

An Object Recognition Method Based on Bag-of-Visual-Words and Fusing Multi-feature

Qi Xueting

Dept of Computer Application
WuHan University of Technology
WuHan, China
Qixueting163@163.com

Chen Tianhuang

Dept of Computer Application
WuHan University of Technology
WuHan, China
thchen57@126.com

Wang Hongxia

Dept of Computer Application
WuHan University of Technology
WuHan, China
whx_green@163.com

Abstract—The traditional bag-of-visual-words(BOV) model only uses one single feature to classify objects, which is difficult to achieve good results when dealing with many object categories. To solve this problem, in this paper, we proposed an object recognition method based on BOV model and fusing multi-feature. First, it extracted Scale Invariant Feature Transform (SIFT) features and Local Binary Pattern(LBP) features in the multi-scale space simultaneously, Second, we utilized SVM classifier to pre-classify separately on these two kinds of features and assigned their weights 0 or 1 according to the scale of pre-classification results. Third, the method fused SIFT and LBP features by introducing their weights, obtaining a fused vector. At last, classified on the fused vector using SVM classifier again, and then achieved the final result of object recognition. The experimental results show that the method proposed in this paper shows good performance and could improve the accuracy of object recognition effectively.

Keywords—object recognition; BOV model; fusing multi-feature; SIFT; LBP

I. INTRODUCTION

Object recognition, which in essence is to classify the object contained in the image or video in the view of computer, has broad application prospects in image retrieval, intelligent video surveillance and bio-medicine. Object recognition is one of the very active research directions in computer vision and pattern recognition fields. However, due to the influence of scale change, illumination problems, partial occlusion and other issues, object recognition is also a very challenging difficult point. In recent years, object recognition has been developing very fast, formed two main methods: appearance-based and structure-based[1]. BOV model[2] is a typical feature representation method based on appearance, while part-based-methods[3] is a typical one based on structure.

BOV model was proposed and applied into the field of computer vision by Csurka et al[2] in 2004. BOV first extracts local feature points of each image, clustering to generate several clustering center. Each cluster center is a visual word and the set of all the cluster center is the visual vocabulary. Then each image can be expressed by a few words in the visual vocabulary. All of these works are

preparations for subsequent image recognition. In BOV model, using of local features to represent the image can effectively enhance the robustness to the partial occlusion, changes within the class and other issues. In addition, due to the characteristics of simplicity, flexibility, BOV model has become a hot research point in recent years.

With the continuous development of the object recognition, fusing multi-feature has gradually become a research trend. Extracting several kinds of features can describe an image to a greater extent, and then improve the object recognition accuracy.

This paper proposed an object recognition method based on BOV model and fusing SIFT(Scale Invariant Feature Transform)[4] feature and LBP(Local Binary Pattern)[5-7] feature. The recognition process consists of four steps: Extract local features, generate visual vocabulary, code the image and classification. The recognition procedure is shown in Fig.1.

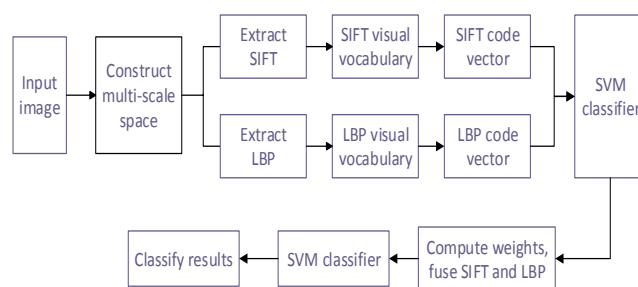


Figure 1. Recognition procedure

II. OBJECT RECOGNITION METHOD BASED ON FUSING MULTI-FEATURE

A. Extract local features

Extracting local features is an important step in BOV models, and the quality of the features directly affects the final classification accuracy. Common local features include SIFT, LBP and HOG (Histogram of Oriented Gradient)^[8] and so on. In this paper, we chose to extract and fuse SIFT and LBP these two features.

SIFT was presented by David G. Lowe in 2004. It is a feature matching algorithm based on multi-scale space, used for detecting and describing the local outstanding features of the image. It finds extreme points in the image's different scale space, and uses gradient direction statistics of a number of pixels within a certain neighborhood to describe the extreme points. Therefore, SIFT algorithm could extract more stable features, and has a strong robustness to the image scaling, rotation, perspective changes, illumination change and other problems. Extracting SIFT feature consists of two steps.

First, extract feature points. Build image's multi-scale space by constructing Gaussian pyramid and DOG (Difference of Gaussian) pyramid. The next work is detecting the extreme points in DOG pyramid and extracting potential feature points, then removing the feature points which are relatively unstable. Here, the construction of multi-scale space aims at extracting image features more effectively, because in many cases, features that is not easy to obtain in a certain scale will be relatively easy to detect in other scales.

Second, describe feature points. In order to enhance the matching ability of SIFT, David suggested describing each point with its regional 16×16 image area. Compute the gradient of each pixel in this area, and make a count on these gradient information in the 4×4 sub-area respectively. Then we can obtain a gradient direction histogram, forming a 128 dimensional vector. The vector is the description of the feature point.

LBP, proposed by Ojala[5], is a kind of feature to describe image's local texture information. It not only has the advantages of simple principle, low computational complexity, but also is image rotation and gray scale invariant. Therefore, LBP has a rapid development in recent years, and has been widely applied in face recognition.

The original LBP is calculated in the neighborhood 8 sampling points around the pixel. Set the gray value of the center pixel threshold, compared with neighborhood pixels' gray value, if the sampling point is small, then it is marked as 0, otherwise marked as 1. Finally, multiply each sampling point's marked value and their own weights, and the sum of these 8 products is the LBP value of the center pixel. But the original LBP only can extract features in the region with fixed radius. In order to improve this situation, Ojala[6] put forward the corresponding improvement scheme, extending the size of neighborhood area into a circular region with any radius R . However, with the expansion of neighborhood size, the number of sampling points also increases, and the dimension of LBP feature vectors shows a sharp growth trend at the same time. So Ojala proposed a uniform mode LBP and rotation invariant LBP again. Both of these two kinds of modes could reduce the dimension of LBP feature vector, besides, the latter also could make LBP rotation invariant.

In view of the multi-scale space's advantages in extracting image features, many researchers introduced multi-scale space theory into LBP, forming multi-scale LBP, in order to extract texture feature more efficiently. The uniform and rotation invariant LBP extracted in the multi-scale space not only has the gray invariance, rotation invariance and scale invariance, meanwhile, it also greatly reduces the dimension of LBP feature vector.

To identify a variety of object categories, using only a single type of feature is difficult to obtain better classification results, so research on the effective fusion of several kinds of features has certain values. This paper fused SIFT and multi-scale uniform and rotation invariant LBP. SIFT has high stability, it describes sampling points' gradient information within the neighborhood area of the key point, while LBP has low computational complexity, expressing the local texture information of the image. Fusing these two kinds of features could complement each other and improve object recognition accuracy effectively. Considering both SIFT and LBP features are extracted in the multi-scale space, therefore, when to extract features, we first constructed the multi-scale space of the image in this paper, then extracted SIFT and LBP features in this space simultaneously. While realizing LBP's scale invariance, it had the aid of Gaussian pyramid used in the process of extracting SIFT features, besides, the computation complexity of LBP algorithm was low, so the computation complexity of extracting these two kinds of features did not significantly improve compared with single SIFT features.

B. Visual vocabulary

Cluster the obtained local features by K-means algorithm, generating K cluster centers and the set of these centers is the visual vocabulary. The essence of K-means is to divide a disordered data sets into K groups, making the difference of the data within the same group minimum and that within the different groups as large as possible. The K-means algorithm consists of four steps.

- Select randomly K elements from N features as the initial cluster centers.
- Calculate the distance between each of the remaining features and every cluster centers, and packet the feature into the group in which its nearest cluster center is located.
- After traversing all objects, so far we have grouped data set initially, then calculate the average of all the data values in each packet, and set the mean value as the group's new cluster center.
- Do iteration on step 2~3, until the variance of the set of features converge.

In this paper, it clustered SIFT and LBP features respectively and generated

two visual vocabulary: SIFT visual vocabulary and LBP visual vocabulary. Through experimental comparison, set the size of SIFT visual vocabulary 1024 and LBP visual vocabulary 100.

C. Coding image

After extracting features, the number of feature points of each image is not entirely consistent, which will bring great difficulties to subsequent operations. Therefore, BOV model codes the image with visual vocabulary and extracted features so as to get vectors with the same dimension. Coding the image is a process of choosing the most appropriate word from the visual vocabulary to replace the feature point of the image.

VQ(Vector Quantization) algorithm is the most basic and simple coded method, and it's a hard-coded method. The principle is that for each feature point of the image, find a nearest word from the visual vocabulary to replace it, and then make a count on the frequency of every word, finally the vector composed of these frequency that is the coded vector. For example, we have obtained a n-vocabulary $V = \{v_1, v_2, \dots, v_n\}$. When coding image I_i , assume that $N(t, i)$ is the frequency of word V_t occurred in the image I_i , then the vector that could mark image I_i uniquely is defined as formula(1).

$$W_i = \{N(1, i), N(2, i), \dots, N(n, i)\} \quad (1)$$

We need to code the image twice in this article, the first time generate a 1024-dimensional SIFT coded vector according to SIFT features and SIFT visual vocabulary, while the second time generate a 100-dimensional LBP coded vector according to LBP features and LBP visual vocabulary. Thus far, we have finished coding the image.

D. SVM classifier based on weighted fused features

After coding, image I_i contains a 1024-dimensional SIFT coded vector S_i and a 100-dimensional LBP coded vector L_i . In this paper, we utilized SVM[9] classifier to pre-classify separately on these two kinds of coded vectors, and assigned the weights of SIFT coded vector w_s and LBP coded vector w_l according to the pre-classification results: SIFT accuracy r_s and LBP accuracy r_l . Then fused these two types of coded vectors in the light of their own weights, generating a 1124-dimensional fused vector. The vector is shown as formula(2).

$$SL_i = w_s * S_i + w_l * L_i \quad (2)$$

Used SVM classifier again on the fused vector, then we could get the final object recognition result. The method of computing weight is defined as formula(3).

$$\begin{cases} w_s = 1 & w_l = 0, & r_s > r_l \\ w_s = 0 & w_l = 1, & r_s \leq r_l \end{cases} \quad (3)$$

Among formula(3), w_s and w_l are the weights of SIFT and LBP coded vector respectively, while r_s and r_l are the object recognition accuracy when classify on the SIFT and LBP coded vector separately.

In summary, the steps of the method presented in this paper are summarized as follows.

- Input training image set T_r and test image set T_s .
- Extract SIFT and LBP features of all the image in T_r and T_s .
- Cluster respectively SIFT and LBP features of the image in T_r , generating SIFT visual vocabulary D_s and LBP visual vocabulary D_l .
- Code all the image in T_r and T_s , obtaining SIFT coded vector 错误!未找到引用源。 and LBP coded vector L_i .
- Classify on the S_i with SVM, getting SIFT's pre-classification result r_s . In the same way, classify on the 错误!未找到引用源。 with SVM, getting LBP's pre-classification result r_l .
- Compute weights according to formula(3), and fuse SIFT and LBP coded vector on the basis of formula(2).
- Classify on the fused vector again, then we can get the final object recognition result.

III. EXPERIMENT

A. Image database

In this paper, we used image database Caltech-101[10]. It is a large scale standard image database applied in general object recognition, covering 101 types of objects. Caltech-101 has a total number of 9146 images and the average size of the image is about 300*200. We chose former 30 images of each category as training samples, while the remaining images were composed of test samples. Part of images in Caltech-101 were shown in Fig.2.



Figure 2. Part of images in Caltech-101

B. Experimental results and analysis

Simulation experiments were carried out according to the steps of the proposed method in this paper. The result based on one single type of the feature and the result

produced by carrying out the proposed method are compared in TABLE I.

TABLE I. THE COMPARISON BETWEEN THE RESULTS BASED ON ONE SINGLE TYPE OF FEATURE AND THE RESULT BASED ON FUSED FEATURES(PROPOSED IN THIS PAPER)

	Results of LBP feature	Results of SIFT feature	Results of fused features
accordion	0.9600	1	1
binocular	0.6667	0.3333	1
camera	0.6500	0.7000	0.8000
watch	0.5354	0.8182	0.8182
windsor_chair	0.4615	0.7692	0.7692
Average accuracy	0.6548	0.7241	0.8775

In addition, there are two methods commonly used in the fusion of feature: The first one fuses several types of features in proportion of 1:1[11], while the second one fuses that in proportion of their own weights[12]. When computing the weights of every type of features in the second method, the larger the contribution rate to the final result, the greater the weight. A comparison was made between the two common fusing methods and the method proposed in this paper. The result is shown in TABLE II.

TABLE II. THE COMPARISON BETWEEN THE RESULTS USING TWO COMMON FUSING METHODS RESPECTIVELY AND THE RESULT USING METHOD PROPOSED IN THIS PAPER

	The first common method	The second common method	Method propose in This paper
accordion	1	1	1
binocular	0.3333	0.6667	1
camera	0.6500	0.6500	0.8000
watch	0.8182	0.7278	0.8182
windsor_chair	0.7692	0.6923	0.7692
Average accuracy	0.7141	0.7473	0.8775

As can be seen from the data in Table 1, the average accuracy computed through the method in this paper is obviously higher than that computed with SIFT features or LBP features separately. Contrasting the recognition rate of each object category, we can see that the accuracy based on the proposed method is also not less than any accuracy based on only SIFT features or LBP features. So we can draw a conclusion that fusing multi-feature could extract

more information of the image and will be more effective to object recognition.

The data in Table 2 can be concluded that compared with those two commonly used fusing method, the average accuracy calculated by the proposed method is improved. As for each object category, the proposed method shows good performance as well. As we can see from Table 2, the average accuracy of the first common method is lower than the second one. The reason is that the former fuses two types of features according to the ratio of 1:1, ignoring their different influences over the final result of recognition. Although the second one considers the difference of SIFT and LBP features by importing their own weights, it doesn't have a great improvement than the first one. As for the method in this paper, unlike the second common method, it only has 0 and 1 these two weights. If the contribution rate of one type of features is bigger, its weight is assigned as 1, otherwise assigned as 0. The experimental results show that, obvious improvement has been achieved with the proposed method.

IV. CONCLUSION

The method proposed in this paper fused SIFT and multi-scale LBP features based on BOV model. It extracted SIFT features and LBP features in the multi-scale space simultaneously, and confused these two features according to their own weights defined as 0 or 1. The experimental results show that this method can get a better effect of object recognition.

ACKNOWLEDGMENT

The work described in this paper was supported by a grant from the Natural Science Foundation of Hubei Province(ZRY1239).

REFERENCES

- [1] HUANG Kai-Qi, REN Wei-Qiang, TAN Tie-Niu. A Review on image object classification and detection. CHINESE JOURNAL OF COMPUTERS[J],2013,16(12):2-18.
- [2] G. Csúrká, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints[C]//ECCV'04 Workshop on Statistical Learning in Computer Vision, [S. l.]:[s. n.], 2004:59-74.
- [3] KUMAR M P, ZISSERMAN A, TORR P H S. Efficient discriminative learning of parts of parts-based models[C]//Proc of International Conference on Computer Vision.2009:552-559.
- [4] David G Lowe. Distinctive Image Features from Scale-Invariant Interest Points[J]. International Journal of Computer Vision,2004,60(2):91-110.
- [5] Ojala T, Pietikainen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition, 1996, 29(1): 51-59.
- [6] Ojala T, Pietikainen M, Maenpää T. Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.
- [7] SONG Ke-Chen, YAN Yun-Hui, CHEN Wen-Hui, ZHANG Xu. Research and Perspective on Local Binary Pattern. ACTA AUTOMATICA SINICA[J]. 2013,39(6):730-744.
- [8] Déniz O, Bueno G, et al. Face recognition using histograms of oriented gradients[J]. Pattern Recognition Letters, 2011, 32(12): 1598-1603.
- [9] Nello Cristianini and John Shawe-Taylor. An Introduction to Support Vector Machines and other kernel-based learning methods[M]. Cambridge University Press, 2000. ISBN 0-521-78019-5.

- [10] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. Proceedings of the Computer Vision and Pattern Recognition (CVPR), Workshop on Generative-Model Based Vision. Washington, DC, USA , 2004:178.
- [11] LI Wei-sheng,ZHAOXiao-xia.Object recognition based optimal bag-of-visual-words[J].Applicati-on Research of Computers, 2011, 28(9):3288-3290.
- [12] LI Wei-sheng,ZHAO Ling-zhi.Object Recognition Method Based on Fusing Muti-features of Interst Points[J].Computer Engineering,2010,36(18):7-9.