Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 24 (2012) 1255 – 1261

Physics Procedia

2012 International Conference on Applied Physics and Industrial Engineering

Improved Bags-of-Words Algorithm for Scene Recognition

Liu Gang[1], Wang Xiaochi[2]

[1]*Department of Information Technology*
Wenzhou Vocational College of Science & Technology Wenzhou, China
[2]*School of Information and Electrical Engineering*
Zhejiang University City College
Hangzhou, China

Abstract

This paper proposes a new bags-of-words (BoW)-based algorithm for scene/place recognition. Current scene recognition works that adopt BoW as the framework usually use a single codeword to represent the clusters obtained by k-means. Further, most of them often assign a hard value to a certain codeword to construct the BoW histogram. Using a single codeword to represent each cluster in fact is very preliminary since different clusters usually have different mean and covariance values. This causes using only mean value-based codeword will lose the covariance information and also makes the hard assignment to the codeword become biased. Considering this, this paper proposes an effective BoW-based technique to perform scene recognition. It first uses k-means algorithm to cluster the feature vectors into a certain number of clusters, in addition with an occurrence matrix. Gaussian mixed model (GMM) is then used to model the distribution of each cluster. Each GMM will be used as the new "codeword" of the codebook. Finally we propose to establish a new soft BoW histogram to represent each image through the soft assignment of the image features to each GMM. Support vector machine (SVM) is used to train these BoW histograms. Experimental results on the 15 categories dataset show that the proposed new BoW-based approach is very effective for scene/place recognition.

Keywords: scene recognition, bags-of-words (BoW), GMM, soft assignment

1. Introduction

The image's scene usually refers to the physical conditions of the environment where the image is taken 0. Many kinds of objects can be called scenes, such as the animals or plants, the outstanding architectures, the tall buildings and other similar environmental settings.

As the development of the camera equipments (e.g., camera phones, digital cameras, and PDA), people become more and more interested to capture the interested scenes that they see. As a result, image scene recognition is increasingly important for the camera users, such as place recognition for tourists, shopping guide (obtaining some related information about a product), and personalized services (purchasing tickets) 0-0 etc. An illustration of a typical scene recognition system is shown in Fig.1.

In Fig.1, for the query scene image, local patches/regions are first obtained by sampling the image, features are then extracted from these patches, such as the visual features (color histogram, wavelet coefficient, etc.) and scale invariant feature transform (SIFT) descriptor 0. These features are then forward to the trained classifiers (such as the support vector machine (SVM)) for matching. Finally, the recognition results are returned to the client, e.g., the category of the scene, the name of the scene, and other information related to the scene.

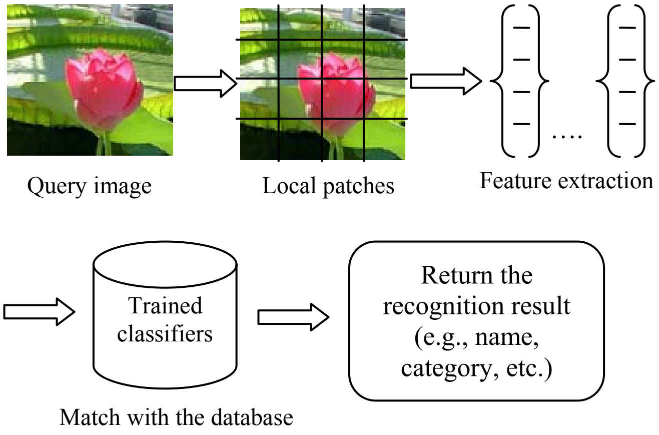

Fig.1 Illustration of a scene recognition system

Currently, many scene recognition works adopt bags-of-words (BoW) method as the framework 0-0. BoW first constructs a codebook using clustering algorithms such as k-means, in which each codeword is represented by the centroid of the cluster. Then the appearance times of each codeword in the image is used to represent the image, namely BoW histogram. This idea was first proposed in 0. From the above we can see that one very important step for BoW is the construction of codebook and the histogram calculation. Many methods have been proposed to design a codebook that has more powerful representation abilities. For example, a combination of adaptive and universal codebook is proposed in 0. It first establishes a universal codebook and then adjusts it to suit each scene category. However, one common shortcoming associated with these conventional codebook learning methods is the hard assignment of visual features to a single codeword. This may be appropriate for text classification but not for sensory data with appearance variances. The hard assignment causes the problem of selecting the correct codeword from two or more relevant candidates. The above codebook learning approaches only select the best representative codeword, without considering the relevance of other candidates.

In view of this, this paper proposes an efficient soft BoW-based algorithm based on GMM model. Instead of using single codeword-based codebook representation, we use GMM to represent the original codeword of each cluster. After generating the GMM-based codebook, the visual feature of each patch in the image is assigned to each GMM with a soft score. These soft scores for all patches in an image are then summarized to obtain a new histogram as the image's representation. Finally, SVM is used to train these histograms for classification.

The rest of this paper is organized as follows. A general overview of the proposed method' framework is first presented in Section 2. The detailed methodology is then discussed in Section 3, which includes feature extraction from the dense patches, GMM-based codebook generation, soft assignment of each patch to trained GMMs and SVM classification. Experimental results and discussions are given in Section 4. Finally, conclusions are given in Section 5.

2. The proposed framework

The proposed method is illustrated in Fig.2. It consists of two processing phases: offline training and online testing.

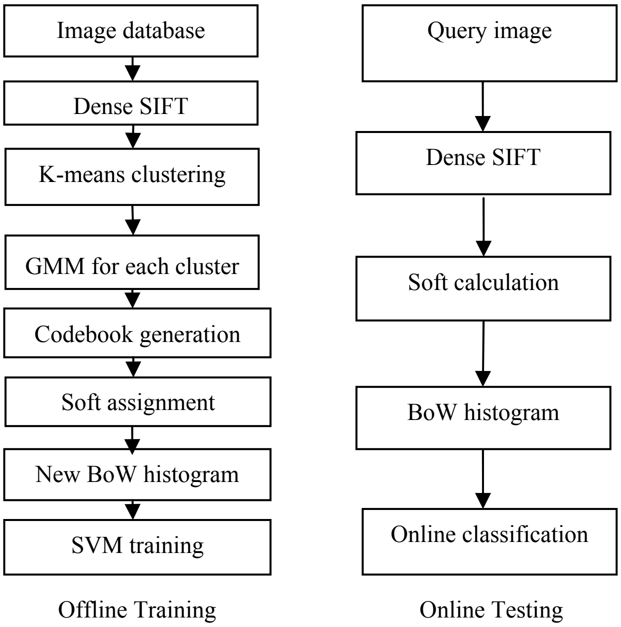

Fig.2 The framework of the proposed method

During offline training process, multi-scale dense SIFT descriptors are first extracted from the images in all scene categories. K-means algorithm is then used to cluster these dense SIFT into different clusters. GMM is used to model the distribution of each cluster as the new codebook representation. Based on this GMM-based codebook, dense SIFT extracted from each patch is assigned to these GMMs, and a set of soft scores are obtained through matching with these GMMs. These soft scores calculated using the same GMM are then summarized to obtain a histogram as the image's representation. The dimension of the histogram is the same as the GMMs number (also is the number of the clusters). Support vector machine (SVM) with histogram intersection kernel is finally used to train these new histograms. During online testing phase, each image will first be sampled into a set of multi-scale patches as in the training process. SIFT descriptors are then extracted, and further matched with trained GMMs to obtain a BoW histogram as calculated using the proposed soft assignment-based BoW algorithm. This histogram is then forwarded to the trained SVMs to determine which scene category the query image belongs to.

3. Proposed BOW-based Algorithm for SCENE recognition

3.1 Dense SIFT Extraction

The strategy of dense sampling has been shown to provide comparable or even better performance than interest points 0. SIFT descriptor is also proved to be robust to illumination, clutter and scale changes in 0. Therefore in this paper we use dense SIFT extracted from the multi-scale patches sampled in the images as the representation.

The reason we adopt multi-scale patches is that different users usually keep different distances from the captured scene. The different capturing distance will cause the scene image have a problem of scale changes. In view of this, a multi-scale decomposition scheme is adopted in this work. It partitions the image into dense patches with multiple scales. In our work, three different scales (l = 1, 2, 3) are used: 16×24 for l = 1, 24×32 and 32×64 for l = 2 and 3 respectively. Next, scale invariant feature transform (SIFT) descriptor is extracted from each dense patch.

3.2 GMM-based Codebook Generation

The reason to adopt GMM to model each cluster for the new codebook representation is that different clusters composed of image dense patches obtained using k-means algorithm usually have different mean and covariance values. Using only the centroid of the cluster as the codeword representation will lose much important information such as the distribution information (covariance) of the cluster. Therefore, GMM is proposed to simulate the distribution of each cluster and more information about this cluster will thus be kept. The proposed GMM-based codebook learning process is summarized in Algorithm 1.

Algorithm 1 GMM-based codebook learning method

1) Sample the training images from C scene categories into multi-scale patches using different resolutions as in Section 3.1.
2) Extract the SIFT descriptor from each image patch p_c^l, $c = 1,2,...$C; $l = 1,2,3$; where l is the image level.
3) Utilize k-means algorithm to cluster these multi-scale patches into M clusters.
4) Calculate the M×N occurrence matrix F(m, n) as below:

$$\begin{array}{c|ccccc|}
 & \text{scene } 1 & ... & \text{scene } n & ... & \text{scene } C \\
\hline
\text{cluster } 1 & f(1,1) & ... & f(1,n) & ... & f(1,C) \\
... & ... & & & & ... \\
\text{cluster } m & f(m,1) & ... & f(m,n) & ... & f(m,C) \\
... & ... & & ... & & ... \\
\text{cluster } M & f(M,1) & ... & f(M,n) & ... & f(M,C) \\
\hline
\end{array}$$

where the item *f*(m, n) represents all training patches from the *n*-th scene category whose SIFT descriptors are clustered as m-*th* cluster.

5) For each cluster, perform Expectation Maximization (EM) to estimate the GMM parameters, to be explained later.

The proposed GMM-based codebook learning scheme is explained as follows. First, a set of dense SIFT descriptors are obtained from the training images in the C scene categories. Then k-means algorithm is used to perform clustering, after which, these dense SIFT will be clustered into M clusters, in addition with an occurrence matrix. GMM is then used to learn the distribution of each cluster. The detailed process is as follows:

First, the initial centroids for GMM are determined by K-means clustering method. Then we use Expectation Maximization (EM) algorithm to train a GMM for these image patches in the cluster. The EM algorithm is briefed as follows:

We assume (i) $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ denotes a set of SIFT vectors (128 dimensional) from the *m*-th cluster; (ii) the distribution of these dense vectors can be approximated by Gaussian mixture models with k components $c_j (j \in [1, k])$; and (iii) the objective of EM is to maximize the log-likelihood $L(\theta|\mathbf{X})$ as follows:

$$\theta^* = argmax_{\theta \in \varphi} L(\theta|P)$$

$$= argmax_{\theta \in \varphi} \sum_{x \in X} log \left\{\sum_{j=1}^{k} w_j \cdot p(\mathbf{x}|c_j)\right\} \quad (1)$$

$$p(\mathbf{x}_i|c_j) = exp\left\{-\tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^T (\boldsymbol{\Sigma}_j)^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j)\right\} / \{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_j|\} \quad (2)$$

where $\theta = \{w_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{\mathrm{j}}\}_{j=1}^{k}$, w_j is the mixing proportion of the component c_j, $\boldsymbol{\mu}_j$ is the mean vector of the component c_j, and $\boldsymbol{\Sigma}_j$ is the covariance matrix of the component c_j.

After randomly selecting the parameters, EM first performs E-step, which estimates the probability that the feature vector $\mathbf{x}_i$ belongs to the component c_j in the *t*-th iteration by the Bayes rule:

$$p^{(t)}(c_j|\mathbf{x}_i) = p(\mathbf{x}_i|c_j)/(\textstyle\sum_{l=1}^{k} p(\mathbf{x}_i|c_l)) \quad (3)$$

$$s_j^{(t)} = \textstyle\sum_{x \in X} p(\mathbf{x}_i|c_j) / (\sum_{l=1}^{k} p(\mathbf{x}_i|c_l)) \quad (4)$$

where $s_j^{(t)}$ is the sum of the probabilities.

Then, the parameters of GMM are updated using Maximization step as follows:

$$w_j^{(t+1)} = s_j^{(t)}/n \quad (5)$$

$$\boldsymbol{\mu}_j^{(t+1)} = \textstyle\sum_{i=1}^{n} \mathbf{x}_i \cdot p^{(t)}(c_j|\mathbf{x}_i)/s_j^{(t)} \quad (6)$$

$$\boldsymbol{\Sigma}_j^{t+1} = \textstyle\sum_{i=1}^{n} \left(p^{(t)}(c_j|\mathbf{x}_i)\right)(\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})^T \quad (7)$$

E-step and M-step are performed recursively until the log-likelihood $L(\theta|\mathbf{X})$ is maximized. Finally, the feature vectors are assigned to the corresponding GMM components.

Finally, the estimated GMM for the *m*-th cluster can be represented as follows:

$$GMM_m(\mathbf{x}) = \textstyle\sum_{i=1}^{k} Gauss_m(\mathbf{x}; w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \quad (8)$$

where

$$\begin{aligned} Gauss_m(\mathbf{x}; w_i, \mu_i, \Sigma_i) &= -\log(w_i) - \log\left(c\left(\mathbf{x}; \mu_i, \Sigma_i\right)\right) \\ &= -\log(w_i) + \frac{1}{2}\log\left(\det\left(\Sigma_i\right)\right) \\ &\quad + \frac{1}{2}\left[\mathbf{x} - \mu_i\right]^T \Sigma_i^{-1}\left[\mathbf{x} - \mu_i\right] \end{aligned} \quad (9)$$

where $GMM_m(\mathbf{x})$ is the tested GMM score for the test SIFT vector **x.**

3.3 Soft Score Calculation based on GMM

After obtaining a set of GMMs for each cluster, we propose a soft score calculation method based on these GMMs to obtain the final BoW histogram for each image. The process is as follow:

For each image, a set of dense SIFT $\mathbf{f}_i(i = 1,2,\dots,q)$ are obtained using the proposed content analysis method as Section 3.2; where q is the number of sampled dense patches from the image. Then, the patch $\mathbf{f}_i$ is matched with all the GMMs of all the clusters to obtain a set of M scores as follows:

$$\mathbf{H}^i(m) = GMM_m(\mathbf{f}_i), m = 1,2,\dots,M \qquad (10)$$

where M is the number of GMMs, the same as the cluster number. $\mathbf{H}^i(m)$ is the calculated GMM score for patch *i* using the m-th GMM. Based on this, the final histogram for the image can be calculated using the summarization process as below:

$$\mathbf{H}(m) = \sum_{i=1}^{q} \mathbf{H}^i(m)\,, m = 1,2,\dots,M \qquad (11)$$

where **H** is the obtained new histogram based on the patch's soft assignment to each GMM.

After obtaining the soft BoW histograms for the training images in each category, one versus all SVM 0 is adopted in this work to train them.

4. Experimental Results

Fifteen scene categories dataset as 0 is used in this work to evaluate the proposed methods. We implement the BoW method in 0-0 as the baseline method for comparison. Further, the comparison with and without GMM-based soft estimation method is also made. We use a support vector machine (SVM) with histogram intersection kernel as a classifier. One versus-all SVM classifier is trained for each scene category in the database. 300 clusters are used as the codebook size. The final results are reported in the Table 1.

Table 1 The experimental results of proposed methods

Codebook learning method	Histogram establishment method	Recognition accuracy (%)
Method in 0	Method in 0	75
Method in 0	Method in 0	79
Proposed GMM-based method	Hard assignment in 0	77
Method in 0	Proposed soft histogram	76
Proposed GMM-based method	Proposed soft histogram	81

From the table, we can see that (i) the recognition rate (accuracy) increases by 2% compared with the standard BoW method in 0 when using the proposed GMM-based codebook learning method to replace the traditional single codeword-based codebook representation method. This shows GMM-based codebook has more powerful representation capabilities for each cluster. This is because that each cluster usually has not only different mean values but also different covariance. Single codeword-based representation will cause serious information loss since the covariance is lost; (ii) the proposed soft histogram calculation method is also better than the original histogram calculation method when using the same codebook leaning method, which increases by 1%. This shows that soft assignment of each patch to the codebook can better describe the patch's characteristics; (ii) when the proposed GMM-based codebook learning method is combined with proposed soft histogram establishment method for scene recognition, the accuracy is further increased to 81%, which is higher than the performance in [6], and is also the highest amongst all the methods. This shows that the proposed GMM-based codebook learning method and soft histogram establishment methods can perform better when combined together. Finally, we can conclude that the proposed new BoW-based approach is very effective for scene recognition.

5. Conclusion

This paper presents an effective BoW-based technique for image scene recognition. A new GMM-based codebook learning approach is proposed to replace the original single codeword-based codebook representation. GMM can better characterize the cluster's distribution than a single mean value and thus obtains better results. Besides, a soft histogram establishment method is also proposed based on the trained GMMs. It assigns the patch's descriptor to each cluster's GMM and obtains a set of soft values. These values in fact give an indication that how similar the test patch is with different GMMs. Compared with original method, such soft-based histogram has better representation ability for the image. Finally, using the one versus all SVM as the classifier, a set of experiments are made to evaluate the proposed methods. The results on the scene 15 dataset show that the proposed method is very effective and reasonable in image scene recognition.

Acknowledgment

This work is supported by *Ministry of Education*, Zhejiang Province, under *Scientific Research* Grant Y200909887.

References

[1]J. Liu and M. Shah, "Scene modeling using co-clustering", *IEEE International Conference on Computer Vision (ICCV),* 2007.

[2]J. Hays and A. A. Efros, "Scene completion using millions of photographs", *ACM Transactions on Graphics*, vol. 26(3), pp.87-94, 2007.

[3]B. Yamauchi and P. Langley, "Place recognition in dynamic environments", *Journal of Robotic Systems*, vol. 14, pp. 107-120, 1997.

[4]A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image classification for content-based indexing", *IEEE Trans. on Image Processing*, vol.10, 2001.

[5]G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray "Visual categorization with bags of keypoints", *In Proc. Of European Conference on Computer Vision*, pp. 59-74, Prague, 2004.

[6]F. Perronnin, C. Dance, G. Csurka and M. Bressan, "Adapted vocabularies for generic visual categorization", *In Proc. Of European Conference on Computer Vision*, Graz, 2006.

[7]Kim-Hui Yap, Tao Chen, Zhen Li, Kui Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intelligent Systems*, vol. 25, no. 1, pp. 48-57, 2010.

[8]D.G. Lowe. "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

[9]S. Lazebnik, C. Schmid and J. Ponce. "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories", *IEEE International Conference on Computer Vision and Pattern Recognition*, 2006.

[10]C.-F. Lin, S.-D. Wang, "Fuzzy support vector machines", *IEEE Transactions on Neural Networks*, vol.13, no. 2, 2002.