CATEGORY-SPECIFIC INCREMENTAL VISUAL CODEBOOK TRAINING FOR SCENE CATEGORIZATION

Jianzhao Qin, Nelson H. C. Yung

Laboratory for Intelligent Transportation Systems Research
Department of Electrical and Electronic Engineering
The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

ABSTRACT

In this paper, we propose a category-specific incremental visual codebook training method for scene categorization. In this method, based on a preliminary codebook trained from a subset of training samples, we incrementally introduce the remaining training samples to enrich the content of the visual codebook. Then, the incremental learned codebook is used to encode the images for scene categorization. The advantages of the proposed method are (1) computationally efficient comparing with batch mode clustering method; (2) the number of visual words is determined automatically in the incremental learning procedure; (3) scene categorization performance is improved using the enriched codebook comparing with using the codebook trained from a subset of training samples. The experimental results show the effectiveness of the proposed method.

Index Terms— scene categorization, incremental learning, visual codebook

1. INTRODUCTION

Scene categorization is a task of automatic labeling a given image to a specific scene category (e.g., coast, highway, office, kitchen, street and etc.). In recent years, the bags of visual words model has been widely used for object recognition and scene categorization [1, 2, 3]. The principle of the bags of visual words model is that it quantizes a set of features extracted from a set of image patches to form a list of visual words. Then, this list of visual words is further used as a codebook for coding other images. Using this model, each patch in a given image is represented by a visual word in the codebook which is most similar to the feature of the patch and this image is coded by a vector which stores the distribution or the existence of the visual words in the codebook.

Obviously, the representative ability and the discriminative ability of the codebook will significantly influence the performance of categorization. The codebook which consists of a list of visual words is created by quantizing the features of the image patches in training set. Due to the large number of these patches and the high dimensional feature vector extracted from the patches, the memory and time cost to perform quantization on these large number of patches in a batch mode is prohibitive. Usually, a subset of the training samples is selected and a batch mode quantization algorithm (e.g. k-means) is employed to construct the codebook [1, 2]. Since only a subset of the training samples is used to form a codebook in the batch mode method, this codebook may not be sufficient to represent some features of the images that belong to a certain visual category. This may result in an inadequate model to represent the visual category, which may adversely affect the generalization ability of the classifier. Thus, on-line clustering method (e.g. on-line k-means [4]) has been employed to generate the codebook. However, one disadvantage of the batch-mode k-means and the on-line k-means is that the best number of visual words can only be determined using time-consuming cross-validation. Although some criteria (e.g. Akaike information criterion (AIC) and Bayesian information criterion (BIC)) are proposed to choose the best number of clusters for clustering, these criteria may not be suitable for choosing the number of visual words in order to optimize the classification performance. Yeh et al [5] proposed an adaptive codebook updating method for updating the content of the codebook. Their method, however, needs to define a capacity parameter (when the number of samples of a cluster exceeds this number, the samples and the samples of other adjacent clusters would be re-clustered) to determine when to add a new visual word. This criterion for re-clustering seems somewhat arbitrary. Moreover, since the updating strategy is based on the number of samples in a cluster, their codebook updating method also can not guarantee the discriminative ability of the generated visual words. Li-Jia Li and Li Fei-Fei [6] proposed an incremental model learning method to update the codebook and latent topics based on a variant of Hierarchical Dirichlet process. However, this method can only be used for generative visual model updating and whether a new visual word is added to the codebook also depends on a number of parameters. In this paper, we propose a category-specific incremental visual codebook training method. Based on a pre-trained category-specific codebook using a subset of the training samples, we extend this code-

book incrementally using the remaining training samples by determining whether the patches of the remaining training images are represented by the visual words from the same category. If the patch is correctly represented by the visual words from the same category, we just update the representation feature vector of the visual words based on current feature otherwise we add a new visual word to the codebook belonging to the category of current image (as depicted in Figure 1). In other words, we generate a codebook in order to ensure that the patches of every training image are correctly represented by the visual words belonging to the same scene category. We believe that this is a more reasonable method to update the codebook for increasing the discriminative ability of the codebook. Based on the updated codebook, each patch in an image is represented by a visual word in the codebook whose feature is most similar to the feature of the patch. We then simply classify an unknown image by summing the probabilities of the visual words that represent the image patches belonging to each scene categories. We tested the proposed method on two datasets consisting of 8 (2688 images) and 13 (3759 images) scene categories respectively using 10-fold cross-validation. The experimental results show that, using the incremental learned codebook, the average accuracy rate is improved by about 3% at coarse scales and about 1% at fine scales comparing with using the codebook generated from a subset of training samples.

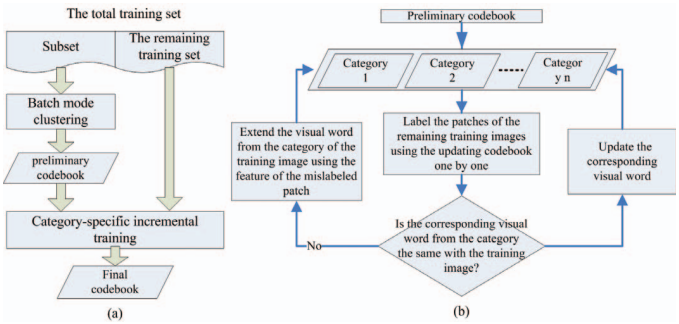

Fig. 1. (a) The framework of category-specific incremental visual codebook training; (b) Procedure of the category-specific incremental visual codebook training.

The paper is organized as follows. Section 2 introduces the proposed method. Section 3 shows the experimental results. And this paper is concluded in Section 4.

2. PROPOSED METHOD

2.1. Multi-scale scene image representation and category-specific contextual visual words

In this section, we briefly review the multi-scale representation of the scene image and how to introduce contextual information to the visual words introduced in our previous work [7]. In order to capture image information from different

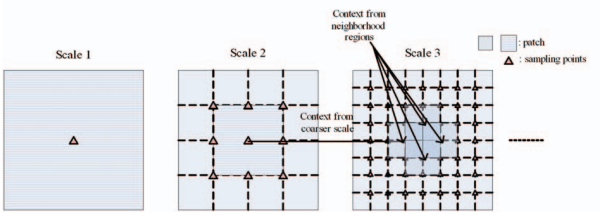

Fig. 2. Sampling points of the image patches at different scales and the regions for context information extraction.

scales, the image is regularly divided into patches at different scales from the coarsest scale (i.e. the whole image) to consecutive finer scales (depicted in Figure 2). Then, the Scale Invariant Feature Transform (SIFT) features [8] are extracted from all these patches. Meanwhile, the contextual information is integrated to describe the region of interest (ROI) [7]. The new contextual information provides useful information or cue about the ROI, which can reduce the ambiguity when employing visual words to represent the local regions. We combine the SIFT feature from the region at coarser scale (but with the same sampling point) and the SIFT features from the neighbor regions at the same scale with the feature of ROI to describe the ROI. That is, let $\mathbf{P}_L \in \Re^{m_L \times n_L}$ denotes the ROI, $\mathbf{P}_C \in \Re^{m_C \times n_C}$ denotes the region having the same sampling point as the ROI but at a coarser scale level and $\mathbf{P}_N \in \Re^{m_N \times n_N}$ denotes the neighbor regions of the ROI at the same scale level. For local visual word, the ROI is represented by $\mathbf{f} = f(\mathbf{P}_L)$ where f denotes the feature extraction function. For the contextual visual word, we represent the ROI by $\mathbf{f} = f(\mathbf{P}_L, \mathbf{P}_C, \mathbf{P}_N)$, and linearly combine them. The feature of the ROI is then represented as

$$\mathbf{f} = [f(\mathbf{P}_L), w_C \cdot f(\mathbf{P}_C), w_N \cdot f(\mathbf{P}_N)], \qquad (1)$$

where w_C and w_N are the weighing parameters that control the significance of features from the coarser scale and the neighborhood regions.

2.2. Category-specific incremental visual codebook training

This section describes the proposed category-specific incremental visual codebook training method, of which the procedure is depicted in Figure 1.

The codebook is formed by incrementally updating a preliminary codebook generated using a batch-mode clustering method. To produce the preliminary codebook, we firstly select a subset of images from the training image samples. Then, the SIFT features of the patches in those images are extracted and combined to form the image features that include the contextual information (see Section

2.1). Next, clustering operation is performed separately on the image features belonging to different scene categories. The centers of the clusters are taken to describe the corresponding visual words. We describe the set of visual words generated from the image features in scene category c as $\{\mathbf{v}_1^c, \mathbf{v}_2^c, \cdots, \mathbf{v}_{n_c}^c\}$. The codebook is formed by concatenating the visual words from different scene categories, $\mathbf{B} = \{\mathbf{v}_1^1, \mathbf{v}_2^1, \cdots, \mathbf{v}_{n_1}^1, \mathbf{v}_1^2, \mathbf{v}_2^2, \cdots, \mathbf{v}_{n_2}^2, \cdots, \mathbf{v}_1^C, \mathbf{v}_2^C, \cdots, \mathbf{v}_{n_C}^C\}$, where n_c is the number of visual words in category c. Details in [9] shows that creating the visual words in this category-specific manner could generate visual words with better discriminative ability that improve the classification performance.

Since the visual words are formed in a category-specific manner, the visual words belong to different categories. Ideally, the image patches of a scene image belonging to a category shall be represented by the visual words of the same category. The following incremental codebook updating criteria is based on this assumption. Given a new image belonging to scene category c, if the image patch (The feature of this image is denoted by $\mathbf{f}$) is wrongly represented by the visual word from other scene categories, we then add a new visual word to the group of visual words that belongs to category c , i.e. $\mathbf{B} = \mathbf{B} \cup \mathbf{v}_{n_c+1}^c, n_c = n_c + 1$. The feature representing this newly added visual word is the feature of the patch being wrongly represented, i.e., $\mathbf{v}_{n_c+1}^c = \mathbf{f}$. Extending the codebook in this manner is trying to guarantee that the wrongly represented patches can be represented by the visual words having the same categories with the labeled images. If the patch of the image is correctly represented by the visual word from the same scene categories, we just update the value of the feature describing the visual word, i.e., $\mathbf{v}_{i(new)}^c = \frac{m_i \mathbf{v}_{i(old)}^c + \mathbf{f}}{m_i + 1}$, where m_i is the number of features to form the visual word.

2.3. Classification

This section describes how to classify an unknown image based on the category-specific incremental trained codebook. Given the incremental trained codebook with a list of visual words $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \cdots, \mathbf{v}_n\}$ and their corresponding probabilities belonging to different scene categories, $\{p_{11}, p_{12}, \cdots, p_{1C}, p_{21}, p_{22}, \cdots p_{2C}, \cdots, p_{n1}, p_{n2}, \cdots, p_{nC}\}$, where C is the number of categories. The steps for classification are as follows:

Step 1: Regularly divide the unknown image into m overlapped patches.

Step 2: Represent each patch with the most similar visual word. We denote the most similar visual word to the ith patch as $\mathbf{v}_i$.

Step 3: Sum up the probabilities of the representation visual words belonging to scene category, $P_c = \sum_{i=1}^{m} p_{ic}, c = 1, \cdots, C$.

Step 4: The predicted category is set as $c_{prd} = \max_c(p_c)$.

Table 1. Average accuracy rates (standard deviation) (%) and p value of the Student's t-test at scales 1, 2 and 3 respectively for Dataset 1 ('BOW' denotes the traditional bags of words model which is taken as the baseline method.)

	Scale 1	Scale 2	Scale 3
BOW	59.25 (4.87) p <0.0001	71.47 (3.13) p <0.0001	73.52 (3.34) p <0.0001
Codebook from subset	67.41 (3.90) p =0.0014	74.21 (3.65) p =0.0004	82.68 (3.00) p =0.31
Incremental learned codebook	**69.80**(3.49)	**77.84**(3.60)	**83.54**(3.14)

Since the classification process is rather simple, compared with SVM classifier or other complex classifiers, the classification of an unknown image can be completed in a much shorter time.

3. EXPERIMENTAL RESULTS

This section reports the experimental results of the proposed method including a comparison of the proposed method using the proposed incremental learned codebook with the method using the batch-mode trained codebook from subset (100 images from the training set for each category). We also show the result of the traditional bags of visual words model based method (in which no category-specific visual words training and contextual information are used) as a baseline for comparison.

Performance of the proposed method is tested on two datasets which have been widely used in previous research [3, 10, 11][1]. Dataset 1 consists of 2688 color images from 8 categories. And Dataset 2 is an extension of Dataset 1 which contains 3759 images from 13 categories. Gray version of the images is used for our experiment.

In the experiments, we perform a 10-fold cross-validation in order to achieve a more accurate performance estimation. Moreover, in order to have a reliable comparison between different methods, we also performed the paired Student t-test on the accuracy rates from 10-fold cross-validation. The experiment is performed on different scale levels in order to see the performance variation with the change of scale levels (ref. Figure 2).

Table 1 shows that, using the category-specific incrementally learned codebook, the classification success rate is improved by 2.39%, 3.63% and 0.86% respectively at scales 1, 2 and 3 comparing with using the codebook trained from the subset of training samples. The p-values of the paired Student's t-test indicate that the improvement at scales 1 and 2 are statistically significant but not significant at scale 3. Comparing with the baseline method, the traditional bags of visual words model [1, 2], and the proposed method improved the

[1]The authors would like to thank Antonio Torralba, Fei-Fei Li, Rob Fergus and Lazebnik for providing their data sets.

Table 2. Average accuracy rates (standard deviation) (%) and p value of the Student's t-test at scales 1, 2 and 3 respectively for Dataset 2

	Scale 1	Scale 2	Scale 3
BOW	55.93 (3.05)	61.84 (3.95)	71.66 (2.77)
	p =0.0002	p <0.0001	p =0.0007
Codebook	59.02 (3.89)	68.90 (4.20)	76.42 (2.49)
from subset	p =0.0052	p =0.0012	p =0.2983
Incremental	**62.29**(3.12)	**72.32**(3.76)	**77.20**(3.29)
learned codebook			

accuracy rates by 10.55%, 6.37% and 10.02% at scales 1, 2 and 3 respectively. The p-values show the improvement is statistically significant at the three scale levels. Table 2 shows that, using the category-specific incrementally learned codebook, the classification success rate is improved by 3.27%, 3.42% and 0.78% respectively at scales 1, 2 and 3 comparing with using the codebook trained from the subset of training samples. The p-values of the paired Student's t-test also reveal that the improvements at scales 1 and 2 are statistically significant but not statistically significant at scale 3. Again, comparing with the baseline method, the traditional bags of words model, the proposed method improved the accuracy rates by 6.36%, 10.48% and 5.54% at scales 1, 2 and 3 respectively. The p-values indicate that the improvement is statistically significant at the three scale levels.

The results on the two datasets suggest that the proposed method is more effective at coarse scales. The reason may be that at coarse scales the visual words are more discriminative than the visual words at fine scales (Because the visual words at coarse scales describe a larger region of image which is more unique to certain scene category). When the visual words are used to describe small region at fine scales, some of them become more similar (the visual words describing the leaves of a tree may exist in 'forest', 'highway', 'coast' and 'inside city'). Thus, at fine scales, the determination of the scene category of an image should put more weights on some unique visual words. In the future, we will consider the distribution of weights across different visual words in the classification process.

4. CONCLUSIONS

In this paper, we have presented a category-specific incremental visual codebook training method for scene categorization. Based on a pre-trained preliminary category-specific codebook using a subset of the training samples, we extend this codebook incrementally using the remaining training samples by determining whether the patches of the remaining training images is represented by the visual words from the same category. Unlike the previous incremental codebook updating method in which the updating of the codebook is somewhat arbitrary, we update the codebook which aims at increasing the discriminative ability of the codebook. And the number of visual words of the proposed method can be determined automatically in the updating process. The experimental results show the proposed method is very effective at coarse scales and slightly useful at fine scales.

5. REFERENCES

[1] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *ICCV 2005*, L. Fei-Fei, Ed., 2005, vol. 2, pp. 1816–1823.

[2] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *ICCV 2003*, 2003, pp. 1470–1477.

[3] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR 2005*, 2005, vol. 2, pp. 524–531.

[4] Eric Nowak, Frdric Jurie, and Bill Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV 2006*, 2006, pp. 490–503.

[5] T. Yeh, J.J. Lee, and T. Darrell, "Adaptive vocabulary forests br dynamic indexing and category learning," in *ICCV 2007*, 2007, pp. 1–8.

[6] L.J. Li and L. Fei-Fei, "Optimol: Automatic online picture collection via incremental model learning," *International Journal of Computer Vision*, 2009.

[7] Jianzhao Qin and Nelson H.C. Yung, "Scene categorization via contextual visual words," *Pattern Recognition*, vol. 43, no. 5, pp. 1874–1888, 2010.

[8] David G. Lowe, "Object recognition from local scale-invariant features," *ICCV 1999*, vol. 2, pp. 1150–1157, 1999.

[9] Jianzhao Qin and Nelson H.C. Yung, "Scene categorization with multiscale category-specific visual words," *Optical Engineering*, vol. 48, no. 4, pp. 047203, 2009.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[11] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.