Neurocomputing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

New image descriptors based on color, texture, shape, and wavelets for object and scene image classification

Sugata Banerji *, Atreyee Sinha, Chengjun Liu

Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, United States

ARTICLE INFO

Article history:
Received 23 August 2012
Received in revised form
5 December 2012
Accepted 1 February 2013
Communicated by Y. Yuan

Keywords:
Three dimensional local binary patterns (3D-LBP) descriptor
H-descriptor
H-fusion descriptor
Scale invariant feature transform (SIFT)
Pyramid histograms of visual words (PHOW)
Object and scene image classification

ABSTRACT

This paper presents new image descriptors based on color, texture, shape, and wavelets for object and scene image classification. First, a new three Dimensional Local Binary Patterns (3D-LBP) descriptor, which produces three new color images, is proposed for encoding both color and texture information of an image. The 3D-LBP images together with the original color image then undergo the Haar wavelet transform with further computation of the Histograms of Oriented Gradients (HOG) for encoding shape and local features. Second, a novel H-descriptor, which integrates the 3D-LBP and the HOG of its wavelet transform, is presented to encode color, texture, shape, as well as local information. Feature extraction for the H-descriptor is implemented by means of Principal Component Analysis (PCA) and Enhanced Fisher Model (EFM) and classification by the nearest neighbor rule for object and scene image classification. And finally, an innovative H-fusion descriptor is proposed by fusing the PCA features of the H-descriptors in seven color spaces in order to further incorporate color information. Experimental results using three datasets, the Caltech 256 object categories dataset, the UIUC Sports Event dataset, and the MIT Scene dataset, show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank, the Hierarchical Matching Pursuit, as well as LBP.

1. Introduction

The color cue is often applied by the human visual system for object and scene image classification. Indeed, color images, which contain more discriminative information than grayscale images, have been shown to perform better than grayscale images for image classification tasks [1–6]. Image descriptors defined in different color spaces usually help improve the identification of object, scene and texture image categories [7,2]. Image descriptors derived from different color spaces often exhibit different properties, among which are high discriminative power and relative stability over the changes in photographic conditions such as varying illumination. Color histogram and global color features and local invariant features often provide varying degrees of success against image variations such as rotation, viewpoint and lighting changes, clutter and occlusions [8,9].

Texture, shape, and local information contribute as well to object and scene image classification. Local Binary Patterns (LBP), for example, has been shown to be promising for recognition and classification of texture images [10–12]. The Histograms of Oriented Gradients (HOG) descriptor [13,14], which represents an image by histograms of the slopes of the object edges in an image, store information about the shapes contained in the image. As a result, the HOG descriptor has become a popular method for content based image retrieval. In addition, wavelets, such as the Haar wavelets have been widely applied for object detection in images [15].

Content-based image classification using large image databases has been a popular research area during recent years. Several state-of-the-art image classification methods involve the use of sparse coding and local coordinate coding [16]. Their use, coupled with efficient learning and classification methods, has been effectively demonstrated in [17] where over a million images belonging to a thousand categories in the ImageNet dataset were encoded and classified with very high speed and accuracy.

We present in this paper new image descriptors that integrate color, texture, shape, and wavelets for object and scene image classification. First, we introduce a new three Dimensional Local Binary Patterns (3D-LBP) descriptor for encoding the color and texture information of a color image. Specifically, the 3D-LBP descriptor produces three new color images from the original color image. Second, we apply the Haar wavelet transform to the three

* Corresponding author. Tel.: +1 732 983 8818; fax: +1 973 596 5777.
E-mail addresses: sb256@njit.edu, sugata.banerji@gmail.com (S. Banerji), as739@njit.edu (A. Sinha), chengjun.liu@njit.edu (C. Liu).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.02.014

new 3D-LBP color images and the original color image. We further calculate the Histograms of Oriented Gradients (HOG) of these Haar wavelet transformed images for encoding shape and local features. Third, we propose a novel H-descriptor, which integrates the 3D-LBP and the HOG of its wavelet transform, to encode color, texture, shape, and local information for object and scene image classification. Finally, we present a new H-fusion descriptor by fusing the Principal Component Analysis (PCA) features of the H-descriptors in the seven individual color spaces. Experimental results using three datasets, the Caltech 256 object categories dataset, the UIUC Sports Event dataset, and the MIT Scene dataset, show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT) [18,19], the Pyramid Histograms of visual Words (PHOW) [20], the Pyramid Histograms of Oriented Gradients (PHOG) [13,2], Spatial Envelope [21], Color SIFT four Concentric Circles (C4CC) [22], Object Bank [23], the Hierarchical Matching Pursuit [24], as well as LBP [10].

2. New image descriptors based on color, texture, shape, and wavelets

We present in this section new image descriptors based on color, texture, shape, and wavelets for object and scene image classification. In particular, first, we introduce a new three Dimensional Local Binary Patterns (3D-LBP) descriptor that produces three new color images for encoding both color and texture information of an image. These three new color images together with the original color image then undergo the Haar wavelet transform with further computation of the Histograms of Oriented Gradients (HOG) for encoding shape and local features. Second, we present a novel H-descriptor, which integrates the 3D-LBP and the HOG of its wavelet transform, for encoding color, texture, shape, and local information for object and scene image classification. Finally, we propose an innovative H-fusion descriptor that fuses the PCA features of the H-descriptors in the seven individual color spaces.

2.1. A new three dimensional local binary patterns (3D-LBP) descriptor

We now introduce a new three Dimensional Local Binary Patterns (3D-LBP) descriptor that produces three new color images for encoding both color and texture information of an image. The Local Binary Patterns (LBP) method derives the texture description of a grayscale image by comparing a center pixel with its neighbors [10,25,26].

In particular, for a 3×3 neighborhood of a pixel $\mathbf{p} = [x,y]^t$, $\mathbf{p}$ is the center pixel used as a threshold. The neighbors of the pixel $\mathbf{p}$ are defined as $N(\mathbf{p},i) = [x_i,y_i]^t$, $i = 0,1,\ldots,7$, where i is the number used to label the neighbor. The value of the LBP code of the center pixel $\mathbf{p}$ is calculated as follows:

$$LBP(\mathbf{p}) = \sum_{i=0}^{7} 2^i S\{G[N(\mathbf{p},i)] - G(\mathbf{p})\} \quad (1)$$

where $G(\mathbf{p})$ and $G[N(\mathbf{p},i)]$ are the gray level of the pixel $\mathbf{p}$ and its neighbor $N(\mathbf{p},i)$, respectively. S is a threshold function that is defined below:

$$S(x_i - x_c) = \begin{cases} 1, & \text{if } x_i \geq x_c \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

LBP tends to achieve grayscale invariance because only the signs of the differences between the center pixel and its neighbors are used to define the value of the LBP code as shown in Eq. (1). Fig. 1 shows a grayscale image on the left and its LBP image on the right. The two 3×3 matrices in the middle illustrate how the LBP code is computed for the center pixel whose gray level is 90. In particular, the center pixel functions as a threshold, and after thresholding the right 3×3 matrix reveals the signs of the differences between the center pixel and its neighbors. Note that the signs are derived from Eqs. (1) and (2), and the threshold value is 90, as the center pixel is used as the threshold in the LBP definition. The binary LBP code is 01001101, which corresponds to 77 in decimal.

LBP, however, does not encode color information, which is an effective cue for pattern recognition such as object and scene image classification [2,27,28]. The motivation for our new three dimensional LBP descriptor, or 3D-LBP descriptor, rests on the extension of the conventional LBP method to incorporate the color cue when encoding a color image. Specifically, given a color image, our 3D-LBP descriptor generates three new color images by applying three perpendicular LBP encoding schemes. Fig. 2 shows a color image, the three perpendicular LBP encoding schemes, and the three encoded color images generated by our 3D-LBP descriptor. The first LBP encoding scheme applies a 3×3 neighborhood, which is shown in pink color in the top row of the second column, to encode the red, green, and blue component images, respectively. The encoded three images then form a new color image that is displayed as the top image in the last column in Fig. 2. The outer pixels are discarded on all sides after performing the LBP operation and hence this image is smaller than the original image by one pixel on all sides. The second LBP encoding scheme utilizes a 3×3 neighborhood shown in pink color in the middle row of the second column to encode the rows across the red, green, and blue component images, and the encoded three images form a new color image that is shown as the middle image in the last column in Fig. 2. The third LBP encoding scheme uses a 3×3 neighborhood shown in pink color in the bottom row of the second column to encode the columns across the red, green, and blue component images, and the encoded three images form a new color image that is displayed as the bottom image in the last column in Fig. 2. Normally, after performing an LBP operation, we need to discard the outer pixels. However, since the number of color planes is just three, we cannot simply discard the top and bottom planes after performing the new LBP operations as shown in the second and third rows of Fig. 2. To solve this problem, we replicate

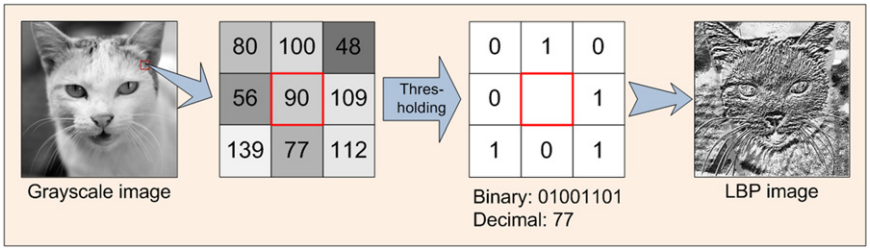

Fig. 1. A grayscale image, its LBP image, and the illustration of the computation of the LBP code for a center pixel with gray level 90.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

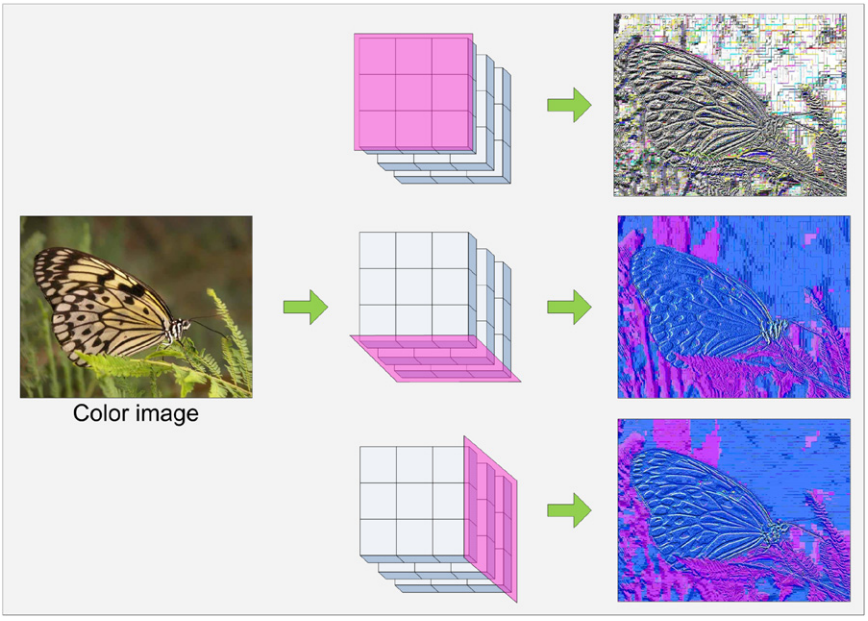

Fig. 2. A color image, the three perpendicular LBP encoding schemes, and the three encoded color images generated by our 3D-LBP descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the existing planes in a manner that puts an extra plane on either side of the three existing planes without copying a plane next to itself. For example, if the image is RGB, our new five-plane matrix will be BRGBR. After the 3D-LBP operation is done, these two new planes, i.e. the first and fifth planes of the five-plane image, are discarded to give us a three plane image again. The 3D-LBP descriptor thus encodes the color and texture information to generate three new color images as shown in the last column in Fig. 2, which will be further processed in order to extract shape and local information.

The extension of LBP to 3D-LBP is based on two reasons. First, the relative values of the intensities of the pixels at the same position in the three component planes of an image determines the color of that particular pixel. 3D-LBP captures this relationship between the pixel intensities across the planes to encode color. Second, 3D-LBP also encodes the relationship between a pixel's intensity value and the next pixel's intensity value, which essentially works as a high-pass filter in selective orientations to enhance local intensity variations in an image. Hence the two new color images produced by the 3D-LBP operation, other than the traditional LBP image, which are shown in the center and bottom right of Fig. 2, have enhanced local features separately in the vertical and horizontal directions, respectively.

2.2. A novel H-descriptor

Our 3D-LBP descriptor improves upon the conventional LBP method by means of encoding both color and texture information of a color image. The motivation for our next new descriptor, the H-descriptor, is based on incorporating additional useful and important features for object and scene image classification, such as shape and local features. Towards that end, we first compute the Haar wavelet transform of the color image and its new 3D-LBP color images. We then derive the Histograms of Oriented Gradients (HOG) [29] of the Haar wavelet transformed images for encoding both shape and local features. And we finally integrate these HOG features corresponding to the Haar wavelet transform of both the original color image and the 3D-LBP color images to form the H-descriptor, which encodes color, texture, shape, and local information for object and scene image classification.

The Haar wavelet transform [30], which extracts local information by means of enhancing local contrast, is applied to every component image of the color image and its three 3D-LBP color images. Haar is chosen over other wavelets due to its simplicity and computational efficiency. The 2D Haar wavelet transform is defined as the projection of an image onto the 2D Haar basis functions, which are formed by the tensor product of the one dimensional Haar scaling and wavelet functions [30,31]. The Haar scaling function $\phi(x)$ is defined below [30,32]:

$$\phi(x) = \begin{cases} 1, & 0 \le x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

A family of functions can be generated from the basic scaling function by scaling and translation [30,32]:

$$\phi_{i,j}(x) = 2^{i/2}\phi(2^i x - j) \tag{4}$$

As a result, the scaling functions $\phi_{i,j}(x)$ can span the vector spaces V^i, which are nested as follows: $V^0 \subset V^1 \subset V^2 \subset \cdots$ [33].

The Haar wavelet function $\psi(x)$ is defined as follows [30,32]:

$$\psi(x) = \begin{cases} 1, & 0 \le x < 1/2 \\ -1, & 1/2 \le x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The Haar wavelets are generated from the mother wavelet by scaling and translation [30,32]:

$$\psi_{i,j}(x) = 2^{i/2}\psi(2^i x - j) \tag{6}$$

The Haar wavelets $\psi_{i,j}(x)$ span the vector space W^i, which is the orthogonal complement of V^i in V^{i+1}: $V^{i+1} = V^i \oplus W^i$ [30,32]. The 2D Haar basis functions are the tensor product of the one dimensional scaling and wavelet functions [31].

Fig. 3 shows a color image, its three component images, their Haar wavelet transformed images, and the color Haar wavelet transformed image. One can see that these Haar wavelet transformed images reveal both local and shape information. The image in the upper left quadrant of the Haar wavelet transformed image is a lower resolution version of the original image while the other three quadrants contain the high-frequency information from the images along separate orientations. As the second step of

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

generating the proposed descriptor, the Haar wavelet transform of each of the three 3D-LBP color images is carried out.

To further encode local and shape information, we compute the HOG of the Haar wavelet transformed images. The idea of HOG rests on the observation that local object appearance and shape can often be characterized well by the distribution of local intensity gradients or edge directions [29]. Since 3D-LBP and Haar wavelet transform both work towards enhancing edges and other high-frequency local features, the choice of HOG as the next step seems logical as an image with enhanced edges is likely to yield more shape information than an unprocessed image. HOG features are derived based on a series of well-normalized local histograms of image gradient orientations in a dense grid [29]. In particular, the image window is first divided into small cells. For each cell, a local histogram of the gradient directions or the edge orientations is accumulated over the pixels of the cell. All the histograms within a block of cells are then normalized to reduce the effect of illumination variations. The blocks can be overlapped with each other for performance improvement. The final HOG features are formed by concatenating all the normalized histograms into a single vector. In our experiments, we divide the image into 3×3 parts and each histogram divides the gradients into nine bins. That makes the HOG vector 81 elements long. In the case of a color image, we repeat this process separately for the three component images and then concatenate the histograms. The length of a color HOG feature vector is 81×3, i.e. 243. Fig. 4 shows how the HOG descriptor is formed by the gradient histograms from a color image.

Specifically, we compute four HOG descriptors from the four quadrants of a Haar wavelet transformed image and then concatenate them to get the HOG descriptor of a Haar wavelet transformed image. Fig. 5 shows a color Haar wavelet transformed image, its four quadrant color images, their HOG descriptors, and the concatenated HOG descriptor. We finally integrate the HOG descriptors from the Haar wavelet transform of the component images of the color image and its 3D-LBP color images to form the H-descriptor, which

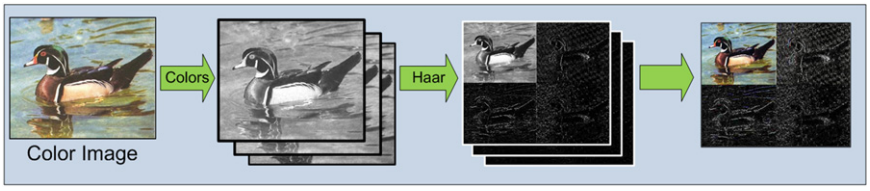

Fig. 3. A color image, its three component images, their Haar wavelet transformed images, and the color Haar wavelet transformed image. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

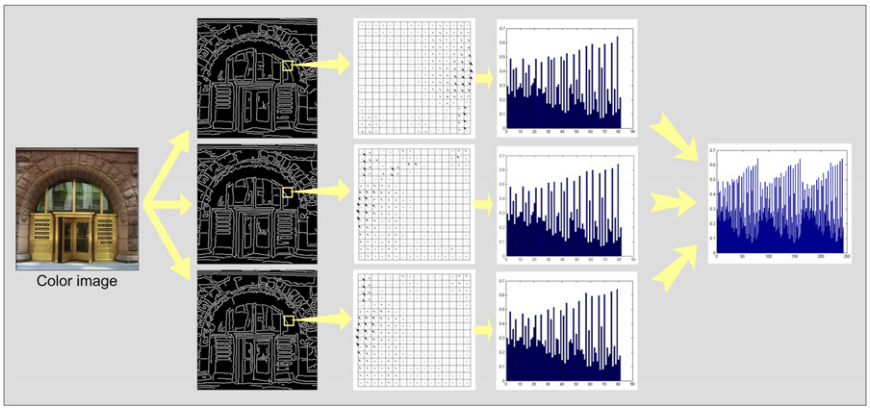

Fig. 4. A color image, the edge images of its three color component images, the orientation gradients of an example small area from every edge image, the three HOG descriptors for the three color component images, and the concatenated HOG descriptor for the whole color image. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

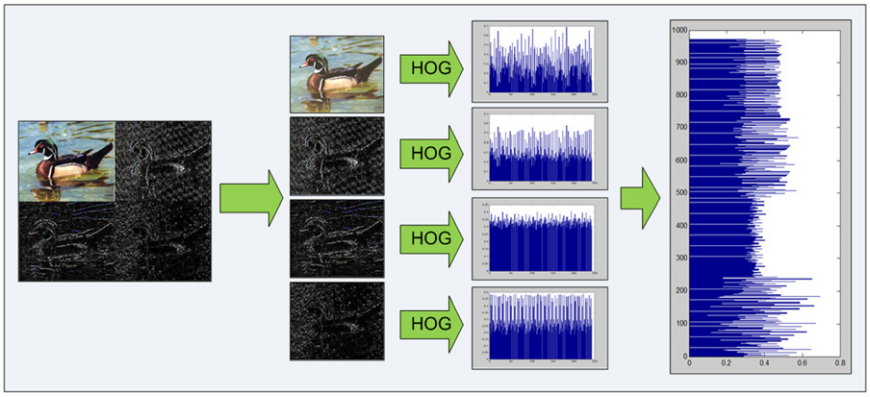

Fig. 5. A color Haar wavelet transformed image, its four quadrant color images, their HOG descriptors, and the concatenated HOG descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

encodes color, texture, shape, and local information for object and scene image classification. In particular, for a color image, our 3D-LBP descriptor first generates three new color images. The Haar wavelet transform then produces twelve wavelet transformed images from the twelve color component images of the color image and its three 3D-LBP color images. The HOG process further generates four HOG descriptors corresponding to each of the Haar wavelet transformed images. The HOG descriptors from all the Haar wavelet transformed images are finally concatenated to form a new descriptor, the H-descriptor. The dimensionality of this descriptor is 3888 which is the product of the size of the grayscale HOG vector and the total number of quadrants from all the twelve component images of the four Haar transformed color images ($81 \times 4 \times 12$). The time taken to compute the H-descriptor from an image is empirically seen to be directly proportional to the number of pixels in the image. For experiments done with a large number of images, the average feature extraction time is found to be 5.5 s per image on an Intel® Core™ i3-2120 3.30 GHz CPU with 8 GB RAM. Fig. 6 shows a color image, its 3D-LBP color images, the Haar wavelet transformed color images, and the H-descriptor derived from the concatenation of the HOG descriptors of the Haar wavelet transformed color images.

2.3. An innovative H-fusion descriptor

Color provides a very important cue for pattern recognition in general and for object and scene image classification in particular [1–7,27,28]. To further incorporate color information, we introduce an H-fusion descriptor that fuses the most expressive features of the H-descriptors in seven different color spaces, where the most expressive features are extracted by means of principal component analysis and the seven color spaces are the RGB, oRGB, HSV, YIQ, YCbCr, $I_1I_2I_3$, and DCS color spaces [27].

Principal component analysis, or PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation. Specifically, let $\mathcal{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is defined as follows [34]:

$$S = \mathcal{E}\{[\mathcal{X} - \mathcal{E}(\mathcal{X})][\mathcal{X} - \mathcal{E}(\mathcal{X})]^t\} \tag{7}$$

where $\mathcal{E}(\cdot)$ represents expectation and t the transpose operation. The covariance matrix S is factorized as follows [34]:

$$S = \Phi \Lambda \Phi^t \tag{8}$$

where $\Phi = [\phi_1 \phi_2 \cdots \phi_N]$ is an orthogonal eigenvector matrix and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ a diagonal eigenvalue matrix with diagonal elements in decreasing order. An important application of PCA is the extraction of the most expressive features of $\mathcal{X}$. Towards that end, we define a new vector $\mathcal{Y}$: $\mathcal{Y} = P^t \mathcal{X}$, where $P = [\phi_1 \phi_2 \ldots \phi_K]$, and $K < N$. The most expressive features of $\mathcal{X}$ thus define the new vector $\mathcal{Y} \in \mathbb{R}^K$, which consists of the most significant principal components.

Next, we briefly review the seven color spaces used to define our H-fusion descriptor. The RGB color space, whose three component images represent the red, green, and blue primary colors, is the common tristimulus space for color image representation. Other color spaces are usually derived from the RGB color space using either linear or nonlinear transformations. The $I_1I_2I_3$ color space is defined by the following linear transformation from the RGB color space [35]: $I_1 = (R+G+B)/3$, $I_2 = (R-B)/2$, $I_3 = (2G-R-B)/4$. The HSV (hue, saturation, and value) color space, however, is derived nonlinearly from the RGB color space [36]:

$$H = \begin{cases} 60\left(\dfrac{G-B}{\delta}\right) & \text{if } MAX = R \\ 60\left(\dfrac{B-R}{\delta} + 2\right) & \text{if } MAX = G \\ 60\left(\dfrac{R-G}{\delta} + 4\right) & \text{if } MAX = B \end{cases}$$

$$S = \begin{cases} \delta/MAX & \text{if } MAX \neq 0 \\ 0 & \text{if } MAX = 0 \end{cases} \qquad V = MAX \tag{9}$$

where $MAX = \max(R,G,B)$, $MIN = \min(R,G,B)$, and $\delta = MAX - MIN$.

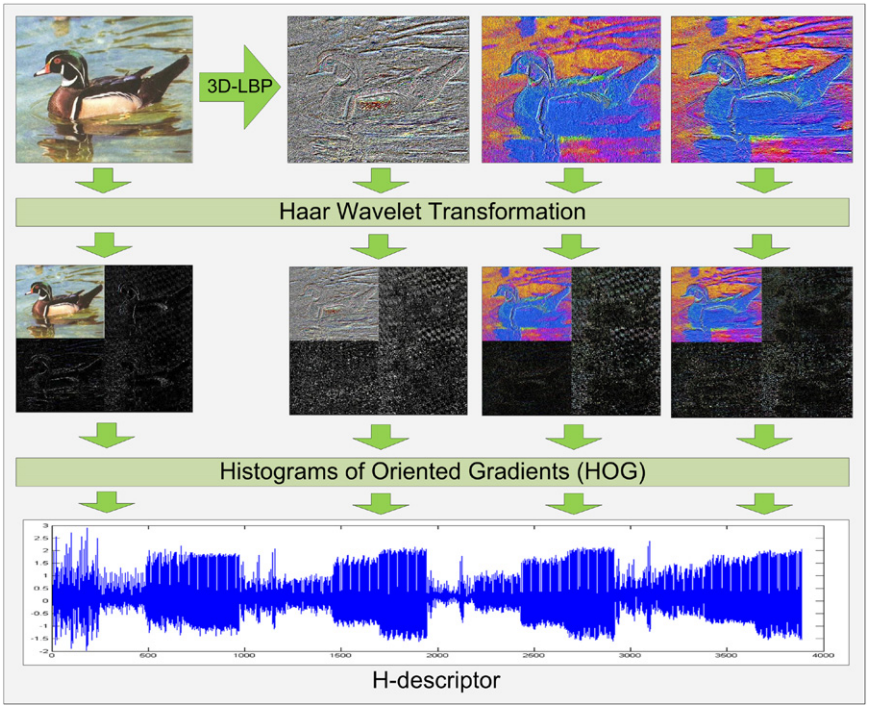

Fig. 6. A color image, its 3D-LBP color images, the Haar wavelet transforms of these color images, and the H-descriptor formed by the concatenation of the HOG descriptors of these Haar transform images. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

The remaining four color spaces are transformed from the RGB color space using linear transformations. The YCbCr color space is defined as follows [37]:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112.000 \\ 112.000 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (10)$$

The YIQ color space is defined as given below [38]:

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5957 & -0.2745 & -0.3213 \\ 0.2115 & -0.5226 & 0.3111 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (11)$$

The three component images L, C_1, and C_2 of the oRGB color space are defined as follows [39]:

$$\begin{bmatrix} L \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5000 & 0.5000 & -1.0000 \\ 0.8660 & -0.8660 & 0.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (12)$$

The discriminating color space or DCS defines discriminating component images by means of a linear transformation from the RGB color space [27]: $[D_1,D_2,D_3]^t = W_D[R,G,B]^t$. The transformation matrix, $W_D \in \mathbb{R}^{3\times3}$, is derived through a procedure of discriminant analysis [27,34]. Fig. 7 shows a color image, its grayscale image, and the color component images in the oRGB, RGB, YIQ, HSV, $I_1I_2I_3$, YCbCr, and DCS color spaces, respectively.

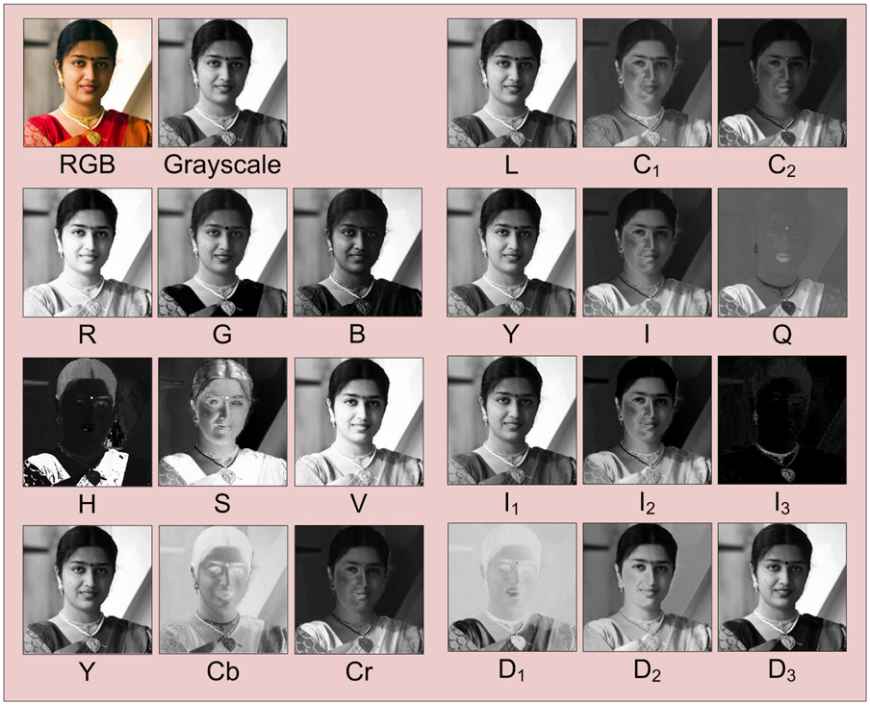

Fig. 7. A color image, its grayscale image, and the color component images in the oRGB, RGB, YIQ, HSV, $I_1I_2I_3$, YCbCr, and DCS color spaces, respectively. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

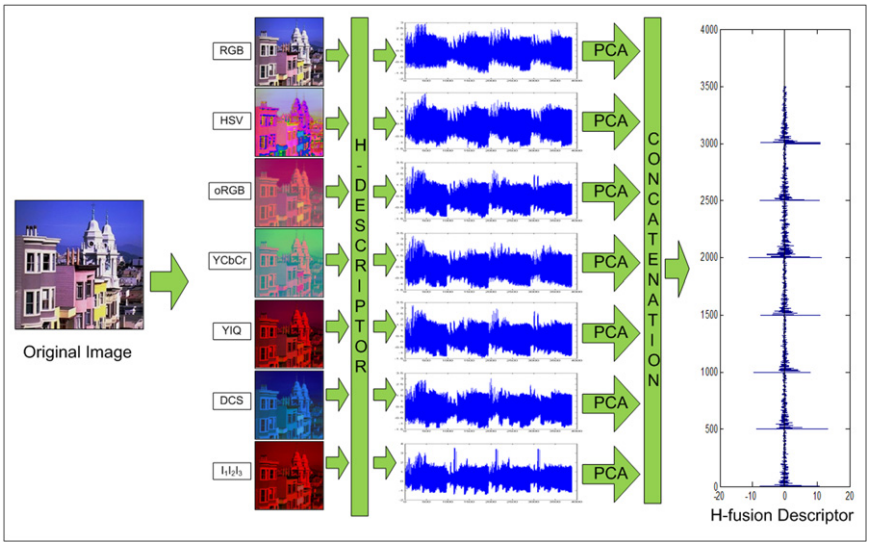

Fig. 8. A color image, its corresponding color images in the seven color spaces, the H-descriptors of the color images, the PCA process, the concatenation process, and the H-fusion descriptor. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

The proposed H-fusion descriptor is derived by first computing the H-descriptors in the seven color spaces, then extracting the most expressive features of the H-descriptors using PCA, and finally concatenating these most expressive features from the seven color spaces. Specifically, we take each color image, convert it to six different color spaces from the RGB color space, and compute the H-descriptor from each of these six color images and the RGB image. Next, we extract the most expressive features from each of these seven H-descriptors using PCA and then concatenate these seven sets of PCA features to get the H-fusion descriptor. The number of PCA features selected from each of the H-descriptors depends on the size of the dataset and the size of the training set. Fig. 8 shows a color image, its corresponding color images in the seven color spaces, the H-descriptors of the color images, the PCA process, the concatenation process, and the H-fusion descriptor. In the example shown in the figure, 500 PCA features are shown from each of the seven H-descriptors to eventually form a 3500-dimensional H-fusion descriptor.

3. The enhanced fisher model for feature extraction and the nearest neighbor classification rule – the EFM-NN classifier

Object and scene image classification using the new descriptors introduced in the preceding section is implemented using the Enhanced Fisher Model (EFM) for feature extraction [40] and the Nearest Neighbor (NN) to the mean classification rule for classification. We call this EFM feature extraction and NN classification procedure the EFM-NN classifier.

In pattern recognition, a popular method, Fisher's Linear Discriminant (FLD), applies first PCA for dimensionality reduction and then discriminant analysis for feature extraction. PCA is discussed in the previous section, and discriminant analysis often optimizes a criterion defined on the within-class and between-class scatter matrices S_w and S_b, which are defined as follows [34]:

$$S_w = \sum_{i=1}^{L} P(\omega_i)\mathcal{E}\{(\mathcal{Y}-M_i)(\mathcal{Y}-M_i)^t \mid \omega_i\} \tag{13}$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i-M)(M_i-M)^t \tag{14}$$

where $P(\omega_i)$ is *a priori* probability, ω_i represent the classes, and M_i and M are the means of the classes and the grand mean, respectively. One discriminant analysis criterion is J_1: $J_1 = tr(S_w^{-1}S_b)$, and J_1 is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1}S_b$ [34]:

$$S_w^{-1}S_b\Psi = \Psi\Delta \tag{15}$$

where Ψ and Δ are the eigenvector and eigenvalue matrices of $S_w^{-1}S_b$, respectively. The discriminating features are defined by projecting the pattern vector $\mathcal{Y}$ onto the eigenvectors of Ψ:

$$\mathcal{Z} = \Psi^t\mathcal{Y} \tag{16}$$

$\mathcal{Z}$ thus contains the discriminating features for image classification.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix [40]. As a result, the Enhanced Fisher Model (EFM) is developed to improve upon the generalization performance of the FLD method [40]. Specifically, the EFM method improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices [40]. The simultaneous diagonalization reveals that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. Since the small eigenvalues tend to encode noise [40], they cause the whitening step to fit for misleading variations, and this leads to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance) [40].

Fig. 9. Some sample images from (a) the Caltech 256 dataset, (b) the UIUC Sports Event dataset, and (c) the MIT Scene dataset.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

4. Experiments

We assess our proposed descriptors for object and scene image classification using three popular datasets, namely the Caltech 256 dataset [41], the UIUC Sports Event dataset [42], and the MIT Scene dataset [21]. Specifically, we first assess the H-descriptor in seven different well-known color spaces, apart from four randomly generated color spaces, and then compare the H-fusion descriptor with other popular descriptors, such as combinations of Scale Invariant Feature Transform (SIFT) [18,19] with other descriptors [42,24], the Pyramid Histograms of visual Words (PHOW) descriptor [20], LBP [10] and Pyramid Histograms of Oriented Gradients (PHOG) [13] based features [2], Spatial Envelope [21], Color SIFT four Concentric Circles (C4CC) [22], and other approaches such as Object Bank [23] and Hierarchical Matching Pursuit [24].

4.1. Datasets

In this section, we briefly describe three publicly available and fairly challenging image datasets. All of these datasets are widely used for evaluating the performance of object and scene image descriptors and classification methods.

4.1.1. The caltech 256 dataset

The Caltech 256 dataset [41] holds 30,607 images divided into 256 object categories and a clutter class. The images have high intra-class variability and high object location variability [41]. Each category contains a minimum of 80 images and a maximum of 827 images. The mean number of images per category is 119. The images represent a diverse set of lighting conditions, poses, backgrounds, and sizes [41]. Images are in color, in JPEG format with only a small percentage in grayscale. The average size of each image is 351×351 pixels. Some sample images from this dataset are shown in Fig. 9(a), which reveals that some classes like lighthouse and minaret have very similar visual appearance and hence their inter-class variability is low.

4.1.2. The UIUC sports event dataset

The UIUC Sports Event dataset [42] contains eight sports event categories: badminton (200 images), bocce (137 images), croquet (236 images), polo (182 images), rock climbing (194 images), rowing (250 images), sailing (190 images), and snowboarding (190 images). The mean image size is 845×1077 pixels. Most of the images are color jpeg images, with a small percentage in grayscale. A few sample images from this dataset are shown in Fig. 9(b). This dataset contains indoor and outdoor scenes and some classes like badminton and bocce contain both. In some of the classes like bocce and croquet, the interclass distance is very low for the image background and the human poses provide the only information for classification.

4.1.3. The MIT scene dataset

The MIT Scene dataset [21] has 2688 images classified as eight categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. All of the images are in color, in JPEG format, and the size of each image is 256×256 pixels. There is a large variation in light, content and angles, along with a high intra-class variation [21]. Fig. 9(c) shows some images from this dataset.

4.2. Comparative assessment of the H-descriptor in different color spaces

We now assess the H-descriptor in seven different color spaces – the RGB, oRGB, HSV, YIQ, YCbCr, $I_1I_2I_3$, and DCS color spaces – for image classification performance using the three datasets. Note that for some large scale images, we resize them so that the larger dimension is reduced to 400 pixels. To derive the H-descriptor from each image, we first compute the 3D-LBP descriptor to produce three new color images. We then calculate the Haar wavelet transform of the 3D-LBP color images and the original color image. We further compute the HOG descriptors from the Haar wavelet transform of the component images of the color image and its 3D-LBP color images. We finally derive the H-descriptor by concatenating the HOG descriptors of the Haar wavelet transformed color images. We transform each image in the seven color spaces and perform the same operations to construct the seven different color H-descriptors.

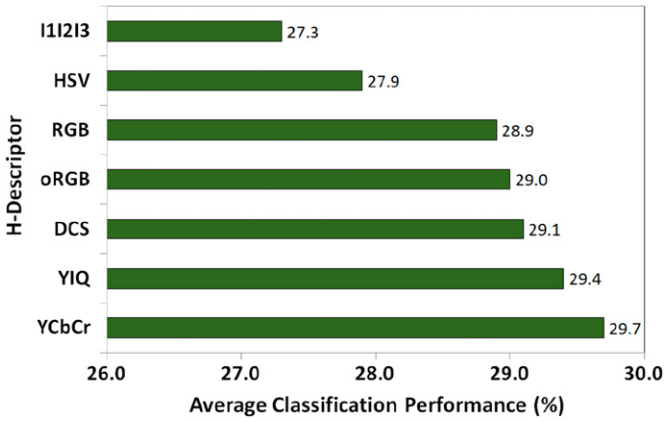

Fig. 10. The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, RGB, oRGB, DCS, YIQ, and YCbCr color spaces using the EFM-NN classifier on the Caltech 256 dataset.

We next apply PCA to reduce the dimensionality of the H-descriptors to derive the most expressive features. For each of the datasets, the set of training images, which do not contain any of the images to be used for testing, are used for doing the PCA. The number of features that are chosen after PCA depends on the size of the training data – for training data matrices with rank less than 2000 we choose (rank-1) PCA features, and for training data matrices with rank greater than 2000 we choose the first 2000 PCA features. We empirically find the number of PCA values that work best with EFM for a particular dataset and then perform EFM to obtain the most discriminatory features for classification. The number of features obtained after EFM is one less than the number of categories in the dataset. For instance, the EFM process produces a 255-dimensional vector for the Caltech 256 dataset and a 7-dimensional vector for the UIUC Sports Event and MIT Scene datasets. We finally use the nearest neighbor rule for image classification on this vector.

For the Caltech 256 dataset, we use a protocol defined in [41]. On this dataset, we conduct experiments for the H-descriptors from seven different color spaces. For each class, we use 50 images for training and 25 images for testing. The data splits are the ones that are provided on the Caltech website [41]. Fig. 10 shows the detailed performance of the H-descriptors using the EFM-NN classifier on the Caltech 256 dataset. The horizontal axis indicates the average classification performance, which is the percentage of correctly classified images averaged across the 256 classes and the five runs of the experiments, and the vertical axis shows the seven different H-descriptors in the seven color spaces. Among the different H-descriptors, the H-descriptor in the YCbCr color space achieves the best average classification performance of 29.7%, followed by the H-descriptors in the YIQ, DCS, oRGB, RGB, HSV and $I_1I_2I_3$ color spaces with the average classification performance of 29.4%, 29.1%, 29.0%, 28.9%, 27.9%, and 27.3%, respectively.

For the UIUC Sports Event dataset, we use a protocol defined in [42], which specifies that for each class in this dataset, 70 images are used for training and 60 images for testing. To achive more reliable performance, we repeat our experiments five times using

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

random splits of the data, and no overlapping occurs between the training and the testing sets of the same split. Fig. 11 shows that the H-descriptor in the YIQ color space is the best descriptor with 82.5% average classification performance followed in order by the H-descriptors in the RGB, oRGB, YCbCr, DCS, HSV and $I_1I_2I_3$ color spaces with 82.3%, 81.8%, 81.7%, 81.6%, 81.6% and 80.7% success rates, respectively. Again the horizontal axis indicates the average classification performance and the vertical axis the H-descriptors in the seven color spaces.

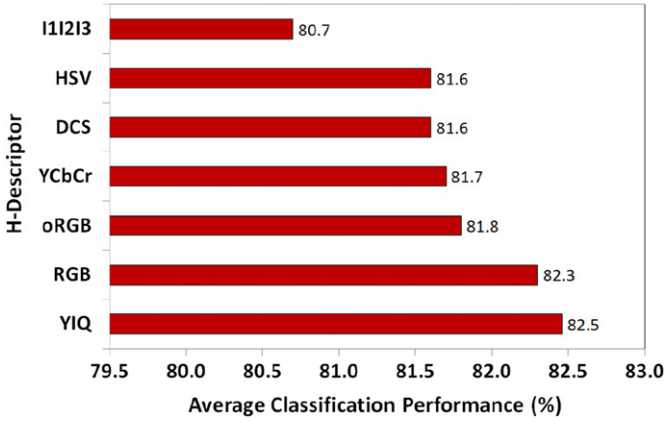

Fig. 11. The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, DCS, YCbCr, oRGB, RGB, and YIQ color spaces using the EFM-NN classifier on the UIUC Sports Event dataset.

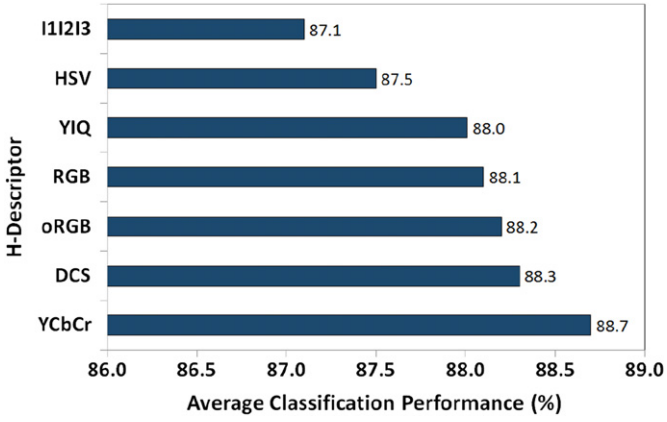

Fig. 12. The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, YIQ, RGB, oRGB, DCS, and YCbCr color spaces using the EFM-NN classifier on the MIT Scene dataset.

For the MIT Scene dataset, we use 250 images from each class for training and the rest of the images for testing. All experiments are performed for five random splits of the data. Fig. 12 reveals that the H-descriptor in the YCbCr color space performs the best with 88.7% average classification rate. The H-descriptors in the DCS, oRGB, RGB, YIQ, HSV and $I_1I_2I_3$ color spaces correctly classify on an average 88.3%, 88.2%, 88.1%, 88.0%, 87.5% and 87.1% of the images respectively. Again the horizontal axis shows the average classification performance and the vertical axis the H-descriptors.

4.3. Random color spaces and performance of the H-descriptor in these color spaces

To further establish the robustness of the proposed H-descriptor for object and scene image classification, we generate four random color spaces and assess the classification performance using our descriptor in these color spaces. To generate a random color space, we create a 3×3 transformation matrix with randomly chosen elements:

$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{17}$$

where R_1, R_2 and R_3 are the three color components in the new random color space, and $W_{ij} \in (-1,1)$ are pseudorandom numbers. The three color components in the new color space are thus given by

$$\begin{aligned} R_1 &= W_{11}R + W_{12}G + W_{13}B \\ R_2 &= W_{21}R + W_{22}G + W_{23}B \\ R_3 &= W_{31}R + W_{32}G + W_{33}B \end{aligned} \tag{18}$$

We next assess the classification performance of the proposed H-descriptor in four random color spaces. In particular, we generate four such random transformation matrices and name the resulting color spaces random color spaces 1, 2, 3 and 4 (RCS1, RCS2, RCS3 and RCS4). We then transform the original images from each of the three datasets mentioned Section 4.1 into each of these color spaces and subsequently generate the H-descriptor and use the same training and testing framework as we use for the other seven color spaces. Fig. 13 shows the component images of the color image shown before in Fig. 7 in the four random color spaces used for our experiments. It should be noted that the images shown here are just four instances of what the components of a random color space could look like.

The results of the classification experiments are shown in Fig. 14 with the performance in the RGB color space for reference. Here, the horizontal axis shows the H-descriptors in different color spaces, and the different datasets while the vertical axis shows the average classification performance. The performance of the H-descriptor in RCS1, RCS2, RCS3 and RCS4 remains, in all cases except one, within 2% of the performance of the H-descriptor in the RGB color space. First,

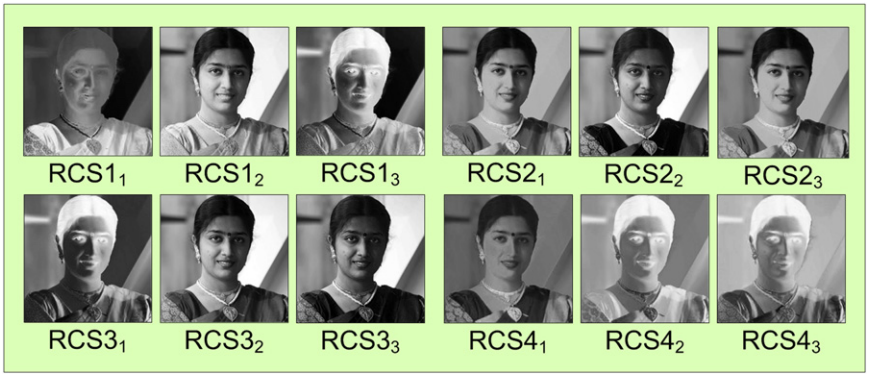

Fig. 13. The color component images of the image from Fig. 7 in the four random color spaces, namely RCS1, RCS2, RCS3 and RCS4 color spaces, respectively.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

these results show that the performance is random and unpredictable. In some cases it is more than the RGB H-descriptor performance and in other cases it is less. This indicates that simply transforming the color space does not increase the performance – the exact nature of the transformation is also important. Second, the results demonstrate that for the H-descriptors in RCS1, RCS2, RCS3 and RCS4 color spaces, the classification success rates stay reasonably close to the classification rate of the H-descriptor in the RGB color space. This indicates that the proposed descriptor is robust enough to yield stable performance under unpredictable changes in the color component values.

4.4. Comparative assessment of the grayscale H-descriptor, the color H-descriptors and the H-fusion descriptor

In this section, we attempt to investigate the importance of using color information for classification, and then try to justify the fusion of H-descriptors in the seven different color spaces to form the H-fusion descriptor. Towards that end, we generate a grayscale H-descriptor and comparatively evaluate its classification performance with the RGB H-descriptor and H-fusion descriptor.

The 3D-LBP operation, which is the first step of generating the H-descriptor, is only defined for a color image, i.e. an image with three component planes. This is because the 3D-LBP captures the variations in pixel intensities across the color planes thus encoding image color information. To generate the H-descriptor for a grayscale image, we first have to convert it to a three-plane image. In particular, for this experiment we take each color image with three planes and convert it to a grayscale image with just one plane by forming a weighted sum of the R, G, and B components:

$$Gray = 0.2990R + 0.5870G + 0.1140B \tag{19}$$

Note that these are the same weights used to compute the Y component of the YIQ color space. Then we replicate that single plane twice to form a three-plane image again. We subsequently generate the H-descriptor from this image and perform classification using the EFM-NN classifier.

To create the H-fusion descriptor, the H-descriptor is computed from each image in each of the seven well-defined color spaces as described in Section 4.2. Then, after reducing the dimensionality of each of these seven feature vectors to min(2000,*rank*−1) PCA features, we concatenate them and form the H-fusion descriptor. Subsequently, we further reduce the dimensionality using PCA and extract the most discriminatory features using EFM. Here also, the final number of features before classification is one less than the number of categories.

Fig. 15 compares the classification performance of the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor. Specifically, for the Caltech 256 dataset, the grayscale H-descriptor yields a success rate of 24.2%. Simply including RGB color information takes the correct classification rate up to 28.9% and the fusion of color spaces increases this by a further 4.7% to correctly classify 33.6% of the images. On the UIUC Sports Event dataset, the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor show classification rates of 77.4%, 82.3% and 86.2% respectively, thus demonstrating a significant advantage of using color. For the MIT Scene dataset, the classification rates obtained for the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor are 83.7%, 88.1%, and 90.8% respectively. Thus the H-fusion descriptor increases classification performance by over 7% from the grayscale H-descriptor, which is a quite high improvement for a dataset of this size and complexity. It should be noted that for the MIT Scene dataset, 250 images from each class are used for training in these experiments.

On comparing Fig. 15 with Figs. 10–12, and 14, we find that the classification performance of the grayscale H-descriptor is not only less than the RGB H-descriptor, but it is also less than the classification performance of the H-descriptor in any other color space as well. This is in accordance with the principle behind the 3D-LBP operation which is the first step of generating the H-descriptor. The 3D-LBP operation has been designed specifically to extract color information from the difference in pixel values in the three color component images, and since this difference is zero in a grayscale image, the H-descriptor does not perform as well for grayscale images as it does for color images. Also, the H-fusion descriptor performs better than the H-descriptor in any of the individual color spaces which justifies the fusion of H-descriptors from different color spaces.

4.5. Comparative assessment of the H-fusion descriptor and some popular state-of-the-art image descriptors

In this section we evaluate the performance of the proposed H-fusion descriptor on the three datasets described in Section 4.1. We first compare our H-fusion descriptor with the popular and robust SIFT-based Pyramid Histograms of visual Words (PHOW) descriptor [20]. For fair comparison, both descriptors apply the EFM-NN classifier for image classification. We then compare our H-fusion descriptor with some other popular state-of-the-art

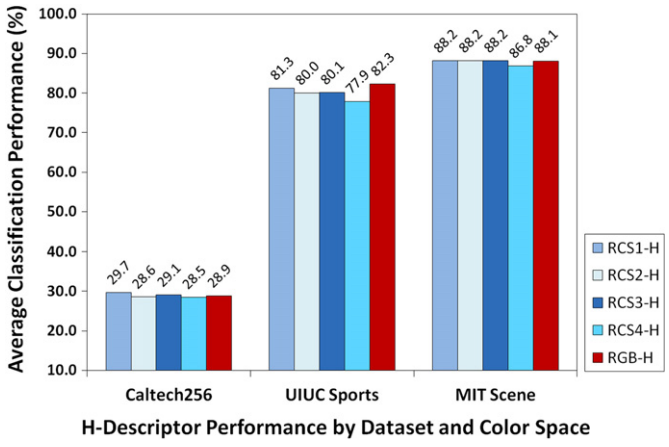

Fig. 14. A comparison of the average classification performances of the H-descriptor in the RGB color space and the four random color spaces RCS1, RCS2, RCS3, and RCS4 on the three image datasets. Note that all the five descriptors apply the EFM-NN classifier.

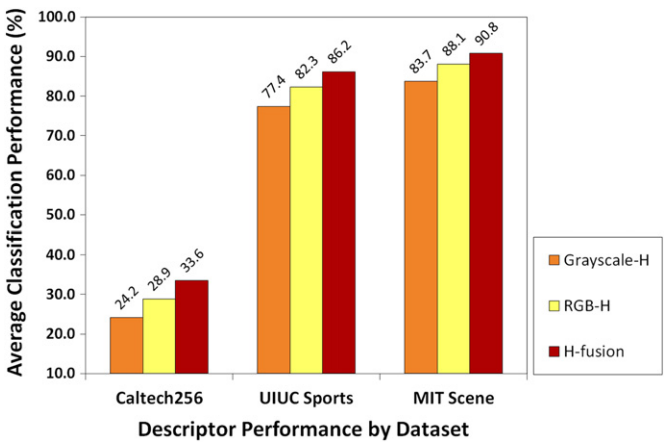

Fig. 15. A comparison of the average classification performances of the H-descriptor in grayscale, in the RGB color space and the H-fusion descriptor on the three image datasets. Note that all the three descriptors apply the EFM-NN classifier.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

descriptors using the image classification performance reported in the published papers.

To make a comparative assessment of the H-fusion descriptor with a popular SIFT-based feature, we generate the Pyramid Histograms of visual Words (PHOW) feature vector [20] using the software package VLFeat [43]. Here feature extraction is a three-step process. First SIFT features are extracted from images using a fast SIFT process. In this algorithm, SIFT descriptors are computed at points on a dense regular grid instead of the SIFT-generated interest points [44,20]. Next, the SIFT features are subjected to K-means clustering with $K=1000$ to form a visual vocabulary. Finally, the images are spatially tiled into 2×2 parts and the histograms of visual words are computed for the SIFT features from each part. These four histograms are concatenated to generate the final PHOW feature vector. For a color image, the same process is repeated for the three color component images and the feature vectors are concatenated. We use the grayscale PHOW and the color PHOW feature vectors with our EFM-NN classifier to compare the classification performance. Please note that the SIFT process applied here is an optimized C code that is 30–70 times faster than the conventional SIFT method [43]. In comparison, our H-descriptor is implemented using the MATLAB code that is not optimized in terms of computational efficiency. However, the vector generation time for the color PHOW is slightly longer than that for the color H-descriptor. For both PHOW and H-fusion descriptors, we apply PCA for dimensionality reduction and the EFM-NN for classification in order to make a fair comparison.

Fig. 16 shows that our H-fusion descriptor has an image classification performance better than both the grayscale and the color PHOW descriptors on the Caltech 256 dataset. Note that the horizontal axis of this graph lists the three descriptors and the three datasets while the vertical axis shows the average classification performance as a percentage. In particular, the H-fusion descriptor achieves the average classification rate of 33.6%, compared to the color-PHOW descriptor with the average classification rate of 29.9% and to the grayscale-PHOW descriptor with the average classification rate of 25.9%, respectively. Note that the classification performance for the Caltech 256 dataset is quite low, because this dataset has a very high intra-class variability and in several cases the object occupies a small portion of the full image.

Fig. 16 also displays the image classification performance on the UIUC Sports Event dataset. Specifically, the H-fusion descriptor correctly classifies 86.2% of the images and performs better than both the grayscale and the color PHOW descriptors, which achieve the average classification performance of 76.4% and 79.0%, respectively. Using this UIUC Sports Event dataset, we further compare our H-fusion descriptor with some popular state-of-the-art descriptors and methods, such as the Hierarchical Matching Pursuit [24], Object Bank approach [23] and variations of the popular Scale Invariant Feature Transform (SIFT) [19] descriptor [24,42]. Note that the performance reported here for the competing methods are from the published papers. Table 1 shows that our H-fusion descriptor achieves the best classification performance of 86.2% compared to HMP [24] with classification performance of 85.7%, to SIFT+SC [24] with classification performance of 82.7%, to Object Bank [23] with classification performance of 76.3% and to the SIFT+GGM [42] method with classification performance of 73.4%.

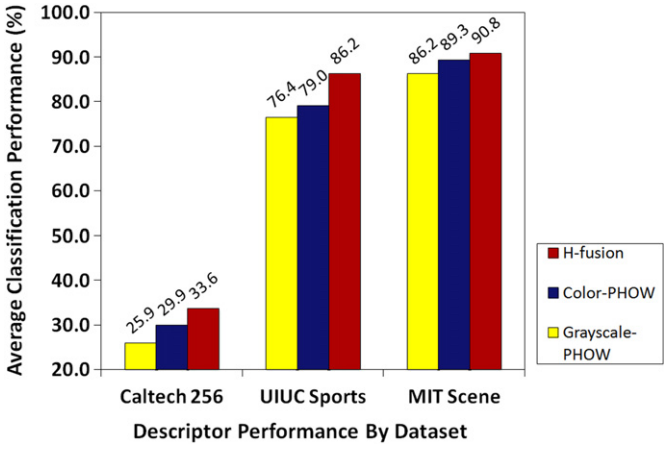

Fig. 16. A comparison of the average classification performances of the color-PHOW descriptor, the grayscale-PHOW descriptor, and the proposed H-fusion descriptor on the three image datasets. Note that all the three descriptors apply the EFM-NN classifier.

On the MIT Scene dataset, we perform two sets of experiments with our H-fusion descriptor. First we use 250 images from each class for training and the rest of the images for testing. In this set of experiments, the proposed H-fusion descriptor yields an average success rate of 90.8% and exceeds the performance achieved by the PHOW descriptors. Fig. 16 shows the image classification performance on this dataset as well. Specifically, the H-fusion descriptor correctly classifies 90.8% of the images and performs better than both the grayscale and the color PHOW descriptors, which achieve the average classification performance of 86.2% and 89.3%, respectively. In the next set of experiments we use 100 images from each class for training and the remaining images for testing. We further compare our descriptor with some widely used state-of-the-art descriptors and classification approaches such as the Spatial Envelope [21], Color SIFT four Concentric Circles (C4CC) [22], Color Grayscale LBP Fusion (CGLF) [2] and Pyramid Histograms of Oriented Gradients (PHOG) [13,2]. Here also, the results achieved by other researchers are reported directly from their published work. Table 2 shows that with 250 training images, the proposed H-fusion descriptor achieves the best classification performance of 90.8% as compared to CGLF+PHOG [2] with a classification performance of 89.5%, to CGLF [2] with a classification performance of 86.6% and to PHOG [13,2] with a classification performance of 79.1%. With 100 training images per class, our H-fusion descriptor again yields the best classification performance of 87.7%, as compared to Color SIFT four Concentric Circles (C4CC) [22] with a classification performance of 86.7%, to CGLF+PHOG [2] with a classification performance of 84.3%, to Spatial

Table 1
Comparison of the classification performance (%) of the H-fusion descriptor with other popular methods on the UIUC Sports Event Dataset.

Method	**#train=560, #test=480**	
H-fusion	**Proposed descriptor**	**86.2**
HMP	[24]	85.7
SIFT+SC	[24]	82.7
OB	[23]	76.3
SIFT+GGM	[42]	73.4

Table 2
Comparison of the classification performance (%) of the H-fusion descriptor with other popular methods on the MIT Scene Dataset.

Method	**#train=2000, #test=688**	
H-fusion	**Proposed descriptor**	**90.8**
CGLF+PHOG	[2]	89.5
CGLF	[2]	86.6
PHOG	[2]	79.1
	#train=800, #test=1888	
H-fusion	**Proposed descriptor**	**87.7**
C4CC	[22]	86.7
CGLF+PHOG	[2]	84.3
SE	[21]	83.7
CGLF	[2]	80

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

Envelope with a classification performance of 83.7%, and to CGLF [2] with a classification performance of 80.0%.

5. Conclusion

We have presented in this paper new image descriptors based on color, texture, shape, and wavelets for object and scene image classification. The contributions of the paper are manifold, and in particular, we have first presented a new three Dimensional Local Binary Patterns (3D-LBP) descriptor for encoding both color and texture information of a color image. We have then proposed a novel H-descriptor, which integrates the 3D-LBP and the HOG of its wavelet transform, to encode color, texture, shape, and local information. We have also comparatively assessed the H-descriptor in seven different well known color spaces – the RGB, the HSV, the YCbCr, the oRGB, the $I_1I_2I_3$, the YIQ, and the discriminating color spaces – for image classification performance. Apart from these, we assessed the classification performance of the H-descriptor in four randomly generated color spaces and grayscale. We have finally presented a new H-fusion descriptor by fusing the PCA features of the H-descriptors in the seven color spaces. Experimental results using three datasets show that the proposed new H-fusion descriptor achieves significantly better image classification performance than the H-descriptors in individual color spaces and grayscale. The H-fusion descriptor also achieves better image classification performance than other popular descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank, the Hierarchical Matching Pursuit, as well as LBP.

Acknowledgments

The authors would like to thank the associate editor and the anonymous reviewers for their critical and constructive comments and suggestions, which help improve the quality of the paper.

References

[1] C. Liu, V. Mago (Eds.), Cross Disciplinary Biometric Systems, Springer, 2012.
[2] S. Banerji, A. Verma, C. Liu, Novel color LBP descriptors for scene and image texture classification, in: Proceedings of the 15th International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, Nevada, 18–21 July 2011, pp. 537–543.
[3] C. Liu, Extracting discriminative color features for face recognition, Pattern Recognition Lett. 32 (2011) 1796–1804.
[4] C. Liu, J. Yang, ICA color space for pattern recognition, IEEE Trans. Neural Networks 2 (2009) 248–257.
[5] C. Liu, The Bayes decision rule induced similarity measures, IEEE Trans. Pattern Anal. Mach. Intel. 6 (2007) 1116–1117.
[6] C. Liu, Enhanced independent component analysis and its application to content based face image retrieval, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 34 (2004) 1117–1127.
[7] A. Verma, S. Banerji, C. Liu, A new color SIFT descriptor and methods for image category classification, in: Proceedings of the International Congress on Computer Applications and Computational Science, Singapore, 4–6 December 2010, pp. 819–822.
[8] G. Burghouts, J.-M. Geusebroek, Performance evaluation of local color invariants, Comput. Vis. Image Understanding 113 (2009) 48–62.
[9] H. Stokman, T. Gevers, Selection and fusion of color models for image feature detection, IEEE Trans. Pattern Anal. Mach. Intel. 29 (2007) 371–381.
[10] T. Ojala, M. Pietikainen, D. Harwood, Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, in: Proceedings of the International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994, pp. 582–585.
[11] C. Zhu, C. Bichot, L. Chen, Multi-scale color local binary patterns for visual object classes recognition, in: Proceedings of the International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010, pp. 3065–3068.
[12] M. Crosier, L. Griffin, Texture classification with a dictionary of basic image features, in: Proceedings of the Proceedings of the Computer Vision and Pattern Recognition, Anchorage, Alaska, 23–28 June 2008, pp. 1–7.
[13] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007, pp. 401–408.
[14] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes, Trainable classifier-fusion schemes: an application to pedestrian detection, in: Proceedings of the Twelfth International IEEE Conference on Intelligent Transportation Systems, vol. 1, St. Louis, 4–7 October 2009, pp. 432–437.
[15] L. Zhang, R. Chu, S. Xiang, S. Liao, S. Z. Li, Face detection based on multi-block LBP representation, in: ICB'2007, pp. 11–18.
[16] K. Yu, T. Zhang, Y. Gong, Nonlinear learning using local coordinate coding, in: Proceedings of the Neural Information Processing Systems, Vancouver, Canada, 7–10 December 2009, pp. 2223–2231.
[17] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, T.S. Huang, Large-scale image classification: fast feature extraction and SVM training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA, 20–25 June 2011, pp. 1689–1696.
[18] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the Proceedings of the International Conference on Computer Vision, Corfu, Greece, 20–25 September 1999, pp. 1150–1157.
[19] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[20] A. Bosch, A. Zisserman, X. Munoz, Image classification using random forests and ferns, in: Proceedings of the Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007, pp. 1–8.
[21] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (2001) 145–175.
[22] A. Bosch, A. Zisserman, X. Munoz, Scene classification via pLSA, in: Proceedings of the Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006, pp. 517–530.
[23] L.-J. Li, H. Su, E.P. Xing, L. Fei-Fei, Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: Proceedings of the Neural Information Processing Systems, Vancouver, Canada, 6–9 December 2010, pp. 1378–1386.
[24] L. Bo, X. Ren, D. Fox, Hierarchical matching pursuit for image classification: architecture and fast algorithms, in: Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–14 December 2011, pp. 2115–2123.
[25] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recognition 29 (1996) 51–59.
[26] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intel. 24 (2002) 971–987.
[27] C. Liu, Learning the uncorrelated, independent, and discriminating color spaces for face recognition, IEEE Trans. Inf. Forensics Secur. 3 (2008) 213–222.
[28] C. Liu, Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance, IEEE Trans. Pattern Anal. Mach. Intel. 28 (2006) 725–737.
[29] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, Washington, DC, USA, pp. 886–893.
[30] C. Burrus, R. Gopinath, H. Guo, Introduction to Wavelets and Wavelet Transforms: a Primer, Prentice-Hall, 1998.
[31] G. Beylkin, R. Coifman, V. Rokhlin, Fast wavelet transforms and numerical algorithms I, Commun. Pure Appl. Math. 44 (1991) 141–183.
[32] P. Porwik, A. Lisowska, The haar wavelet transform in digital image processing: its status and achievements, Mach. Graph. Vis. 13 (2004) 79–98.
[33] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intel. 11 (1989) 674–693.
[34] K. Fukunaga, Introduction to Statistical Pattern Recognition, second edition, Academic Press, 1990.
[35] Y. Ohta, Knowledge-Based Interpretation of Outdoor Natural Color Scenes, Pitman Publishing, London, 1985.
[36] A. Smith, Color gamut transform pairs, Comput. Graph. 12 (1978) 12–19.
[37] R. Gonzalez, R. Woods, Digital Image Processing, third edition, Pearson Prentice Hall, 2008.
[38] P. Shih, C. Liu, Comparative assessment of content-based face image retrieval in different color spaces, Int. J. Pattern Recognition Artif. Intell. 19 (2005).
[39] M. Bratkova, S. Boulos, P. Shirley, oRGB: a practical opponent color space for computer graphics, IEEE Comput. Graph. Appl. 29 (2009) 42–55.
[40] C. Liu, H. Wechsler, Robust coding schemes for indexing and retrieval from large face databases, IEEE Trans. Image Process. 9 (2000) 132–137.
[41] G. Griffin, A. Holub, P. Perona, Caltech-256 Object Category Dataset, Technical Report 7694, California Institute of Technology, 2007.
[42] L.-J. Li, L. Fei-Fei, What, where and who? classifying event by scene and object recognition, in: Proceedings of the Proceedings of IEEE International Conference in Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007, pp. 1–8.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014

[43] A. Vedaldi, B. Fulkerson, Vlfeat — an open and portable library of computer vision algorithms, in: Proceedings of the Proceedings of the 18th Annual ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010, pp. 1469–1472.
[44] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, Washington, DC, USA.

Sugata Banerji received his Bachelor of Engineering in Information Technology from the West Bengal University of Technology, Kolkata, India in 2005. After working for a few years as a programmer in a software company, he is currently pursuing his Ph.D. degree in the Department of Computer Science at the New Jersey Institute of Technology, Newark, USA. His research interests include color image processing, content based image classification, image search and retrieval, texture and shape-based feature extraction and pattern recognition.

Atreyee Sinha received her B. Tech. degree in Computer Science in 2010 from the West Bengal University of Technology, India. She is currently a Ph.D. candidate in Computer Science at the New Jersey Institute of Technology, Newark, USA. Her present work includes developing image descriptors and designing robust image classification models by applying computer vision concepts and statistical learning theory. Her research interests lie in the fields of Image Processing, Computer Vision, Pattern Recognition and Image Search with applications to object and scene image classification.

Chengjun Liu is the Director of the Face Recognition and Video Processing Lab at New Jersey Institute of Technology. His research is mainly in Machine Learning, Pattern Recognition, Computer Vision, Image and Video Analysis, and Security. He currently serves as an editorial board member for the International Journal of Biometrics.

Please cite this article as: S. Banerji, et al., New image descriptors based on color, texture, shape, and wavelets for object and scene image classification, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.02.014