Improvement the Bag of Words Image Representation Using Spatial Information

Mohammad Mehdi Farhangi, Mohsen Soryani, and Mahmood Fathy

Department of Computer Engineering, Iran University of Science and Technology

Abstract. Bag of visual words (BOW) model is an effective way to represent images in order to classify and detect their contents. However, this type of representation suffers from the fact that, it does not contain any spatial information. In this paper we propose a novel image representation which adds two types of spatial information. The first type which is the spatial locations of the words in the image is added using the spatial pyramid matching approach. The second type is the spatial relation between words. To explore this information a binary tree structure which models the is-a relationships in the vocabulary is constructed from the visual words. This approach is a simple and computationally effective way for modeling the spatial relations of the visual words which shows improvement on the visual classification performance. We evaluated our method on visual classification of two known data sets, namely 15 natural scenes and Caltech-101.

Keywords: BOW Representation, Spatial Information, N-gram Model, Spatial Pyramid Matching.

1 Introduction

As the acquiring and storing of images and multimedia data is becoming fast and easy, the databases of these data become very large. In this situation the necessity for developing methods to manage these databases becomes more and more important. Classifying images based on their content is one of these methods that finds the category of an image among several categories. However, this task is a challenging problem in the real world because we encounter a number of difficulties in the images where there exists occlusion, background clutter and lighting changes.

Many of the recent methods for classifying the images represent each image as sets of patches or regions, described by various descriptors. Based on this description, an image can be represented as a bag of visual words [1]. To achieve this representation, the first step is the extraction of the local patches from the image. Several methods proposed to extract local patches in the literature. While some researchers obtained local regions using regular grids which segment images by horizontal and vertical lines [2], others used various interest point detectors such as difference of Gaussian [3], Harris affine region detector [4] and Hessian matrix [5] to detect patches that contain local information of an image. After detecting the patches, a feature descriptor method like SIFT [3], SURF [5], etc is used to describe them. Previous studies have

N. Meghanathan et al. (Eds.): Advances in Computing & Inform. Technology, AISC 177, pp. 681–690.
springerlink.com

shown that, the SIFT descriptor extracts robust features from the image which are invariant to affine transformations more than other descriptors [6]. After that, similar patches are clustered in the same groups and each of these groups is treated as a visual word. At this point, the vocabulary which consists of cluster centers is generated.

After the vocabulary construction, an image can be represented as a bag of visual words by assigning each local descriptor to one or several visual words with different weights [7]. Previous studies showed that by assigning the local descriptors to more than one word, the classification accuracy is increases [8].

Despite all of the successes in image classification based on the BOW, this type of representation does not consider the spatial information and this is because of the fact that the histogram representation naturally neglects the spatial location of visual words and spatial relations between them. One of the first attempts in order to utilize spatial information was proposed by Lazebnik et al. [9]. Their work was based on partitioning an image into increasingly finer grids. For each grid cell the frequency of visual words was computed. The BOWs from each cell were concatenated to each other and thus a representation of image which conveys the spatial location of visual words was obtained. In [10] a visual language model using training images was constructed. This model represents three kinds of relations including unigram, bigram and trigram between visual words and captures the proximity information of visual words. In [11] a new representation based on utilizing the informative adjacent word pairs were proposed. To find the informative word pairs, they measured the confidence that neighboring visual words are relevant. Visual words with high confidence were used to add to BOW representation.

In this paper we propose a new representation for images which adds the spatial information to bag of word representation. For this purpose we explore two types of spatial information. First, it is important to know where a certain visual word occurs in the image. For example a blue patch which is located above the image is probably representing a piece of sky while if this patch be in the bottom of the image, it may represent a part of a sea. Words adjacency is the second type of spatial information which although conveys important information about the content of the image, it is neglected in BOW model. For example a white patch can be part of a sheep, cloud or moon if it is surrounded by green grass, blue sky or dark area respectively. To consider this relation, we calculate number of times that each pair combination of words occurs in a certain neighborhood and construct the bag of N-grams inspired by Li et al. [10] and concatenate it to BOW representation.

The remaining sections of this paper are organized as follows. In section 2 we propose details of our method. Section 3 presents experimental results. And section 4 concludes the paper.

2 The Proposed Method

The new image representation which is called spatial bag of words (SPBOW) is constructed in two stages. In the first stage we use the spatial pyramid matching approach [9] and partition the image into fine sub regions and obtain the histogram of local features inside each sub regions. In the second stage the numbers of occurrences of visual word pairs are obtained and concatenated to the BOW representation as new features. The following subsections present these stages in details.

2.1 BOW Representation

In order to represent an image by bag of visual words, local patches are extracted from an image and every patch is described using SIFT descriptor. Since previous studies have shown that sampling on a regular grid outperforms other approaches like interest point detectors, we use the SIFT descriptor, sampled on a regular grid.

After that, each local descriptor of the image should be assigned to one or more visual words. If $\{r_1, r_2, \dots, r_n\}$ represents local descriptors in the image and $V = \{\omega_1, \omega_2, \dots, \omega_k\}$ represents the vocabulary, the hard histogram of visual words is computed as

$$\mathrm{HBOW}(\omega_j) = \sum_{i=1}^{n} \begin{cases} 1 & \text{if } \omega_j = \underset{\omega\in W}{\arg\min}\left(\mathrm{dist}(\omega_j, r_i)\right) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Where r_i is the i-th patch in the image and ω_j is the j-th word in the vocabulary.

2.2 Spatial Pyramid Matching

The BOW representation described above ignores some useful information of the image. For example in this representation there is no way to find out how many times a certain visual word takes place in a specific part of the image. To combine this information with BOW, we use the spatial pyramid matching proposed in [9] and partition the image into rectangular regions.

In details, pyramid matching works by placing a sequence of increasingly finer grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two points which fall into the same cell are matched. Matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. More specifically, a sequence of grids is constructed at resolutions 0… L, such that the grid at level l has 2^l cells along each dimension, for a total of $D = 2^{dl}$ cells. Let H_X^l and H_Y^l denote the histograms of X and Y at this resolution, so that $H_X^l(i)$ and $H_Y^l(i)$ are the numbers of points from X and Y that fall into the i-th cell of the grid. Then the histogram intersection function finds the number of matches at level l.

$$I(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min\left(H_X^l(i), H_Y^l(i)\right) \tag{2}$$

In this equation, the number of new matches found at level l is given by $I^l - I^{l+1}$ for $l = 0, \dots L - 1$. The weight associated with level l is set to $\frac{1}{2^{L-l}}$, which is inversely proportional to the cell width at that level. Intuitively, since the matches found in larger cells involve dissimilar features, they should be weighted lower. So the following definition was obtained for the pyramid match kernel:

$$\kappa^l(X, Y) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}\left(I^l - I^{l+1}\right) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{3}$$

In order to combine this pyramid matching kernel with spatial location of words in the image, the elements of X and Y are used for representing the coordinates of a certain word in the image. Therefore, by placing an increasingly fine grid on this feature space, the spatial information is combined with BOW representation.

2.3 Spatial Relation Modeling

Although the relations between visual words in an image convey essential information about its content, this information is neglected in traditional BOW. In text categorization area the relations between words are obtained using the N-gram model and the conditional probability of word sequences are estimated using this model. However this relation is not considered in the image representation. One reason for neglecting this relation in previous studies is the fact that considering N-gram model for images consumes too much memory space which makes it impractical. To deal with this problem inspired by [12], we propose a method based on the visual ontology construction. An example of such ontology is shown in Fig. 1. The leaves of this tree are the visual words and the internal nodes are used for words adjacencies modeling. In details, after constructing the vocabulary which includes k visual words we employ the agglomerate clustering algorithm to hierarchically group word pairs. The leaves of this ontology are the visual words and the internal nodes are the ancestors of the words. We refer to these internal nodes as general words since they are constructed from two child nodes and contain features which are similar to the features of their child. We will use the internal nodes of this ontology at level l for word adjacencies modeling.

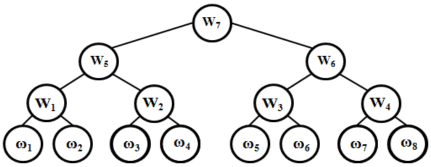

Fig. 1. An example of visual ontology. The leaves are visual words and represented by ω_i, the internal nodes are general words represented by W_j.

After defining the ontology, we use the general words to construct Bag of Bi-grams. We use general words for this purpose instead of visual words because if we directly use visual words the dimension of the features vector will be too high and can't be computed effectively. For example if the vocabulary consists of 200 words the number of Bi-grams will be more than 20000 which is very high and it is not suitable to consider it in BOW representation. In order to reduce the dimension of the feature vector, we can consider just 25 general words which are the ancestors of these words and obtain 325 Bi-grams. The diagram of this model is illustrated in Fig 2. For constructing the bag of Bi-grams we traverse the image from top-left to bottom-right and for each patch we consider the right, bottom and diagonal neighbors and assign each of them to one general word and count the number of general word pairs which are adjacent.

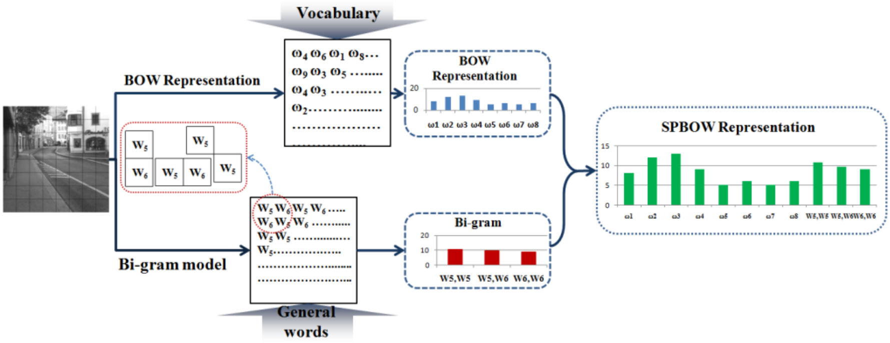

Fig. 2. Adding spatial information to BOW representation

Finally, we summarize our method for representing an image as follow:

1. Construct the visual ontology using agglomerative clustering of visual words and obtain the general words.
2. Calculate the frequencies of each individual word in the image. The histogram of occurrences of these words is referred to as BOW.
3. Count the number of times that every two general words are adjacent and concatenate these numbers to the BOW representation. We refer to this representation as SPBOW.

3 Experimental Results

In this section, we evaluate our proposed method for image classification on two data sets: 15 natural scene categories [9], and Caltech-101. Although these data sets contain color images, all the experiments are performed in grayscale. For experimental set up we follow Lazebnik et al. [9], and select randomly subsets of the data set to create train and test images. An SVM classifier with Laplacian radial basis function (LRBF) kernel is chosen to classify the images. To obtain image features, we extract SIFT features on a regular grid. The patches of this grid are 16*16 pixels and the sampling rate is set to 8 pixels. For constructing the vocabulary we employ the k-means algorithm on the features extracted from the training image and visual words are generated.

First experiments are performed on the 15 natural scenes which consists of 15 classes. Some samples of these dataset are shown in Fig. 3. We randomly select 100 images for training set and use the rest of the images in each class for testing. For all experiments a one level spatial pyramid is used. So, each image is partitioned to 2*2 sub images.

Table 1 shows the classification result of our method on this data set. For this experiment we use a vocabulary consisting of 256 words. The number of general words to serve as spatial relation modeling varies from 4 to 64. We see that when the number of general words is increased, the classification accuracy also increases. On the other hand, as the number of general words increases the algorithm becomes more complex because the number of Bi-grams becomes very high and therefore the algorithm needs more time and memory. So, it is not practical to use all of the general words and a

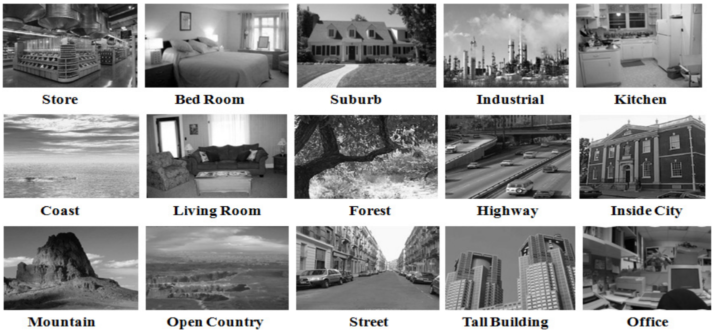

Fig. 3. Example images from the 15 natural scenes data set

tradeoff between the number of general words (algorithm complexity) and the classification accuracy is needed.

Table 1 shows that the classification accuracy doesn't change very much when we increase the number of general words from 16 to 32 and higher. For example when we use 64 general words the accuracy increases only 0.29 percent in comparison with the case we use 16 general words but we have to add almost 2000 value to feature vector in this case (There are 2080 and 136 pair combination for 64 and 16 general words respectively). Such behavior of the algorithm may be because the information content of the words adjacencies is limit. To illustrate this behavior, consider two white and blue patches occur in vicinity, showing a part of the sky. To realize that these two patches represent the sky, there is no need to quantize the blue color into several different blues and count the number of Bi-grams for every blue color. So, we can limit the number of general words to a predefined threshold. The value of this threshold depends on various parameters like number of classes, vocabulary size and the accuracy that we require.

Table 1. Classification results of the SPBOW representation

Number of general words	Number of Bi-grams	Classification accuracy
0	0	75.01 ± 0.4
4	10	75.86 ± 0.3
8	36	76.28 ± 0.4
16	136	76.35 ± 0.6
32	528	76.44 ± 0.5
64	2080	76.64 ± 0.3

Fig. 4 compares our method against BOW representation [2] and spatial pyramid matching (SPM) [9]. This figure plots the relationship between the classification accuracy and vocabulary size. In this experiment we used 16 general words for all vocabulary sizes. We see that our method which adds words adjacencies information to the BOW outperforms other methods which neglect this information. This supremacy can be seen for all vocabulary sizes but for small vocabularies this is more obvious. This behavior is because we use 16 general words for all vocabulary sizes. So, when we use small vocabularies the general words and visual words are more similar to each other in comparison with case we use larger vocabularies. For example when we use a vocabulary which consists of 16 visual words, the general words are the same as visual words. Furthermore, the number of visual words that each general word is a candidate for them increases when the size of vocabulary increases. For example, when the vocabulary consists of 512 words each of the 16 general words is a candidate for 32 of the visual words. In contrast each general word is a candidate for only 2 visual words when the size of the vocabulary is 32. So, as we model the spatial relation using general words, more information of visual words is neglected and we observe less improvement in classification accuracy for large vocabularies.

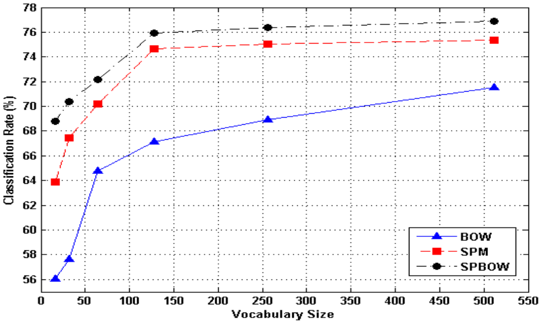

Fig. 4. Classification result on natural scene data set. The horizontal axis shows the vocabulary size and vertical axis represent the classification accuracy.

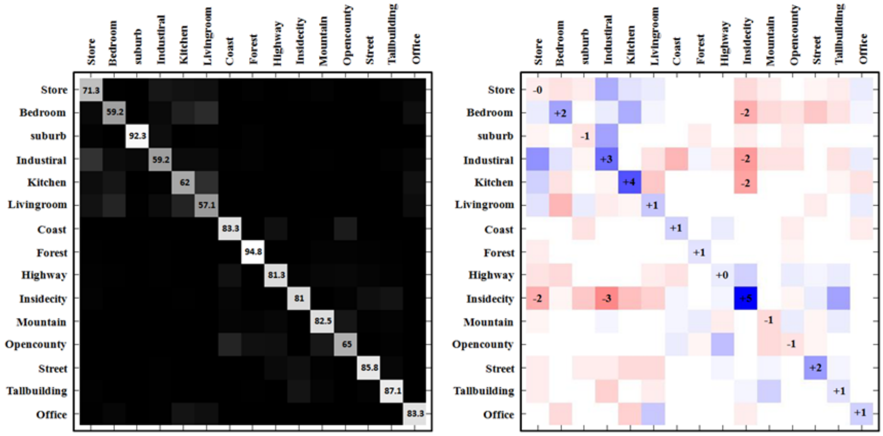

Fig. 5. (a) Confusion matrix of the 15 natural scene data set. The value at position (i,i) shows the classification rate for the class i. (b) Relative confusion matrix of natural scenes. The value at row i and column j which has been scaled, represents the difference between SPBOW and SPM to classify the images of class i as class j. We show positive and negative entries in blue and red respectively.

In Fig. 5 we show the confusion matrix of SPBOW representation and relative confusion matrix of 15 natural scene dataset. The relative confusion matrix illustrates the relation between confusion matrices of SPBOW and SPM representations. Every entry in this matrix denotes the absolute differences between entries in the confusion matrix of SPBOW and confusion matrix of SPM [9]. For this experiment the vocabulary size and number of general words are set to 256 and 16 respectively. We set up this experiment to observe the impact of words adjacency information on each class more clearly. The entries on the main diagonal of the relative matrix that shows the instances that correctly classified are mostly increased. As can be seen the classification accuracy of inside city, kitchen and industrial classes increase more than others. The non diagonal elements of this matrix show the misclassification rate and we see that the confusion declines for most of the class pairs. We clearly observe this improvement in the confusion between inside city as industrial, kitchen as inside city and industrial as inside city.

The second data set used for experiments is the Caltech-101. This data set consists of 101 objects and contains a broad range of objects. Fig. 6 shows some samples in this data set.

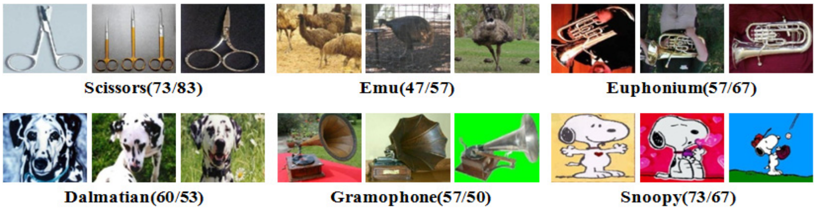

Fig. 6. Example images from Caltech-101 data set. Three top classes are those sample classes which our method has performed well compared to SPM and three bottom are samples which our method did not perform well. (SPM classification accuracy/SPBOW classification accuracy)

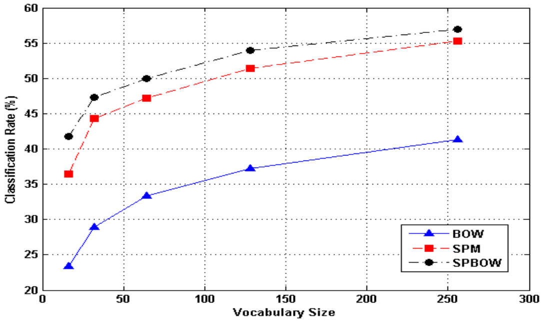

Fig. 7. Classification results on Caltech-101 data set. The horizontal axis shows the vocabulary size and the vertical axis represents the classification accuracy.

To construct train and test sets we randomly select 30 images per class for training and 30 images for testing. In Fig. 7 we show the classification accuracy of various image representation methods on Caltech-101 data set. In this experiment the number of general words used for words relation modeling is set to 16. We observe that spatial relations which are considered in our algorithm have positive effect on classification rate for all vocabulary sizes. In Fig. 6 some example classes are shown which our method has the best and worst performance on them.

4 Conclusion

In this paper we addressed the problem of neglecting the spatial information in BOW representation. For this purpose a new image representation based on modeling the words adjacencies using a tree structure was constructed and spatial relation between words were added to BOW. The experimental results on two known data sets showed that this new representation outperforms other representations and the spatial information plays an important role in detecting the content of the images. In this study the number of general words for modeling the relation between words was chosen based on the classification accuracy and algorithm complexity. However, an interesting future work is to find ways for selecting sub sets of the general words based on feature selection methods and using these sub sets in order to model the spatial relation.

References

[1] Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. ICCV (2003)

[2] Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2005)

[3] Lowe, K.D.: Distinctive Image Features from Scale-Invariant Keypoints. J. of Computer Vision 2(60), 91–110 (2004)

[4] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conf., pp. 147–151 (1988)

[5] Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)

[6] Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: Proc. CVPR 2003, Madison, WI, pp. 257–263 (June 2003)

[7] van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Trans. Pattern Analysis and Machine Intelligence 32(7), 1271–1283 (2010)

[8] Jiang, Y.G., Yang, J., Ngo, C.W.: Representation Of KeyPoint-Based Semantic Concept Detection: A Comprehensive Study. IEEE Trans. Multimedia 2(1), 42–53 (2010)

[9] Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2169–2178 (2006)

[10] Wu, L., Li, M., Li, Z., Ma, W.-Y., Yu, N.: Visual language modeling for image classification. In: ACM Multimedia Workshop on Multimedia Information Retrieval, pp. 115–124 (2007)

[11] Mei, L., Kweon, I., Hua, X.: Contextual Bag-of-Words for Visual Categorization. IEEE Trans. Circuits and Systems for Video Technology 21(4), 381–392 (2011)

[12] Jiang, Y.G., Ngo, C.W.: Bag-of-visual-words expansion using visual relatedness for video indexing. In: ACM Conf. on Research & Development on Information Retrieval, pp. 769–770 (2008)