2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Scene Image Classification using a Wigner-Based Local Binary Patterns Descriptor

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

***Abstract*—This paper introduces a new local feature description method to categorize scene images. We encode local image information by exploring the pseudo-Wigner distribution of images and the Local Binary Patterns (LBP) technique and make four major contributions. In particular, we first define a multi-neighborhood LBP for small image blocks. Second, we combine the multi-neighborhood LBP with the pseudo-Wigner distribution of images for feature extraction. Third, we derive the innovative WLBP feature vector by utilizing the frequency domain smoothing, the bag-of-words model and spatial pyramid representations of an image. Finally, we perform extensive experiments to evaluate the performance of the proposed WLBP descriptor. Specifically, we test our descriptor for classification performance using a Support Vector Machine (SVM) classifier on three fairly challenging publicly available image datasets, namely the UIUC Sports Event dataset, the Fifteen Scene Categories dataset and the MIT Scene dataset. Experimental results reveal that the proposed WLBP descriptor outperforms the traditional LBP technique and yields results better than some other popular image descriptors.**

I. Introduction

Content-based image classification, search and retrieval is an active and growing research area due to the presence of an increasingly large volume of uncategorized user-generated images over the Internet. The area of scene image classification in particular, has seen a steady series of improvements in the recent years [1], [2], [3], [4]. Since the advent of the bag of visual words model [5], there have been notable contributions to enhance recognition performance by developing new and robust image descriptors as well as effective classification frameworks that have resulted in reduced quantization loss and improved recall performance [6].

This paper addresses the problem of recognizing scene images by encoding local image information. The goal of this work is to design a new feature descriptor that can lead to an effective classification performance. To this end, we start by choosing the computationally efficient Local Binary Patterns (LBP) descriptor that captures the variation in intensity between neighboring pixels to encode texture from images [7], [7]. The LBP method has been found suitable for scene classification tasks [8] and hence has been used alone or along with other features to develop new image descriptors [9], [10]. The Wigner distribution has been extensively used in signal processing. In this paper, we have applied the pseudo-Wigner distribution of images as a part of our feature extraction framework and used a multi-neighborhood LBP to derive the innovative bag-of-words based WLBP descriptor that significantly improves classification performance.

II. Background

This section gives a brief outline of the concepts that have been used for generating our proposed descriptor.

A. Pseudo-Wigner Distribution

The Wigner distribution, also known as Wigner-Ville distribution is a generalized time-frequency representation proposed by Wigner [11] and Ville [12] in 1932 and 1948 respectively. Although it has been extensively used in signal processing area, its applications in image processing are limited. Jacobson and Wechsler [13] were the first researchers to apply the Wigner distribution to solve image processing problems. A family of Wigner distributions is called the pseudo-Wigner distribution [14].

In order to use the Wigner distribution function for image processing applications, it needs to be extended to two-dimensional space. Thus Wigner distribution of a two dimensional image is a four-dimensional distribution function which has two space domain variables and two frequency domain variables. The concept of windows is also applied here, which allows applying a sliding window to the original function in the time domain.

In this work, we have used pixel-wise pseudo-Wigner distribution for grayscale images, calculated with a N-pixels-one dimensional oriented square window where N is the operational window size. To compute the pixel-wise Wigner-distribution (W) of an image X, the algorithm takes an array of N pixels arranged in direction θ. For our purposes, we have chosen the function to be periodic which takes the $(N+1)$ pixel value to be equal to the value determined by the image in position $N = 1$. Hence, for each pixel (i, j) of an image X, $W(i, j, k)$ is the pseudo-Wigner distribution of that pixel in the image, where $1 \leq k \leq N$. We have only chosen the first plane of W to design our WLBP descriptor.

B. Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) method encodes the texture features from a grayscale i.e. intensity image by comparing each pixel with its neighboring pixels [15], [7]. Specifically, 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 $N(\mathbf{p}, i) = [x_i, y_i]^t$, $i = 0, 1, \cdots, 7$, 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:

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

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

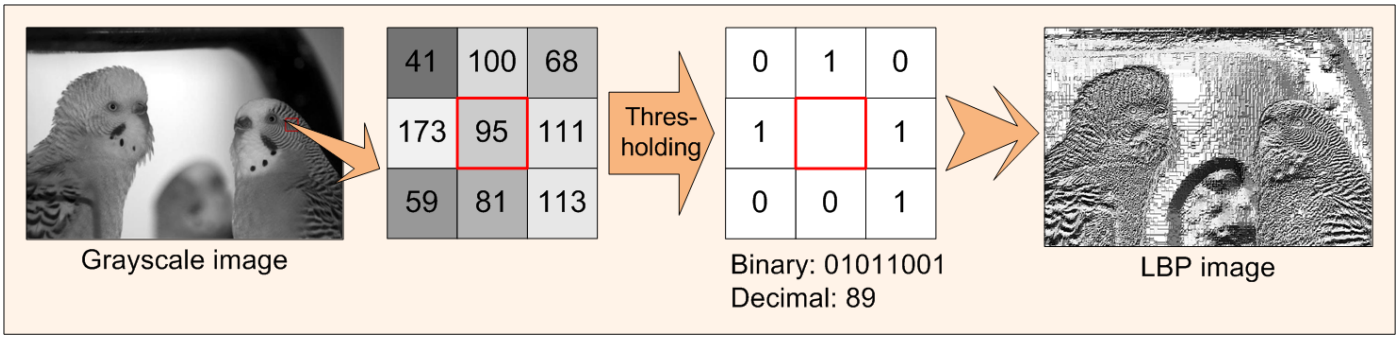

Fig. 1. A grayscale image, its LBP image, and the illustration of the computation of the LBP code for a center pixel with gray level 95.

where $G(\mathbf{p})$ and $G[N(\mathbf{p}, i)]$ are the gray levels 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 achieves grayscale invariance to a large extent because only the signs of the differences in intensities between the center pixel and its neighbors are used to define the value of the LBP code, rather than their absolute intensities. This is shown in Equation 1. Figure 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 95.

III. Feature Description and Classification

In this section, we will present the methodology adopted for developing our WLBP image descriptor.

A. Sampling and Bag of Features

In order to derive the WLBP descriptor, we first start with sampling the image. Popular descriptors like SIFT [16] use multiscale keypoint detectors such as Laplacian of Gaussian or Harris-affine to select regions of interest within the image. This sampling method is appropriate for object recognition, but it has been found that dense sampling often outperforms the keypoint-based sampling methods [17]. This is particularly true of images with large uniform regions, where SIFT does not detect any keypoints. Scene images, such as the ones used for this work, often have such homogeneous regions depicting the sky or walls. For this purpose, we have used a dense sampling approach in which the image is divided into a number of equal sized overlapping square blocks or patches using a uniform grid and each block is used as a separate region for extracting features. We have sampled the scene images using 40×40 pixel overlapping blocks, each block offset by 10 pixels from the next. Such patches are extracted from all training images and then the patches are clustered to form visual words.

B. Multi-Scale WLBP Features for Small Image Patches

We now discuss the feature extraction of the sampled image regions. First, the pixel-wise pseudo-Wigner distribution for each of the small image patches is computed as described in Section II-A in three different directions. For our experiments, we have used the parameter values to be $N = 2$, $\theta = 0, \pi/4, \pi/2$, and have only retained the first planes of each of the three Wigner distributions for the image blocks for subsequent feature extraction.

We then extract multi-neighborhood LBP features from the image patch and the three images produced as a result of applying the Wigner-distribution on it. Different researchers have chosen various neighborhoods of different styles for extracting LBP features from an image [18], [8], [19]. The conventional 8-neighborhood LBP mask assigns one out of 2^8 possible intensity values to each pixel, resulting in a 256-bin histogram. However, since we are dealing with small image patches, we have chosen 4-pixel neighborhood LBP masks to reduce the sparseness of the features. These LBP masks produce a dense 16-bin histogram, and eight such histograms from different neighborhoods and four sub-images are fused to design the 128-dimensional WLBP feature vector describing each image block. Figure 2 depicts the two 4-pixel neighborhoods used for generating the multi-neighborhood LBP descriptor used here.

The Discrete Cosine Transform (DCT) is a well-known technique of transforming an image to the frequency domain for various applications like compression, smoothing, etc. [20], where an image is decomposed into a combination of various uncorrelated frequency components. Specifically, the DCT of an image with the spatial resolution of $M \times N$, $f(x, y)$, where $x = 0, 1, \cdots, M-1$ and $y = 0, 1, \cdots, N-1$, transforms the image from the spatial domain to the frequency domain [21]. DCT is thus able to extract the features in the frequency domain to encode different image details that are not directly accessible in the spatial domain. Due to these

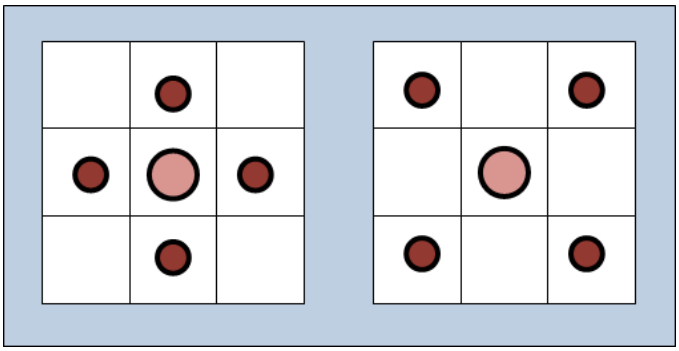

Fig. 2. The two 4-neighborhood LBP masks used for computing the proposed WLBP descriptor.

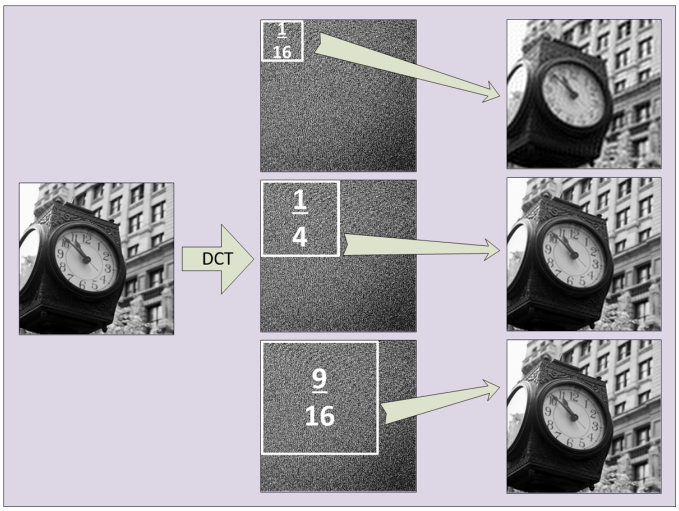

Fig. 3. DCT can be used for smoothing out the image. The original image is transformed to the frequency domain and the lowest 1/16, 1/4 and 9/16 parts are used for regenerating the image, respectively, resulting in three output images with various degrees of smoothing.

specific properties, DCT has been successfully applied to face recognition [22], [23], [20]. In the proposed method, DCT is used to eliminate higher frequencies from an image, resulting in a form of smoothing. To achieve image smoothing for capturing textures at different scales, we apply DCT to transform the original image to frequency domain and use the lowest 6.25%, 25% and 56.25% of frequencies to regenerate the image. This process is explained in Figure 3. The original image and the three images thus formed undergo the same process of dense sampling and WLBP feature extraction. All these features together form a bag of features, as shown in Figure 4, that needs to be clustered into distinct visual words to form a visual vocabulary.

Figure 5 illustrates the complete process of generating the WLBP features from a grayscale image.

C. Quantization and Pyramid Representation

The next stage is to quantize the bag of WLBP features extracted from the training images into a visual vocabulary with discrete visual words. For this step, we have used the popular k-means algorithm. The vocabulary size used by researchers vary from a few hundreds [24], [25] to several thousands [5], [26]. In our work, we have performed experiments with vocabularies of varying sizes and empirically chose a 1000-word vocabulary. After the creation of the visual vocabulary, each scene image is represented by a histogram of visual words. This is explained in Figure 6(a).

The image pyramid representation proposed by [24] allows a descriptor to represent local image features and their spatial layout. Here, at each level, an image is tiled into its successively smaller blocks and the feature vectors are computed for each block. These features from each pyramid level are then weighted accordingly, which are finally concatenated to form a pyramid histogram. This technique is explained in Figure 6(b). It should be noted that the histograms shown in Figure 6 are for illustration purposes only. For this work, only the second level of this pyramid has been utilized to keep the computational complexity low. Finally, a 4000 dimensional feature vector is constructed for each image.

D. Classifier Used

After all training and test images have been processed and the feature vectors have been generated, an SVM classifier is used for classification. It is a known fact in texture and other image classification that for comparing histograms, using χ^2 or Hellinger distance measures usually yields better results than Euclidean distance [27] . The use of the Hellinger kernel has been shown to benefit SIFT [27]. Since the proposed WLBP descriptor is also a histogram, intuitively it seems that it should yield better classification results with the Hellinger kernel and it is empirically seen that using the Hellinger kernel does indeed improve the classification results greatly.

If x and y are n-vectors with unit Euclidean norm ($|x|_2 = 1$), then the Euclidean distance $d_E(x, y)$ between them is related to their similarity (kernel) $S_E(x, y)$ as

$$d_E(x,y)^2 = |xy|_2^2 = |x|_2^2 + |y|_2^2 - 2x^t y = 2 - 2S_E(x,y) \quad (3)$$

where $S_E(x, y) = x^t y$, and the last step follow from $|x|_2^2 = |y|_2^2 = 1$. The Euclidean similarity/kernel here needs to be replaced by the Hellinger kernel.

The Hellinger kernel, which is also known as the Bhattacharyya's coefficient, is defined for two L1 normalized histograms, x and y (i.e. $\sum_{i=1}^{n} x_i = 1$ and $x_i \geq 0$) as:

$$H(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i} \quad (4)$$

Arandjelović et al. suggest a simple algebraic manipulation to compare SIFT vectors by a Hellinger kernel [27]. Since WLBP vectors are also based on histograms of words, the same technique can be applied to the WLBP vectors as well. This can be done in two steps: (i) L1 normalize the WLBP vector (originally it has unit L2 norm); (ii) square root each element. It then follows that $S_E(\sqrt{x}, \sqrt{y}) = \sqrt{x}^t \sqrt{y} = H(x, y)$, and

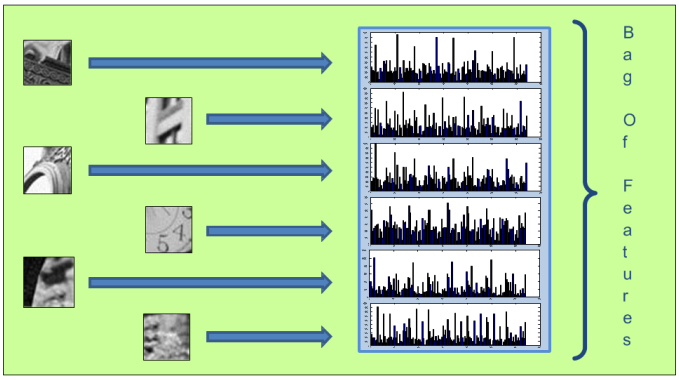

Fig. 4. The features are computed from a large number of image patches from all training images and form a bag of features from which a visual vocabulary can be created.

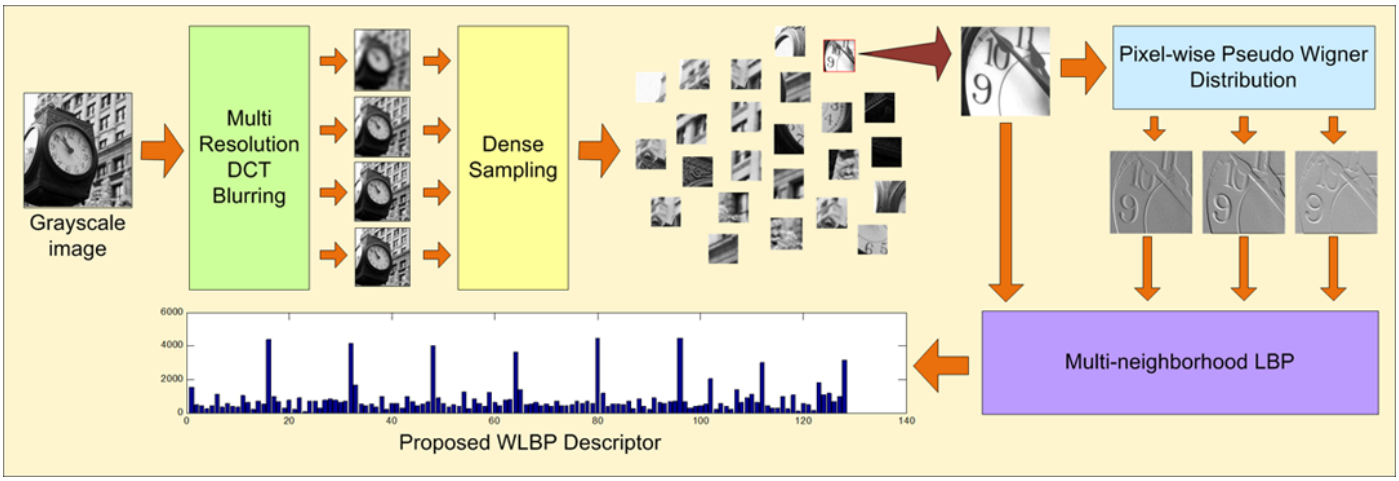

Fig. 5. The process of computing the proposed WLBP descriptor has been simplified in this schematic diagram.

the resulting vectors are L2 normalized since $S_E(\sqrt{x}, \sqrt{y}) = \sum_{i=1}^{n} = 1$ [27].

The key point is that comparing the square roots of the WLBP descriptors using Euclidean distance is equivalent to using the Hellinger kernel to compare the original WLBP vectors:

$$d_E(\sqrt{x}, \sqrt{y})^2 = 2 - 2H(x, y) \tag{5}$$

For the classification process, an SVM is trained independently for each class (one-vs-all classification). This is repeated for each category separately and the precision rates from all the iterations give the average precision which is the mean classification accuracy. A similar configuration has been successfully used by other researchers like [28] in recent works. The SVM implementation used here is the one that is distributed with the VlFeat package [29].

IV. Experiments

In this section, we first briefly introduce the three scene image datasets used for evaluating the classification performance of the WLBP descriptor, and then we make a comparative assessment of the classification performances of the LBP and the WLBP descriptors. Finally we compare the classification performance of the WLBP descriptor with that of some popular image descriptors used by other researchers on these datasets. It should be noted that the results of other researchers are reported directly from their published work.

A. Datasets Used

Three publicly available and widely used image datasets are used in this work for assessing the classification performance of the proposed descriptor.

1) The UIUC Sports Event Dataset: The UIUC Sports Event dataset [30] contains 1,574 images from eight sports event categories: 250 rowing, 200 badminton, 182 polo, 137 bocce, 190 snowboarding, 236 croquet, 190 sailing, and 194 rock climbing. The mean image size in this dataset is 966×1156 pixels. These images contain both indoor and outdoor scenes where the foreground contains elements that define the category. The background is often cluttered and is similar across different categories like rowing and sailing, or croquet and polo. Some sample images from this dataset are displayed in Figure 7(a).

2) The MIT Scene Dataset: The MIT Scene dataset (also known as OT Scenes) [2] has 2,688 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. The sources of the images vary (from commercial databases, websites, and digital cameras) [2]. Figure 7(b) shows a few sample images from this dataset.

3) The Fifteen Scene Categories Dataset: The Fifteen Scene Categories dataset [24] is composed of 15 scene categories: thirteen were provided by [1], eight of which were originally collected by [2] as the MIT Scene dataset, and two were collected by [24]. Each category has 200 to 400 images, most of which are grayscale. Figure 7(c) shows a few images from this dataset.

B. Comparative Assessment of the LBP, the WLBP and Other Popular Descriptors

We now evaluate the classification performance of our proposed WLBP descriptor by comparing it with the traditional LBP feature and some other popular image descriptors on the three scene image datasets. To that end, we first derive the WLBP feature vector from each image in the dataset. To compute the WLBP descriptor, first each color image is converted to grayscale and then all the training images are divided into overlapping uniform image patches. Please note that the large scale images are resized in such a way that their largest dimension does not exceed 256 pixels. The WLBP features are extracted from all the image patches generated from the grayscale image and the three DCT-smoothed images to generate a bag of features which is quantized using the k-means algorithm to form a visual vocabulary with 1000 words. Next each training and test image is represented as

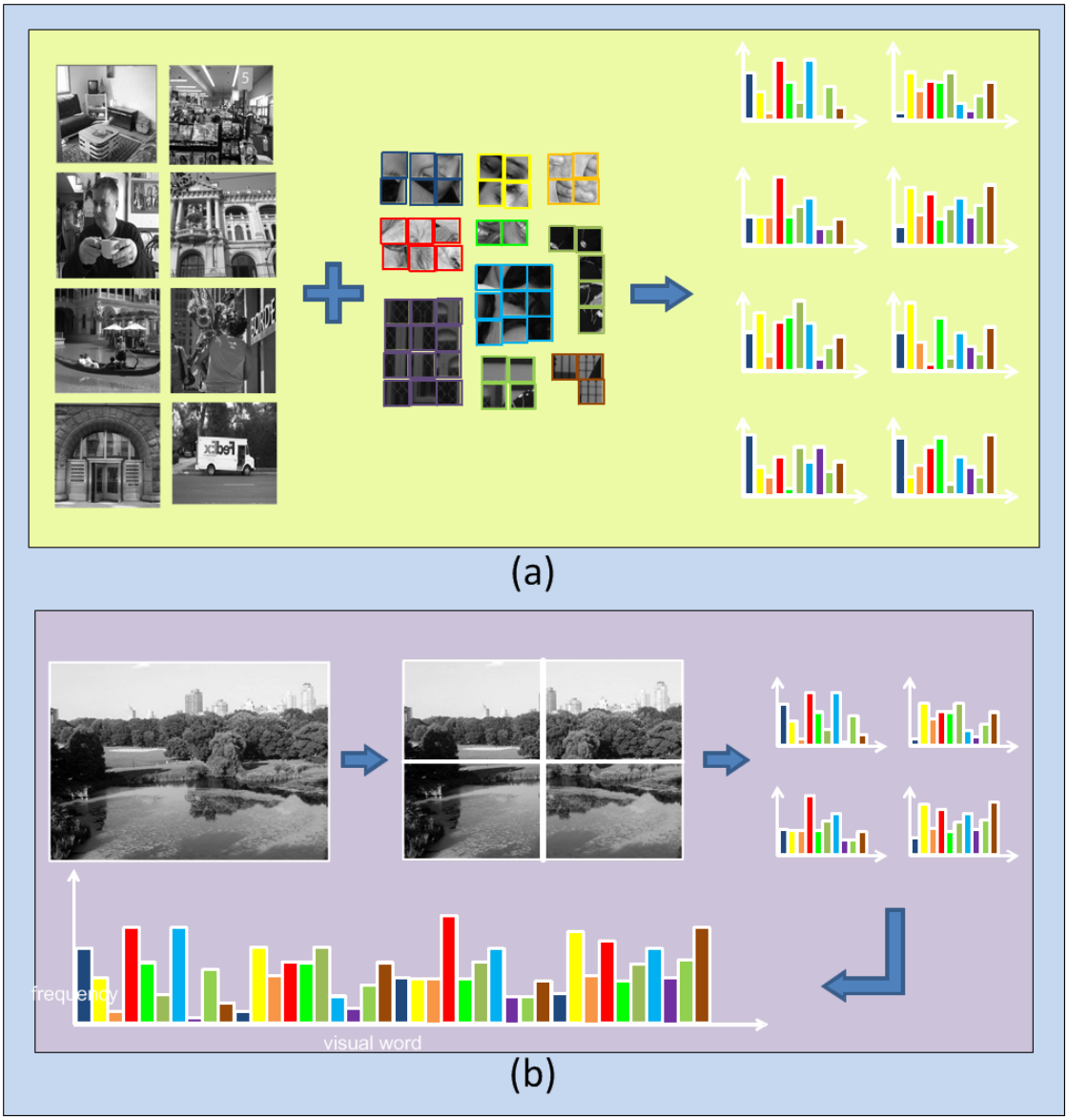

Fig. 6. (a) All images are converted to histograms of visual words using the visual vocabulary created from the training images. (b) For the spatial pyramid representation, a full image is broken down into multiple spatial tiles. Then histograms of visual words are computed from each tile and concatenated.

a pyramid histogram of these visual words. We use an SVM classifier with a Hellinger kernel [31], [29] for evaluating the relative classification performances of the LBP and the WLBP descriptors.

For the UIUC Sports Event dataset, we use 70 images from each class for training and 60 from each class for testing both the LBP and the WLBP descriptors. The results are obtained using five random splits of data where there is no overlap between the training and testing images of the same split. Figure 9 shows the relative average precisions achieved by the LBP and the WLBP descriptors on this dataset. Note that here, the horizontal axis shows the two descriptors and the three datasets, and the vertical axis shows the classification performance measured by average precision as percentage. Here, the WLBP descriptor outperforms the LBP by over 14%. The proposed WLBP vector also produces better results than other SIFT-based and state-of-the-art methods on this dataset, which is listed in Table I.

For the MIT Scene dataset, we used the protocol defined in [2] where 100 images from each class are used for training and the remaining images for testing the performance. Here also, the WLBP significantly improves over the LBP feature, by a margin of 18%, and achieves an average precision of 92.17% (as shown in Figure 9) which is a very good result for this dataset. Table I shows a comparative evaluation of results obtained by other methods and by our proposed descriptor on this dataset.

On the Fifteen Scene Categories dataset, we use 100 training images from each category and rest for testing and the results are measured from five runs of experiments. Here, the overall performance of the WLBP is 85.13% which is again, much better than the traditional LBP as is evident

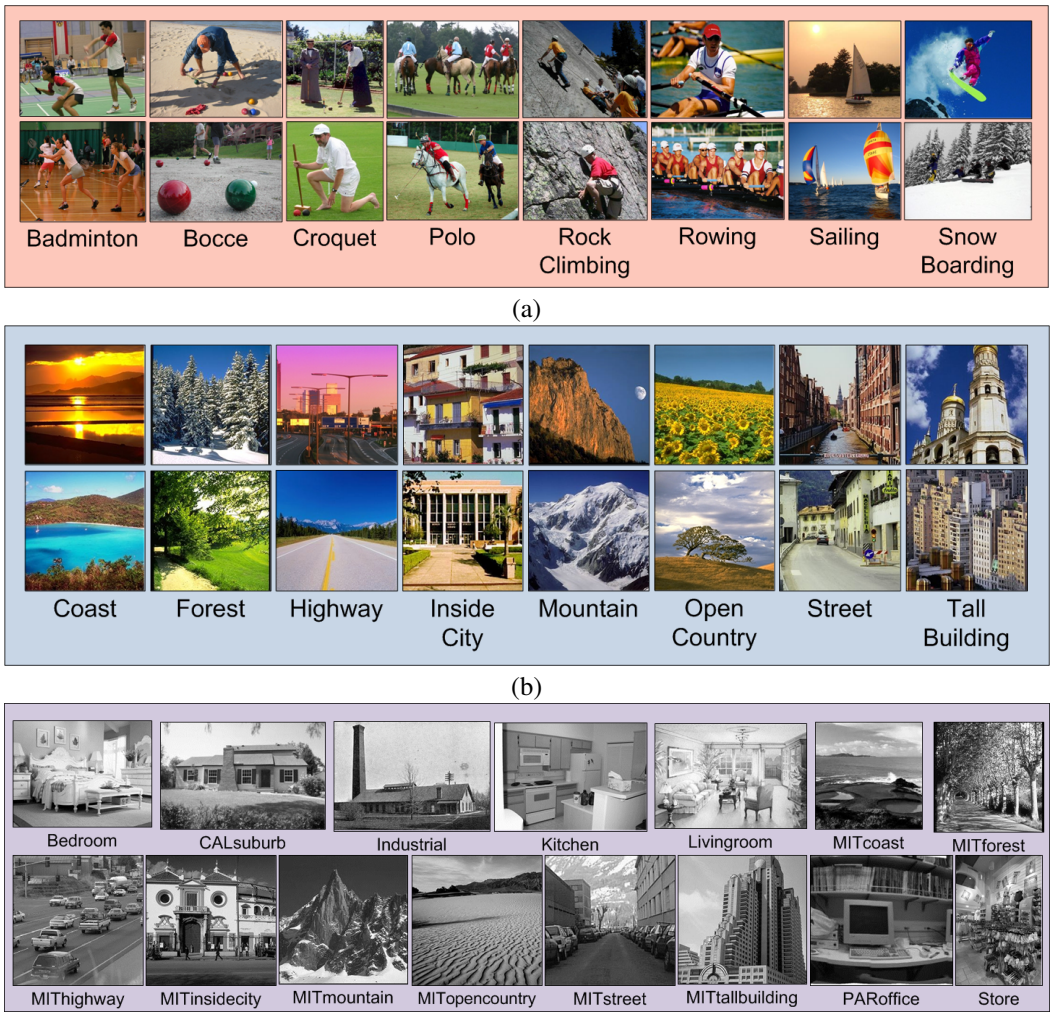

Fig. 7. Some sample images from (a) the UIUC Sports Event dataset, (b) the MIT Scene dataset, and (c) the Fifteen Scene Categories dataset.

from Figure 9. In addition, the category-wise classification performances of the grayscale LBP and the proposed WLBP features is displayed in Figure 8. Here, the horizontal axis reveals the fifteen scene categories, and the vertical axis displays the classification performance. A detailed comparison of the WLBP and other competitive methods on this dataset is given in Table I.

V. Conclusion

We have presented a new local image descriptor for recognizing scene images by applying the Wigner distribution and a multi-neighborhood LBP technique on image patches. Combined with the DCT-based smoothing technique, the bag-of-visual words model and the spatial pyramid image representation and coupled with the SVM classifier, our new image descriptor significantly improves image classification

TABLE I. Comparison of the Classification Performance (%) of the Proposed Grayscale WLBP Descriptor with Other Popular Methods on the Three Image Datasets

Method		UIUC Sports Event	MIT Scene	Fifteen Scene Categories
SIFT+GGM	[30]	73.4	-	-
OB	[32]	76.3	-	-
KSPM	[33]	-	-	76.7
KC	[4]	-	-	76.7
CA-TM	[34]	78.0	-	-
ScSPM	[33]	-	-	80.3
SIFT+SC	[3]	82.7	-	-
SE	[2]	-	83.7	-
HMP	[3]	85.7	-	-
C4CC	[35]	-	86.7	-
WLBP+SVM (Proposed)		**86.2**	**92.2**	**85.1**

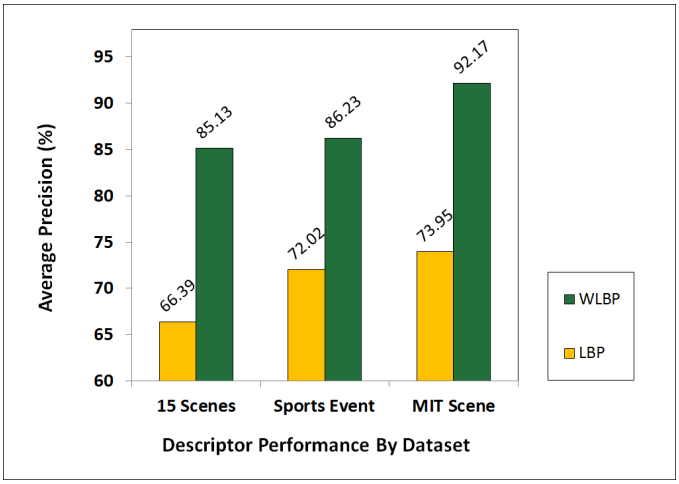

Fig. 9. Comparison of the classification performance of the LBP and the proposed WLBP descriptors using an SVM classifier with a Hellinger kernel on the three datasets.

performance over LBP. Experimental results on three popular scene image datasets show that the WLBP descriptor yields better classification performance than several recent state-of-the-art methods used by other researchers, such as the popular nonlinear Kernel Spatial Pyramid Matching (KSPM), SIFT Sparse-coded Spatial Pyramid Matching (ScSPM) and the Kernel Codebook (KC).

REFERENCES

[1] 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.

[2] 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.

[3] L. Bo, X. Ren, and D. Fox, "Hierarchical matching pursuit for image classification: Architecture and fast algorithms," in *Neural Information Processing Systems*, Granada, Spain, 2011, pp. 2115–2123.

[4] J. Van Gemert, C. Veenman, A. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271–1283, 2010.

[5] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Ninth IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 1470–1477.

[6] R. Arandjelović and A. Zisserman, "All about VLAD," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[7] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[8] 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.

[9] A. Sinha, S. Banerji, and C. Liu, "Novel color gabor-lbp-phog (glp) descriptors for object and scene image classification," in *The Eighth Indian Conference on Vision, Graphics and Image Processing*, Mumbai, India, December 16-19 2012, p. 58.

[10] S. Banerji, A. Sinha, and C. Liu, "New image descriptors based on color, texture, shape, and wavelets for object and scene image classification," *Neurocomputing*, vol. 117, no. 0, pp. 173–185, 2013.

[11] E. Wigner, "On the Quantum Correction For Thermodynamic Equilibrium," *Physical Review Online Archive (Prola)*, vol. 40, no. 5, pp. 749–759, 1932.

[12] J. Ville, "Theorie et Applications de la Notion de Signal Analytique," *Cables et Transmission*, vol. 1, pp. 61–74, 1948.

[13] L. Jacobson and H. Wechsler, "Derivation of optical flow using a spatiotemporal-frequency approach," *Computer Vision, Graphics, and Image Processing*, vol. 38, no. 1, pp. 29–65, 1987.

[14] V. G. Vaidya and R. M. Haralick, "Wigner distribution for 2d motion estimation from noisy images," *Journal of Visual Communication and Image Representation*, vol. 4, no. 4, pp. 281–297, 1993.

[15] 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.

[16] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *9th European Conference on Computer Vision*, Graz, Austria, 2006, pp. 490–503.

[18] 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.

[19] J. Gu and C. Liu, "Feature local binary patterns with application to eye detection," *Neurocomputing*, vol. 113, no. 0, pp. 138–152, 2013.

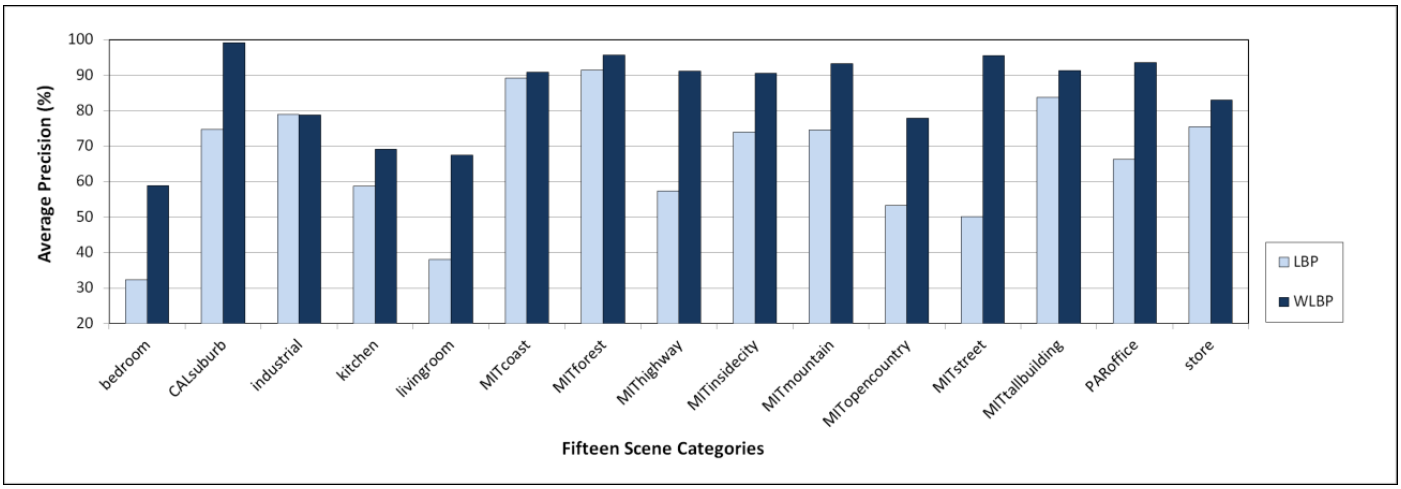

Fig. 8. The comparative average classification performance of the LBP and the WLBP descriptors on the 15 categories of the Fifteen Scene Categories dataset.

[20] Z. M. Hafed and M. D. Levine, "Face recognition using the discrete cosine transform," *International Journal of Computer Vision*, vol. 43, no. 3, pp. 167–188, July 2001.

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

[22] Z. Liu and C. Liu, "Fusion of the complementary discrete cosine features in the YIQ color space for face recognition," *Computer Vision and Image Understanding*, vol. 111, no. 3, pp. 249–262, 2008.

[23] W. Chen, M.-J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 2, pp. 458–466, 2006.

[24] 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.

[25] 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, June 2007.

[26] W. Zhao, Y. Jiang, and C. Ngo, "Keyframe retrieval by keypoints: Can point-to-point matching help," in *The Fifth International Conference on Image and Video Retrieval*, Tempe, AZ, USA, 2006, pp. 72–81.

[27] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Singapore, December 4-6 2009, pp. 1794–1801.

[34] 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.

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