Nearest-Neighbors Based Weighted Method for the BOVW Applied to Image Classification

Mengxi Xu*, Quansen Sun[†], Yingshu Lu*** and Chenming Shen**

Abstract – This paper presents a new Nearest-Neighbors based weighted representation for images and weighted K-Nearest-Neighbors (WKNN) classifier to improve the precision of image classification using the Bag of Visual Words (BOVW) based models. Scale-invariant feature transform (SIFT) features are firstly extracted from images. Then, the K-means++ algorithm is adopted in place of the conventional K-means algorithm to generate a more effective visual dictionary. Furthermore, the histogram of visual words becomes more expressive by utilizing the proposed weighted vector quantization (WVQ). Finally, WKNN classifier is applied to enhance the properties of the classification task between images in which similar levels of background noise are present. Average precision and absolute change degree are calculated to assess the classification performance and the stability of K-means++ algorithm, respectively. Experimental results on three diverse datasets: Caltech-101, Caltech-256 and PASCAL VOC 2011 show that the proposed WVQ method and WKNN method further improve the performance of classification.

Keywords: Bag of visual words, K-means++ algorithm, Weighted vector quantization, Weighted K-nearest-neighbors classifier

1. Introduction

Automated image classification (i.e. object recognition or scene categorization) is a vital component in a variety of image processing and computer vision applications, and the objective of this work is to classify an image by the object category that it contains. Conventional image processing technologies (i.e. bilateral filtering, feature driven methods, locality constraints, etc.) can improve the classification precision for a specific parameter to within a certain precision [1-3]. However, over the past decade, several novel methods have been proposed which are uniquely suited for specific types of classification. One such technique is the Bag of Visual Words (BOVW) model, which has been widely researched and applied commercially since its introduction [4].

BOVW originated from the Bag of Words (BOW) model which was initially used in document processing, but has recently been introduced into the imaging community [5]. Since then, several derivatives and refinements have been proposed to adapt to different applications (i.e. scene classification, spatial pyramid matching, weighted feature trajectories, etc.) [6-10]. In this model, scale-invariant feature transform (SIFT) features, proposed by D.G. Lowe

[11], are extracted from images and clustered to construct a visual dictionary so that each image can be represented by quantizing features. After that, a specific classifier (i.e. SPM-SVM, Random Forests and Ferns, Boosting, etc.) [7, 12, 13] is chosen to complete the task of classification. However, images are commonly represented by statistical histograms of the visual dictionary or named vector quantization (VQ) [8], in which the value of a word bin simply increases if the visual word is closest to a feature. This implies that this histogram, to some extent, fails to describe the different feature expressions, resulting in poor precision assessment.

Moreover, when an image dataset contains similar levels of background noise, it is of little effect to the classification system if VQ method is adopted to describe the images, because image features are usually mixed together in a relatively small area. For instance, in PASCAL VOC 2011 dataset, nearly all the images from different object categories mainly contain people image that is rational to be regarded as the similar background noise. As a result, it is difficult to distinguish a few discriminative features from the numerous noises. Similarly, a classification system in which object images have non-uniform resolutions leading to a large gap in the number of present image features (from roughly 20 to 7000 in Caltech-256 dataset), the traditional histogram for a low resolution image representation would have little positive value, so the expressive performance of histogram would be significantly unbalanced.

With consideration for the problems above, a weighted VQ (WVQ) method is proposed to enhance the expressive

[†] Corresponding Author: School of Computer Science and Technology, Nanjing University of Science & Technology, China. (sunquansen@mail.njust.edu.cn)

* School of Computer Science and Technology, Nanjing University of Science & Technology, China. (mxu26@126.com)

** School of Computer Engineering, Nanjing Institute of Technology, China.

*** College of Computer and Information, Hohai University, China.

Received: November 27, 2014; Accepted: April 13, 2015

performance of the histogram in the BOVW based image classification models, as it is to make use of those severely discarded information in VQ method through a kind of weighted reconstruction. Simultaneously, with the aim of reducing the randomness of the generated visual dictionary to obtain a relatively stable categorization performance, the K-means++ [15] algorithm is included in the clustering process. This clustering algorithm takes the distances between points (visual words) into account to ensure the differences among the visual words. Moreover, the work shows that the stability of the K-means++ algorithm is better than that of K-means algorithm.

Furthermore, conventional image classification methods include an intensive parametric learning stage, while non-parametric Nearest-Neighbors based image classifiers require no training time and have other favorable properties (described by Boiman et al. [13]). Therefore, in order to deal with the case that similar levels of background noise are present in the classification system, a weighted K-Nearest-Neighbors classifier (WKNN) algorithm is proposed and implemented to strengthen the categorization performance for the classification task. Notably, the classification task in this study is to classify an image by the object category that it contains. In short, our technique can make use of the discarded information to more effectively describe images and improve the stability of the effectiveness of the visual dictionary. The models based on the proposed methods further become more discriminative as WKNN is proposed and adopted in the classification process.

The rest of this paper is organized as follows. In Section 2, we give a brief overview of the K-means++ algorithm and describe its implementation in reducing dependence on hardware performance. In Section 3, we then describe the method for WVQ to effectively represent the images. The KNN and WKNN classifiers are discussed in Section 4. Section 5 demonstrates the performance and provides an analysis of the proposed methods which were implemented in three datasets: Caltech-101, Caltech-256, and PASCAL VOC 2011. Finally, Section 6 concludes the paper.

2. Acquiring Visual Words

In conventional BOVW, SIFT features are extracted to produce a preliminary description of images. In the traditional method, the visual dictionary is established by a K-means algorithm which is used as an effective method for clustering features. The basic K-means algorithm is organized as follows [16]:

- (i) Select k data points (features) to serve as initial centroids.
- (ii) (Re) Assign all points to their nearest centroids.
- (iii) Recalculate the centroid of each newly assembled cluster.
- (iv) Repeat steps 2 and 3 until the centroids do not

change.

This algorithm, however, randomly chooses k features to be the initial centers in the features clustering process. That normally leads to negative impacts on the performance of the classification system as the final visual words are merely local optimum. Because of the constraints on the K-means algorithm, described by Wagstaff et al. [17], we adopt a modified version called the K-means++ algorithm [16]. The difference between these two algorithms is that the latter initialize centers based on probability which is dependent on the distances between points. If a given feature is a significant distance away from the others, it should be picked as an initial center with higher probability. The reasoning behind this convention is that visual words should be far away from each other to ensure their effectiveness as global optimum.

In this way, the SIFT features are denoted as $F = \{f_1, f_2, f_3, \dots, f_n\}$, at which point one of them is randomly selected from F , and set as f_i to be the first initial center. After this, the distances between this center and all of the other features are computed and set as $D = \{d_1, d_2, d_3, \dots, d_n\}$. Following [15], the probability $P = \{p_1, p_2, p_3, \dots, p_{n-1}\}$ is then defined as

$$p_j = \frac{d_j}{\sum_{i=1}^{n-1} d_i}, \quad (1)$$

Once the probabilities used to initialize k centers are generated in turns using (1), the remainder of the process works as K-means.

In order to reduce its dependence on hardware, the clustering algorithm is applied to each category to generate relatively large numbers of primarily homogeneous visual words, resulting in the fewer and more effective features remained. After that, it is carried out again to cluster all of the primary visual words for the purpose of obtaining relatively few final words.

3. A Weighted Representation

In [7], the conventional vector quantization (VQ) is adopted in the BOVW based spatial pyramid matching model (SPM), then utilizing support vector machine (SVM) to complete the classification tasks. However, VQ algorithm is a hard encoding method and leads to severe information loss of SIFT features. In order to overcome the shortcoming, Yang et al. [8] proposed the ScSPM model, in which sparse coding (SC) algorithm is introduced into the conventional SPM model to replace K-means clustering algorithm to obtain a sparser vector representation. Although the SC has achieved excellent performance, the locality of the obtained codes is much more essential compared to the sparsity of the codes for features in an

Fig. 1. Schematic diagram of the proposed WVQ method. The top hexagonal element is a visual word, and the rest with different shapes lying in the curves is the features of an image. The lines with different thickness represent their weights.

images [3]. At the same time, Lee et al. [20] pointed out that learning large, highly over-complete sparse representations is extremely expensive and Wang et al. [21] showed that the codebook initialized with some heuristics can still achieve excellent performance. Therefore, we adopt the K-means++ algorithm to cluster the features as the fixed visual dictionary for reducing the computational cost.

According to the analysis above, based on the Nearest-Neighbors framework and VQ encoding method, this study proposed a weighted local vector quantization method, which is to rectify the drawbacks of VQ and SC methods mentioned above to enhance expressiveness. In this method, as shown in Fig. 1, only the features (the black solid elements between the curve lines) nearest to a visual word (hexagonal element) are given with relatively large weight values (thick lines), while the rest are given with much smaller ones (thin lines). At the same time, the contributions of features for the visual word bin are also evaluated by their distances to the visual word. This implies that if a given feature has few similarities with a visual word, it will have little influence on the value of the corresponding visual word bin. If multiple features are close to the same visual word, their contributions to the visual word bin can be further determined from their distances to the visual word.

In this way, each visual word bin will be directly obtained to be regarded as the final image vector representation through calculating the weighted similarities from each visual word to their own local Nearest-Neighbors, respectively. The experimental results show that the models using WVQ method outperform the ones using VQ method.

The training images are denoted as $M_a = \{m_i^a, F_i^a, r_i^a \mid 1 \leq i \leq n^a\}$, where m_i^a is the i th image pixel data, $F_i^a = \{f_{i1}^a, f_{i2}^a, f_{i3}^a, \dots, f_{in_i}^a\}$ represents the SIFT features of i th image, r_i^a is the WVQ representation of the i th image and n^a is the number of training images.

Similarly, $M_e = \{m_i^e, F_i^e, r_i^e \mid 1 \leq i \leq n^e\}$ is used to represent the testing images. $F_w = \{f_1^w, f_2^w, f_3^w, \dots, f_k^w\}$ is the visual dictionary which includes k visual words. The

value of the f_j^w bin is denoted as v_j . In this way, a certain training image m_i^a can be represented by $r_i^a = \{v_1, v_2, v_3, \dots, v_k\}$ using the method of WVQ. The distances from the SIFT features f_i^a of the training image to the word f_j^w are denoted as

$$D_j = \text{sorted}(d_{j1}, d_{j2}, d_{j3}, \dots, d_{jn_i}),$$

$$d_{jn} = d(f_j^w, f_{in}^a), \quad \forall n \in n_i^a, \quad (2)$$

Where $\text{sorted}(\cdot)$ is the sort function (from small to large) and $d(\cdot, \cdot)$ is used to calculate the Euclidean distance.

Assumed that the number of features in an image is much larger than M , v_j can then be defined as follows:

$$v_j = \sum_{n=1}^M \frac{s(f_j^w, f_{in}^a)}{w_n}, \quad (3)$$

where $s(\cdot, \cdot)$ represents the cosine distance between visual word and feature and w_n represents the weight of feature f_{in}^a and can be defined as: $w_n = 2^{1+(n-1)/M}$ ($1 \leq n \leq M$). The index of D_j can be used to find the top M features in F_i^a to get v_j . By implementing (2) and (3) for each visual word, the WVQ of image m_i^a can be constructed and the testing images can be represented in the same way.

4. Weighted Classifiers

The Nearest-Neighbors based classifier algorithms are occasionally considered to be overly simplistic and hence ill-suited for complex classification tasks [14]. However, it has no training process and works well in various classification systems. Moreover, KNN classifier can flexibly combine other algorithms to enhance the performance for particular classification tasks, such as SVM-KNN [18] and ML-KNN [19]. Therefore, KNN classifier is chosen as the basic classifier for our classification tasks.

In the case of experimental images with significant background noise, according to the above analysis, a weighted KNN (WKNN) classifier is proposed to improve the categorization precision by about 1% in our experiments. The reason is that, in this case, due to similar background noise, features are interwoven in a relatively small area. If the similarities among the testing data and the k closest training data can be taken into account, the classification precision would be rationally higher than it would be for the conventional method.

The training data are listed as $R^a = \{r_1^a, r_2^a, r_3^a, \dots, r_{n^a}^a\}$ and the testing data are $R^e = \{r_1^e, r_2^e, r_3^e, \dots, r_{n^e}^e\}$. The labels of classes are represented by $C = \{c_1, c_2, c_3, \dots, c_c\}$ and the labels to which the training data belong are known. In the classic KNN classifier, in order to predict the classes of the testing data r_m^e , distances to all of the training data, represented by $D_m^e = \{d_1^e, d_2^e, d_3^e, \dots, d_{n^e}^e\}$, must be computed. The k nearest neighbors are then

acquired and the appropriate class to which these neighbors belong is determined and set as $val = \{v_1^c, v_2^c, v_3^c, \dots, v_k^c\}$,

$$v_j^c = \sum_{i=1}^{k_j} 1, \quad (4)$$

where k_j is the number of training data (neighbors) which belong to class c_j . The maximum value in val is the class of the testing data r_m^e .

In the case of the proposed WKNN classifier, instead of simply counting the class number each neighbor belongs to, the distances are regarded as weighted values and used to compute their similarities. Therefore, the value of the j th class can be calculated as follows,

$$v_j^c = \sum_{i=1}^{k_j} \frac{H}{d_i^c}$$

$$d_i^c = 1 - s(r_j^e, r_i^a), \quad (5)$$

where H is a constant and $s(\cdot, \cdot)$ represents cosine distance. After this is done, in the same way, the maximum value is regarded as the predicted consequence.

5. Experimental Results

In this section, three benchmark datasets, Caltech-101, Caltech-256, and PASCAL VOC 2011, are used as our experimental datasets. In particular, six classes are randomly selected from each dataset separately. We perform all processing with grayscale to extract SIFT features, even when color images are available. For Caltech-101 and Caltech-256 datasets, four description models are implemented for the experimental comparisons, which are the BOVW model using VQ algorithm (BVQ), the BOVW model using the proposed WVQ encoding method (BWVQ), the conventional SPM in [7] and the SPM model using our WVQ method (WSPM), respectively. The SPM framework is the BOVW model added with spatial information. In this framework, spatial information of image features are added through dividing one image into increasing sub-regions in the increasing spatial-pyramid layers. The BOVW model is then performed to obtain vector representation for each sub-region. After that, all vectors are put in a row following their layers to get a longer vector representation. We adopt the top three layers in which each image is divided into 1×1 , 2×2 , 4×4 sub-regions in increasingly layer, respectively. Suppose that the size of dictionary is set to be K , and then an image can be encoded as $v \in R^{1 \times (21 \times K)}$ after the implementation of SPM framework and coding phase.

Notably, KNN and the proposed WKNN classification algorithm work as their classifiers in turns. In addition, the experimental results of the K-means and K-means++ algorithms, such as the classification precision and the stability of classification systems, are also described in

the text.

5.1 Caltech-101

This dataset, which was collected by Li Fei-Fei et al. [22], consists of images from 101 object categories, with each class containing between 31 and 800 images. Most images are of medium resolution, about 300×300 pixels.

We randomly select 20 images for training per class (Helicopter, Dalmatians, Hawksbill, Faces, Watch, and Soccer ball) as shown in Fig. 2, 40 images are used for testing per class. The K-means++ algorithm is then adopted to cluster features of the training data to get visual words whose numbers are between 100 and 300 with an interval of 50. The description models of BVQ, BWVQ, SPM and WSPM are used to describe the images in turns. We set

Fig. 2. Examples of each category selected from Caltech-101.

$M=20$ and $m=5$ for the proposed BVQ method. Finally, the KNN classifier is implemented to accomplish the classification task and the number of nearest neighbors in it was set to be 3. In addition, we calculate the average precision (AP) [23] of each experiment to assess the properties of different image representations. AP is given by:

$$AP = \frac{\sum_{i=1}^{n^s} a_i}{n^s}, \quad (6)$$

Where n^s is the number of classes and a_i is the classification precision of the i th category. This definition (6) is implemented for each experiment, respectively. The acquired AP results are listed in Table 1 and displayed graphically in Fig. 3. Due to spatial information added in the SPM based models, the performance of SPM based algorithms is rationally better than that of the BOVW based models [8].

From Fig. 3, we can find out that WSPM achieves the best performance and BWVQ also outperforms BVQ. Furthermore, when the number of visual words is fixed as 300, we compare the classification precision of each

Table 1. The AP results of BVQ, BWVQ, SPM and WSPM (Caltech-101).

No. of visual words	100	150	200	250	300
BVQ	0.7792	0.7875	0.7792	0.7750	0.7667
BWVQ	0.7917	0.7958	0.8042	0.7883	0.8125
SPM	0.8589	0.8545	0.8695	0.8631	0.8737
WSPM	0.8725	0.8797	0.8832	0.8743	0.8932

Table 2. The classification precision of each category using BVQ, BWVQ, SPM and WSPM (Caltech-101).

Class Name	Accuracy (BVQ)	Accuracy (BWVQ)	Accuracy (SPM)	Accuracy (WSPM)
Helicopter	0.6500	0.8500	0.8640	0.8830
Dalmatians	0.6750	0.8500	0.9050	0.8610
7Hawksbill	0.9000	0.9000	0.9230	1.0000
Faces	0.9250	1.0000	0.9530	1.0000
7Hawksbill	0.6500	0.5750	0.7320	0.8050
Faces	0.8000	0.7000	0.8650	0.8220

Fig. 3. The AP comparisons for four description models.

category using the description models of BVQ, BWVQ, SPM and WSPM, respectively. The results are shown in Table 2. According to this table, compared with the performance of the models using VQ encoding method (BVQ, SPM), the AP results of the corresponding models using the proposed WVQ encoding method (BWVQ, WSPM) increase by about 1-5%, although this dataset from Caltech-101 has little background noises. Therefore, it is reasonable to conclude that our proposed WVQ encoding method can capture more discriminative information to effectively describe images than using VQ method in the coding phase.

5.2 Caltech-256

The second dataset experiment is the Caltech-256. This dataset collected by Griffin et al. consists of images from

Fig. 4. Examples of each category selected from Caltech-

Table 3. The AP results of BVQ, BWVQ, SPM and WSPM (Caltech-256).

No. of visual words	100	150	200	250	300
BVQ	0.4458	0.5167	0.4958	0.5292	0.5125
BWVQ	0.5417	0.6042	0.6167	0.6292	0.6125
SPM	0.6189	0.6520	0.6646	0.6521	0.6637
WSPM	0.6703	0.6920	0.6803	0.6943	0.6955

Table 4. The classification precision of each category using BVQ, BWVQ, SPM and WSPM (Caltech-256).

Class Name	Accuracy (BVQ)	Accuracy (BWVQ)	Accuracy (SPM)	Accuracy (WSPM)
AK	0.5500	0.6250	0.6520	0.721
Baseball bat	0.6750	0.7500	0.7550	0.7920
Computer mouse	0.4250	0.6000	0.6320	0.6280
Eiffel	0.4250	0.4250	0.5550	0.6200
Fish	0.2000	0.7000	0.7200	0.7290
Skeleton	0.8000	0.5750	0.6640	0.6830

Fig. 5. The AP comparisons for four description models.

256 object categories [24] and is an expansion of Caltech-101. Each of its categories contains between 80 and 827 images. There is also no alignment among the object categories. The resolution of images varies more violently than it does for Caltech-101 so the number of SIFT features extracted from different images is roughly between 20 and 7000 in our experiments.

In the same way, we randomly select 20 images per category for training and 40 images per class for testing. The selected categories are summarized as follows: *AK*, *Baseball bat*, *Computer mouse*, *Eiffel*, *Fish*, and *Skeleton*, as shown in Fig. 4. The values of M and m are set to be 20 and 5 in the experiments, respectively. All of the above procedures for Caltech-101 are then implemented again.

The AP results of four description models are shown in Table 3 and graphically displayed in Fig. 5.

When experiments with the Caltech-256 dataset include the generation of 300 visual words as a visual dictionary, the classification precision of each category is compared in Table 4. Similarly, we find that almost all of the precision results are improved, which is more obvious here than in Table 2. It is worth noting that the experimental results of

Caltech-101 differ from the AP results of Caltech-256 using different representations and the results of the latter have a larger gap. For the Caltech-101 dataset, the AP of each experiment is improved by about 1-5%, while for Caltech-256, the AP results are generally raised by about 3-10%, as the description model of BWVQ and WSPM using the proposed WVQ method can get balanced representations regardless of the number of features extracted from them. Therefore, we can conclude that our WVQ encoding method can raise categorization precision in our classification system, especially when the experimental dataset features a large gap in the resolution of the images.

5.3 PASCAL VOC 2011

PASCAL VOC 2011 is a popular benchmark dataset

Fig. 6. Examples of each category selected from PASCAL VOC 2011.

for computer vision [25]. As we did before, we use 6 categories (*Running, Phoning, Riding Bike, Riding Horse, Shooting, and Playing Guitar*) which are shown in Fig. 6.

These are randomly selected and split into training and testing data respectively. Because of the similarities in background noise which exists in the dataset, a greater number of training images are needed to extract the similarities and describe the category. We employ 40 images per category for training and 20 images per category for testing.

As we have experimentally proved that our proposed WVQ encoding method outperforms the conventional VQ method using four description models in Sections 5.2 and 5.3, BVQ and BWVQ are chosen as the image description model in this section for simplicity. In addition to implementing the procedures performed on the above two datasets, the performance of KNN and WKNN classifiers is compared and analyzed. The values of M , m and H are set to be 20, 5, and 10 in the experiments, respectively.

In the experiments with the PASCAL VOC 2011 dataset, it is determined that the classification precision is not sensitive to the number of closest neighbors in KNN or WKNN classifier. As a result, instead of using AP as the metric in the comparisons, we compute the mean AP (MAP) results [26] of the models with different odd numbers from 1 to 19 in WKNN classifier to assess classification performance. The number of visual words is between 100 and 300 with an interval of 50. The MAP results for BVQ and BWVQ are shown in Table 5 and graphically displayed in Fig. 7. Most notably, these experiments include the WKNN classifier which further adds about 1% to the classification precision.

Table 5. The MAP results of BVQ and BWVQ using KNN and WKNN (PASCAL VOC 2011).

No. of visual words	100	150	200	250	300
BVQ+KNN	0.3203	0.3085	0.3090	0.3337	0.3230
BVQ+WKNN	0.3222	0.3117	0.3080	0.3330	0.3262
BWVQ+KNN	0.3787	0.3697	0.4046	0.4157	0.4105
BWVQ+WKNN	0.3803	0.3742	0.4067	0.4228	0.4150

Fig. 7. MAP comparisons of the experiments.

From Table 5, we draw the conclusion that, for the PASCAL VOC 2011 dataset, when using the KNN classifier, the proposed WVQ encoding method can improve the categorization precision by 5% to 10%. Moreover, in our experiments, the performance of WKNN is better than or equal to KNN when using VQ method to describe images. Inspired by this conclusion, our WKNN is performed for the classification system in which WVQ method is used to represent images. In Fig. 6, it is evident that when using WVQ method, the performance of WKNN classifier is generally better than that of KNN classifier, while these improvements of the performance using VQ method is not much obvious, the reasons are that the proposed WVQ method is able to make use of the discarded information to more effectively describe images and WKNN classifier can take the similar values between images into account. Therefore, the overlap of KNN and WKNN classifier means that WKNN is basically better than or equal to KNN when using VQ method to describe images.

Thus, from the above analysis, we can conclude that our WVQ method can obviously improve the performance of the classification system when each category in the object dataset includes similar levels of background noise. WKNN classifier can then make further efforts to enhance the performance of the categorization system.

5.4 The performance of K-means++

As previously mentioned, the K-means++ algorithm initializes its centers based on probability which is dependent on the distances between each point. It is thereby rational to believe that the performance of K-means++ algorithm is more stable than that of the K-means algorithm. Given this consideration, we design the experiments to compare the stability of AP with these two clustering methods.

The experiments are implemented on two datasets (Caltech-101, PASCAL VOC 2011). First of all, the number of visual words is set to 100 and all images are represented by the model of BWVQ. The KNN classifier is then utilized to carry out the classification task. The experiment is repeated 5 times on these two dataset. Using K-means++ and the K-means algorithm to cluster features, respectively, their AP results and average AP values are shown in Table 6.

The absolute changing degree (ACD) is proposed to evaluate the stability of these two clustering algorithms given in Table 7:

$$ACD = \frac{\sum_{i=1}^{n^d} \sum_{j=1}^{n^i} (AP_j^i - u^i)^2}{n^d \times n^i}, \quad (7)$$

where n^d is the number of datasets, n^i represents the time of the experiment and u^i is the average value of AP

Table 6. The results of five experiments using K-means and K-means++ algorithm.

The <i>i</i> th experiment	Caltech-101		PASCAL VOC 2011	
	K-means	K-means++	K-means	K-means++

1	0.7709	0.7917	0.3610	0.3912
2	0.7833	0.7917	0.3712	0.3690
3	0.7917	0.7708	0.3753	0.3957
4	0.7875	0.7750	0.3882	0.3882
5	0.7583	0.7875	0.3707	0.4012
Average	0.7783	0.7833	0.3732	0.3890

Table 7. The ACD results of K-means and K-means++ algorithms, respectively.

-----	K-means	K-means++
ACD(10^{-5})	1.136	0.983

performed on the *i*th dataset. The essence of ACD is to compute the variance value of AP results. Therefore, it is rational to conclude that the performance of an algorithm is more stable if its ACD (or variance value) is smaller. Table 7 shows the ACD results of these two clustering algorithms.

From this Table 7, in contrast with the K-means, the ACD value of the K-means++ declines by about 13.5%, therefore, we can be convinced that the stability of the K-means++ algorithm is superior to the K-means algorithm.

6. Conclusions

In this paper, we proposed using WVQ encoding method in BOVW based description models to represent images for classification task. Through the experimental comparisons and analysis, this new method of representation can remarkably improve the precision assessment for the three datasets and is extremely suitable to the classification tasks in which similar levels of background noise are present or the resolution of images varies violently. The K-means++ algorithm was applied in the processing and enhances the stability of the performance of the visual words. In addition, the proposed WKNN classifier is introduced into the image classification models to further boost the classification precision in the case of object images with similar levels of background noise. Finally, Experimental results demonstrate the effectiveness of our approach.

Acknowledgements

This work is supported partly by the National Natural Science Foundation of China (No. 61401195, 61273251), and the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province, China (No. 13KJB520009).

References

- [1] A. Shi, L. Xu, F. Xu, et al., "Multispectral and panchromatic image fusion based on improved bilateral filter," *Journal of Applied Remote Sensing*, 5(1): 053542-1-053542-17, 2011.
- [2] F. Xu, T. Fan, C. Huang, et al., "Block-Based MAP Super-resolution Using Feature-Driven Prior Model," *Mathematical Problems in Engineering*, 48(1):331-350, 2014.
- [3] J. Wang, J. Yang, K. Yu, et al., "Locality-constrained linear coding for image classification," *Computer Vision Pattern Recognition*, 3360-3367, 2010.
- [4] W. Tao, Y. Zhou, L. Liu, et al., "Spatial adjacent bag of features with multiple super pixels for object segmentation and classification". *Information Sciences*, 281: 373-385, 2014.
- [5] M. Zhang, A. A. Sawchuk, "Motion primitive-based human activity recognition using a bag-of-features approach," *Proceedings of the 2nd ACM SIGHT International Health Informatics Symposium*, 631-640, 2012.
- [6] L. Zhou, Z. Zhou, D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognition*, 46(1): 424-433, 2013.
- [7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Computer Vision Pattern Recognition*, 2169-2178, 2006.
- [8] J. C. Yang, K. Yu, Y. H. Gong and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *Computer Vision Pattern Recognition*, 1794-1801, 2009.
- [9] J. Yu, M. Jeon, W. Pedrycz, "Weighted feature trajectories and concatenated bag-of-features for action recognition", *Neurocomputing*, 131: 200-207, 2014.
- [10] A. Plinge, R. Grzeszick, G. A. Fink, "A bag-of-features approach to acoustic event detection", *Acoustics, Speech and Signal Processing*, 3704-3708, 2014.
- [11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, 60(2): 91-110, 2004.
- [12] A. Bocsh, A. Zisserman, et al., "Image Classification using Random Forests and Ferns," *Computer Vision (ICCV)*, 1-8, 2007.
- [13] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer, "Weak hypotheses and boosting for generic object detection and recognition," *Computer Vision (ECCV)*, 71-84, 2004.
- [14] O. Boiman, E. Shechtman, M. Irani, "In defense of Nearest-Neighbor based image classification," *Computer Vision Pattern Recognition*, 1-8, 2008.
- [15] D. Arthur, S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms*, 1027-1035, 2007.

- [16] S. Agarwal, S. Yadav, K. Singh., "K-means versus K-means++ clustering technique," Students Conference on Engineering and Systems, 1-6, 2012.
- [17] K. Wagstaff, C. Cardie, S. Rogers, et al., "Constrained K-means clustering with background knowledge," International Conference on Machine Learning, 1: 577-584, 2001.
- [18] H. Zhang, A. Berg, M. Maire, et al., "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," Computer Vision and Pattern Recognition, 2: 2126-2136, 2006.
- [19] M. L. Zhang, Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern recognition, 40(7): 2038-2048, 2007.
- [20] H. Lee, A. Battle, R. Raina and A. Y. Ng, "Efficient sparse coding algorithms," Advances in neural information processing systems, 801-808 2006.
- [21] S. Gao, I. W. H. Tsang, L. T. Chia, "Sparse representation with kernels", Image Processing, 22(2): 423-434, 2013.
- [22] L. Fei-Fei, R. Fergus, P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," Computer Vision and Image Understanding, 106(1): 59-70, 2007.
- [23] L. Bourdev, S. Maji, J. Malik, "Describing people: A poselet-based approach to attribute classification," Computer Vision (ICCV), pp. 1543-1550, 2011.
- [24] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset. Technical Report USB/CSD-04-1366," California Institute of Technology, 2007.
- [25] C. Vondrick, A. Khosla, T. Malisiewicz, et al., "Hoggles: Visualizing object detection features," Computer Vision (ICCV), 1-8, 2013.
- [26] T. Malisiewicz, A. Gupta, A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," Computer Vision (ICCV), 89-96, 2011.

Quansen Sun He is a professor of School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received his PhD degree from Nanjing University of Science and Technology, China. Currently, his research area includes theory and application of pattern recognition, image processing, analysis and recognition, remote sensing information intelligent processing.

Yingshu Lu He received B.S degree in electronic information science and technology from Jiangsu University of Science and Technology. His research interests are image processing, pattern recognition and signal processing.

Chenming Shen He received B.S degree in computational mathematics from Nanjing Normal University. His research interests are database development and data mining.

Mengxi Xu She received the MSc degree from UNESCO-IHE, the Netherlands. She is currently a lecturer in School of Computer Engineering, Nanjing Institute of Technology, China. Her current research interests include image processing, information processing system and its application.