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection" in *Proc.2005 IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp.886-893.

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Image Processing (IJIP)* aims to be an effective forum for interchange of high quality theoretical and applied research in the Image Processing domain from basic research to application development. It emphasizes on efficient and effective image technologies, and provides a central forum for a deeper understanding in the discipline by encouraging the quantitative comparison and performance evaluation of the emerging components of image processing.

We welcome scientists, researchers, engineers and vendors from different disciplines to exchange ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJIP.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with volume 7, 2013, IJIP is appearing with more focused issues. Besides normal publications, IJIP intends to organize special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

LIST OF TOPICS

The realm of International Journal of Image Processing (IJIP) extends, but not limited, to the following:

- Architecture of imaging and vision systems
- Character and handwritten text recognition
- Chemistry of photosensitive materials
- Coding and transmission
- Color imaging
- Data fusion from multiple sensor inputs
- Document image understanding
- Holography
- Image capturing, databases
- Image processing applications
- Image representation, sensing
- Implementation and architectures
- Materials for electro-photography
- New visual services over ATM/packet network
- Object modeling and knowledge acquisition
- Photographic emulsions
- Prepress and printing technologies
- Remote image sensing
- Autonomous vehicles
- Chemical and spectral sensitization
- Coating technologies
- Cognitive aspects of image understanding
- Communication of visual data
- Display and printing
- Generation and display
- Image analysis and interpretation
- Image generation, manipulation, permanence
- Image processing: coding analysis and recognition
- Imaging systems and image scanning
- Latent image
- Network architecture for real-time video transport
- Non-impact printing technologies
- Photoconductors
- Photopolymers
- Protocols for packet video
- Retrieval and multimedia

- Storage and transmission

- Video coding algorithms and technologies for ATM/p

CALL FOR PAPERS

Volume: 7 - Issue: 5

i. Paper Submission: November 30, 2013

ii. Author Notification: December 25, 2013

iii. Issue Publication: December 2013

CONTACT INFORMATION

Computer Science Journals Sdn Bhd

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6207 1607

006 03 2782 6991

Fax: 006 03 6207 1697

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2013
COMPUTER SCIENCE JOURNALS SDN BHD
B-5-8 PLAZA MONT KIARA
MONT KIARA
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6207 1607
006 03 2782 6991

FAX: 006 03 6207 1697
EMAIL: cscpress@cscjournals.org