2012 IEEE International Conference on Systems, Man, and Cybernetics
October 14-17, 2012, COEX, Seoul, Korea

Scene Image Classification: Some Novel Descriptors

Sugata Banerji[1], Atreyee Sinha and Chengjun Liu
New Jersey Institute of Technology,
Newark, NJ 07102, USA
Email:{sb256, as739, chengjun.liu}@njit.edu

***Abstract*—This paper introduces several novel color, shape and texture-based image descriptors for scene image classification with applications to image search and retrieval. Specifically, first, a new 3-Dimensional Local Binary Pattern (3DLBP) descriptor is proposed for color image local feature extraction. Second, a new shape descriptor (HaarHOG) is introduced by combining Haar wavelet transformation and Histogram of Oriented Gradients (HOG). Third, these descriptors are fused using an optimal feature representation technique to generate a robust 3-Dimensional LBP-HaarHOG (3DLH) descriptor that can perform well on different scene image categories. Finally, the Enhanced Fisher Model (EFM) is applied for discriminatory feature extraction and the nearest neighbor classification rule is used for image classification. The proposed descriptors and fusion technique are evaluated using three grand challenge datasets: the MIT Scene dataset, the UIUC Sports Event dataset, and a part of the Caltech 256 dataset.**

***Index Terms*—The HaarHOG descriptor, the 3-Dimensional Local Binary Pattern (3D-LBP) descriptor, the 3DLH descriptor, Enhanced Fisher Model (EFM), image search, scene classification**

I. Introduction

A color image contains much more information than a grayscale image and hence color images are more suited for image classification tasks. In recent years, use of color as a means to face recognition [1], [2] and object and scene retrieval [3], [4] has gained popularity. Color features can be derived from various color spaces and they exhibit different properties. Recent work on color based image search appears in [1], [5], [4] that propose several new color spaces and methods for face, object and scene category recognition. The HSV color space is used for scene category recognition in [6], and the evaluation of local color invariant descriptors is performed in [7]. The UCS and DCS colorspaces have been discussed in [8] and the rgb and $I_1I_2I_3$ color spaces have been shown to possess certain advantages over other color spaces in [2]. Fusion of color models, color region detection and color edge detection has been investigated for representation of color images [9]. Key contributions in color, texture, and shape abstraction have been discussed in Datta et al. [10].

Lately, several methods based on LBP features have been proposed for image representation and classification [11], [12], [13]. Extraction of LBP features is computationally efficient and with the use of multi-scale filters invariance to scaling and rotation can be achieved [12], [3]. LBP features are also invariant to illumination changes and fusion of different color features has been shown to achieve a good image retrieval success rate [3], [13], [14]. Local image descriptors have also been shown to perform well for texture based image retrieval [3], [14]. Several researchers have used the Haar wavelet transform for object detection in images and LBP has also been combined with Haar-like features for face detection [15]. The Histograms of Oriented Gradients (HOG) descriptor [16] is able to represent an image by storing information about its local shape. The generation of the HOG vector is explained in Figure 1.

Efficient retrieval requires a robust feature extraction method that has the ability to learn meaningful low-dimensional patterns in spaces of very high dimensionality [17]. Low-dimensional representation is also important when one considers the computational aspect. PCA has been widely used to perform dimensionality reduction for image indexing and retrieval [18]. The EFM feature extraction method has achieved good success for the task of image representation and retrieval [19].

In this paper, we employ three masks in three perpendicular planes to generate a novel 3D-LBP feature that contains more information than the traditional LBP. We also subject the image to a Haar wavelet transformation and then generate the HOG of the resultant image to create a robust HaarHOG feature vector. We fuse these two feature vectors in the PCA space to form the 3-Dimensional LBP-HaarHOG (3DLH) feature. We extend this concept to different color spaces and propose several new 3DLH feature representations, in particular the DCS-3DLH, oRGB-3DLH and YCbCr-3DLH feature representations, and then integrate them with other color 3DLH features to produce the novel Fused Color 3DLH (FC-3DLH) descriptor. Feature extraction applies the Enhanced Fisher Model (EFM) [18], [19] and image classification is based on the cosine similarity measure and the nearest neighbor classification rule. The effectiveness of the proposed descriptors and classification method is evaluated using three datasets: a part of the Caltech 256 grand challenge image dataset, the UIUC Sports Event dataset and the MIT Scene dataset.

II. Implementation details

We first review in this section eight color spaces in which our new descriptor is defined, and then discuss the 3D-LBP descriptor which is an improvement upon the traditional LBP

[1] Corresponding author

978-1-4673-1714-6/12/$31.00 ©2012 IEEE

descriptor. Next we present the HaarHOG descriptor and finally, the combination of these two descriptors, the 3DLH descriptor, in several individual color spaces and fused color FC-3DLH descriptor.

A. *Color spaces*

A color image contains three component images, and each pixel of a color image is specified using three values which show varying degrees of correlation. 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 calculated from the RGB color space by means of either linear or nonlinear transformations. To reduce the sensitivity of the RGB images to luminance, surface orientation, and other photographic conditions, the rgb color space is defined by normalizing the R, G, and B components. The HSV color space is motivated by human vision system because humans describe color by means of hue, saturation, and brightness. Hue and saturation define chrominance, while intensity or value specifies luminance [20]. The YCbCr color space is developed for digital video standard and television transmissions. In YCbCr, the RGB components are separated into luminance, chrominance blue, and chrominance red:

$$\begin{aligned}\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} &= \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \\ &+ \begin{bmatrix} 65.4810 & 128.5530 & 24.9660 \\ -37.7745 & -74.1592 & 111.9337 \\ 111.9581 & -93.7509 & -18.2072 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}\end{aligned} \tag{1}$$

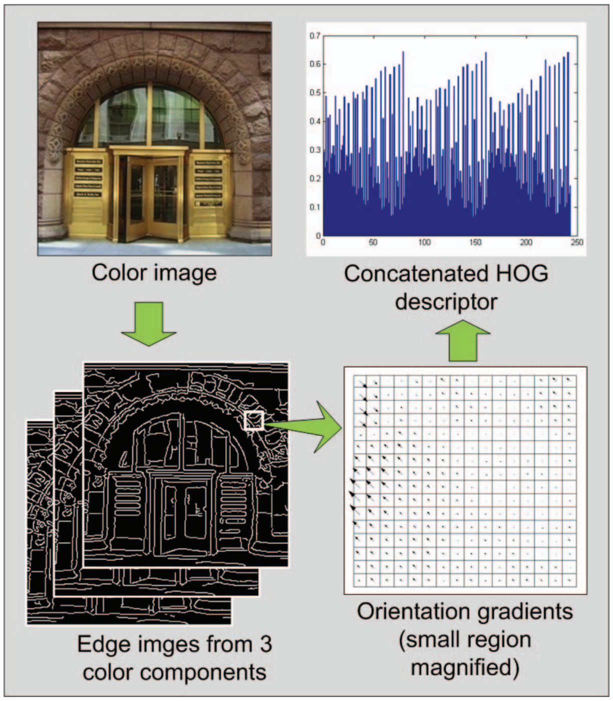

Color image
Concatenated HOG descriptor
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
Edge imges from 3 color components
Orientation gradients (small region magnified)

Fig. 1. The Histograms of Oriented Gradients (HOG) descriptor.

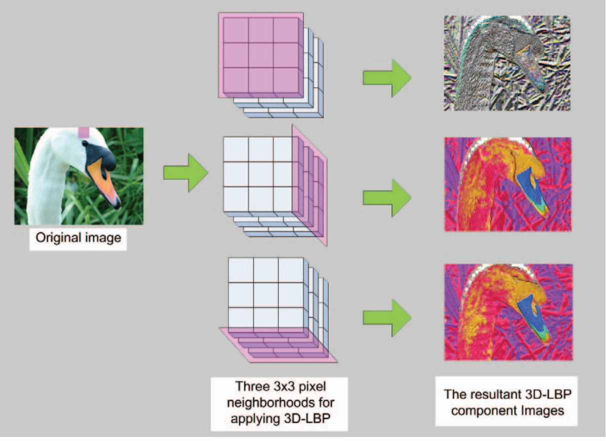

Original image
Three 3x3 pixel neighborhoods for applying 3D-LBP
The resultant 3D-LBP component Images

Fig. 2. The proposed 3D-LBP descriptor. A $3\times3\times3$ pixel region of the original image is magnified to show the 3D-LBP neighborhoods and the resulting LBP images.

where the R,G,B values are scaled to $[0,1]$.

The oRGB color space [21] has three channels L, C_1 and C_2. The primaries of this model are based on the three fundamental psychological opponent axes: white-black, red-green, and yellow-blue. The color information is contained in C_1 and C_2. The value of C_1 lies within $[-1,1]$ and the value of C_2 lies within $[-0.8660, 0.8660]$. The L channel contains the luminance information and its values ranges between $[0,1]$:

$$\begin{bmatrix} L \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5000 & 0.5000 & -1.0000 \\ 0.8660 & -0.8660 & 0.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2}$$

Another approach to stabilize RGB images is to decorrelate the RGB components [2]. The $I_1I_2I_3$ color space proposed by Ohta et al. [22] applies a Karhunen Loeve transformation to achieve this. The linear transformation is defined as follows:

$$\begin{aligned} I_1 &= (R+G+B)/3 \\ I_2 &= (R-B)/2 \\ I_3 &= (2G-R-B)/2 \end{aligned} \tag{3}$$

The UCS, is derived from the RGB color space using a decorrelation method, such as PCA [23]. In the RGB color space, a color image with a spatial resolution of $m \times n$ contains three color component images R, G, and B with the same resolution. Each pixel (x,y) of the color image thus contains

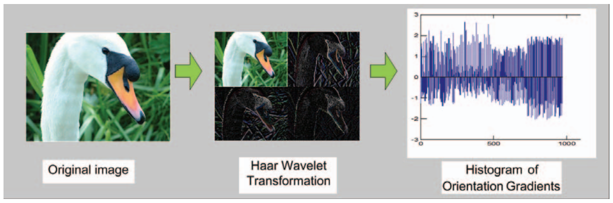

Original image
Haar Wavelet Transformation
Histogram of Orientation Gradients
3
2
1
0
-1
-2
-3
0
500
1000

Fig. 3. The proposed HaarHOG descriptor. The original image undergoes Haar Wavelet Transformation and then HOG is generated for each component of the resulting image and concatenated.

three elements corresponding to the red, green, and blue values from the R, G, and B component images which are correlated as well. The UCS decorrelates its component images via a linear transformation $W_U \in \mathbb{R}^{3\times 3}$ from the RGB color space

$$\begin{bmatrix} U_1(x,y) \\ U_2(x,y) \\ U_3(x,y) \end{bmatrix} = W_U \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \tag{4}$$

where $U_1(x,y)$, $U_2(x,y)$, and $U_3(x,y)$ are the values of the uncorrelated component images U_1, U_2, and U_3 in the UCS, $x = 1,2,\ldots,m$ and $y = 1,2,\ldots,n$. The transformation matrix may be derived using PCA. Let $\mathscr{X}$ be the 3-D vector in the RGB color space

$$\mathscr{X} = \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \tag{5}$$

The covariance matrix of $\mathscr{X}$ can be factorized in the following form:

$$\Sigma_{\mathscr{X}} = W_U^t \Lambda W_U \tag{6}$$

where $W_U^t \in \mathbb{R}^{3\times 3}$ is an orthonormal eigenvector matrix of the covariance matrix of the random vector $\mathscr{X}$, and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ a diagonal eigenvalue matrix with diagonal elements in decreasing order. The value of N here is 3.

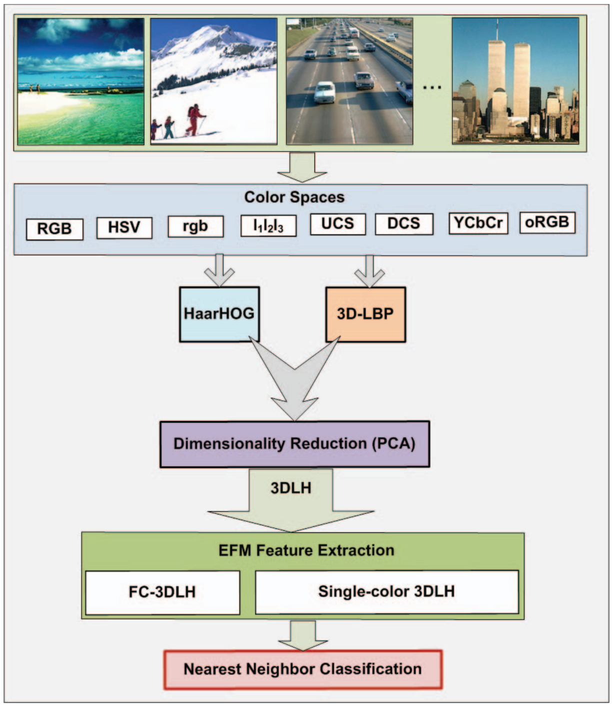
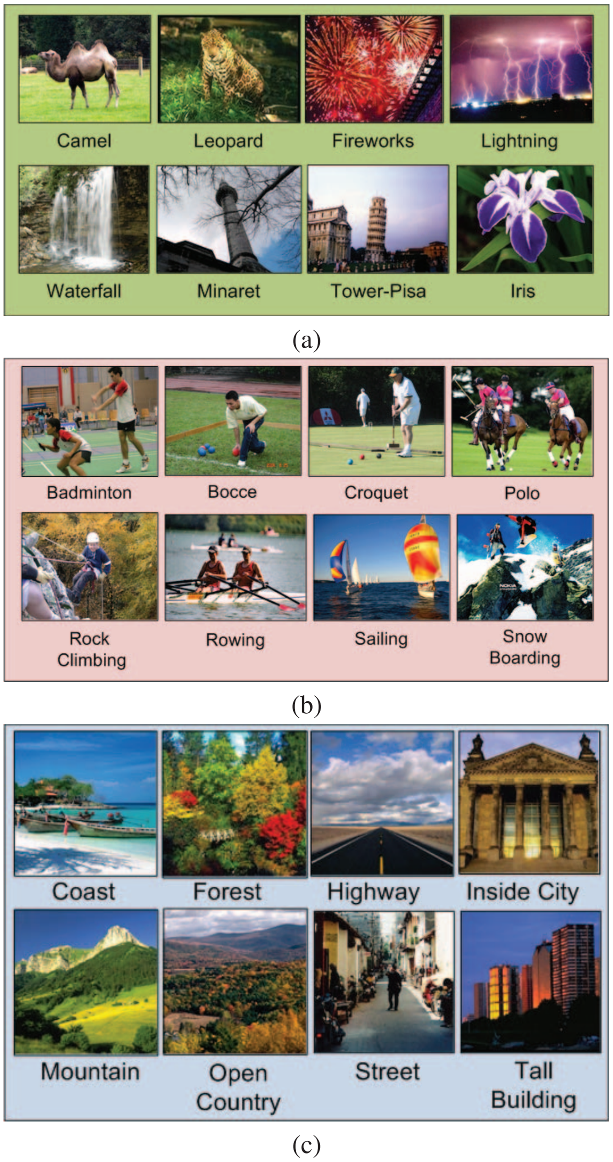


Fig. 4. An overview of multiple features fusion methodology, the EFM feature extraction method, and the classification stages.



(a)

(b)

(c)

Fig. 5. Example images from the (a) Caltech 25 scene dataset, (b) the UIUC Sports Event dataset and (c) the MIT Scene dataset.

The Discriminating Color Space (DCS), is derived from the RGB color space by means of discriminant analysis [23]. The DCS defines discriminating component images via a linear transformation $W_D \in \mathbb{R}^{3\times 3}$ from the RGB color space

$$\begin{bmatrix} D_1(x,y) \\ D_2(x,y) \\ D_3(x,y) \end{bmatrix} = W_D \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \tag{7}$$

where $D_1(x,y)$, $D_2(x,y)$, and $D_3(x,y)$ are the values of the discriminating component images D_1, D_2, and D_3 in the DCS, $x = 1,2,\ldots,m$ and $y = 1,2,\ldots,n$. The transformation matrix $W_D \in \mathbb{R}^{3\times 3}$ may be derived through a procedure of discriminant analysis [23]. Let S_w and S_b be the within-class and the between class scatter matrices of the 3-D pattern vector $\mathscr{X}$ respectively. $S_w, S_b \in \mathbb{R}^{3x3}$. The discriminant analysis procedure derives a projection matrix W_D by maximizing the criterion $J_1 = tr(S_w^{-1} S_b)$ [23]. This criterion is maximized when W_D^t consists of the eigenvectors of the matrix $S_w^{-1} S_b$ [23]

$$S_w^{-1} S_b W_D^t = W_D^t \Delta \tag{8}$$

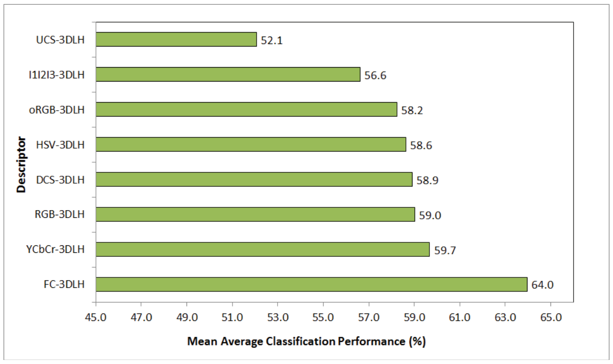


Fig. 6. The mean average classification performance of the proposed 3DLH descriptor in the RGB, HSV, YCbCr, oRGB, $I_1I_2I_3$, UCS, DCS and fused color spaces using the EFM-NN classifier on the Caltech 25 scene dataset.

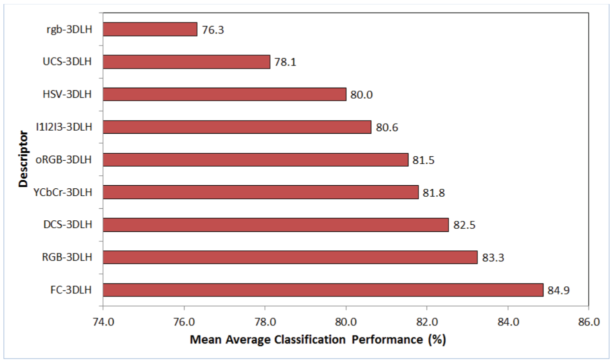


Fig. 7. The mean average classification performance of the proposed 3DLH descriptor in the RGB, rgb, HSV, YCbCr, oRGB, $I_1I_2I_3$, UCS, DCS and fused color spaces using the EFM-NN classifier on the UIUC sports event dataset.

where W_D^t, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1}S_b$, respectively.

B. The 3D-LBP, HaarHOG and 3DLH descriptors

The LBP descriptor [11] assigns an intensity value to each pixel of an image based on the intensity values of the eight neighboring pixels using a 3×3 mask. Since a color image is represented by a three dimensional matrix, we extended this concept to assign an intensity value to each pixel based on its neighboring pixels not only on the same color plane but on other planes as well. This method is explained in Figure 2.

For doing this operation, we replicate the first and third image planes on opposite sides of the three existing planes to create a five-plane matrix. After the 3D-LBP operation, only the three middle planes are retained. The 3D-LBP method produces three images. We concatenate the dense histograms of these three images to obtain the 3D-LBP feature vector.

To form the HaarHOG feature vector we apply the Haar wavelet transformation to each component of the original image to divide each component image into four distinct regions that separate the local features of each image. We then generate the HOG descriptor for each of these regions and then concatenate them for all the components to get our final HaarHOG feature vector. This process is illustrated in Figure 3. It should be noted that the generation time for both the 3DLBP and HaarHOG descriptors is linear with respect to the number of pixels.

Finally, we extract the most expressive features from both these vectors and fuse them in the PCA space to form the 3DLH feature vector which outperforms both 3D-LBP and HaarHOG based classification.

TABLE I
COMPARISON OF THE CLASSIFICATION PERFORMANCE (%) WITH OTHER METHODS ON THE UIUC SPORTS EVENT DATASET

#train	#test	3DLH		[24]		[25]	
560	480	DCS	82.5	SIFT+SC	82.7	OB	76.3
		RGB	83.3	HMP	**85.7**		
		Fused	84.9				

C. The EFM-NN Classifier

We perform learning and classification using Enhanced Fisher Linear Discriminant Model (EFM) [18], [19]. The EFM method first applies Principal Component Analysis (PCA) to reduce the dimensionality of the input pattern vector. A popular classification method that achieves high separability among the different pattern classes is the Fisher Linear Discriminant (FLD) method. The FLD method, if implemented in an inappropriate PCA space, may lead to overfitting. The EFM method, which applies an eigenvalue spectrum analysis criterion to choose the number of principal components to avoid overfitting, improves the generalization performance of the FLD. The EFM method thus derives an appropriate low dimensional representation from the 3DLH descriptor and further extracts the EFM features for pattern classification. We compute similarity score between a training feature vector and a test feature vector using the cosine similarity measure and the nearest neighbor classification rule. Figure 4 gives an overview of multiple feature fusion methodology, the EFM feature extraction method, and the classification stages. Note that PCA is used twice - once before fusion for dimensionality reduction and again as the first step of EFM.

III. EXPERIMENTAL RESULTS

A. Caltech 25 Scene Dataset

The Caltech 256 dataset [26] holds 30,607 images divided into 256 object categories and a clutter class. We have selected 25 scene image categories from this dataset to form the Caltech 25 scene dataset. This subset contains 110 Camel, 104 Canoe, 87 Duck, 83 Eiffel Tower, 101 Elk, 110 Fern, 100 Fireworks, 80 Golden Gate Bridge, 201 Grapes, 93 Hawksbill, 120 Ibis, 108 Iris, 111 Ketch, 91 Killer whale, 190 Leopards, 136 Lightning, 130 Minaret, 202 Mushroom, 103 Palm tree, 102 Rainbow, 95 Skyscraper, 105 Tennis court, 90 Tower Pisa, 95 Waterfall and 96 Zebra. The images have high intra-class variability and high object location variability. The images represent a diverse set of lighting conditions, backgrounds and sizes. As can be seen from Figure 5(a), some classes like Minaret and Tower-Pisa have a very similar visual appearance

TABLE II
CATEGORY WISE DESCRIPTOR PERFORMANCE (%) ON THE MIT SCENE DATASET. NOTE THAT THE CATEGORIES ARE SORTED ON THE FC-3DLH RESULTS

Category	Fusion	YCbCr	HSV	oRGB	DCS	UCS	$I_1I_2I_3$	rgb	RGB
forest	**97**	96	**97**	96	96	95	94	94	95
street	**94**	90	91	90	93	90	88	89	88
coast	**93**	91	90	91	92	90	87	89	92
mountain	**93**	89	86	89	88	88	87	88	88
tall building	**91**	90	88	89	87	84	85	85	87
inside city	89	**91**	89	89	90	87	88	87	90
highway	88	84	**90**	88	84	84	88	84	82
open country	**78**	76	74	74	72	71	69	68	72
	90.4	**88.4**	**88.3**	**88.2**	**87.9**	**86.1**	**85.8**	**85.5**	**85.3**

and classification can be very challenging. Most images are in color JPEG format with only a small percentage in grayscale.

On this dataset, we conduct experiments for 3DLH descriptors from seven different color spaces and their fusion. For each class, we make use of 50 images for training and 25 images for testing. Figure 6 shows the detailed performance of our EFM-NN classification technique on this dataset. This is the average result obtained by training and testing over five random data splits. The best recognition rate that we obtain is 64.0%, which is a very respectable value for a dataset of this size and complexity. Since intra-class variability is very high for the Caltech 256 dataset, and in several cases the object occupies a small portion of the full image, SIFT-based methods usually achieve better classification success for this dataset than dense histogram-based methods. The proposed method is faster than the SIFT-based methods and the YCbCr-3DLH alone achieves a success rate of 59.7% on the 25 classes selected. Fusion of color spaces improves our result further by over 4%. Due to the nature of the 3D-LBP descriptor, it is not defined for grayscale images and so we did not conduct any experiments for grayscale. Also, conversion to the rgb color space is undefined for grayscale images and we did not use the rgb color space on this dataset as it contains some grayscale images.

B. UIUC Sports Events Dataset

The UIUC sports events dataset [28] contains 8 sports event categories: 250 rowing, 200 badminton, 182 polo, 137 bocce, 190 snowboarding, 236 croquet, 190 sailing, and 194 rock climbing. Some images from this dataset can be seen in Figure 5(b). From each class, we use 70 images for training and 60 images for testing the performance, and we do this for five random splits. Here RGB-3DLH is the best single-color descriptor at 83.3% followed by DCS-3DLH, YCbCr-3DLH and oRGB-3DLH respectively. The combined descriptor FC-3DLH gives a mean average performance of 84.9%. See Figure 7 for details. Table I compares our result with that of other methods. It should be noted that this method is much faster than the SIFT-based methods used by other researchers. We tested our descriptor in the rgb color space as well for this dataset and the fused descriptor contains RGB, HSV, rgb, YCbCr, oRGB, UCS, DCS and $I_1I_2I_3$-3DLH descriptors.

C. MIT Scene Dataset

The MIT Scene dataset [27] has 2,688 images classified as eight categories: 360 coast, 328 forest, 374 mountain, 410 open country, 260 highway, 308 inside of cities, 356 tall buildings, and 292 streets. All of the images are in color JPEG format. There is a large variation in light and angles, along with a high intra-class variation. A few sample images from this dataset can be seen in Figure 5(c).

From each class, we use 250 images for training and the rest of the images for testing the performance, and we do five-fold cross-validation. Here YCbCr-3DLH is the best single-color descriptor at 88.4% followed closely by HSV-3DLH and oRGB-3DLH. The combined descriptor FC-3DLH gives a mean average performance of 90.4%. See Figure 8 for details. Table III compares our result with that of other methods. For this dataset also we tested our descriptors in all eight colorspaces. Table II shows the category wise descriptor

TABLE III
COMPARISON OF THE CLASSIFICATION PERFORMANCE (%) WITH OTHER METHODS ON THE MIT SCENE DATASET

#train	#test	3DLH		[3]		[27]
2000	688	YCbCr	**88.4**	CLF	**86.4**	-
		HSV	**88.3**	CGLF	**86.6**	
		Fused	**90.4**	CGLF+PHOG	**89.5**	
800	1888	YCbCr	**85.8**	CLF	79.3	
		HSV	**85.7**	CGLF	80.0	
		Fused	**87.5**	CGLF+PHOG	84.3	83.7

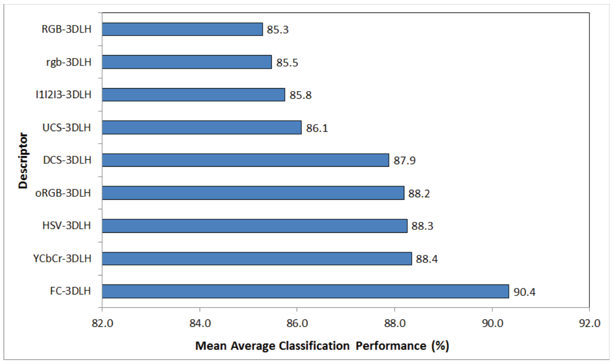


Fig. 8. The mean average classification performance of the proposed 3DLH descriptor in the RGB, rgb, HSV, YCbCr, oRGB, $I_1I_2I_3$, UCS, DCS and fused color spaces using the EFM-NN classifier on the MIT scene dataset.

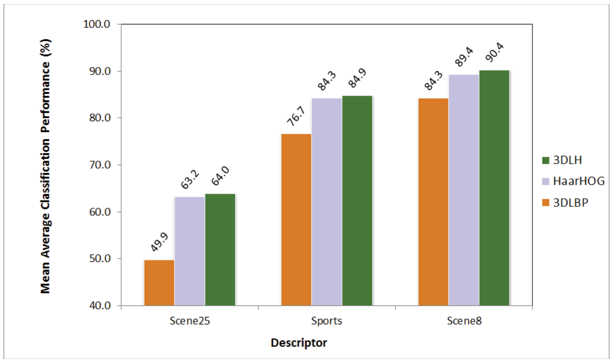


Fig. 9. The comparative mean average classification performance of the FC-3DLBP, FC-HaarHOG and FC-3DLH descriptors on the Caltech scene 25, UIUC Sports Event and MIT Scene datasets.

performance on the MIT Scene dataset.

Figure 9 gives a comparison of the two descriptors and their fusion for image classification in the three datasets used for our experiments.

IV. Conclusion

We have proposed a new LBP-based color and texture feature extraction method for images and combined it with Haar wavelet features and HOG features to generate several new descriptors for color scene images: the RGB-3DLH descriptor, the oRGB-3DLH descriptor, the YCbCr-3DLH descriptor, the DCS-3DLH descriptor, the HSV-3DLH descriptor and the FC-3DLH descriptor for scene image classification. Results of the experiments using three challenging datasets show that our oRGB-3DLH, HSV-3DLH, DCS-3DLH and YCbCr-3DLH descriptors improve recognition performance over conventional color and grayscale LBP descriptors. Different colorspaces perform differently for different image sets, and further, the fusion of multiple color 3DLH descriptors (FC-3DLH) shows improvement in the classification performance, which indicates that various color 3DLH descriptors are not fully redundant for image classification tasks.

References

[1] C. Liu and J. Yang, “ICA color space for pattern recognition,” *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 248–257, 2009.

[2] P. Shih and C. Liu, “Comparative assessment of content-based face image retrieval in different color spaces,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 7, 2005.

[3] S. Banerji, A. Verma, and C. Liu, “Novel color LBP descriptors for scene and image texture classification,” in *15th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, Las Vegas, Nevada, July 18-21 2011.

[4] A. Verma, S. Banerji, and C. Liu, “A new color SIFT descriptor and methods for image category classification,” in *International Congress on Computer Applications and Computational Science*, Singapore, December 4-6 2010, pp. 819–822.

[5] J. Yang and C. Liu, “Color image discriminant models and algorithms for face recognition,” *IEEE Transactions on Neural Networks*, vol. 19, no. 12, pp. 2088–2098, 2008.

[6] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification using a hybrid generative/discriminative approach,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[7] G. Burghouts and J.-M. Geusebroek, “Performance evaluation of local color invariants,” *Computer Vision and Image Understanding*, vol. 113, pp. 48–62, 2009.

[8] C. Liu, “Learning the uncorrelated, independent, and discriminating color spaces for face recognition,” *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 213–222, 2008.

[9] H. Stokman and T. Gevers, “Selection and fusion of color models for image feature detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 371–381, 2007.

[10] R. Datta, D. Joshi, J. Li, and J. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” *ACM Computing Surveys*, vol. 40, no. 2, pp. 509–522, 2008.

[11] T. Ojala, M. Pietikainen, and D. Harwood, “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions,” in *International Conference on Pattern Recognition*, Jerusalem, Israel, 1994, pp. 582–585.

[12] C. Zhu, C. Bichot, and L. Chen, “Multi-scale color local binary patterns for visual object classes recognition,” in *International Conference on Pattern Recognition*, Istanbul, Turkey, August 23-26 2010, pp. 3065–3068.

[13] M. Crosier and L. Griffin, “Texture classification with a dictionary of basic image features,” in *Proceedings of Computer Vision and Pattern Recognition*, Anchorage, Alaska, June 23-28, 2008, pp. 1–7.

[14] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[15] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, “Face detection based on multi-block LBP representation,” in *ICB’2007*, 2007, pp. 11–18.

[16] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 1*, Washington, DC, USA, 2005, pp. 886–893.

[17] C. Liu and H. Wechsler, “Independent component analysis of Gabor features for face recognition,” *IEEE Transactions on Neural Networks*, vol. 14, no. 4, pp. 919–928, 2003.

[18] ——, “Robust coding schemes for indexing and retrieval from large face databases,” *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 132–137, 2000.

[19] ——, “Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition,” *IEEE Transactions on Image Processing*, vol. 11, no. 4, pp. 467–476, 2002.

[20] R. Gonzalez and R. Woods, *Digital Image Processing*. Prentice Hall, 2001.

[21] M. Bratkova, S. Boulos, and P. Shirley, “oRGB: A practical opponent color space for computer graphics,” *IEEE Computer Graphics and Applications*, vol. 29, no. 1, pp. 42–55, 2009.

[22] Y. Ohta, *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*. Pitman Publishing, London, 1985.

[23] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

[24] L. Bo, X. Ren, and D. Fox, “Hierarchical Matching Pursuit for Image Classification: Architecture and Fast Algorithms,” in *Advances in Neural Information Processing Systems*, December 2011.

[25] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei, “Object bank: A high-level image representation for scene classification & semantic feature sparsification,” in *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2010.

[26] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” California Institute of Technology, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694

[27] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[28] L. Fei-Fei and L.-J. Li, “What, Where and Who? Telling the Story of an Image by Activity Classification, Scene Recognition and Object Categorization,” *Studies in Computational Intelligence- Computer Vision*, pp. 157–171, 2010.