HaarHOG: Improving the HOG Descriptor for Image Classification

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

***Abstract*—The Histograms of Oriented Gradients (HOG) descriptor represents shape information by storing the local gradients in an image. The Haar wavelet transform is a simple yet powerful technique that can separately enhance the horizontal and vertical local features in an image. In this paper, we enhance the HOG descriptor by subjecting the image to the Haar wavelet transform and then computing HOG from the result in a manner that enriches the shape information encoded in the descriptor. First, we define the novel HaarHOG descriptor for grayscale images and extend this idea for color images. Second, we compare the image recognition performance of the HaarHOG descriptor with the traditional HOG descriptor in four different color spaces and grayscale. Finally, we compare the image classification performance of the HaarHOG descriptor with some popular descriptors used by other researchers on four grand challenge datasets.**

***Index Terms*—HaarHOG descriptor, Haar wavelets, Histograms of Oriented Gradients descriptor, shape descriptor, object and scene image classification**

I. Introduction

The field of content-based image classification, search and retrieval has expanded greatly in recent years with millions of color images being stored and shared over the Internet each day. Creation of the feature descriptor is one of the first steps in the image search and classification process and this paper introduces a novel descriptor for grayscale as well as color images.

Shape and high-frequency local information contribute heavily to object and scene image recognition, and hence, descriptors based on such features are frequently used for image classification. The Histograms of Oriented Gradients (HOG) descriptor [1], which represents an image by histograms of the slopes of the object edges in an image, stores information about the shapes contained in the image. As a result, HOG has become a popular descriptor for object tracking in images and videos, and content based image retrieval. Wavelets are known to selectively enhance high frequency local information in selected orientations. That is why wavelets, such as the Haar wavelets have been widely applied for object detection in images [2], [3].

The human visual system often uses color information for object and scene image classification. In fact, color images contain much more discriminative information than grayscale images and have been shown to perform better than grayscale images for image classification tasks [4], [5], [6], [7], [8]. The 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.

This paper introduces a novel image descriptor based on shape and local high-frequency features from an image, and then extends it to include the benefits of using multiple color spaces. Specifically, first, a new HaarHOG feature vector is defined that extracts shape as well as other local features from a grayscale image by combining the Haar wavelet transform with the Histograms of Oriented Gradients (HOG). This is intuitively based on the idea that scene recognition is often based on the presence of certain objects in a scene and hence the Haar wavelets and HOG would both help scene recognition. Next, we extend the definition of the new descriptor for use in color images.

To assess the classification performance of the proposed descriptor, a Support Vector Machine (SVM) classifier with a linear kernel is used on several widely used and publicly available image datasets. In these experiments, it is shown to achieve a significantly better classification performance than the conventional HOG descriptor, as well as some other popular image descriptors, such as Scale Invariant Feature Transform (SIFT) based methods, Spatial Envelope (SE), Object Bank (OB), as well as Local Binary Patterns (LBP).

This paper is organized in the following manner. Section II discusses the background work by other researchers that have been used in this paper. Section III explains the new HaarHOG descriptors introduced here. Section IV evaluates the performance of the HaarHOG descriptor on four different image datasets and compares the performance with the HOG descriptor and some other popular descriptors. Finally, Section V summarizes the contributions and findings of this paper.

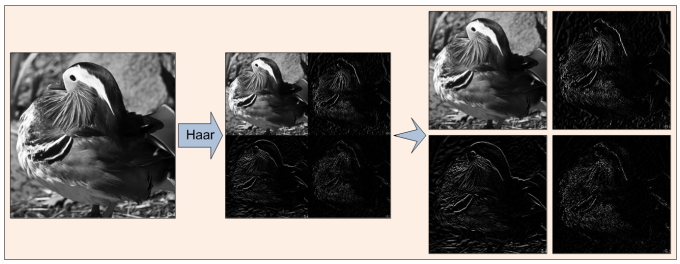

Fig. 1. A grayscale image and its Haar wavelet transform.

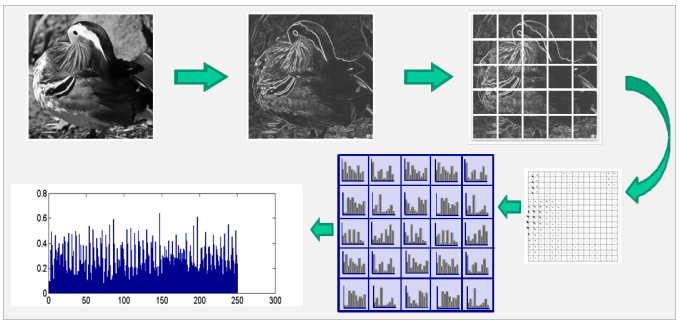

Fig. 2. A grayscale image and the formation of the HOG descriptor.

II. Background

This section first discusses the theoretical background work related to the concepts used in this work, and also discusses the different color spaces in which our new descriptor is tested.

A. Haar Wavelet Transform

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 [9], [10]. The Haar scaling function $\phi(x)$ is defined below [9], [11]:

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

A family of functions can be generated from the basic scaling function by scaling and translation [9], [11]:

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

As a result, the scaling functions $\phi_{i,j}(x)$ can span the vector spaces V^i, which are nested as follows: $V^0 \subset V^1 \subset V^2 \subset \cdots$ [12].

The Haar wavelet function $\psi(x)$ is defined as follows [9], [11]:

$$\psi(x) = \begin{cases} 1, & 0 \le x < 1/2 \\ -1, & 1/2 \le x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The Haar wavelets are generated from the mother wavelet by scaling and translation [9], [11]:

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

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}: $V^{i+1} = V^i \oplus W^i$ [9], [11]. The 2D Haar basis functions are the tensor product of the one dimensional scaling and wavelet functions [10].

Figure 1 shows a grayscale image of a Mandarin duck and its Haar wavelet transformed image. The right side of the figure displays an enlargement of the four quadrants of the Haar wavelet transformed image which shows that different sub-images enhance high-frequency local features in different orientations.

B. Histograms of Oriented Gradients (HOG)

The idea of Histograms of Oriented Gradients (HOG) rests on the observation that local features such as object appearance and shape can often be characterized well by the distribution of local intensity gradients in the image [1]. HOG features are derived from an image based on a series of normalized local histograms of image gradient orientations in a dense grid [1], [13].

Figure 2 demonstrates the formation of the HOG vector for a grayscale image. The image of a duck at the top right is the original grayscale image. The first step is the calculation of the gradient magnitudes at every pixel. The gradient magnitude image is shown in the middle figure of the top row. Next, the image window is divided into a number of blocks as shown in the last image in the first row of Figure 2. In the original implementation by [1], dividing the image into 3×3 blocks was found to be optimal for pedestrian detection. For our experiments, however, we found the classification performance increasing for 5×5 blocks and so we used 5×5 blocks for our implementation. Next, the orientation of each pixel in each block is put in one of 10 orientation bins weighted by its magnitude and thus a weighted histogram is formed for each block of cells. There is an overlap of half the block size between consecutive blocks to increase accuracy. Finally, the histograms from the individual cells are normalized and concatenated to form the HOG vector. This whole operation of forming histograms and concatenating them is shown in the bottom row of Figure 2.

C. Color Spaces

We now briefly review the four color spaces which we have used in this work for assessing the performance of the proposed HaarHOG descriptor. The RGB color space is the common tristimulus space used for representing color images on a computer. It represents an image as three component images that represent the intensities of the red, green, and blue primary colors. Other color spaces can be derived from the RGB color space by linear or nonlinear transformations.

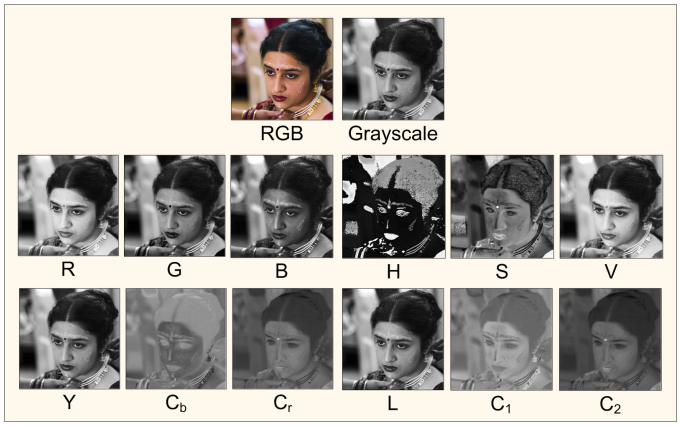

Fig. 3. An RGB color image, its grayscale image, and the color component images in the RGB, HSV, YCbCr and oRGB color spaces, respectively.

In this paper, we are using three very popular color spaces in addition to the RGB color space: the HSV color space [14], the YCbCr color space [15] and the recently defined oRGB color space [16].

The HSV (hue, saturation, and value) color space is based on the way humans perceive color. Hue and saturation define chrominance, while intensity or value specifies luminance [15]. The HSV color space is derived nonlinearly from the RGB color space [14]. The other two color spaces are derived from the RGB color space using linear transformations. The YCbCr color space was originally developed for digital video standard and television transmissions. In the YCbCr color space, the image is split into luminance (Y), chrominance-blue (Cb) and chrominance-red (Cr) components [15]. The recently introduced oRGB color space [16] has three channels L, C1 and C2. The primaries of this color model are based on the three fundamental psychological opponent axes: white-black, red-green, and yellow-blue. In the oRGB color space, the color information is contained in the C1 and C2 channels. The value of C1 lies within [-1, 1] and the value of C2 lies within [-0.8660, 0.8660]. The L channel contains the luminance information and its value ranges within [0, 1].

Figure 3 shows a color image, its grayscale image, and its color component images in the RGB, HSV, YCbCr and oRGB color spaces, respectively.

D. The Classifier: Support Vector Machine

We use a Support Vector Machine (SVM) classifier with a linear kernel for the classification task. SVM is a particular realization of statistical learning theory. The approach described by SVM, known as structural risk minimization, minimizes the risk functional in terms of both the empirical risk and the confidence interval [17]. SVM is built from two ideas: (i) a nonlinear mapping of the input space to a high-dimensional feature space, and (ii) designing the optimal hyperplane in terms of the maximal margin between the patterns of the two classes in the feature space. SVM is very popular and has been applied extensively for pattern classification, regression, and density estimation since it displays a good generalization performance.

For our experiments, we trained an SVM with a linear kernel independently for each class (one-vs-all classification). A similar configuration has been successfully used by other researchers like [18] in recent works. The SVM implementation used is the one that is distributed with the VlFeat package [19].

III. The Novel HaarHOG Descriptor for Image Classification

This section first introduces the new HaarHOG descriptor for grayscale images as an improvement over HOG and explains the proposed technique in detail. Then it extends this concept to color images to define the new color HaarHOG descriptor.

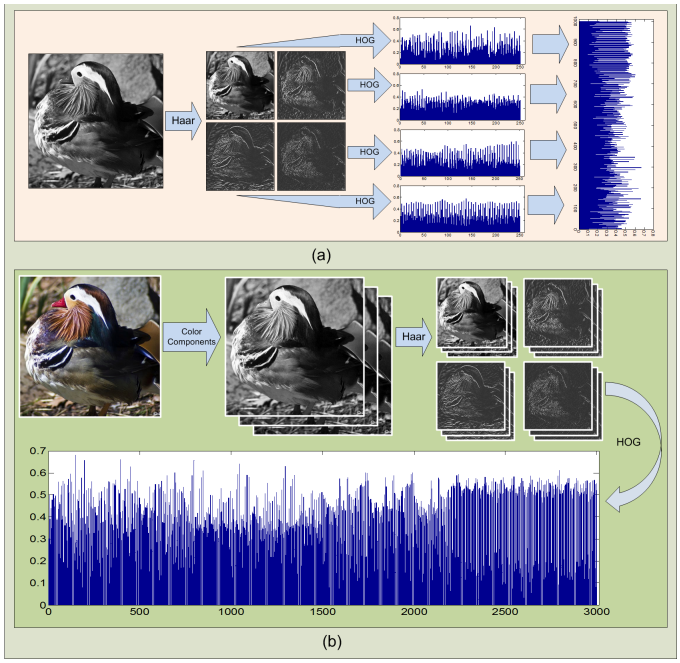

Fig. 4. The formation of (a) the grayscale HaarHOG feature vector from a grayscale image and (b) the color HaarHOG feature vector from a color image. In both cases, the four quadrants of the Haar wavelet transformed image is shown separated for clarity. The HOG operation shown at the extreme right of (b) represents one HOG operation on each of the 12 images generated in the previous step.

A. The Novel Grayscale HaarHOG Descriptor

The motivation for the proposed new descriptor, the HaarHOG descriptor, is based on enhancing useful and important local high-frequency features before extracting shape for object and scene image classification. Towards that end, the Haar wavelet transform of a grayscale image is first computed. This process divides the grayscale image into four grayscale sub-images. One of these sub-images contains the low frequency information from the original image and the other three contain the high frequency information in different orientations. Each of these sub-images are one-fourth the size of the original image.

To generate the new HaarHOG descriptor, the HOG is next calculated the four quadrants of a Haar wavelet transformed image and then concatenated to get the HaarHOG descriptor. The size of the grayscale HaarHOG feature vector thus obtained is four times the size of one HOG vector. For the parameters used in our implementation, the size of the grayscale HaarHOG feature vector is $4\times5\times5\times10$ i.e. 1000. This method is explained in Figure 4(a).

B. The Innovative Color HaarHOG Descriptor

The process described above is applicable only to grayscale images. Since color images contain more discriminatory information than grayscale images, we can incorporate this information into our descriptor by calculating a HaarHOG vector from each color component image, and then concatenating the three vectors. Figure 4(b) shows this method. Specifically,

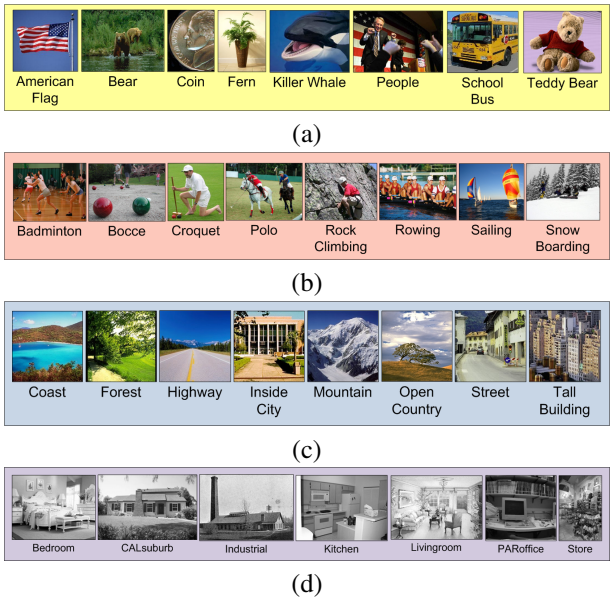

Fig. 5. Some sample images from (a) the Caltech 256 dataset, (b) the UIUC Sports Event dataset, (c) the MIT Scene dataset, and (d) the Fifteen Scene Categories dataset.

the color image on the upper left of the figure undergoes the Haar wavelet transformation on each of its color component images to generate a color Haar transformed image with four sub-images each. The four quadrants of these Haar wavelet transformed images are shown at the upper right of the figure. These twelve images, i.e. the three component images from the four quadrants of the color Haar transformed image, undergo the HOG operation, and their vectors are concatenated to form the innovative color HaarHOG descriptor. The color image may be converted to the HSV, the YCbCr or the oRGB color space from the RGB color space to obtain the color HaarHOG descriptor in the desired color space as the end result. The length of the color HaarHOG feature vector is 3000.

IV. Experiments

In this section we first introduce the datasets used for testing our new image descriptors and then do a comparative assessment of the classification performance of the HOG and the HaarHOG descriptors. Finally we compare the classification performance of the HaarHOG descriptor with some popular image descriptors used by other researchers.

A. Datasets Used

This section briefly introduces the four publicly available and widely used image datasets used for assessing the classification performance of our descriptor.

1) The Caltech 256 Dataset: The Caltech 256 dataset [20] holds 30,607 images divided into 256 object categories and a clutter class. Each category contains a minimum of 80 and a maximum of 827 images. The images, which are mostly color, represent a diverse set of lighting conditions, poses, backgrounds, and sizes [20]. The average size of each image is 351 $\times$ 351 pixels. Some sample images from this dataset are shown in Figure 5(a).

2) The UIUC Sports Event Dataset: The UIUC Sports Event dataset [21] 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 $\times$ 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 Figure 5(b). This dataset contains indoor and outdoor scenes and some classes like badminton and bocce contain both.

3) The MIT Scene Dataset: The MIT Scene dataset [22] has 2,688 images classified as eight scene 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 and in JPEG format, and the size of each image is 256 $\times$ 256 pixels. There is a large variation in light, content and angles, along with a high intra-class variation [22]. Figure 5(c) shows some images from this dataset.

4) The Fifteen Scene Categories Dataset: The Fifteen Scene Categories dataset [23] is composed of 15 scene categories: thirteen were provided by [24], eight of which were originally collected by [22] as the MIT Scene dataset, and two were collected by [23]. Each category has 200 to 400 images, most of which are grayscale. Figure 5(d) shows one image each from the newer seven classes of this dataset.

B. Comparative Assessment of the HOG and HaarHOG Descriptors on the Different Datasets

In this section, we make a comparative assessment of the HOG and our proposed HaarHOG descriptors in four different color spaces – RGB, HSV, oRGB, and YCbCr color spaces, as well as in grayscale, using the four datasets described earlier to evaluate classification performance. Note that we do not propose the new descriptor as a stand-alone state-of-the-art solution for different image classification problems, but as an improvement over HOG.

From the Caltech 256 dataset, we use 50 images per class for training and 25 images per class for testing. The experiment is done for five random splits of the data with no overlap between training and testing sets of the same split. As can be seen in Figure 6(a), the HaarHOG significantly outperforms the HOG in all four color spaces as well as in grayscale. The horizontal axis shows the proposed descriptors in four different color spaces and in grayscale, and the vertical axis denotes the mean average classification performance, which is the percentage of correctly classified images averaged across all the 256 classes and five runs of experiments.

For the UIUC Sports Event dataset, we use 70 images from each class for training and 60 for testing. Figure 6(b) shows the mean average classification performance obtained over five random splits of the data. Here also, the HaarHOG outperforms the HOG by a big margin that varies from about 3% to over 7%. Indeed, on this dataset the HaarHOG not only outperforms

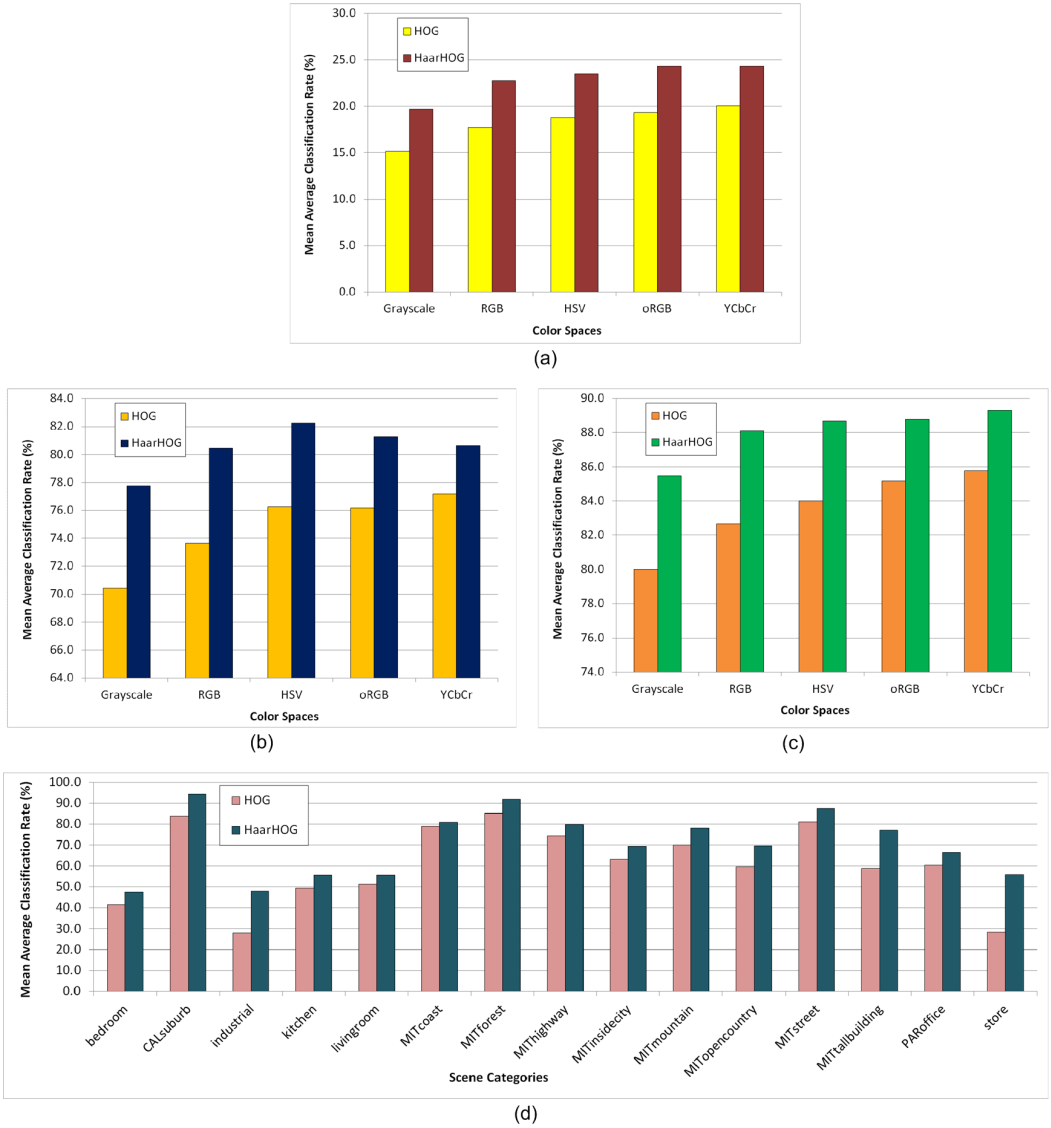

Fig. 6. (a), (b), (c) show the mean average classification performance of the HOG and proposed HaarHOG descriptors in the grayscale, RGB, HSV, oRGB, and YCbCr color spaces using the SVM classifier with linear kernel on the Caltech 256 dataset, the UIUC Sports Event dataset, and the MIT Scene dataset respectively. (d) shows the comparative mean average classification performance of the grayscale HOG and grayscale HaarHOG descriptors on the 15 categories of the Fifteen Scene Categories dataset.

the HOG, but also provides a decent classification performance by itself.

From both the MIT Scene dataset and the Fifteen Scene Categories dataset we use five random splits of 100 images per class for training, and the rest of the images for testing. Again, the HaarHOG produces decent classification performance on its own apart from beating the HOG by a fair margin. Figure 6(c) displays these results on the MIT Scene dataset. Here also, the horizontal axis shows the different descriptors in the four color spaces and in grayscale, and the vertical axis shows the mean average classification performance. The highest classification rate for this dataset is as high as 89.3% for the HaarHOG descriptor in the YCbCr color space which is a very respectable value for a dataset of this size and complexity. On the Fifteen Scene Categories dataset we conduct experiments only in grayscale. The overall success rate for HOG on this dataset is 60.9% and for HaarHOG it is 70.5%. In Figure 6(d), we display the category wise classification rates of the grayscale HOG and HaarHOG descriptors for all 15 categories of this dataset. Here, the horizontal axis reveals the fifteen scene categories, and the vertical axis displays the mean average classification performance. The HaarHOG here is shown to better the HOG classification performance in each scene category.

While the HaarHOG descriptor is proposed as an improvement over HOG, it shows some good classification performance even when used alone on some of these four datasets. We compare the classification performance of the proposed

TABLE I
COMPARISON OF THE CLASSIFICATION PERFORMANCE (%) OF THE PROPOSED HAARHOG DESCRIPTOR WITH OTHER POPULAR METHODS ON THE UIUC SPORTS EVENT AND MIT SCENE DATASETS

Descriptor		UIUC Performance (%)	MIT Performance (%)
SIFT+GGM	[21]	73.4	-
OB	[25]	76.3	-
CA-TM	[26]	78.0	-
LBP		-	77.9
CGLF	[5]	-	80.0
SE	[22]	-	83.7
CGLF+PHOG	[5]	-	84.3
C4CC	[27]	-	86.7
HOG		76.3	85.8
HaarHOG	**(proposed)**	**82.2**	**89.3**

HaarHOG descriptor with some popular image classification techniques used by other researchers. The detailed comparison is shown in Table I. The first column contains the different descriptors used for classification, the second column contains the classification performance in the UIUC sports event dataset and the third column shows the classification performance in the MIT Scene dataset. As can be seen from this table, the proposed HaarHOG descriptor (shown in bold on the bottom row) yields the best classification performance on both these datasets. It should be noted that the results of other researchers are reported directly from their published work.

V. CONCLUSION

We have presented in this paper a new image descriptor based on shape and local features for object and scene image classification that improves upon the popular HOG descriptor. We have first presented a new HaarHOG descriptor for a grayscale image. We then extended this definition for color images. We have also comparatively assessed the HaarHOG descriptor in four different color spaces — the RGB, the HSV, the YCbCr, and the oRGB — for image classification performance. Experimental results using four datasets show that the proposed new HaarHOG descriptor not only achieves significantly better image classification performance than the conventional HOG descriptor, but can also beat other popular descriptors, such as the Scale Invariant Feature Transform (SIFT), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank (OB), Context Aware Topic Model (CA-TM), as well as LBP on some popular scene image datasets.

REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *The 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 886–893.

[2] S. Vural, Y. Mae, H. Uvet, and T. Arai, "Multi-view fast object detection by using extended haar filters in uncontrolled environments," *Pattern Recognition Letters*, pp. 126–133, 2012.

[3] L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li, "Face detection based on multi-block LBP representation," in *The 2007 International Conference on Advances in Biometrics*, Seoul, Korea, 2007, pp. 11–18.

[4] C. Liu, "Effective use of color information for large scale face verification." *Neurocomputing*, pp. 43–51, 2013.

[5] 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, USA, July 18-21 2011, pp. 537–543.

[6] C. Liu, "Extracting discriminative color features for face recognition," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1796–1804, 2011.

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

[8] C. Liu, "The Bayes decision rule induced similarity measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 29, pp. 1116–1117, 2007.

[9] C. Burrus, R. Gopinath, and H. Guo, *Introduction to wavelets and wavelet transforms: A Primer*. Prentice-Hall, 1998.

[10] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical algorithms I," *Communications on Pure and Applied Mathematics*, vol. 44, no. 2, pp. 141–183, 1991.

[11] P. Porwik and A. Lisowska, "The haar wavelet transform in digital image processing: Its status and achievements," *Machine graphics & vision*, vol. 13, no. 1, pp. 79–98, 2004.

[12] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[13] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *12th International IEEE Conference On Intelligent Transportation Systems*, vol. 1, St. Louis, USA, 2009, pp. 432–437.

[14] A. Smith, "Color gamut transform pairs," *Computer Graphics*, vol. 12, no. 3, pp. 12–19, 1978.

[15] R. Gonzalez and R. Woods, *Digital Image Processing*, 3rd ed. Pearson Prentice Hall, 2008.

[16] 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.

[17] Y. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[18] J. Sanchez, F. Perronnin, and T. Campos, "Modeling the spatial layout of images beyond spatial pyramids," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2216 – 2223, 2012.

[19] A. Vedaldi and B. Fulkerson, "Vlfeat — an open and portable library of computer vision algorithms," in *The 18th Annual ACM International Conference on Multimedia*, Firenze, Italy, 2010, pp. 1469–1472.

[20] 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

[21] L.-J. Li and L. Fei-Fei, "What, where and who? classifying event by scene and object recognition." in *IEEE International Conference in Computer Vision*, 2007, pp. 1–8.

[22] 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.

[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006, pp. 2169–2178.

[24] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 524–531.

[25] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Neural Information Processing Systems*, Vancouver, Canada, 2010, pp. 1378–1386.

[26] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 16-21 2012, pp. 2743–2750.

[27] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *The European Conference on Computer Vision*, Graz, Austria, 2006, pp. 517–530.