Concept-Specific Visual Vocabulary Construction for Object Categorization

Chunjie Zhang, Jing Liu, Yi Ouyang, Hanqing Lu, and Songde Ma

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
{cjzhang,jliu,youyang,luhq}@nlpr.ia.ac.cn, mostma@gmail.com

Abstract. Recently, the bag-of-words (BOW) based image representation is getting popular in object categorization. However, there is no available visual vocabulary and it has to be learned. As to traditional learning methods, the vocabulary is constructed by exploring only one type of feature or simply concatenating all kinds of visual features into a long vector. Such constructions neglect distinct roles of different features on discriminating object categories. To address the problem, we propose a novel method to construct a concept-specific visual vocabulary. First, we extract various visual features from local image patches, and cluster them separately according to different features to generate an initial vocabulary. Second, we formulate the concept-specific visual words selection and object categorization into a boosting framework. Experimental results on PASCAL 2006 challenge data set demonstrate the encouraging performance of the proposed method.

Keywords: Visual vocabulary, object categorization, SIFT descriptor, k-means.

1 Introduction

Recently, a popular representation of image content for object categorization is the bag of words [1] (BOW) model, in which one image is represented by a histogram of the occurrences of visual words. The idea behind the BOW representation for object categorization is to quantize the continuous high-dimensional space of local image features (*e.g.*, SIFT [2] descriptors) to a vocabulary of "visual words". However, compared with textual document-categorization, there is no available vocabulary for image-based object categorization and it has to be learned from a training image set. Accordingly, how to construct a suitable visual vocabulary becomes an important task for object categorization.

Nowadays, the visual vocabulary is typically constructed by leveraging local image features (e.g., SIFT descriptors). Sivic and Zisserman [3] originally proposed to cluster the SIFT descriptors of local image patches with *k*-means algorithm and treated the center of each cluster as a visual word. Farquhar *et al.* [4] and Perronnin [5] proposed the Gaussian Mixture Model (GMM) to perform clustering with the SIFT descriptors of training images. Winn *et al.* [6] used textons of training images to

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 936–942, 2009.

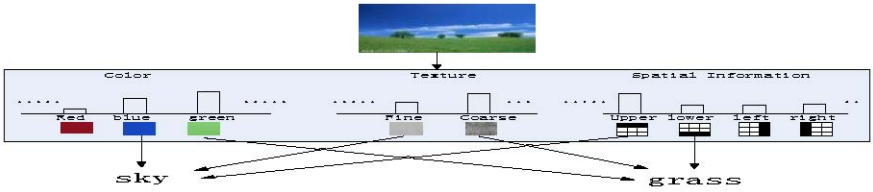

Fig. 1. A toy example. Visual words with explicit semantic meanings will help make correct categorization of images and different visual words should be chosen for different concepts.

generate an initial codebook, and the information bottleneck principle is then used to pair-wisely merge these visual words into a more discriminative codebook. Moosmann [7] proposed a fast codebook construction method which used different types of features, e.g., color descriptor, color wavelet descriptor and SIFT descriptor, but in a separate way. Hsu and Chang [8] concatenated different types of visual features into a high dimensional vector to represent each key frame of videos.

From above introduction, we can observe that most of previous works construct codebooks by considering only one type of feature or simply concatenating all kinds of visual features into a long vector. Obviously, it is insufficient to represent all concepts with only one type of feature. And, the concatenated representation maps images into a more complex space and usually cannot be explained in an explicit manner. In fact, any visual feature is extracted to reflect a specific visual property and can be explained semantically to some extent. For instance, the color histogram describes the color distribution in an image or a patch and the dominant colors ("red", "green", or "blue"…) can be deduced from the color feature. If visual words can be endowed with relatively explicit semantic meanings, the aggregated vocabulary will be helpful to discriminate various images. Besides, images from different categories are expected to have distinctive visual words distribution. Thus, a compact and discriminative BOW representation is necessary in the task of object categorization.

This can be easily understood from the toy example illustrated in Fig. 1. When color, texture and spatial features are clustered separately we will probably get visual words of "red", "green", and "blue" for color; "fine" and "coarse" for texture; and "upper", "left", "right", "lower" for spatial property. Then, the concept of "sky" can be clearly expressed with visual words of "blue" in color, "fine" in texture and "upper" in spatial property. Similarly, to the visual words of "grass", "green", "coarse" and "lower" are used. In contrast, it is hard to understand the semantic meanings of visual words generated by the concatenated features. Although visual words in reality may not be as meaningful as the toy example, we believe clustering different types of features separately will generate more semantic meaningful visual words than directly clustering the concatenated features.

The remainder of this paper is organized as follows. Section 2 shows the framework of the proposed concept-specific visual vocabulary construction method for object categorization. The details of concept-specific visual vocabulary construction and visual words selection are described in Section 3. We give the experimental results in Section 4.

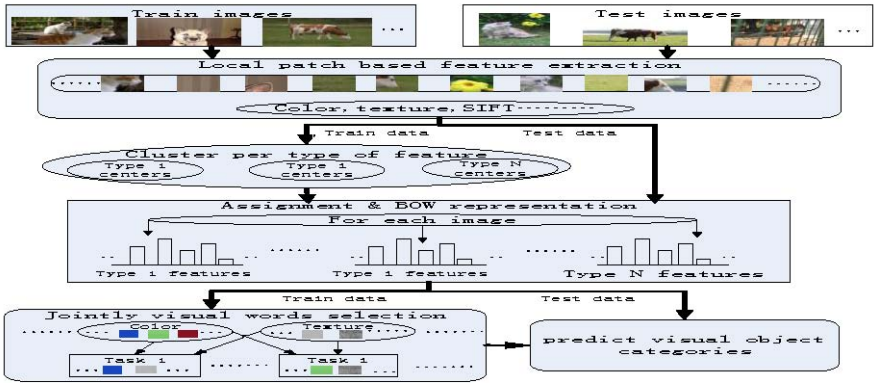

Fig. 2. Processing flow of the proposed concept-specific visual vocabulary construction method for object categorization

2 Overview of Concept-Specific Visual Vocabulary Construction Framework

Fig. 2 shows the major steps of our concept-specific visual vocabulary construction method for object categorization. There are mainly five components: extracting raw features, clustering each type of raw features, bag-of-words based image representation, selecting visual words, and making predictions.

We choose local patches of images and represent them with different types of features (e.g., color, texture, SIFT feature…). These raw features of different types are clustered separately using *k*-means. In this way, we try to take the advantages of different feature types and generate visual words with more explicit semantic meanings than that of clustering the concatenated features. As to the BOW representation of images, each type of features is assigned to the nearest visual words of the corresponding type. For each object categorization task, we choose some discriminative visual words under a boosting framework, and simultaneously the corresponding classifier to the object category is obtained. Finally, we make predictions of image categories.

3 Concept-Specific Visual Vocabulary Construction

3.1 General Visual Vocabulary Construction

Recently, local patch based image representation becomes popular. Typically, a local image patch is firstly identified and then SIFT descriptor is used to represent this local image patch. However, it is insufficient to represent all concepts with only one type of feature, and different types of features should be jointly considered. We adapt a unified process to both take the advantages of local patch based image representation

and leverage the representativeness of different types of features. For each image, we firstly choose local image patches and then extract different types of features from the local patches. Suppose we have extracted n types of features from the same local patch and denote them as $x_1, x_2, \dots, x_n$ respectively. The i-th type of feature x_i is of d_i dimensions. The concatenated feature is of size $\sum_{i=1}^{n} d_i$, and can be written as $x = \{x_{11}, x_{12}, \dots, x_{1d_1}, x_{21}, \dots, x_{2d_2}, \dots, x_{n1}, \dots, x_{nd_n}\}$.

We propose to cluster each type of features separately. Let $C_1, C_2, \dots, C_n$ denote the cluster centers of each type of features and the number of cluster centers are denoted as $l_1, l_2, \dots, l_n$ respectively. The total number of cluster centers is $\sum_{i=1}^{n} l_i$.

3.2 Choosing Optimal Visual Words for Object Categorization

After the visual vocabulary has been constructed, we can characterize an image by a histogram of visual words occurrences. Since our visual vocabulary is constructed by clustering different types of features separately, the resulting visual words represent different semantic meaning of images and should be treated separately; besides, the dimensions also vary from one type of visual words to another. Hence we adapt a different way and assign each type of features only to the corresponding nearest visual words of the same type.

We can concatenate these BOW representations of different types of features into a long vector to represent images. However, there are two drawbacks of this concatenated BOW representation. First, for a particular object categorization task, not all of the visual words are useful for object categorization. Second, the dimension of this concatenated representation will be high if we use many types of features. If we could choose some discriminative visual words for each object categorization task, we will be able to solve the two problems of the concatenated BOW representation of images.

We can choose each visual word independently, however, since visual words are correlated, the performance of choosing visual words independently will not be so good. Considering the correlations of visual words, we adapt to choose the next visual word by jointly considering the influences of previously chosen visual words.

Let $\{(x^j, y^j)\}_{j=1}^{m}$ be the set of training images, where each x^j is the concatenated BOW representation of images and is of dimension $\sum_{i=1}^{n} l_i$. y^j is the class label which belongs to a finite label space $\{-1,1\}$. Our aim is to choose the optimal visual words which can help make correct categorization of images. Besides, for different object categorization tasks, different visual words should be chosen.

Boosting with stumps [9] jointly considers the influences of different features by adapting a reweighting scheme and fits our problem well. For the BOW representation of images, constructing a stump learner on dimension k can be viewed as choosing the corresponding visual word k to represent images. By boosting stumps, we can consider the correlations of different visual words and choose the optimal visual words to represent images. Table 1 shows our jointly visual words selection algorithm. The corresponding classifier to the object category is simultaneously obtained. Different visual words will be chosen for different tasks.

Table 1. Jointly visual words selection by boosting with stumps algorithm

- Require: training examples$\{(x^j, y^j)\}_{j=1}^m$, initial weights of examples $w_1(j) = 1/m$ for all $j \in \{1,2,\dots,m\}$. P. Chosen visual words set $V = \emptyset$. $p = 0$.
- Do for $p = 1,2,\dots,P$:

  $p = p + 1$;

  1. Choose an optimal visual word v_p by training stump learner h_p with the weights $\{w_p(j)\}_{j=1}^m$.
  2. Calculate the error of choosing visual word v_p: $\varepsilon_p = \sum_{j:h_p(x_j)\neq y_j} w_p(j)$.
  3. Set $\beta_p = \frac{\varepsilon_p}{1-\varepsilon_p}$.
  4. Update weights of training examples: $w_{p+1}(j) = \frac{w_p(j)}{Z_p} \times \begin{cases} \beta_p & h_p(x_j) = y_j \\ 1 & otherwise \end{cases}$

     Where Z_p is normalization constant.
  5. Add v_p to V and exclude duplicate visual words.

  end for
- Output: the final chosen visual words set V.

4 Experiments

We evaluate the proposed visual vocabulary construction method on the PASCAL VOC Challenge 2006 data set [10]. Our training set (common across all methods) consists of all the PASCAL VOC Challenge 2006 training images. The AP is calculated based on the predictions of all the testing images provided by the data set. For image representation, we randomly select image patches from a pyramid with regular grids in position and densely sampled scales between 10 to 50 pixels. For each selected image patch, 36-dimensional color histograms, 144-dimensional color correlogram, 24-dimensional Polynomial Wavelet Tree (PWT), and 128-dimensional SIFT descriptors are extracted. Considering the diverse of images, we did not use the spatial information as shown in Figure 1 for robustness. For consistency, we ensure that the same set of low-level image features are used by all of the methods in our experiments.

To test the performance of the proposed method, we implement a few relevant methods. We show the performances of using only one type of features (abbreviated to Color, Corr*, Texture and SIFT respectively) and the concatenated features (abbreviated to Con[1]). Figure 3 shows the influences of visual vocabulary sizes of the four feature types and the concatenated features respectively. By jointly considering the vocabulary size and the average AP, we choose to cluster the color feature, the color correlogram feature, the texture feature, the SIFT feature and the concatenated feature into 500, 600, 600, 500 and 800 centers respectively. Gentle Adaboost [11] using stumps as weak learners are used for all methods and five-fold cross validation is adapted to find the optimal number of iterations. We use the same parameter settings for all methods for fair comparison.

Table 2 shows the average precision (AP) for these methods. For the four feature types and the concatenated features, we also give the results of using linear SVM classifiers. The optimal cluster center numbers (600, 900, 600, 600 and 1000 for the color, color correlogram, texture, SIFT and the concatenated features respectively) and parameter settings are found in the same way as the boosting framework does.

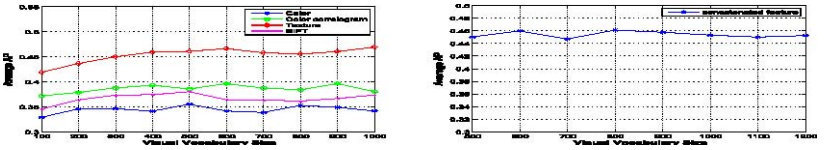

Fig. 3. The influences of visual vocabulary sizes for the four types of features and the concatenated features

Table 2. Average precision comparison on PASCAL 2006 Challenge among different representations (Corr* □color correlogram). "Con[1]" concatenates the four types of features and clusters them by *k*-means. "Con[2]" concatenates the BOW representations of the four features and trains SVM classifier to make predictions.

Class	Color		Texture		Corr*		SIFT		Con[1]		Con[2]	Ours
	Boost	SVM	Boost	SVM	Boost	SVM	Boost	SVM	Boost	SVM	(SVM)	
bicycle	0.287	0.327	0.529	0.544	0.349	0.454	0.507	0.562	0.430	0.561	0.589	**0.596**
bus	0.260	0.261	0.628	0.659	0.336	0.314	0.344	0.382	0.437	0.481	0.675	**0.685**
car	0.586	0.413	0.791	0.722	0.631	0.565	0.705	0.679	0.757	0.775	0.731	**0.812**
cat	0.336	0.288	0.431	0.432	0.357	0.353	0.399	0.398	0.379	0.428	0.382	**0.541**
cow	0.328	0.306	0.315	0.314	0.394	0.410	0.337	0.329	0.488	0.492	0.450	**0.524**
dog	0.263	0.232	0.306	0.298	0.275	0.284	0.251	0.251	0.308	0.303	**0.352**	0.343
horse	0.190	0.164	0.227	0.203	0.187	0.219	0.227	0.266	0.250	0.209	0.315	**0.328**
motor-bike	0.374	0.393	0.566	0.689	0.449	0.384	0.197	0.274	0.489	0.563	**0.604**	0.578
person	0.403	0.338	0.437	0.436	0.441	0.412	0.388	0.412	0.464	0.498	0.463	**0.506**
sheep	0.529	0.474	0.437	0.522	0.546	0.562	0.448	0.479	0.611	**0.614**	0.606	0.610
average	0.356	0.474	0.467	0.484	0.397	0.396	0.380	0.403	0.461	0.492	0.517	**0.552**

The "Con[2]" method concatenates the BOW representations of the four types of features into a 2700 dimensional vector to represent images and trains linear SVM classifiers to make predictions.

We can see from the results that for the four feature types alone, neither can performs better than the other three on the ten concepts. This shows that it is insufficient to represent images with only one feature type and different types of features should be combined. However, the performance of simply concatenating different types of features into a long vector is not so good. It is better to cluster different types of features separately to generate more semantic meaningful visual words. Moreover, by jointly choosing the optimal visual words for different tasks, we will be able to further improve the performances of object categorization.

We also show the histograms of chosen visual word numbers per feature type in Figure 4. The chosen visual words exhibit some similarities for concepts of the same category. For example, "cat", "cow", "horse", "sheep" are all animals; the chosen visual word histograms of these four concepts are very similar. The same thing also happens for the concepts of "bicycle", "bus", "car" and "motorbike", which all belong to vehicle.

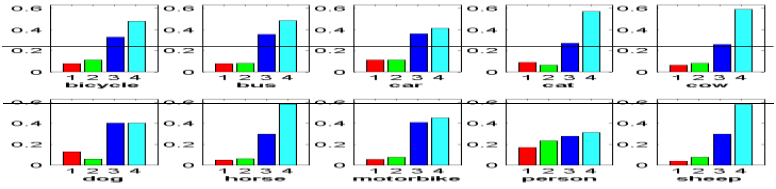

Fig. 4. The histograms of chosen visual word numbers per type of features on PASCAL VOC 2006 Challenge. The numbers of 1, 2, 3 and 4 stand for color feature, color correlogram feature, texture feature and SIFT feature respectively.

Acknowledgement

This work is supported by Natural Science Foundation of China (Grant No. 60835002, 60723005 and 60675003).

References

1. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1981)
2. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
3. Sivic, J.S., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: 9th IEEE International Conference on Computer Vision, Nice, pp. 1470–1477 (2003)
4. Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: Improving "bag-of-keypoints" Image Categorization: Generative Models and PDF-Kernels. Technical report, University of Southampton (2005)
5. Perronnin, F., Dance, C., Csurka, G., Bressan, M.: Adapted Vocabularies for Generic Visual Categorization. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 464–475. Springer, Heidelberg (2006)
6. Winn, K., Criminisi, A., Minka, T.: Object Categorization by Learned Universal Visual Dictionary. In: 10th IEEE International Conference on Computer Vision, Beijing, pp. 1800–1807 (2003)
7. Moosmann, F., Triggs, B., Jurie, F.: Fast Discriminative Visual Codebooks Using Randomized Clustering Forests. In: 20th Annual Conference on Neural Information Processing Systems, Hyatt, pp. 985–992 (2006)
8. Hsu, W.H., Chang, S.-F.: Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation. In: 4th International Conference on Image and Video Retrieval, Singapore, pp. 82–91 (2005)
9. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: 13th International Conference on Machine Learning, Italy, pp. 148–156 (1996)
10. Everingham, M., Zisserman, A., Williams, C., Gool, L.: The 2006 PASCAL visual object classes challenge (2006)
11. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: A Statistical View of Boosting. Annals of Statistics 28(2), 337–407 (2000)