The Extended Bag of Words Model for Visual Recognition and Categorization

Miaomiao Liu, Xinde Li*, Xiaobin Jin and Xiao Zhang

School of Automation, Southeast University, Si Pai Lou 2, Nanjing, 210096, China

Abstract: With the development of science and technology, more and more images need to be recognized and categorized. Although the classical Bag of Words (BoW) model has played a great role in the past, there are still many limitations about it, i.e. low precision and accuracy, high complexity of computation, etc. In this paper, it is improved and extended from four ways. Firstly, the features filtered from the background are sampled to reduce the influence of background noise. Secondly, the spatial relationship among all features is integrated with the classical BoW vector to improve the accuracy of recognition and categorization. Thirdly, vocabulary tree is constructed by applying hierarchical K mean value, in order to obtain more reasonable vocabulary list and greatly reduce the clustering time. Fourthly, a weighted visual word histogram is considered, in order to stand out the essential difference among images. At last, some experiments are conducted to show the advantage of the proposed method.

Keywords: BoW, spatial relationship, visual recognition, visual categorization, generic object recognition.

1. INTRODUCTION

With the development of science and technology, more and more images need to be recognized and categorized. For example, with the improvement of living, indoor mobile robot is confronted with the unstructured environment with diversity, personality. Traditional artificial landmarks might destroy the harmony of indoor decoration, however, natural landmarks do not need to change the environment, where the robot navigates by recognizing and categorizing some existing objects with distinct features. Obviously, natural landmarks will replace artificial landmarks without question, especially, on some spots where it is unfit to place artificial landmarks. Internet has gotten growing up to be an information window for everyone in the world, which has caused an exponential increase in the amount of online video data. Visual categorization for indexing, filtering, searching, mining, storing and analyzing becomes increasingly significant and even necessary. Since image indexing based on content from a great deal of video or image collections is a challenge for a computer vision system, queried by keyword provides an attractive and interesting way to search for appearance. For visual categorization, an image is usually represented by using the low level global features in conventional methods in the past. However, the current alternatives focus on abstraction and representation of the semantic feature.

Bag of Words (BoW) model is an effective arithmetic of object recognition and categorization from this point of view, because of its simple strategy and its robustness for object position and deformation in image. However, every feature is independent with each other in this model, that is, there is no spatial relationship to be considered. Actually, the spatial relationship between features could be useful to describe the internal structure of objects or to highlight the importance of contextual visual information for these objects. For example, Teng Li, *et al*. [1] proposed a contextual bag-of-words (CBoW) representation to model two kinds of typical contextual relations between local patches, i.e., a semantic conceptual relation and a spatial neighboring relation, since the conventional BoW neglects the contextual relations between local patches due to its Naïve Bayesian assumption. In their study, in order to describe the semantic conceptual relation, visual words were grouped on multiple semantic levels in terms of the similarity of class distribution, which were associated with different local patches and global image. In order to describe the spatial neighboring relation, N-gram language model were adopted to measure the confidence that the neighboring visual words are relevant. Tinglin Liu, *et al*. [2] focused on discovering the dependency relationship among all the visual words through exploiting co-occurrence information in spatial domain. Agarwal *et al*. [3] proposed a two-step approach. First, some object parts were detected in images based on a previously generated vocabulary. Then, for these detected parts, the spatial relations were described by quantizing their relative distances and orientations. The final image signature was a two-part feature vector containing parts occurrences on one side and quantized relations on the other side. Although their idea is similar to ours, the second step - quantized relations is different from

*Address correspondence to this author at the School of Automation, Southeast University, Si Pai Lou 2, Nanjing, 210096, China; Tel: +86-025-83790871; Fax: +86-025-83792724; E-mail: xindeli@seu.edu.cn

E-ISSN: 2409-9694/14

ours. Xianglong Liu *et al*., [4] proposed a soft match and score method to consider the spatial relationships among visual words. Xiangang Cheng *et al*. [5] proposed a structure propagation technique to build more reasonable co-occurrence matrices of visual words to describe the spatial information. That is, if two patches lie in the same object part, and then, a higher weight is assigned to the co-occurrence over them. Meng Sun, *et al*. [6] proposed a graph regularized non-negative matrix factorization (NMF) model for image pattern discovery, which preserved the spatial closeness of visual code words in the obtained patterns, thus improved the main short-coming of bag-of-words representation. Ismail Elsayad [7] proposed a new spatial weighting *scheme* for bag-of-visual-words based on a mixture of n Gaussians in the feature space. On the elicitation of space pyramid matching, Lazebnik *et al* [8] proposed a Bag-of-words based on space pyramid, which firstly partitioned image into a series of subfield stepwise, and then compute the histogram of Bag-of-words of each subfield, in order to establish a histogram representation of pyramid, at last, classify image. However, with the increment of layers, the advantage of it over the original Bag-of-words will disappear because it is very sensitive to pose. On the elicitation of Bag-of-words based on space pyramid, Zhang Linbo, *et al.* [9] proposed a Bag-of-phrases model, whose effect is only a little better than that of Bag-of-words based on space pyramid.

In addition, others also improved the classical BoW model aiming to its other shortcoming. For example, according to the limitation of conventional model that *scheme*s are mostly migrated from text retrieval domain and don't take into account fundamental differences between textual words and visual words. Wassim Bouachir, *et al*. [10] proposed a new weighting *scheme*, where they used a fuzzy representation to index images with a more robust signature. Since the k-means algorithm commonly used to construct a visual vocabulary for quantizing the extracted 3D interest points from videos has two major drawbacks: sensitive to the vocabulary size and the initialization; unable to capture the salient properties of the videos. Changhong Liu *et al*. [11] proposed to construct a visual vocabulary and represent a video by sparse coding followed by the max pooling. Jingyan Wang, *et al*. [12] proposed an assignment method of novel quadratic programming for reconstruction weights.

In this paper, we further extended BoW model, in order to make it more helpful for visual recognition and categorization. Therefore, this paper can be organized as follows: the classical BoW model is reviewed in section 2. It is improved and extended in section 3. For example, a sampling method of feature filtered from background is proposed in section 3.1, in order to reduce the disturbance of background noise. And then, we integrate spatial relationship among all features on the base of classical BoW in section 3.2. In section 3.3, the vocabulary tree is constructed by applying hierarchical K mean value, in order to obtain the more reasonable vocabulary list and greatly reduce the clustering time. In section 3.4, in order to stand out the essential difference among images, we carry out a weighted processing of visual word histogram. In section 4, some experiments are conducted to show the advantage of new method. At last, a conclusion is given in section 5.

2. REVIEW OF BOW MODEL

For example, there are some different objects described by BoW model, where the upper layer refers to the image described by BoW, the bottom layer refers to the set of visual words from image set, i.e. vocabulary list. Most of these visual words have some definite semantic information, for example, the image of human face includes nose, eye and mouth, etc. The middle layer refers to the statistical histograms of visual words from three different images, which are distinctive with each other through observation. Therefore, they belong to the different category. However, there are also some distinctions among different objects, which belong to the same category, we can still find some common features. For example, although there are great distinctions among human faces from different people, we can't find great distinctions on some minor organs from big scale space, i.e. eyes, mouth, nose, etc., that is to say, visual words relative to image semanteme take big proportion in histogram.

Generic object recognition is more difficult than special object recognition, such as human face, air-plane, and car, etc. Some reasons are given as follows:

1. Category diversity. Since there might be not great distinction among different categories, visual vocabulary must accommodate enough visual words, in order to distinctly describe so many categories. However, with the increment of vocabulary, the computation amount becomes greater and greater, leading to the retrieval speed lower.

2. Great distinction among the same category. Since the diversity of object, the same category

has different distortions, evolvements and so on. So it is very necessary to extract some visual words with the common or representative features. For example, there are many different kinds of chairs, which shows the challenge to find a recognition arithmetic like human brain to recognize them.

3. Effect angle of view. The object seen from different angle will appear differently. Therefore, visual words must own the ability of affine invariability.

4. Disturbance of illumination. That is to say, visual recognition will suffer from the illumination, so visual words must tolerate certain difference of illumination.

5. Occlusion. For a whole object, even if its partial appearance is blocked, or has some changes, it still belongs to that category. For example, for a bike, if one of its wheels is blocked, it shouldn't be regarded as a wheelbarrow.

It is the core to describe object with visual words in BoW model. This is because an image is divided into many patches with semantic features. And then, vocabulary is made of these visual words. That is, the image is represented finally to be a histogram vector. Since these visual words from image are not similar to those in a text, we must extract some independent ones. The main steps are listed as follows:

- Sampling from image features.
- Acquire the neighboring features vector from lots of image features, and regard them as a mathematics expression of alternative visual words.
- Extract the most representative visual words from those candidates, in order to construct vocabulary.
- Count up the weight of words in visual vocabulary, and optimize the vocabulary.
- Image is abstracted as the histogram of visual words.
- After the supervised training, establish a classifier for the histogram of visual words from lots of training images.
- Extract the histogram of visual words from image to be recognized, and recognize it with classifier.

3. IMPROVED BOW MODEL

3.1. Sampling Method for Background Filter

In classical BoW model, there are two limitations from the sampling of features:

1. In the course of quantizing in classical model, the local features from each image are described by nearest neighbor in vocabulary. Although each feature vector can be represented by certain a visual word as nearest neighbor in vocabulary, there is no similarity computation and evaluation among them, so that some visual words with low similarity are also intermingled into vocabulary, which have some bad effects on classifier.

2. In the real environment, generally speaking, there are always some disturbances near target, which isn't helpful for object recognition and leads to the decline of classifying ability.

So it is necessary to reduce the influence of background, the main arithmetic is given as follows:

1. According to the constructing method of visual words in classical BoW model, establish visual vocabulary after sampling features from lots of images.

2. Before the description of image to be recognized, compute the similarity between each feature point in an image to be recognized and each visual word in vocabulary.

Suppose the visual vocabulary is $Q = \{Q_j, j = 1,2,\ldots k\}$ after clustering,

Q_j refers to the *j* th visual word in visual vocabulary; there are M feature points in one image, P_i is the *i* th SIFT feature vector with 128 dimensions.

The similarity between Q_j and P_i is defined in Eq. (1).

$$s(P_i, Q_j) = \frac{\left|P_i - Q_j\right|}{\left|P_i\right| \times \left|Q_j\right|} \tag{1}$$

$$s'(P_i) = \min_{j=1,2,\ldots k} (s(P_i, Q_j)) \tag{2}$$

If $s'(P_i) < T_s$, remove P_i. Here T_s refers to the threshold. That is, if the similarity is within the threshold, we regard this feature point as a valid one.

By far, we believe the left local features belong to the image, but among these left local features, there are some coming from the background. Therefore, some extra actions need to be taken to delete them according to the following reasons: 1) the number of local features from object is much more than that from the background after similarity computation; 2) we can reduce the background's disturbance further according to the density distribution of local features in an image.

Suppose we get T local features from M local features after computing similarity, we may get the object in the rectangle in spite of some disturbances from the background. In order to do that, we may use RANSAC [13] to reduce the negative influence on the later image description. For convenience, we use a circle to cover the area where the density of features is very high. All other features outside the circle field belong to background, which will be filtered out. The procedure is shown in Figure **1**. The pseudo codes are given as follow:

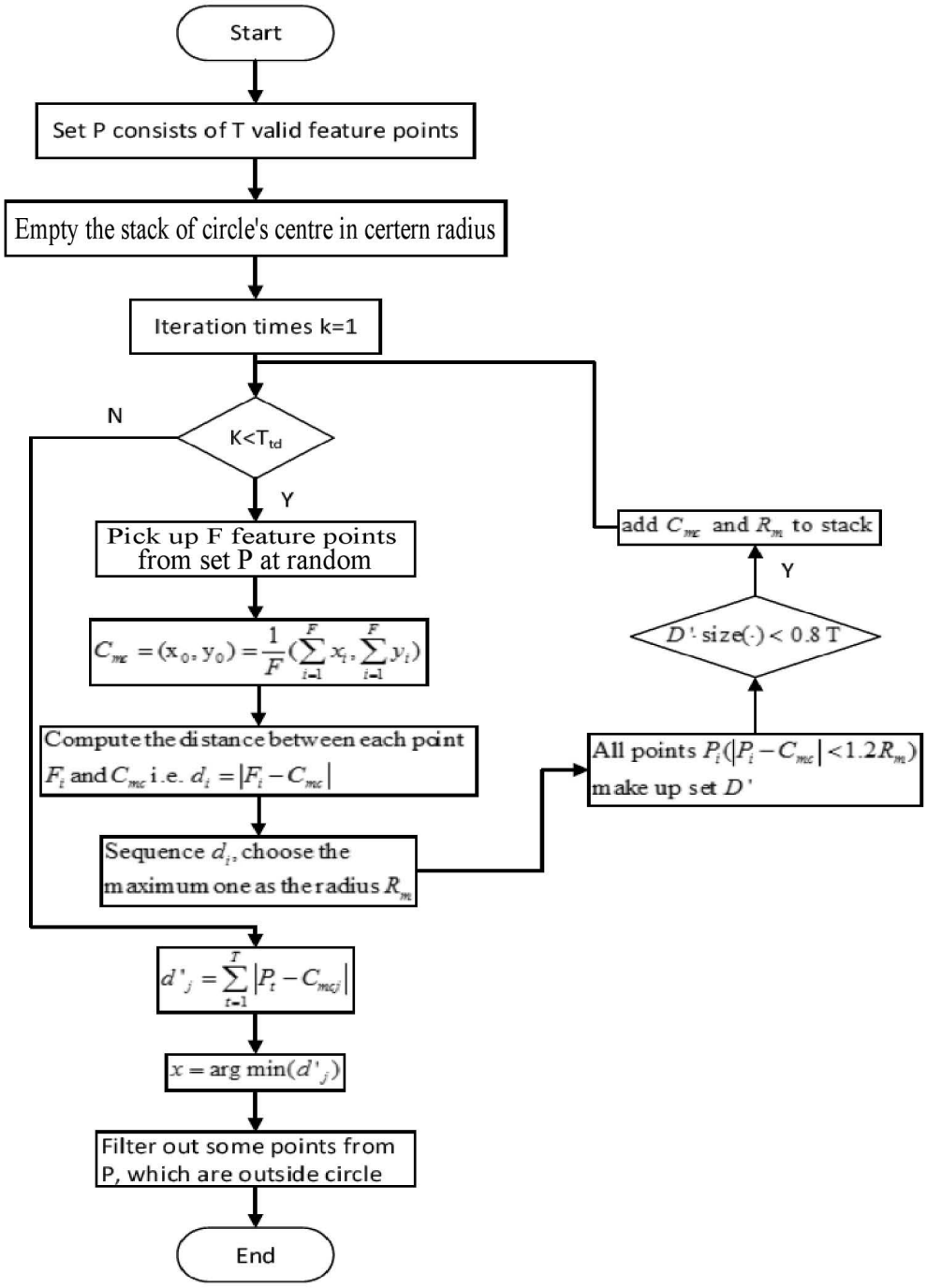

Figure 1: The flowchart of sampling feature after background filter.

// While iteration times $k < T_{td}$ (here it takes 50)

// F key points randomly selected from T data inside model;

// possible centre of a circle is:

$$C_{mc} = (x_0, y_0) = \frac{1}{F}(\sum_{i=1}^{F} x_i, \sum_{i=1}^{F} y_i,) \quad (3)$$

//Sequence the distances between the key points inside model and the possible centre of a circle, choose the maximum one as the radius R_m;

//For every key point which doesn't belong to model, if the distance is less than $R_m * 1.2$, believe this key point to be in this model, and add 1 to then umber of key points in this model;

// If the number of key points in this model is larger than $E(E = 80\% * T)$; Believe this model is right, save the possible centre of the circle and key points in this model;

//If $k > T_{td}$

$$d'(j) = \sum_{t=1}^{T} \left| P_t - R_{mcj} \right| \quad (4)$$

//If $d'(r, r \in [1, T])$ is the minimum value among all distances, save the possible radius R_{mcr}), save these 80% key points in this model which is the nearest to the possible center, believe these key points belong to the image.

After that, we compare the results before and after sampling for background filter. It is obvious that most of feature points gather on the target after sampling method of background filter, which are ready for the latter job.

3.2. Integrate Spatial Relationship

Since the classical BoW model origins from auto-indexing document, the order among visual words isn't considered. That is, the classical BoW only focuses on what there are in a bag, instead of considering their spatial relationship. Although this simple model is highly efficient, it also brings up a serious problem. For example, we only change the position of image patches, so that it represents a different object. However, according to the BoW model, it represents the same one. This is because the BoW model ignores the spatial relationship of patches. Actually, spatial relationship of patches or features is very important to describe an object.

In this paper, we integrate the spatial relationship of visual words, in order to improve the performance of classical BoW. Although we adopt the descriptive way of features according to the classical BoW, that is, each image is represented as a vector with a stable length, the vector is different from that in the classical BoW, which is divided into two categories: 1) the statistics of occurrence of visual words. That is, how many times visual words in vocabulary occur in an image. Suppose there are P words in vocabulary, the histogram dimension of visual words in each image is P, noted $X = (x0, x1, \ldots x_{P-1})$, here x_i represents the occurrence times of visual words. 2) the description of spatial relationship of visual words. The position of each visual word is described or represented by distance and angle, which are relative to the geometrical center of target. After section 3.1, the new geometrical center of target is

$$O = (x_c, y_c) = \frac{1}{m}(\sum_{i=1}^{m} x_i, \sum_{i=1}^{m} y_i) \quad (5)$$

Here m refers to the number of left features after sampling method of background filter. The geometrical center is shown in Figure **2**, where the visual words after quantization are around the center, the same shape signs represent the same visual word occurring in different position. In order to describe the spatial relationship of visual words, we define the following distance and angle.

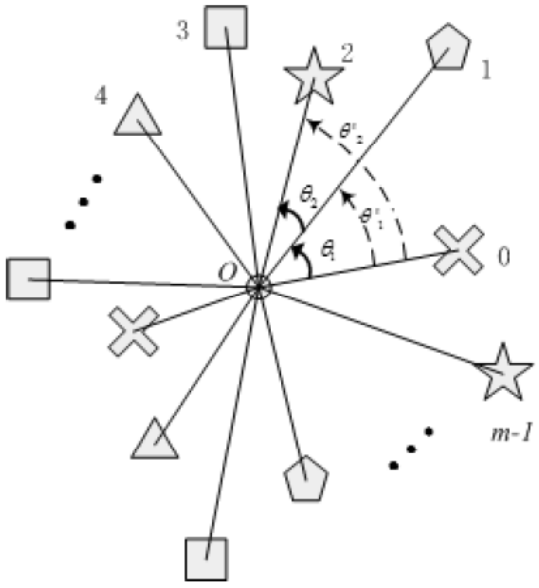

Figure 2: The description of spatial relationship of features.

For distance, compute the Euler distance between each feature and the geometrical center (x_c, y_c), i.e.

$(L_1, L_2, \ldots L_m)$ and take the mid-value as the unit length L, and divide other distances into four zones according to the ratio L_i / L, i.e. $0 \sim 0.5L$, $0.5L \sim L$, $L \sim 1.5L$, $1.5L \sim MAX$. Therefore, the distance of each feature point is quantized as the component of distance histogram. For angle, in Figure **2**, the angle θ is formed by two half lines, arbitrarily choose one feature point as the initial point F_0, and counter-clockwise choose other points as $F_1, F_2, \ldots F_{m-1}$. And then, compute the angle $\theta_i^{'}$ between $\overrightarrow{OF_i}$ and $\overrightarrow{OF_0}, i = 0,1,\ldots m-1$, O is the geometrical center. According to Eq. (6), we can easily get θ_i between $\overrightarrow{OF_i}$ and $\overrightarrow{OF_{i-1}}$.

$$\theta_i = \begin{cases} \theta'_i - \theta'_{i-1} & i = 1,2,\ldots m-1 \\ 360 - \theta'_{m-1} & i = m \end{cases} \quad (6)$$

Generally speaking, we can pick up hundreds of SIFT feature points from one image, moreover, most of them concentrate on target, the angle θ between two feature points should be not too great, so we divide angle θ of different points into 5 zones: $0^o \sim 30^o, 30^o \sim 60^o, 60^o \sim 90^o, 90^o \sim 120^o, 120^o \sim MAX$.

Since both distance and angle are relative, this kind of feature description has good characters of rotational and scale invariance. So each image can be represented as a vector:

$$\{P_i\}_{i=0}^{P-1} + \{Q_i\}_{i=0}^{Q-1} = \{H_i\}_{i=0}^{P+Q-1} \quad (7)$$

Here the former P vectors represent the histogram of visual words, the latter

'Q' ones represent the histogram of spatial relationship of visual words.

3.3. Vocabulary Tree Based on Hierarchical K Mean (HKm) Value Clustering

The concept of vocabulary tree was proposed by John. J. Lee [14], which is an efficient data structure of image retrieval through visual keywords. This is because when the number of words in visual vocabulary becomes great, it doesn't need to scan all keywords to search the matching image, but scan part of them. In addition, for the dynamic environment, the image is always updated, vocabulary tree can be conveniently extended by adding leaf node. In traditional K mean (Km) value clustering method, the clustering number k is not easy to be decided, which suffers from the disturbance of noise around. However, for the layered clustering method, we don't need to decide the clustering number, so we propose a vocabulary tree based on hierarchical K mean value clustering. Here leaf node in hierarchical K mean value clustering method refers to visual words.

In the hierarchical K mean value clustering method, we only need to decide the layer L, which has something to do with data size and decides the depth of layer partition. At first, regard all datum as a clustering class C_1, and continuously partition layers by applying the clustered division way. The procedure of division is explained as follows:

- For each image $p_i, i = 1,2,\ldots n$, we extract its SIFT feature set, noted $F = f_i$, here f_i is a $m_i * 128$ feature vectors. m_i refers to the SIFT feature number on image p_i.
- Construct a vocabulary tree for F, the whole feature set is regarded as a clustered class, which is the root node of vocabulary tree T.
- On the first layer of T, have a K mean value clustering, the feature set F is divided into k copies $\{F_i \mid 1 \le i \le k\}$, and compute the center vector C_i of F.
- Similarly, for each new F_i, we always divided it into k copies according to K mean value. Repeat this procedure until that the depth reaches L. If the vector number of certain a clustered class is less than k, stop the division.

The total node number except the root node is $s = \sum_{l=1}^{L} k^l = \frac{k^{L+1} - k}{k-1}$, each new clustered class is $\{F_{li} \mid 1 \le l \le L, 1 \le i \le k^l\}$, the leaf node in a vocabulary tree is visual word, the maximum number of visual words represented is k^L. However, the actual number will be less than k^L, because some clustered classes stopped dividing further.

The course of constructing vocabulary tree is actually unsupervised, which is prepared for quantization of feature. Vocabulary tree is an efficient index way by computing the similarity between feature vector input and tree node. In fact, the hierarchical K mean value clustering is a Voronoi division for the sampling space. That is, the feature space is divided into many disjoint subsets, and the corresponding node of each feature point can be found in the vocabulary tree. The quantization always starts from the root node.

We compute the similarity between the clustering center and each feature of the index image on each layer, and put the feature into the closest clustering center, till the leaf node of last layer. By far, we complete the quantization course of local image feature input.

Vocabulary tree based on the hierarchical K mean value clustering method has some advantages i.e. high-speed clustering, good scalability and faster sponse. In addition, by contrast of the K mean value method, the hierarchical clustering not only construct vocabulary tree, but also establish an index mechanism, so that it reduces some operational steps and improve the efficiency.

3.4. Weighted Histogram of Visual Words

When the histogram of classical BoW is quantized, the frequency or normalized frequency of visual word is used to represent weight W. Since the frequency is only counted from the single image, it is not enough to stand out the natural distinction of images. For example, there are some images with the same background, which takes big proportions, in this case, it is difficult to distinct these images only according to frequency, so it is necessary to reduce some visual words without great distinctions.

TF-IDF (term frequency-inverse document frequency) is an efficient weighted *scheme* used in text index [15]. TF-IDF weight *scheme* is composed of two parts: 1) TF, term frequency, which is used to weigh how a word describes a document. If a word has high frequency in a document, then, this word plays an important role in it. 2) IDF, inverse document frequency, which weighs the distinction capability of a word in the whole training set. If a word occurs in most of the documents, then this word has the low capability of distinction.

For a document set $D = \{d_j\}$, key words set $K = \{k_i\}, i = 1,2,\ldots t$, w_{ij} is extracted from it. Weight in a document is used to describe the degree of correlation between k_i and d_j. For a key word k_i without occurring in d_j,we define $w_{ij} = 0$. Otherwise, TF_{ij} is defined as follows:

$$TF_{ij} = \frac{n_{ij}}{N_j} \tag{8}$$

Here n_{ij} refers to the frequency that k_i occurs in d_j. N_j refers to the number of key words in d_j. IDF_i is defined in Eq. (9)

$$IDF_i = \log(\frac{|D|}{n'_i}) \tag{9}$$

Here n'_i refers to the frequency that k_i occurs in the document set D. $|D|$ is a constant value, and the document number. w_{ij} is given in Eq. (10)

$$w_{ij} = TF_{ij} \times IDF_i \tag{10}$$

Document d_j may be represented by weight vector in Eq. (11)

$$d_j = [w_{1j}, w_{2j}, \ldots, w_{tj}] \tag{11}$$

Similarly, w'_{ij} refers to the degree of correlation between visual word F_i and image p_j. The frequency of visual word is regarded as the frequency of key word in document. w'_{ij} is defined in Eq. (12)

$$w'_{ij} = m_{ij} \times \lg \frac{N}{n_i} \tag{12}$$

Here m_{ij} refers to the frequency that visual word F_i occurs in image p_i, n_i refers to the frequency that visual word F_i occurs in image set. Image set $P = \{p_i\}$ is expressed as a matrix vector in Eq. (13) through TF-IDF.

$$W = \begin{pmatrix} w_{1,1} & w_{2,1} & \cdots & w_{t,1} \\ w_{1,2} & w_{2,2} & \cdots & w_{t,2} \\ \vdots & \vdots & \vdots & \vdots \\ w_{1,t} & w_{2,t} & \cdots & w_{t,\mathrm{t}} \end{pmatrix} \tag{13}$$

IDF_i is a representation of entropy of information. It may be associated with the visual word in a vocabulary tree. In phrase of training, complete the solution of entropy of information, which reduces the amounts of weight computation.

4. EXPERIMENT RESULTS

For convenience, the method proposed in this paper is named as SBoW (Spatial Bag of words). Object recognition model is established according to the flowchart in Figure **3**.

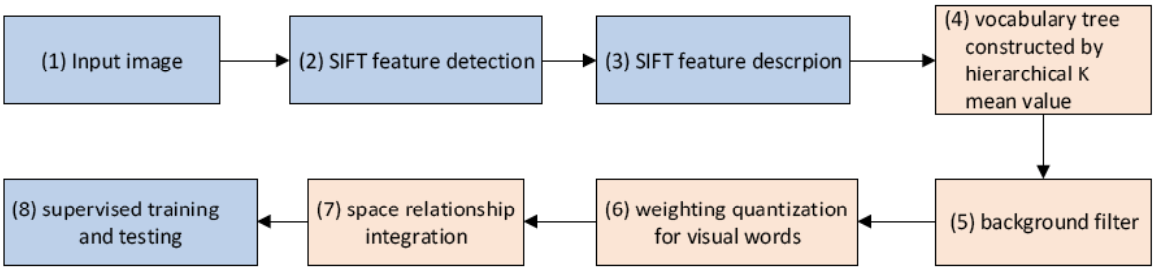

Figure 3: The work flowchart of SBoW.

In order to compare the efficiency between before and after improvement, we choose 4 kinds of objects as samples from the *caltech*256 database, i.e. motorbike, car, human face and airplane [16] (partial images are shown in Figure **4**). Since in the *caltech*256 database, there are less than 120 pieces of images about car, we replenish them from the voc2007 [17] database. There are more than 400 pieces of images for any one of other 3 kinds. After numbering them, for each kind, we choose the former 120 pieces as training samples, the other 100 pieces as testing images. Under the same condition, i.e. training samples and testing samples, we adopt K-mean clustering BoW arithmetic, noted BoW+Km, hierarchical K mean clustering BoW, noted BoW+HKm and SBoW proposed in this paper respectively. This experiment is conducted on a PC with Intel E7200 2.53GHz CPU and the compiling environment of OpenCV2.1 and C++.

In the experiment, we adopt the support vector machine (SVM) to realize the classification. Since the SVM is a diaschistic and supervised classifier, the Multi-Class SVM classifier may be constructed by one-vs-one way. That is, here arbitrarily choose 2 from 4 training sample classes as a suite, and get $C_4^2 = 6$ SVMs. The sample in every SVM is represented to be vector (x, y). If $y = +1$, and it belongs to class A; If $y = -1$, and it doesn't belong to class A.

Figure 4: The partial sample images for experiment.

Table 1: Hybrid Matrix Comparison from Different Methods, i.e. a=BoW+Km, b=BoW+HKm, c=SBOW

	motorbikes			car side			faces			airplanes		
	a	b	c	a	b	c	a	b	c	a	b	c
motorbikes	80	91	95	0	5	4	0	0	0	0	1	1
car side	12	7	4	89	80	86	6	5	4	7	9	4
faces	8	1	0	3	3	0	90	90	91	8	0	2
airplanes	0	1	1	8	12	10	4	5	5	85	90	93
accuracy	0.8	0.91	0.95	0.89	0.8	0.86	0.9	0.9	0.91	0.85	0.9	0.93

Shown in Figure **4**, there is a SVM for any two classes. When we are carrying out a test, we input the testing sample (x, y) into the SVM. That category given the most votes is regarded as the real category of the testing sample.

We all adopt linear kernel functions in the 3 methods (i.e. BoW+Km, BoW+HKm and SBoW) and divide all samples into 8 copies by applying cross validation method [18], where the parameters are automatically optimized. In order to produce the vocabulary with the same size, we choose $L = 3, k = 8$ in HKm and $k = 512$ in Km. The classifying results are given in hybrid matrix form, where the element H_{ij} represents that there are H_{ij} j-category objects justified as i-category. The hybrid matrix results are given in Table **1** by applying three methods (i.e. BoW+Km, BoW+HKm and SBoW), where SBoW has higher accuracy than the other two.

It is important for an object recognition system to evaluate the performance, which will decide whether this system satisfies the requirement of application. In this paper, performance evaluation is done with R(recall) in Eq. (14) and P(precision) in Eq. (15).

$$R(i) = \frac{n_c}{N_i} \tag{14}$$

$$P(i) = \frac{n_c}{n_c + n_f} \tag{15}$$

Here N_i refers to the sample amount which belongs to i-category assuredly. n_c refers to the sample amount which is rightly justified as i-category. n_f refers to the sample amount which is falsely justified as i-category. R (recall) reflects how many samples to be recognized are rightly recognized. P(precision) reflects the reliable degree, that is , the rate that sample to be justified as some certain category is rightly justified. The R (recall)and P(precision) by comparing three methods are shown in Figure **5** and Figure **6**. Seen the experimental result, SBoW has a more distinct improvement than classical BoW in both R (recall) and P (precision), especially, there is a5% improvement in R(recall).

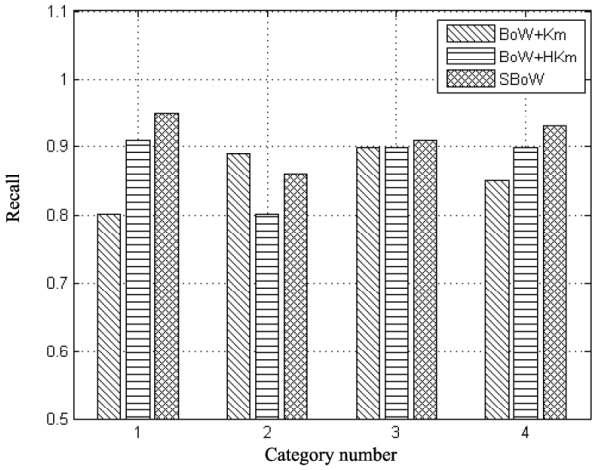

Figure 5: The comparison of recall from different methods.

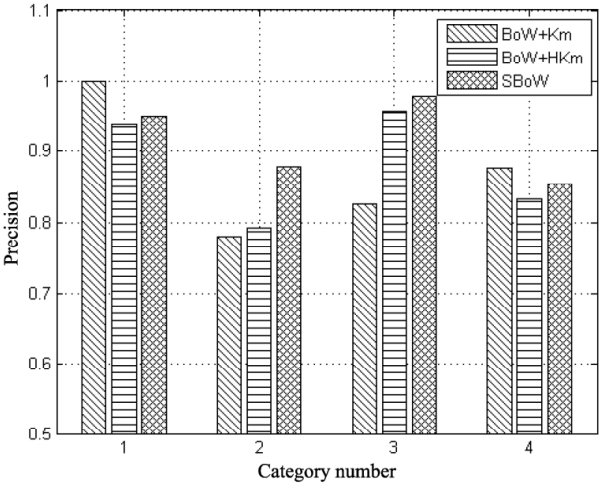

Figure 6: The comparison of precision from different methods.

SBoW not only has the more recognition rate, but also greatly reduce the time of producing vocabulary. The time-consuming is compared between HKm and

Km under the same condition (i.e. sample number and clustering number) through 3 times experiments in Figure **7**. Since HKm adopts the hierarchical clustering way, it only needs 8 clusters on every branch. With the depth of layer becoming more, the clustering amount reduces distinctly. However, Km needs 512 clusters. More the clustering amount is, more the time-consuming of clustering optimism is. In addition, when the sample number is very great, Km might fall into suboptimum.

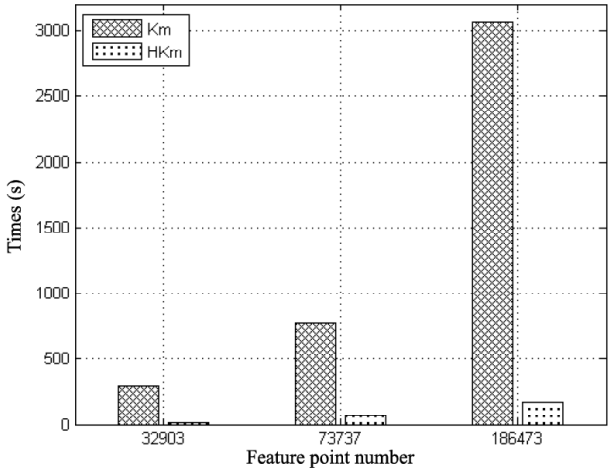

Figure 7: The comparison of clustering time from different methods.

Analysis of results: Seen from these experiments, SBoW has less consuming-time than classical BoW, since it adopts HKm. In addition, SBoW avoids the deficiency of misrecognition, which is caused by the same components, but the distinctive spatial pose of components, because it considers spatial relationship of components. And even, it has strong ability in resisting on the distribution of background, because it adopts the sampling feature of background filter.

5. CONCLUSION

In this paper, in order to make BoW model play a greater role in visual recognition and categorization, we improve and extend it from four ways, noted SBoW. Experimental results show SBoW has higher accuracy, efficiency and stronger ability in resisting on the distribution of background noise than classical BoW. This work is very significant for visual recognition and categorization of generic objects in the field of pattern recognition and artificial intelligence.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 60804063, 61175091), Natural Science Foundation of Jiangsu Province under Grant (No. BK2010403), the Public Funds of Image Processing and Intelligent Control Key Laboratory of Chinese Education Ministry under Grant (No. 200902) and The Science and Technology Innovation Foundation in Southeast University under Grant (No. 3208000501). The authors are very grateful to the anonymous reviewers for their valuable remarks which have permitted to clarify and improve the overall quality of this paper.

REFERENCES

[1] Teng L, Tao M, So KI, Sheng HX. Contextual Bag-of-Words for Visual Categorization. IEEE Transactions on Circits and Systems for Vidio Technology 2011; 21(4): 381-392. http://dx.doi.org/10.1109/TCSVT.2010.2041828

[2] Tinglin L, Jing L, Qinshan L, Hangqing L. Expanded bag of words representation for object classification. 2009 16th IEEE International Conference on Image Processing 2009; 297-300: 7-10 Nov.

[3] Agarwal S, Awan A, Roth D. Learning to detect objects in images *via* a sparse. part-based representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2004; 26(11): 1475-1490. http://dx.doi.org/10.1109/TPAMI.2004.108

[4] Xianglong L, Yihua L, Wei YA, Bo L. Search by mobile image based on visual and spatial consistency. 2011 IEEE International Conference on Multimedia and Expo 2011; pp. 1-6.

[5] Xiangang C, Jingdong W, Liangtien C, Xiansheng H. Learning to combine multi-resolution spatially-weighted co-occurrence matrices for image representation. 2010 IEEE International Conference on Multimedia and Expo 2010; pp. 631-636.

[6] Meng S, Van H. Image pattern discovery by using the spatial closenessof visual code words. 2011 18th IEEE International Conference on Image Processing 2011; pp. 205-208.

[7] Elsayad I, Martinet J, Urruty T, Djeraba C. A new spatial weighting *scheme* for bag-of-visual-words. 2010 International Workshop on Content-Based Multimedia Indexing 2010; pp. 1-6.

[8] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE Computer Vision and Pattern Recognition. New York, USA: IEEE 2006; pp. 2169-2178.

[9] Zhang L, Wang C, Xiao B, Shao Y. Image Representation Using Bag-of-phrases. ACTA AUTOMATICA SINICA 2012; 38(1): 46-54. http://dx.doi.org/10.3724/SP.J.1004.2012.00046

[10] Bouachir W, Kardouchi M, Belacel N. Improving Bag of Visual Words Image Retrieval: A Fuzzy Weighting *Scheme* for Efficient Indexation, 2009 Fifth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS) 2009; pp. 215-220. http://dx.doi.org/10.1109/SITIS.2009.43

[11] Liu C, Yang Y, Chen Y. Constructing Visual Vocabularies Using Sparse Coding for Action Recognition, ICIECS 2009. International Conference On Information Engineering and Computer Science 2009; pp.1-4.

[12] Wang J, Li Y, Zhang Y, Wang C, Xie H, Chen G, Gao X. Bag-of-Features Based Medical Image Retrieval *via* Multiple Assignment and Visual Words Weighting. IEEE Transactions on Medical Imaging 2011: 30(11): 1996-2011. http://dx.doi.org/10.1109/TMI.2011.2161673

[13] Martin A, Fischler, Bolles RC. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Comm of the ACM 1981; 24(6): 381-395. http://dx.doi.org/10.1145/358669.358692

[14] Yeh T, Lee J, Darrell T. Adaptive Vocabulary Forests for Dynamic Indexing and Category Learning. In Proceeding of IEEE 11th International Conference on Computer Vision 2007; 1-8.

[15] SpÄarck Jones Karen. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1972; 28(1): 11-21. http://dx.doi.org/10.1108/eb026526

[16] Caltech 256 dataset [EB/OL] http://www.vision.caltech.edu/ImageDatasets/Caltech 256/.

[17] VOC 2007 dataset [EB/OL] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc 2007/.

[18] Chung CC, Jen LC. LIBSVM: a library for support vector machines ACM Transactions on Intelligent Systems and Technology 2011; 2(27): 1-27.

Received on 25-10-2014 Accepted on 05-11-2014 Published on 09-01-2015

https://doi.org/10.15377/2409-9694.2014.01.02.3