Scene image classification with biased spatial block and pLSA

Zhong Ji, Peiguang Jing, Jing Wang, Yuting Su
School of Electronic and Information Engineering, Tianjin University, China.
jizhong@tju.edu.cn

Abstract

Scene image classification is a fundamental problem in the fields of computer vision and image understanding. A novel scene image classification method based on biased spatial block information and an improved coding approach in bag-of-visual-words (BOW) model is proposed. The spatial constraints biased to central object regions are employed to achieve better discrimination power for image classification. Locality-constrained linear coding approach instead of traditional vector quantization is adopted. And then, we apply probabilistic Latent Semantic Analysis (pLSA) to a BOW representation for each image to obtain more compact and semantic features. Finally, scene images are classified by support vector machine with these features. Experimental results show that our method has satisfactory classification performances on a popular dataset.

Keywords: scene image classification, bag-of-visual-words, probabilistic Latent Semantic Analysis

1. Introduction

Recent years have witnessed an explosion of work in the area of scene image classification and object recognition. Many multimedia and computer vision applications, such as image retrieval, automatic color correction, video surveillance and robotics path planning, are based on successful scene image classification. It is evident that the performance of these algorithms is directly linked to the performance of the classification. The goal of scene image categorization is to classify a collection of unlabeled images into a set of predefined classes according to semantic meanings, which is a fundamental problem in the fields of computer vision and image understanding. Though extensive researches have been made to improve the categorization accuracy, it is still a challenging task given the presence of intra-class variation, lighting changes, view angles, occlusion, background clutter, and complex spatial variations.

Bag-of-visual-Words (BoW) strategy [1, 2, 14] is one of the most popular and effective method in scene image classification recently. The idea of BoW is inspired by the success of textual words in natural language processing (NLP), where documents are represented by a bag of textual words from a constructed dictionary. Analogously, to represent an image using BoW model, an image is treated as a collection of unordered appearance descriptors extracted from local patches, and then these patches are quantized into discrete “visual words” by the pre-generated codebooks, and then the image can be represented with a compact histogram which are employed as the feature for classification. Despite BOW model achieves great success and wide adoption, however, it is still suffering the following main drawbacks [2]. 1) It ignores the spatial relationships between the local features, and then missing some discriminative power since the spatial layout of the features may be almost as important as the features themselves. 2) The code words quantization process usually results large quantization errors, thus, the generated visual words might be not semantically reasonable, and many noisy visual words can be generated.

Many works have been proposed to analyze these shortages and improve the descriptive power of visual words. For the direction of spatial information utilization, Lazebnik et al. [1] presented a Spatial Pyramid Matching (SPM) based method to exploit the spatial information, where the distribution of the visual words were calculated at multi-spatial resolutions to form a spatial pyramid representation, and a SPM kernel was used to measure the similarity between pyramids. Although it is easy and simple, it has been demonstrated to be very effective. [3] proposed a spatial string matching method, which represented an image as an unordered distribution of local features under some spatial constraints. Further, Gemert et al. [4] investigated different soft weighting schemes in assigning visual words to BoW representations to capture spatial relationships. Moreover, Visual phrase was proposed in [5] as descriptive concurrent patterns of visual words, which captured the spatial information among visual
words and presented better discriminative ability. Generally, group the visual words into meaningful phrases could capture the spatial configuration among them, and afford more discriminative ability than the classical visual words in image classification tasks.

For the improvement of codebook quantization stage, researchers have spent more effort to design more powerful quantizer to reduce quantization errors and preserve more information of the feature descriptor. Local coding [6] and sparse coding [7, 8] emerged recently as effective alternatives to the previous vector quantization approach. These methods optimize a linear combination of few visual words to approximate a local feature and code it with the optimized coefficients. For example, Wang et al. [6] employed Locality-constrained Linear Coding (LLC) to replace the VQ coding in traditional SPM, which is simple but effective. In detail, they utilized the locality constraints to project each descriptor into its local-coordinate system, and the projected coordinates are integrated by max pooling to generate the final representation. Yang et al. applied sparse coding to replace $k$-means in the SPM framework, and then improved it with supervised hierarchical sparse coding model [7], achieving state-of-the-art performances on several benchmarks. Further, Coates [8] investigated the difference between sparse coding and vector quantization through the training phrase (vocabulary construction) and the encoding phrase (feature quantization) respectively, and argued that it is more important to choose a good quantizer than building an exquisitely-designed vocabulary. And the authors also pointed out that the main strength of sparse coding appears to arise from its non-linear encoding scheme.

In this paper, we aim to the effective combination of the steps of spatial information and codebook generation. The contributions are threefold. First, we propose a novel biased spatial block information utilization approach to incorporate spatial information into the BOW model. Second, we combine the Locality-constrained Linear Coding approach and the bias spatial information utilization in BOW model, which makes the generated visual words more semantic. Third, we apply probabilistic Latent Semantic Analysis (pLSA) to a BOW representation for each image to obtain more compact and semantic features. Experimental results show that our method can achieve satisfactory classification accuracy on popular dataset.

The rest of the paper is organized as follows. In section 2, we present the framework of our proposed algorithm and the detailed implementation information. And the experimental results are provided in section 3. Section 4 concludes our paper.

2. The proposed framework

Figure 1 illustrates the overall framework of the proposed method. Image patches are first sampled in a dense way, and then a biased spatial division method is applied to the image to stress the importance of center regions. Next, the patches are described by lots of local Scale-Invariant Feature Transform (SIFT) descriptors [9], which construct a codebook dictionary with a linear coding approach. In this way, all the training and test images can be represented by a histogram of visual words by mapping the local features of an image back into the codebook dictionary. And then, probabilistic Latent Semantic Analysis is adopted to obtain more compact and semantic features. Finally, support vector machines are used as classifiers to classify each image into its category. In the following, we will give a detailed description for the stages of image division, codebook generation, pLSA model and the classifiers.

2.1 Image division

We divide each image into four overlapped blocks ($f_1$~$f_4$), as can be seen from figure 2. The width and length of each block is 0.75 times of the original one, then, the center region ($f_0$) is considered by four times. The aim of such a biased spatial block mechanism is to highlight the importance of center region. This kind of mechanism plays two roles in our framework.

On one hand, it is used for enhancing the number of center region descriptors in an image. In image description step, we first employ 128-dimensional SIFT descriptors in each of the four blocks, and then unite them together as a whole. We can notice that the center region is described by four times, while the other regions are only described once. In such a biased treatment way, both the center and the other regions are all used, however, the importance of center region is highlighted. This is because that the object information is usually in the center
of an image, which has the most discriminative nature to represent an image. It should be noticed that more complicated methods such as object detection and saliency detection can also be employed to locate the object regions. However, in our daily life, we are inclined to put the interest object in the center of a photo or picture. Thus, it is a reasonable but efficient shortcut to locate object region with our biased block division method. And we also require the background information, since they may provide useful contextual clues. This is why we do not get rid of these regions, but utilizes them in a subordinate way.

On the other hand, the biased spatial block mechanism is employed in the histogram generation stage. Each block is represented by an $N$-dimensional histogram of visual words, and the final BOW feature is the combination of the four blocks, that is a $4 \times N$-dimensional vector. The idea behind this operation is similar with the above one, which is to stress the importance of the center object region.

Figure 1. Flowchart of the proposed scene image classification approach

Figure 2. Biased spatial block division

2.2 Locality-constrained Linear coding
In the codebook construction step, a codebook is generated by using unsupervised \( k \)-means method. And as indicated by [8], it is more important to design a quantizer than a codebook. In this section, we use Locality-constrained Linear Coding (LLC) to replace the traditional vector quantization (VQ) coding method. Let \( X = [x_1, x_2, \ldots, x_N] \) be a set of local descriptors extracted from an image, \( B = [b_1, b_2, \ldots, b_N] \) be the codebook, then VQ minimizes the variance between the clusters and the descriptors and maps the descriptors \( (x_i) \) into a new feature space. It solves the following constrained least square fitting problem:

\[
\begin{align*}
\arg\min_C \sum_{i=1}^{N} \| x_i - B_c \|^2 \\
\text{s.t.} \| c_i \|_0 = 1, \| c_i \|_1 = 1, c_i \geq 0, \forall i
\end{align*}
\]

(1)

where \( C = [c_1, c_2, \ldots, c_N] \) is the set of codes for \( X \). The cardinality constrain \( \| c_i \|_0 = 1 \) means that there will be only one non-zero element in each code \( c_i \), corresponding to the quantization id of \( x_i \). The non-negative \( \ell_1 \) constraint \( \| c_i \|_1 = 1, c_i \geq 0 \) means that the coding weight for \( x \) is 1. in practice, the single non-zero element is found by searching the nearest neighbor.

This method is simple and easy implementation, however, it tends to allot more clusters to high frequency occurring features, which may not informative enough. Sparse coding improves VQ by relaxing the restrictive cardinality constraint \( \| c_i \|_0 = 1 \) and abandoning the non-negative constraint \( \| c_i \|_1 = 1, c_i \geq 0 \) in Eq.(1).

\[
\begin{align*}
\arg\min_C \sum_{i=1}^{N} \| x_i - B_c \|^2 + \lambda \| c_i \|_1
\end{align*}
\]

(2)

Further, to improve the scalability, LLC was presented in [6] to incorporate locality constraint instead of the sparsity constraint in Eq.(2), since locality must lead to sparsity but not necessary vice versa. The criteria used in LLC are as follows:

\[
\min_C \sum_{i=1}^{N} \| x_i - B_c \|^2 + \lambda \| d_i \|_1 c_i^T
\]

\[
\text{s.t.} \| c_i \|_1 = 1, \forall i
\]

(3)

where \( \varnothing \) represents the element-wise multiplication, and \( d_i \in R^v \) is the locality adaptor that gives different freedom for each basis vector proportional to its similarity to the input descriptor \( x_i \). LLC has some meaningful properties of better reconstruction, local smooth sparsity, analytical solution and fast implementation, thus it is employed in our approach to replace the VQ method.

### 2.3 Probabilistic latent semantic analysis model

With LLC, codebook dictionary can be generated and images can be transformed as Bag-of-visual Words (BoW) representation. To get a more compact and discriminative representation, we apply probabilistic Latent Semantic Analysis (pLSA) to BoW features, which had been demonstrated effective in literatures [10, 11, 15]. pLSA is a statistical technique for the analysis of two-mode and co-occurrence data.

Let \( W = \{w_i\}_{i=1}^V \) be an image feature space, and \( D = \{d_j\}_{j=1}^N \) be the image data set. We can estimate an image in a \( V \times N \) co-occurrence matrix \( M_{V \times N} \in R^{V \times N} \), where each element \( M_{ij} \) is the frequency of features \( w_i \) appearing in an image \( d_j \). Let \( Z = \{z_j\}_{j=1}^Z \) be the latent variable set in our pLSA model. Each latent variable \( z_j \in Z \) corresponds to a certain cluster. A joint probability model \( P(w, d) \) over \( V \times N \) is defined by the mixture:
where $P(w|d)$ is the topic specific distribution and $P(z|d)$ is a mixture of topics for each image.

### 2.4 Non-linear SVM classifiers

Once images are represented by vector features, we use support vector machines (SVM) for classification. The key idea is to separate two classes with an optimal decision hyper-plane that has maximum margin using the training samples. For two-class SVM, the decision function for a test sample $x$ has the following form:

$$g(x) = \sum_i \alpha_i y_i \kappa(x_i, x) - b$$

where $\kappa(x_i, x)$ is the value of a kernel function for the training sample $x_i$ and the test sample $x$, which measures the similarity between the two data samples; $y_i \in \{-1, +1\}$ is the class label of $x_i$, $\alpha_i$ is the learned weight of the training sample $x_i$, and $b$ is a learned threshold. The training samples with $\alpha_i > 0$ are called support vectors.

For multi-class classification applications, multi-class SVM can be adopted by training several classifiers and combining their results. The dominant approach for doing so is to reduce the single multi-class problem into multiple binary classification problems, of which the one-versus-all rule is one of the most popular one. It employs a winner-takes-all strategy, in which a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

### 3. Experimental results

We report our results based on Caltech 101, which is widely used in scene image classification community. Caltech101 contains 9144 images in total from 102 different categories, including 101 object categories and one additional background category. The number of images per category ranges from 31 to 800. For each image, it contains only one single object, and its resolution is about $300\times 200$ pixels. In the experiments, we follow the dense sampling, in which an evenly sampled grid spaced at $10 \times 10$ pixels is adopted for a given image, and the size of the patch is randomly sampled between scale 10 to 30 pixels. The size of codebook $N$ is 1000, and so the final dimension of BOW is 4000. The topic number $Z$ is 80.

We follow the usual approach for evaluation [12]. For each category, we randomly select 5, 15, 25 and 30 images respectively for training and up to 25 images in a disjoint set for test. Each test image is assigned a predicted label, and mean classification rate is the average of the diagonal elements of the confusion matrix. Table 1 shows the performance comparisons with other methods reported on the dataset. Because many researchers have reported their results on Caltech-101, we directly compare our algorithm to the existing ones. From this table we can see that our method outperform methods using spatial pyramid BoW [12], multi-feature fusion model [13], LLC method [6] and SP-pLSA method [11]. Figure 3 shows the confusion matrix with the training number of 15.
Table 1. The performance comparison

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train_num</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>KSPM [12]</td>
<td>44.2</td>
</tr>
<tr>
<td>Multi-Feature Fusion [13]</td>
<td>45.7</td>
</tr>
<tr>
<td>LLC [6]</td>
<td>51.2</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>51.9</strong></td>
</tr>
</tbody>
</table>

Figure 3. Confusion matrix

4. Conclusions

In this paper, we presented a novel scene image classification approach in the framework of BOW. By dividing the image in a biased way, we can sample and represent the image with the highlight of center region, which not only employs the contextual information between the object and background regions, but stresses the importance of object information. To overcome the drawback of VQ method, we adopted Locality-constrained Linear Coding in the codebook construction stage. Moreover, we also introduce probabilistic Latent Semantic Analysis model to compact the representation of BOW. Experiment results show promising performance for image classification task. In the future work, we will further address the effective utilization of contextual information by employing some object detection or saliency detection methods [16].
Acknowledgment

This paper is supported by Innovation Foundation of Tianjin University (No.60302019), Tianjin Municipal Natural Science Foundation (No.09JCYBJC00900), and National Natural Science Foundation of China (No. 61170239)

References