# A New Bag of Words LBP (BoWL) Descriptor for Scene Image Classification

Sugata Banerji*, Atreyee Sinha, and Chengjun Liu

Department of Computer Science,
New Jersey Institute of Technology,
Newark, NJ 07102, USA
{sb256,as739,cliu}@njit.edu

**Abstract.** This paper explores a new Local Binary Patterns (LBP) based image descriptor that makes use of the bag-of-words model to significantly improve classification performance for scene images. Specifically, first, a novel multi-neighborhood LBP is introduced for small image patches. Second, this multi-neighborhood LBP is combined with frequency domain smoothing to extract features from an image. Third, the features extracted are used with spatial pyramid matching (SPM) and bag-of-words representation to propose an innovative Bag of Words LBP (BoWL) descriptor. Next, a comparative assessment is done of the proposed BoWL descriptor and the conventional LBP descriptor for scene image classification using a Support Vector Machine (SVM) classifier. Further, the classification performance of the new BoWL descriptor is compared with the performance achieved by other researchers in recent years using some popular methods. Experiments with three fairly challenging publicly available image datasets show that the proposed BoWL descriptor not only yields significantly higher classification performance than LBP, but also generates results better than or at par with some other popular image descriptors.

**Keywords:** BoWL descriptor, Bag of Words, LBP, Scene Image Classification, Spatial Pyramid.

## 1 Introduction

Content-based image classification, search and retrieval is a rapidly-expanding research area. The large volume of digital images taken worldwide every year necessitates the development of automated classification systems. Apart from classifying large volume of uncategorized images, image recognition has a variety of uses such as weather forecasting, medical diagnostics and robot vision.

The Local Binary Patterns (LBP) descriptor, which captures the variation in intensity between neighboring pixels, was originally introduced to encode the texture from images [1]. Due to its computational efficiency, the LBP feature has been used alone or in conjunction with other features to develop new image descriptors suitable for content-based classification tasks [2], [3], [4].

* Corresponding author.

R. Wilson et al. (Eds.): CAIP 2013, Part I, LNCS 8047, pp. 490–497, 2013.

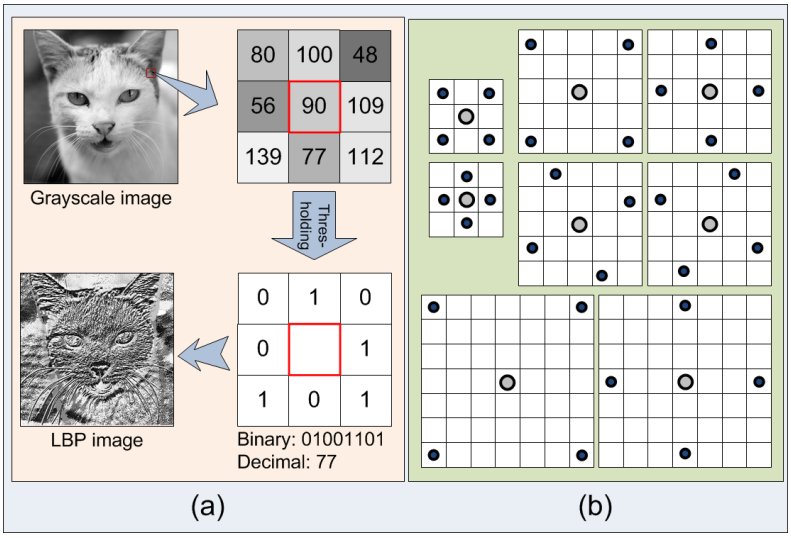


**Fig. 1.** (a) shows a grayscale image, its LBP image, and the illustration of the computation of the LBP code for a center pixel with gray level 90. (b) shows the eight 4-neighborhood masks used for computing the proposed BoWL descriptor.

Lately, part-based methods have been very popular among researchers due to their accuracy in image classification tasks [5]. Here the image is considered as a collection of sub-images or parts. After features are extracted from all the parts, similar parts are clustered together to form a visual vocabulary and a histogram of the parts is used to represent the image. This approach is known as a ”bag-of-words model”, with features from each part representing a ”visual word” that describes one characteristic of the complete image [6].

This paper explores a new bag-of-words based image descriptor that makes use of the multi-neighborhood LBP concept from [7], but significantly improves the classification accuracy.

## 2 An Innovative Bag of Words LBP (BoWL) Descriptor for Scene Image Classification

In this section, we review the LBP descriptor, and then describe the process of computing the proposed Bag of Words LBP (BoWL) descriptor from an image.

### 2.1 Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) method derives the texture description of a grayscale i.e. intensity image by comparing a center pixel with its neighbors [1]. 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. Figure 1(a)

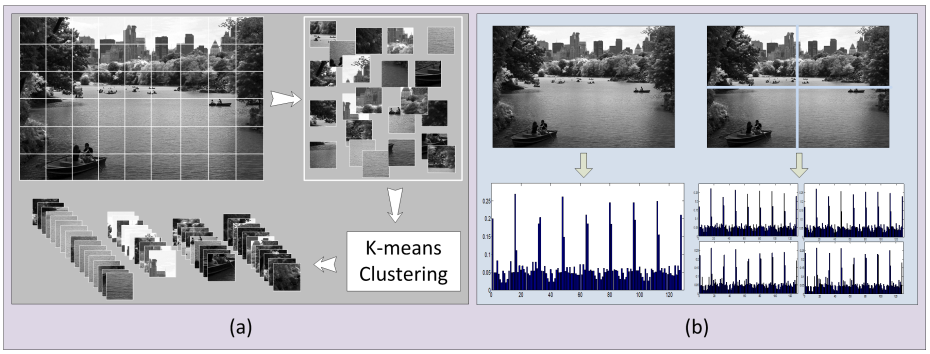


**Fig. 2.** (a) A grayscale image is broken down into small image patches which are then quantized into a number of visual words and the image is represented as a histogram of words. (b) The spatial pyramid model for image representation. The image is successively tiled into different regions and features are extracted from each region and concatenated.

shows a grayscale image on the top left and its LBP image on the bottom left. The two $3 \times 3$ matrices on the right illustrate how the LBP code is computed for the center pixel whose gray level is $90$.

### 2.2 Dense Sampling: Image to Bag of Features

The first step while computing the new BoWL descriptor is sampling. Some image descriptors like SIFT [8] use multiscale keypoint detectors to select regions of interest within the image, but dense or even random sampling often outperforms the keypoint-based sampling methods [9]. In the method proposed here, the image is divided into a large number of equal sized blocks using a uniform grid and each block is used as a separate region for feature extraction. To increase classification performance, overlapping image blocks are used. This process is explained in Figure 2(a).

### 2.3 A Modified LBP for Small Image Patches

Different forms of the LBP descriptor have resulted from different styles of selecting the neighborhood by different researchers [10], [7], [11]. Figure 1(b) shows the eight 4-pixel neighborhoods used for generating the multi-neighborhood LBP descriptor used here. The traditional LBP process assigns one out of $2^8$ possible intensity values to each pixel forming a 256 bin histogram. However, if this technique is applied to a small image patch with $\sim$256 pixels the histogram becomes sparse. To solve this problem, eight smaller neighborhoods of four pixels each are used. These neighborhoods produce a more dense 16-bin histogram, and eight such histograms from different neighborhoods are concatenated to generate the 128-dimensional feature vector describing each image patch.

The Discrete Cosine Transform (DCT) can be used to transform an image from the spatial domain to the frequency domain. 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. In the proposed method, the original image is transformed to the frequency domain and the highest 25%, 50% and 75% frequencies are eliminated, respectively. The original image and the three images thus formed undergo the same process of dense sampling and eight-mask LBP feature extraction.

### 2.4 Bag of Features to Histogram of Visual Words

As demonstrated in the lower part of Figure 2(a), the bag of features extracted from the training images are quantized into a visual vocabulary with discrete visual words using the popular k-means clustering method. The vocabulary size used by other researchers varies from a few hundreds [12] to several thousands and tens of thousands [13]. For the BoWL features, experiments were performed with vocabularies of varying sizes and a 1000-word vocabulary was found to be optimum. After the formation of the visual vocabulary, each image patch from each training and test image is mapped to one specific word in the vocabulary and the image, therefore, can be represented by a histogram of visual words.

Using the image pyramid representation of [12], a descriptor is able to represent local image features and their spatial layout. In this method, an image is tiled into successively smaller blocks at each level and descriptors are computed for each block and concatenated. This technique is explained in Figure 2(b). For this work, only the second level of this pyramid has been used to keep the computational complexity low. This creates a 4000 dimensional BoWL feature vector for each image.

For classification, a Support Vector Machine (SVM) with a Hellinger kernel is trained independently for each class (one-vs-all). The SVM implementation used here is the one that is distributed with the VlFeat package [14].

## 3 Experiments

This section first introduces the three scene image datasets used for testing the new BoWL image descriptor and then does a comparative assessment of the classification performances of the LBP, the BoWL and some other popular descriptors.

### 3.1 Datasets Used

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

**The UIUC Sports Event Dataset.** The UIUC Sports Event dataset [15] contains 1,574 images from eight sports event categories. 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. Some sample images are displayed in Figure 3(a).

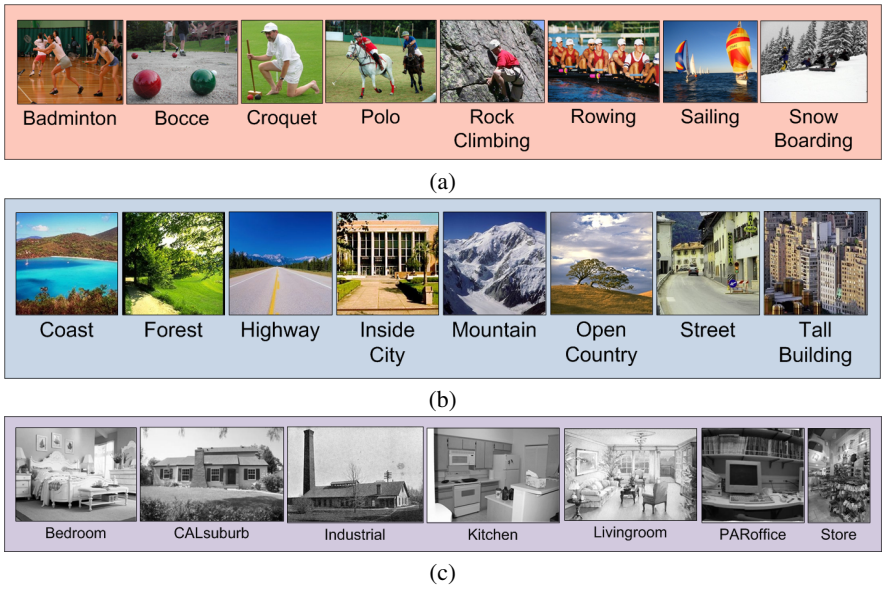


**Fig. 3.** Some sample images from (a) the UIUC Sports Event dataset, (b) the MIT Scene dataset, and (c) the Fifteen Scene Categories dataset

**The MIT Scene Dataset.** The MIT Scene dataset (also known as OT Scenes) [16] has 2,688 images classified as eight categories. There is a large variation in light, content and angles, along with a high intra-class variation [16]. Figure 3(b) shows a few sample images from this dataset.

**The Fifteen Scene Categories Dataset.** The Fifteen Scene Categories dataset [12] is composed of 15 scene categories with 200 to 400 images: thirteen were provided by [5], eight of which were originally collected by [16] as the MIT Scene dataset, and two were collected by [12]. Figure 3(c) shows one image each from the newer seven classes of this dataset.

### 3.2 Comparative Assessment of the LBP, the BoWL and other Popular Descriptors on Scene Image Datasets

In this section, a comparative assessment of the LBP and the proposed BoWL descriptor is made using the three datasets described earlier to evaluate classification performance. To compute the BoWL and the LBP, first each training image, if color, is converted to grayscale. For evaluating the relative classification performances of the LBP and the BoWL descriptors, a Support Vector Machine (SVM) classifier with a Hellinger kernel [17], [14] is used.

For the UIUC Sports Event dataset, 70 images are used from each class for training and 60 from each class for testing of the two descriptors. The results are obtained over five random splits of the data. As shown in Figure 4, the BoWL outperforms the LBP

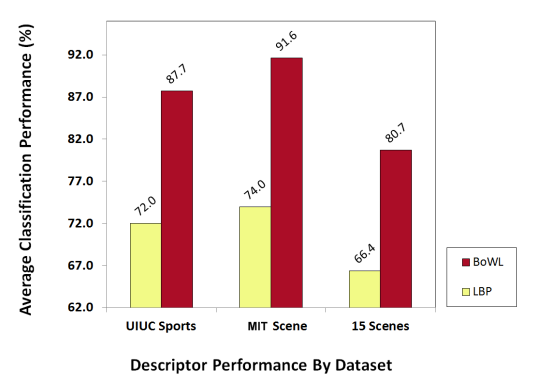


**Fig. 4.** The mean average classification performance of the LBP and the proposed BoWL descriptors using an SVM classifier with a Hellinger kernel on the three datasets

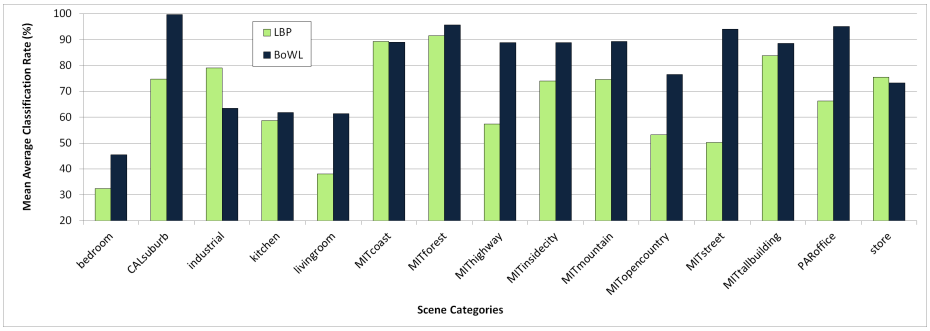


**Fig. 5.** The comparative mean average classification performance of the LBP and the BoWL descriptors on the 15 categories of the Fifteen Scene Categories dataset

by a big margin of over 15%. In fact, on this dataset the BoWL not only outperforms the LBP, but also provides a decent classification performance on its own.

From both the MIT Scene dataset and the Fifteen Scene Categories dataset five random splits of 100 images per class are used for training, and the rest of the images are used for testing. Again, the BoWL produces decent classification performance on its own apart from beating the LBP by a fair margin. Figure 4 displays these results on the MIT Scene dataset and Fifteen Scene Categories dataset. The highest classification rate for the MIT Scene dataset is as high as 91.6% for the BoWL descriptor. The classification performance of BoWL beats that of LBP by a margin of over 17%.

On the Fifteen Scene Categories dataset, the overall success rate for BoWL is 80.7% which is again over 14% higher than LBP. This is also shown in Figure 4. In Figure 5, the category wise classification rates of the grayscale LBP and the grayscale BoWL descriptors for all 15 categories of this dataset are shown. The BoWL here is shown to better the LBP classification performance in 12 of the 15 scene categories.

**Table 1.** Comparison of the Classification Performance (%) of the Proposed Grayscale BoWL Descriptor with Other Popular Methods on the Three Image Datasets

| Method | | UIUC | MIT Scene | 15 Scenes |
|---|---|---|---|---|
| SIFT+GGM | [15] | 73.4 | - | - |
| OB | [18] | 76.3 | - | - |
| KSPM | [19] | - | - | 76.7 |
| KC | [20] | - | - | 76.7 |
| CA-TM | [21] | 78.0 | - | - |
| ScSPM | [19] | - | - | 80.3 |
| SIFT+SC | [22] | 82.7 | - | - |
| SE | [16] | - | 83.7 | - |
| HMP | [22] | 85.7 | - | - |
| C4CC | [23] | - | 86.7 | - |
| **BoWL+SVM (Proposed)** | | **87.7** | **91.6** | **80.7** |

The classification performance of the proposed BoWL descriptor is also compared with some popular image descriptors and classification techniques as reported by other researchers. The detailed comparison is shown in Table 1.

## 4 Conclusion

In this paper, a variation of the LBP descriptor is used with a DCT and bag-of-words based representation to form the novel Bag of Words-LBP (BoWL) image descriptor. The contributions of this paper are manifold. First, a new multi-neighborhood LBP is proposed for small image patches. Second, this multi-neighborhood LBP is coupled with a DCT-based smoothing to extract features at different scales. Third, these features are used with a spatial pyramid image representation and SVM classifier to prove that the BoWL descriptor significantly improves image classification performance over LBP. Finally, experimental results on three popular scene image datasets show that the BoWL descriptor also yields classification performance better than or comparable to several recent methods used by other researchers.

## References

1. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29(1), 51–59 (1996)
2. Banerji, S., Sinha, A., Liu, C.: New image descriptors based on color, texture, shape, and wavelets for object and scene image classification. Neurocomputing (2013)
3. Banerji, S., Sinha, A., Liu, C.: Scene image classification: Some novel descriptors. In: IEEE International Conference on Systems, Man, and Cybernetics, Seoul, Korea, October 14-17, pp. 2294–2299 (2012)
4. Sinha, A., Banerji, S., Liu, C.: 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, p. 58 (2012)

5. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Conference on Computer Vision and Pattern Recognition, pp. 524–531 (2005)
6. Yang, J., Jiang, Y., Hauptmann, A., Ngo, C.: Evaluating bag-of-visual-words representations in scene classification. In: Multimedia Information Retrieval, pp. 197–206 (2007)
7. Banerji, S., Verma, A., Liu, C.: 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, pp. 537–543 (2011)
8. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
10. Zhu, C., Bichot, C., Chen, L.: Multi-scale color local binary patterns for visual object classes recognition. In: International Conference on Pattern Recognition, Istanbul, Turkey, August 23-26, pp. 3065–3068 (2010)
11. Gu, J., Liu, C.: Feature local binary patterns with application to eye detection. Neurocomputing (2013)
12. Lazebnik, S., Schmid, C., Ponce, J.: 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)
13. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Ninth IEEE International Conference on Computer Vision, pp. 1470–1477 (2003)
14. Vedaldi, A., Fulkerson, B.: Vlfeat – an open and portable library of computer vision algorithms. In: The 18th Annual ACM International Conference on Multimedia (2010)
15. Li, L.J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: IEEE International Conference in Computer Vision (2007)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
17. Vapnik, Y.: The Nature of Statistical Learning Theory. Springer (1995)
18. Li, L.J., Su, H., Xing, E.P., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Neural Information Processing Systems, Vancouver, Canada (December 2010)
19. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, Singapore, December 4-6, pp. 1794–(1801)
20. Van Gemert, J., Veenman, C., Smeulders, A., Geusebroek, J.M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(7), 1271–1283 (2010)
21. Niu, Z., Hua, G., Gao, X., Tian, Q.: Context aware topic model for scene recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, pp. 2743–2750 (2012)
22. Bo, L., Ren, X., Fox, D.: Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In: Advances in Neural Information Processing Systems (December 2011)
23. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)