Image Retrieval Based on Spatial Organization of Patches

Yan Yu[1,2]

[1] *Hubei Province Key Laboratory of Systems Science in Metallurgical Process, Wuhan University of Science and Technology, China*

[2] *School of Science, Wuhan University of Science and Technology, China*

Abstract

This paper presents a novel approach for image retrieval by utilizing the spatial organization of image patches. Firstly, each image is presented as a set of patches and each patch is descripted as a bag-of-visual-words (BoW) feature. Secondly, the matrix of matching costs between all pairs of patches that come from the two images respectively is computed by utilizing both the appearance feature and spatial information. Finally, the distance between two images is obtained by minimize the total cost of matching using Hungarian algorithm. The resulting distance is applied to image retrieval. The experimental results indicate that the proposed method outperforms the traditional BoW method in image retrieval scenario.

Keywords Image retrieval; Bag-of-visual-words; Image patch; Image distance measure; Hungarian algorithm.

1 Introduction

The most popular approaches for content-based image retrieval nowadays are based on the model of bag-of-visual-words (BoW). They represent image as an orderless collection of local features and demonstrate impressive performance because of their compactness and invariance. The traditional approach based on BoW contains the following steps. Firstly, local features are extracted from a training set of images by interesting point sampling or dense sampling. Secondly, all the feature vectors are quantized to construct a dictionary of visual words by k-means clustering. Each cluster tends to contain image patches with similar appearance and its centroid represents a visual word. Thirdly, each image is descripted as a histogram indicating the probability distribution of visual words by hard assignment or soft assignment. Lastly, the distances between histograms are used for image retrieval.

The major limitation of such approaches is that their image representation ignores encoding the spatial information or geometrical relationships among visual words, which has been discovered very important for understanding image content [1]. For solving the problem, some existing approaches leave the spatial verification as a post-processing step [2-3], however, they are only suitable for partial-duplicate image retrieval, as for semantic retrieval they are limited. Spatial Pyramid Matching [4] is successful for representing the spatial information of visual words, which subsquently turn into a general framework combined with various feature encoding methods [5-6] for image classification. For analogous approaches like SPM, we uniformly call them patch-based methods. In methods like these, image is divided into several non-overlapping rectangular patches, and the feature histogram of each patches is computed and combined as the image feature. Consequently, the distance between images is the weighted sum of distances between corresponding patches locating in the same spatial positions. But we know, the foreground objects of similar images maybe lie in different spatial locations, poses and orientations, which make the distance measure is not accurate in many situations.

In this work, we present a more flexible patch-based distance measure method integrating with BoW, aiming at increasing the precision of the traditional BoW for semantic retrieval. Each patch of the first image is assigned to exactly one patch of the second one considering both the BoW feature and the spatial location of patches in the image, and the distance measure problem of images is transformed into square assignment problem. Our idea is similar to [7] which works in pixel level and only makes use of image color feature. However, our method works in patch level and utilize image local features which have stronger descriptive power.

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the image feature representation as a set of patches. Section 4 details the proposed image distance measure. Section 5 gives the experimental results on image retrieval application. Section 6 concludes the paper.

2 Related Work

There are several recent works in the content-based image retrieval (CBIR) area which present interesting advances for encoding spatial information of visual words. In the early days, researchers present color correlogram [8] to encode the spatial arrangement of colors. Now the study object is turned into local feature rather than pixel color. The most popular approach to encode the spatial information of visual words is Spatial Pyramid Matching, which involves repeatedly subdividing the image and computing histograms of local features at increasingly fine resolutions. It is succesful for image classification, however, it is not robust for some image transformations such as roation.

Many of the existing approaches leave the spatial verification as a post-processing step. They find the matching points between images, then compute the spatial representation which is used to identify and remove the false matches, compute the similarity between images by using the number of matched features and a penalty term from spatial verification. RANSAC exploits full geometric verification [9-10], which can greatly improve retrieval performance, but it is computationally expensive. Another recent approach explicitly encodes the spatial relationship among each pair of matching points by utilizing binary spatial maps [2]. This method is efficient and effective, but its representation is very specific for partial-duplicate search and it is not suitable for the semantic-search application.

Another recent approah learns the postions of visual words occurrences in an image dataset, called lp-norm pooling [11]. The method first transforms all the images into the same resolution, and use a dense sampling which generates m points in each image. Each visual word w has a mD vector, in which the m-th value representis the m-th point activate the visual word w. However, the method is designed for image classification and it represent the absolute postions of visual words, which is not robust to translation or rotation of the object in the image.

A recent image content description is proposed for describing the spatial layout of visual words, called $\Delta-$TSR [12]. Triangle is used to describe the relationship among arbitrary three interesting points, which is represented as a 7D signature composed of labels of correspoind viusal words, angles of vertices, relative orientations and scales of points. After the strategy of triangle selection is applied, the similarity between images is defined as the ratio of the sum of similairity between the most similar triangles and the cardinality of the set of triangles. This method explicitly encodes the spatial relationship among visual words,but it only works with hard assignment. More recently, graph-based approaches for encoding spatial information are proposed. The nested multi-layered local graphs [13] is proposed by building upon sets of SURF points with Delaunay triangulation. A BoW framework is applied on these graphs, giving birth to a Bag-of-Graph-Words representation, visual dictionary is built for each layer of graphs and L1 distance function is used for image retrieval. This method explicitly encodes the spatial relationship of image points, however it has higher computational cost and only work with hard assignment.

In this work, the content representation and similarity measure method of images can work with both hard assignment and soft assignment, furthermoe, it is suitable for semantic-retrieval.

3 Proposed Method

3.1 Image as a Set of Patches

For a given image, the set of patches can be represented as

$$\{f_i\}_{i=1,\ldots,n}, f_i \in R^m \tag{1}$$

where f_i is the feature vector in the i-th image patch, n is the total number of image patches and m is the length of the feature vector. The pixel sizes of images in our dataset are various. In order to realize the assignment relation between

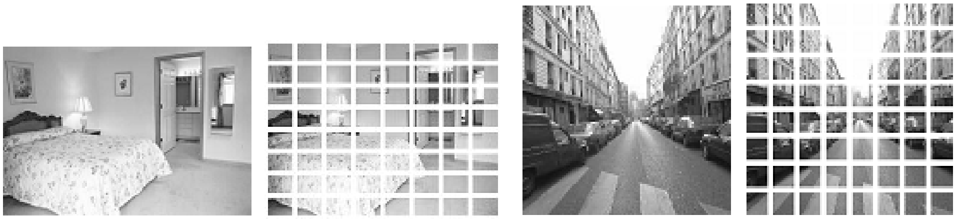


Fig.1 Original Images and their sets of patches.

patches of two images, all the images are divided into patches of the same amount, rather than limited to the same size of each patch. For example, Figure1 shows that two images are divided into 8 rows and 8 columns respectively. Although they have different aspect ratio, they all have a set of 64 patches in total.

In our method, we encode the appearance feature for each patch rather than the global image. The feature vector of each patch is extracted via BoW model according to the following steps. Firstly, the images were all preprocessed into gray scale and local features are extracted from images in the training set by dense sampling. In this work, SIFT descriptor [14] is adopted as local feature descriptor, because it has averagely the best performance among local feature descriptors [15]. Dense sampling is realized by a sliding window in the size of 16*16 pixels, which sliding from top to down and left to right in an image by a step of 8 pixels. These regions are overlapping for obtaining enough abundant local features. Secondly, all the feature vectors are quantized to construct a visual dictionary by standard k-means clustering algorithm [16]. Lastly, each patch of each image is descripted as a histogram indicating the probability distribution of visual words by hard assignment. Many more effective dictionary construction method such as hierarchical clustering, and more advanced coding schemes, such as soft assignment [17], sparse coding [5], fisher coding [18] et.al., all can be utilized in our method. However, they are not key for evaluating our method, so we adopt the easiest clustering algorithm and coding method.

Consequently, the signature for representing the content of an image is composed of n patch features. The i-th patch feature is represented as $f_i = \{A_i, P_i\}$, where A_i codes the appearance feature in the patch, whose dimension is the same as the size of visual dictionary. $P_i = (x_i, y_i)$ is composed of the horizontal and vertical spatial positions of the patch in the 2D plane space. Note that x_i and y_i both are normalized to [0,1] for each patch.

3.2 Distance between Images

The method to measure the distance between images is critical for retrieval precision. In this work, a flexible patch-based distance measure method is proposed. Given the i-th patch of the first image u and the j-th patch of the second image

v, the distance between patches is given by

$$d(f_i^u, f_j^v) = w_A \cdot \frac{d(A_i^u, A_j^v)}{\delta_A} + w_p \cdot \frac{d(P_i^u, P_j^v)}{\delta_P} \tag{2}$$

where $d(A_i^u, A_j^v)$ is the distance between appearance features which can be computed by Chi-square distance as (3)

$$d(A_i^u, A_j^v) = \sum_{k=1}^{D} \frac{(A_i^u(k) - A_j^v(k))^2}{A_i^u(k) + A_j^v(k)} \tag{3}$$

where D is the length of the appearance feature vector,$A_i^u(k)$ represent the k-th component value of the vector. And $d(P_i^u, P_j^v)$ is the distance between the spatial positions of the two patches, then we have

$$d(P_i^u, P_j^v) = \sqrt{(x_i^u - x_j^v)^2 + (y_i^u - y_j^v)} \tag{4}$$

Parameters w_A and w_p are weight coefficients, which balance appearance and s-patial contributions respectively and the sum of them always is 1. The proportion of them has a great influence on the performance of the method, which will be discussed in the next section. Parameters δ_A and δ_P are chosen according to the appearance and spatial dynamics. Typical values used in this paper are $\delta_A = 2$ and $\delta_P = \sqrt{2}$.

The distance $d(f_i^u, f_j^v)$ between two patches can be seen as matching cost, which is denoted as C_{ij}. Given the set of costs between all pairs of patches in the first image and in the second image, the minimum of the total matching cost can be seen the distance between two images, subject to the constraint that the matching be one-to one. Then the distance between image u and image v is given by

$$d(u, v) = \underset{\pi}{min} \sum_{i=1}^{n} d(f_i^u, f_{\pi(i)}^v) \tag{6}$$

where π is a permutation of $\{1, ..., n\}$ to guarantee the constraint. The input to the assignment problem is a square cost matrix with entries C_{ij}. The result is a permutation π such that total matching cost is minimized. This is an instance of the square assignment problem, which can be solved using the Hungarian algorithm [19]. Given a cost matrix to find an optimal assignment using the Hungarian algorithm is presented as follow:

1) Subtract the smallest entry in each row from all the entries of its row.

2) Subtract the smallest entry in each column from all the entries of its column.

3) Draw lines through appropriate rows and columns so that all the zero entries of the cost matrix are covered and the minimum number of such lines is used.

4) Test for optimality:

(i) If the minimum number of covering lines is n, an optimal assignment of zeroes is possible and we are finished.

(ii) If the minimum number of covering lines is less than n, an optimal assignment of zeroes is not yet possible.In that case, proceed to 5).

5) Determine the smallest entry not covered by any line. Subtract this entry from each uncovered row, and then add it to each covered column. Return to 3).

Compared with the traditional patch based distance measure methods, our method can deal with the position change of the foreground objects in the image automatically. We should note that the distance that we proposed is not a rigid mathematical distance. However, it dose not matter because the distance can be used to represent the divergence between images and it can be utilized in ranking retrieval results.

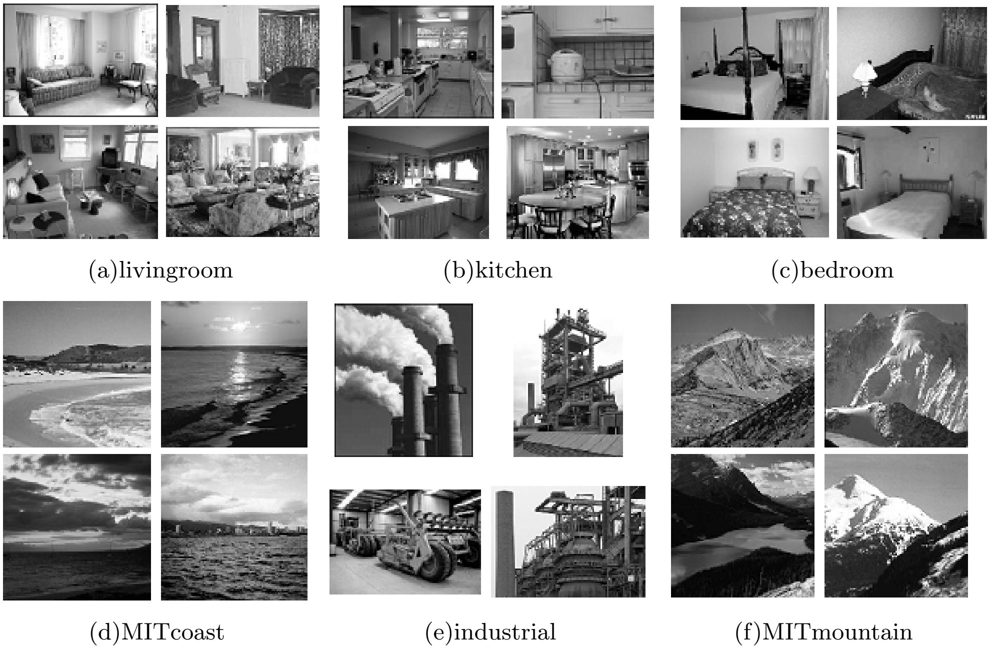
4 Experiments

In this section, we evaluate the performance of our proposed method in image retrieval scenario.

4.1 Experimental Setups

Experiments are performed on a personal laptop in Windows 8.1, with quad Intel Core i7-4500U running at 1.8GHz. Most of our programs are implemented in Matlab except dense sift and Hungarian algorithm for which we use open source code in C. Our dataset is composed of fifteen scene categories, including bedroom, kitchen, livingroom, store, coast, forest, et.al.. Each category contains 200 to 400 images, and average image size is 300×250 pixels. The dataset is popular in image classification and retrieval scenario. Fig.2 show examples of the dataset. We use 100 images per class randomly selected for generating visual words dictionary and the rest for testing. We use dictionary size 200 and hard assignment for testing our method. In practical application, the soft assignment can be used in our framework for improving retrieval performance.

The experimental results are presented in terms of mean average precision (MAP) and precision for the top N retrieved images (P@N). Although MAP is a very popular measure to assess the effectiveness of CBIR methods, it dose not reflect the ranking quality in the first positions. The practical Web users usually pay more attentions to the good set of top 10 or 20 retrieved images, so we are more interested in good P@N values than MAP values. Precision measures the retrieval accuracy. It is defined as the ratio between the number of relevant images retrieved and the total number of images retrieved. Relevant images are in the same category with the query example.



(a)livingroom (b)kitchen (c)bedroom

(d)MITcoast (e)industrial (f)MITmountain

Fig.2 Sample images from the dataset, highlighting 6 categories (livingroom, kitchen, bedroom, MITcoast, industrial and MITmountain).

Table 1 Comparison between two kinds of partition patterns based on 10 queries.

Partition Pattern	P@5(%)	P@10(%)	Average Retrieval Time(ms)
4×4	62.00	63.00	23.64
8×8	78.00	69.00	185.77

Table 2 Comparison between two kinds of weighting ratios based on 100 queries.

Weighting Ratio	P@5(%)	P@10(%)	P@15(%)	P@20(%)
1:1	67.00	59.10	55.47	52.50
2:1	67.40	59.90	55.80	53.65
4:1	**68.60**	**60.00**	**57.00**	**55.15**
8:1	66.40	59.80	56.33	54.60

4.2 Influence about Size of Patches

In this section, the impact of the patch size is illustrated. We use n to present the total number of patches in each image. We divide each images into several rows and several columns, including two partition patterns which are 4×4 and 8×8 in our experiments. To compare the two pattern patterns, 10 images are randomly selected from the dataset and each of them is iteratively selected as a query example, so the total number of queries in each pattern is 10. It is obvious that the size of each patch is decreasing when the total number of the patches is increasing. The details can be seen in table 1. It can be seen that increasing the total number of the patches has a improvent on the retrieval results, however, the computation become more time-consuming. In consideration of the efficiency, in the rest of the paper patches of n=16 are always used.

4.3 Appearance-spatial weighting

The weighting parameters w_A and w_P balance appearance feature and spatial contributions respectively. In this section, the impact of ratio of the weighting parameters is illustrated. To compare the influence of weighting parameters, 100 images are randomly selected from the dataset and each of them is iteratively selected as a query example, so the total number of queries in each experiment is 100.

If we wish each patch to be matched with the patch which is close to it in spatial position, w_P must be set at a high value. Otherwise, if we wish each patch in the first image can be matches with any patch in the second image, w_P must be set at a low value. We test the weighting parameters including four kinds of ratio patterns which are 1:1, 2:1, 4:1 and 8:1. The details can be seen in Table2. It can be seen that increasing the proportion of w_A has a slight improvement on results. But if the ratio keep going up and reach 8:1, the precision begin to decline. In the rest of the paper the ratio 4:1 are always used.

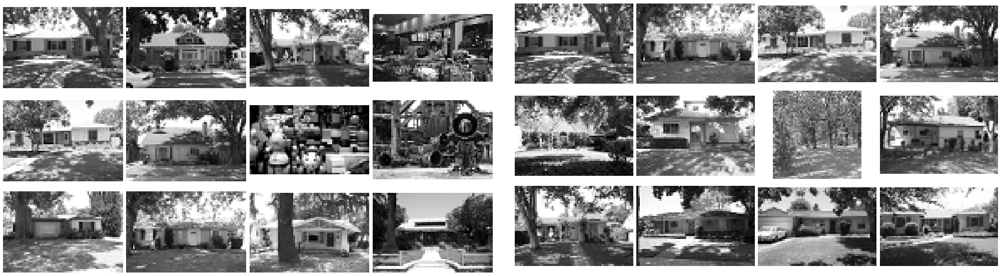
4.4 Comparison with the traditional method

The traditional BoW method is also implemented for comparison. To compare our proposed method with it fairly, we iteratively select each image in the dataset as query example, so the total number of queries is 2985 in each query experiment. In Table 3, it can be observed that the precision decreases while the number of retrieved images increases. In our proposed method, the P@N is always better than the traditional BoW method. Therefore, our proposed method is more suitable for image retrieval.

Table 3 Comparison with the traditional Bow based on 2985 queries.

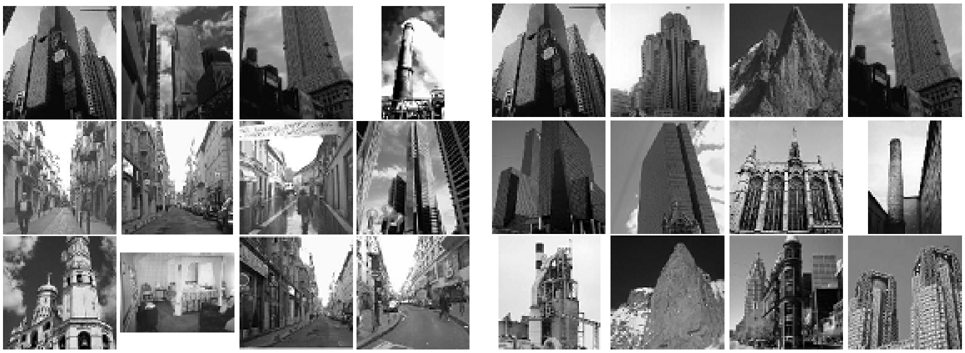
Methods	P@5(%)	P@10(%)	P@15(%)	P@20(%)	MAP(%)
Traditional BoW	57.45	50.67	47.92	46.08	29.28
our method	**66.11**	**58.87**	**55.30**	**53.15**	**32.54**

In the end, we give some retrieval examples of the query image random selected from the dataset via the two kinds of methods in Fig. 3, Fig. 4 and Fig.5. The query images come from “CALsuburb”, “MITtallbuilding”and “MITmountain” category respectively. In each Fig., the query image lies on the top left corner. Retrieved top 12 images are arranged from left to right and top to bottom according to the increasing distances between the retrieved images and the query example. It can be seen that the vast majority of the retrieved images are in the same category with the query image in our proposed method.



(a)traditional BoW (b)our method

Fig.3 Retrieval example for a query image taken from the "CALsuburb" category.

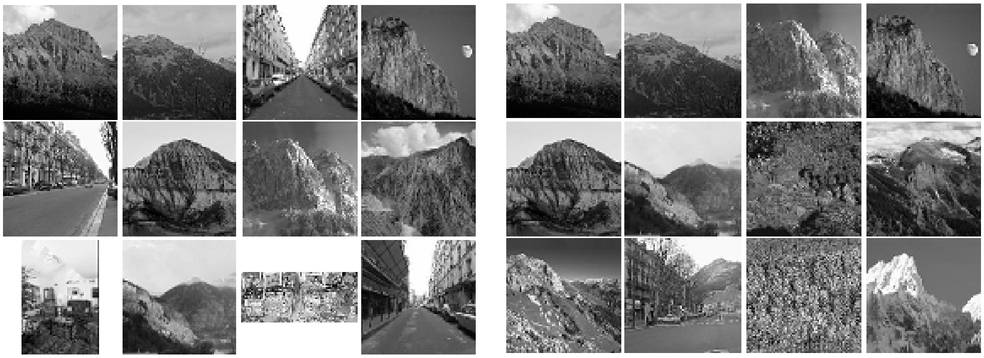


(a)traditional BoW (b)our method

Fig.4 Retrieval example for a query image taken from the "MITtallbuilding" category.

5 Conclusions

In this paper we present a flexible patch-based distance measure method for images, which utilizes the abundant local feature of images and the spatial information of them. The proposed method is tested in image retrieval application. Experiments show better retrieval results than the traditional BoW. We use hard assignment for testing our method. In the future, we will test the retrieval performance of integrating the soft assignment with our method. In addition, we only utilize the absolute position of visual words in our method. The relative relationship of visual words is more robust to image transformation, which will be our next research work.



(a)traditional BoW

(b)our method

Fig.5 Retrieval example for a query image taken from the "MITmountain" category.

Acknowledgments

This work was supported by Hubei Province Key Laboratory of Systems Science in Metallurgical Process (Wuhan University of Science and Technology):Y201406, and the Natural Science Foundation of Hubei Province: 22013CFA131.

References

[1] J. Philbin, O. Chum, M. Isard, et al. (2007), "Object retrieval with large vocabularies and fast spatial matching", *IEEE Conference on Computer Vision and Pattern Recognition.*

[2] W. Zhou, Y. Lu, H. Li, et al.(2010), "Spatial coding for large scale partial-duplicate web image search", *18th ACM International Conference on Multimedia*, pp.511-520.

[3] H. Jégou, M. Douze and C. Schmid.(2010), "Improving bag-of-features for large scale image search" *International Journal of Computer Vision*,Vol.87,pp.316-336.

[4] S. Lazebnik, C. Schmid and J. Ponce. (2006), "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2169-2178.

[5] J. Yang, K. Yu, Y. Gong et al. (2009), "Linear spatial pyramid matching using sparse coding for image classification", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.1794-1801.

[6] C. C. Yan, L. Li, Z. Wang, et al. (2014), "Fusing multi-cues description for partial-duplicate image retrieval" *Journal of Visual Communication and Image Representation*, Vol.25, No.6, pp. 1726-1731.

[7] T. Hurtut, Y. Gousseau and F. Schmitt. (2008), "Adaptive image retrieval based on the spatial organization of colors", *Computer Vision and Image Understanding*, Vol.112, pp.101-113.

[8] J. Huang, S. R. Kumar, M. Mitra, et al. (1997), "Image indexing using color correlograms", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.762-768.

[9] H. E. J. Egou, M. Douze and C. Schmid. (2008), "Hamming embedding and weak geometric consistency for large scale image search", *10th European Conference on Computer Vision*, pp.304-317.

[10] O. Chum, M. Perdoch and J. Matas. (2009), "Geometric min-hashing: finding a (thick) needle in a haystack", *IEEE Conference on Computer Vision and Pattern Recognition*, pp.17-24.

[11] J. Feng, B. Ni, Q. Tian, et al. (2011), "Geometric lp-norm feature pooling for image classification",*IEEE Conference on Computer Vision and Pattern Recognition*, pp.2609-2704.

[12] N. V. Hoàng, V. Gouet-Brunet, M. Rukoz, et al. (2010), " Embedding spatial information into image content description for scene retrieval ", *Pattern Recognition*, Vol.43, pp.3013-3024.

[13] S. Karaman, J. Benois-Pineau, R. Megret, et al. (2012), "Multi-layer local graph words for object recognition ", *18th International Conference on Multimedia Modeling*, pp.29-39.

[14] D. G. Lowe. (1999), "Object recognition from local scale-invariant features", *7th IEEE International Conference on Computer Vision*, pp.1150-1157.

[15] K. Mikolajczyk and C. Schmid. (2005), "A performance evaluation of local descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27, pp.1615-1630.

[16] C. M. Bishop. (2006), *Pattern Recognition and Machine Learning*, Springer, pp.424-430.

[17] J. C. Van Gemert, J. Geusebroek, C. J. Veenman, et al. (2008), "Kernel codebooks for scene categorization ", *10th European Conference on Computer Vision*, Marseille, pp.696-709.

[18] F. Perronnin, J. S A Nchez and T. Mensink. (2010), “Improving the fisher kernel for large-scale image classification”, *11th European Conference on Computer Vision*, pp.143-156.

[19] C. H. Papadimitriou and K. Stieglitz. (1982), “Combinatorial Optimization: Algorithms and Complexity”, Prentice Hall.

Corresponding author
Yan Yu can be contacted at: yuyan_wust@163.com