# Spatial Pyramid Local Keypoints Quantization for Bag of Visual Patches Image Representation

Yousef Alqasrawi, Daniel Neagu, Peter Cowling
*School of Computing, Informatics and Media (SCIM)*
University of Bradford
*{Y.T.N.Al-qasrawi, D.Neagu, P.I.Cowling}@bradford.ac.uk*

***Abstract*—Bag of visual patches (BOP) image representation has been the main research topic in computer vision literature for scene and object recognition tasks. Building visual vocabularies from local image feature vectors extracted automatically from images have direct effect on producing discriminative visual patches. Local image features hold important information of their locations in the image which are ignored during quantization process to build visual vocabularies. In this paper, we propose Spatial Pyramid Vocabulary Model (SPVM) to build visual vocabularies from local image features at pyramid level. We show, with experiments on multi-class classification task using 700 natural scene images, that the spatial pyramid vocabulary model is suitable and discriminative for bag-of-visual patches semantic image representation compared to using universal vocabulary model (UVM).**

***Keywords-Bag of visual patches; image classification; Pyramid visual vocabulary; semantic modelling***

## I. INTRODUCTION

The availability of low-cost image capturing devices, popularity of internet, wide use of multimedia-sharing social networks such as *Facebook* and *Myspace*, have led to an increase in the size of image collections. For efficient use of such large image collections, image categorization, searching, browsing and retrieval techniques are required for users from different domains [1-3].

Scene classification has been investigated and analysed in two complementary research areas: human visual perception of scenes and developing computer vision techniques for automatic scene categorization. From the computer vision point of view, scene classification is the task of automatically assigning an unlabelled image into one of several predefined classes. It provides contextual information to help other processes such as object recognition, content-based image retrieval and image understanding [4].

However, designing and implementing algorithms that are capable of successfully recognizing image categories is still a challenging problem [5]. This is because of illumination changes, scale variations, occlusions and large variation between images belonging to the same class and also small variations between images in different classes. This makes the problem of describing (representing) images more complicated.

Early work in scene image classification was based on low-level image features, like colour, shape and texture, extracted automatically from the whole image or from image regions [3, 6-9]. Methods that are based on global image features failed to represent the high-level semantics of user perception which is recognised as a semantic gap in content based image retrieval (CBIR) systems [1, 3].

Semantic modelling refers to the intermediate semantic level representation between low-level image features and image classification to narrow the semantic gap between low-level features and high-level semantic concepts [5, 10]. The simplest way to represent semantic concept is to partition an image into blocks and then to label them manually by human subjects into semantic concepts [5, 11] such systems, though, need time and human work which is time consuming and monetarily expensive.

In recent years, local invariant features or local semantic concepts [12] and the bag of visual patches (BOP) became very popular in computer vision field and have shown impressive levels of performance in scene image classification task [4, 10,13-17]. There are two main parts to build any image classification system within the BOP framework. The first relates to the extraction of features that characterise image content, and the work described in this paper relates to this part. The second part is the classifier. The subsequent elements needed to build bag of visual patches are: feature detection, feature description, visual vocabulary construction and image representation. Each step is performed independently from each other.

In more details, salient point detectors are used to find the location of interest points from the image data set, and then local descriptors are used to extract information from these locations. That means each interest point is characterized by a feature vector, and different images will have different number of feature vectors. Hence, vector quantization techniques are used to build visual vocabularies by clustering feature vectors of all training images on the data set. Each image keypoint is assigned to the index of the closest visual patch in vocabulary. An image is then represented as a histogram counting the number of keypoints that belongs to specific vocabulary index, allowing a classifier to be trained to recognize the categories based on histograms representing images.

Very recently, we have witnessed many successful methods to improve the performance of the conventional BOP paradigm. We can classify these attempts into three main categories. The first category attempts to improve the construction of visual vocabulary [18-22]. The second

978-1-4244-8136-1/10/$26.00 ©2010 IEEE

category suggests using multiple cues with weighting techniques to combine them using early and/or late fusion approaches [16, 23-26]. In the third category, techniques that add spatial information over the BOP have been proven to improve the performance of scene classification tasks [27-30].

In this paper, we propose Spatial Pyramid Vocabulary Model (SPVM) to build visual vocabularies from local image features at pyramid level. This approach is an extension of the Universal Vocabulary Model (UVM) which is used in building BOP image representation. The difference between the two models is that in UVM visual patches are learned by clustering over features at image level and no spatial information is added. In our SPVM approach, visual patches are learned at pyramidal level where local features located in the same feature subspace are clustered together. We studied the effectiveness of the proposed SPVM on natural scene image classification task. Experiments are conducted on six natural scene image categories. Moreover, our approach is compared to the UVM as well as to the spatial pyramid BOP approach [27].

The rest of this paper is organized as follows. In section II we give an overview of related work. In section III we describe the main steps needed to represent and summarise image contents. We describe our approach to build spatial pyramid vocabulary in section IV. We present our experimental setup and results in section V and we conclude in section VI.

## II. Previous Work

Researchers in computer vision observed that text-based techniques can also be extended to represent image contents in image classification and retrieval systems [31]. The bag of words approach is one of these techniques and is very common in text-based information retrieval systems. The analogy between document and image is that both contain information but the main obstacle is how to extract semantic words from image content (i.e., visual patches). In literature, much work has been done in image/object classification and retrieval based on the bag of visual patches. We will review only most related and common approaches that cover the three main categories discussed in section I.

The bag of visual patches started to be very popular and widely used with the development of robust salient features detectors and descriptors such as Scale Invariant Feature Transform (SIFT) [15] features. Csurka et al. [17] and Sivic et al. [32] were the first to use bag of visual patches by clustering the low-level features with the K-means algorithm where each cluster corresponds to a visual patch. Perronnin et al. [21] proposed to build adapted vocabularies by combining universal vocabularies with class specific vocabularies. The universal visual vocabulary describes the visual features of all considered image classes. Adapted visual vocabularies extracted from the universal vocabulary are combined to the universal vocabulary using specific data. It added interesting contribution to the computation of distinctive visual vocabularies. However, their adapted vocabulary does not show the differences between classes and it handles only one kind of image feature.

Another contribution to build discriminative visual vocabularies has been investigated in [33]. They proposed a clustering algorithm to build a visual vocabulary. The algorithm produces an ordered list of centers where a quantization rule assigns patches to the first center in the list that lies within a fixed radius of them, or leaves them unlabeled if there is no such center. In Nilsback and Zisserman [34], several visual vocabularies learned from different aspects (shape, texture and colour) are combined to distinguish 17 flower image classes. In [20], local features are hierarchically quantized in a vocabulary tree, showing improvement in retrieval quality. The visual vocabulary compactness has been investigated recently in [35].

## III. BOP Image Representation

The following subsections details all steps required to build BOP image representation.

*A) Local invariant point detection and description*: we chose to use the Difference of Gaussian (DOG) point detectors and SIFT descriptors [15] to catch and describe local interest points or patches from images. They showed good performance compared to other methods in the literature [13, 36]. The DOG detector has properties of invariance to translation, rotation, scale and constant illumination changes. Once local invariant points are defined, we need to describe them to discriminate their characteristics. SIFT descriptors capture the structure of the local image patches and are defined as local histograms of edge directions computed over different parts of the patch. Each patch is partitioned into 4x4 parts and each part is represented by a histogram of 8 orientations (bins) that gives a feature vector of size 128-d [15]. In this paper we use the binaries provided at [37] to detect DOG local points and to compute the 128-d real valued SIFT descriptors from them.

*B) Visual Vocabulary Construction*: In this section, we describe how to build the universal visual vocabulary UVM. To obtain visual vocabulary, we use the feature vectors (SIFT features) stored in Features Database. All feature vectors from all training images on the data set are quantized, using the k-means algorithm, to obtain $k$ centroids which represent the visual patches. The $k$ visual patches constitute the universal visual vocabulary.

*C) Summarizing image content*: To build the BOP histogram, each image SIFT descriptor is assigned to the index of nearest cluster in the visual vocabulary. The visual patches in the context of this paper refer to the cluster centers (centroids) produced from k-means clustering algorithm.

## IV. Spatial Pyramid Vocabulary Model

Single-resolution visual vocabulary (Universal vocabulary) is commonly used in literature to build BOP image representation. It restricts the information to a certain level of resolution. This leads to our approach to include more information about locations of local keypoints in images. In this section, we illustrate using spatial pyramid vocabulary model shown in Fig. 1. After detecting and extracting local image features from images, training images are divided into 4 spatial tiles $T_1$, $T_2$, $T_3$ and $T_4$. K-means

algorithm is applied to all feature vectors in each tile and result in 4 different visual vocabularies of the same size. A new visual vocabulary $V_{sp}$ is constructed by linearly merging the 4 visual vocabularies learned from each tile. Let $V_{sp}$ denote the set of all visual patches (centroids) produced from the clustering step over a set of local point descriptors as follow:

$$V_{sp} = \left\{ \begin{matrix} v_{T_{11}}, v_{T_{12}}, \ldots, v_{T_{1|v|}}, v_{T_{21}}, v_{T_{22}}, \ldots, v_{T_{2|v|}}, \\ v_{T_{31}}, v_{T_{32}}, \ldots, v_{T_{3|v|}}, v_{T_{41}}, v_{T_{42}}, \ldots, v_{T_{4|v|}} \end{matrix} \right\} \quad (1)$$

where:

$v_{T_{ij}}$ is the j-th visual patch produced from clustering tile $T_i, i = 1..4$ and $|v|$ is vocabulary size. For each visual vocabulary $v_{T_i, i=1..4}$, we adopt fixed vocabulary of size 200 as suggested in [27]. This will result in 800 centroids visual vocabulary ($V_{sp}$).

The set of all SIFT descriptors for each image $d$ is mapped into a histogram of visual patches $h(d)$ at image-level and not at pyramid level, such that:

$$h_j(d) = \sum_{k=1}^{N_d} f_{d_k}^{(j)} \quad , j = 1,..,|V_{sp}| \quad (2)$$

$$f_{d_k}^{(j)} = \begin{cases} 1 & , \ \|u_k - v_j\| \le \|u_k - v_l\| \ , \ l = 1,..,|V_{sp}| \text{ and } j \ne l \\ 0 & , \ \text{otherwise} \end{cases} \quad (3)$$

where:

$h_j(d)$ is the number of descriptors in image $d$ having the closest distance to the *j*-th visual patch $v_j$ and $N_d$ is the total number of descriptor in image $d$. $f_{d_k}^{(j)}$ is equal to one if the *k*-th descriptor $u_k$ in image $d$ is closest to visual patch $v_j$ among other visual patches in the vocabulary $V_{sp}$.

An image $d$ is represented as a histogram $h(d)$ of frequency of each visual patch in that image.

## V. Experiments And Results

In this section we show our image dataset, experimental setup, the classifier we use and the final results we obtain.

*Image dataset:* Most of benchmarks image datasets that are available to the public are dedicated for object image detection and recognition.

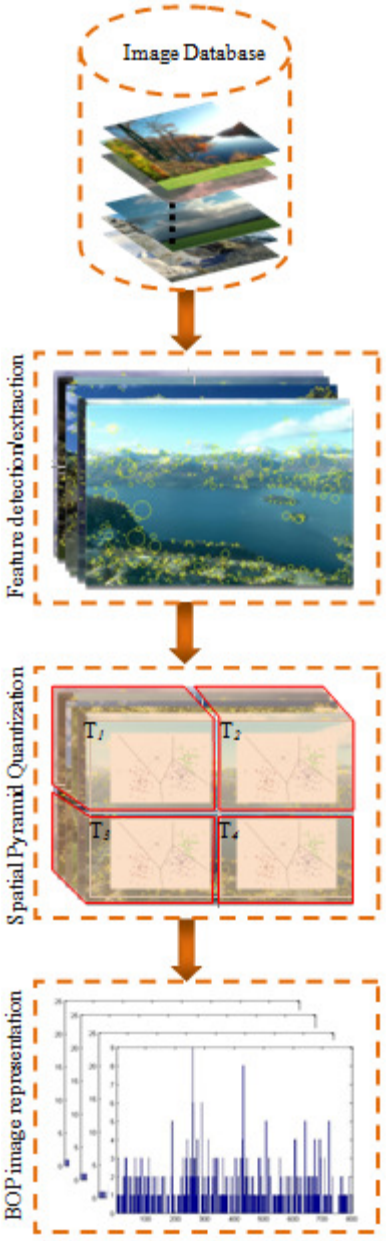


Fig. 1 Spatial Pyramid vocabulary construction model and BOP image representation. Different visual vocabularies are constructed from each tile $T_i$, $i$=1..4.

In our research we aim at recognizing different number of natural scene that has no objects which makes recognition task more challenging Recently, Vogel [5] has built a dataset of 700 natural scene images constituted of six diverse categories. The categories and number of images used are: coasts, 142; rivers/lakes, 111; forests, 103; plains, 131; mountains, 179; sky/clouds, 34. One challenge in this image dataset is the ambiguity and diversity of inter-class and intra-class which makes the classification task more challenging. Fig. 2 shows some examples of each image category.

*Classifier:* For learning recognition of natural scene image classes using BOP image representation, we adopted support vector machine (SVM) [38] as a learning algorithm with default kernel function (Radial Basis Function RBF). We use SVM in our study as they have been empirically proved to yield higher classification accuracy in scene and text classification [7, 8, 13].

All experiments have been validated using 10-fold cross validation where 90% of all images are selected randomly for learning the SVM and the remaining 10% are used for testing. The procedure is repeated 10 times such that all images are actually tested by the SVM classifier. The average of the results over the 10 splits reports the overall classification accuracy of the experiment. We use one-against-one multi-classification approach that results in $M(M\text{-}1)/2$ two-class SVMs for $M$ scene classes.

To evaluate the effectiveness of SPVM approach using the above mentioned experimental setup, we use UVM and spatial pyramid model [27] as baseline methods. We only experiment with visual vocabulary of size 200 in order to compare with Lazebnik et al. approach [27].

Table I shows the classification accuracy of using our model with a vocabulary size 800 (200 for each tile).

To see the effectiveness of our approach, table II demonstrate the accuracy of different baseline methods. Apart from the classification results, data dimensionality is an important factor that affects memory usage, classifier learning time, etc. In BOP+UVM, classification accuracy is (57%) and based on our previous work this accuracy did not improve as we increase vocabulary size. What is interesting in table II is that our SPVM classification accuracy is comparable to the spatial pyramid approach [27], despite the fact that in spatial pyramid approach dimensionality of data increases as we increase the pyramid level. The BOP+SPBOP (L=1 and 2) produces higher dimensional data (1000-d and 4200-d) than our approach (800-d) although classification accuracy is nearly same.

## VI. Conclusion

In this paper, we presented spatial pyramid vocabulary model (SPVM) to study the feasibility and effectiveness of adding spatial information of local keypoint image features in visual vocabulary construction step. The resultant new visual vocabulary is used to build BOP image representation. Experiments on well known natural scene image dataset have shown that constructing BOP using our approach improve classification accuracy over universal vocabulary model. Moreover, our approach achieved comparable accuracy compared to Lazebnik [27] spatial pyramid approach but with much lower dimensional data.

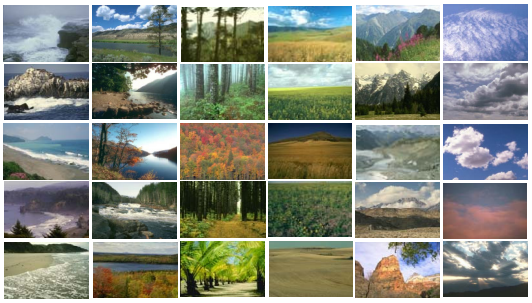

Fig. 2 Examples of the six classes with 5 randomly selected examples per scene class. From left to right: coasts, river/lakes, forests, plains, mountains and sky/clouds.

TABLE I. CLASSIFICATION RATE AND CONFUSION MATRIX FOR THE SIX CLASSES USING BOP WITH SPATIAL PYRAMID VOCABULARY

| Ground Truth | Classification | | | | | | # of img |
|---|---|---|---|---|---|---|---|
| | c | r/l | f | p | m | s/c | |
| coasts | **0.54** | 0.15 | 0.02 | 0.07 | 0.20 | 0.02 | 142 |
| river/lakes | 0.24 | **0.29** | 0.08 | 0.11 | 0.25 | 0.03 | 111 |
| forests | 0.05 | 0.05 | **0.76** | 0.06 | 0.09 | 0.00 | 103 |
| plains | 0.15 | 0.03 | 0.05 | **0.65** | 0.10 | 0.02 | 131 |
| mountains | 0.09 | 0.07 | 0.06 | 0.03 | **0.74** | 0.01 | 179 |
| sky/clouds | 0.15 | 0.00 | 0.00 | 0.15 | 0.03 | **0.68** | 34 |
| Overall performance **61%** | | | | | | | |

TABLE II. COMPARISONS BETWEEN UNIVERSAL AND SPATIAL PYRAMID VOCABULARY MODELS AND SPATIAL PYRAMID APPROACH OF LAZEBNIK [27] IN TERMS OF ACCURACY AND BOP DEIMENSIONALITY

| | Dimensionality | Classification Accuracy |
|---|---|---|
| BOP+UVM | 200 | 57% |
| SPBOP+UVM L=1 | 1000 | 61% |
| SPBOP+UVM L=2 | 4200 | 60% |
| BOP+SPVM | 800 | 61% |

## VII. Acknowledgements

The first author acknowledges the financial support received from the Applied Science University in Jordan. The authors would like to thank Dr. Julia Vogel for providing us access to the natural scene image dataset and for valuable discussion.

## REFERENCES


[1] Ritendra Datta , Dhiraj Joshi , Jia Li , James Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age", ACM Computing Surveys, v.40 n.2, p.1-60, 2008.

[2] Rui, Yong and Huang, Thomas S. "Image retrieval: Current techniques, promising directions and open issues," Journal of Visual Communication and Image Representation, 1999.

[3] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma, "A survey of content-based image retrieval with high-level semantics," Pattern Recognition 40(1): pp. 262-282, 2007.

[4] P. Quelhas, et al., "Modeling scenes with local descriptors and latent aspects," in: Proceedings of IEEE Computer Society International Conference on Computer Vision ICCV, 2005, pp. 883–890.

[5] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," Lecture notes in computer science, 2004, pp. 195-203.

[6] J. Wang, et al., "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," IEEE Transactions on pattern analysis and machine intelligence, 2001, pp. 947-963.

[7] A. Vailaya, et al., "Image classification for content-based indexing," IEEE Transactions on Image Processing, vol. 10, no. 1, 2001, pp. 117-130.

[8] M. Szummer and R. Picard, "Indoor-outdoor image classification," in: Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, Bombay, India, Jan. 1998, pp. 42–51.

[9] 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, 2001, pp. 145-175.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, 2005, pp. 524-531

[11] A. Bosch, et al., "Object and scene classification: what does a supervised approach provide us," in: Proceedings of International Conference on Pattern Recognition, IEEE Computer Society, ICPR, 2006, pp. 773-777.

[12] A. Bosch, et al., "Which is the best way to organize/classify images by content?," Image and vision computing, vol. 25, no. 6, 2007, pp. 778-791.

[13] P. Quelhas, et al., "A thousand words in a scene," in: IEEE Transactions on pattern analysis and machine intelligence, vol. 29, no. 9, 2007, pp. 1575-1589.

[14] D. Gokalp and S. Aksoy, "Scene classification using bag-of-regions representations," in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2007, pp.1-8.

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

[16] P. Quelhas and J. Odobez, "Natural scene image modeling using color and texture visterms," in: Proceedings of International Conference on Image and Video Retrieval, CIVR, Lecture Notes in Computer Science, vol.4071, 2006, pp. 411-421.

[17] G. Csurka, et al., "Visual categorization with bags of keypoints," in: Proceedings of ECCV workshop on Statistical Learning in Computer Vision, 2004, pp. 59–74.

[18] P. Quelhas and J. Odobez, "Multi-level local descriptor quantization for bag-of-visterms image representation," in: in Proceedings of the 6th ACM international Conference on Image and Video Retrieval, 2007, pp. 242-249.

[19] Z. Wu, et al., "A Multi-Sample, Multi-Tree Approach to Bag-of-Words Image Representation for Image Retrieval," in: Proceedings of 12th IEEE International Conference on Computer vision, ICCV, 2009.

[20] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in: Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 2006,pp. 2161-2168

[21] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 7, 2008, pp. 1243-1256.

[22] J. Wu and J. Rehg, "Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," *ICCV09* 2009.

[23] Y. Jiang, et al., "Towards optimal bag-of-features for object categorization and semantic video retrieval," in: Proceedings of the 6th ACM international Conference on Image and Video Retrieval, CIVR, 2007, pp. 494 501.

[24] Y. Alqasrawi, et al., "Natural Scene Image Recognition by Fusing Weighted Colour Moments with Bag of Visual Patches on Spatial Pyramid Layout," *Proc. ISDA09*, IEEE Computer Society, 2009.

[25] J. Yang, et al., "Evaluating bag-of-visual-words representations in scene classification," in: Proceedings of the international Workshop on Workshop on Multimedia information Retrieval ACM MIR, 2007, pp. 197-206.

[26] F. Khan, et al., "Top-Down Color Attention for Object Recognition," in: Proceedings of 12th IEEE International Conference on Computer vision, ICCV, 2009.

[27] S. Lazebnik, et al., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, CVPR, 2006, pp. 2169-2178.

[28] A. Bosch, et al., "Representing shape with a spatial pyramid kernel," in: Proceedings of the 6th ACM international Conference on Image and Video Retrieval, CIVR, 2007, pp. 401-408.

[29] C. Lampert, et al., "Beyond sliding windows: Object localization by efficient subwindow search," in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, 2008, pp. 1-8.

[30] S. Battiato, et al., "Spatial Hierarchy of Textons Distributions for Scene Classification," *Proc. Eurocom Multimedia Modeling*, 2009, pp. 333-342.

[31] L. Zhu and A. Zhang, "Theory of keyblock-based image retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 2, 2002, pp. 224-257.

[32] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," *Proc. ICCV*, 2003, pp. 1470–1477.

[33] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," ICCV, 2005.

[34] M. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," *Proc. CVPR*, IEEE Computer Society, 2006.

[35] J.C. van Gemert, et al., "Comparing compact codebooks for visual categorization," *Computer Vision and Image Understanding*, vol. 14, issue 4, pp. 450-462.

[36] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on pattern analysis and machine intelligence*, 2005, pp. 1615-1630.

[37] http://lear.inrialpes.fr/people/mikolajczyk/.

[38] C.-C. Chang, and Ling, C.-J., "LIBSVM: a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/," 2001