CATEGORY SENSITIVE CODEBOOK CONSTRUCTION FOR OBJECT CATEGORY RECOGNITION

Chunjie Zhang[1], Jing Liu[1], Yi Ouyang[1], Qi Tian[2], Hanqing Lu[1], Songde Ma[1]

[1]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
[2]University of Texas at San Antonio, One UTSA Circle, San Antonio Texas, 78249-USA
qitian@cs.utsa.edu

ABSTRACT

Recently, the bag of visual words based image representation is getting popular in object category recognition. Since the codebook of the bag-of-words (BOW) based image representation approach is typically constructed by only measuring the visual similarity of local image features (e.g., k-means), the resulting codebooks may not capture the desired information for object category recognition. This paper proposes a novel optimization method for discriminative codebook construction that considers the category information of local image features as an additional term in traditional visual-similarity-only based codebook construction methods. The category sensitive codebook is constructed through solving an optimization problem. Therefore, the category sensitive codebook construction method goes one step beyond visual-similarity-only methods. Besides, the proposed category sensitive codebook construction method can be implemented with k-means clustering very efficiently and effectively. Experimental results on PASCAL VOC Challenge 2006 data set demonstrate the effectiveness of our method.

Index Terms—image processing, pattern recognition, image analysis, feature extraction, visual recognition

1. INTRODUCTION

Object category recognition is a pattern classification problem whose aim is to assign one or multiple labels to an image based on its semantic content. Recently, a popular representation of image content for object category recognition is the bag of visual words model. It was inspired by the *bag-of-words* approach to text-categorization [1]. A text document is viewed as a histogram of the number of word counts. Similarly, one image can be represented by a histogram of the number of visual words occurrences. These visual words provide a very efficient representation of images and help to bridge the semantic gap between the low-level image features and the high-level concepts to be classified. The idea behind the BOW representation for object category recognition is to quantize the continuous high-dimensional space of local image features (*e.g.*, SIFT descriptors [2]) to a vocabulary of "visual words". Several studies [3-8] have shown promising performance of this approach in object category recognition. However, compared with text-categorization, there is no given codebook for the visual categorization problem and the codebook has to be learned from a training set. Therefore, how to construct codebooks becomes an important problem for object category recognition.

Nowadays, the codebook for object category recognition is typically constructed by only considering the visual similarities of local image features. Sivic and Zisserman [3] originally proposed to cluster the local image features with *k*-means algorithm and treat the center of each cluster as a visual word. Hsu and Chang [4] and Winn *et al*. [5] use the information bottleneck principle to obtain more discriminative vocabularies. A compact visual codebook is learned by pair-wise merging of visual words of an initially large codebook. Since this merging process cannot be reversed, it is probably to be error propagation. To make use of the category information of image features, Farquhar *et al.* [6] and Perronnin [7] proposed the Gaussian Mixture Model (GMM) to perform clustering. [6] assigned a local image feature not to one visual word but to all visual words probabilistically which results in a continuous histogram representation of images. [7] considered the category information of local image features and represented an image by a set of histograms; each histogram describes whether the image content is best modeled by the universal vocabulary or the corresponding class vocabulary. However, the estimation of parameters for GMM is a challenging problem if the number of Gaussian distributions is large. Representing an image by a set of histograms also costs a lot of computation.

From above introduction, we can observe that most of previous works focus on finding more useful clustering methods or more suitable visual representation in order to get a suitable codebook including visually similar words. However, visually similar features may come from different categories of images and have different semantic meanings. Naturally, it is a better idea to generate a visual codebook with visual words consistent both on visual appearance and semantics. From this perspective, a novel method for the codebook construction is proposed in this paper. We incorporate the category information as an additional term into the traditional visual-similarity-only based codebook construction methods so as to generate more discriminative codebooks for object category recognition, just as [6, 7] did. The proposed method is very effective and can be combined with other codebook construction methods. Experimental results on PASCAL VOC Challenge 2006 data set [8] demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 describes how to construct category sensitive codebook by considering the category information of local image features. The category information serves as an additional term in traditional visual-similarity-only codebook construction methods. Experimental results are shown in Section 3. Finally, we conclude in Section 4.

2. CATEGORY SENSITIVE CODEBOOK CONSTRUCTION

In this section, we will first motivate the use of category information for discriminative codebook construction with a simple two-class problem and then propose our category sensitive codebook construction method.

2.1 Motivation

Let us first motivate the use of category information of local image features for the codebook construction with a simple problem of distinguishing dogs from cats.

Given some images mixed with “dogs” and “cats”, usual solution is to cluster visual features of some regions or points in these images in order to obtain a codebook for the BOW representation. For clarity, we assuming the obtained codebook is an ideal one whose items can reflect specific semantics such as “eye”, “ear” and “nose”. The visual words for ‘dogs’ and ‘cats’ images may be confused each other, since the two animals are similar on some body parts. That is, the words with different semantics are perhaps clustered into the same cluster. Accordingly, the BOW representations of images based on such a codebook are difficult to discriminate dogs from cats.

However, if we can utilize the semantic information as well as the visual appearance over local regions or points in images, a more discriminative codebook can be built, which includes some visual words consistent both on visual appearance and semantics. Ideally, “dog’s eye”, “dog’s ear”, “dog’s nose” and “cat’s eye”, “cat’s ear”, “cat’s nose” are given. Obviously, such visual words are more discriminative to classify dogs and cats.Although visual words may not be as meaningful as this simple two-class example, we believe the incorporation of semantic information as well as visual appearance into the codebook construction is potential to generate more discriminative codebook and hence bring better performance on the task of object category recognition.

2.2 Category Sensitive Codebook Construction

As the *k*-means approach has been widely applied in the codebook construction in the literatures, we will introduce our improvement on the typical codebook construction, in which category information as a kind of crude semantics is considered to build a more discriminative codebook. Note that our category sensitive codebook construction method can also be combined with other codebook construction methods besides the *k*-means approach. Firstly, we will show how the *k*-means clustering works [9] and then propose our category sensitive codebook construction method.

Suppose we have a data set $\{x_1,...,x_N\}$ consisting of N observations of a random D-dimensional Euclidean variable x. The goal is to partition the data set into K clusters. Let $\{\mu_k\}$ be a set of D-dimensional vectors where $k=1,...,K$, in which μ_k is a prototype associated with the k^{th} cluster. For each data point x_n, a corresponding indicator variable $r_{nk}\in\{0,1\}$ describes which of the K clusters the data point x_n is assigned to, so that if data point x_n is assigned to cluster k then $r_{nk}=1$ and $r_{nj}=0$ for $j\neq k$. The K-means clustering defines an objective function, given by

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\left\|x_n-\mu_k\right\|^2 \tag{1}$$

which represents the sum of the squares of the distances of each data point to its assigned vector μ_k. By minimizing J, *k*-means clustering finds the optimal $\{r_{nk}\}$ and $\{\mu_k\}$ as

$$r_{nk}=\begin{cases}1 & k=\arg\min_j\left\|x_n-\mu_k\right\|^2\\ 0 & otherwise\end{cases} \tag{2}$$

$$\mu_k=\frac{\sum_n r_{nk}x_n}{\sum_n r_{nk}} \tag{3}$$

Since the *k*-means clustering is unsupervised, the resulting visual words may not capture the desired information for object category recognition. This is just as the example of distinguishing dogs from cats given above.

We propose an improved formulation that regards the category information of local image features as an additional term in traditional visual-similarity-only based codebook construction methods (*e.g. k*-means). Our category sensitive codebook construction method combines the categories information and visual similarities among local image features and hence results in the more discriminative codebook than only considering the visual similarity.

Using the same symbols as above, we define the label y_i of low-level image feature x_i as the label of images from which classes this feature is extracted. y_i is a $D^{'}$ -dimensional vector of 0/1 where $D^{'}$ is the number of image categories. That is, if a low-level image feature is extracted from one category of images, the corresponding dimension of y_i will be assigned the label of 1 and the other dimensions of y_i will be assigned the label of 0. We define an objective function as

$$J^{'} = \sum Dis(x_i) + \sum \overline{\alpha} \times d(y_i) \qquad (4)$$

where the first item of Eq. 4 measures the visual dissimilarities of local image features and $d(y_i)$ measures the label dissimilarities. $\overline{\alpha}$ is a parameter vector adjusting the influence of label dissimilarities.

Note that the two items in Eq. 4 can be of any dissimilarity measures. Here, we adopt the most usual solution and utilize Euclidean distance to measure the distance or dissimilarity among visual features and label information. For the first item, we replace it with the typical objective function as in Eq. 1. Similarly, the label dissimilarity is calculated as the sum of Euclidean squared distance between the label of each data point and the label prototype of the cluster. Then we will get：

$$J^{'} = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|x_n - \mu_k\right\|^2 + \sum_{i=1}^{N}\sum_{j=1}^{K} \overline{\alpha}_j \times r_{ij}\left\|y_i - \overline{\pi_j}\right\|^2 \quad (5)$$

where $\overline{\pi_j}$ is the label prototype associated with the j^{th} cluster

In this way, we can implement the proposed category sensitive codebook construction method in a very efficient way. First, we extend the D -dimensional variable x with extra dimensions of category information y multiplied by a parameter $\sqrt{\overline{\alpha}_i}$. For simplicity, we set the parameters $\overline{\alpha}_i$ to be equal, and this yields a $D + D^{'}$ -dimensional variable $\overline{x}$; Eq. 5 then equals

$$J^{'} = \sum_{n=1}^{N}\sum_{k=1}^{K} \overline{r}_{nk}\left\|\overline{x}_n - \overline{\mu}_k\right\|^2 \qquad (6)$$

where $\overline{\mu}_k$ is a $D + D^{'}$ -dimensional vector. By minimizing $J^{'}$, the optimal $\{\overline{r}_{nk}\}$ and $\{\overline{\mu}_k\}$ are found as

$$r_{nk} = \begin{cases} 1 & k = \arg\min_j \left\|\overline{x}_n - \overline{\mu}_k\right\|^2 \\ 0 & otherwise \end{cases} \qquad (7)$$

$$\overline{\mu}_k = \frac{\sum_n r_{nk}\overline{x}_n}{\sum_n r_{nk}} \qquad (8)$$

The first *D*-dimensional of $\overline{\mu}_k$ is used to form the codebook. In this way, the proposed category sensitive codebook construction method goes one step further beyond the traditional visual-similarity-only based codebook construction methods by considering the category information of local image features.

If we replace the first item of Eq. 4 with different dissimilarity measures of local image features, we will get different kinds of codebook construction methods. The visual-similarity-only based codebook construction methods can be viewed as a special case of this method where no category information of local image features is used.

The advantages of the proposed categorization sensitive codebook construction method lie in two aspects. Firstly, it considers the category information of local image features as a complement to improve the traditional visual-similarity-only based methods and can yield more discriminative codebooks. This helps especially when two clusters of visual words from different categories lie nearby but are distinct. Secondly, by choosing proper label dissimilarity measure and clustering method, the proposed method can be implemented very efficiently.

3. EXPERIMENTS

We evaluate the proposed category sensitive codebook construction method on the PASCAL VOC Challenge 2006 data set [9]. This challenging dataset contains 5304 images with 9507 annotated objects. Ten annotated object classes are provided: bicycle, bus, car, motorbike, cat, cow, dog, horse, sheep and person. The training/validation and test sets are well balanced. For the classification task, the goal is to predict the presence or absence of at least one object of that class in a test image for each of the ten object classes. The area under the ROC curve (AUR) is adopted as a quantitative measure of the binary classification performance for each object class.

Our evaluation focuses on the problem of object category recognition given a limited number of training images. Our training set (common across all methods) consists of 10 randomly selected images per concept and we exclude images that are labeled with more than one concept (for example, an image can be labeled with “person” and “motorbike” simultaneously) in the training process. The AUR is calculated based on the prediction of 50 randomly selected testing images per concept. We repeated the experiment for ten times, and the averaged AUR over these trials is reported.

Table 1. AUR on PASCAL 2006 with 10 training images and 50 testing images per class. Numerical values in this table denote mean ± standard deviation.

Class	k-means	ACAC	Ours
bicycle	0.643 ± 0.058	0.632 ± 0.073	**0.657 ± 0.059**
bus	0.686 ± 0.058	0.684 ± 0.066	**0.709 ± 0.062**
car	0.741 ± 0.039	0.735 ± 0.028	**0.752 ± 0.043**
cat	0.690 ± 0.051	0.691 ± 0.053	**0.723 ± 0.052**
cow	0.701 ± 0.067	0.722 ± 0.056	**0.723 ± 0.045**
dog	0.604 ± 0.054	0.603 ± 0.057	**0.623 ± 0.083**
horse	0.536 ± 0.030	**0.575 ± 0.046**	0.537 ± 0.046
motorbike	0.586 ± 0.066	0.597 ± 0.051	**0.625 ± 0.048**
person	0.579 ± 0.052	**0.582 ± 0.039**	0.581 ± 0.043
sheep	0.711 ± 0.063	0.697 ± 0.061	**0.715 ± 0.076**

For image representation, we randomly select image patches from a pyramid with regular grids in position and densely sampled scales. SIFT descriptor [2] is used to describe these local image patches. We ensure that the same set of local image features are used by all of the three methods in our experiments for consistency.

To test the performance of the proposed method, we implement two relevant methods. The one is a standard method that constructs visual codebook using *k*-means as in [3]. The other agglomerates each class-codebooks into an all-class associated codebook, abbreviated to ACAC. The codebook size of the two baseline methods and our method are all set to 800 for fair comparison. Each low-level image feature is quantized to its closest center, and each image is represented as an 800-bin histogram over these centers. One vs. all SVM classifier with RBF kernel is used for the two baseline methods and our category sensitive codebook construction method. Five-fold cross validation is used to find the optimal parameters. The α in Eq. 5 is empirically set to 0.2 in our experiment.

Table 1 shows the AUR results for the three methods. Except “horse” and “person”, our method outperforms the two baseline methods. This is because the *k*-means method only utilizes visual similarities and does not consider the semantic information of the local features while the ACAC method constructs the class-codebooks in a separate way: each class-codebook is generated without considering the information of other classes. Our category sensitive codebook construction method jointly considers the category information and visual dissimilarity of local image features. The effectiveness of our method is demonstrated by the experimental results. Besides, the results for “horse” and “person” are not so good, that is because the label assignment of our method is somewhat contaminated with noise. However, the adding of category information of low-level image features does help to construct more discriminative codebooks.

4. CONCLUSION

This paper proposes a novel optimization method that regards the category information of local image features as an additional term in traditional visual-similarity-only based codebook construction methods. By choosing proper label dissimilarity measure, the proposed category sensitive codebook construction method can be implemented very efficiently and effectively.

Our future work will focus on how to assign the label to low-level image features more effectively. The influence of different codebook sizes and soft visual word assignment methods will also be considered.

5. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation of China (Grant No. 60675003, 6072300 and 560835002).

6. REFERENCES

[1] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[2] D. G. Lowe. “Distinctive image features from scale- invariant keypoints,” In *International Journal of Computer Vision*, 60(2):91-110, 2004.

[3] J. S. Sivic and A. Zisserman. “Video google: A text retrieval approach to object matching in videos,” In *Proc. of ICCV*, volume 2, pages 1470-1477, 2003.

[4] W. H. Hsu and S.-F. Chang. “Visual cue cluster construction via information bottleneck principle and kernel density estimation,” In *Proc. of CIVR*, 2005.

[5] K. Winn, A. Criminisi, and T. Minka. “Object categorization by learned universal visual dictionary,” In *Proc. of ICCV*, pages 1800-1807, 2005.

[6] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. “Improving “bag-of-keypoints” image categorization: generative models and PDF-kernels,” Technical report, University of Southampton, 2005.

[7] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. “Adapted vocabularies for generic visual categorization,” In *Proc. of ECCV*, pp. 464-475, 2006.

[8] M. Everingham, A. Zisserman, C. Williams, and L. Gool. The 2006 PASCAL visual object classes challenge.

[9] C. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2006.

[10] D. Larlus and F. Jurie. “Latent mixture vocabularies for object categorization,” In *British Machine Vision Conference*, pages 959-968, 2006.

[11] J. Yuan and Y. Wu. “Context-aware clustering,” In *Computer Vision and Pattern Recognition,* 2008.