Bag of Visual Words Method based on PLSA and Chi-Square Model for Object Category

Yongwei Zhao[1], Tianqiang Peng[2], Bicheng Li[1], Shengcai Ke[1]
[1]China National Digital Switching System Engineering and Technological R&D Center
Zhengzhou, Henan 450002 - P. R. China
[2]Department of Computer Science and Engineering, Henan Institute of Engineering
Zhengzhou, Henan 451191 -P. R. China
[e-mail: zhaoyongwei369@163.com,pengtianqiang@zzjinhui.com, {lbclm, keshengcai0705}@163.com]
*Corresponding author: Yongwei Zhao

Received January 29, 2015; revised May 5, 2015; revised May 28, 2015; accepted June 8, 2015; published July 31, 2015

Abstract

The problem of visual words' synonymy and ambiguity always exist in the conventional bag of visual words (BoVW) model based object category methods. Besides, the noisy visual words, so-called "visual stop-words" will degrade the semantic resolution of visual dictionary. In view of this, a novel bag of visual words method based on PLSA and chi-square model for object category is proposed. Firstly, Probabilistic Latent Semantic Analysis (PLSA) is used to analyze the semantic co-occurrence probability of visual words, infer the latent semantic topics in images, and get the latent topic distributions induced by the words. Secondly, the KL divergence is adopt to measure the semantic distance between visual words, which can get semantically related homoionym. Then, adaptive soft-assignment strategy is combined to realize the soft mapping between SIFT features and some homoionym. Finally, the chi-square model is introduced to eliminate the "visual stop-words" and reconstruct the visual vocabulary histograms. Moreover, SVM (Support Vector Machine) is applied to accomplish object classification. Experimental results indicated that the synonymy and ambiguity problems of visual words can be overcome effectively. The distinguish ability of visual semantic resolution as well as the object classification performance are substantially boosted compared with the traditional methods.

Keywords: Bag of Visual Words Method; Probabilistic Latent Semantic Analysis; K-L divergence; Chi-Square Model; Object Category

This research was supported by the National Natural Science Foundation of China [Grant No. 60872142 and Grant No. 61301232].

http://dx.doi.org/10.3837/tiis.2015.07.017 **ISSN : 1976-7277**

1. Introduction

With the rapid development and widespread use of computer and communication technologies, the environment for huge image information is generated. Object image classification is a fundamental problem in the fields of computer vision and image understanding. Its intention is to categorize unlabeled images into the pre-defined classes according to their semantic meanings. In the last decades, BoVW model based methods [1-5] have achieved good classification performances on many image data sets. It consists of four major steps shown as in **Fig. 1**, namely:1) descriptor extraction; 2) feature coding; 3) spatial pooling; and 4) support vector machine (SVM) classification, to classify an image into its semantic category. In a typical setup, gradient-based local image descriptors, such as scale-invariant feature transform (SIFT) [6], PCA-SIFT [7], SURF [8] and so on. These descriptors are all invariant to various image degradations, such as geometric and photometric transformations, which is essential when addressing image categorization problems. The experiment results of literature [9] indicated that SIFT presents its stability in most situations and SURF is the fastest one. Since the environment of object classification experiments is complicated, the stability of the image features is especially important. So, in our paper ,we choose SIFT feature.

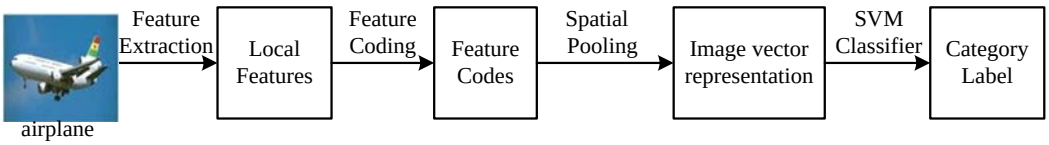

Fig. 1. The BoVW framework for object category

There are several different ways to encode the local features (SIFT) and generate feature codes, which can be seen as a visual dictionary, such as vector quantization(VQ) [1-5], sparse coding [10, 11], or Fisher kernels [12]. The vector quantization coding methods treat an image as a collection of unordered appearance descriptors extracted from local patches, quantizes them into discrete “visual words”, and then computes a compact histogram representation for image classification. However, there exists some synonymy and ambiguity problems in visual words [13-15] as well as the seriously quantization error of compact histogram representation. In addition, due to the existence of image background noise and the limitation of clustering algorithms [16, 17], some visual words generated from SIFT features may be less meaningful to express useful image content, but decrease the semantic resolution ability of visual dictionary. These noise words are similar to stop words, such as “the” “and” “is”, existing in text documents. In this context, we call them “visual stop-words”.

Sparse coding computes a spatial-pyramid image representation based on sparse codes of SIFT features, instead of the K-means vector quantization in the traditional BoVW. Yang et al. [10] proposed the ScSPM method where sparse coding was used instead of vector quantization to obtain nonlinear codes. Wang et al. [11] presented a Locality-constrained Linear Coding (LLC), which can be seem as a fast implementation of LCC that utilizes the locality constraint to project each descriptor into its local-coordinate system. Moreover, He et al. [18] proposed a spatial pyramid pooling in deep convolutional networks. Furthermore, unlike the original BoVW model that performs spatial pooling by computing histograms, the sparse coding

approaches always use max spatial pooling [10, 11] that is more robust to local spatial translations and more biological plausible

The Fisher kernel [12] is a powerful tool to transform an incoming variable-size set of independent samples into a fixed size vector representation, assuming that the samples follow a parametric generative model estimated on a training set. This description vector is the gradient of the sample's likelihood with respect to the parameters of this distribution, scaled by the inverse square root of the Fisher information matrix. Jegou et al. [19] proposed VLAD representation derived from both original BoVW and Fisher kernel, that aggregates SIFT descriptors and produces a compact representation. Although there are several feature coding approaches and spatial pooling schemes, it should be pointed out that the main work of our article is aim at solving the synonymy and ambiguity of visual words, "visual stop-words" etc problems of the traditional vector quantization coding based BoVW model.

2. Related Work

In order to overcome the influence of these negative factors brought by synonymy and ambiguity of visual words, many researchers have made lots of explorations and attempts. Philbin et al. [20] presented a kind of BOVW model based on soft-assignment to build the visual vocabulary histogram. In which, a SIFT feature point is assigned to several nearest visual words, and each word is weighted according to the distance. Gemert et al. [15] established a visual word uncertainty model, where some kernel functions were carried out to complete soft-mapping between local features and visual words. This model can efficiently decrease the quantization error, as well as further verify the effectiveness of soft assignment method in solving the synonymy and ambiguity problem of visual words. Li et al. [21] constructs the histograms using a kind of context information strategy to improve the mapping accuracy. To some extent, it can also reduce the quantization error caused by words synonymy and ambiguity. Weinshall et al. [22] proposed a Latent Dirichlet Allocation model based soft-assignment method (LDA+SA). Considering the ambiguity effect of visual words, Danilo et al. [23] introduced a fuzzy clustering algorithm to complete the soft-assignment of visual words, which achieved good results.

In comparison with the conventional hard-assignment based BoVW model (BoVW+HA) [14], the above mentioned methods, all can overcome the problem of synonymy and ambiguity on visual words, reduce the quantization error, and enhance the semantic expression ability of histograms. However, the weakness for these methods is that they all measure the semantic distance between words in low feature space. Due to the inconsistency of metric space, the visual words are not relatively close in semantic space as they are in feature space. Furthermore, all these methods [20-23] assign each SIFT feature with the same number of visual words, which will lead to new noisy and redundant information for the reason that some local features without ambiguity are mandatorily mapped to multiple visual words. Therefore, once semantic relevance of visual words is accurately measured, and the number of soft-assignment words is adaptively selected according to different categories of SIFT features, both problems of synonymy and ambiguity in visual words as well as the serious quantization error, could be overcome significantly.

Moreover, The removal of "visual stop-words" will not cause a significant content loss but improve the classification accuracy significantly. Considering the relationship between the capacity of the words information and their appearance frequency, Sivic et al. [1] proposed a method to eliminate "visual stop-words" based on term frequency. Yuan et al. [24] proposed a

solution using the "visual phrase" technique based on an improved frequency itemset mining algorithm and a likelihood ratio test method, but this method only considers the co-occurrence information among visual words and ignores the spatial information of visual words. Chen et al. [25] gave a high discriminative visual phrase (DVP) method which can filter noise efficiently overcoming the problem of feature information loss in traditional visual phrases construction methods [26]. Roman et al. [27] proposed a new methodology for the automatic estimation of the optimal amount of visual words that can be removed from a visual dictionary. This method relies on a special definition of the entropy of each visual word when considered as a random variable, and a new definition of the overlap of class models computed with a normalized Bhattacharyya coefficient.

However, these methods all ignore the interrelationships between the visual words and semantic concepts of different images. Therefore, some visual words with less occurrence frequency but with high discrimination are easily mistaken for "visual stop-words".

For the purpose that identifying the semantic relevance more accurately among visual words, selecting soft assignment words number for different local features adaptively, as well as eliminating "visual stop-words" effectively, a novel bag of visual words method based on PLSA and chi-square model for object category is proposed in this paper. The main contribution is to mine image semantic topics using PLSA model and K-L divergence, accurately measuring the semantic distance between words. Meanwhile, the analyzing of the ambiguity of SIFT features perform soft-assignment more accurately between features and homoionym. Based on that, some "visual stop-words" are eliminated by chi-square model, Therefore, the method described in this paper can effectively solve the problem of synonymy and ambiguity of visual words, and improve the image classification accuracy by enhancing distinguishing ability of visual dictionary.

3. Bag of visual words method based on PLSA and chi-square model for object category

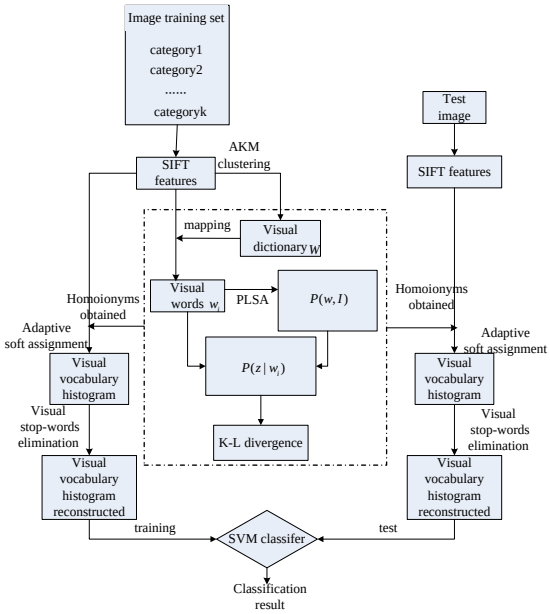

Fig. 2. The flow of bag of visual words method based on PLSA and chi-square model for object category

For training image dataset $I = \{I_1, I_2, \ldots I_k\}$, the method proposed in [6] is used to extract SIFT features and approximate K-Means algorithm (AKM) is adopt to generate visual dictionary. The entire process of bag of visual words method based on PLSA and chi-square model for object category is shown as **Fig. 2**. Firstly, PLSA is used to analyze the semantic co-occurrence probability of visual words and infer the latent image semantic topics, and the conditional probability of a specific word w given the unobserved latent topic z can be obtained. Secondly, Bayesian estimation is utilized to infer the latent topic distribution induced by individual word probability $P(z \mid w)$ and K-L divergence is used to measure the semantic distance between visual words and get the homoionyms which have similar semantic. Then, the mapping between SIFT features and words with similar semantic is completed by adaptive soft-assignment. Finally, chi-square model is introduced to analyze the correlations between visual words and various image categories and a number of "visual stop-words" with weak correlation is eliminated to reconstruct the visual vocabulary histograms, and the object classification is completed through SVM classifier at last.

3.1 Visual semantic concept expression and measurement

The reasons leading to the problem of "Semantic Gap" widely exists in the field of computer vision, which could be explained by the inconsistency between the feature space and semantic space on distance measuring. Therefore, the traditional methods [15, 22], which measure the semantic distance in Euclidean space, could not accurately reflect the actual semantic relevance between visual words. The method described in reference [21] represents the semantic concepts by getting the conditional probability distribution of a specific word w given the image category and hence achieves satisfied classification accuracy. But a precondition that unable to contain the same semantic concept in different categories of images is enforced in this method. Obviously, the precondition is difficult to ensure in real-world applications. While using PLSA model, the conditional probability distribution of a specific word w given the unobserved latent topic z can be obtained, and the semantic concept implied in the word can be expressed more accurately. In the following section, the method of PLSA model based visual words expression will be introduced.

1 Visual semantic concept expression based on PLSA model

PLSA proposed by Hoffman et al in [28] is a topic generated model for the latent semantic analysis, which is widely used in machine learning and information retrieval. For training image set $I = \{I_1, I_2, \ldots I_k\}$ and the visual words $W = \{w_1, w_2, \ldots, w_n\}$ generated by AKM clustering, we can get the co-occurrence frequency matrix $N = [n(w_i, I_j)]$ of images and visual words, where, $n(w_i, I_j)$ is the number of times w_i appeared in image I_j. The joint probability of (w, I) can be calculated as Equation (1):

$$\begin{aligned} P(w, I) &= P(I)P(w \mid I) \\ &= \sum_{z \in Z} P(w \mid z)P(z)P(I \mid z) \end{aligned} \tag{1}$$

Where Z indicates all topics in the latent semantic space. According to the maximum likelihood principle, the $P(z)$, $P(w \mid z)$ and $P(I \mid z)$ can be obtained by Equation (2) through EM algorithm.

$$L = \log P(\mathrm{I}, W) = \sum_{I \in \mathrm{I}} \sum_{w \in W} n(w, I) \log P(w, I)$$
$$s.t. \sum_{z \in Z} P(z)=1, \sum_{w \in W} P(w|z)=1, \sum_{I \in \mathrm{I}} P(I|z)=1 \tag{2}$$

Then, Bayesian estimation is utilized to infer the word w occurrence probability $P(w)$ and the latent topic distribution induced by individual word $P(z \mid w)$ as:

$$P(w) = \sum_{z \in Z} P(w|z) P(z) \tag{3}$$

$$P(z \mid w) = \frac{P(z, w)}{P(w)} = \frac{P(w \mid z) P(z)}{\sum_{z \in Z} P(w \mid z) P(z)} \tag{4}$$

It should be noted that the number of topics in present PLSA model is mostly a fixed value which is a set artificially from experience [29]. The topic model is trained and finally gets the image semantic representation about the fixed topic set. This method with setting topic numbers artificially overlooks the situation that the content ranges from sample to complex among different image categories. In view of this, here we use a density based adaptive topic number selection method of the PLSA model [30]. When building this topic model for semantic content of different image categories, the method could get much better topic analyzing results for its automatically setting the number of semantic topic according to the complexity of image content.

2 Semantic distance measurement with K-L divergence

For the reason that the same image may contain more than one latent semantic topic, and for semantic topics' difference, they have different contributions to express the semantic content of images. In consequence, we need to weight these semantic topics adaptively. Inspired by the literature [25], here we use K-L divergence [21] to measure semantic distance between visual words, and the conditional entropy H is applied to measure discrimination of each topic. H could be represented as,

$$H(\mathrm{I} \mid z \in Z) = -\sum_{I \in \mathrm{I}} P(I, z) \log (P(I|z)) \tag{5}$$

It is easy to see from Equation (5) that the higher value of H presents the latent topic z has less discriminative capability. Then, a Gaussian function is used for normalization of $H(\mathrm{I} \mid z \in Z)$ to generate the discrimination weight $\omega(z)$ as,

$$\omega(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} H^2(\mathrm{I}|z)} \tag{6}$$

However, K-L divergence is asymmetric meaning that, it cannot always guarantee $d(w_i, w_j)$=$d(w_j, w_i)$. Hence, the semantic distance $d(w_i, w_j)$ between the visual words w_i and w_j can be calculated as,

$$d(w_i, w_j) = \frac{P(w_i)}{P(w_i) + P(w_j)} \cdot \sum_{z \in Z} \omega(z) P(z \mid w_i) \log \frac{P(z \mid w_i)}{P(z \mid w_j)} + \frac{P(w_j)}{P(w_i) + P(w_j)} \cdot \sum_{z \in Z} \omega(z) P(z \mid w_j) \log \frac{P(z \mid w_j)}{P(z \mid w_i)} \tag{7}$$

According to Equation (3) - Equation (7), we can calculate the semantic distance between visual words and get the semantically related homoionym.

3.2 Adaptive soft-assignment

After getting the homoionyms through PLSA model and K-L divergence, to realize constructing visual vocabulary distribution histogram with adaptive soft-assignment, we first need to analyze the fuzziness caused by mapping the SIFT features. The fuzziness diagram is shown in **Fig. 3**. As **Fig. 3** shows, the little dot represents SIFT feature, oval represents visual word and diamond and square denote two SIFT features with different fuzziness. For diamond feature, if it is only closest to visual word w_1 and farther from other visual words, we can assume its semantic content can be expressed by visual word w_1 and the feature point has no or little fuzziness defined as the first class feature. For square feature, if it has a close range with the distance between visual words w_2 and w_3 (or among more words), we can assume its semantic content should be expressed by w_2 and w_3 or more visual words together. That is, this kind of feature point is fuzzier and is defined as the second class feature.

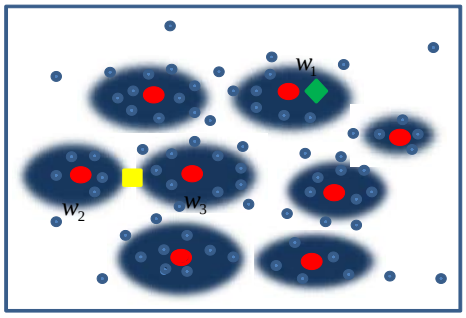

Fig. 3. The sketch map of SIFT features fuzzification

Suppose that the visual dictionary is defined by $W = \{w_1, w_2, \ldots, w_n\}$, where n denotes the size of visual dictionary. The different classes of SIFT features can be adaptively mapped to a certain number words by adaptive soft-assignment strategy by calculating the distance between each SIFT feature to the homoionym based on the method mentioned in section 2.1 and distinguishing different kind of SIFT features. The entire process can be described as follows.

Step1: For image $I = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_i, ..., \boldsymbol{r}_T\}$, T denotes the number of SIFT features in image I. We first calculate the closest visual word w_k^1 to SIFT feature $\boldsymbol{r}_i$ in visual dictionary W;

Step2: According to Equation (7) described in section 2.1, we can get the m semantic closest visual words $w_k^j, (1 \le j \le m)$ to word w_k^1 in visual dictionary. Then, calculate the Euclidean distance between SIFT features and each of the m words respectively, and sign the distance from smallest to largest as $d = \{d_1, d_2, ..., d_j, ..., d_m\}$, here d_j indicates the j closest distance between words and feature points.

Step3: Based on principle $N_{adp} = \arg\max_i \{d_i \le \alpha \cdot d_1\}$, $(i = 1,2,\ldots,m)$, the number of visual words N_{adp} assigned to $\boldsymbol{r}_i$ can be determined adaptively. And each word is weighted with $e^{-\frac{d_l^2}{2\sigma^2}} \left(l = 1,2,...,N_{adp} \le m\right)$, where α is the "adaptive soft-assignment factor" that used to control the assigned words number, usually $\alpha \ge 1$. Repeat the above-mentioned process, visual vocabulary distribution histograms can be constructed with adaptive soft-assignment.

3.3 "Visual stop-words" elimination

Chi-Square model [31] is often used to measure the independence of two random variables. Here, we utilize chi-square model to perform statistical analysis on the correlations between visual words and each category of images, as well as to discover visual stop words and eliminate them. The smaller chi-square value means the less correlation between the visual word and each image categories, indicating the weaker discrimination, and vice versa. Therefore, combined with chi-square model and term frequency, the "visual stop-words" can be eliminated more efficiently. Here, assuming the appearance frequency of visual word w is independent of any image category I_j, $I_j \in I$ $(1 \le j \le k)$ where $I = \{I_1, I_2, \ldots I_k\}$ is the training image set. The interrelationship between visual word w and image category in training set I can be described by **Table 1**.

Table 1. The statistical relationships between visual word w and each image category

categories / images	I_1	I_2	…	I_k	**Total**
w appear	n_{11}	n_{12}	…	n_{1k}	n_{1+}
w not appear	n_{21}	n_{22}	…	n_{2k}	n_{2+}
Total	n_{+1}	n_{+2}	…	n_{+k}	N

where, N is the total number of images in I, n_{1j} denotes the number of images that contain word w in category I_j, n_{2j} is the number of images that don't contain w in category I_j, n_{+j} denotes the total number of images in category I_j, n_{i+} $(i=1,2)$ is the total number of images that contain w and the total number of images that don't contain w in image training set I respectively. Here, the chi-square value between visual word w and each image category can be calculated as,

$$x^2 = \delta = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(Nn_{ij} - n_{i+}n_{+j})}{Nn_{i+}n_{+j}} \tag{8}$$

The chi-square value indicates the different degrees of statistical correlation between w and each image category. Moreover, considering the influence of word frequency, the chi-square value is given a corresponding weight as,

$$\tilde{x}^2 = \frac{x^2}{tf(w)} \tag{9}$$

Where $tf(w)$ indicates the term frequency of w. It's easy to indicate that Equation (9) accounts for both the word frequency of w and the statistical correlation between w and each image categories. Therefore, a percentage of (recorded as S) visual dictionary can be removed as "visual stop words" according to the chi-square values from small to large. And the corresponding dimensions of words would be eliminated when constructing the visual vocabulary histograms.

4. Experiments

4.1 Experimental dataset setup and evaluation

In this experiment, we use the standard test image collections Caltech-256 [32] and Pascal Voc 2007[33] to evaluate object classification performance. The Caltech-256 dataset holds 29,780 images falling into 256 categories with much higher intra-class variability and higher object location variability compared with Caltech-101. Each category contains at least 80 images. The Pascal Voc 2007 involves 9,963 images in 20 categories. 5,011 images are for training, and the rest are for testing. Firstly, we do some experiments on the whole Caltech-256 dataset to evaluate the effectiveness of PLSA, adaptive soft-assignment and the chi-square model respectively. 50 images are choose in each category to construct training image set for generating visual dictionary and the remaining are as testing set. The visual dictionary size is 15K. The SVM classifier is employed here, particularly LIBSVM [34] which kernel function is Radical Basis Function. To obtain reliable experimental results, all object classification experiments are run 10 times and then averaged to produce the final average precision. The hardware configuration for experiment is a desktop with Core 3.1G×4 CPU and 4G of Ram. The performance criteria of object classification are recall rate, accuracy rate, and confusion matrix based on recall rate and Average Precision (AP). The related definitions are as follows,

$$\text{Recall} = \frac{\text{correctly classify image numbers}}{\text{the total image numbers of one category}} \times 100\% \tag{10}$$

$$\text{Precision} = \frac{\text{correctly classify image numbers}}{\text{the total classify image numbers}} \times 100\% \tag{11}$$

$$\text{Average precision} = \frac{\text{sum of precision}}{\text{the total number of image categories}} \tag{12}$$

4.2 Experimental results

First of all, in order to evaluate the effectiveness of PLSA based on Soft-Assignment method (PLSA+SA) on overcoming the synonymy and ambiguity problem of visual words, we compare it to the traditional soft- assignment (SA) method [35] and Hard-Assignment (HA) method [14] respectively. **Fig. 4** depicts the relationship between average precision of different methods and the number of soft-assignment words. It can be concluded that from **Fig. 4**, the average precision of SA method and PLSA+SA method presented in this article is higher than HA method. As Hard-Assignment method is to assign each SIFT feature to the nearest

single word, its AP values are always 62.7% and does not change with the soft assignment numbers. In contrast, the AP scores of SA method and PLAS+SA method increase the soft-assignment word numbers firstly; however, when the number exceeds a certain value, the average precision is decreasing. The reason is that too few soft-assignments will not be adequate to express the content of feature points, while too many of them will lead to excessive assignment and introduce new redundant information. Moreover, the PLSA+SA method in this paper can analyze the similarity between words from the semantic concepts content, and then assign the corresponding feature points to a number of visual words that with similar semantic concepts. Therefore, the PLSA+SA method proposed in this paper can better overcome the quantization error and other problems that brought by synonymy and ambiguity of visual words, as well as the average precision is also superior to the traditional SA method.

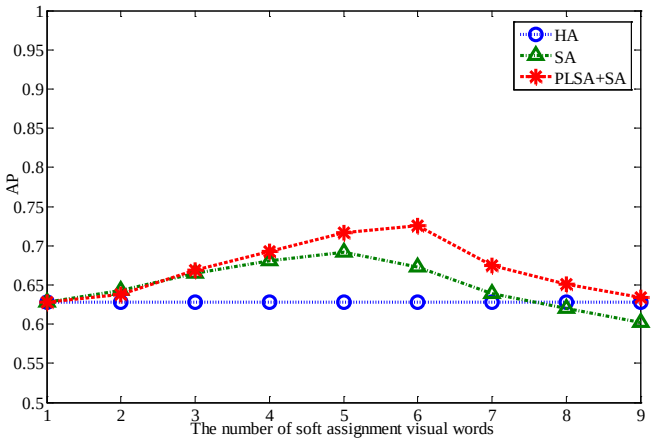

Fig. 4. The AP values comparison of different methods

Note that in the experiment of **Fig. 4**, we assign the same number of words to each SIFT feature point without considering the differences of SIFT features. This will inevitably make some unambiguity local features map to multiple visual words, and introduce new noise and redundant information. By the content of section 3.2, it can overcome this problem by analyzing the ambiguity category of SIFT features and then implementing the adaptive soft-assignment method. Hence, after obtaining homoionym using PLSA model, to verify the effectiveness of this adaptive soft-assignment and analyze how it changes over adaptive soft-assignment factor α , we make object classification experiments with the traditional soft-assignment method (PLSA+SA) and adaptive soft-assignment method (PLSA+ASA), respectively. Setting the value m of adaptive soft-assignment method in section 3.2 is equal to 20, and for PLSA+SA method, we set the soft-assignment word numbers as 6 and the average precision is 71.89%. The results of AP values for object classification are shown as **Fig. 5**. From **Fig. 5** it can be seen that when the factor α gets larger, the SIFT features with different fuzzy category can be more accurately assigned to a number of homoionym, and the average classification accuracy of PLSA+ASA method also increases. When $\alpha = 2.2$, the AP score reaches a maximum 75.47%, which is superior to that of PLSA+SA method. However, when the value of α increases to a certain degree, its AP score tends to decrease to some extent, for the reason that too large of α can also cause over assignment problem which is usually occurred in traditional soft-assignment method.

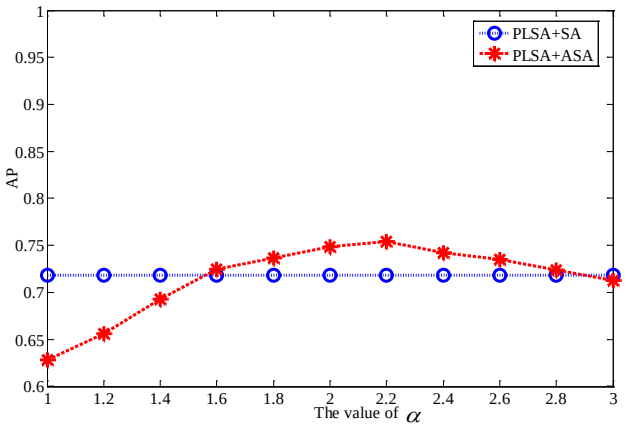

Fig. 5. The influence for AP of factor α

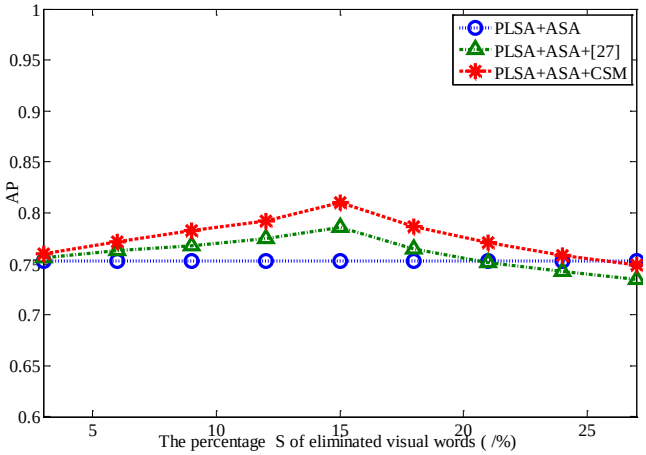

Fig. 6. The influence for AP of the percentage S

Then, in order to evaluate the impact of the percentage S for AP result and compare it to PLSA+ASA without "visual stop-words" elimination as well as PLSA+ASA based on literature [27]. The AP values of different methods are shown in **Fig. 6**. From **Fig. 6** it can be seen that eliminate a certain percentage of visual words by chi-square model and the method of [27] both can improve the average precision of object classification. Moreover, their AP values will be maximized when the percentage of removed visual words are S=15% and the performance of our method PLSA + ASA + CSM is better than the method PLSA+ASA based on literature [27]. However, it inevitable eliminate some words with strong representation when the percentage of removed visual words is too big, that will greatly reduce the classification performance.

Furthermore, **Fig. 7** shows the confusion matrix of 15 categories in Caltech-256 obtained by PLSA+ASA method with non- eliminated visual stop-words. And **Fig. 8** depicts the confusion matrix of PLSA+ASA+CSM with visual words elimination percentage $S=15\%$. Both of **Fig. 7** and **Fig. 8** results are based on the whole Caltech-256 dataset. From **Fig. 7** and **Fig. 8**, we can conclude that the "visual stop-words" elimination method in our paper can improve the recall rate of all object classification efficiently.

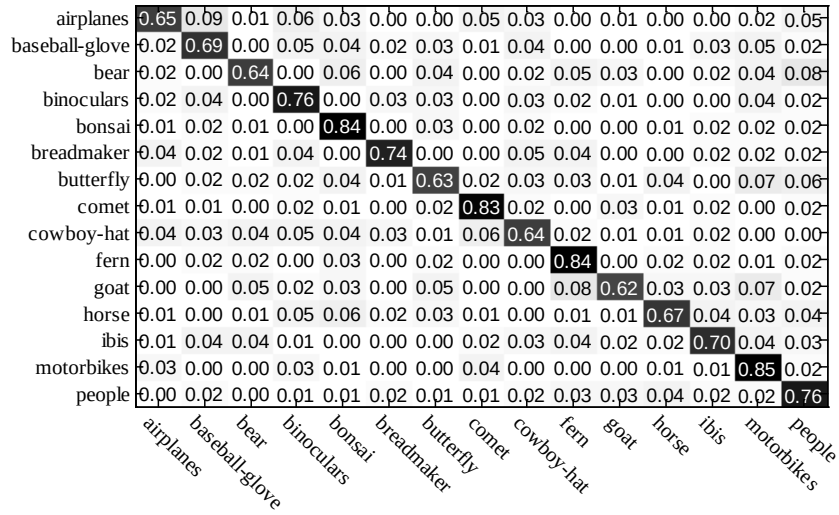

Fig. 7. The confusion matrix of non- eliminated visual stop-words

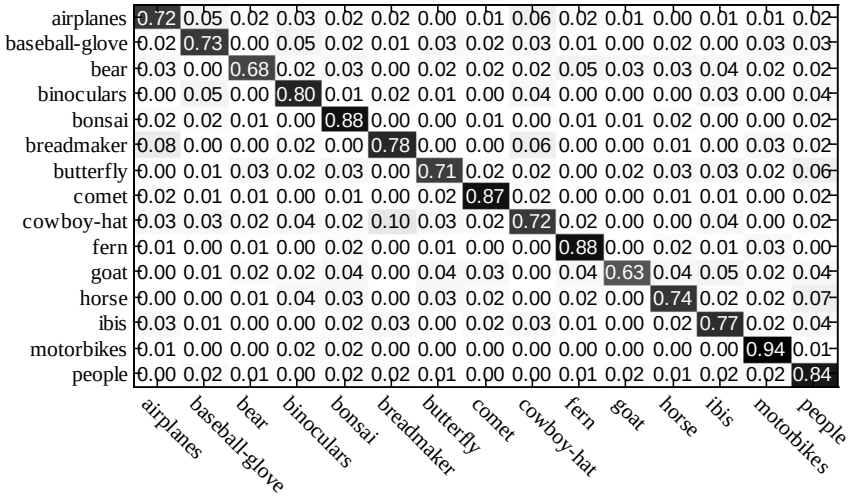

Fig. 8. The confusion matrix of the percentage of eliminated visual words $S = 15\%$

Finally, we do experiment on Pascal Voc2007 dataset to further evaluate the effectiveness of our method in large image dataset environment, in which the values of each parameter are $\alpha = 2.2, m = 20$， $S = 15\%$ and the dictionary size is 10K. We use the same classifier to compare the average precision of our method-PLSA+ASA+CSM with that of hard-assignment based bag of visual words model (HA) [14], Soft-Assignment based bag of visual words model（SA）[35], Contextual information based bag of visual words model [21] and LDA model based soft-assignment method（LDA+SA）[22], respectively. The average precisions of different methods are shown as **Table 2**. From **Table 2**, we can conclude that both SA method and contextual-BoVW methods both introduce some strategies to overcome quantization error caused by synonymy and ambiguity of visual words, hence the object classification accuracy is obviously better than HA method. Meanwhile, due to the combining with LDA model, the LDA+SA method can express the image content more accurately, and the average precision could be further improved. Compared with previous methods, the average precision of the method proposed in this paper is highest.

Table 2. The object classification results of different methods on Pascal Voc2007 database

Object categories	HA(%)	SA(%)	Contextual-BoVW(%)	LDA-SA(%)	PLSA+ASA+CSM(%)
airplanes	71.3	76.5	79.6	81.7	84.6
bicycle	67.1	73.8	77.1	79.5	82.9
bird	62.5	67.5	69.4	72.1	77.1
boat	66.7	73.1	78.2	79.5	84.2
bottle	46.7	55.7	63.1	66.4	70.8
bus	70.2	74.9	77.8	80.4	83.8
car	73.8	79.6	83.1	85.8	88.9
cat	62.7	68.6	73.6	76.4	78.6
chair	67.8	70.8	74.2	77.1	80.5
cow	68.1	74.3	77.6	80.4	84.3
diningtable	66.4	71.4	75.3	76.8	85.1
dog	54.5	64.5	69.1	74.2	81.2
horse	79.6	84.7	86.4	88.3	92.7
motorbike	70.6	75.0	77.6	78.5	83.1
person	85.9	90.1	91.6	91.4	95.6
pottedplant	58.7	65.0	72.8	75.4	78.1
sheep	62.4	68.1	73.2	76.1	77.9
sofa	61.9	68.2	71.6	73.2	80.1
train	82.6	89.5	92.4	92.6	95.5
Tvmonitor	61.4	66.6	70.3	73.4	72.8
Average	66.85	72.89	75.65	78.96	82.89

5. Conclusion

We have proposed a novel bag of visual words method based on PLSA and chi-square model for object category. First of all, in view of the serious quantization error problem caused by synonymy and ambiguity on visual words during constructing the visual vocabulary histograms, we use PLSA model to get the probability distributions of semantic topics on some visual words, then measure the semantic distance between visual words through K-L divergence, thus to obtain the homoionym in semantic space. Secondly, according to different fuzziness categories of SIFT features, the adaptive soft-assignment strategy is proposed for mapping the SIFT features to a number of homoionym adaptively, which can reduce the quantization error efficiently. Finally, chi-square model is adopt to analyze the relativity between each visual word and image category, and based on that, the “visual stop-words” induced by the limits of clustering algorithm or image background noises are eliminated to reconstruct the histograms. Finally, object classification is implemented through SVM classifier. The experimental results show that our method can overcome the synonymy and ambiguity of visual words as well as the quantization error problem to some degree. Moreover, the method can effectively eliminate “visual stop-words” in visual dictionary, which can improve the object classification performance substantially. It should be noted that our method cannot measure the semantic distance between SIFT feature and visual word while analyzing the distance between visual words on the semantic level as well as there are several different image vector representation ways, such as sparse coding, Fisher kernel coding etc. Therefore,

how to make the distance in feature space much closer to the real semantic distance through distance metric learning and construct a more efficient image vector representation are important research keys that need to be concerned in the future.

References

[1] J. Sivic, A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *Proc. of 9th IEEE International Conference on Computer Vision*, pp. 1470-1477, 2003. Article (CrossRef Link)

[2] H. Jegou, M. Douze, C. Schmid, "Packing bag-of features," in *Proc. of IEEE 12th International Conference on Computer Vision*, pp. 2357-2364, 2009. Article (CrossRef Link)

[3] Y. Z. Chen, A. Dick, X. Li, et al., "Spatially aware feature selection and weighting for object retrieval," *Image and Vision Computing*, vol. 31, no. 6, pp. 935-948, 2013. Article (CrossRef Link)

[4] Z. Ji, J. Wang, Y. T. Su, et al., "Balance between object and background: Object-enhanced features for scene image classification," *Neuro computing*, vol. 120, no. 40, pp. 5-23, 2013. Article (CrossRef Link)

[5] A. B. Penatti Otávio, B. Silva Fernanda, Eduardo Valle, et al., "Visual word spatial arrangement for image retrieval and classification," *Pattern Recognition*, vol. 47, no. 1, pp. 705-720, 2014. Article (CrossRef Link)

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004. Article (CrossRef Link)

[7] Y. Ke, R. Sukthankar, "PCA-SIFT: a more distinctive representation for local image descriptors," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 21, no. 2, pp.506-513, 2004. Article (CrossRef Link)

[8] H. Tuytelaars, T. Van Gool, "SURF: Speeded Up Robust Features," in *Proc. of 9th European Conference on Computer Vision*, pp. 452-461, 2006. Article (CrossRef Link)

[9] L. Juan, O. Bwun, "A Comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing*, vol. 3, no. 4, pp. 1-10, 2009. http://www.cscjournals.org/manuscript/Journals/IJIP/volume3/Issue4/IJIP-51.pdf

[10] J. Yang, K. Yu, Y. Gong et al., "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794-1801, June 22-24, 2009. Article (CrossRef Link)

[11] J. Wang, J. Yang, Yu, K. Lv et al., "Locality-constrained linear coding for image classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3360-3367, 2010. Article (CrossRef Link)

[12] F. S. Perronnin, J. Mensink, "Improving the fisher kernel for large-scale image classification," in *Proc. of 11th European Conference on Computer Vision*, pp. 143-156, 2010. Article (CrossRef Link)

[13] D. Nister, H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2161-2168, 2006. Article (CrossRef Link)

[14] J. Philbin, O. Chum, M. Isard, et a1., "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007. Article (CrossRef Link)

[15] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, et al., "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1271-1283, 2010. Article (CrossRef Link)

[16] Y. Su, F. Jurie, "Visual word disambiguation by semantic contexts," in *Proc. of IEEE International Conference on Computer Vision*, pp. 311-318, 2011. Article (CrossRef Link)

[17] Liu S, Bai X, "Discriminative features for image classification and retrieval," *Pattern Recognition Letters*, vol. 33, no. 6, pp. 744-751, 2012. Article (CrossRef Link)

[18] H. Kaiming, Z. Xiangyu, R, Shaoqing et al., "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 12, pp. 1-16, 2015. Article (CrossRef Link)
[19] H. Jegou, M. Douze, C. Schmid et al., "Aggregating local descriptors into a compact image representation," in *Proc. of 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304-3311, 2010. Article (CrossRef Link)
[20] J. Philbin, O. Chum, M. Isard et al., "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 278-286, 2008. Article (CrossRef Link)
[21] T. Li, T. Mei,I. S. Kweon et al., "Contextual bags-of-words for visual categorization," *IEEE Transactions on Circuits System Video Technology*, vol. 21, no. 4, pp. 381-392, 2012. Article (CrossRef Link)
[22] D. Weinshall, G. Levi, D. Hanukaev, "LDA topic model with soft assignment of descriptors to words," in *Proc. of the 30th International Conference on Machine Learning*, pp. 711-719, 2013. Article (CrossRef Link)
[23] D. Danilo, G. Carneiro, T. J. Chin et al., "Fuzzy clustering based encoding for Visual Object Classification," in *Proc. of IFSA World Congress and NAFIPS Annual Meeting*, pp. 1439-1444, 2013. Article (CrossRef Link)
[24] J. Yuan, Y.Wu, M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2012. Article (CrossRef Link)
[25] T. Chen, K. H. Yap, D.J. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 612-622, 2014. Article (CrossRef Link)
[26] J. B. Yeh, C. H. Wu, "Extraction of robust visual phrases using graph mining for image retrieval," in *Proc. of IEEE Conference on Multimedia and Expo*, pp. 3681-3684, 2010. Article (CrossRef Link)
[27] E. Roman-Rangel, S. Marchand-Maillet, "Automatic Removal of Visual Stop-Words," in *Proc. of the ACM International Conference on Multimedia*, pp. 1145-1148, 2014. Article (CrossRef Link)
[28] T. Hoffmann, "Probabilistic Latent Semantic Analysis," in *Proc. of Uncertainty in Artificial Intelligence*, pp. 289-296, 1999. Article (CrossRef Link)
[29] E. Emrah, A. Nafiz, "Scene Classification Using Spatial Pyramid of Latent Topics," in *Proc. of IEEE 20th International Conference on Pattern Recognition*, pp. 3603-3606, 2010. Article (CrossRef Link)
[30] Z. Ruije, L. Bicheng and W. Fushan, "Image Scene Classification Based on Multi-Scale and Contextual Semantic Information," *Acta Electronica Sinica*, vol. 42, no. 4, pp. 646-652, 2014. Article(CrossRef Link)
[31] K. Kesorn, S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 211-222, 2012. Article (CrossRef Link)
[32] G. Griffin, Holub, P. AD. Perona, "Caltech-256 object category dataset," *Technical Report 7694*, http://resolver.caltech.edu/CaltechAUTHORS:CNS-TR-2007-001, 2012.
[33] M. Everingham, L. van Gool and C.K.I. Williams, et al., "The Pascal visual object classes challenge 2007(VOC 2007) results," http://pascalin.ecs.soton.ac.uk/challenges/VOC/voc2007/results/ 2014-05.
[34] "LIBSVM-A library for support vector machines," http://www.csie.ntu.edu.tw/cjlin/libsvm/ 2014-04.
[35] P. Koniusz, K. Mikolajczyk, "Soft Assignment of visual words as linear coordinate coding and optimisation of its reconstruction error," in *Proc. of 18th IEEE International Conference on Image Processing*, pp. 2413-416, 2011. Article(CrossRef Link)

Yongwei Zhao received his B.S. in Electronic Information Engineering from Shandong University, Jinan, China, in 2009, and received his M.S. in China National Digital Switching System Engineering and Technological R&D Center in 2012, and is currently pursuing the Ph.D. degree. His research interests include image processing and classification.

Tianqiang Peng received his Ph.D. degree in China National Digital Switching System Engineering and Technological R&D Center in 2008. He is currently an Associate Professor in department of Computer Science and Engineering, Henan Institute of Engineering. His research interests is image processing and pattern recognition.

Bicheng Li received his M.S. and Ph.D. degrees in China National Digital Switching System Engineering and Technological R&D Center in 1995 and 1998, respectively. He is currently a Professor with Department of Information Science. His research interests include text processing, image processing and pattern recognition.

Shengcai Ke received his BS. degrees in Signal and Information Processing from Zhengzhou Information Science and Technology Institute in 2013, and is currently pursuing the M.S. degree in China National Digital Switching System Engineering and Technological R&D Center. His research interests include image processing.