2009 Ninth International Conference on Intelligent Systems Design and Applications

Natural Scene Image Recognition by Fusing Weighted Colour Moments with Bag of Visual Patches on Spatial Pyramid Layout

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*— The problem of object/scene image classification has gained increasing attention from many researchers in computer vision. In this paper we investigate a number of early fusion methods using a novel approach to combine image colour information and the bag of visual patches (BOP) for recognizing natural scene image categories. We propose keypoints density-based weighting method (KDW) for merging colour moments and the BOP on a spatial pyramid layout. We found that the density of keypoints located in each image sub-region at specific granularity has noticeable impacts on deciding the importance of colour moments on that image sub-region. We demonstrate the validity of our approach on a six categories dataset of natural scene images. Experimental results have proved the effectiveness of our proposed approach.**

Keywords-scene image classification; features fusion; semantic modelling

I. INTRODUCTION

The availability of low-cost image capturing devices, wide use of the Internet and popularity of photo-sharing websites such as *Flicker* and *Facebook* hosting hundreds of millions of pictures has led to an increase in 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, 2, 3]. Much research has been done on scene classification recently. Moreover, scene image classification is considered an important task in computer vision community which helps to provide contextual information to guide other vision tasks such as object recognition for organizing personal and professional images and videos [4].

Early work in scene image classification was based on low-level image features, like colour and texture, extracted automatically from the whole image or from image regions [2, 6, 7]. Methods that are based on global image features failed to represent the high level semantic of user perception which is recognised as a semantic gap in content based image retrieval (CBIR) systems [2].

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 [8, 9]. 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 [8, 10]. 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 [11] 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 [4,5,9, 12-15].

Spatial pyramid matching was proposed by Lazebnik et al [14] as an extension to the orderless BOP. Most work that used BOP and spatial pyramid matching focussed mainly on texture analysis but discount image colour information, which we believe has an equal significant importance in recognizing natural scene image categories.

In this paper we propose a simple yet effective weighting method, namely *keypoints density-based weighting* (KDW) method, which is based on the density of quantized local invariant image features over all images sub-regions, to control the fusion of image colour information (colour moments) and BOP histograms on a spatial pyramid layout. Moreover, we use a number of baseline methods (colour histogram, colour moments and the BOP) to represent image content separately. A linear combination of these baseline methods is also conducted to compare their results to our proposed approach.

The rest of this paper is organized as follows: the next section discusses three main steps needed to represent image contents. The spatial pyramid layout and our proposed approach for merging colour information with BOP are explained in section III. The experiments and results are listed in section IV and we conclude the paper in section V.

II. LOCAL IMAGE SEMANTIC REPRESENTATION

In this section we briefly explain the main steps needed to construct the BOP.

A. Local invariant points detection and description

In this work we chose to use the Difference of Gaussian (DOG) point detectors and SIFT (Scale Invariant Feature Transform) descriptors [12] to catch and describe local interest points or patches from images. They showed good performance compared to other methods in the literature [16]. The DOG detector

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.134

IEEE computer society

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 orientation (bins) that gives a feature vector of size 128-D [12]. In this paper we use the binaries provided at [18] to detect DOG local points and to compute the 128-D real valued SIFT descriptors from them.

B. *Summarizing image content (BOP)*

Bag of visual patches provides a summary of image contents. To build the BOP histogram, the first step is to construct the vocabulary of visual patches from SIFT descriptors of training images. The vector quantization is carried out by *k*-means clustering algorithm. The result of the quantization process is a set of clusters that constitute the vocabulary of visual patches. Each image SIFT descriptor is assigned to the index of nearest cluster in the vocabulary. The visual patches in the context of this paper refer to the cluster centers produced from k-means clustering algorithm. Let V denote the set of all visual patches (vocabulary) produced from the clustering step over a set of local point descriptors $V = \{v_i | i = 1,..,|V|\}$, where v_i is the *i-th* visual patch (or cluster) and $|V|$ is the size of the vocabulary. We select to use a vocabulary of size 200 since there have not been observed improvements in performance beyond 200 [14]. The set of all SIFT descriptors for each image d is mapped into a histogram of visual patches $h(d)$ at image-level, such that:

$$h_i(d) = \sum_{j=1}^{N_d} f_{dj}^{(i)} \quad , i = 1,..,|V| \tag{1}$$

$$f_{dj}^{(i)} = \begin{cases} 1 & , \ \|u_j - v_i\| \le \|u_j - v_l\| \ , \quad l = 1,..,|V| \text{ and } i \ne l \\ 0 & , \quad \text{otherwise} \end{cases} \tag{2}$$

where:

$h_i(d)$ is the number of descriptors in image d having the closest distance to the *i*-th visual word v_i and N_d is the number of descriptor in image d.

$f_{dj}^{(i)}$ is equal to one if the j-th descriptor u_j in image d is closest to visual word v_i among other visual words in the vocabulary V.

III. SPATIAL PYRAMID LAYOUT

In this section we briefly review the idea of spatial pyramid matching followed by a description of our proposed approach.

A. *Spatial pyramid matching*

Despite the fact that the orderless bag of visual words approach is widely used and has made a notable performance in object/scene image modelling it overpasses the spatial information and context needed to improve recognize image visual information Spatial pyramid matching [14] works by repeatedly subdividing an image into increasingly coarser sub-regions and then computing histograms of local patches found inside each sub-region An image is represented as a concatenation of weighted histograms at all levels of divisions. Based on this approach, three different hierarchical subdivisions on image regions were recently proposed for recognizing scene categories [17]. In this paper, spatial pyramid layout refers to represent images by placing a sequence of increasingly coarser grids over an image. Here we didn't penalize local histograms of BOP as described in [14, 17] since it decreases the performance of our system.

B. *Proposed approach*

In this section, we describe the proposed approach for modelling image semantic information based on merging BOP and colour moments on spatial pyramid layout.. The motivation of our approach is that most techniques that use BOP rely only on intensity information extracted from local invariant points and neglect colour information which indeed helps in the performance of recognizing scene image categories. We can see in (Fig. 1) an image with circles around interest points detected densely by DOG detectors. What is interesting is that local patches do not cover the complete image, but they seize salient regions in the scene. In natural scene images, colour information has a significant effect in discriminating image areas such as sky, water and sand. Subsequently, we believe that merging colour information and the BOP will be significant in modelling image visual content.

Figure 1. Sample image with circles around interest points. Sky and water contain little information of interest.

The fusion of colour and intensity information on the BOP paradigm is proposed by Quelhas et al. [15]. Quantized colour information and the BOP are computed over local interest regions. Although this approach has shown an improvement on the classification performance, it has two main limitations:

1) Colour information is computed over interest regions only 2) No spatial information is implemented.

Motivated by these facts we adopt the spatial pyramid matching [14] and propose keypoint density-based weighting method KDW for merging colour information and BOP over image sub-regions at all granularities on spatial pyramid layout. The KDW method aims to regulate how important colour information is in each image sub-region before fusing it with BOP. The spatial pyramid layout (Fig. 2) works by splitting an image into increasingly coarser grids over spatial locations of image local points. That is an image with L=2, will have three different representations with an overall of 21 sub-regions ($2^0 + \sum_{l=1}^{L} (2^l)^2$) where the first sub-region is the whole image area. Each image sub-region is represented by a combination of BOP and weighted colour moments vector of size 6 on the HSV colour space (2 for hue, 2 for saturation and 2 for value). Both colour moments and BOP histogram are normalized to unit vector before the merging process. An image with L=2 and visual vocabulary of size 200 will produce 4326-D vector.

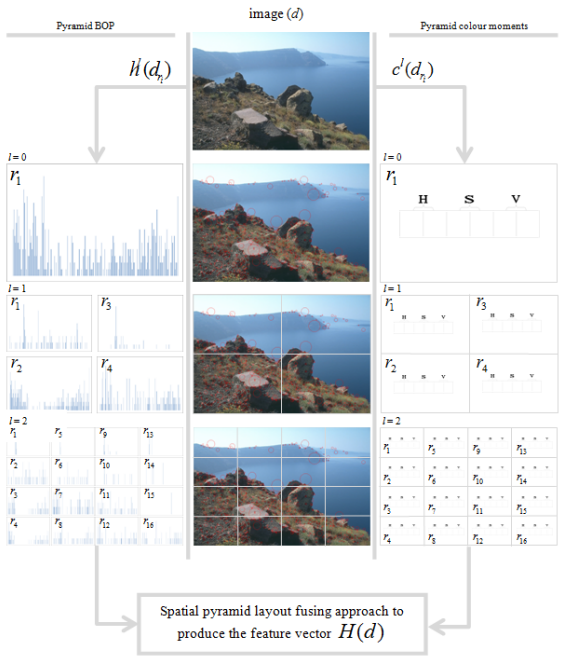

Figure 2. Features fusion model on spatial pyramid layout (L=2). The left column represents histograms of BOP. The right column represents colour moments for the HSV colour space bands. The middle column represents an image at different levels overlaid with circles around interest points.

To formulate our proposed approach we introduce the problem formulation followed by a description of KDW method:

Let L denote the number of levels, $l = 0,1,..,L$, needed to represent an image d on the spatial pyramid layout, i.e., an image d will have a sequence of L grids of increasingly finer granularity. Let $h^l(d_{r_i})$ and $c^l(d_{r_i})$ denote a histogram vector of BOP computed using equation (1) and colour moments vector respectively. Both are computed from an image d at level l and sub-region r_i, $i = (2^l)^2$.

The concept of Keypoint Density-based Weight (KDW): Colour moment vector $c^l(d_{r_i})$ is assigned a high weight on image sub-regions that have a number of local interest points (keypoints density) below threshold $T^l_{r_i}$. Furthermore, colour information will be less important in image sub-regions with high number of local interest points. The threshold T is a real valued vector. Each component represents the average density of keypoints (number of keypoints) at specific image sub-region over all training images. We propose the keypoint density-based weight as:

$$T^l_{r_i} = \frac{1}{m} \sum_{j=1}^{m} h^l(d^j_{r_i}) \qquad (3)$$

where m is the number of images in the training image dataset. The components of the threshold vector, which is the average keypoints density of all images at specific sub-regions and granularity, help in making a decision about the importance of color information at specific image sub-region. The unified feature vector $H(d)$ for image d is a concatenation of weighted colour moments and BOP at all levels and over all granularities:

$$H(d) = \left\langle \begin{matrix} (h^0(d_{r_1}), w^0_{r_1} C^0(d_{r_1})), (h^1(d_{r_1}), w^1_{r_1} C^1(d_{r_1}), ..., h^1(d_{r_4}), w^1_{r_4} C^1(d_{r_4})), \\ (h^L(d_{r_1}), w^L_{r_1} C^L(d_{r_1}), ..., h^L(d_{r_{(2^L)^2}}), w^L_{r_{(2^L)^2}} C^L(d_{r_{(2^L)^2}})) \end{matrix} \right\rangle \qquad (4)$$

$$w^l_{r_i} = \begin{cases} 1 & , \ \sum_{j=1}^{|r|} h^l_j(d_{r_i}) < T^l_{r_i} \\ 0.5 & , \ \text{otherwise} \end{cases} \qquad (5)$$

We should notice that the values of weights w are a non-negative numbers in the range (0 to 1) to indicate the importance of colour information. We aim to force images from the same category to be close, and images from different categories to be far away in the new image representation. Weight values have been obtained empirically during learning the SVM classifiers. In section IV (Fig. 4) shows parts of our experiments to choose the best weighting values. We should notice that weight values are highly dependent on the threshold vector obtained from equation (3). We use the proposed weighted colour moments with the BOP to improve the performance of BOP on spatial

pyramid layout. We applied the idea to recognize natural scene image categories, and the results are presented in section IV.

IV. Experiments And Results

The first part of this section presents the SVM classifier. Next, we demonstrate the dataset we use in our experiments. We also implemented a number of baseline image representations and their early fusion in different ways. We use the confusion matrix to assess the performance of all experiments.

A. Scene classifier

Multi-class classification is done using the support vector machine (SVM) with a RBF kernel. We use SVMs in our study as they have been empirically proved to yield higher classification accuracy in scene and text classification [4, 7, 8]. All experiments have been validated using 10-fold cross validation. That is 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. Our six categories classification problem therefore requires 15 classifiers. To implement the SVM method we used the publicly available LIBSVM software [19] where all parameters are selected based on 10-fold cross validation on each training fold.

B. Image dataset

There are many image datasets available in computer vision literature but most of them are dedicated for object image detection and categorization. Recently, Vogel [8] 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.

For baseline methods, we use colour histogram, the first and second colour moments and the BOP to represent local image contents on gridded and spatial pyramid layout separately. Next, we use baseline methods to get novel fused features which help us to compare and find the best configuration for our proposed KDW approach. All experiments are conducted using 10-fold cross validation and the SVM classifier. For the BOP implementation, we built ten different vocabularies, one for each of the ten training folds.

C. Single and fused features baseline methods

We extract image colour information in the HSV colour space since it is quite similar to the way which humans perceive colour [8]. In this paper we are not quantizing colour information as proposed in [15]. Specifically, we use three types of features to represent image visual content: colour histogram, the first and second moments of each colour channel of an image and the BOP. These features are extracted on gridded and spatial pyramid layout. Each feature type is used separately to represent images and we fuse them in different ways to find suitable features for our proposed approach. The performance of baseline methods is reported in Table I.

Gridded Colour Moments (GCM): the local image colour information is extracted from a regular grid of 10x10 regions as proposed in Vogel [8]. The first two moments of colour for each channel are computed in each region. An image is represented as a normalized colour moments vector (600-D) over all image regions. This GCM method has reported 57% classification accuracy.

Pyramid Colour Moments (PCM): instead of using regular grid, we extract colour moments on spatial pyramid layout at two different levels (L=1 and L=2). An image at level L would be represented by a feature vector of size $S = \sum_{l=0}^{L} \left(2^{l}\right)^{2} * 6$. We achieved an average classification performance of 57% using L=2. We can observe from previous experiment that the PCM achieved equal results to the GCM despite that the GCM representation needs 600-D vector whereas in PCM representation we need only 126-D vector for L=2.

Pyramid Colour Histogram (PCH): we use the traditional colour histogram to represent images on spatial pyramid layout. We also experimented with two different levels (L=1 and L=2). We use 32-bins, 16-bins and 8-bins for the hue, saturation and value respectively. The histograms are then concatenated and normalized to unit vector of size 56-D. This is repeated for all parts of an image and at all granularities. The overall classification performance achieved at L=2 is 59% which is better than the result achieved using PCM.

Bag Of visual Patches (BOP): We use a vocabulary of 200 (as suggested by Lazebnik [14]) and L=1 and L=2. The BOP has an overall accuracy of 58% which is a bit lower than the PCH as well as a bit higher than GCM and PCM (1%). This give us an indication that bag of visual patches model alone is not suitable to represent local image contents.

Pyramid Bag of Visual Patches (PBOP): in this representation, BOP is modelled on a spatial pyramid layout. An improvement (3%) is achieved compared with the BOP. We can conclude from this experiment that spatial pyramid is indeed helpful to incorporate spatial information and increase classification performance. To make our proposed approach rational, we conduct another set of experiments based on direct fusion of previous implementations (BOP+GCM, BOP+PCM, PBOP+GCM, PBOP+PCM). The direct merging has produced higher classification than using

single features alone. This indicates the importance of adding colour information to the BOP model.

D. Proposed approach

In this section we demonstrate the experimental results of our KDW approach to fuse image colour moments and the BOP. It worth mentioning here that this is the first paper that addresses the importance of the density of keypoints in image sub-regions, as well as providing empirical evidence that giving more or less importance to image colour information before the features fusion process can improve classification accuracy. We compare our performance with the methods proposed by Vogel et al. [8] and Quelhas et al. [15] since they used the same dataset. The resulting confusion matrix and the comparison methods results are shown in Table II. When fusing colour information and the BOP without the weighting scheme, for example PBOP+PCM approach, we clearly notice improvements in the classification accuracy for some image classes (such as *forest* class) and decreasing in other classes (such as *sky/clouds* class). Looking at Fig. 3, we observe that our proposed approach behaved constantly in improving classification performance over most image categories compared with other proposed fusion approaches. We reported a classification performance of 69.3% which outperforms approaches proposed in [8, 15]. This suggests that there is the potential for significant improvements in classification accuracy using keypoints density to decide on the importance of colour information.

V. CONCLUSION

We have presented a number of simple yet effective approaches for merging image colour information with the bag of visual patches. Our proposed KDW approach relies on the idea that image colour information should have an increasing importance in image sub-regions where keypoints density is below the average. Also, the colour information has low importance in image sub-regions with keypoints density above the average. Results of our experiments showed an improvement on classification performance applied to a well known benchmark dataset.

TABLE I. CONFUSION MATRIX SUMMARY OF THE EXPERIMENTS COMPARING SIMPLE BASELINE METHODS (THE FIRST 8 COLUMNS) AND THE RESULTS FROM THE FUSION PROCESS WE PROPOSE (THE LAST 4 COLUMNS). THE COLUMNS ARE THE DIAGONAL ENTITIES OF THE CONFUSION MATRIX CORRESPONDING TO A GIVEN EXPERIMENT.

	GCM	**PCM L=1**	**PCM L=2**	**PCH L=1**	**PCH L=2**	**BOP**	**PBOP L=1**	**PBOP L=2**	**BOP + GCM**	**BOP + PCM**	**PBOP + GCM**	**PBOP + PCM**
coasts	0.61	0.53	0.61	0.58	0.64	0.57	0.54	0.54	0.66	0.64	0.66	0.68
river/lakes	0.39	0.46	0.39	0.42	0.41	0.26	0.33	0.34	0.40	0.34	0.41	0.43
forests	0.75	0.70	0.75	0.74	0.73	0.79	0.82	0.78	0.83	0.82	0.86	0.85
plains	0.53	0.48	0.50	0.47	0.48	0.56	0.50	0.55	0.56	0.57	0.59	0.59
mountains	0.58	0.60	0.61	0.68	0.68	0.65	0.77	0.74	0.74	0.76	0.77	0.77
sky/clouds	0.62	0.38	0.50	0.44	0.41	0.71	0.76	0.71	0.68	0.65	0.74	0.71
Average Accuracy	**0.57**	**0.54**	**0.57**	**0.57**	**0.59**	**0.58**	**0.61**	**0.61**	**0.65**	**0.64**	**0.67**	**0.67**

TABLE II. THE FIRST PART OF THIS TABLE SHOWS THE CONFUSION MATRIX OF OUR PROPOSED APPROACH. THE DIAGONAL BOLD VALUES ARE THE AVERAGE CLASSIFICATION RATE OF EACH IMAGE CATEGORY. THE OVERALL SYSTEM ACCURACY IS 69.3% AND IS CLEARLY OUTPERFORMS THE WORK OF VOGEL [8] AND QUELHAS [15].

		Classified as								
		c	**r**	**f**	**p**	**m**	**s**	**Recall**	**Vogel**	**Quelhas**
Correct class	**coasts**	**0.70**	0.11	0.01	0.03	0.14	0.01	0.704	0.599	0.690
	river/lakes	0.23	**0.43**	0.09	0.08	0.16	0.01	0.432	0.416	0.288
	forests	0.03	0.04	**0.86**	0.04	0.03	0.00	0.864	0.941	0.854
	plains	0.11	0.06	0.08	**0.60**	0.15	0.01	0.600	0.438	0.626
	mountains	0.07	0.05	0.03	0.05	**0.79**	0.01	0.793	0.843	0.777
	sky/clouds	0.03	0.03	0.00	0.12	0.00	**0.82**	0.824	1.000	0.765
Overall performance								**69.3%**	67.2%	66.7%

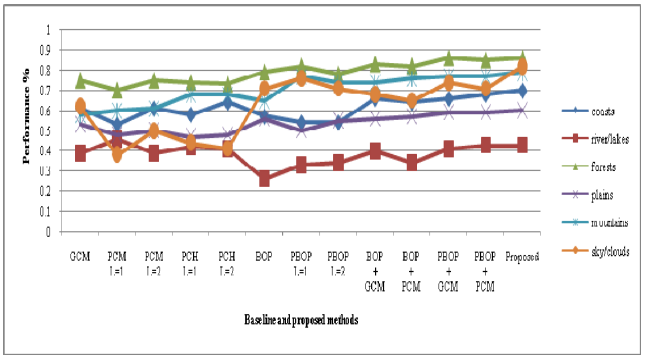

Figure 3. A comparison of the average performance accuracy of different baseline methods and the proposed approach. It is clear that our proposed approach outperforms different baseline methods in most image categories. Each line represents the performance of different methods on a specific image category.

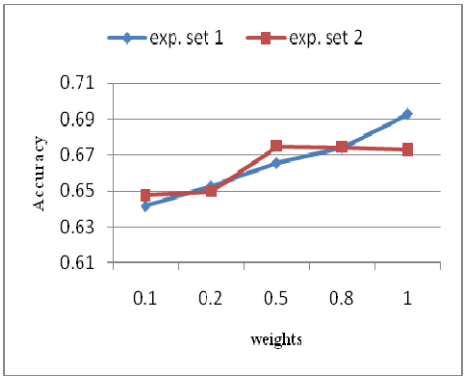

Figure 4. Weights selection experiments. In the first experiment set we used different weights for the high importance with fixed weight (w=0.5) for low importance. In the second experiment, we conducted different weights for the low importance with fixed weight (w=1) for high importance.

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

[2] 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.

[3] 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.

[4] Pedro Quelhas, Florent Monay, Jean-Marc Odobez, Daniel Gatica-Perez, Tinne Tuytelaars, "A Thousand Words in a Scene," IEEE Transactions on PAMI, vol. 29, no. 9, pp. 1575-1589, 2007.

[5] Gokalp, D.; Aksoy, S., "Scene Classification Using Bag-of-Regions Representations," Computer Vision and Pattern Recognition, CVPR, IEEE Conference on, pp.1-8, 2007.

[6] J. Z. Wang, J. Li, and G. Wiederhold, “SIMPLIcity: Semantics-sensitive integrated matching for picture libraries”, IEEE Transactions on PAMI, 23(9):947-963, 2001.

[7] A. Vailaya, A. Figueiredo, A. Jain, H. Zhang, “Image classification for content-based indexing”, IEEE Transactions on Image Processing 10, pp. 117–129, 2001.

[8] J. Vogel and B. Schiele, “Natural Scene Retrieval Based on a Semantic Modeling Step”, Proc. Int'l Conf. Image and Video Retrieval, July 2004.

[9] Fei-Fei, L.; Perona, P., "A Bayesian hierarchical model for learning natural scene categories," Computer Vision and Pattern Recognition, CVPR. IEEE Computer Society Conference on, vol.2, pp. 524-531, 2005.

[10] A. Bosch., X. Munoz, A. Oliver, and R. Marti, “Object and Scene Classification: what does a Supervised Provide us?”, In ICPR, 2006.

[11] Bosch, A., Munoz, X., and Marti, R, “A Review: Which is the best way to organize/classify images by content?”, Image and Vision Computing, vol. 25 no. 6. pp. 778-791, June 2007.

[12] Lowe, D. G., “Distinctive Image Features from Scale-Invariant Keypoints”, Int. J. Computer Vision, pp. 91-110, 2004.

[13] E. Nowak et al., “sampling strategies for bag-of-features image classification”, In ECCV, 2006.

[14] Lazebnik, S.; Schmid, C.; Ponce, J., "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", CVPR, vol.2, no., pp. 2169-2178, 2006.

[15] Quelhas P. and Odobez, J., “Natural Scene Image Modeling Using Color and Texture Visterms”, Proc. Int'l Conf. Image and Video Retrieval, pp. 411-421, 2006.

[16] Mikolajczyk, K.; Schmid, C., "A performance evaluation of local descriptors", IEEE Transactions on PAMI, vol.27, no.10, pp.1615-1630, 2005.

[17] Battiato, S., Farinella, G. M., Gallo, G., and Ravì, D., “Spatial Hierarchy of Textons Distributions for Scene Classification”, In Proceedings of the 15th international Multimedia Modeling Conference on Advances in Multimedia Modeling, pp. 333-343, 2009.

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

[19]http://www.csie.ntu.edu.tw/~cjlin/libsvm/