EXPANDED BAG OF WORDS REPRESENTATION FOR OBJECT CLASSIFICATION

Tinglin Liu, Jing Liu, Qinshan Liu, Hanqing Lu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Science, Beijing 100190, China
{tlliu, jliu, qsliu, luhq}@nlpr.ia.ac.cn

ABSTRACT

Currently, the bag of visual words (BOW) representation has received wide applications in object categorization. However, the BOW representation ignores the dependency relationship among visual words, which could provide informative knowledge to understand an image. In this paper, we first design a simple method to discover this dependency through computing the spatial correlation between visual words in overlapped local patches. Obtaining the dependency relationship, we further propose a novel update strategy to modify the BOW representation. The modification is motivated by the idea of Query Expansion applied successfully in text retrieval. We implement our approach on challenging PASCAL 2006 database, and the experimental results show its improved performance against the BOW representation.

Index Terms—object classification, bag of words, spatial correlation, query expansion

1. INTRODUCTION

Object categorization is a fundamental task in computer vision, and it has been extensively investigated these years. The bottleneck of the task is the lack of an understandable image representation for machines. Due to the presence of large variations in appearance, such as scale, illumination, pose, and background clutter, how to present a reasonable representation is still an open problem.

Currently, the popular and efficient bag of visual words (BOW) image representation approaches for object categorization adopted the statistics of local appearances, their idea is to quantify the continuous high-dimensional space of image local features, such as SIFT [1], to form a finite clusters, which are described as vocabulary of 'visual words'. According to this vocabulary, each feature extracted from a new image can be mapped into its closest visual word, and then the image is represented as a histogram over the vocabulary of visual words. The visual vocabulary could provide a 'mid-level' representation, which helps to bridge the semantic gap between the low-level visual features and the high-level concepts to be categorized [2].

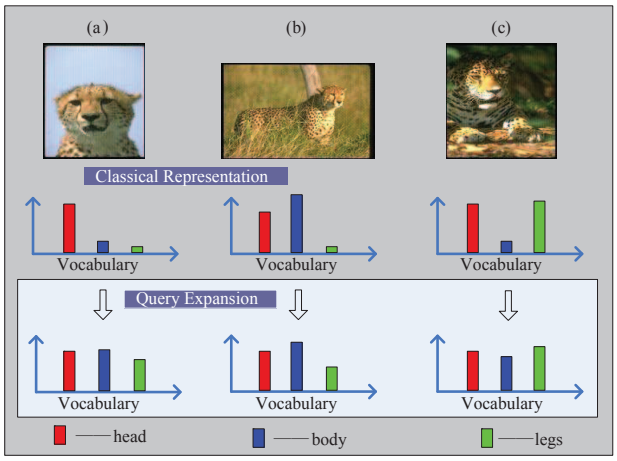

Fig 1: Illustration of the comparison between the classical BOW representation (shown in the second row) and the updated representation after Query Expansion (shown in the third row).

However, the classical BOW representation assumes that the local features within each image are independent to each other and purely explores the statistics of individual visual words. Actually, the dependency among visual words is valuable to understand an image, since the discriminative items to recognize an object are several associated visual words, not a set of independent ones. From this perspective, various methods were proposed, such as Visual Phrase [3], Visual Synset [4], and local shape correspondences [5]. However, the high computational complexity makes them difficult to apply.

In this paper, we design a simple and computationally efficient method to computer the spatial correlation between any two visual words. Next, we explore the obtained contextual information to update the classical BOW representation. The basic idea behind our approach is inspired by the Query Expansion method used successfully in text retrieval [6], which expands a seed query based on the semantic relationship among terms to improve the retrieval performance. For example, given a query of 'Jaguar xk', the search engine ought to catch a more precise description through expanding the query with 'Jaguar xk

car' and achieves more precise results accordingly. The similar process would be suitable in the image domain. Specially, when we view an image as a histogram of visual words, the expansion by the visual word correlations would bring more robust performance. To be clarity, we take an example of the 'leopard' images as illustrated in Fig. 1. First, we assume that an ideal vocabulary for the BOW representation is obtained, in which each visual word has specific semantics to indicate a body part of leopard. The classical BOW representation would have poor discriminative ability because it suffers from the lack of certain visual words brought by the large variance of leopard appearance. Nevertheless, if we explore the dependency among the visual words to modify it, the missing visual words can be offset. Obviously, the subsequent categorization will benefit from it.

The rest of this paper is organized as follows. The detailed procedure of discovering the dependency among visual words is described in Section 2, and the Section 3 will present the proposed updated strategy. Experiments and conclusions are reported in Section 4 and Section 5 respectively.

2. DISCOVER THE DEPENDENCY

Let $\Omega=\{W_1,W_2,...,W_m\}$ denotes the vocabulary generated by an unsupervised clustering method, such as the k-means. Then the classical BOW model represents an image as a geometric-free unordered set of these visual words, which discards all spatial information. However, the contextual information could be very useful. For instance, the foot pedal of bicycle always appears with the wheels in the neighborhood, the eyes always with the nose, and the tires of cars always with the road, etc. The first phase of our approach focuses on discovering the dependency relationship among all the visual words through exploiting co-occurrence information in spatial domain.

A reasonable assumption that most of the related visual words would appear in the neighborhood is given. We place a sequence of windows sliding over the two-dimensional image space to extract local patches. The occurrence of every visual word is counted in each local patch, and then each visual word i corresponds to a L-dimensional vector H_i, where $H_i(l)$ is the number of visual word i that falls into the *l'th* local patch and L is the total number of local patches. Obtaining the occurrence, the co-occurrence between any two visual words can be calculated by the histogram intersection function, which sums up the matches at all local patches:

$$S(i,j)=\sum_{l=1}^{L}\min(H_i(l),H_j(l)) \tag{1}$$

where $S(i,j)$ denotes the co-occurrence between the visual words of i and j.

We make the windows slide with the 50 percent overlap so as to keep the matching more robust between those visual words at the edge of the sliding window. Also in this way those matches in a closer neighborhood will be given more weights, since such matches repeatedly appear in many local patches and they are recounted many times as the window sliding.

Based on the co-occurrence relationship, we further define the dependency as follows:

$$D(i,j)=\frac{S(i,j)}{(S(i)+S(j))/2} \tag{2}$$

where

$$S(i)=\sum_{l=1}^{L}H_i(l) \tag{3}$$

Essentially, the process as in Eq. (2) is an implementation of normalization. Similar to the TF-IDF idea in text retrieval, the word with high occurrence in all local patches will be brought negative effect to the corresponding co-occurrence calculation.

Note that the dependency relationship is computed over images in one category. This is because the same visual words could render different dependency relationship in different categories. We can easily understand this through the following example. Assuming that the visual word 'A' represents visually similar tires, in bicycle images it may have high dependency with the visual word 'B' representing the foot pedal; however in motorbike images it would be relative to the visual word 'C' representing vent-pipe.

3. QUERY EXPANSION

For the classical BOW representation, an image can be described as a histogram of occurrence of visual words over the global vocabulary. Let $F_c=[f(1),f(2),...,f(m)]$ denotes the classical representation of the c^{th} image, where $f(i)$ shows the weight of the i^{th} visual word in the image. This representation gives the visual word different weights simply according to their respective amount in an image, without taking their mutual dependency into account. The second phase of our approach is to update the classical representation using the dependency relationship among all the visual words.

3.1. The expanded representation

The update strategy is inspired by the idea of Query Expansion successfully used in text domain, which could eliminate the mismatch through formulating the query to more related terms, and could provide more precise information for the effective search. The expansion is also suitable to the object recognition based on the BOW representation because the terms in a document are analogous to the visual words in an image. Our approach aims to expand each visual word with its

correlated neighbors to obtain a refined representation of image appearance. As illustrated in Fig 1, the visual words standing for the head, body and legs of a leopard are highly correlated. When we use image (a) as a query, in which the head of leopard occupies most room but no body and legs appear, to search the 'leopard' images, the better performance can be achieved through expanding the BOW representation with the visual words standing for the body and legs of a leopard. This is also true for the application to recognize the object in an image.

Specifically, for each visual word *i* in an image, we would enhance the weights of its related ones. Supposing the visual word *j* is relevant to the visual word *i*, then during the expansion of visual word *i*, the weight of *j* is updated:

$$f^*(j) = f(j) + \alpha \cdot D(i,j) \cdot f(i) \quad (4)$$

where α is a factor that controls the extent of the expansion. Only the top *n* related ones could get enhanced during the expansion of each visual word, for it could not be relative to all of the others. Besides, the updated representation is obtained after the expansion of each of the total *m* visual words.

As shown in equation (4), the extent of expansion is charged by the factor α and the dependency between the visual words. As the dependency relationship is diverse over all categories, the value of α should also be determined per category. Specially, when $\alpha = 0$, this approach would reduce to the classical BOW representation.

Note that the final representation of our approach also involves normalization. Easily to understand, in an image the visual words frequently appearing could certainly indicate some image content. Therefore, after the overall expansion the visual words related with such highly-weighted ones will get extended; otherwise, the visual words inconsistent with the image content will be penalized for it receives relatively less expansion from others.

3.2. Classification strategy

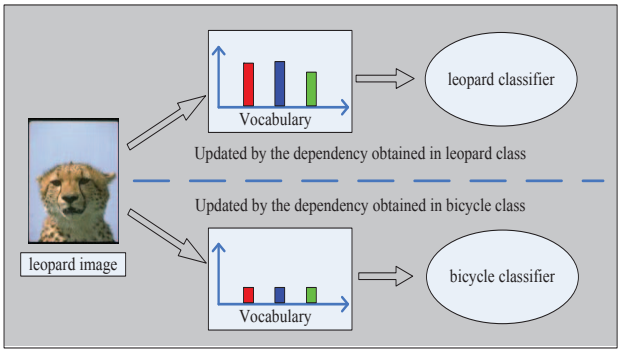

Fig 2: Generate one expanded representation per category. Each representation is subsequently fed to a respective classifier.

As discussed above, dependency relationships among visual words within different categories are different. Therefore, the expanded representations of each image, which is built on the different dependency relationships, are varying. To classify various images, we attempt to learn a SVM classifier per category, which is trained in a one-vs-all manner. For training and testing, each classifier is fed with corresponding expanded representation computed on the dependency in the respective category. Also taking the leopard for instance, the classifiers are trained within each category, suppose these are 'leopard' classifier and 'bicycle' classifier. During testing period, the representation updated on dependency in leopard category will be fed to the 'leopard' classifier; meanwhile, the representation updated on the dependency in bicycle category will be fed to the corresponding 'bicycle' classifier, as shown in Fig 2.

4. EXPERIMENT

We evaluate our approach on the PASCAL VOC 2006 data set. In the challenging database ten annotated object classes are provided, which are bicycle, bus, car, cat, cow, dog, horse, motorbike, person and sheep respectively. As a multi-object classification task, for each of the ten object categories, the goal is to predict the presence/absence of at least one object given a test image. The binary classification performance for each object category is measured quantitatively by the area under the ROC curve (AUC) as used in [7].

4.1. Parameter Settings

We randomly choose 30 training images per category to build the training database. The AUC is computed based on the prediction for the 500 randomly selected PASCAL testing images. Each experiment is repeated ten times, and the reported results are the average and standard deviation of the AUC results over these trials.

For image features, we randomly selected 1000 local regions with different scales from an image, in conjunction with the SIFT descriptor to extract low-level features from every local regions [8].

The baseline method is a very standard BOW model [2], which constructs a visual vocabulary using k-means with 1000 cluster centers. Thus there are 1000 visual words, and each low-level feature is hard-assigned to its nearest visual word; e.g. an image is represented by a 1000-bin histogram over the vocabulary. The RBF kernel is used for the SVM classifiers.

In our approach, there are three parameters: the number of sliding windows, determining the size of local patches, is set to be 12*12. If the size of image is 600*400, then the local patches are equally 92*63; number *n*, determining how many related ones each visual word has to expand with, is set to be 5; and α in equation (4), determining the extent of

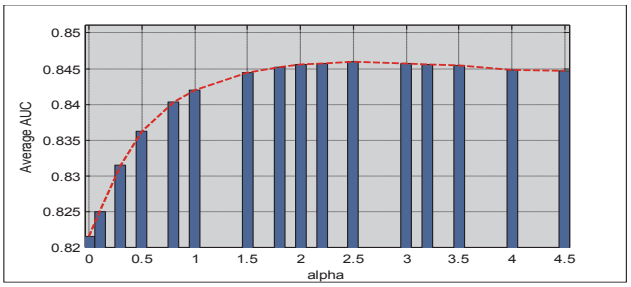

Fig 3: Average AUC of the 'sheep' classifier with different value of α .

expansion, is set to be [1, 1, 0.3, 4.5, 0.1, 1, 4.5, 4.5, 1.8, 2.5] per category.

Among these three parameters, in opposite to the first two, the value of α determines, to a great extent, the final results. We perform different values of α in each category and the optimal one is selected finally for each category. Fig 3 shows the change of the average AUC with the different values of α in the sheep category, and we can see α equal to 2.5 is the best.

4.2. Results

Table 1 summarizes the AUC results of our method against classical BOW representation. Our method demonstrates improvements over the classical BOW representation on every category. Especially, large improvement is made in the 'horse' category, whose AUC jumps from 0.603 to 0.670, and some efficient improvements on 'motorbike' from 0.737 to 0.782, 'dog' from 0.621 to 0.666, 'sheep' from 0.821 to 0.846, and 'cat' from 0.720 to 0.763. On 'cow' and 'bus', the baseline is outperformed slightly by the proposed approach.

In some categories, the improved results are still not as good as we expected, which may be due to the relatively scarce semantics in visual words generated on such challenging database. We also implement our approach on some style-uniform databases selected from Caltech 101, the improvement would be obvious. Anyway, the improved results on this challenging database could indicate the success of our expanded BOW representation.

Table 1: Average and standard deviation of AUC results on PASCAL 2006.

Class	Classical	Our method
bicycle	0.805 ± 0.023	0.830 ± 0.022
bus	0.746 ± 0.036	0.758 ± 0.040
car	0.812 ± 0.019	0.837 ± 0.017
cat	0.720 ± 0.019	0.763 ± 0.017
cow	0.871 ± 0.020	0.872 ± 0.020
dog	0.621 ± 0.032	0.666 ± 0.035
horse	0.603 ± 0.032	0.670 ± 0.024
motorbike	0.737 ± 0.021	0.782 ± 0.024
person	0.572 ± 0.031	0.601 ± 0.029
sheep	0.821 ± 0.022	0.846 ± 0.027

5. CONCLUSION

This paper proposed a simple method to calculate the spatial correlation between every two visual words, and designed a novel update strategy of expanding each visual word with its most related neighbors, inspired by the idea of Query Expansion. Obtaining the expanding representation, SVM is applied to classify different images from various categories. The experimental results on the PASCAL 2006 database demonstrate the effectiveness of the dependency information among visual words and we can also believe that the proposed solution for object categorization is an enhancement over the classical BOW models.

6. ACKOWLEDEGEMENT

The research was supported by National Natural Science Foundation of China (Grant No.60675003, 60723005, 60835002).

7. REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints" , In International Journal of Computer Vision, 2004.

[2] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, " Local features and kernels for classification of texture and object categories: A comprehensive study", In International Journal of Computer Vision, 2007.

[3] J. Yuan, Y. Wu, and M. Yang. "Discovery of collocation patterns: from visual words to visual phrases", In proceedings of the international conference on Knowledge discovery and data mining, 2007.

[4] Y. T. Zheng, M. Zhao, and S. Y. Neo. "Visual Synset: Towards a Higher-level Visual Representation", In In proceedings of Conference on Computer Vision and Pattern Recognition, 2008.

[5] T. B. A. C. Berg and J. Malik, "Shape matching and object recognition using low distortion correspondences", In Proceedings of Computer Vision and Pattern Recognition, 2005.

[6] Y. Qiu, and H. P. Frei, "Concept Based Query Expansion", In proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, 1993.

[7] Y. Liu, J. Rong, S. Rahul, and J. Frederic, "Unifying Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition", In proceedings of Conference on Computer Vision and Pattern Recognition, 2008.

[8] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification", In European Conference on Computer Vision, 2006.