Generic Object Regions Matching Based VLAD Model for Image Retrieval

Hongyan Zhai

Guangdong Industry Polytechnic, Guangzhou 510300, China

teacherzhai@126.com

Keywords: Image Retrieval, Multi-threshold Segmentation, Gaussian Mixture Model, VLAD, SURF

Abstract. In the popularly used BoVW and VLAD models for image retrieval, feature extraction is easily affected by the color and textures of images. Also, the clustering results of K-means algorithm used in these two models is usually affected by the initial cluster centroids. In order to solve these problems, a generic object regions matching based VLAD model for image retrieval is proposed. In this model, multi-threshold for image segmentation is proposed for the extraction of SURF features. Then the location information of SURF features are utilized for the Gaussian Mixture Model clustering instead of K-means algorithm. Finally, VLAD descriptors are calculated according to the features in each cluster and used for similar image searching. Our proposed Multi-Threshold Segmentation and SURF Feature Location based Clustering algorithm can improve the matching accuracy of features and obtain a better codebook such that the feature distribution can be better expressed. Experimental results on the Holidays dataset show that the mAP of our algorithm is higher than the 5 mainstream image retrieval algorithms, which efficiently improves the image retrieval accuracy.

1. Introduction

With the development of internet and multimedia techniques, how to retrieve similar images from large image dataset conveniently, quickly and accurately becomes more and more important. For example, in the public security area, image retrieval can be used to find crime evidences from massive surveillance videos, which will significantly improve the accuracy and efficiency for the case detection. In the recent years, Content-Based Image Retrieval (CBIR) achieves great attentions in the areas of multimedia processing, information retrieval, artificial intelligent, database and deep learning etc. "Content-based" means that the search analyzes the contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image. The term "content" might refer to colors, shapes, textures, or any other information that can be derived from the image itself, which is independent with annotation quality and completeness. Thus, CBIR can be more objective and accurate.

In the CBIR, color and texture are two of the commonly used global features. Also, recently, local features such as SIFT and SURF show their efficiency and robustness for object deformation and illumination variance. Moreover, the fusion of global and local features is a good choice for similar image search and object search[1]. The popularly used method for image retrieval includes three steps: 1. Feature extraction; 2. Bag-of-Visual-Words(BoVW) or Vector of Locally Aggregated Descriptors (VLAD)[2] schemes is used to obtain the local feature expression; 3. Similarity functions are used to calculate the similarities between the query image and images in the dataset. Then the similar images in the dataset are retrieved according to their similarity values.

In the BoVW scheme, image feature information will be lost by using the local feature representation method. Also, the codebook size is very large such that it is difficult for feature clustering[5]. VLAD scheme overcomes these disadvantages. It has become a hot issue in the image retrieval[2-8]. However, VLAD scheme only extract feature directly, not considering the relations of feature locations. Thus, Liu et.al. [8] employed the K-means algorithm to cluster the features based on their locations. Then a circle with radius r is selected based on the Cluster centroid. Features located in the circle are represented as a vector for image retrieval. But how to choose a

suitable *r* is an obstacle. In [9], Wang et.al. proposed a method for image retrieval based on the salient region segmentation. It used the Fingerprint algorithm to segment the object area of an image. Then the object area is used for matching such that the matching accuracy can be improved. But its performance will be affected when the object cannot be segmented completely.

In this paper, an image retrieval algorithm is proposed, which imposes multiple thresholds segmentation and SURF feature clustering based on the feature locations. First, an image is segmented by multiple thresholds and SURF visual features are extracted. Second, based on the locations of SURF features, Gaussian mixture model(GMM) is used to cluster SURF features into different clusters. Finally, VLAD scheme is used for quantization of each cluster of SURF features.

Image segmentation by multiple thresholds can normalize the image color and texture, which reduces the interferences of the color and texture for the extraction of SURF features. Clustering based on the feature locations can quantize the features falling into a same image area together, which will reinforce the matching of useful information and reduce the interferences of useless information. In the stage of retrieval, the quantized VLAD descriptors obtained from each cluster of the query image is used to calculate the similarity scores with the quantized VLAD descriptors obtained from the dataset images. Then the retrieval results can be obtained according to the similarity scores. The main contributions of our paper includes:

（1）Multi-threshold segmentation of the images is proposed, which can normalize the color and texture features such that the extracted SURF features can be more robustness.

（2）Feature locations based clustering by GMM is proposed, which will cluster the neighbor features of a same object into one cluster such that these features can be quantized into one descriptor vector. It transfers the image matching into the object matching.

2. Generic object regions matching based VLAD model for image retrieval

SIFT and SURF are two kinds of commonly used features in the image retrieval. Compared with the SIFT feature, SURF feature can be extracted more quickly and its dimension is only a half of the SIFT feature. Also, the performance of SURF feature is comparable with SIFT feature. Thus SURF feature is selected in our algorithm. The general process of the image retrieval is shown in Fig. 1. It is divided into two parts: off-line and online processes. In the off-line process, SURF features of the dataset images are extracted and clustered. Each cluster centroid is a visual word and all of the centroids consist the codebook. All of the SURF features of an image in the dataset are quantized into a vector according to their nearest visual words in the codebook. In the online process, all of the SURF features of the query image are also quantized into a vector according to their nearest visual words in the codebook. Then the similarity scores can be calculated by the quantization vectors of the query and dataset images. Finally, retrieval result is obtained according to the values of the similarity scores.

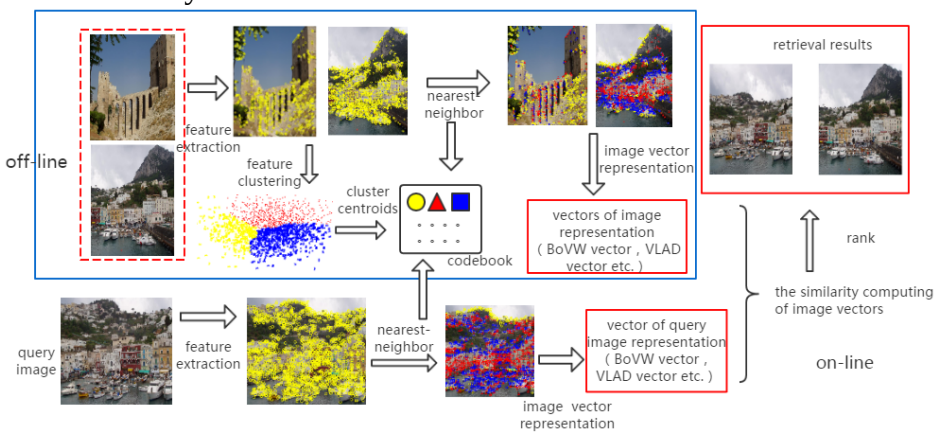


Fig. 1 The general image retrieval process

BoVW is a classical image retrieval model and achieved a lot of attentions in applications. It quantizes the SURF features of an image into a histogram vector by the codebook. The histogram vector only records the frequency of the features falling into a bin. The information of the original

SURF feature values is totally lost. Moreover, cluster process for the codebook construction is very slow since there are too many features. In general, the codebook size is more than 20k. Also, a codebook with large size will lead to a slow quantization speed. These shortcomings make BoVW hard to be implemented in the hardware. Thus, in the recent years, VLAD model is proposed and applied for image retrieval widely.

The main difference between BoVW and VLAD is the quantization process. Different with the BoVW, the quantization result of VLAD is a vector of the residuals of the SURF features and the visual words, which contains more local feature information. In addition, the codebook size of VLAD is much smaller than BoVW. In general, the features of the dataset images are clustered into 64 classes, i.e., there are only 64 visual words in the codebook, which significantly reduces the clustering time and quantization time. The quantization process of VLAD model is as follows: Suppose that there are *K* visual words in the codebook. For each visual word c_k and each local feature d_i in the image, we have

$$v_k = \sum_{d_i \in \zeta_k} (d_i - c_k), \ 1 \le k \le K \qquad (1)$$

where $\zeta_k = \{d_i | ||d_i - c_k||^2 < ||d_i - c_{k'}||^2, \forall k' \neq k\}$ denotes the nearest-neighbor of the feature d_i in the codebook. c_k is the *k*th visual word in the codebook. v_k is the residual vector corresponding to the visual word c_k. The conjunction of all of the *K* residuals forms the VLAD descriptor, i.e., $V = [v_1, v_2, \cdots, v_K]$, Then V is normalized by $V = V / \| V \|$, where $\| V \|$ denotes the 2-norm of V.

Although the above mentioned VLAD model significantly reduces the codebook size and the feature information is added into the final VLAD descriptor, it still quantizes all of the SURF features obtained from a image into one descriptor vector. As we know, the generic object regions in an image are very important for the understanding of the image. In many cases, the generic object regions in an image imply the key information of this image. Thus, we consider clustering the SURF features extracted from a same image according to their locations in the image such that the features belong to a same generic object region can be clustered into one cluster. All features in a same cluster are quantized by the codebook to form a VLAD descriptor. Then the similarity scores among the query image and images in the dataset can be calculated by the VLAD descriptors of the query image and each image in the dataset.

The generic object regions matching based VLAD model for image retrieval proposed in this paper is shown in Fig. 2. In the off-line process, SURF features are extracted from the images obtained by the multi-threshold segmentation. After the codebook is constructed, GMM model is used to cluster the SURF features based on the feature locations for each image in the dataset. Then the features in each cluster are quantized into a VLAD descriptor by (1). Also, the VLAD descriptors of the query image can be obtained in a same way. The similarity score S_I between the query image and the *I*th image in the dataset is calculated as follows.

$$S_I = \sum_{k=1}^{K} \min_{j=1}^{K} ||Q_k - I_j||_2, \qquad (2)$$

where *K* denotes the VLAD descriptor number of an image, i.e., the number of clusters obtained by the GMM algorithm. Q_k denotes the *k*th VLAD descriptor of the query image. I_j denotes the jth VLAD descriptor of the *I*th image in the dataset. $\min_{j=1}^{K} ||Q_k - I_j||_2$ denotes the minimal distance among Q_k and the all the VLAD descriptors of the *I*th image in the dataset. Let

$$s_{kI} = \min_{j=1}^{K} ||Q_k - I_j||_2, \qquad (3)$$

then (2) can be simplified as

$$S_I = \sum_{k=1}^{K} s_{kI} \qquad (4)$$

Finally, we sort the values of S_I in a decent order and the image corresponding to the biggest 10 similarity scores are output as the retrieval results.

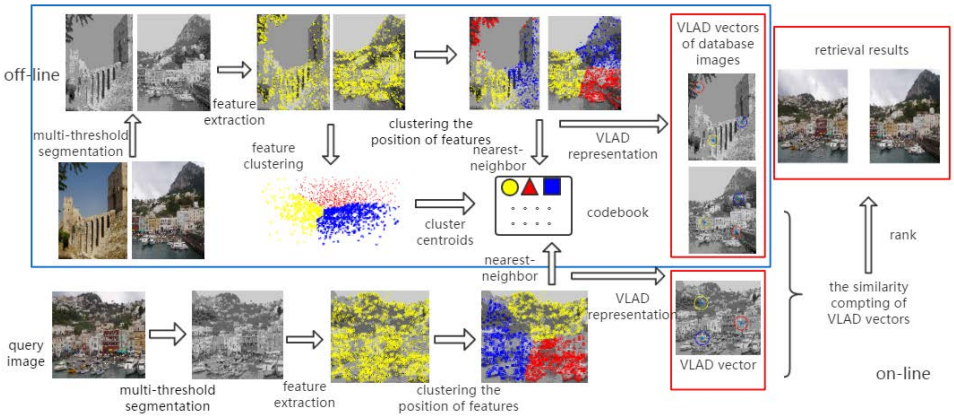


Fig. 2 The generic object regions matching based VLAD model for image retrieval proposed in this paper.

3. Multi-threshold segmentation and GMM clustering

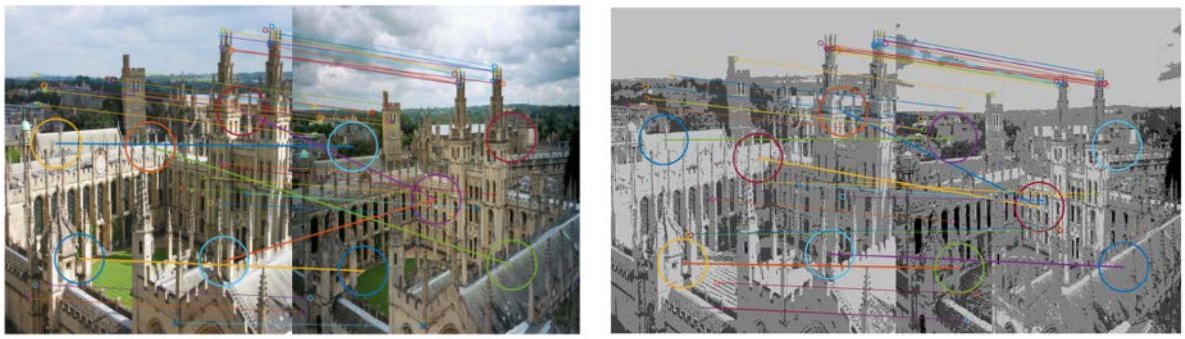
In order to efficient distinguish the image's outline, multi-threshold segmentation is proposed in our algorithm. Also normalization is applied to the image pixels to diminish the noise of local features. According to the local SURF features clustering based on the feature locations, the neighbor features will be clustered into a same cluster such that the generic object regions in the image can be represented more accuracy[8].

3.1 Multi-threshold segmentation

The weakness of the method proposed in [9] is that the salient region of the image need to be detected in the first. But it is difficult to detect and segment an object from the image accurately. This may cause missing parts of the object region and lost a lot of object information, which will affect the image matching results significantly. Also, we find that the color and texture etc. of an image may affect the feature extraction and matching. Thus, in order to normalize the image color and texture information and reduce the noise of local SURF features, multi-threshold segmentation method is proposed, which is simple but will not lose the object outline in the image. According to the experiments, we find that in general there are very few local SURF features in the background region. These background features will disappear after gray-threshold segmentation, which can significantly diminish the noise of features that exist in background region and enhance the performance of similar image matching. Our proposed multi-threshold segmentation method is described as follows.

First, the color image is transformed into a gray image with gray values falling in the interval [0, 1]. Second, [0, 1] is divided into several small intervals with multiple thresholds. Third, for every gray value, if it belongs to a small interval, it is substituted by the lower-bound of this interval. For example, if the threshold values are $(0, t_1, \cdots, t_n, 1)$, then all of the gray values falling into the interval $[0, t_1)$ will be set as 0, and pixel values belonging to the interval $[t_{n-1}, t_n)$ will be set as t_{n-1}. Finally the multi-threshold segmentation result can be obtained. The intrinsic of this method is reducing the gray levels and obtain the distinguishable outline of each image. At the same time, this method can reduce or eliminate the noise of the background region for each image. Figure 3 shows an example of the multi-threshold segmentation result.

After the multi-threshold segmentation, local SURF features are extracted and their corresponding coordinates are recorded. Based on the locations of the local SURF features, GMM algorithm is applied for feature clustering and VLAD descriptor of each cluster can be obtained by the VLAD model. If the number of clusters is K, then the number of VLAD descriptors of each image is also K. For example, in Fig. 3, the features of each image are clustered into 5 clusters. Thus 5 VLAD descriptors are obtained for each image. The locations of big circles are the VLAD descriptors computed by Eq. (1) after GMM clustering.



(a) (b)

Fig. 3 Comparison of the matching results. (a) Matching result of the VLAD descriptors. GMM clustering is applied without multi-threshold segmentation. (b) Matching result of our proposed method. Both GMM clustering and multi-threshold segmentation are applied.

It can be seen that after the multi-threshold segmentation and GMM clustering, the matching results are more accurate compared with the results obtained without multi-threshold segmentation.

3.2 GMM clustering

The method of Liu [8] has two weaknesses. Fist, the initial cluster centroids have a decisive effect for the clustering results. Different initial cluster centroids may result in different clusters for a same set of samples. Second, after the clustering algorithm is performed, a circle centering in the cluster centroid is selected as the outline of objects. This is under the hypothesis that the outlines of objects in every image are circles. Also, it is difficult to select an appropriate value of *r*.

In order to overcome these two problems, firstly we add GMM clustering algorithm in the clustering step. The initial cluster centroids have a decisive influence for the GMM clustering as well as the K-means clustering algorithm. However, after K-means algorithm is performed, the initial cluster centroid values can be set as the cluster centroids of GMM clustering algorithm, which can effectively solve the cluster centroids initialization problem. In the meanwhile, due to the GMM clustering algorithm is implemented according to the distribution of samples, the clustering results are more accurate than the K-means algorithm. But GMM clustering algorithm is significantly slow when the data set is too large. Due to the number of SURF features extracted from one image is not very large (the average number is 2000) and the cluster number is less than 10, thus it is suitable by using GMM clustering algorithm in our proposed method.

Secondly, in our algorithm, the approximate outlines of the image is obtained after the multi-threshold segmentation and all of the features in each cluster are represented as a VLAD descriptor. This can express the SURF features distribution much better than the method mentioned in [8].

In the GMM clustering algorithm, the probability of each sample is computed and the clustering results are obtained according to the probability. It assumes that the samples follow Gaussian mixture distribution. Each Gaussian mixture model is consisted by multiple Gaussian distributions. When it is used as clustering, each Gaussian distribution corresponding to a cluster. GMM computes the probability of each sample that clustered into each cluster by the Expectation Maximization (EM) algorithm according to the mean μ_k, covariance Σ_k and mixing coefficient π_k of each Gaussian distribution. Mathematically. the probability of x_n clustered into the *k*-th cluster is calculated as:

$$\mathrm{r}(z_{nk}) = \mathrm{p}(z_k = 1|x_n) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)},$$

where $N(x_n|\mu_k,\Sigma_k)$ is the distribution with mean μ_k and the covariance Σ_k.The iterative steps for calculating μ_k, Σ_k and π_k are illustrated as follows.

$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\mathrm{r}(z_{nk})x_n;$

$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\mathrm{r}(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T;$

$\pi_k = \frac{N_k}{N}.$

Here $N_k = \sum_{\mathrm{n}=1}^{N}\mathrm{r}(z_{nk})$. The stop criterion is that the difference of the likelihood function values in

the two adjacent iterations is less than a given constant or the iteration time reaches the maximal. The value of likelihood function is computed by:

$$\ln\big(\mathrm{P}(\mathrm{x}|\pi,\mu,\Sigma)\big)=\sum_{n=1}^{N}\ln[\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)]$$

The cluster centroids obtained by GMM are the mean μ_k of every Gaussian distribution. The probability of data x_n belonging to the k-th cluster is $\mathrm{r}(z_{nk})$. The corresponding cluster of the maximal probability of x_n is the final cluster that x_n belongs to.

GMM clusters the samples by estimating the probability of each sample belonging to every Gaussian distribution. If the selection of initial values of the mean vectors of Gaussian distributions are suitable, the cluster results will be very stable. An effective method is using the K-means cluster centroids as the initial centers for GMM. Although the results of K-means clustering is unstable, the stability of cluster results will be significantly enhanced after GMM clustering algorithm is performed.

4. Experimental results

Based on the proposed method, we use the open SURF source code of MATLAB to extract the SURF features. Image retrieval system is implemented on a laptop with 2.5GHz Inter i5 CPU, 8G CPU and 64 bits Windows operating system. Image retrieval experiment is realized on INRIA Holidays[10] dataset. This dataset contains 1491 images with 500 query images and the corresponding ground truth.

We use average recall and mean Average Precision (mAP) as the measurement indicators. The recall is computed by:

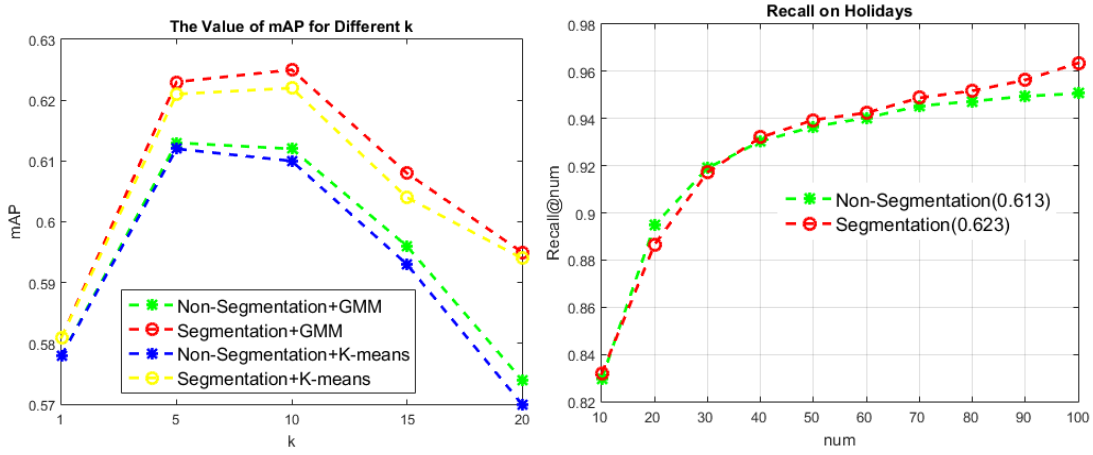
$$\text{Recall}=\frac{the\ number\ of\ retrieved\ relevant\ images}{the\ number\ of\ relevant\ images\ in\ the\ dataset}$$

Average recall value equals to the division result of the sum of all the recalls of all the query images divided by the number of query images. The Average Precision is computed by:

$$\text{AP}=\frac{sum\ of\ the\ precisions\ of\ retrieved\ relevant\ images}{the\ number\ of\ relevant\ images\ in\ the\ dataset}$$

The precision of a retrieved relevant image is defined as: the number of all *retrieved relevant images* rank ahead this image divided by the number of all *retrieved images* rank ahead this image.

After a plenty of experiments we find that the segmentation result is the best while the gray level is set as 5. Thus, in our experiments, the gray threshold partition intervals are set as [0,0.2), [0.2,0.4), [0.4,0.6), [0.6,0.8), [0.8,1). If the gray level is less than 5, then the information will be lost too much, otherwise the retrieval performance will not be improved. Actually, most of the images just contain up to 5 generic object regions, thus it is reasonable to scale the gray level into 5 levels from a practical perspective.



(a) (b)

Fig. 4 Retrieval results comparison on the Holidays database. (a) The mAP curves of

non-segmentation and multi-threshold segmentation corresponding to the number of clusters for K-means and GMM, respectively. The horizontal axis is the number of clusters and the vertical axis is the mAP corresponding to the horizontal axis. (b) The recall curves of non-segmentation and multi-threshold segmentation. The horizontal axis is the number of top num images after sorted. The vertical axis is the recall@num corresponding to the horizontal axis.

The cluster centroids of K-means algorithm are used as the initial centroids of the GMM algorithm. In order to obtain an appropriate cluster number, we tested the effects of the cluster number for the retrieval performance. Fig. 4(a) is the mAP curves before and after multi-threshold segmentation of K-means and GMM clustering algorithms, respectively. It can be seen that the retrieval results of GMM are subtly superior to the results of K-means. With the number of clusters ranges from 5 to 10, the time consumption will be increased, but the retrieval performance is only slightly increased(for the Segmentation+GMM and Segmentation+K-means). Considering the time consumption and retrieval performances, k=5 will be a better choice.

Our retrieval results on the Holidays database are shown in Fig. 4(b). the mAP is 0.613 before multi-threshold segmentation, while the mAP is 0.623 after multi-threshold segmentation. Meanwhile, from the recall curves shown in Fig. 4 (b), we notice that the recall value of non-segmentation is lower than the multi-segmentation when num>30. The recall of multi-threshold segmentation is 0.965(Recall@100), which is higher than the recall of the non-segmentation, i.e., 0.95(Recall@100). Thus, the retrieval performance improved by the multi-threshold segmentation, for mAP and recall, 1% and 1.5% (Recall@100), respectively. Fig. 5 and Fig. 6 are two examples of our image retrieval results. It can be seen that the top 3 images of the retrieval results are more accuracy.

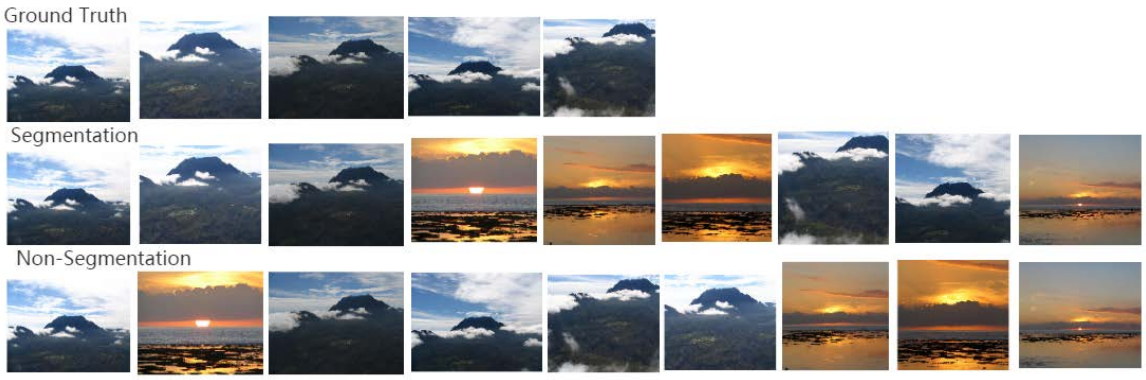


Fig. 5 The retrieval results of a scenery image, the first image of each row is the query image. Images in the first row are the ground truth, the second row is the retrieval results obtained by multi-threshold segmentation and the third row is the retrieval results obtained by non-segmentation. The AP of multi-threshold segmentation is 0.7679 and non-segmentation is 0.6792.

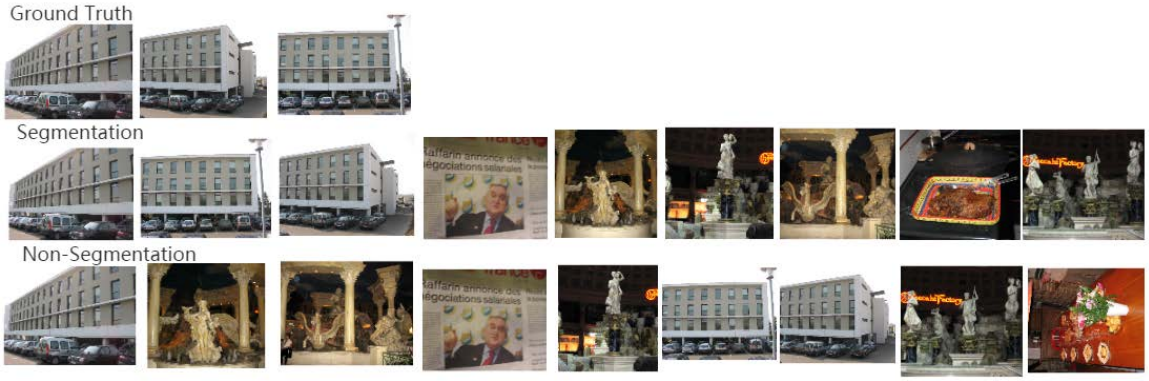


Fig. 6. Retrieval results of a building image. The first image of each row is the query image. The first row is the ground truth. The second row is the retrieval results obtained by multi-threshold segmentation and the third row is the retrieval results obtained by non-segmentation. The AP of multi-threshold segmentation is 1 and non-segmentation is 0.2667.

Moreover, we compare our experimental results with the methods of VLAD+SIFT[4] ,

VLAD+RootSIFT[4]， BoVW+SIFT[5] and Fisher+SIFT[5]. The comparison results are shown in the Table 1, where D is the length of vectors obtained by the quantization of local features.

Tab. 1 Comparison results of our method and methods proposed in [4] and [5] on the Holidays database.

Method	D	mAP
VLAD + SIFT [4]	8192	0.561
VLAD + RootSIFT [4]	8192	0.589
BoVW + SIFT [5]	20000	0.437
Fisher + SIFT [5]	8192	0.595
VLAD + SURF	4096	0.578
VLAD+ SURF+GMM	4096	**0.613**
VLAD+SURF+GMM+Seg	4096	**0.623**

From Tab. 1, it can be seen that the mAP of VLAD+SURF+GMM+Seg is the highest. This mAP is 18.6% higher than the mAP of BoVW+SIFT, 6.2% higher than the VLAD+SIFT, 3.4% higher than the VLAD+RootSIFT and 2.8% higher than the Fisher+SIFT. Meanwhile, it is 4.5% higher than the mAP of VLAD+SURF in our experiment.

5. Conclusions

In the image retrieval, in order to using the object outline, multi-threshold segmentation is proposed. Then the SURF features are extracted from the segmented image. Considering the spatial locations of SURF features, an generic object regions matching based retrieval technique by using GMM clustering algorithm with SURF features location is proposed. Finally, VLAD model is used to represent each cluster of SURF features and the similar scores among query image and images in the database are computed. Compared with the methods mentioned in [4], [5] and the method of non-segmentation and GMM clustering, experimental results show that our proposed VLAD+SURF+GMM+ Seg method achieves a better retrieval result.

Fund Project

Training program for excellent young teachers in Colleges and universities of Guangdong province (project number: Yq2013186); Guangzhou Education Science "12th Five-Year" planning project (project number: 12A160)

References

[1] Zheng L, Wang S, Tian L, et al. Query-adaptive late fusion for image search and person re-identification[C]Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1741-1750.

[2] Jegou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]. IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2010:3304-3311.

[3] Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris I, et al. An Empirical Study on the Combination of SURF Features with VLAD Vectors for Image Search[J]. Image Analysis for Multimedia Interactive Services International Workshop on, 2012:1-4.

[4] Spyromitros-Xioufis E, Papadopoulos S, Kompatsiaris I Y, et al. A Comprehensive Study Over VLAD and Product Quantization in Large-Scale Image Retrieval[J]. IEEE Transactions on Multimedia, 2014, 16(6):1713-1728.

[5] Hervé J, Florent P, Matthijs D, et al. Aggregating Local Image Descriptors into Compact Codes[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2012, 34(9):1704-1716.

[6] Peng X, Wang L, Qiao Y, et al. Boosting vlad with supervised dictionary learning and high-order

statistics[M]Computer Vision–ECCV 2014. Springer International Publishing, 2014: 660-674.

[7] Arandjelovic R, Zisserman A. All about VLAD[C]Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 1578-1585.

[8] Liu Z, Li H, Zhou W, et al. Uniting Keypoints: Local Visual Information Fusion for Large Scale Image Search[J]. IEEE Transactions on Multimedia, 2015, 17(4):1-1.

[9] Wang J, Shou L, Li X, Chen G. Bundling features with multiple segmentations for object-based image retrieval[J]. Journal of Zhejiang University(Engineering Science), 2011(2):259-266.

[10] Jegou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search[M]. Computer Vision–ECCV 2008. Springer Berlin Heidelberg, 2008: 304-317.