Image classification method based on improved bag-of-words model

Li Li[1*], Zhou Yan

School of Computer, Henan Institute of Engineering; Henan Zhengzhou 451191, China

Received 1 October 2014, www.cmnt.lv

Abstract

Image classification is one of the basic problems of image analysis and understanding. An improved SIFT algorithm is proposed for BoW model, which includes feature extraction and generation of visual dictionary. Caltech 256 database and Caltech 101 database are used for experiment to test the classification accuracy of proposed scheme. The experiment results show that the proposed scheme has higher classification accuracy than BoW model based on SIFT.

Keywords: image classification, Bag-of-Words, SIFT, feature extraction.

1 Introduction

In recent years, with the rapid development of network technology, more and more digital images turns up in the people's life. How to classify the massive image data into different categories based on image content is an important and meaningful task. Image classification techniques are proposed to deal with this problem [1]. Image classification is one of the basic problems of image analysis and understanding. How to classify these huge amounts of image information fast and accurately, has gradually become one of research hot spots [2]. Bag of words (BoW) model was applied in the field of document classification and was widely used because of its simple and effective advantages. Researchers in the field of computer vision try to apply the same idea in the field of image processing and recognition, and transition from the text processing technology to image processing is established.

Semantic text on forests for image categorization and segmentation was written by J. Shotton [3]. Learning discriminative spatial representation for image classification was written by G. Sharma [4]. Visual language modeling for image classification was written by L. Wu [5]. Spatial pyramid co-occurrence for image classification was written by Y. Yang [6]. Visual vocabulary optimization with spatial context for image annotation and classification was proposed by Z.G. Yang [7]. Improved bag-of-features for large scale image search was proposed by H. JEgou [8]. Linear spatial pyramid matching method using sparse coding for image classification was proposed by J. Yang [9]. Locality-constrained linear coding for image classification was also proposed by J. Yang [10]. Spatial pyramid matching for recognizing natural scene categories was proposed by S. Lazebnik [11].

When BoW model is applied in the domain of image recognition and classification, the image feature extraction and description should be carried out firstly, which requires appropriate feature extraction method and description to make the extracted characteristics describe the image as accurately as possible. SIFT [12] is a traditional descriptor, which has the invariance of rotation, scale variation, and has good robustness to noise and brightness change, and was successfully applied to target recognition, image restoration, image stitching, and other fields. But the algorithm still has some problems, such as longer feature matching time resulting from large number of feature points, high computational complexity and slow speed. Bag of hierarchical co-occurrence features for image classification was given by Takumi Kobayashi [15]. Ensembles of novel keywords descriptors for image categorization was written by Abdullah [16]. Extraction of image semantic features with spatial mean shift clustering algorithm was given by Mengyue Wang [17]. Visual object recognition method using DAISY descriptor was proposed by Chao Zhu [18].

Aimed at some defects existing in the Bags of words model, an improved scheme is improved. An improved SIFT descriptor is proposed, which is suitable for BoW model. The paper is organized as follows. In the next section, feature extraction based on SIFT for BoW model is investigated. In Section 3, an improved SIFT is proposed for BoW, including feature extraction and generation of visual dictionary. In Section 4, in order to test the classification accuracy of proposed algorithm, Caltech 256 database and Caltech 101 database are used for experiment. Section 5 gives some conclusions.

2 Feature extraction based on SIFT

Scale space theory can be used to simulate the multi-scale feature of image data, and it is proved the Gaussian convolution kernel can realize scale transformation. The two-dimensional image scale space expression is

[*] ***Corresponding author's*** e-mail: LL_mm2005@163.com

$$L(x,y,\sigma)=G(x,y,\sigma)\cdot \mathrm{I}(x,y) \tag{1}$$

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x_2+y^2)/2\sigma^2}$$

represents Gaussian function, (x,y) is the axis of the image, and σ represents image space factor. Gaussian difference function has high computing efficiency and is similar to scale normalized gauss Laplace function. Its expression is

$$D(x,y,\sigma)=(G(x,y,k\sigma)- G(x,y,\sigma))\cdot I(x,y) \tag{2}$$

Each feature point consists of three parts: the position, scale and direction. Each point in the scale space of Gaussian difference function is compared with adjacent points to get the location of the feature points. Use local features of each image to allocate one direction for each feature point, and then the feature descriptor has the rotational invariance. According to gradient distribution characteristics of neighborhood pixels of feature points, gradient modulus value is expressed as Equation (3).

$$m(x,y)=\sqrt{\begin{matrix}(L(x+1,y)-L(x-1,y))^2\\ +(L(x,y+1)-L(x,y-1))^2\end{matrix}} \tag{3}$$

Use direction distribution characteristic, Equation (4) is obtained.

$$\theta(x,y)=\tan^{-1}((L(x,y+1)-L(x,y-1))/ L(x+1,y)-L(x-1,y))) \tag{4}$$

L represents scale of the feature point.

SIFT feature descriptor is shown in Figure 1. Neighborhood selection method is as follows. Feature point is taken as the centre and the pixel size is 16×16. Size of each child area is 4×4 pixels. For each child area, calculate gradient direction histogram of eight directions. Finally, 8 direction gradient histograms of 16 number of child areas are ordered according to its location to constitute a 128-dimensional feature vector. This vector is taken as descriptor of SIFT feature.

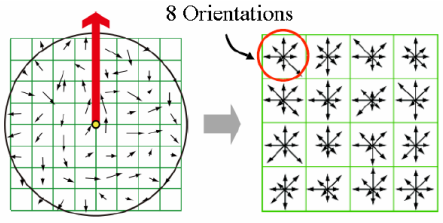

FIGURE 1 SIFT feature descriptor

3 BOW model based on improved SIFT

It can be seen from the above discussion that SIFT descriptor only describes the stable feature points of image, so it inevitably has the problem of information loss and omission. In the process of feature point detection and description, it is not hard to see, its complexity is very high, which needs to consume a large amount of computing time and which is also the downside of image recognition and classification. In BoW model, after feature extraction process, the clustering method is applied to generate visual words. If the feature extraction process does not provide enough information, it will directly affect the representation of the generated visual words. In turn, it affects the subsequent classification accuracy. Therefore, based on the comprehensive consideration of the above factors, an improved SIFT feature extraction method is proposed.

The improved SIFT descriptor adopts uniform sampling approach, and extracts image feature by the same pixel interval. The sampling density is controlled by a parameter step. Thus intensive feature points are obtained, which can guarantee the better use of the abundant information of the image. After feature point extraction based on sampling interval, each key point is assigned an uniform scale S, which avoids a lot of complicated scale calculation operation. In order to ensure the characteristics of rotation invariance, take the key point as the centre and construct circle area with pre-assigned scale S as radius. The pixels falling in the circular area is divided into 4×4 non-overlapping subareas, and the cumulative value of gradient in eight directions is calculated.

SIFT adopts Gaussian window function for weighted accumulation of the gradient, and in the improved SIFT, rectangular window is used to replace Gaussian window function. After completing gradient accumulation on key points, the mean value of Gaussian function of its own unit is used to weight adjacent area. This approximate method can not only improve the speed, but also ensure the performance. Each feature area is still represented by a 128-dimensional vector.

After the SIFT feature extraction, calculate sparse coding based on SIFT to get visual words. All collection of visual word forms the initial visual dictionary. For a given image I, an interested area set $P=\{p_1,p_{2,},...,p_M\}$ can be obtained by SIFT algorithm. V is obtained in the training process, given an interested area p_m, response of each descriptor a_m is calculated by Equation (5).

$$\min_{a,V}\sum_{m=1}^{M}\left\|p_m-a_mV\right\|^2+\lambda\left|a_m\right| \quad s.t.\ \left\|v_k\right\|\le 1,\ \forall k=1,2,...,K \tag{5}$$

V is a matrix of $K\times D$, K is the number of visual words, D is dimension of vector of visual word. In a_m, the maximum response element is id_m,

$$id_m=\max\left\{\left|a_{m1}\right|,\left|a_{m2}\right|,...,\left|a_{mK}\right|\right\} \tag{6}$$

v_{id_m} represents visual word corresponding to perceptual area. For each local feature p_m and space coordinate vector l_m. R Number of corresponding adjacent features

can be found which is $\left\{nb_m^{(1)},...,nb_m^{(R)}\right\}$. $\left\{(p_m,nb_m^{(1)}),...,(\mathrm{p}_m,nb_m^{(R)})\right\}$ is used to characterize local space information of feature p_m.

$$nb_m^{(1)}=\left\{p_j \mid \max_{m\neq j} dis(l_m,l_j)\right\},$$

$$nb_m^{(r)}=\left\{\begin{array}{l} p_j \mid \max dis(l_m,l_j),\ p_j\neq \underset{m\neq j}{nb_m^{(1)}},...nb_m^{(r-1)}, \\ 2\le r\le R \end{array}\right\}. \qquad (7)$$

$dis(l_m,l_j)$ is distance function, which calculates spatial location of two local characteristics. After visual dictionary is built, each image will be represented by a vector of statistical histogram. Then classifier is trained to determine image categories according to histogram vector of the training and testing images. SVM algorithm [13,14] is used for image classification here.

4 Experiment and analysis

In order to verify the validity of the method proposed in this chapter, the classic Caltech 101 database is adopted and Caltech 256 database is also adopted for experiment. In order to make the result of the experiment more persuasive, ten types of images are selected from Caltech 101 and Caltech 256 for the experiment respectively. Images samples of database Caltech 101 is shown in Figure 2 and Image samples of database Caltech 256 is shown in Figure 3.

Caltech 101: accordion, elephant, helicopter, llama, minaret, menorah, pyramid, schooner, sea-horse, stegosaurus.

Caltech 256: billiards, binoculars, birdbath, blimp, bonsai, boom-box, bowling-bell,
bowling-pin, bowling-glove, brain.

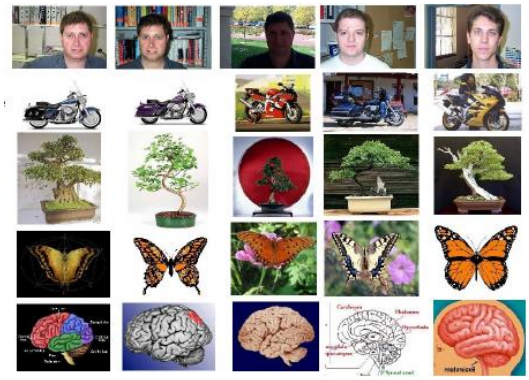

FIGURE 2 Images samples of database Caltech 101

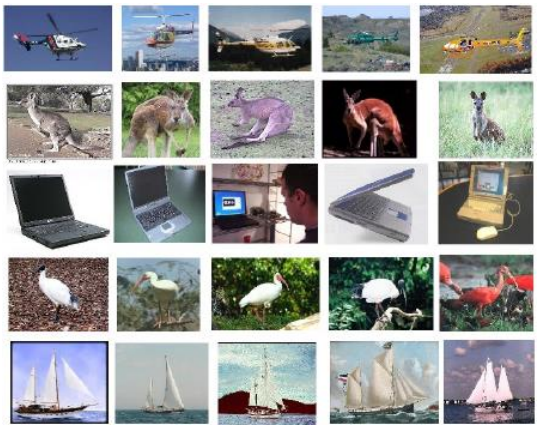

FIGURE 3 Image samples of database Caltech 256

The number of training images is set to 5, 10, 15, 20, 25, and 30 respectively, and 20 numbers of images are selected for testing randomly. The experimental results are the statistical average of 20 times. Classification accuracy comparison of Bow based on SIFT and improved scheme is shown in Table 1. The proposed BoW model has higher classification accuracy than BoW model based on SIFT.

TABLE 1 Classification accuracy comparison

The number of training images	Caltech 101		Caltech 256	
	BoW	Proposed scheme	BoW	Proposed scheme
5	60.0%	85.5%	26.5%	46.8%
10	62.5%	85.0%	35.0%	52.0%
15	68.5%	85.5%	36.5%	55.6%
20	70.0%	96.0%	37.5%	57.0%
25	71.5%	95.5%	37.5%	58.6%
30	71.5%	95.0%	40.5%	59.0%

Then 101 numbers of categories of the whole Caltech 101 database are used for the experiment. The results are compared with the existing relevant methods; analyse its advantages and disadvantages. In order to be convenient, the number of training images is 15 and 30 respectively.

Firstly, 15 pieces of images are selected randomly from each type of image to be as training images and another 15 pieces of images are selected randomly from each type of image to be as testing images. Figure 4 shows experimental results under this condition, and the classification accuracy is obtained by computing confusion matrix. Different colour represents the numerical value in corresponding confusion matrix. The result is statistical average of 10 number of experiment results and classification accuracy is 80.51%.

Then 30 pieces of images are chosen randomly from each type of image as the training images, and other random 15 images are chosen as testing images. The same confusion matrix is calculated to get the classification results and the result is shown in Figure 5. Classification accuracy is 83.16%.

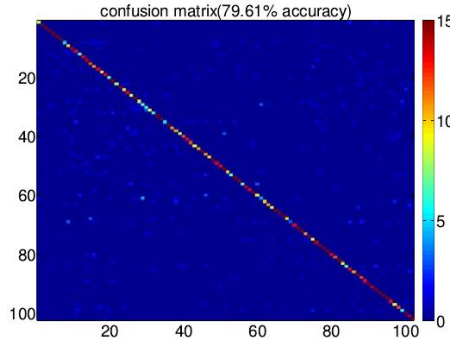

FIGURE 4 Classification result when the number of training sample is 15

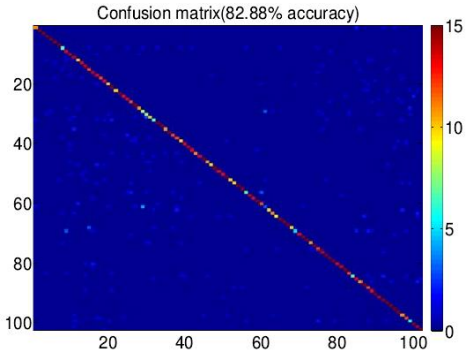

FIGURE 5 Classification result when the number of training sample is 30

In order to verify the superiority of this method, experimental results of the entire database are compared with the other existing methods. To be fair, Caltech 101 is selected as experiment database and BoW model is adopted. The results of different algorithms are shown in Table 2. It can be seen from the comparison of results, the proposed method is superior to other similar methods.

TABLE 2 The results of different algorithms

Algorithms	**N_train=15**	**N_train=30**
Takumi	Kobayashi[15]	59.8%
Azizi	Abdullah[16]	66.8%
Mengyue	Wang[17]	Not
Chao	Zhu[18]	56%
The proposed algorithm	80.51%	83.16%

In order to further analysis advantages and dis-advantages of the methods in this article, the experiment results are analyzed in detail. Figure 6 and Figure 7 are the best and worst categories of classification. Figure 8 shows images which are classified falsely. The classification accuracy of images in Figure 6 is 100%. By analyzing the image characteristics of these categories, it can be found that the images have little change within the class. For example they have the same rotation angle, and are less affected by the background factors, etc. For the category with the worst classification performance, through the analysis, it can be found that these categories of images tend to have larger change within the class. They are greatly influenced by artificial factors, objects are hidden in the background, and is very similarity to other classes.

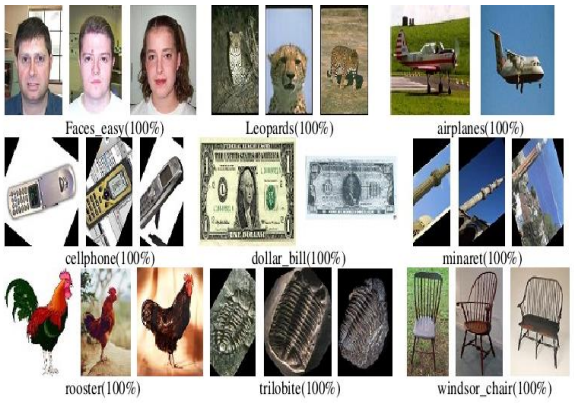

FIGURE 6 The best categories of classification

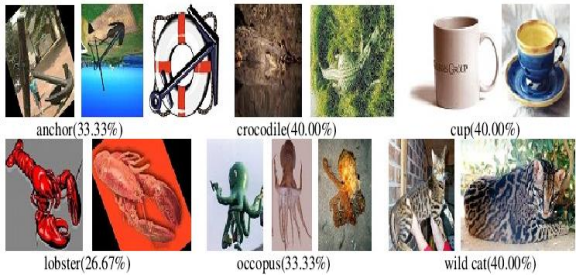

FIGURE 7 The worst categories of classification

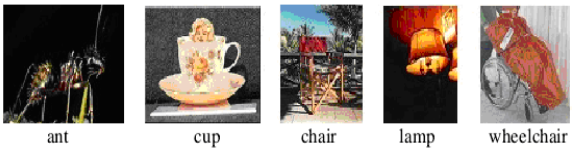

FIGURE 8 Images which are classified falsely

In order to further analysis advantages and disadvantages of the methods in this article, the experiment results are analyzed in detail. Figure 6 and Figure 7 are the best and worst categories of classification. Figure 8 shows images which are classified falsely. The classification accuracy of images in Figure 6 is 100%. By analyzing the image characteristics of these categories, it can be found that the images have little change within the class. For example they have the same rotation angle, and are less affected by the background factors, etc. For the category with the worst classification performance, through the analysis, it can be found that these categories of images tend to have larger change within the class. They are greatly influenced by artificial factors, objects are hidden in the background, and is very similarity to other classes.

Based on the above analysis, it can be found that the present method has high robustness for the change of the scale of the image, as well as position change and the influence of background. The falsely classified images are also analyzed. One reason is that the image is seriously influenced by light, and the second reason is that the images are greatly influenced by artificial factors. Its possible reason is that the method of this article does not adopt light pre-treatment. Artificial factors have stronger uncontrollability and how to avoid the impact of these factors greatly becomes one of the open problems to be solved.

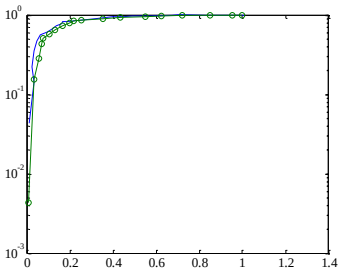

FIGURE 9 ROC curve of proposed algorithm and algorithm proposed by Takumi when the number of training sample is 15

ROC curve is adopted (also known as relevant operating characteristic curve) to measure the merits of the classification performance. ROC curve used true positive rate (sensitivity) as horizon axis and the false positive rate as vertical axis. ROC curve of proposed algorithm and algorithm proposed by Takumi is shown in Figure 9 when the number of training sample is 15. ROC curve of proposed algorithm and algorithm proposed by Abdullah is shown in Figure 10 when the number of training sample is 30. The blue line represents the proposed algorithm. It can be seen that the proposed algorithm has better classification accuracy.

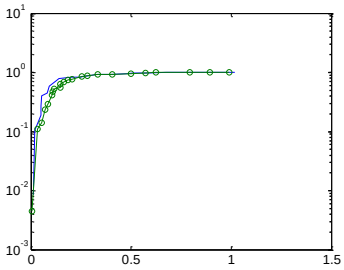

FIGURE 10 ROC curve of proposed algorithm and algorithm proposed by Abdullah when the number of training sample is 30

5 Conclusions

With the development of internet technology, large amount of images spring up. An improved SIFT algorithm including feature extraction and generation of visual dictionary is proposed for BoW model. The experiment results show that the proposed scheme has higher classification accuracy than BoW model based on SIFT.

References

[1] Cao Y, Wang C H, Li Z W. 2010 Spatial-bag-of-features *The IEEE Conference on Computer Vision and Pattern Recognition San Francisco USA* 3352-9

[2] Takumi Kobayashi, Nobuyuki Ostu 2010 Bag of hierarchical co-occurrence features for image classification *IEEE International Conference on Pattern Recognition* 3882-5

[3] Shotton J, Johnson M, Cipolla R 2008 Semantic text on forests for image categorization and segmentation *IEEE Conference on Computer Vision and Pattern Recognition, Anchorage USA* 1-8

[4] Sharma G, Jurie F 2011 Learning discriminative spatial representation for image classification *The 22nd British Machine Vision Conference Dundee Britain* 1-11

[5] Wu L, Li M J, Li Z ,W 2007 Visual language modeling for image classification *The International Workshop on Multimedia Information Retrieval Augsburg Germany* 115-24

[6] Yang Y, Newsam S 2011 Spatial pyramid co-occurrence for image classification *The 13th IEEE International Conference on Computer Vision Barcelona Spain* 1465-72

[7] Yang Z G, Peng Y X, Xiao J G 2012 Visual vocabulary optimization with spatial context for image annotation and classification *The 18th International Conference on Advances in Multimedia Modeling Klagenfurt Austria* 89-102

[8] JEgou H, Douze M, Schmid C 2010 Improving bag-of-features for large scale image search *International Journal of Computer Vision* **87**(3) 316-36

[9] Yang J,Yu K, Gong Y, Huang T 2009 Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)* 1794-801

[10] Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y 2010 Locality-constrained Linear Coding for Image Classification I*EEE Conference on Computer Vision and Pattern Recognition(CVPR)* 3360-7

[11] Lazebnik S, Schmid C, Ponce J 2006 Beyond bags of features: spatial pyramid matching for recognizing natural scene categories *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2169-78

[12] Lowe D G 2004 Distinctive image features from scale-invariant key points *International Journal of Computer Vision* **60**(2) 91-110

[13] Burges C J C 1998 A Tutorial on support vector machines for pattern recognition *Data Mining and Knowledge Discovery* (2) 121-67

[14] Chang C-C, Lin C-J 2013 LIBSVM-A Library for SupportVectorMachines http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[15] Takumi Kobayashi, Nobuyuki Ostu 2010 Bag of hierarchical co-occurrence features for image classification *IEEE International Conference on Pattern Recognition* 3882-5

[16] Abdullah A, Veltkamp R C, Wiering M A 2010 Ensembles of novel keywords descriptors for image categorization *IEEE 11th International Conference on Control Automation Robotics* 1206-11

[17] Wang M, Zhang C, Songl Y 2010 Extraction of image semantic features with spatial mean shift clustering algorithm *IEEE 10th International Conference on Signal Processing* 906-9

[18] Chao Z, Bichot C E, Liming C 2011 Visual object recognition using DAISY descriptor *IEEE International Conference on Multimedia and Expo* 1-6

Authors

LI LI, 1979.07, Zhengzhou, Henan, P.R. China.

Current position, grades: the lecturer of School of Computer, Henan Institute of Engineering, China.
University studies: MSc from Huazhong University of Science and Technology in China.
Scientific interest: computer application, multimedia technology.
Publications: 10 papers.
Experience: teaching experience of 11 years, 5 scientific research projects.

YAN ZHOU, 1981.04, Zhengzhou, Henan, P.R. China.

Current position, grades: the lecturer of School of Computer, Henan Institute of Engineering, China.
University studies: MSc from Huazhong University of Science and Technology in China.
Scientific interest: computer application, computer software and theory.
Publications: 5 papers.
Experience: teaching experience of 10 years, 3 scientific research projects.