Springer Journal of Signal, Image and Video Processing manuscript No.
(will be inserted by the editor)

Scene Recognition by Semantic Visual Words

Elahe Farahzadeh · Cham Tat-jen · Andrzej Sluzek

Received: date / Accepted: date

Abstract In this paper we propose a novel approach to introduce semantic relations into the bag-of-words framework. We use the latent semantic models, such as LSA and pLSA, in order to define semantically-rich features and embed the visual features into a semantic space. The semantic features used in LSA technique are derived from the low-rank approximation of word-image occurrence matrix by SVD. Similarly, by using the pLSA approach, the topic-specific distributions of words can be considered dimensions of a concept space. In the proposed space, the distances between words represent the semantic distances which are used for constructing a discriminative and semantically meaningful vocabulary. Position information significantly improves scene recognition accuracy.Inspired by this, in this paper we bring position information into the proposed semantic vocabulary frameworks. We have tested our approach on the 15-Scene and 67-MIT Indoor datasets and have achieved very promising results.

Keywords scene recognition · semantic vocabulary · visual words

1 Introduction

The *bag-of-words* (BOW) framework has been shown to be useful in various computer vision applications like object recognition [3], scene recognition [6]. The framework builds a visual vocabulary by vector quantization of raw features extracted from

E. Farahzadeh
Center of Computational Intelligence, School Of Computer Engineering, Nanyang Technological University, Singapore 639798 E-mail: elah0001@ntu.edu.sg

T-J. Cham
Center For Multimedia And Network Technology, School Of Computer Engineering, Nanyang Technological University, Singapore 639798 E-mail: astjcham@ntu.edu.sg

A. Sluzek
Department of Electrical and Computer Engineering, Khalifa University of Science Technology and Research, Abu Dhabi, UAE 127788 E-mail: andrzej.sluzek@kustar.ac.ae

local image patches. The vector quantization essentially involves clustering of the raw features by k-means and choosing a cluster's mean as the codebook or visual word. However, an important drawback of k-means clustering is that it is based on the appearance of the image or video as represented in the raw features, as opposed to being based on the semantic relations between features. Utilizing the semantics inherent in visual content improves image/video categorization and understanding.

There have been several attempts to incorporate semantics into the BOW model so that a more discriminative visual vocabulary is realized. Generative methods use latent variable models like Probabilistic Latent Semantic Analysis (pLSA) [33, 22] and Latent Dirichlet Allocation (LDA) [6, 22] to obtain models for each category and subsequently to fit the query to one of the models in an unsupervised manner. Although these methods are efficient, their unsupervised nature limits their performance. Moreover, the number of topics in these methods is equal to the number of categories. This too limits their efficiency. Discriminative methods which incorporate label information have also been explored. Among the recent methods is the notable work of Liu and Shah which finds a semantic visual vocabulary via Maximization of Mutual Information (MMI) between visual words and images [17] or videos [18]. The algorithm starts with singleton clusters and in each iteration, merges two clusters which result in the minimum loss in mutual information. This procedure continues until a certain threshold in the information loss or in the number of clusters is achieved. This approach is effective in discovering the optimum number of clusters, but the formed clusters do not necessarily represent topics or synonym words which is required for constructing discriminative histograms.

Liu et al. [19] use Diffusion Map (DM) to construct a semantic visual vocabulary. Unlike geodesic distance which is based on the shortest path between points, diffusion distance considers all paths between two points to measure the shortest distance, and, hence, is not sensitive to noise. However, considering connectivity in measuring the semantic distance is not appropriate in the presence of polysemy.[1] For example assume that word B is a polyseme with two distinct meanings: 1 & 2. If word B is connected to word A based on meaning 1 (they both have the same meaning 1) and also word B is connected to word C based on meaning 2, then words A and C will be connected in the diffusion distance framework, but they convey different meanings. So, diffusion distance does not always represent semantic distance.

Considering these drawbacks, we propose a method for scene recognition based on a semantic visual vocabulary that uses latent aspect models to embed visual words into a rich semantic space which we call the concept space. Using LSA (Latent Semantic Analysis) or its probabilistic version pLSA, the synonym words which convey the same meanings are embedded close to each other so that they can be clustered together into the same semantic cluster. The distance in the proposed concept space is actually based on the meanings of the words and, thus, it represents the semantic relations. Consequently, the formed histograms based on these semantic clusters are efficient and discriminative for scene recognition.

[1] Polysemy is the existence of words which convey different concepts in different images. For instance in text domain, the word *table* can either be interpreted as *a piece of furniture* or *an arrangement of data*.

In contrast with generative methods that do not make use of category labels, our method trains a classifier using the histograms from the training set. Moreover, in our method the number of topics can be changed as opposed to the unsupervised framework where this number is fixed and equal to the number of classes. This will allow us to analyze the semantic relations in more details and to consider as many topics as appropriate. On the other hand, pLSA is able to handle polysemy which is very effective in cases when different categories share the same topics e.g. *office*, *livingroom* and *kitchen* categories include similar visual tokens such as chairs and tables.

2 Related Works

In this section, we review some methods that attempt to introduce semantic relations in the BOW framework for object and scene recognition. These approaches can be broadly divided into generative and discriminative methods. Generative methods usually involve hidden variables. These methods [33, 6] try to model each image as a mixture of hidden concepts using either pLSA or LDA. On the other hand, discriminative methods are only based on observed variables. These approaches usually incorporate a classifier. Among these methods, Vogel and Schiele [37] define a set of concept classes (visual words) like *sand*, *sky* and *sea* to label image regions. In this method, image regions are represented by a combination of color and texture features and classified into concept classes. Thus for each image, a concept occurrence vector is constructed and classified for scene retrieval. In labeling the databases containing ambiguous images, this approach claims that obtaining the ground truth for local semantic concepts is easier than for the whole image. However, this approach suffers from the large amount of manual work needed to annotate local regions. Randomized clustering forests have been used by Moosmann et al. [21] for image classification. Ensembles of decision trees are constructed based on the image class labels. Subsequently, visual words are assigned to each leaf. After building the trees, a bottom-up pruning process is done to reach a threshold of number of leaves to control the codebook size. They use randomized forests for clustering and quantization. Randomized forests have also been used in [10] for object classification and segmentation. In their work decision trees are used directly on image pixels to save time for extracting descriptors. In contrast to [21], which uses forests for clustering only, they use forests for both clustering and classification purposes. Randomized clustering forest is fast, yet discriminative, when compared to conventional k-means clustering, However, it tends to overfit, especially when applied in noisy situations. Moreover the model is complex and it is hard to understand the relation between the predictor variables.

Quelhas et al. [30] have used the pLSA model to extract image-specific distributions of topics in order to represent images. This is followed by a SVM classifier in order to classify scenes. Bosch et al. [1] have also used a similar framework. Liu et al. [19] use diffusion distance to build a semantic dictionary. They construct a graph on the visual words in which the weights between points reflect the similarity. Visual words are represented by pointwise mutual information. By applying the diffusion

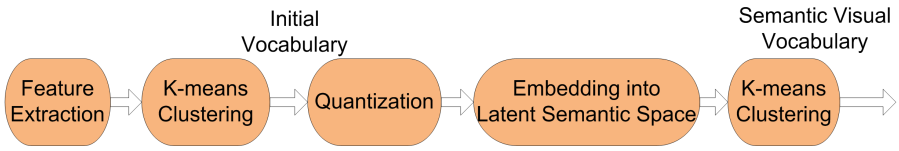


Fig. 1 Constructing the semantic visual vocabulary.

map, points are embedded into a lower dimensional space in which Euclidean distance is equal to diffusion distance.

The approach proposed in this paper is a discriminative embedding method that projects visual words into a concept space where the dimensions are discovered in an unsupervised manner by latent topic models.

3 Overview of the Proposed Framework

In the proposed framework after projecting images into the histograms of visual words, these histograms are embedded into less dimensional space using LSA/pLSA. Our main contribution is co-clustering the visual words which are semantically closed together at this new space to build semantic vocabulary. Afterwards we propose two approaches to build spatial pyramids over the semantic visual words. Figure 1 shows the flowchart for constructing the semantic visual vocabulary via embedding into concept space. We first extract features from patches in the images. The initial vocabulary is constructed by performing k-means clustering on the extracted features and choosing the cluster centers as the codewords. The feature vectors are quantized based on the initial codebook to form the word-image matrix which describes the occurrences of words in images. The codewords are then embedded into the concept space by latent semantic models - we demonstrate the embedding both by LSA as well as by pLSA. Finally, the embedded codewords in the concept space are again clustered using k-means to obtain the desired semantic visual vocabulary.

The contributions of this paper are threefold:

Using word space All of the methods that project images into latent semantic space by pLSA work in new *semantic document space*. This means that, in order to classify a test image, the image is projected into the semantic document space. However, the focus in our framework is the *semantic word space*. The similar visual words in the word space are co-clustered together to form a semantic visual vocabulary. In our framework, in order to classify a test image, it is directly quantized based on the semantic visual vocabulary without embedding the image.

Investigating the changes in the number of topics In the generative frameworks employing pLSA, the number of topics is considered the same as the number of categories. This is in contrast to our method which analyzes the changes in the number of topics and empirically fixes the best fit for representing the semantic relations in the scene.

Using LSA embedding Our framework applies both LSA and pLSA semantic embedding. To the best of our knowledge, there is no other method that uses LSA as

semantic embedding although, according to our experiments, the results of LSA and pLSA are almost on par. However, the time and memory complexity for LSA are substantially less since it uses a simple Singular Value Decomposition (SVD) compared to pLSA which uses an expensive expectation maximization algorithm. This is the reason that using LSA is really advantageous.

4 Concept Space

We obtain the initial vocabulary by performing k-means clustering on the extracted visual features. This initial codebook forms a reasonable-sized set representing all features, but they are not semantically clustered, i.e. the features in a cluster may convey different concepts. Thus, the formed histograms will not be semantically discriminative for classification. Therefore, we need a space in which semantically related words are adjacent. In order to find such a space, we use latent semantic models that find the underlying latent semantics given the occurrence matrix of word-image. These models are the well known LSA and pLSA, which we briefly review in the following sections. We use *tf-idf* instead of the normal count in the occurrence matrix for a higher efficiency.

4.1 Embedding into Concept Space using Latent Semantic Analysis

LSA [4] finds a low-rank approximation for the word-image matrix. The word-image matrix itself delivers semantics since synonym words appear in similar images resulting in similarities among their occurrence vectors. However, the original word-image matrix is noisy, sparse and large. Hence, the low-rank approximation to the original matrix is desirable. The consequence of this dimension reduction is that the dimensions relating to synonym words (e.g. *see* and *look* in text domain) are merged. It means the LSA hypothesis is that the synonym terms will have the same direction in the latent semantic space.

Let X be the occurrence matrix whose rows correspond to words and columns correspond to images. Decomposing of X using SVD, i.e.

$$X = U\Sigma V^T \tag{1}$$

gives the orthogonal matrices U and V and the diagonal matrix Σ that contains the singular values of X. By Selecting the L largest singular values and their corresponding singular vectors, we find the rank-L approximation of X by:

$$X \approx U_L \Sigma_L {V_L}^T. \tag{2}$$

This L-rank is optimal in a sense of l_2 matrix norm. The column vectors of U_L span the concept space of words and the columns of V_L span the concept space of images, so we can consider $M \times L$ matrix U_L the word space and $L \times N$ matrix V_L the concept space. The SVD decomposition process is illustrated in Figure 2. The i^{th} row of X, t_i, describes the i^{th} word. Consequently, the i^{th} row of U_L is the description of the i^{th} word in the concept space with L concepts and we refer to it as $\hat{t}_i$. In fact, each of

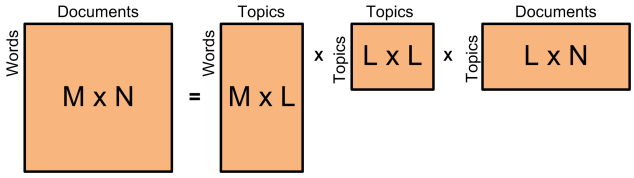


Fig. 2 SVD decomposition, word space and concept space.

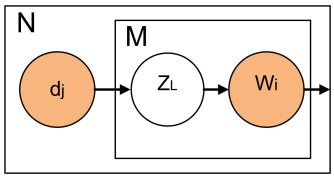


Fig. 3 Graphical view of pLSA model.

the L dimensions in the low dimensional vector $\hat{t}_i$ shows the projection of the words along one of the concepts. It is expected that synonym words are close in the concept space.

4.2 Embedding into Concept Space using Probabilistic Latent Semantic Analysis

pLSA [9] is the statistical version of LSA which defines a generative model on the data. It is assumed that there is a latent topic variable z_l associated with occurrence of the word w_i in the image d_j. It is expected that the joint probability $P(w_i, d_j, z_l)$ follow the form of the graphical model in Figure 3. The observed variables are w_i and d_j while z_l is latent. The probability of observation pair $P(w_i, d_j)$ is:

$$P(w_i, d_j) = P(w_i|d_j)P(d_j). \tag{3}$$

Since the occurrences of w_i and d_j are assumed to be independent, we can marginalize over latent topics z_l in order to find the conditional probability $P(w_i|d_j)$, i.e.

$$P(w_i|d_j) = \sum_{l=1}^{L} P(w_i|z_l)P(z_l|d_j), \tag{4}$$

where $P(z_l|d_j)$ is the probability of occurrence of topic z_l in the document d_j and $P(w_i|z_l)$ is the probability of occurrence of word w_i given the topic z_l. L is the total number of latent topics. Equation 4 is a decomposition of the word-document matrix, similar to LSA, but with the condition that the values are normalized to be probability distributions. We fit the model by determining $P(z_l|d_j)$ and $P(w_i|z_l)$ given the observation occurrence matrix. Maximum likelihood estimation of the parameters is performed using Expectation Maximization (EM) algorithm. Assuming a vocabulary of M words and N documents, the likelihood function to be maximized is:

$$\prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}, \tag{5}$$

where $n(w_i, d_j)$ is the number of words w_i in the document d_j and $P(w_i|d_j)$ is obtained by Equation 4.
The original pLSA algorithm in the unsupervised learning framework tries to categorize the query document given the learned parameters [9]. However, we use the pLSA algorithm only to determine the probabilities $P(w_i|z_l)$. In fact $P(w_i|z_l)$ is equivalent to the l^{th} dimension of $\hat{t}_i$ in the LSA framework. Therefore, using pLSA we obtain the concept space embedded vector $\hat{t}_i$ as:

$$\hat{t}_i = \left[\, p(w_i|z_1)\ p(w_i|z_2)\ \ldots\ p(w_i|z_L) \,\right]^T . \tag{6}$$

It should be noted here that L, the dimension of the concept space, does not need to be equal to the number of classes; this enables us to define arbitrary latent concepts. In fact the number of semantic topics can be much more than the number of classes. In other words, classes are wider concepts that may include some finer and more detailed concepts which are referred to as topics. For instance "computer" and "pen" are topics related to the class of office. Note that in the unsupervised framework in which pLSA is used (e.g, in [33]), the dimension of the concept space must be equal to the number of classes.

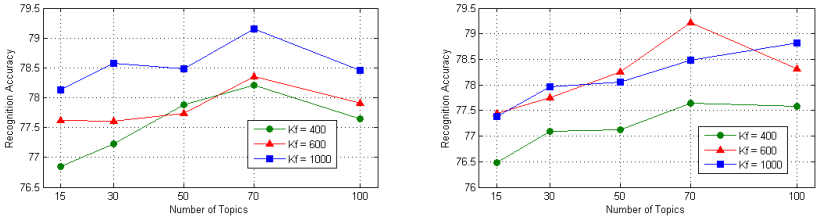
In LSA, each word is projected into a single point in the concept space so that each word can refer to a single meaning only. Instead pLSA is able to capture polysemy. Thus, given a word w observed in two different documents d_i and d_j, the topics associated with the word in d_i and d_j can be different or, in other words, $argmaxp(z|d_i, w)$ can be different from $argmaxp(z|d_j, w)$ [9], [16]. The advantage of LSA compared to pLSA is the faster and easier implementation. LSA needs a simple SVD, while pLSA uses the iterative EM algorithm which is only guaranteed to find a local maximum of the likelihood function [9].

5 Experimental Results

To demonstrate the efficiency of our method, we have evaluated our method on two challenging scene datasets: the *15-Scene [6, 13]* dataset and the *MIT 67-Indoor Scenes [29]* .

The features applied in these experiments use the patch of 16×16 size sampled densely with the M=8 pixel spacing. The feature descriptor applied on each patch is SIFT [20]. For 15-Scene dataset we randomly select 100 images per category as training images and the rest for testing. The results are averaged over five times random splitting of training and testing images. For MIT 67-Indoor Scenes we used the exact same partitions as used in [29] which contain 80 images for training and 20 images for testing, so that all results are directly comparable.

Support Vector Machine (SVM) with Histogram Intersection kernel (HIK) is used as the classifier, while the size of the initial vocabulary (K_i) is fixed to 1500 during the experiments.



(a) LSA (b) pLSA

Fig. 4 Performance of proposed method on 15 scene dataset with different number of topics using LSA/pLSA.

5.1 Results and Analysis

One of the advantages of the proposed method is that it allows the number of topics to be varied, in contrast to pLSA using unsupervised framework (where the number of topics is constrained to be the same as the number of classes). Figure 4(a) and Figure 4(b) show how the number of topics L affects the recognition accuracy with LSA and pLSA as the embedding method. The experiments have been performed using three different semantic vocabulary sizes, K_f. As the number of topics is increased above $L = 15$ (which is the number of classes) the recognition rate increases since the increased number of topics enables a better discrimination between classes. However, after around $L = 70$ topics, the recognition accuracy decreases. This is mainly because adding more dimensions to the concept space implies further division into semantic units that are not meaningful. The recognition accuracy has a variance of about $1\% - 3\%$ as L varies.

According to the experiments, the number of topics for obtaining the best accuracy for the three different values of K_f is the same. Thus, the number of topics is independent of the final vocabulary size.

To verify the efficiency and discriminative ability of the method in scene classification, we have compared it with the classic BoW framework for different vocabulary sizes. The results are shown in Figure 5. The number of topics is chosen to be 70. According to the figure, our method outperforms the classic BOW framework in all cases. This shows the efficiency of the method proposed. Apart from the instabilities in the initial parts of the curves, we can say that the behaviors of LSA and pLSA are consistent. For small vocabulary sizes, pLSA outperforms LSA by a small margin due to its ability to handle polysemy. However, as vocabulary size increases, LSA performs better than pLSA (at approximately $Kf = 900$). Based on the differences between pLSA and LSA embedding, the possible cause of this effect is that the larger vocabulary size brings in more details and compensates for the effect of polysemy. However, pLSA takes into account every possible meaning of a word, even the rare ones, which results in confusion in larger vocabularies, thus reducing the accuracy. Also it should be noted that LSA has always a shorter implementation time compared to pLSA. This is due to a time-consuming iterative EM process for pLSA compared to the straightforward SVD in LSA.

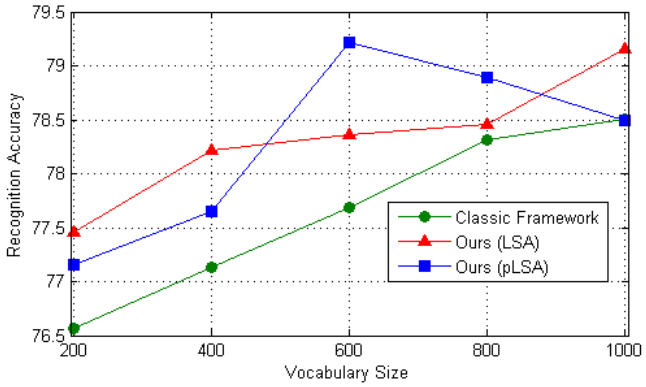


Fig. 5 Comparison of results with the classic framework for different sizes of vocabulary in 15 scene dataset.

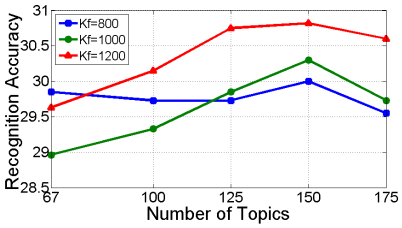


Fig. 6 Performance of proposed method on 67-MIT Indoor dataset using LSA concept space by changing the number of topics.

The best result achieved on *15-Scene* dataset with our method is **79.22** using pLSA model and the semantic codebook size of 600.

Figure 6 shows how the number of topics L affects the recognition accuracy with LSA as the embedding method in MIT 67-Indoor Scenes. According to the experiments, the number of topics for obtaining the best accuracy for all three different values of K_f is the same. As the number of topics is increased above $L = 67$ the recognition rate increases since the increased number of topics enables a better discrimination between classes. However, after around $L = 150$ topics, the recognition accuracy decreases. As mentioned before this is because adding more dimensions to the concept space implies further division into semantic units that are not meaningful.

Figure 7 shows the efficiency of our framework using either pLSA or LSA embedding in comparison with classic bag-of-words.

To confirm the advantage of the semantic visual vocabulary compared to the original visual vocabulary, in Table 1 we present the results of applying both vocabularies to four different scene images. For clearer illustration, we have only shown a subset (40) of the visual words. Each color is associated with one visual word. As seen in the tables third column, which is associated with the semantic visual vocabulary, the areas are more uniform, especially in the marked regions by the rectangles in the scene image.

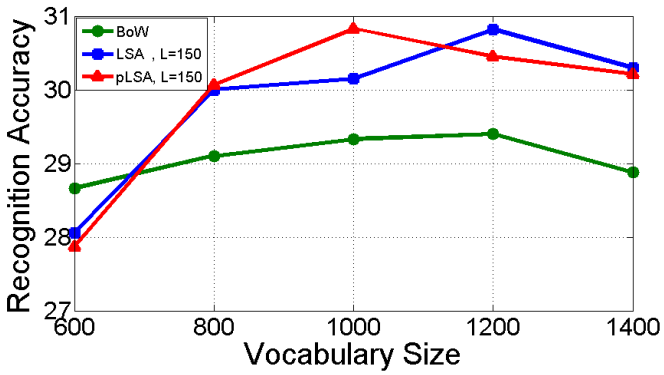


Fig. 7 Comparison of results with the classic framework for different sizes of vocabulary in 67-MIT Indoor dataset.

6 Capturing Image Spatial Information in Scene Recognition Systems

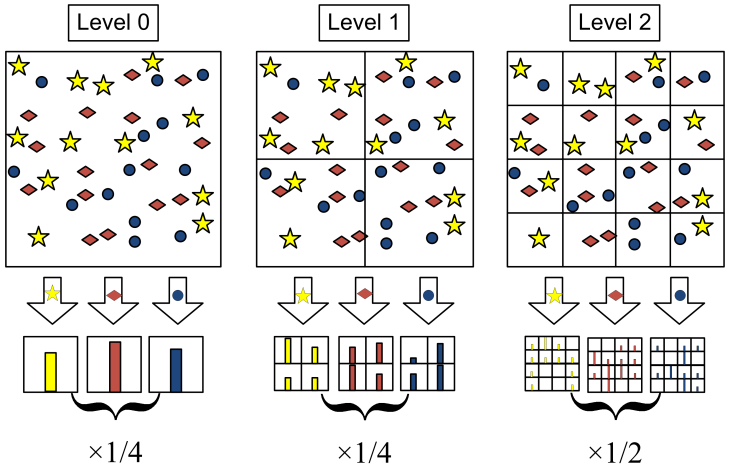
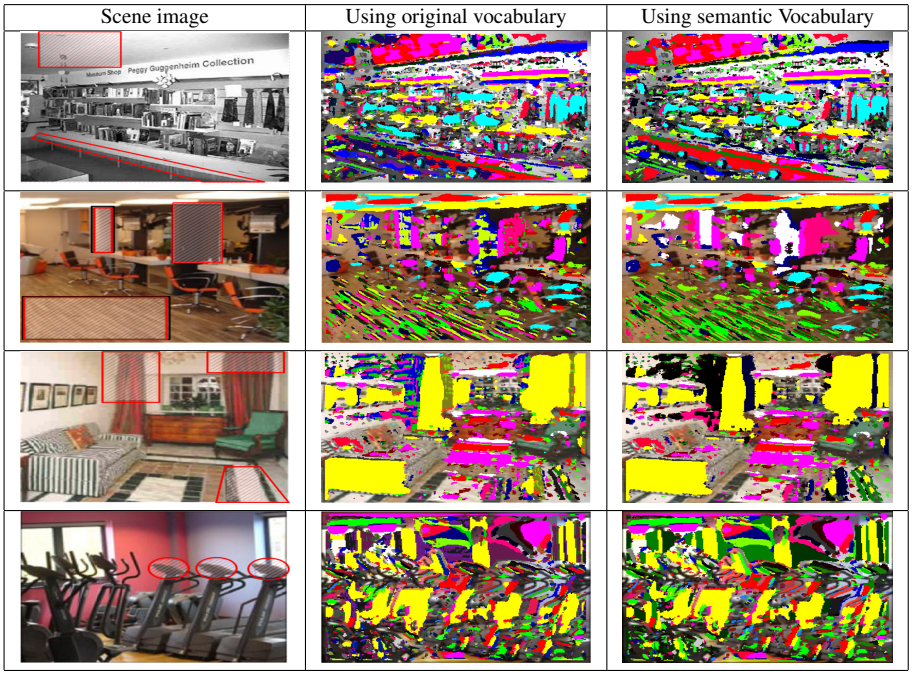
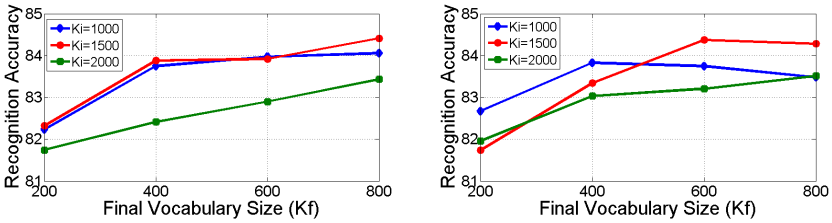


Fig. 8 Example of constructing a three-level pyramid.

In the proposed methods in Section 3, the spatial information was not being considered. In this section we try to make use of spatial information to increase the efficiency. Spatial arrangements of visual features have a significant impact on image classification systems. Lazebnik et al. [13] developed a spatial version of the Pyramid Match Kernel (PMK) [7] to overcome the lack of spatial information in the *bag-of-words* framework. The Pyramid Match Kernel was proposed by Grauman and Darrell [7] on feature space to find the correspondence between feature sets (such as bag-of-words feature vectors) using a hierarchical quantization. Lazebnik et al. [13] have used Spatial Pyramid Matching on the image space, i.e. instead of dividing the

Table 1 Semantic visual words vs. Original visual words. The first column shows the scene images, the second column illustrates the results of applying the original visual vocabulary to the scene image and the third column illustrates the results of quantizing dense SIFT visual features based on the semantic visual words. Each visual word is illustrated with one color. Less colors in the marked areas shows that visual words in the semantic areas are clustered together.





(a) LSA (b) pLSA

Fig. 9 Performance of proposed global method using LSA/pLSA.

feature space into hierarchical levels, the pyramid levels are built over the image sub-regions. Spatial Pyramid Matching successfully accomplishes the goal of incorporating spatial information into the *bag-of-words* framework. There are frameworks which incorporate spatial pyramid structure and better classification accuracy was reported [2, 38, 11].

In this section, we impose spatial information on the proposed framework in Section 3. We propose two methods (*global* and *region-based*) to capture location information when using a semantic vocabulary. Both methods build spatial pyramids over

the image's blocks. The methods differ in that the *Global* method initially forms the semantic vocabulary and then divides the image into spatial sub-regions, while the *Region-based* method forms the semantic vocabulary over each region individually.

6.1 Capturing Spatial Information with Spatial Pyramid Matching (SPM)

The spatial pyramid matching technique is a simple yet efficient framework. SPM gives a multi-resolution representation of the image by dividing it into increasingly finer sub-regions. It was first proposed by Lazebnik et al. [13] and inspired by the Pyramid Match Kernel (PMK). PMK defines different levels of resolution in the feature space by defining increasingly coarser grids [7]. Although PMK is very precise, it ignores the spatial location of individual features. SPM has been successfully used as an extension of the *bag-of-words* framework. The bag-of-words framework gives an orderless representation of the image, while SPM uses a spatial pyramid representation of the image. The matching score of this pyramid representation is obtained by a weighted combination of histogram intersections at multi-spatial resolutions.

The image local feature f is denoted by $f = (x, y, d)$ where (x, y) shows the feature's location coordinates and d represents the feature descriptor. In the BoW framework, there are K discrete cluster centroids where each centroid represents one of the visual words. Feature f is quantized into one of these centroids with respects to its descriptor, but the location coordinates (x, y) are completely ignored.

On the other hand, in the SPM framework these location coordinates are used to enhance the power of BoW, e.g. in Figure 8, the image is divided into $2^l \times 2^l$ uniform sub-regions at level l of resolution with 2^l evenly sized partitions in each dimension (horizontal or vertical). At higher levels of resolution, the image is divided into finer regions. At resolution l $(0 \leqslant l \leqslant L)$, feature f is assigned to one of the $2^l \times 2^l$ sub-regions based on its location coordinates, while its descriptor is quantized into one of the K centroids.

Given the two-dimensional vectors X_k and Y_k as the set of coordinates for channel k, $H^l_{X_k}$ and $H^l_{Y_k}$ are our histograms at level l of resolution, where the dimension of this histogram is D. To find the number of matches at level l of channel k, the Histogram Intersection Kernel (HIK) is applied to the H_{l_k} histograms:

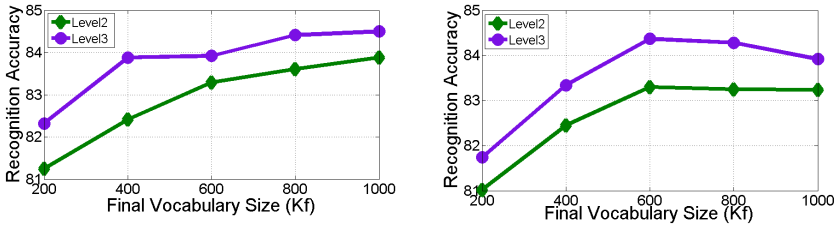
$$I(H^l_{X_k}, H^l_{Y_k}) = \sum_{i=1}^{D} min(H^l_{X_k}(i), H^l_{Y_k}(i)). \tag{7}$$

The matching score in channel m is measured by applying PMK:

$$K^L(X_k, Y_k) = \frac{1}{2^L} I^0 + \sum_{l=0}^{L} \frac{1}{2^{L-l+1}} I^l. \tag{8}$$

The pyramid matching formulation ($K^L(X_k, Y_k)$) shows that the weights are inversely proportional to the size of the sub-regions. The final match kernel is sum of the match scores from all of the K channels:

$$K^L(X, Y) = \sum_{1}^{K} k^L(X_k, Y_k). \tag{9}$$



(a) LSA (b) pLSA

Fig. 10 Recognition accuracy using pLSA and LSA concept space in different levels of resolution.

6.2 Spatial Content Capture using Global and Region-based Models

In this section, the spatial pyramid matching schema is incorporated into the new concept space proposed in Section 3. To perform the spatial content capture, we suggest two methods: global and region-based. In the *global* method, after projecting the word space into the concept space, the k-means algorithm is applied to co-cluster the synonymous words. The image is divided into increasingly finer blocks. Afterward, the visual features within each sub-region are quantized into histograms of bag-of-words based on the appearance-based vocabulary and then the histogram bins of the synonymous words are merged together.

The final spatial representation based on the semantic vocabulary is obtained after applying the spatial pyramid weighting for each level of resolution.

In the *region-based* method, there is a separate concept space for every spatial block of the image, i.e. the latent semantic models are applied to each sub-region after it is quantized into the bag-of-words histogram according to the appearance-based vocabulary. K-means clustering is applied to these new regional word-topic (concept) spaces and, consequently, the histogram bins for the synonymous words in the *bag-of-words* representation are bound together.

We argue that the region-based method is more effective than the global method because the spatial locations are considered while co-clustering the synonym words in the region-based method; this is important because semantic visual words differ according to the visual features located on that partition of the image.

Although the experiments show that the region-based method outperforms the global method, the improvements in the results are not very significant and it is due to the essence of visual word vocabularies. Since we have only one semantic vocabulary for the whole dataset while building this vocabulary, the location of the features is not important; rather, the most important issue is which part of the image we project by this vocabulary.

6.3 Experimental Evaluation on 15-Scene dataset

In these experiments the number of topics is fixed to 70 which, according to [31] results in the highest accuracy. In the first experimental series, we apply the global method to impose spatial information. As shown in Figures 9(a) and 9(b), the results

are very promising and the diagrams prove the efficiency of our method. In these figures we study the influence of 3 different initial vocabulary sizes (K_i) on the recognition accuracy while the final semantic vocabulary size (K_f) changes. In Figure 9(a), the LSA method is used as the embedding method. According to this figure, the best accuracy is achieved with $K_i = 1500$ and a final vocabulary size of $K_f = 800$. We use the pLSA model to project the word-document space into the concept space, the results are illustrated in Figure 9(b), and the best accuracy is achieved at $K_i = 1500$ and $K_f = 600$.

To implement the proposed global method, we apply pyramid matching on different levels of resolution. The results of the global method in level 1 and 2 of resolution for the LSA and pLSA models are demonstrated in Figure 10.

The time complexity of implementing region-based model by applying pLSA embedding is very high, therefore to perform the experiments for this model we just used LSA embedding. The experimental results for the region-based approach using LSA are demonstrated in Figure 11 while the initial vocabulary size is fixed to 1500, $K_i = 1500$. This figure shows the efficiency of the region-based method in using LSA. According to this plot diagram the accuracy increases until we reach $K_f = 800$, then it slightly decreases.

To compare the global method versus the region-based method using LSA model, the results of these methods are illustrated in Figures 11, with the initial vocabulary size fixed to $K_i = 1500$. According to this demonstration although the region-based method outperforms the global method, their results are almost on par. However, since it is necessary to project each spatial block separately into a new concept space in the region-based method, the time complexity is very high. Therefore, the difference in time complexity between the global and region-based methods justifies using the global method.

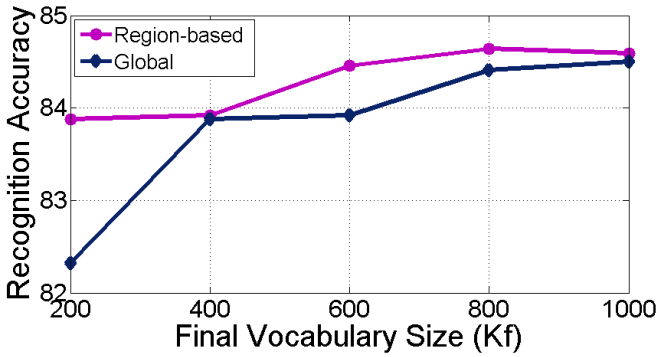
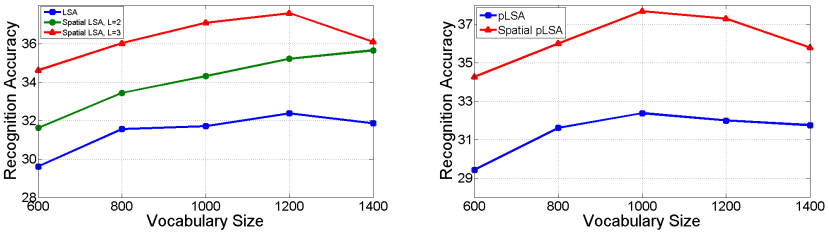


Fig. 11 Comparing the performance of proposed global method vs. region-based method using LSA.

Table 2 summarizes recognition accuracy of our method and some notable related works. As seen from the table, the proposed method is the best. The work of Fei-Fei and Perona [6] which has used LDA in a generative framework, has a lower performance compared to the methods using a semantic vocabulary but incorporating

Table 2 Comparison with recently reported results for 15-Scene.

Method	Accuracy(%)
Our method	**84.82**
Wu&Rehg [38]	83.88
Bosch et al. [2]	83.7
Lazebnik et al. [13]	81.4
Li et al. [14]	80.9
Saghafi et al. [31]	79.22
Parizi et al. [28]	78.6
Liu&Shah [17]	75.16
Liu et al. [19]	74.9
Oliva&Torralba [25]	74.10
Bosch et al. [1]	73.30
Quelhas et al. [30]	71.24
Fei-Fei&Perona [6]	65.2



(a) Spatial-LSA (b) Spatial-pLSA

Fig. 12 Performance of our method using Spatial-LSA, Spatial-pLSA concept space.

category labels like Liu (DM) [19] and Liu (MMI) [17]. Also the works of Quelhas et al. [30] and similarly Bosch et al. [1] have lower accuracy compared to the works using co-clustering to obtain semantic vocabulary like Liu (MMI) [18], Liu (DM) [19] and ours. This is mainly because in contrast to former methods, which use a histogram of topics equal to the number of categories, latter methods perform the clustering step in the semantic space to further group the semantically related words together and to construct more discriminative histograms with the actual number of topics.

6.4 Experimental Evaluation on 67-MIT indoor dataset

We have evaluated our global spatial method on 67-MIT indoor dataset in LSA/pLSA concept space. The result of this evaluation is illustrated in Figures 12(a), 12(b). As seen in these figures, recognition accuracy by applying spatial-semantic vocabulary is remarkably higher than the original semantic vocabulary in both LSA and pLSA concept spaces.

In Figure 13 the recognition accuracy for original Spatial Pyramid Matching method is compared to spatial-pLSA and spatial-LSA. In this evaluation the level of resolution is set to L=3 for the three methods.

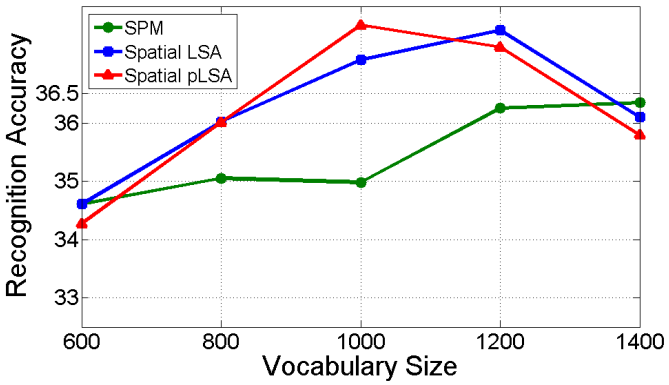


Fig. 13 Performance of different spatial content capture methods

Table 3 Comparison with recently reported results for 67-Indoor Scenes.

Method	Accuracy(%)
Our method	**37.68**
Singh et al. [32]	38.1
Parizi et al. [28]	37.93
Li et al. [14]	37.6
Wu&Rehg [38]	36.9
Lazebnik et al. [13]	34.4
Pandey&Lazebnik [27]	30.8
Quattoni&Torralba [29]	26.00
Quelhas et al. [30]	21.17
Oliva&Torralba [25]	22.0
Bosch et al. [1]	20

There are approaches that try to deal with scene recognition challenges by imposing high level concepts [14, 15, 29, 27, 28, 32, 8, 5]. In [14], the images are represented by a vocabulary of objects called *object-bank*, despite the high computational complexity, the method does not offer much increase in recognition accuracy. Li and Gua [15] improve scene recognition performance by capturing objects' co-occurrence and their geometric correlations. They build a three levels (super-pixel, object, scene) hierarchical model, the operation in all of the levels is performed automatically.

In [1] the images are considered as a mixture of semantic topics. If followed by SVM classification, the classification rate is lower than our method. Oliva and Torralba [25] and Wu and Rehg [38] use gist and CENTRIST global features respectively, bypassing object-centered and local information. Although the recognition rates in these holistic methods were higher than those for purely local methods, they were still lower than our reported results. Capturing the spatial location of local patches in [13] and [38] significantly improved the recognition accuracy for scene recognition. In [27] Pandey and Lazebnik used the popular Deformable Part-based Model(DPM) [26] for scene recognition and achieved the accuracy of **30.08%**. Subsequently they combined DPM results with color, gist and SIFT spatial pyramids in-

formation and achieved an accuracy of 43.1%. The discriminative mid-level patches in [32] achieved **38.1%** accuracy while combining these mid-level patches with color, gist, DPM and SIFT spatial pyramids they obtained **49.4%** recognition accuracy.

7 Conclusion

In this paper, we have proposed a novel approach for using semantic relations in BoW framework. We have used the latent aspect models such as LSA and pLSA to map the visual words into a semantic space. Under the LSA framework this mapping is done by a low-rank decomposition of the word-document occurrence matrix using SVD. Also, the topic-specific distributions of words are considered (using pLSA) as the projections words onto different concepts. The distances in the proposed concept space reveal the semantic relations. Clustering is done in the concept space to capture the semantic structures. Also our method performs better compare to some similar methods for constructing semantic vocabularies.

In this paper also a spatial method is proposed. The proposed spatial method incorporate the pyramid matching technique. In semantic vocabulary, pyramid levels are constructed directly over the latent semantic space using either one concept space for the entire scene or each image sub-scene separately projected into the concept space. The experimental evaluation shows remarkable improvements in both the 15-Scene and 67-Indoor Scenes datasets.

References

1. Bosch A, Zisserman A, Muñoz X (2006) Scene classification via pLSA. In: European Conference on Computer Vision (ECCV)
2. Bosch A, Zisserman A, Muoz X (2008) Scene classification using a hybrid Generative/Discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(4):712–727
3. Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: International Workshop on Statistical Learning in Computer Vision, ECCV
4. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6):391–407
5. Farahzadeh E, Cham TJ, Li W (2013) Incorporating local and global information using a novel distance function for scene recognition. In: IEEE Workshop on Robot Vision, Winter Vision Meetings (WVM)
6. Fei-Fei L, Perona P (2005) A Bayesian hierarchical model for learning natural scene categories. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
7. Grauman K, Darrell T (2005) The pyramid match kernel: Discriminative classification with sets of image features. In: IEEE International Conference on Computer Vision (ICCV)

8. Gupta S, Arbelaez P, Malik J (2013) Perceptual organization and recognition of indoor scenes from rgb-d images. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
9. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1-2):177–196
10. J Shotton RC M Johnson (2008) Semantic texton forests for image categorization and segmentation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
11. Kwitt R, Vasconcelos N, Rasiwasia N (2012) Scene recognition on the semantic manifold. In: European Conference on Computer Vision (ECCV)
12. Lazebnik S (2006) Local, semi-local and global models for texture, object and scene recognition. PhD thesis, University of Illinois at Urbana-Champaign
13. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
14. Li LJ, Su H, Xing EP, Fei-Fei L (2010) Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: Neural Information Processing Systems (NIPS)
15. Li X, Guo Y (2012) An object co-occurrence assisted hierarchical model for scene understanding. In: British Computer Vision Conference (BMVC)
16. Liu D, Chen T (2007) Unsupervised image categorization and object localization using topic models and correspondences between images. IEEE International Conference on Computer Vision (ICCV)
17. Liu J, Shah M (2007) Scene modeling using co-clustering. IEEE International Conference on Computer Vision (ICCV)
18. Liu J, Shah M (2008) Learning human action via information maximization. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
19. Liu J, Yang Y, Shah M (2009) Learning semantic visual vocabularies using diffusion distance. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
20. Lowe D (1999) Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision (ICCV)
21. Moosmann F, Triggs B, Jurie F (2006) Fast discriminative visual codebooks using randomized clustering forests. In: Neural Information Processing Systems (NIPS)
22. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79(3):299–318
23. Nowak E, Jurie F, Triggs B (2006) Sampling strategies for bag-of-features image classification. In: European Conference on Computer Vision (ECCV)
24. Odone F, Barla A, Verri A (2005) Building kernels from binary strings for image matching. IEEE Transactions on Image Processing 14(2):169–180
25. Oliva A, Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3):145–175

26. P F Felzenszwalb DM R B Girshick, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence
27. Pandey M, Lazebnik S (2011) Scene recognition and weakly supervised object localization with deformable part-based models. In: IEEE International Conference on Computer Vision (ICCV)
28. Parizi S, Oberlin J, Felzenszwalb P (2012) Reconfigurable models for scene recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
29. Quattoni A, Torralba A (2009) Indoor scene recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
30. Quelhas P, Monay F, Odobez Jm, Gatica-perez D, Tuytelaars T, Van Gool L (2005) Modeling scenes with local descriptors and latent aspects. In: IEEE International Conference on Computer Vision (ICCV)
31. Saghafi B, Farahzadeh E, Rajan D, Sluzek A (2010) Embedding visual words into concept space for action and scene recognition. In: British Machine Vision Conference (BMVC)
32. Singh S, Gupta A, Efros AA (2012) Unsupervised discovery of mid-level discriminative patches. In: European Conference on Computer Vision (ECCV)
33. Sivic J, Russell BC, Efros AA, Zisserman A, Freeman WT (2005) Discovering objects and their location in images. In: IEEE International Conference on Computer Vision (ICCV)
34. Swain MJ, Ballard DH (1991) Color indexing. International Journal of Computer Vision 7(1):11–32
35. Szummer M, Picard R (1998) Indoor-outdoor image classification. In: IEEE International Workshop on Content-Based Access of Image and Video Database
36. Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. International Journal of Computer Vision 62(1):61–81
37. Vogel J, Schiele B (2004) Natural scene retrieval based on a semantic modeling step. In: ACM International Conference on Image and Video Retrieval (CIVR)
38. Wu J, Rehg JM (2011) CENTRIST: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8):1489–1501