Proceedings of the 2012 International Conference on Wavelet Analysis and Pattern Recognition, Xian, 15-17 July, 2012

A SCENE CLASSIFICATION METHOD BASED ON ENSEMBLE SVM RESULTS

HUI-LAN LUO, LIAN-PING DU

Institute of Information Engineering, Jiangxi University of Science and Technology, Ganzhou, Jiangxi, 341000, China
E-MAIL: 16beyonder@sina.cn, luohuilan@sina.com

Abstract:

In classification problem, single classifier may not fully catch the dataset's information. Thus, an ensemble method based on Support Vector Machine (SVM) is proposed in this paper for image scene classification. First, Scale Invariant Feature Transform (SIFT) is used to extract the features of the images, and the SIFT features are clustered to form a visual vocabulary. Then, the SIFT features of each image are compared with this visual vocabulary to calculate the appearance frequencies of the visual words, which consist of the Bag-of-Words (BOW) model descriptions of the image. Probabilistic Latent Semantic Analysis (PLSA) is used to exploit the latent semantic features on the basis of the BOW model, and SVM classifier is then trained by these latent semantic features. These processes repeat N times and N different SVM classifiers are trained. Finally, they are used to classify the testing images, and the ensemble of the N different classification results are calculated as the final result. Experiments show that our method can be effectively applied to the scene classification problem, and the accuracy could be improved with a certain degree of robustness.

Keywords:

BOW; SIFT; SVM; Scene classification; PLSA; Ensemble

1. Introduction

With the advancement of computer technology, image acquisition and processing have become increasingly important. At the same time, the image database is developed rapidly which is widely used in various fields. In face of large amount of images, especially the scene images, the traditional manual processing cannot further met the practical needs. How to efficiently process image information and implement image classification has become a hot research topic in the field of image processing.

Currently, the research of scene classification mainly consists of the following two aspects: 1) the low-level visual features of images are used to implement scene classification. This method directly combines the common visual characteristics of image (such as color, texture and shape, etc.) with supervised training method to implement scene classification and discrimination [1-5]. For example, a color complexity weighted color histogram-based method for images retrieval is proposed in literature [2]; Logarithmic polar coordinates based on edge histogram is used to describe the sub-block shape characteristics and implement images recognition [4]. The low-level characteristics of images are valid for relatively simple scene classification, but the verdict is not ideal when the scene is too complex. 2) The mid-level semantic of images is used to implement scene classification [6-10]. This method uses the middle-level semantic of images to construct model, which considers the correspondence between the low-level visual features of images and the high-level semantics of scene, and this method is suitable for more complex scene classification and discrimination. For instance, the global semantic features of the scene are constructed [6], and the words such as natural, open, rough, vast, steep, etc. are used to describe the global structure of image which implements scene classification. However, in semantic-level classification, it is subjective to set the global symbol only. In literature [7], the local area of the image was given by a certain semantic concept. It overcame the above-mentioned subjectivity and implemented scene classification by calculating the frequency of the local semantic concept. Besides, a local semantic modeling approach based on Bayesian networks for natural scene classification is proposed in literature [8].

Generally speaking, the generation of mid-level semantics requires a lot of manual labeling samples. Under this circumstance, a method based on Gaussian statistical model [9] is proposed which implements scene classification without manually labeling samples. In summary, in order to better solve the problem of scene classification and improve the classification accuracy, the visual features of the images and semantic models should be combined.

By considering the aforementioned issues, we do some researches on the scene classification problem based on ensemble SVM. First, we analyze the size of visual words, which affects the classification performance. Second, we investigate different SVM ensemble schemes which may lead to different ensemble results. The experiments show that our method can obtain the state of the art performance effectively with a certain degree of robustness.

978-1-4673-1535-7/12/$31.00 ©2012 IEEE

2. The proposed method

Figure 1 is the framework of our scene classification method. The classification process has two phases: the training phase and the testing phase. In the training phase, different SIFT features, visual vocabularies, BOW models and latent semantic features are obtained. N different SVM classifiers can be separately trained on the different latent semantic features. In the testing phase, we use the N trained SVM classifiers to classify the testing images. The ensemble of the N different classification results are decided as the final result. The detailed information can be acquired from figure 1 and the following content.

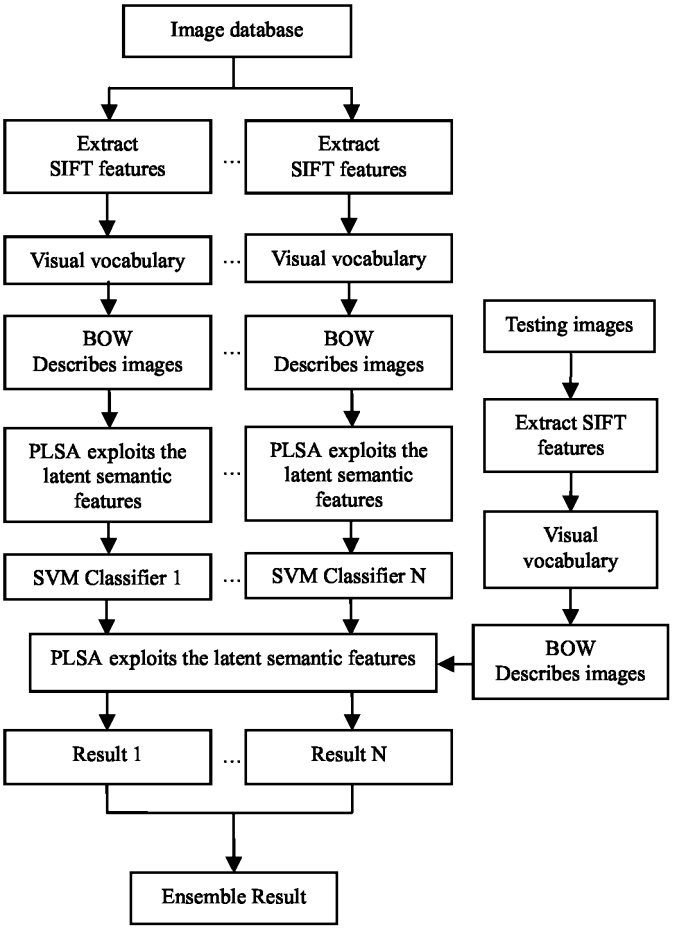

Figure 1.Image classification framework of our method

2.1. The generation of the visual vocabulary

Before the PLSA is used to exploit the images' latent semantic features, visual vocabulary needs to be formed first. The following is the process of forming visual vocabulary:

1) SIFT is used to extract all the images' key features.

2) K-means is used to cluster the images' SIFT features. Each cluster center corresponds to a visual word. A visual vocabulary holds K visual words.

2.2. BOW describes images

An image is consist of a set of visual words and this is similar to the case that text is made of words [11]. So BOW models can be applied to describe the images. Meaningful regions play an important role in scene classification. There are many ways to obtain the region of images [12-14], such as image segmentation, image blocking and so on. We will use the BOW model to describe the images' region. The specific process is as follows:

1) Each images' SIFT features are compared with the words of the visual vocabulary.

2) For each image, the key visual words' occurrence frequencies are calculated to build the BOW description of each image, which means that the images are quantified to construct the BOW descriptions.

2.3. PLSA exploits the latent semantic features

Probabilistic Latent Semantic Analysis (PLSA) was proposed by Hofmann[15, 16] in 1999, which is derived from the processing of natural language. Different from Latent Semantic Analysis of documents which solves the "ambiguity" and "a righteous word"[9], PLSA has a more solid mathematical foundation.

Given a documentation set $D=\{d_1,d_2,\cdots,d_n\}$, let $W=\{w_1,w_2,\cdots,w_n\}$ be the word set, and $A=[a_{ij}]_{n\times m}$ be the document-word co-occurrence matrix, where a_{ij} represents the weight of the word w_j in the document d_i. Hidden in the document words' back is latent variable $Z=\{z_1,z_2,\cdots,z_k\}$, where k is the number of the latent variables. PLSA is the statistical models that make latent variables Z associated with visual co-occurrence data on the frequency of occurrence in the document. PLSA model can be expressed as the joint probability of the vocabulary and the document:

$$p(d_i,w_j)=p(w_j)p(w_j\mid d_i) \quad (1)$$

$$p(w_j\mid d_i)=\sum_k p(w_j\mid z_k)p(z_k\mid d_i) \quad (2)$$

According to the maximized value of the likelihood function, we obtain the joint probability maximum likelihood

estimation. Likelihood function is:

$$L=\sum_{i=1}^{m}\sum_{j=1}^{n}n(d_i,w_j)\log p(d_i,w_j) \quad (3)$$

where $n(d_i,w_j)$ is the number of times that word w_j appears in the document d_i. For PLSA, the Expectation Maximization (EM) algorithm is used to maximize the log likelihood function to re-estimate the value of the model parameters $p(w_j|z_k)$, $p(z_k|d_i)$, $p(z_k|d_i,w_j)$. Expectation Maximization algorithm is divided into two steps:

(1) E-step, the current values of the parameters are used to calculate the posterior probability of the implicit latent variable z_k :

$$p(z_k|d_i,w_j)=\frac{p(w_j|z_k)p(z_k|d_i)}{\sum_{l=1}^{k}p(w_j|z_l)p(z_l|d_i)} \quad (4)$$

(2) M-step, the estimations of the parameters in the model are re-obtained based on the posterior probability:

$$p(w_j|z_k)=\frac{\sum_{i=1}^{n}n(d_i,w_j)p(z_k|d_i,w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{n}n(d_i,w_j)p(z_k|d_i,w_j)} \quad (5)$$

$$p(z_k|d_i)=\frac{\sum_{j=1}^{m}n(d_i,w_j)p(z_k|d_i,w_j)}{n(d_i,w_j)} \quad (6)$$

The EM algorithm repeats with the above two steps alternately. It is not stopped until the likelihood function L is less than a threshold. PLSA can calculate the conditional probability distribution $p(z_k|d_i)$ of latent variables in the document.

PLSA is used to exploit the latent semantic in the two following steps:

(1) In the training process, the use of probabilistic latent semantic models can get the visual theme of the training set's distribution $p(w|z)$. Each image can be expressed as a Z-dimensional vector $p(z|d_{train})$, Where Z is the number of visual themes. For the training set, PLSA model is used to obtain $p(w|z)$ and $p(z|d_{train})$. According to BOW model, $p(z|d_{train})$ is the image features quantified as the classifiers' input, and the vote of the multiple classifiers' outputs on the image is the final classification result, where $n(d,w)$ is the number of visual word w appears in the image document d .

(2) During the testing process, we need to calculate the testing images' $p(z|d_{test})$. Testing images' distribution $p(w|z)$ is the visual themes of the training set. The calculation of $p(z|d_{test})$ is similar to the PLSA training process, except that $p(w|z)$ is fixed.

For each testing image, when the number of the latent semantic features is T, then we can obtain a T-dimensional vector:

$$[p(z_1|d),p(z_2|d),\cdots,p(z_T|d)]^T$$

This T-dimensional vector is defined as the images' latent semantic features. In order to effectively implement image classification, SVM classifiers can be constructed by processing the T-dimensional vector [17].

2.4. Ensemble SVM results for scene classification

Step1: SIFT is used to extract the features of the images.

Step2: SIFT features are clustered by K-means to form visual vocabulary.

Step3: The SIFT features of all of the images are compared with visual vocabulary. Then, the appearance frequencies of the visual words are calculated to form the BOW model.

Step4: PLSA is used to find the latent semantic features on the basis of the BOW model.

Step5: SVM classifier is trained by the latent semantic features.

Step6: Step1 ~ Step5 are repeated N times and N different SVM classifiers are trained.

Step7: The N trained SVM classifiers are used to classify the testing images which lead to N different classification results.

Step8: The ensemble of the N different classification results is calculated as the final result.

3. Experiments and experimental results

In order to verify and assess the effectiveness of our method, we conduct some experiments on six categories of scene images as Airplane, Background, Bike, Car, Face, and Motorbike which are obtained from Caltech datasets. Each type of scene holds 200 to 1000 images and our experimental platform is MATLAB. The image samples are shown in Figure 2.

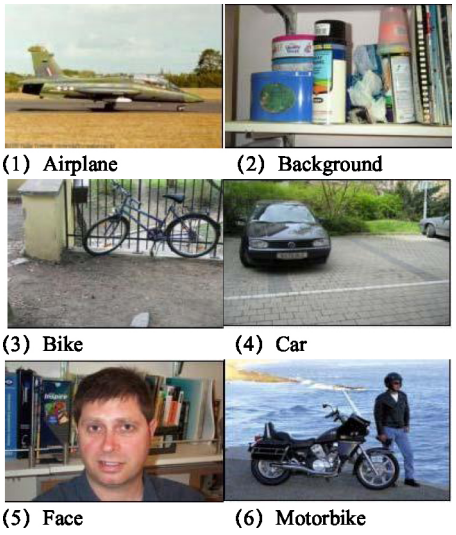

Figure 2.Sample of image dataset

The original images were normalized to fit a 200x150 window. By randomly selecting 100 images from each category, 50 images are taken as the training set and the other 50 as the testing set. The maximum number of interested points is 60, and the maximum number of iterations of the K-means is set as 50.

First, we analyze the size of visual words, which affect the classification performance. We use eight different sizes of visual words to do the experiments which are 50, 100, 200, 300, 500, 600, 800, 1000, and their affection on the classification performance is shown in figure 3. It is not difficult to see that different sizes of visual words do affect the classification performance. When the size of the visual words is too small, different visual semantic features may be assigned the similar visual words, which lead to poor performance, while the performance is improved with the size increasing. However, the performance stops increasing when the visual words' size reaches to a certain level. Besides, the larger the size is, the more time the experiment costs. Considering both the classification performance and the time complexity, we fix the size of the visual words as 300.

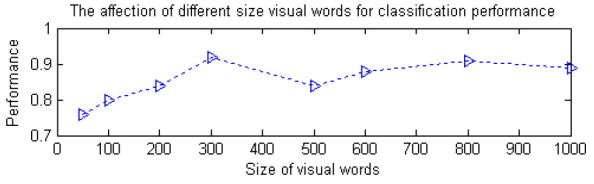

Figure 3.The affection of different size visual words for classification performance

Second, different ensemble schemes may lead to different results. We repeat Step1～Step5 5,10,15,20,25,30 times respectively, and 5,10,15,20,25,30 different SVM classifiers are trained. These classifiers are used to classify the testing images. Taking car and background for instance, the testing accuracies are shown in figure 4. We can see that a relatively higher accuracy is obtained when the number of classifier is 10. More classifiers may result in better performance, but the time cost will be much higher. In view of this circumstance, 10 different SVM classifiers are trained in each experiment.

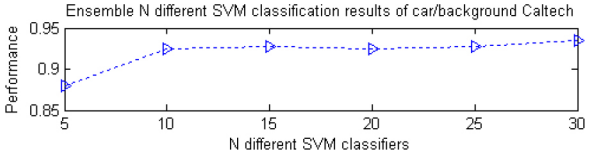

Figure 4.The performance of ensemble N different SVM classification results of car/background Caltech

Finally, we conduct some experimental comparisons among the method of literature [17], the method of literature [18], and our proposed method. Note that the methods of [17] and [18] do not use ensemble technology. The testing accuracies of the six categories are reported in Table 1. It is not difficult to find that our method has better performance which improves the classification accuracy to some extend.

TABLE 1. THE CLASSIFICATION PRECISION COMPARATION BETWEEN OUR METHOD AND OTHER METHODS

Class	The Classification Precision of Different Methods		
	Literature [17]	*Literature [18]*	*Our method*
Car	0.7778	0.8080	0.9240
Airplane	0.8643	0.8540	0.9500
Background	0.8467	0.8440	0.9240
Bike	0.7827	0.8020	0.9020
Face	0.9132	0.9080	0.9680
Motorbike	0.9388	0.9160	0.9700

4. Conclusion

In this paper, we have done some researches on scene classification problem based on the BOW model, PLSA method, and SVM classifier. The experimental results show that our proposed method can be effectively applied to the scene classification problem. Of course, it has some shortcomings which need to be further concerned, and the proposed method is unable to obtain good performances on some categories and some complex scenes. Therefore, how to ensemble different classifiers (such as KNN, SVM and so on) and use ensemble learning technologies to improve the accuracy of complex image classification will be our future research directions.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) project (No.61105042), Natural Science Foundation of Jiangxi Province, China (Project No. 2010GZS0075), and educational Commission of Jiangxi Province, China (Project No.GJJ11464 and GJJ11465).

References

[1] Foster David H, Marin Franch Ivfin, Amano Kinjiro, "Approaching ideal observer eficiency in using color to retrieve information from natural scenes", Optics and Image Science, and Vision(S1084-7529), Vol. 11, No. 26, pp. 14-24, 2009.

[2] Yongjian Yu, Xiangyang Wang, Junfeng Wu, "Weighted Histogram Color Image Retrieval Based on Color Complexity Measure", Journal of Chinese Computer Systems, Vol. 3, No. 30, pp. 507-511, 2009.

[3] Junding Sun, Xiaosheng Wu, "Content-Based Image Retrieval Based on Texture Spectrum Descriptors", Journal of Computer-Aided Design and Computer Graphics, Vol. 3, No. 22, pp. 513-520, 2010.

[4] Belongie S, Malik J, Puzicha J, "Shape matching and object recognition using shape contexts", Transaction on Pattern Analysis and Machine Intelligence(S0162-8828), Vol. 24, No. 24, pp. 509-522, 2002.

[5] Lin Chuen Horng, Chen Rong Tai, "A smart content-based image retrieval system based on color and texture feature", Image and Vision Computing (S0262—8856), Vol. 6, No. 27, pp. 658-665, 2009.

[6] Oliva A, Torralba A, "Modeling the shape of the scenel: A holistic representation of the spatial envelope", International Journal of Computer Vision (S0920—5691), Vol. 3, No. 42, pp. 145-175, 2001.

[7] Vogel J, Schiele B, "Natural scene retrieval based on a semantic modeling step", International Conference on Image and Video RetrievaI (CIVR' 04) Dublin, Ireland, July 21-23, pp. 207-215, 2004.

[8] Huanhuan Cheng, Runsheng Wang, "Bayesian Network Based Local Semantic Modeling for Categorization of Natural Scenes", Signal Processing, Vol. 2, No. 26, pp. 234-240, 2010.

[9] Zhao Xie, Jun Gao, "A Novel Method for Scene Categorization with Constraint Mechanism Based on Gaussian Statistical Model", Acta Electronica Sinica, Vol. 4, No. 37, pp. 733-738, 2009.

[10] Feifei Li, "Perona R.A bayesian hierarchical model for learning mtural scene categories", CVPR, pp. 524-531, June, 2005.

[11] Hofmann T, "Probabilistic latent semantic indexing",Proceedings of the 22nd annual international ACM SIGIR conference, New York: ACM Press, pp. 50-57, 1999.

[12] Lang K, "News weeder: Learning to filter news", In Proc. ICML, Vol. 95, pp. 331-336, 1995.

[13] Huohong Liu, Tao Fang, "Visual Words Ambiguity Analysis in BOW Model", Computer Engineering, Vol. 37, No. 19, pp. 204-209, 2011.

[14] David G Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision,Vol. 60, pp. 91-100, 2004.

[15] Hofmann T, "Probabilistic Latent Semantic Analysis ", New York: Elsevier Science Pub, 1999.

[16] Hofmann T, "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Machine Learning, Vol. 42, 2001.

[17] Ping Wang, Jingwen Zhang, Shunong Yang, "Image Retrieval Method Based on Salient Region and PLSA ", Journal of Chinese Computer Technology and Development, Vol. 21, No. 10, pp. 5-9, 2011.

[18] Pu Zen, Lingda Wu, Jun Wen, "Scene Classification Using Spatial Pyramid Blocks and PLSA", Journal of Chinese Computer System, Vol. 30, No. 6, pp. 1133-1136, 2009.