Informative visual words construction to improve bag of words image representation

Mohammad Mehdi Farhangi, Mohsen Soryani, Mahmood Fathy

School of Computer Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
E-mail: mehdi.farhangi@gmail.com

Abstract: Bag of visual words model has recently attracted much attention from computer vision society because of its notable success in analysing images and exploring their content. This study improves this model by utilising the adjacency information between words. To explore this information, a binary tree structure is constructed from the visual words in order to model the is – a relationships in the vocabulary. Informative nodes of this tree are extracted by using the χ^2 criterion and are used to capture the adjacency information of visual words. This approach is a simple and computationally effective way for modelling the spatial relations of visual words, which improves the image classification performance. The authors evaluated our method for visual classification of three known datasets: 15 natural scenes, Caltech-101 and Graz-01.

1 Introduction

As the acquiring and storing of images and multimedia data is becoming fast and easy, the related databases become very large. In this situation, developing methods to manage such databases becomes more important. Classifying images based on their content is one of the methods which determine the category of an image among several categories. However, this task is a challenging problem in the real world since we encounter a number of difficulties, where there exists occlusion, background clutter and lighting changes in images. To deal with these difficulties, numerous new methods have been developed to describe images based on local features. Some of the recent methods for classifying images represent each image as a set of patches or regions, described by various descriptors. This representation is called bag of visual words (BoW) and provides a set of local descriptions for an image [1].

The process of BoW construction starts by extraction of local patches from the image. From this point of view, several methods have been proposed in the literature. For example, some researchers obtain local regions by using regular grids that segment images by horizontal and vertical lines [2, 3], or use variable size rectangular patches based on the complexity of the regions [4]. Furthermore, some may use blob detection algorithms such as difference of Gaussian [5], Harris detector [6] or Hessian matrix [7] to find salient points, which are placed on the corners and edges of the objects [1, 8, 9].

Salient patch detection is followed by its description using local descriptors like scale invariant feature transform (SIFT) [5], speeded up robust features (SURF) [7] and so on. Previous studies have shown that the SIFT descriptor extracts robust features from an image, which are more invariant to affine transformations than others [10]. Besides, some studies showed that other features like colour, texture and edge histograms, or a combination of them, could produce desirable results in special environments [11]. However, this method creates difficulty for the learning process, mostly because it uses different vectors dimension for each image [12]. To deal with this problem, similar patches are clustered in the same groups to constitute a cluster. Centres of these clusters are called visual words and the set of cluster centroids is treated as a vocabulary.

To represent image patches by visual words, each local description of the image should be assigned to one or more visual words. At last, by constructing the histogram of visual words distribution in the image, a holistic representation of the image, known as BoW model, is obtained.

Despite the success of BoW representation in image classification, 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. In text categorisation studies, this problem has been solved by introducing the *N*-gram terms; a model that considers the relation between the words of a text. On the other hand, the huge number of terms produced by the *N*-gram model, creates difficulty for BoW image representation and further processes; however, these relations convey important information about the content of the image. For example, a white patch can be part of a sheep, cloud or moon if it be surrounded by green grass, blue sky or dark area, respectively. To consider these relationships, we calculate the number of times each pair of the combination of words occurs in a certain neighbourhood and construct the bag of *N*-grams inspired by Li *et al*. [13]. To reduce the number of words combinations, we tried to generate new words from the vocabulary and use them to model the adjacencies.

The number of times that each pair of new words occurs in the vicinity is a new feature, which is concatenated to the BoW representation. The classification performance of this representation is verified on three datasets: 15 natural scenes [14], Caltech-101 [15] and Graz-01 [16]. Experimental results on these datasets confirm that our method, which includes spatial information, outperforms the traditional BoW and demonstrates the importance of spatial information in image categorisation tasks. Furthermore, since instead of visual words, fewer terms are used for modelling words adjacencies, it is computationally effective and does not need much memory space.

The remaining sections of this paper are organised as follows. The related works are reviewed in Section 2. In Section 3, the proposed method is explained in detail. Section 4 presents the experimental results and Section 5 concludes the paper.

2 Related works

As of today, many studies are trying to improve BoW representation with respect to several aspects. For example, although the traditional approach for feature vectors quantisation is to employ *k*-means clustering on the feature space [1, 2, 17, 18], many researchers explored new methods for this purpose. 'Randomly sampled codebooks' proposed by Nowak *et al*. [19], is one of these methods, in which some of the feature vectors are randomly selected as visual words. Although this method does not show superior performance against the *k*-means, but because of its simplicity and computational effectiveness, it can be used when the number of feature vectors is very high. Another study proposed is radius-based clustering, which finds visual words that each represents a distinct part of the feature space [3]. In this algorithm, all the features within a fixed radius of similarity are assigned to a certain cluster. This algorithm outperforms the standard *k*-means since it generates an even distribution of visual words over the feature space compared with the *k*-means algorithm, which generates most of the clusters in the high-frequency area of the feature space.

One problem of the clustering algorithms for vocabulary construction is that many of them fail to converge because of the large amount of image patches. Therefore, in [20], a hierarchical *k*-means algorithm was proposed, in which a vocabulary tree is constructed by applying *k*-means within each partition of the tree. In another study and in order to obtain more informative details of local patches, Yang *et al*. [21] utilise two vocabularies instead of one. One vocabulary is constructed by colour features whereas the other is obtained by quantising local binary pattern features.

Another aspect of the BoW construction, investigated by researchers, is the way that local descriptors are assigned to visual words. This procedure is known as the weighting scheme and can be performed in several ways. Some authors have used hard weighting methods, in which each local descriptor of the image is assigned to its nearest visual word in the vocabulary [1, 2]. On the other hand, recent methods use soft weighting approaches, where each local feature can be assigned to more than one word with different weights [12, 22, 23]. These weights are usually determined by the distances between the local descriptor and visual word vectors [12, 22]. Previous studies have shown that soft weighting approaches outperform hard weighting ones since they effectively model the correlation between visual words, where different visual words can represent one local descriptor with various weights [22].

Since BoW representation naturally neglects spatial information of images, this information has been considered separately in some previous works. One of the first attempts in this area was proposed by Lazebnik *et al*. [14], which is based on partitioning an image into increasingly finer grids. After computing the frequencies of visual words in each grid cell, the BoWs of the cells are concatenated to each other and thus a representation of the image, conveying the spatial location of visual words, is obtained. The importance of this method becomes obvious when we note that the location of visual words in the image conveys essential information about its content. For example, a blue patch, located above the image, is probably representing a piece of sky whereas if this patch be in the bottom of the image, it may represent part of a sea.

Tirilly *et al*. [24] presented another work in this area by introducing visual sentences. They order the words of an image in relation to a certain axis in the image and construct visual sentences, which contain the spatial relations between words. One problem in their method is that since it projects all the local descriptors to the main axis, it does not consider the spatial relations in all directions and only includes the relations in the main axis direction.

Wu *et al*. [25], introduced another approach to capture the proximity information. Their model represents three kinds of relation including unigram, bigram and trigram between visual words. However, this model creates difficulty for the learning process, because it uses all the visual words to construct bigram and trigram and produces a huge number of terms. This problem is solved in [26], where just a small fraction of all words combinations are used for modelling the proximity information. To find these word pairs, the authors measured confidence values that each show by how much two neighbouring visual words are relevant. Word pairs with high confidences were concatenated to BoW representation. However, the constraint of this algorithm appears when the size of the vocabulary becomes large since for informative word pairs selection, we have to construct all the words combinations. In this case, counting all the words combinations, even for small vocabularies, is very time-consuming. To deal with this problem, we propose a new approach, based on extending our preliminary work [27] with a method to construct informative words, producing results on a new dataset and comparison with other image representations. The informative nodes are obtained from the nodes of an ontology structure and are appropriate for words relation modelling. This claim is verified by Section 4, where we tested our algorithm on three known datasets: 15 natural scenes, Caltech-101 and Graz-01.

3 Proposed method

This work focuses on spatial information modelling and is performed in two steps. First, we use the spatial pyramid matching (SPM) approach [14] and partition the image into fine subregions to obtain the histogram of local features inside each subregion. Next, we obtain the number of occurrences of visual word pairs and concatenate that to the BoW representation as new features. The following subsections present this procedure in detail.

3.1 BoW representation

As described in Section 2, after extracting the local patches, every patch has to be described by using SIFT descriptors. Since the previous studies have shown that sampling on a regular grid outperforms other approaches such as interest point detectors, we use the SIFT descriptor, sampled on a regular grid [2].

Next, each local descriptor of the image should be assigned to one or more visual words. In this paper, we use the hard weighting approach by mapping every local patch to one visual word [1]. To do so, assume that $\{r_1, r_2, \ldots, r_n\}$ represents the local descriptors in the image and $V = \{\omega_1, \omega_2, \ldots, \omega_k\}$ represents the vocabulary. With these assumptions the hard histogram of visual words is computed as

$$\mathrm{HBoW}\left(\omega_j\right) = \sum_{i=1}^{n} \begin{cases} 1, & \text{if } \omega_j = \left(\mathrm{dist}\left(\omega_j, r_i\right)\right) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where r_i is the ith patch of the image and ω_j is the jth word in the vocabulary. Clearly, soft weighting approaches provide better performance in classification, but since we intend to compare our method with classical methods based on the importance of spatial information, we choose hard weighting for our algorithm.

3.2 Spatial pyramid matching

The BoW representation described above ignores some useful information of the image. For example, 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 the BoW, we use SPM, proposed by Lazebnik *et al.* [14], to partition the image into rectangular regions.

In detail, 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 given higher weight compared with matches found at coarser resolutions. More specifically, a sequence of grids is constructed at resolutions $0\ldots L$, 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 ith cell of the grid. Then, the histogram intersection function finds the number of matches at level l.

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

Note that the number of matches found at level l also includes all the matches found at the finer level $l+1$. Therefore the number of new matches found at level l is given by $I^l - I^{l+1}$ for $l = 0, \ldots, L-1$. The weight associated with level l is set to $1/2^{L-l}$, which is inversely proportional to the cellwidth at that level. Intuitively, since the matches found in larger cells involve dissimilar features, they should be weighted lower. Hence, the following definition was obtained for the pyramid match kernel

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

To combine the pyramid matching kernel with the spatial location of words in the image, the elements of X and Y are used to represent the coordinates of the visual words in the image. Therefore, by placing increasingly fine grids on the coordinates of the visual words, the spatial information is combined with the BoW representation.

3.3 Spatial relation modelling

The relations between the visual words in an image convey important information about its content; however, this information has been neglected in the traditional BoW model. In the text categorisation area, relations between words are obtained by using the N-gram model and the conditional probability of the word sequences is estimated by using this model. Nonetheless, this model was not applied to the image representation in previous studies because of the fact that considering the N-gram model for images consumes too much memory space, which makes it impractical. To deal with this problem inspired by Jiang and Ngo [28], we proposed a method based on the visual ontology construction.

3.3.1 Visual ontology construction: As mentioned before, spatial relations between visual words can be used as additional information to improve the classification performance. However, considering this information similar to what is used in the text categorisation area and constructing N-gram terms is impractical. For example, if the size of the vocabulary is 200, the number of bigrams will be >20 000, which is very high and it is not practical to consider it in the BoW representation.

To effectively model the spatial relation, we first construct a tree from the visual words and use the informative internal nodes of this ontology tree for adjacency modelling. An example of such ontology, which consists of eight visual words, is shown in Fig. 1. The leaves of this tree are the visual words and the internal nodes, which constitute much fewer pairs and are used for words adjacency modelling. For instance, instead of using 200 visual words we consider only 25 nodes of their ancestors and obtain 325 bigrams.

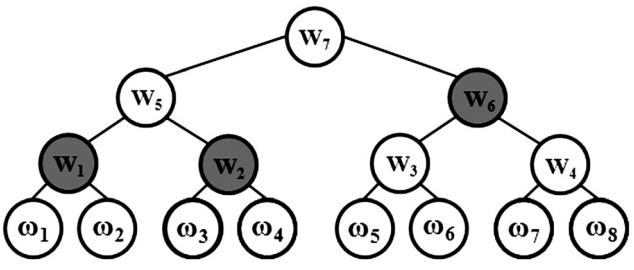

Fig. 1 *Example of visual ontology*

Leaves are visual words and represented by ω_i, and the internal nodes are general words represented by W_j

In general, the number of bigrams obtained by the n words is computed as follows

$$\text{number of Bi-grams} = \binom{n}{2} + n = \frac{n(n+1)}{2} \quad (4)$$

To construct a visual ontology like the one shown in Fig. 1, we use an agglomerative clustering algorithm starting with the visual words of the vocabulary [29]. For example, in Fig. 1, the algorithm starts with $\{\omega_1, \omega_2, \ldots, \omega_8\}$ and tries to find the most similar nodes to combine them. The similarity between two nodes is computed by measuring the Euclidean distance between the feature vectors of those nodes. Now, let us assume that the most similar nodes in Fig. 1 are ω_1 and ω_2. W_1 is obtained by combining these nodes and its feature vector is computed by averaging the feature vectors of its children. The same procedure employs ω_3, ω_4 to construct W_2. This process continues until just one node remains, that is, the root of the ontology (W_7 in Fig. 1).

The described method for ontology construction slightly differs from agglomerative clustering where in the latter case each node is a group of training samples and the distance between the nodes is defined as single linkage, complete linkage or average linkage [29]. In our algorithm, however, each node contains a single sample and the distance between them is the Euclidean distance between their feature vectors. Furthermore, our algorithm forces the nodes of the same depth to merge, although this restriction does not exist in the classical version of agglomerative clustering. This restriction prevents us from generating an unbalanced tree in height, which helps to approximate the visual words by fewer internal nodes.

As mentioned earlier, the leaves of the ontology are the visual words (ω_1, ω_2, …, ω_8) and the internal nodes (W_1, W_2, …, W_8) are the ancestors of the visual words. We refer to the internal nodes as general words since they are constructed from two child nodes and contain features which are similar to the features of their children. In the next subsection, we will show how the informative node pairs of this tree structure are used to effectively capture the spatial information in the images.

3.3.2 Finding informative general word pairs: To obtain informative general word pairs, we start with an initial set of general words (e.g. W_1, W_2, W_3, W_4). We measure a fitness for every general word of this set to determine how much is appropriate to be associated for adjacency modelling. Based on the fitness values of the nodes, a best first approach is used to determine which node should be chosen to be expanded [30]. In other words, the fitness values are computed for all the words in the set and the word with the lowest fitness is replaced by its children hoping that they are more informative than their parent. For example, if we recognise W_n as the least appropriate general word, it will be replaced by its children W_{2n} and W_{2n+1}. This procedure is iterated until we obtain a predefined number of general words. For example, the highlighted words in Fig. 1 are assumed as the most informative words and are candidates for all the visual words. According to this method, the fitness value of every node is determined by feature selection based on χ^2 criterion [31]. In detail, we use the χ^2 statistic to measure the lack of independence between two general words W_1 and W_2 based on the following equation

$$\chi^2(W_1, W_2) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (5)$$

In this equation, A is the number of times that W_1 and W_2 occur in vicinity. B is the number of times that W_1 occurs without W_2, C is the number of times that W_2 occurs without W_1, D is the number of times that neither W_1 nor W_2 occur and N is the number of pair combinations of general words. This criterion is equal to zero, if W_1 and W_2 are independent.

After computing the χ^2 distribution for all the pair combinations, the average goodness of a general word such as W_i is defined as follows

$$\chi^2_{\text{avg}}(W_i) = \sum_{j=1}^{m} P(W_j)\chi^2(W_i, W_j) \quad (6)$$

where $P(W_j)$ is the probability of the observation of W_j in the training images. The expansion of a general word is based on (6). In other words, in each stage of ontology construction, the node with the lowest criterion $(\chi^2_{\text{avg}}(W_i))$ is replaced by its children.

At last, the candidate general words are used to construct bigrams. The diagram of this model is illustrated in Fig 2. To construct bigrams, we traverse the image from the top-left to the bottom-right and for each patch we consider right, below and diagonal neighbours. Next, every patch is assigned to one general word and the numbers of adjacent general words are considered as new features. Concatenating these new features with standard BoW representation provides a new image representation which we refer to as spatial BoW (SPBoW).

This method is computationally effective because of the fact that it dissociates the time complexity of adjacency modelling from the size of the vocabulary and relates it to the number of general words. Precisely, the number of general word pairs, which are examined to find the most appropriate ones to expand, is proportional to the time complexity. In each step of tree expansion, each pair combination of leaf nodes is investigated to find the most appropriate node for expansion. Hence, the time complexity is proportional to the sum of pair leaves in all the generated trees.

$$\text{Sum of pair leaves} = \sum_{i=2}^{N_g} \binom{i}{2}; \quad N_g = \text{number of general words} \quad (7)$$

Comparing this number with all the possible words combinations, which were used in the contextual Bag of Words (CBoW) representation [13], shows that our method outperforms the CBoW representation, when time complexity is investigated.

Finally, we summarise our method for generating SPBoW image representation as follows:

1. Construct the visual ontology by using agglomerative clustering of visual words and obtain the general words using best first search on the ontology space.

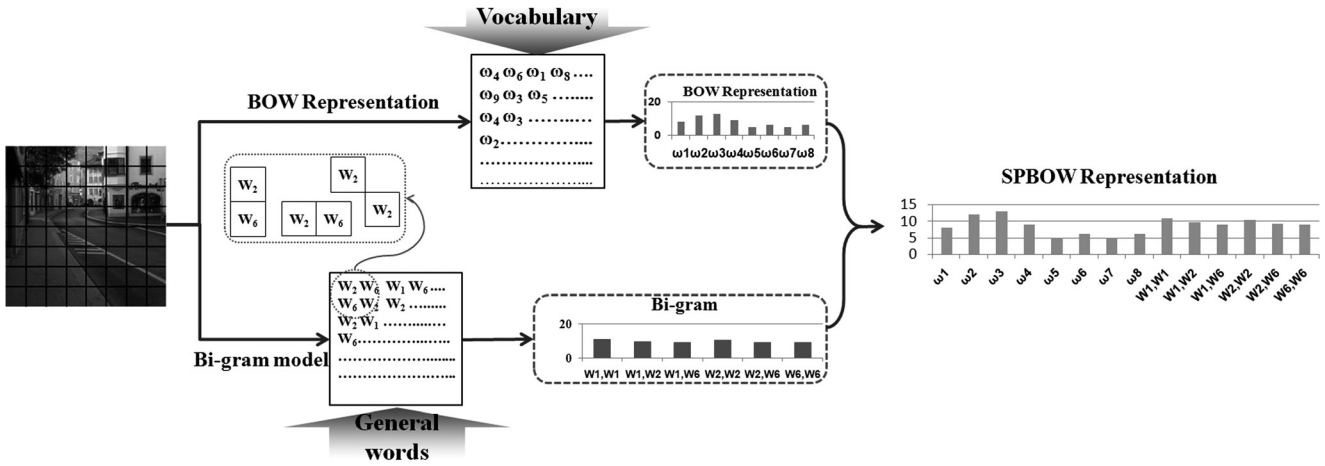

Fig. 2 *Adding spatial information to the BoW representation*

2. Calculate the frequencies of each individual word in the image. The histogram of occurrences of these words is referred to as BoW.
3. Construct the visual ontology and find the most informative general words using χ^2 criterion.
4. Count the number of times that every two informative general words are adjacent and concatenate these numbers to the BoW representation.

4 Experimental results

In this section, we evaluate our proposed method for image classification on three datasets: 15 natural scenes [14], Caltech-101 [15] and Graz-01 [16]. Although these datasets contain colour images, all the experiments have been performed in greyscale. For experimental set-up we follow Lazebnik *et al*. [14] and randomly select subsets of the dataset to create train and test images. However, because of small implementation differences, our implementation of [14] performs slightly lower than their reported results. An support vector machine (SVM) classifier with linear kernel was chosen to classify the images based on SIFT features extracted on a regular grid. The patches of the grid are 16 × 16 pixels and the sampling rate is set to 8 pixels. Hence, each patch shares some pixels with its neighbours. At last, *k*-means was chosen to quantise feature space.

4.1 Fifteen natural scenes

First, experiments were performed on the 15 natural scenes dataset. Some of these dataset samples are shown in Fig. 3. The number of images in the classes of this dataset varies from 210 to 410. We randomly selected 100 images for the training set and use the rest of the images in each class for testing. In all the experiments, a one level spatial pyramid was used, so each image was partitioned into 2 × 2 sub images.

Fig. 4 compares our method against the BoW representation [2] and the SPM [14] by plotting the

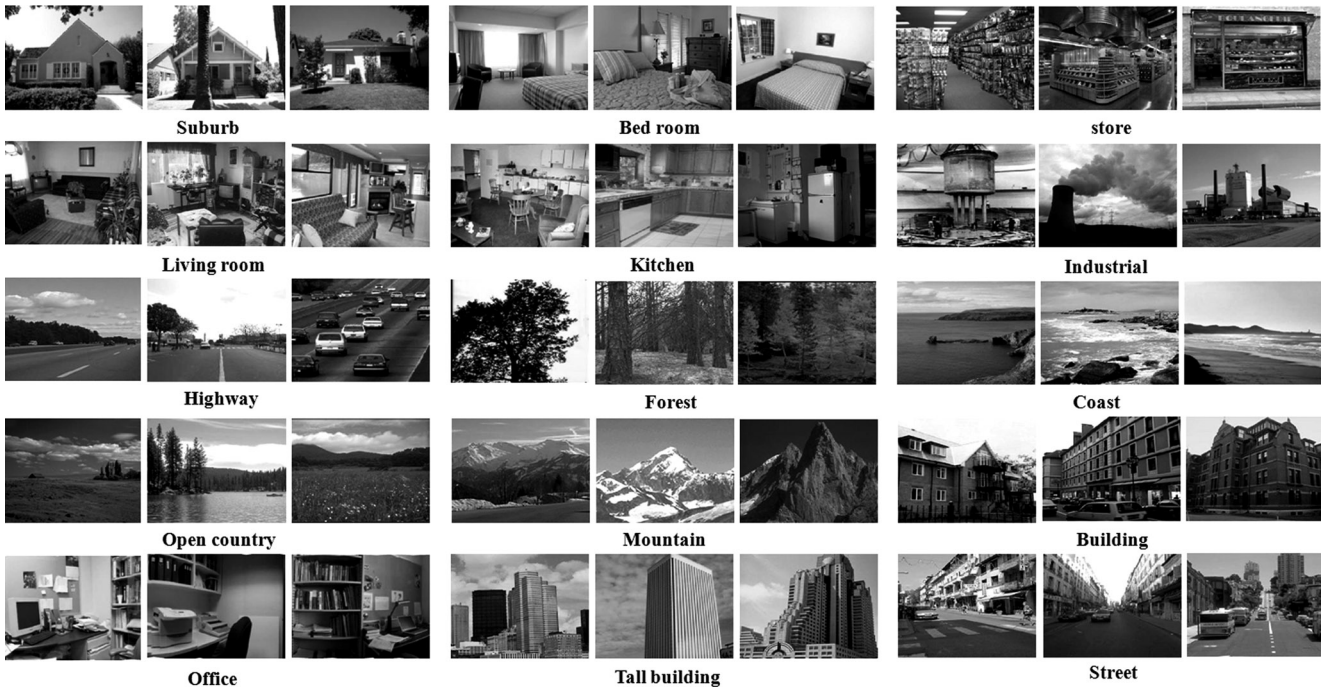

Fig. 3 *Example images from the 15 natural scenes dataset*

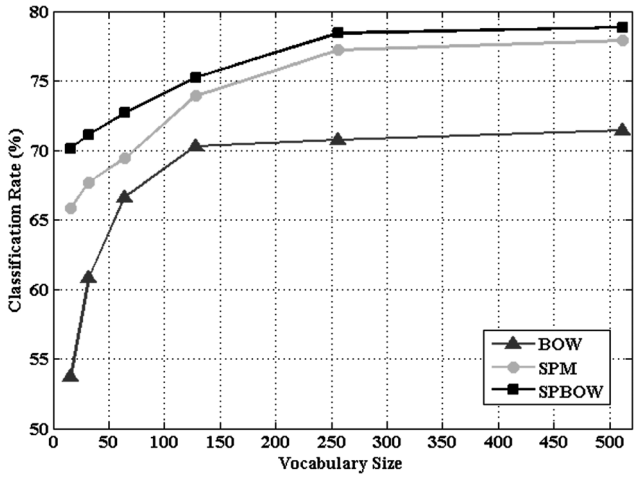

Fig. 4 *Classification results for the 15 natural scenes dataset*

Horizontal axis shows the vocabulary size and the vertical axis represents the classification accuracy

relationship between classification accuracy and vocabulary size. In this experiment, we used 16 general words for all the vocabulary sizes and observed that our method, which adds words adjacencies information to the BoW, outperforms the other representations. This supremacy can be seen for all vocabulary sizes, but for small vocabularies this is more obvious. For example, the difference between the classification accuracy of SPBoW and SPM is $>4\%$ when the size of the vocabulary is equal to 16, whereas this difference is $<1\%$ for a vocabulary of 512 words. The fixed number of general words (16 general words for all vocabulary sizes) used for SPBoW representation is the cause of this behaviour because when we use small vocabularies, the general words and visual words are more similar to each other in comparison with cases where larger vocabularies are used. For example, when we use a vocabulary that consists of 16 visual words, the general words are the same as the visual words. Furthermore, the number of visual words, that each general word is a candidate of, increases when the size of the vocabulary becomes larger. For instance, when the vocabulary consists of 512 words, each of the 16 general words is a candidate for 32 visual words. In contrast, each general word is a candidate for only 2 visual words when the size of the vocabulary is 32. Thus, as we model the spatial relationship using general words, more information of visual words is neglected and we observe less improvement in classification accuracy for larger vocabularies.

Table 1 provides some results to emphasise the importance of spatial information in various representations. For each representation, the number of visual words is changed from 16 to 1024. In this experiment, we use 136 most informative word pairs for CBoW representation and 16 informative general words for SPBoW representation. The last row of this table shows the classification accuracy for another image representation, which is provided for comparison and is called BoW_AR. To obtain this representation, we concatenated the standard BoW representations with the bigram terms of SPBoW representation. Comparing this representation with others, reveals the significant effect of spatial and adjacency information in classification accuracy. Besides, the last column of this table shows a restriction of CBoW representation, where it cannot extract informative pairs from a large vocabulary, because it has to construct all the words combinations and obviously it is not practical for large vocabularies. This restriction does not exist in our representation because it models the spatial relationships in a hierarchical manner and does not need to obtain all the words combinations.

In addition, our representation outperforms CBoW in classification accuracy for all vocabulary sizes. Results of another experiment which supports this claim are shown in Fig 5. In this experiment, we use the bag of bigrams for image classification, in which the informative word pairs (the lower curve) and informative general word pairs (the upper curve) serve as the feature vectors. The vocabulary size in this experiment is 256 and the number of word pairs is varied from 36 to 820. The straight line, which is on top of the figure, is the classification accuracy based on all the 256 words combinations and shows the upper bound for the classification accuracy based on words adjacencies.

Again, we can see that the word pairs generated based on our method are more effective than the informative word pairs obtained by CBoW representation. The cause of this supremacy is revealed when we note carefully the feature vectors of these representations. Although the CBoW

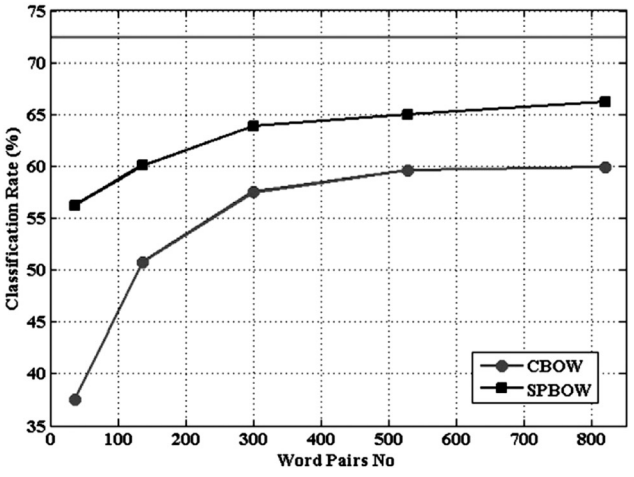

Fig. 5 *Classification results for the 15 natural scenes dataset based on word pairs*

Horizontal axis shows the number of informative word pairs and the vertical axis represents the classification accuracy

Table 1 Performance of various image representations with different vocabulary sizes

Image representation	Vocabulary size = 16	Vocabulary size = 32	Vocabulary size = 64	Vocabulary size = 128	Vocabulary size = 256	Vocabulary size = 512	Vocabulary size = 1024
CBoW	70.1 ± 0.2	71.1 ± 0.7	71.8 ± 0.7	75.1 ± 0.3	78.2 ± 0.2	78.5 ± 0.4	X
SPBoW	70.1 ± 0.2	71.1 ± 0.4	72.7 ± 0.5	75.2 ± 0.3	78.4 ± 0.4	78.8 ± 0.2	79.1 ± 0.2
BoW	53.7 ± 0.7	60.8 ± 0.4	66.6 ± 0.6	70.3 ± 0.6	70.8 ± 0.7	71.4 ± 0.6	71.9 ± 0.6
BoW_AR	65.4 ± 0.7	67.1 ± 0.6	70 ± 0.5	71.6 ± 0.8	72.4 ± 0.6	73.3 ± 0.7	73.9 ± 0.6

Table 2 Performance of the BoW and CBoW representations with different word pairs number

Image representation	Word pairs no. = 36	Word pairs no. = 136	Word pairs no. = 300	Word pairs no. = 528	Word pairs no. = 820
CBoW	77.8 ± 0.2	78.2 ± 0.2	78.1 ± 0.1	78.2 ± 0.1	78.5 ± 0.5
SPBoW	78.3 ± 0.4	78.4 ± 0.4	78.4 ± 0.2	78.5 ± 0.3	78.9 ± 0.4

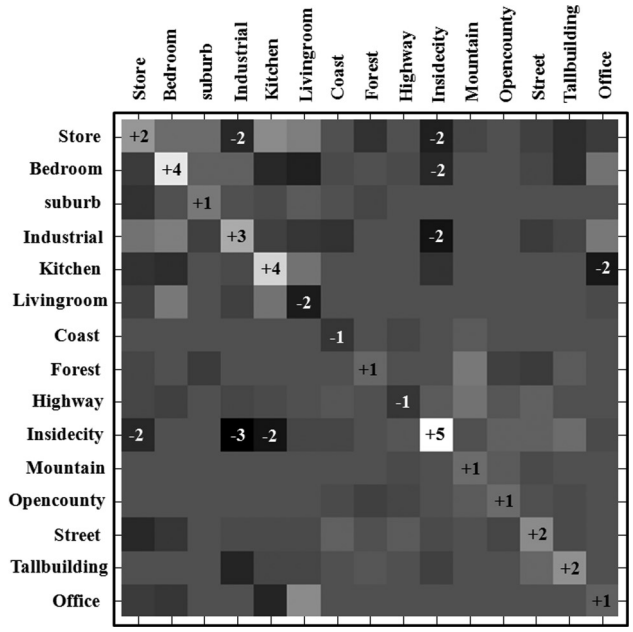

Fig. 6 *Relative confusion matrix of the 15 natural scenes*

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. The brightness of regions is proportional to the values of the matrix entries

representation tries to use the most informative word pairs for adjacency modelling, the bag of bigram generated by this model is extremely sparse because it uses only a small fraction of all possible words combinations and neglects other word pair occurrences. On the other hand, our method utilises all the word pairs by assigning them to general words.

Another interesting behaviour of the algorithm is shown in Table 2. In this experiment, the classification accuracies of CBoW and SPBoW representations based on 256 visual words and different number of bigrams are reported. We can see that the classification accuracy does not change very much when we increase the number of general word pairs from 36 or 136 to higher values. This behaviour of the algorithm may be because of the limitation of the information content of the words adjacencies. To illustrate this behaviour, consider when two white and blue patches occur in vicinity, showing a part of the sky. To realise that these patches represent the sky, there is no need to quantise the blue colour into several different blues and counting the number of bigrams for every blue colour.

To compare the classification accuracy on each class separately for two representations, the relative confusion matrix is shown in Fig. 6. For this experiment, the vocabulary size and the number of general words are set to 256 and 16, respectively. This matrix illustrates the relationships between the confusion matrices of SPBoW and SPM representations. Every entry denotes the absolute difference between the entries in the confusion matrix of SPBoW and confusion matrix of SPM [14]. The entries on the main diagonal of the matrix, which shows the correctly classified instances, are mostly increased. As can be seen, the classification rate of the inside city, kitchen and industrial classes increased more than others. The non-diagonal elements of this matrix show the misclassification rate and we can see that the confusion declines for most of the class pairs. We clearly observe this improvement in the confusion between inside city as industrial, industrial as inside city, kitchen as office and some other class pairs which can be seen in Fig. 6.

4.2 Caltech-101

The second dataset used for the experiments is Caltech-101. This dataset consists of 101 object classes and the number of images in each class varies between 31 and 800. This dataset contains a broad range of objects that are usually placed in the centre of images. Some samples of this dataset are shown in Fig. 7. To construct the train and test sets, we randomly selected 30 images per class for training and 30 images for testing.

The classification performance of the BoW, SPM, CBoW and SPBoW representations on this dataset are shown in Table 3. The vocabulary size for this experiment is set to 256 and we use 136 bigrams for the CBoW and SPBoW representations.

4.3 Graz dataset

Graz-01 is the third dataset, which is chosen to evaluate the proposed method. This dataset consists of two object classes: bike, person and a background class. Two important properties of Graz-01 convinced us to evaluate

Table 3 Result of different representations on Caltech-101

Representation	BoW	SPM	CBoW	SPBoW
accuracy	43.7 ± 0.2	56.2 ± 0.3	56.4 ± 0.1	57.2 ± 0.1

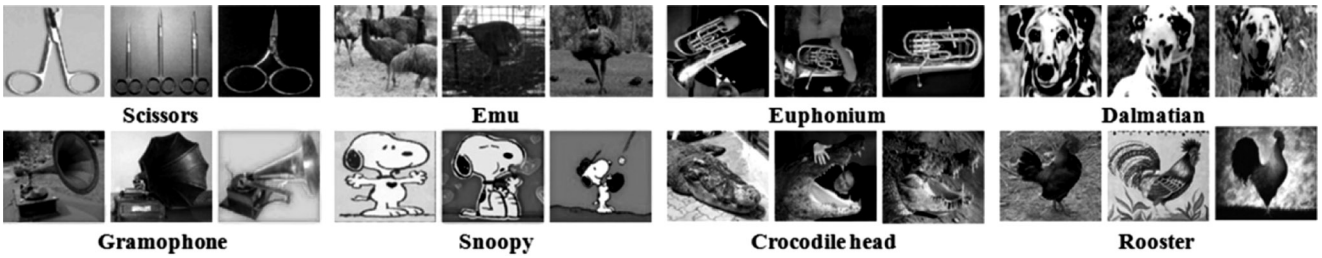

Fig. 7 *Example images from the Caltech-101 dataset*

Fig. 8 *Graz-01 dataset*

First row: people class; second row: bike class. The first five images of each row are those which our method classified incorrectly

Table 4 Result of different representations on the Graz-02 dataset

Class	BoW	SPM	SPBoW	Opelt [32]
bike	82.7 ± 1.4	85.1 ± 1	85.4 ± 2	86.5
people	79.8 ± 3.2	81.2 ± 1.8	81.1 ± 1.8	80.8

our method on this dataset. First, as Fig. 8 presents, the images in each class appear at different scales, viewpoints and positions. In addition, there are significant occlusions, background clutter and lighting changes in the images of this dataset, which make it an appropriate dataset to model real-world images.

In this experiment, we train detectors to detect persons and bikes on 100 positive and 100 negative examples. Half of the samples of negative examples are chosen from the other object class and the rest of them are randomly selected from the background class. This experiment was designed in such a way as to be consistent with the earlier study by Opelt *et al*. [32]. We generate receiver operating characteristics (ROC) curves by thresholding the raw SVM output and report the ROC equal error rate averaged over ten runs.

The results of this experiment are shown in Table 4. As can be seen, the word adjacency information is still useful and causes improved performance on this database. Another notable result is that the deviation between different representations is quite high and that is because of the high intra-class variations in this dataset. However, considering the spatial information in the SPM, CBoW and SPBoW representations adjusts this situation and causes less variation in the classification performance. Although consideration of the spatial information in the SPM, CBoW and SPBoW representations causes better performance and less variation in accuracy, however, it cannot completely overcome the problem of occlusion, background clutter and lighting changes since most of the misclassified images are the ones which are suffering from these disorders. For example, the five images shown on the left of the first and second rows of Fig. 8 are some of the samples of this dataset which our method could not classify correctly. As can be seen, most of these samples contain significant occlusion, background clutter or lighting changes. On the other hand, the algorithm proposed by Opelt *et al*. [32] performs slightly better in these conditions since its aim was to handle this situation based on combining weak classifiers each of which utilises different set of features.

5 Conclusions

In this paper, we modelled the words adjacencies to improve BoW representation. For this purpose, we considered informative nodes of an ontological tree structure to model words adjacencies and the spatial relations of these new words were added to the BoW representation. The experimental results showed that this representation outperforms the other representations and the spatial relations between the words play an important role in detecting the contents of images. This claim was verified by detecting the image contents of the three known datasets.

Like other BoW-based methods, our model suffers from poor analysis when there exists occlusion, background clutter, lighting changes or change in scale. We illustrated this behaviour in Fig. 8, where our algorithm was not able to recognise the objects in small scales or when they were not the dominant object of the scene.

One of the advantages of our method is that it can be easily applied to the field of video analysis to capture temporal information of consecutive frames. Most of the video analysis works, based on the BoW representation, confine their method to the key frames of the shots and disregard the temporal information. An interesting future work is to apply the proposed method to the field of temporal modelling, in which the sequences of general words occurring in consecutive frames can be utilised for temporal information capturing. Since different numbers of general words can be used for this purpose, various accuracies, time and space complexities can be acquired.

6 References

1 Sivic, J., Zisserman, A.: 'Video google: a text retrieval approach to object matching in videos'. Proc. Int. Conf. Computer Vision (ICCV'03), Nice, France, October 2003, pp. 1470–1477

2 Fei-Fei, L., Perona, P.: 'A Bayesian hierarchical model for learning natural scene categories'. Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, June 2005, pp. 524–531

3 Jurie, F., Triggs, B.: 'Creating efficient codebooks for visual recognition'. Proc. Int. Conf. Computer Vision (ICCV'05), Beijing, China, October 2005, pp. 604–610

4 Krishnamoorthy, R., Punidha, R.: 'An orthogonal polynomials transform-based variable block size adaptive vector quantization for color image coding', *IET Image Process.*, 2012, **6**, (6), pp. 635–646

5 Lowe, K.D.: 'Distinctive image features from scale-invariant keypoints', *Int. J. Comput. Vis.*, 2004, **60**, (2), pp. 91–110

6 Harris, C., Stephens, M.: 'A combined corner and edge detector'. Proc. Int. Conf. Alvey Vision, 1988, pp. 147–151

7 Bay, H., Tuytelaars, T., Gool, L.V.: 'Surf: speeded up robust features'. Proc Ninth European Conf. Computer Vision (ECCV'06), Graz, Austria, May 2006, pp. 404–417

8 Wu, X., Zhao, W.L., Ngo, C.W.: 'Near duplicate keyframe retrieval with visual keywords and semantic context'. Proc. Int. Conf. Image and Video Retrieval (CIVR'07), Amsterdam, The Netherlands, July 2007, pp. 162–169

9 Bouachir, W., Kardouchi, M., Belacel, N.: 'Fuzzy indexing for bag of features scene categorization'. Proc Int. Symp. Visual Computing (ISVC'10), Rabat, September 2010, pp. 1–4

10 Milkojcyk, K., Schmid, C.: 'A performance evaluation of local descriptor', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (10), pp. 1615–1630

11 Vogel, J., Shiele, B.: 'Semantic modeling of natural scenes for content based image retrieval', *Int. J. Comput. Vis.*, 2007, **72**, (2), pp. 133–157

12 Jiang, Y.G., Ngo, C.W.: 'Visual word proximity and linguistic for semantic video indexing and near duplicate retrieval', *Int. J. Comput. Vis. Image Underst.*, 2009, **113**, (2), pp. 405–414

13 Li, T., Mei, T., Kweon, I.S., Hua, X.S.: 'Contextual bag-of-words for visual categorization', *IEEE Trans. Circuits Syst. Video Technol.*, 2011, **21**, (4), pp. 381–392

14 Lazebnik, S., Schmid, C., Ponce, J.: 'Beyond bags of features: spatial pyramid matching for recognizing natural scene categories'. Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR'06), New York, NY, June 2006, pp. 2169–2178

15 Fei-Fei, L., Fergus, R., Perona, P.: 'Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories'. Proc. Workshop Generative-Model Based Vision, 2004

16 Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: 'Weak hypotheses and boosting for generic object detection and recognition'. Proc. European Conf. Computer Vision (ECCV'04), Prague, Czech Republic, May 2004, vol. 2, pp. 71–84

17 Herve, N., Boujemaa, N.: 'Visual word pairs for automatic image annotation'. Proc. IEEE Int. Conf. Multimedia and Expo (ICME'09), Cancun, Mexico, June 2009, pp. 430–433

18 Jegou, H., Douze, M., Schmid, C.: 'On the burstiness of visual elements'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1169–1176

19 Nowak, E., Jurie, F., Triggs, B.: 'Sampling strategies for bag of features image classification'. Proc. European Conf. Computer Vision (ECCV'06), Graz, Austria, May 2006, pp. 490–503

20 Nister, D., Stewenius, H.: 'Scalable recognition with a vocabulary tree'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR'06), New York, NY, June 2006, pp. 2161–2168

21 Yang, F., Lu, H., Zhang, W., Yang, G.: 'Visual tracking via bag of features', *IET Image Process.*, 2012, **6**, (2), pp. 115–128

22 Jiang, Y.G., Yang, J., Ngo, C.W.: 'Representation of keypoint-based semantic concept detection: a comprehensive study', *IEEE Trans. Multimedia*, 2010, **12**, (1), pp. 42–53

23 Gemert, J.C.V., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: 'Visual word ambiguity', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, (7), pp. 1271–1283

24 Tirilly, P., Claveau, V., Gros, P.: 'Language modeling for bag of visual words image categorization'. Proc. Int. Conf. Image and Video Retrieval (CIVR'08), Niagara Falls, Canada, July 2008, pp. 249–258

25 Wu, L., Li, M., Li, Z., Ma, W.Y., Yu, N.: 'Visual language modeling for image classification'. Proc. ACM Multimedia Workshop on multimedia information retrieval, Germany, September 2007, pp. 115–124

26 Mei, L., Kweon, I., Hua, X.: 'Contextual bag-of-words for visual categorization', *IEEE Trans. Circuits Syst. Video Technol.*, 2011, **21**, (4), pp. 381–392

27 Farhangi, M.M., Soryani, M., Fathy, M.: 'Improvement the bag of words image representation using spatial information'. Proc. Second Int. Conf. Advances in Computing and Information Technology (ACITY'12), Chennai, India, July 2012, pp. 681–690

28 Jiang, Y.G., Ngo, C.W.: 'Bag-of-visual-words expansion using visual relatedness for video indexing'. Proc. ACM SIGIR Conf. Research and Development on Information Retrieval, Singapore, July 2008, pp. 769–770

29 Alpaydin, E.: 'Introduction to machine learning' (Cambridge Massachusetts, The MIT Press, 2004)

30 Russell, S.J., Norvig, P.: 'Artificial intelligence: a modern approach' (Upper Saddle River, New Jersey, Prentice-Hall, 2003, 2nd edn.)

31 Yang, Y., Pedersen, J.O.: 'A comparative study on feature selection in text categorization'. Proc. Int. Conf. Machine Learn (ICML'97), 1997, pp. 412–420

32 Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: 'Weak hypothesis and boosting for generic object detection and recognition'. Proc. European Conf. Computer Vision (ECCV'04), Prague, Czech Republic, May 2004, vol. 2, pp. 71–84