Scene Classification Using Spatial Pyramid of Latent Topics

Emrah ERGUL
Computer Engineering Department
Turkish Naval Academy
Istanbul, Turkey
eergul@dho.edu.tr

Nafiz ARICA
Computer Engineering Department
Turkish Naval Academy
Istanbul, Turkey
narica@dho.edu.tr

***Abstract*—We propose a scene classification method, which combines two popular methods in the literature: Spatial Pyramid Matching (SPM) and probabilistic Latent Semantic Analysis (pLSA) modeling. The proposed scheme called Cascaded pLSA performs pLSA in a hierarchical sense after the soft-weighted BoW representation based on dense local features is extracted. We associate spatial layout information by dividing each image into overlapping regions iteratively at different resolution levels and implementing a pLSA model for each region individually. Finally, an image is represented by concatenated topic distributions of each region. In performance evaluation, we compare the proposed method with the most successful methods in the literature, using the popular 15-class-dataset. In the experiments, it is seen that our method slightly outperforms the others in that particular dataset.**

Keywords-spatial pyramid matching; probabilistic latent semantic analysis; scene classification; bag of words

I. INTRODUCTION

Scene classification problem addresses analyzing and classifying the images into semantically meaningful categories. The release of many challenging datasets with multiple classes which are supported by recently published papers [1, 2] has proved itself how it is hard and interesting research area. It comprises of many sub-problems including image representation, clustering image features, learning to generate representative models. Among all, image representation is the most important aspect in scene classification problems. The earliest studies have used low-level image features such as color and texture histograms [3]. Later, the Bag of Words (BoW) approach, in which the image is represented as a histogram of local features, has demonstrated remarkable performances [2, 5, 6]. The approaches based on the detection and classification of objects in the images have been proposed with considerable computation cost and evaluated on the databases containing images with single object and background [8]. Recently, the approaches employing the scene prototypes and region of interest have been introduced for combining both the local and global discriminative information [10]. Unfortunately, they are only applicable for some scene types such as indoor scenes.

BoW approaches are still the most promising methods in the previous works for general scene classification [6, 9]. It is improved by Latent Semantic Analysis, which introduces intermediate latent topics over visual words [5]. Another improvement has been achieved by Spatial Pyramid method by partitioning the image into sub-regions and computing the histogram of visual words at each sub-region [2].

In this paper, we combine the Latent Semantic Analysis and Spatial Pyramid methods in representation of images. In addition, we propose to use SIFT-like descriptors [4] in different patch sizes and grid levels for more discriminative low-level feature extraction. Thereafter, we achieve a soft weighted BoW representation similar to [6] followed by intermediate semantic representation using probabilistic Latent Semantic Analysis (pLSA) algorithm. The spatial layout information is given by dividing the image into multiple half-size overlapping regions in a pyramidal division scheme. We then employ pLSA for each region to generate a new probabilistic model which refers to mixture of topics in each corresponding region. After all, we achieve an intermediate semantic representation, which is more robust against the geometric and photometric changes. In the experiments performed on a popular scene dataset containing 15 classes of [2], the proposed method achieves satisfactory results compared to the other generic scene classification methods in the literature.

II. LOCAL IMAGE FEATURES

The local image features are extracted at regular grids all over the image because the comparative results show that utilization of dense keypoints over whole image surface work better than sparsely detected keypoints for scene classification [2, 5]. For this purpose, SIFT-like descriptors are computed on the 16x16 patches with 8-pixel spacing. The proposed descriptor aims to increase the distinctiveness of SIFT method by computing the orientation histograms in a spatial pyramid with three levels. At the first level of the pyramid, the central part with 8x8 pixels of the patch is represented by an orientation histogram with eight bins. The second level describes a larger part with a 12x12 set of pixels using a 2x2 array of histograms. The third level, which covers the whole patch (i.e. 16x16 pixels), is the SIFT descriptor containing the histograms from 4x4 sub-regions. The final patch descriptor is constructed by concatenating the feature vectors from all levels sequentially. It results in a descriptor with a dimension of 168, i.e. 8 + 32 + 128. The illustration of local image feature on a sample patch is depicted at Fig. 1.

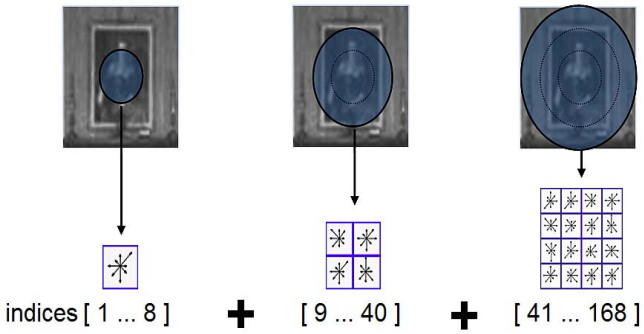

Figure 1. Illustration of local image feature.

III. SOFT-WEIGHTING BOW REPRESENTATION

Next stage after local image description is the image representation based on BoW model, which is simply defined as a histogram of visual words. The visual vocabulary is determined using the k-means clustering algorithm by examining all the local image features in the training images. The center of each cluster is chosen as the representative of the features in that cluster and considered to be the visual words, which generate the visual vocabulary.

The main question in BoW model is how to vote while mapping each local feature to a visual word. The approaches in the literature [2, 6, 8] mostly assign a feature vector to its nearest word and count the votes of features equally. We think that text words are static in every textural corpus provided that they are at the correctly spelled forms. This can lead to semantic approach in the naturally sampled language context. On the other hand, we use data clustering algorithm to find visual words. Thus our visual words are not static and changing due to feature samples and number of cluster centroids. Besides, two features assigned to the same visual word may not equally similar to that visual word since their distances are different.

We provide another weighting scheme similar to that of [6] in co-occurrence table construction. We first find the nearest four visual words for each feature vector in every image with histogram intersection metric. Thereafter, we vote to these four visual words with the weights calculated by multiplying their inversed distances and the factor of $8/2^{i-1}$, where i is the index of the words from 1 to 4, at increasing distance range.

At this point, each image is converted into a list of visual words with their weights at that image derived from distances between features and visual words accordingly.

IV. SPATIAL PYRAMID OF LATENT TOPICS BY CASCADED PLSA

The basic idea in pLSA, which has been first used in text analysis, is to map high dimensional count vectors of documents to a reduced dimensional representation, so-called "Latent Semantic Space" [7]. For a set of documents d_i, $i=1...N$, and words w_j, $j=1...W$, given the conditional probability distributions of words over documents, pLSA generates unobserved topic variables z_k, $k=1...K$. We can formulate the probability of an observation pair as $P(w,d)=P(d,w)=P(d)P(w|d)$ in Naïve Bayesian approach, thus we get the conditional probability distribution of words over documents as:

$$P(w \mid d)=\sum_{k=1}^{K} P(w \mid z_k)P(z_k \mid d) \qquad (1)$$

In order to compare documents in latent space, however, we must work this process in reverse; starting with a term-document matrix, we generate a table which represents $P(z|d)$ in order to describe each document as a mixture of topics.

Our aim in this study is to discover latent topics in sub-regions at different resolution levels for associating spatial layout information. We first construct a sequence of overlapping grids at resolution levels, L, where we have $(2^{L+1}-1)^2$ girds at each level. In the training stage, the topic conditional visual word distributions $P(w|z)$ found at resolution level 0, where the image is intact, is utilized in other resolution levels.

We use this assumption at pLSA modeling and split the training images into nine half-size overlapping sub-regions at $L=1$. That will generate a new word distribution for each sub-region as an input to pLSA by using the same visual vocabulary. The topic distribution over each sub-region, $P(z|d)$, is then computed individually. Specifically, each sub-region is projected onto the triple axis space (Visual word-Image-Topic) by the $P(w|z)$ learnt in $L=0$. This is achieved by updating only the topic distributions vector $P(z|d)$ in each M-step while the learnt $P(w|z)$ kept fixed at Expectation Maximization (EM) iterations of pLSA. We follow the same procedure in finer resolution levels. The process is repeated at the higher levels of the spatial pyramid.

After calculating the topic distributions of each sub-region at each resolution level, we concatenate them all with a proper weight factor $1/2^{L-l}$. This formulation is inversely proportional to the sub-region width at that level, means a finer resolution level is weighted more highly than a coarser level. An illustration of scene classification based on Cascaded pLSA method is illustrated at Fig. 2.

Note that the cascaded pLSA differs from the previous pLSA based approaches utilizing spatial layout information. The ABSolute position pLSA (ABS-pLSA) and its improved version Spatial Pyramid pLSA (SP-pLSA) [5] apply a single pLSA after concatenating the BoW representations calculated from sub-regions. This is equal to employing a different pLSA at each sub-region, which generates a separate topic space for that sub-region. However, we fix the topic space all over the image by using the same word distribution over topic. Therefore, we try to locate the same set of topics at the spatial pyramid. In addition, the overlapping sub-regions smoothes the location changes in the images of the same category.

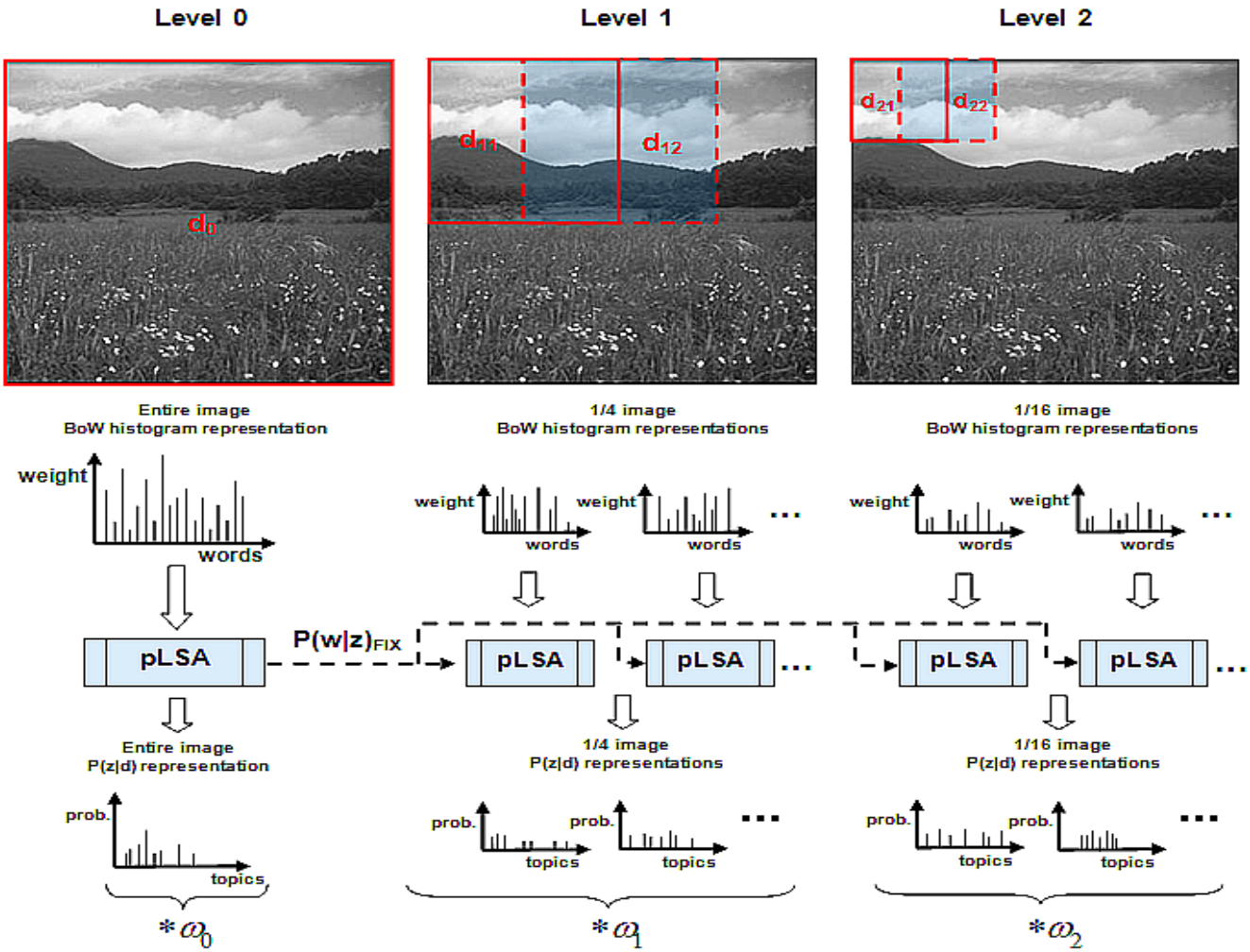

Figure 2. Illustration of Cascaded pLSA based method. *BoW* = Bag-of-Words histogram, *P(z|d)* = image specific probability distribution over latent variable space, *P(w|z)* = probability of a specific word conditioned on a latent variable.

The representation of an unknown image is similar to the training process. The pLSA model fitting calculates the topic distributions, $P(z|d_{test})$, by keeping the word distributions over topic, $P(w|z)$ which has been already computed at $L=0$ in training images. After extracting the cumulative feature vector of topic mixture coefficients for each image, the categorization of an unobserved image is performed in this semantic representation by using discriminative learning algorithms.

V. PERFORMANCE EVALUATION

The dataset on which we evaluate our classification algorithm contains 4.486 images from 15 scene categories. This dataset is first used in [2] and widely referenced by other scene classification works, available on [11]. Most of the scenes in the dataset display large intra-class variability, meaning that object contents within a scene category are different. Also note that indoor scenes (i.e. kitchen, bedroom, office living-room and store) have similar structures, indicating low inter-class variability. Besides, variances in visual domain such as lighting condition, viewpoint, scale, and perception ambiguities make the scene classification problem very hard while working with this dataset. As the same setup of [2, 5], we select 100 images per class randomly for the training set which consists of 1500 images totally. The rest of whole dataset is used as the testing set with varying number of images from each category. In the preprocessing stage, we first resize each image into 256x256 pixels and normalize it to have intensity with zero mean and unit standard deviation. After extracting dense SIFT-like descriptors, we select randomly 40 images from each scene category, and input their local descriptors into k-means clustering algorithm. The vocabulary size and the number of topics are set as $W=400$ and $T=100$, respectively. In the classification stage, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) algorithms are investigated. SVM is trained using the one-versus-all rule with Radial Basis Function (RBF) kernel. The classification results displayed at

Table-1 shows the performance rates at single and multiple levels of the pyramid similar to the format used in [2].

As it is noted in [2], the spatial information improves the performance when we divide an image into finer sub-regions. In Spatial Pyramid of BoWs at Table-1, the classification accuracy is 76.10 at *L=0* while at the highest pyramid level it increases to 81.44. As speaking for comparison of KNN vs. SVM, the performance of SVM is higher (i.e. about 7 percent at *L=0*) than that of KNN. The performance improvement provided by the proposed local features and soft-weighting BoW representation is evaluated approximately 3% for this particular dataset. In addition, the cascaded pLSA approach improves the classification rates about 2%.

TABLE I. CLASSIFICATION RESULTS FOR SPATIAL PYRAMID OF BOWS DESCRIBED IN [2] AND CASCADED PLSA

	Spatial Pyramid of BoWs				**Spatial Pyramid of Cascaded pLSA**			
	W = 400				**T = 100**			
	Single Level (%)		**Pym. (%)**		**Single Level (%)**		**Pym. (%)**	
Level	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**
0 (1x1)	69.54	**76.10**	---	---	69.12	76.39	---	---
1 (3x3)	72.04	79.42	73.55	79.46	72.73	80.68	73.37	81.83
2 (7x7)	72.33	80.49	73.85	**81.44**	73.92	82.12	75.03	**83.31**

Eventually, we compare the performance of our method at Table-2 to Spatial Pyramid Matching (SPM) [2] and Spatial Pyramid pLSA (SP-pLSA) [5] using the same experimental set-up. We implement their algorithms as described in their papers except using their own SVM toolbox. It is shown that our method outperforms the other methods. Note that the performances of SPM and SP-pLSA differ from the ones reported in the papers [2] and [5]. We think that it is due to the resizing all the images into 256x256 and the different implementation of SVM. Similar performance decrement is also evaluated in [10] by cropping the images to be square with 256x256 pixels.

TABLE II. COMPARISON OF OUR ALGORITHM WITH OTHER METHODS USING THE SAME EXPERIMENTAL SETUP

SPM of [2]	**SP-pLSA of [5]**	**Cascaded pLSA**
77.76	79.17	**83.31**

VI. CONCLUSION

In this paper, we focus on scene classification problem with a new image representation scheme called Cascaded pLSA. To achieve more discriminative features, we extract varying-grid SIFT descriptors at concentric support regions densely and concatenate them directly. Thereafter, feature descriptors are clustered into visual words with a soft weighting scheme. We associate location information with the conventional BoW/pLSA algorithm, where the spatial information is actually lost. This is accomplished by dividing each image into half-size overlapping regions iteratively at different grid levels and implementing a pLSA model for each region individually. We claim that overlapping regions enhance the performance in pLSA modeling as in densely feature extraction stage. Hence each region produces its own mixture of topics while staying coherent to the whole image where they belong to; since word-topic distributions of the whole image is kept fixed. Eventually, we concatenate these sub-region specific topic distributions with a weighting scheme to obtain a new semantic image representation.

REFERENCES

[1] Fei-Fei L. and Perona P. "A bayesian hierarchical model for learning natural scene categories," Proc. IEEE CVPR, 2005, pp. 524-531.

[2] Lazebnik S., Schmid C. and Ponce J. "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," Proc. IEEE CVPR, vol. 2, 2006, pp. 2169-2178.

[3] Szummer M. and Picard R W. "Indoor-outdoor image classification," Proc. Int'l Workshop on Content-Based Access of Image and Video Databases, 1998.

[4] Lowe D. "Distinctive image features from scale invariant keypoints," Int'l J. Computer Vision, vol. 60, no. 2, 2004, pp. 91-110.

[5] Bosch A., Zisserman A. and Munoz X. "Scene classification using a hybrid generative/discriminative approach," IEEE Transections on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, 2008.

[6] Jiang J., Ngo C.W. and Yang J. "Towards optimal bag-of-features for object categorization and semantic video retrieval," ACM Int'l CIVR, 2007.

[7] Hofmann T. "Unsupervised learning by probabilistic latent semantic analysis," Machine Learning, vol. 41, no. 2, 2001, pp. 177-196.

[8] Zhang J., Marszalek M., Lazebnik S. and Schmid C. "Local features and kernels for classification of texture and object categories: an in-depth study," INRIA Technical Report RR-5737, 2005.

[9] Nowak E. "Sampling strategies for bag-of-features image classification," ECCV, 2006.

[10] A. Quattoni and A.Torralba. "Recognizing indoor scenes," IEEE CVPR, 2009, pp. 413-420.

[11] http://www-cvr.ai.uiuc.edu/ponce_grp/data/