Neurocomputing 119 (2013) 478–488

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Weighted visual vocabulary to balance the descriptive ability on general dataset

Yi Xie [a], Shuqiang Jiang [a,*], Qingming Huang [a,b]

[a] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, No.6 Kexueyuan South Road, Haidian District, Beijing 100190, China

[b] University of Chinese Academy of Sciences (UCAS), #80 Zhongguancun East Road, Haidian District, Beijing 100190, China

ARTICLE INFO

Article history:
Received 8 October 2012
Received in revised form
14 February 2013
Accepted 8 March 2013
Communicated by Xiaofei He
Available online 17 April 2013

Keywords:
Bag of feature
Visual word
Weighted visual word
Image retrieval

ABSTRACT

Many image retrieval and search systems take visual words as the key to represent images. Extracted feature descriptors are assigned to best matching visual words in the vocabulary. However, the visual word based retrieval method is not satisfactory in real applications as the descriptive ability of visual vocabulary is not strong enough to describe image local features. In this paper, we investigate on the problem of the descriptive ability of visual vocabulary. There are mainly two contributions in this paper. Firstly, we make a comprehensive analysis on the retrieval performance of standard bag-of-features (BOF) method by using different codebooks. The codebook is generated on one dataset and used for retrieval on the same dataset and on different datasets, which is called “homology codebook” and “non-homology codebook”, respectively. Experimental results show that there exists performance drop when a non-homology codebook is used for retrieval on one dataset, compared to the homology codebook. This phenomena is usually neglected by most of the retrieval tasks and systems. Secondly, in order to abate the influence of non-homology codebook, a weighting based method is proposed to balance the descriptive ability of codebook for different datasets. Experimental evaluations show that the proposed method outperforms standard BOF method, and it can prevent the performance drop to some extent when non-homology codebooks are used.

1. Introduction

The Bag-of-features (BOF) method is a widely used method in image search and retrieval systems [1–9]. The BOF representation has some important characteristics which are necessary for large-scale image retrieval, such as adopting scale and rotation invariant descriptors by Lowe's SIFT approach [10], and storing database in inverted files to reduce memory consumption and speed up retrieval. More specifically, transform invariant descriptors are extracted from images, and are assigned to discrete and sparse symbols. These symbols are called visual words, which represent the descriptors compactly. The assignment reduces the computation cost by turning Cartesian product distance computation into vector distance computation, as images are described as the distributions over the visual vocabulary or codebook. The visual vocabulary is so large (several millions) compared to the number of descriptors in one image that the vector descriptions of images are very sparse, which makes it possible to store the description of a very huge dataset in a compact inverted file, and to achieve a fast query.

Visual vocabulary can be learned [6] or generated by simple clustering methods from the original feature space. Usually k-means clustering is adopted as the clustering method to produce a visual vocabulary, for its simple and effectiveness. However, this method is not suitable for large scale clustering samples. According to [2], k-means tends to move centers of the clusters to denser regions in descriptor space. It is so slow to apply a k-means clustering on hundreds of millions of images where K~1 million. Nistér and Stewénius [3] introduced a hierarchical k-means clustering to accelerate vocabulary generation. The hierarchical quantization increases the speed of clustering, but at the same time the moving of centroid is magnified.

Visual vocabulary is convenient for image representation and efficient for retrieval, but these benefits come with some drawbacks. Quantization error is produced when different features in the same cluster are represented by the only centroid of the cluster. Moreover, the elements in the codebook reflect some characteristics of the dataset which the codebook is generated from, therefore using a codebook generated on one dataset for retrieval task on another dataset may meet performance drop. This is verified by our experiments. To the best of our knowledge, this

* Corresponding author. Tel.: +86 10 6260 0573; fax: +86 10 6260 0523.
E-mail addresses: yxie@jdl.ac.cn (Y. Xie), sqjiang@ict.ac.cn, sqjiang@jdl.ac.cn (S. Jiang), qmhuang@jdl.ac.cn (Q. Huang).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.03.004

has not been fully explored yet. Some papers [11,12] which are focused on image classification has mentioned this issue, but there are some differences between our work and those works, and we make a more intense research on this problem. Further details are expanded in the next section.

In this paper, the weighted visual word is proposed to measure the importance of different visual words by using the statistics information of descriptors in the sub-regions of the feature space which is divided by the visual vocabulary. This method can solve the two aforementioned problems: (1) How to reduce the trend that the centroids move towards the denser regions when the size of codebook increases; (2) How to prevent the performance worsen and increase the descriptive ability of codebook for different datasets. It consists in measuring the importance of centroids, and assigning a weight for every visual word, which depends on the distribution of the partitioning in the feature space. The harmful impact of k-means clustering is reduced by introducing weight for the visual words, and the experimental results show that the descriptive ability of codebook for different datasets is enhanced by our method. It is simple and efficient and could be easily integrated into standard BOF method.

The organization of this paper is structured as follows. In the second section, we introduce some related works about image retrieval based on bag-of-features method. In Section 3, we introduce a new perspective on the descriptive ability of codebook for different datasets, which is neglected by most of retrieval systems and tasks. In Section 4, our WV method is introduced, and experimental results are shown in Section 5.

2. Related work

Standard bag-of-feature method. The bag-of-features (BOF) framework was firstly introduced by Sivic and Zisserman [1] from text information retrieval, originally called bag-of-words model. In the information retrieval field, a document is represented by an unordered collection of words or terms, and stop words are removed. In [1], images are treated as documents and feature points as terms. SIFT descriptor [10] is chosen for its scale and rotation invariant, and it is robust to changes in viewpoint, illumination and affine distortion, which are helpful in large-scale image retrieval. Then a visual vocabulary (codebook) is generated from all the images, usually by clustering method. By using the codebook, a very large number of descriptors in the feature space are partitioned, and all the descriptors in the same cluster are considered as identical. Then one image can be described by a fixed length of vector, the length of which equals to the size of codebook, and the vector records how many descriptors are assigned to the corresponding cluster centers. Thus the collection of images can be stored in an inverted file to achieve less memory storage and better retrieval efficiency.

Term-frequency (TF) and inverse-document-frequency (IDF) are two frequently used strategies which improve the representation of documents in BOW model [13]. We assign one weight for each term in the document, which depends on the number of occurrences of the term [13]. Inverse document frequency is introduced to abate the influence of terms which occur too often in the whole collection of documents. The larger the number of documents in the collection that contain a term is, the lesser the IDF weight of the term will be.

The mean average precision (mAP) is used to evaluate the performance of all the retrieval methods [13]. The BOF method is applied to the training dataset together with a ground-truth collection and each image in the ground-truth collection is retrieved. Along with the increase of returned images, the ratio of true-positive (TP) images returned by the system is calculated and the mean value of the ratio is called mAP. mAP is widely adopted as the measurement of performance in the information retrieval field.

Codebook generation. The way of generating a codebook is various. The simplest way is to use direct k-means clustering, but it is difficult to generate a very large scale vocabulary which contains millions of visual words. Usually a vocabulary tree method [3] is utilized to generate huge vocabularies. By using it, traditional k-means clustering is divided into several smaller clustering steps. A k-tree is built top-down by hierarchical clustering, and each of the leaf node in the tree is a visual word. This method speeds up the step of codebook generation and makes it possible to generate a huge enough codebook to improve the retrieval performance. As a drawback of the k-means clustering, quantization error is brought in, and the vocabulary tree method exaggerates it. Another drawback of the vocabulary tree method is that k-means clustering tends to move the centroid of clusters to more denser regions, and hierarchical k-means clustering aggravates it [2,14]. To reduce the impact of the denser regions for “pushing” the centroid towards the high density regions, Jurie and Triggs [2] propose an approach in which the more denser regions of descriptors in feature space are eliminated after the labeling operation. Ji et al. [14] propose a density based metric learning algorithm to refine the similarity metric in the hierarchical k-means clustering.

There are many other ways of generating a vocabulary representation. Mikulik et al. [6] propose a new similarity measurement which is different from the commonly used Euclidean distance measurement and a new kind of vocabulary built on the new similarity measurement. Zhang et al. [15] propose another similarity measurement which considers the spatial contexts of local features. Wang et al. [16] consider the statistics information of descriptors on the visual vocabulary tree and spatial information of descriptors in the images. Jégou et al. [4] consist in considering the burstiness of visual words and giving the more popular visual words appeared in one image a lower weight. The method to measure the importance of visual words proposed in this paper is different from the methods mentioned above, which takes advantage of the distribution information of features in the sub-feature space divided by visual words.

The issue that one visual vocabulary constructed on one dataset and used on another dataset has been discussed in the literature [11,12]. These two papers are both focused on the problem of image classification. Depend on the experiment in [11], this case would not cause the problem of performance drop. We have performed some detailed experiments, and observed a similar result (Fig. 1) on a small visual vocabulary. Nevertheless, the experiment in that paper has used a visual vocabulary of maximum 1000. When the size of visual vocabulary grows up to hundred thousands and millions, the result is not coming out that way (Fig. 1). More importantly, for a common image retrieval task, the size of visual vocabulary is usually far more than 1000. So the experimental environment in that paper is not suitable for our work. Alqasrawi et al. [12] have mentioned this issue in their work either. There are several differences between [12] and ours. Firstly, we build a generic visual vocabulary on an uncategorized dataset, and that paper builds specific visual vocabularies on different classes and integrates them. The diversity of strategies on building visual vocabularies may cause the differences on the experimental results. Secondly, the experiments in that paper proves the conclusion that the visual vocabulary learnt on one dataset and used for retrieval on another dataset has a better performance than the baseline, but meanwhile, the authors also admit that vocabulary is not as good as the visual vocabulary learnt on one dataset and used for retrieval on the same dataset. We have performed more detailed experiments to verify it, and propose specific method to solve this problem. See Section 3 and Fig. 7 for more details.

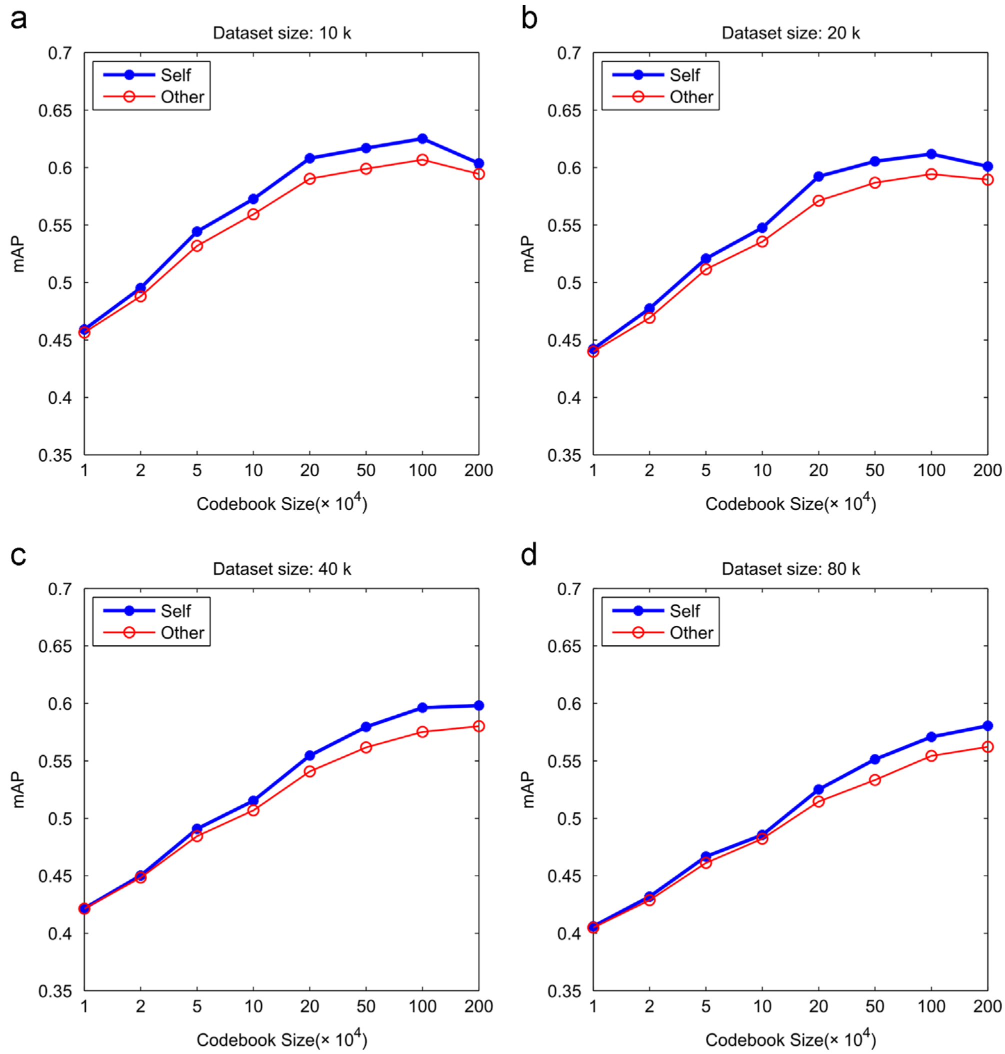

Fig. 1. The impact of codebook size on the performance of retrieval. The blue lines are higher than the red ones, as one codebook learnt on one dataset usually contains some unique characteristics got from the dataset and those characteristics are not adapted to a new dataset. This is depicted in the performance drop. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Image representation. The visual words strategy improves the speed of computing the similarity between different images by transforming the Euclidean distance computation between vectors of descriptors to Euclidean distance computation between vectors of images, and the tradeoff is that there are performance losses due to the quantization error produced when descriptors are assigned to the nearest discrete visual words. Refs. [17,5,18] consist in addressing this problem.

Philbin et al. [17] propose a soft-assignment approach rather than the hard-assignment in the standard BOF method, i.e., descriptors which are assigned to the same cluster are treated as exactly the same. In the soft-assignment method, each descriptor is assigned to n visual words with weight to reduce the quantization error brought in when really matched descriptors are assigned to different clusters. Gemert et al. [5] investigate four kinds of soft-assignment methods. The soft-assignment method outperforms well than the hard-assignment method, but by using the soft-assignment method the retrieval system needs more memory and squared more time in query, as the inverted file stores more entries and the features in query image are represented with more visual words. Some of the other practical techniques in BOF framework are discussed in [19], like vocabulary size, weighting schemes, feature selection, spatial information and visual bi-gram. The difference between our work and that paper is that we apply weighted strategy on the visual words when they are used for evaluating the similarity between images, while the weighted schemes in that paper are used for soft assignment, which is very similar to [5].

Another approach called Hamming embedding [18] is proposed to address the problem mentioned before. In [18] standard BOF method is combined together with a binary signature method. The binary signature called Hamming embedding is generated from originally SIFT descriptor, and the much more larger SIFT descriptor space is mapped into a smaller Hamming space. The measurement for the distance in the Hamming embedding space is Hamming distance, which is performed efficiently with a binary *xor* operation. In fact this method makes a tradeoff between visual words method and the simple pairwise comparison between SIFT descriptors.

The performance is improved when combining the standard BOF method with Hamming embedding, but a larger extra space is required to store the binary signature and inverted file, as one entry per descriptor rather than one entry per image is stored in the inverted list. Longer inverted list construction is also required and the query speed slows down.

Fisher vector has been acting a very important role in image retrieval and classification recently [20–25]. Bag-of-features can be regarded as a restricted Fisher vector method. Fisher vector consists in applying Fisher kernel into visual vocabularies, which is represented by a means of Gaussian mixture models (GMMs). Compared to BOF method, the Fisher vector also considers the first and second order information of visual words assignment, rather than only the word counting. Refs. [22,24] consist in generating very short signatures of images. One image can be compressed to a several hundred bits of signature without loosing too much retrieval performance, which makes it possible to apply image retrieval of tens of millions on a personal computer.

3. A new perspective on the descriptive ability of codebook for different datasets

We pay attention on the descriptive ability of codebooks for different datasets. Usually the codebook of BOF representation is generated from a very large image set, in which the number of images can be several millions or tens of millions. We evaluate the influence of codebook size on the performance of the retrieval system in Section 3.1. Given the codebook, the images in the image set are converted into vectors of histograms which describe the distribution of features assigned to the same visual words in the codebook. Nevertheless, the descriptive ability of codebook for different datasets does not get enough attention in this commonly used method. In other words, whether the codebook generated on one dataset can be used on another dataset and how it can be modified to adapt to the new dataset is not well explored. Moreover, the codebook generated on one dataset which contains a ground-truth collection is used to generate the histogram vectors of the same dataset and test the mAP performance of the same ground-truth collection, which reflects the attributes of the dataset to some extent and undeservedly increase the mAP. We address this problem, and the experimental results in Figs. 3, 4 and 1 show that there is a drop in the performance of retrieval system when mAP is calculated on a different ground-truth collections, but it truly reflects how well the system really performs.

3.1. The size of codebook

In this section, we discuss the influence of codebook size on the performance of the retrieval system. We randomly choose five 80k-size sub-datasets from the Image-Net dataset [26] and make the distribution of classes in the subset the same with the distribution in

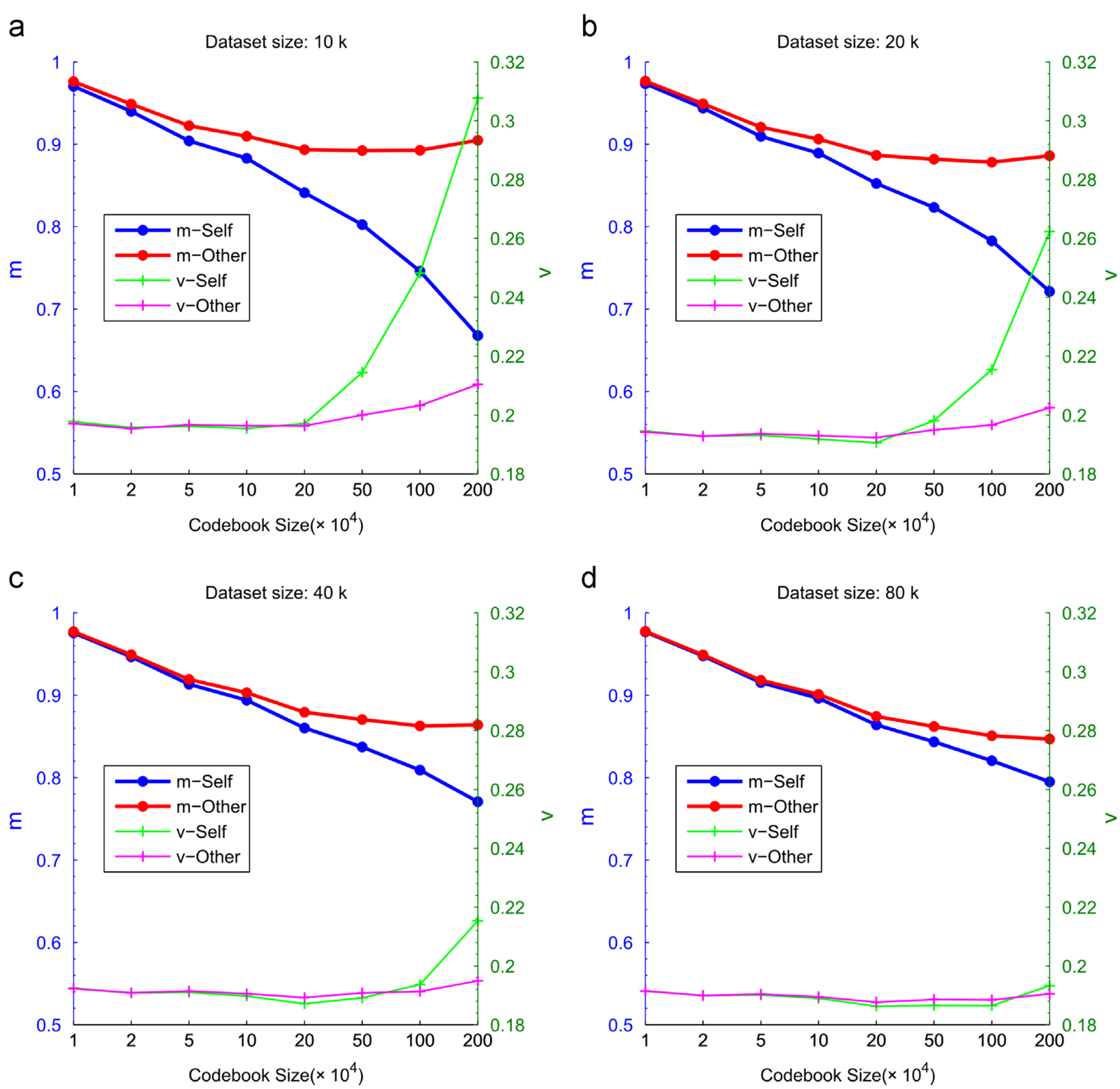

Fig. 2. The mean and variance values with respect to increasing the size of codebook and changing different size of datasets. The blue and red lines stand for the mean values, and the green and magenta ones stand for the variance values. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

the whole Image-Net dataset, and then mAP is computed on different size of subsets. We use the IPDID [27] dataset as the ground-truth dataset, which includes 10 categories and 2000 images. Codebook generated on one subset is used to compute the mAP on the other four same level subsets, i.e., codebook generated on the 20k-size subset is used on another subset if and only if the number of images in the subset is 20k. When one codebook learnt on one dataset is used to compute the mAP on the same dataset, the mAP is named homology-mAP. The homology mAPs and non-homology mAPs are averaged separately. As shown in Fig. 1, when the size of codebook increases, mAP increases first and then decreases, which reaches its peak value when the size of codebook is about one million. Compared between the sub-figures we can also come to the conclusion that as the size of dataset increases, the retrieval performance decreases.

3.2. The influence of codebook size and dataset size on the distribution of cluster

In this section, we discuss the influence of codebook size and dataset size on the distribution of cluster. Codebooks are generated on different sizes of sub-sets selected from the original datasets. By using different codebooks, descriptors are partitioned and assigned to different sub-feature spaces. We collect the statistics information of descriptors in the sub-feature spaces, which is illustrated in Fig. 2. In the legend, “m” stands for average means, “v” stands for average variance, “Self” means that the codebook used to partition the feature space is learnt from the same feature space and “Other” means that the codebook used to the feature space is learnt from another feature space.

Fig. 2 shows that the mean value decreases when the size of codebook increases. This is partly because the nature of clustering method: the number of clustering centers k increases and the size of clusters become smaller. Another reason is that the centroids of clusters move towards the denser regions, which make the clusters more compact. The blue lines drop much faster than the red lines. We use the mean value to measure whether a visual word is important or not. A visual word which gets a low weight means the cluster which takes that visual word as the centroid is more compact than other clusters. That means an overfitting may happen on that cluster, and our method reduces its adverse effect. Comparing between sub-figures, we can also see that the gaps between red lines and blue lines become smaller when the size of dataset increases. That means codebooks generated on larger dataset are more general than those on smaller dataset. We also find an interesting phenomenon when the size of codebook increases, the variance calculated on the same feature space as the one which produces the codebook becomes larger than that on a different feature space, and it also increases along with the increase on the size of codebook. The reason may be the overfitting and the quantification error grows larger when the size of codebook is more than a certain value.

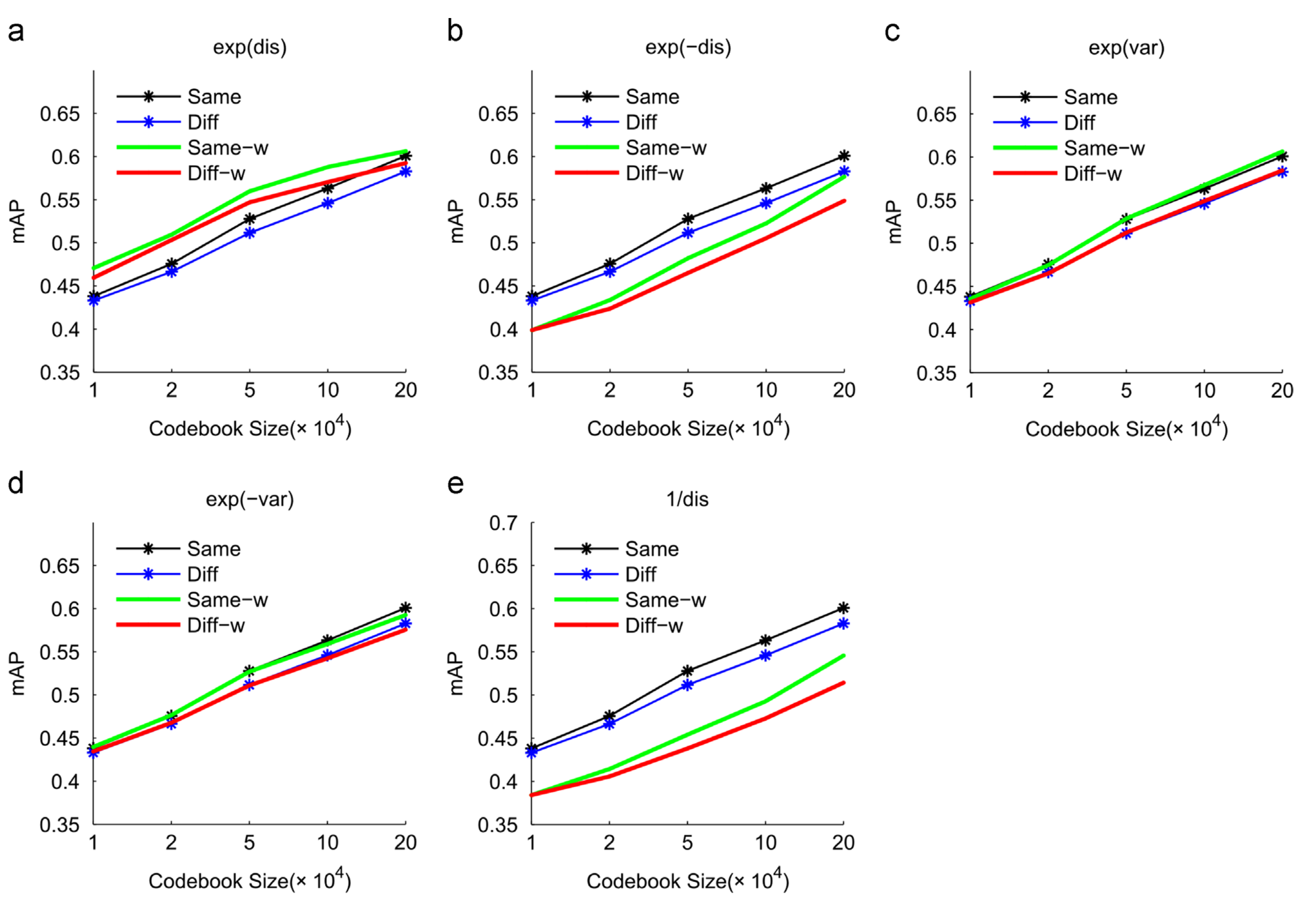

Fig. 3. The comparison of several metric for the importance of different visual words. The blue and black lines with star markers indicate the performance of original BOF method in which all the visual words are treated the same. The wider green and red lines indicate the influence of different distribution vectors on the performance of reformed BOF method. The difference between the blue lines and the black lines is the same with the difference between the green lines and the red lines, which shows that there is a performance drop when a codebook generated on one dataset is used to calculate inverted file and make retrieval on another dataset. Each sub-figure depicts the performance of one D-choosing strategy. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

4. A weighted visual word method

In the traditional visual word method, features in one image are assigned to the approximate nearest neighbors in the codebook, and the image is represented by a histogram vector $\overrightarrow{X}$, where $\overrightarrow{X} \in \mathbb{R}^k$, and k is the size of the codebook. Our method is inspired by [2,14], which figure out that there is an over-adaption procedure when the number of centroids in k-means clustering gets larger, which means that the centroids will move towards the denser regions of the feature space along with the increase of k. In other words, the sparser the distribution of features assigned to the same centroid is, the better the discrimination of centroid is. We calculate data about the distribution of descriptors in the same cluster, and use them to measure whether a centroid is "good" or not. The trend of the features' distribution in the same cluster changes along with the size of codebook and the size of image set as discussed in Section 3.2. More specifically, a vector $\overrightarrow{D} \in \mathbb{R}^k$ is calculated and each item in vector $\overrightarrow{D}$ and $\overrightarrow{X_l}$s is multiplied

$$\overline{X_j} = D_j \times X_{j,l} \tag{1}$$

$\overline{X_j}$, D_j and $X_{j,l}$ are the j-th item in vector $\overrightarrow{\overline{X}}$, $\overrightarrow{D}$ and $\overrightarrow{X_l}$, respectively, and $\overrightarrow{X_l}$ is the l-th vector of histogram that represents the l-th image in the dataset. We will show how to calculate $\overrightarrow{D}$ and how to get data for vector $\overrightarrow{D}$ in the next several paragraphs. The result vector $\overrightarrow{\overline{X}}$ is used to describe the images. The method we adopt is different from the method described in [14], which uses a metric learning method and the distance between each feature and the centroid is changed. We name it the weighted-vocabulary (WV) method.

Several strategies consisting in how to choose the distribution vector D are compared in Fig. 3 (The larger distracter image dataset and ground-truth dataset are introduced in Section 5.1.). We have tried five kinds of strategies on how to choose the best distribution vector $\overrightarrow{D}$. Let $d_{i,j}$ be the distance between the i-th feature in the j-th cluster and the centroid of the j-th cluster, which i, $j \in \mathbb{N}$, $i = 1, 2, \ldots, M_j$, $j = 1, 2, \ldots, N$, M_j is the number of features in the j-th cluster and N is the size of codebook. Now we can make several calculations on the distances

$$m_j = \frac{1}{M_j} \sum_{i=1}^{m_j} d_{i,j} \tag{2}$$

$$v_j = \frac{1}{M_j} \sum_{i=1}^{m_j} (d_{i,j} - m_j)^2 \tag{3}$$

In fact, m_j and v_j indicate the mean and variance of distances, respectively. The mean vector $\overrightarrow{m} \in \mathbb{R}_+^N$ and the variance vector $\overrightarrow{v} \in \mathbb{R}_+^N$ are used as the weighted feature vector $\overrightarrow{D}$ in the following forms: $\exp(\overrightarrow{m})$, $\exp(-\overrightarrow{m})$, $\exp(\overrightarrow{v})$, $\exp(-\overrightarrow{v})$ and $\overrightarrow{m}^{-1}$. In Fig. 3, from left to right and from top to bottom, these vectors are used in the WV

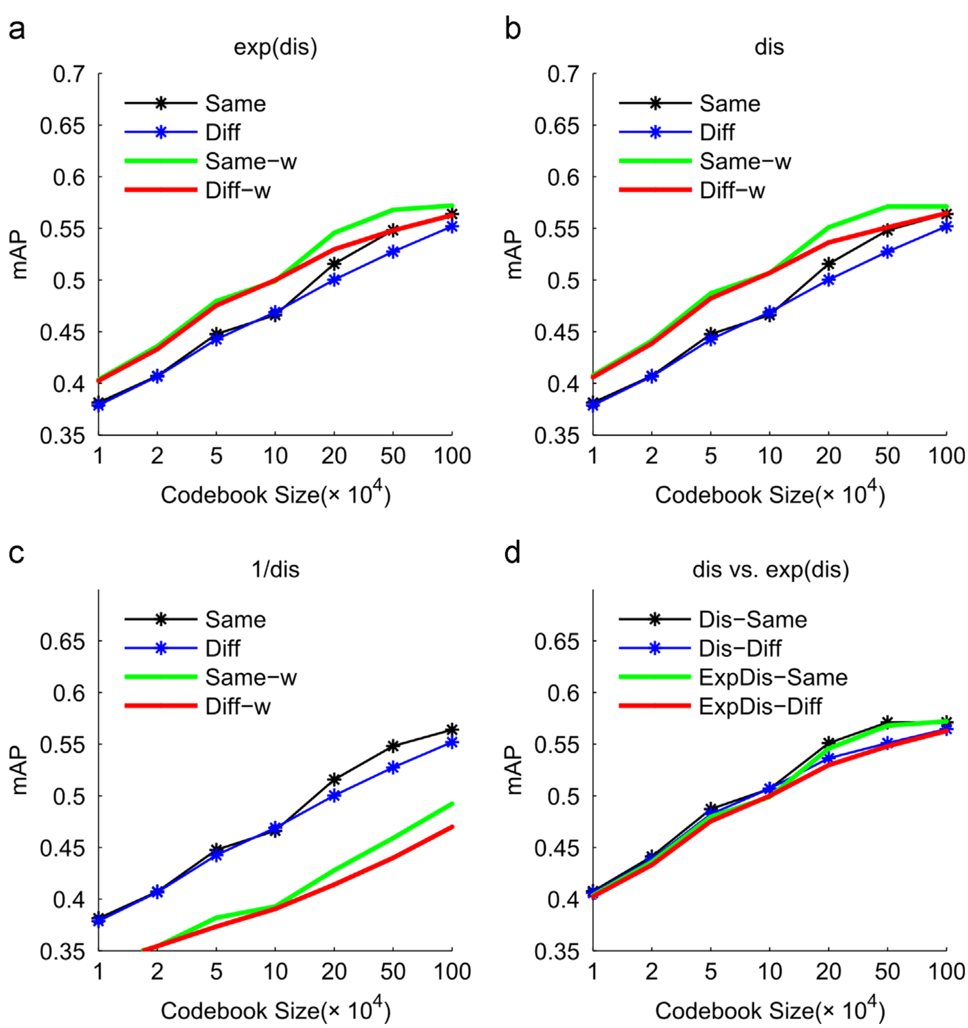

Fig. 4. The comparison of several metric for the importance of different visual words on a larger dataset (with 80 000 images). The meaning of colored lines is the same as the lines in Fig. 3. The sub-figure (d) illustrates the comparison on the performance of two best strategies. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

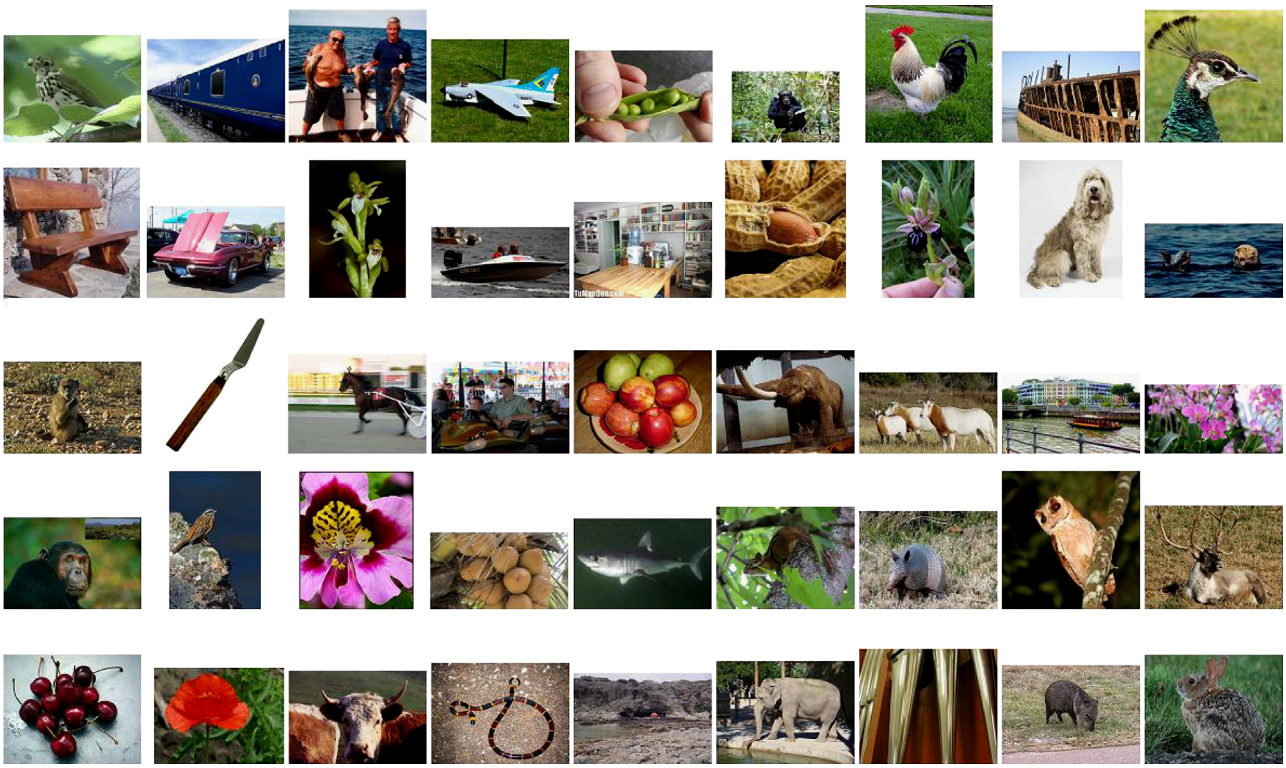

Fig. 5. Some representative sample images from the subset of ImageNet dataset.

Fig. 6. Some sample images from the IPDID dataset.

method, and the performance of each one is showed in the sub-figures. The number of images in all the distracter dataset is 10 000. The black lines show the performance of original BOF method, and the blue ones show the performance of retrieval by using the same codebook and ground-truth dataset but testing on a different distracter dataset. The difference between the green lines and the red lines is the same. Green and red ones show the impact of vector $\overrightarrow{D}$ on the standard BOF method. The following conclusions can be drawn from the figure:

1. The performance becomes better along with the increase in the size of codebook.
2. The variances of feature distribution have a negligible influence on the performance.
3. There is a drop in the performance when retrieval is made on a difference dataset.
4. A monotone increasing function of the mean vector may be used to improve the efficiency of retrieval, which decreases the performance drop when different datasets are retrieved, and it even works better than the original method.

We test the straightforward mean vector $\overrightarrow{m}$ on a larger dataset (with 80 000 images), and it performs on a par with the $\exp(\overrightarrow{m})$ method, which is depicted in Fig. 4. Sub-figures (a)–(c) show that the $\exp(\overrightarrow{m})$ and $\overrightarrow{m}$ strategies still show a high performance when the $\overrightarrow{m}^{-1}$ works bad. Sub-figure (d) depicts a comparison of performance on the two well-running strategies. The performance gap between them is really very small. We choose $\overrightarrow{m}$ as the kernel vector $\overrightarrow{D}$ for the computational efficiency.

5. Experimental results

5.1. Dataset

A subset from ImageNet dataset. Our experiment is set up on the subsets where images are randomly chosen from the 1000 categories in ILSVRC2011 (ImageNet Large Scale Visual Recognition Challenge 2011) competition of ImageNet dataset [26,28]. Different random selecting parameters are taken to produce subsets with a fixed size of images, such as 10 000 and 80 000. We choose four size levels of subsets, which contain 10k, 20k, 40k and 80k of images, and the former one is the subset of the latter one, which means the comparisons of mAP computed on these different datasets are meaningful. We choose several 80k-size subsets (Fig. 5) from the original ImageNet dataset, one for training and another for testing. In other words, we choose 160 images from each of the 1000 categories, thus the subset is composed of the collection of the 160k images. Fig. 5 shows some representative sample images, which are selected from different categories.

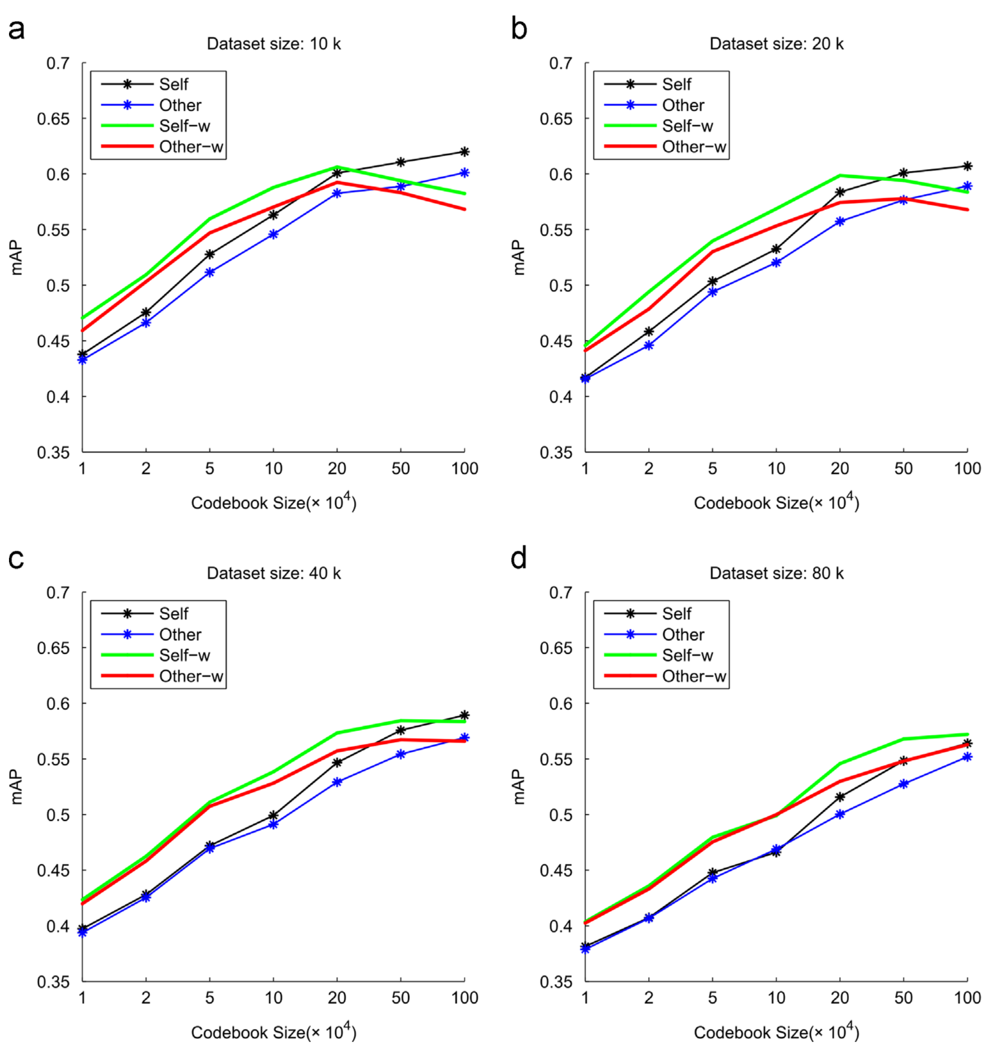

Fig. 7. The retrieval accuracy when our method is applied to different size of datasets ($\overrightarrow{m}$ is chosen as the kernel vector $\overrightarrow{D}$). The meaning of colored lines is the same as the lines in Fig. 3. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

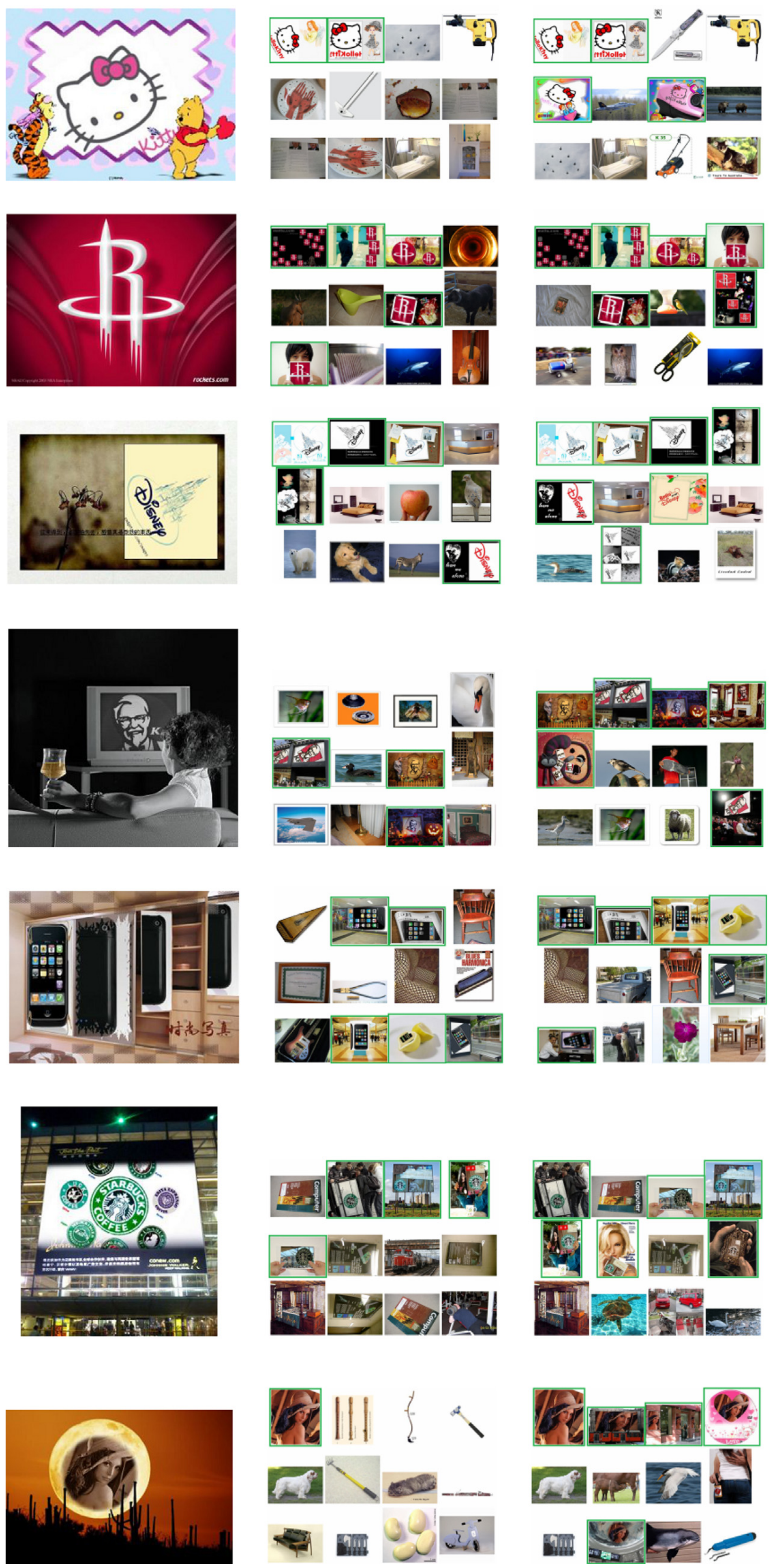

Fig. 8. Some query images and corresponding retrieved results. Query images are from the IPDID dataset and the distracter images are from the subset of the ImageNet dataset, see Section 5.1 for more details. Each row shows a query image and 12 most similar images from the dataset which are returned by two different methods. The middle column shows the retrieval results of baseline method, and the right column shows the retrieval results of our method. Images with green borders are true positive results. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

The ground-truth dataset. The ground-truth dataset is from [29] and can be downloaded from [27], which contains 10 image collections and 2000 images (Fig. 6). Equal division is made on the ground-truth dataset, and the subsets are used for codebook generation and mAP calculation: If a codebook is a homology-codebook, the codebook is generated on a ground-truth subset with a larger distracter dataset, and used for mAP calculation on the same ground-truth subset. Otherwise, the codebook is generated and used for mAP generation on the same ground-truth subset, but the distracter dataset is changed.

Descriptors are detected and extracted by using the SIFT method [10] and the code provided by Rob Hess [30]. Codebooks are generated by using hierarchical *k*-means clustering [3]. mAP (mean average precision) is used for evaluation [31], more specifically, each image in the ground-truth folder is taken as a query. We obtain a precision-recall curve and an average precision (AP) for each image, and the mean value of AP on all the images is taken as the evaluation of retrieval performance.

5.2. Evaluation of our WV method

Our WV method has advantages not only of boosting performance up on the standard BOF method but also of low computational cost. It can be easily integrated into the standard BOF method without making much modification, which means any algorithms and applications using standard BOF method can take our method as an optional add-on. The parameter calculation can be performed off-line and is with very low computational complexity, compared to the codebook generation. The memory storage of the parameters is also very small, 4 bytes for each visual word, and 4 millon bytes when we use a one-millon-size codebook for retrieval, which is a negligible number compared to the memory consumption of the inverted file.

Fig. 7 shows the retrieval accuracy when our method is applied to different size of datasets, and $\vec{m}$ is chosen as the kernel vector $\vec{D}$. When the size of dataset is very small (10 k images as shown in sub-figure (a)), our WV method works better than the standard BOF method when the size of codebook is below 200 k. Nevertheless, the size of dataset increases to 80 k, and our WV method outperforms the standard BOF method on all sizes of codebooks. The cause of this phenomenon may be the over segmentation of the feature space and the bad clustering. In this situation our method assign higher weights for those features in a cluster which has a sparser distribution and worsen the clustering. As the size of dataset increases, the impact of same size codebook lessens, and our method works better. It is expected that our method will work better on a larger dataset.

As shown in Figs. 3, 4 and 7, our method got about 3% of performance increase on the standard BOF method with very little memory consumption, when our method is applied on 80 k size dataset and the size of codebook is 500 k. Moreover, our method can reduce the performance drop when codebook is used on another dataset.

Fig. 8 shows some retrieval results for our method and the baseline method using the images from the ground truth dataset. We add green borders around the true positive images that returned. We can see that the rank of true positives is improved in our method, and the retrieval system returns more true positive images by using the proposed approach.

6. Conclusion

This paper provides a new perspective on the descriptive ability of codebook for different datasets, and proposes a weighting based method to increase the performance of image retrieval. The importance of different visual words is measured by the distribution information of features in the partitioning by the codebook. The experimental results show that the descriptive ability of codebook for different datasets is enhanced by our method. Our method is simple and efficient, and could be easily integrated into standard BOF method to increase retrieval performance, with very little memory usage. We currently use the distribution information of features in the clusters. In the future, we hope to find more effective measurements of visual words, and apply the weighting based method to the clustering step.

Acknowledgments

This work was supported in part by National Basic Research Program of China (973 Program):2012CB316400, in part by National Natural Science Foundation of China: 61070108, 61025011, and 61035001, in part by the Key Technologies R&D Program of China under Grant no. 2012BAH18B02.

References

[1] J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in: ICCV, IEEE Computer Society, Washington, DC, USA, 2003, pp. 1470.
[2] F. Jurie, B. Triggs, Creating efficient codebooks for visual recognition, in: ICCV, vol. 1, 2005, pp. 604–610.
[3] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: CVPR, IEEE Computer Society, Washington, DC, USA, 2006, pp. 2161–2168.
[4] H. Jégou, M. Douze, C. Schmid, On the burstiness of visual elements, in: CVPR, 2009, pp. 1169–1176.
[5] J. Van Gemert, C. Veenman, A. Smeulders, J.-M. Geusebroek, Visual word ambiguity, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1271–1283.
[6] A. Mikulík, M. Perdoch, O. Chum, J. Matas, Learning a fine vocabulary, in: ECCV, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–14.
[7] H. Jégou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, Int. J. Comput. Vision 87 (2010) 316–336.
[8] L. Li, S. Jiang, Q. Huang, Learning hierarchical semantic description via mixed-norm regularization for image understanding, IEEE Trans. Multimedia 14 (2012) 1401–1413.
[9] L. Li, S. Jiang, Z. Zha, Z. Wu, Q. Huang, Partial-duplicate image retrieval via saliency-guided visually matching, IEEE Multimedia, in press.
[10] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2004) 91–110.
[11] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, A thousand words in a scene, IEEE Trans. Pattern Anal. Mach. Intell. 29 (9) (2007) 1575–1589.
[12] Y. Alqasrawi, D. Neagu, P. Cowling, Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification, Signal Image Video Process. 23 (2011) 1–17.
[13] C.D. Manning, P. Raghavan, H. Schtze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.
[14] R. Ji, H. Yao, X. Xie, Q. Tian, Vocabulary hierarchy optimization and transfer for scalable image search, IEEE Multimedia 18 (3) (2011) 66–77.
[15] S. Zhang, Q. Huang, G. Hua, S. Jiang, W. Gao, Q. Tian, Building contextual visual vocabulary for large-scale image applications, in: Proceedings of the International Conference on Multimedia, ACM, New York, NY, USA, 2010, pp. 501–510.
[16] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, T. Han, Contextual weighting for vocabulary tree based image retrieval, in: ICCV, 2011, pp. 209–216.
[17] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: improving particular object retrieval in large scale image databases, in: CVPR, 2008, pp. 1–8.
[18] H. Jégou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: ECCV, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 304–317.
[19] Y. Jiang, J. Yang, C. Ngo, A. Hauptmann, Representations of keypoint-based semantic concept detection: a comprehensive study, IEEE Trans. Multimedia 12 (1) (2010) 42–53.
[20] H. Jégou, M. Douze, C. Schmid, Packing bag-of-features, in: ICCV, 2009, pp. 2357–2364.
[21] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: CVPR, 2007, pp. 1–8.
[22] F. Perronnin, Y. Liu, J. Sándnchez, H. Poirier, Large-scale image retrieval with compressed Fisher vectors, in: CVPR, 2010, pp. 3384–3391.
[23] F. Perronnin, J. Sánchez, T. Mensink, Improving the Fisher kernel for large-scale image classification, Comput. Vision ECCV 2010 (2010) 143–156.
[24] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: CVPR, 2010, pp. 3304–3311.
[25] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Pérez, C. Schmid, Aggregating local images descriptors into compact codes, IEEE Trans. Pattern Anal. Mach. Intell. PP 99 (2011) 1.

[26] Image-net online, ⟨http://www.image-net.org⟩.
[27] Internet partial-duplicate image database (ipdid), URL ⟨http://www.jdl.ac.cn/en/project/mrhomepage/IPDID.htm⟩.
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: CVPR, 2009, pp. 248–255.
[29] Z. Wu, Q. Xu, S. Jiang, Q. Huang, P. Cui, L. Li, Adding affine invariant geometric constraint for partial-duplicate image retrieval, in: ICPR, IEEE Computer Society, Washington, DC, USA, 2010, pp. 842–845.
[30] R. Hess, Sift library, URL ⟨http://blogs.oregonstate.edu/hess/code/sift⟩.
[31] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: CVPR, 2007, pp. 1–8.

Yi Xie is a master student of Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. He received B.S. degree in electronic engineering in Tsinghua University, in 2010. His research interests are visual information processing and retrieval.

Shuqiang Jiang is an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. He is also with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 100 papers on the related research topics. Dr. Jiang was supported by the New-Star Program of Science and Technology of Beijing Metropolis, in 2008. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences, in 2012, and the CCF Award of Science and Technology, in 2012, respectively. He is the senior member of IEEE, member of ACM, CCF, and YOCSEF. Dr. Jiang is the executive committee member of ACM SIGMM China chapter. He has been serving as the guest editor of the special issues for PR and MTA. He is the program chair of ICIMCS2010, special session chair of PCM2008, ICIMCS2012, area chair of PCIVT2011, publicity chair of PCM2011 and proceedings chair of MMSP2011. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICIP, and PCM.

Qingming Huang (SM'08) received the B.S. degree in computer science and Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively. He is currently a professor with the University of the Chinese Academy of Sciences (CAS), Beijing, China, and an adjunct research professor with the Institute of Computing Technology, CAS. He has been granted by China National Funds for Distinguished Young Scientists in 2010. He has authored or coauthored more than 200 academic papers in prestigious international journals and conferences. His research areas include multimedia computing, image and video processing, computer vision, and pattern recognition.