AN EFFICIENT HIERARCHICAL IMAGE RETRIEVAL METHOD FOR LARGE SET OF IMAGES USING LEARNING-BASED GLOBAL AND LOCAL IMAGE FEATURES

Zhiwei Wang, Zongqian Zhan*, Gaofeng Zhou, Xin Wang

School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China
zwwangsgg@whu.edu.cn, (zqzhan, xwang)@sgg.whu.edu.cn, gfzhou9608@163.com

Commission II, WG II/6

KEY WORDS: Learning-based Features, Local Image Features, Global Image features, Hierarchical Image retrieval

ABSTRACT:

Image retrieval is one of the supporting technologies for (near) real-time photogrammetry and loop closure detection in visual SLAM, the conventional retrieval strategy is to firstly obtain the image features of the query image and database images, and search for the resulted images based on nearest features retrieval. However, the image retrieval method based on traditional hand-crafted features (SIFT, SURF, GIST) are hard to guarantee both the efficiency of time and precision in practical applications. Nowadays, learning-based features have shown superior performance in ample computer vision tasks. Thus, this paper investigates several popular learning-based global features (ResNet101, VGG16+NetVLAD, Yolov3+VGG16+NetVLAD) and local features (SuperPoint), to take care of both time efficiency and precision, we present hierarchical image retrieval solutions that combines these two kinds of features, in which global feature is for accelerating searching speed and local feature is for precision. Specifically, three sets of hierarchical retrieval solutions are designed by various combinations of learning-based global feature and local feature. Their precision and time efficiency are compared on different public benchmarks (one contains more than 10,000 images), the experimental results show that among the proposed solutions, VGG16+NetVLAD+SuperPoint has the best performance in efficiency, but the precision is slightly lower than the solution preprocessed with Yolov3.

1. INTRODUCTION

Image retrieval is to find the most similar images (typically, regarding the image content) for one specific target image. Over the last years, due to the developments of sensors and computer machines, images are quite easy to access such as images from some practical surveying tasks or the crowdsource images (shared via social media applications, e.g., Flickr, Instagram etc.), and it is much less costly to process these images even with a common consumer computer (Wang et al., 2017). Therefore, it is an ongoing challenging topic to determine similar images for large set of images with high level of time and accuracy efficiency, it is very crucial in many fields, e.g., the detection of mutual image overlapping relationship for image orientation in photogrammetric dataset without prior knowledge (such as GPS). For (near) real-time photogrammetry, a fast and accurate image retrieval solution plays an important role for quickly and correctly determining the overlapping frames of the current frame (such as dealing with arbitrary flight path, as Fig. 1 shows).

Content-based image retrieval (CBIR) is one of the relevant technologies, which needs to first obtain the image features of the database images and the query image, then, generate feature descriptors that characterize the input images, and finally execute the query operation based on the feature description index. Since the 1990s, content-based image retrieval technology has begun to be researched, but, the corresponding methods often adopted some traditional features, such as GIST (Torralba et al., 2003), HOG (Dalal and Triggs, 2005),SIFT(Lowe, 2004.), HARRIS (Harris and Stephens, 1988.),ORB(Rublee et al., 2011) and so on, these traditional features are in general difficult to meet the requirements of practical applications in terms of time efficiency and accuracy, where GIST needs lots of computer memory for retrieving images, SIFT is too computationally expensive regarding the procedure of extraction and matching, HARRIS is not invariant to scale changes.

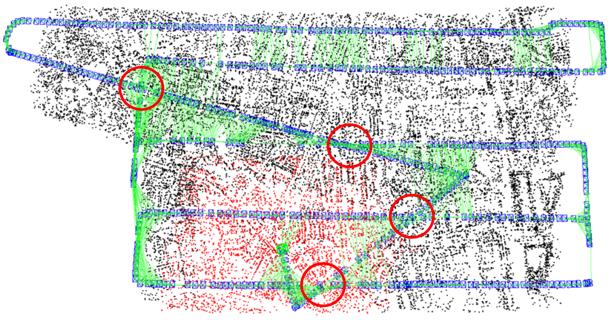


Figure 1. Arbitrary UAV flight for real-time photogrammetry. The red circles denote the locations where image retrieval should work to find the loop closure.

Since the last decade, the methods based on deep-learning, especially convolutional neural networks (CNN), have seen a tremendous success in many fields including photogrammetry and remote sensing, and can be used to as a candidate technique to advantage also for the image retrieval method. In general, CNN can be regarded as a series of nonlinear functions in essence. A network includes convolution, pooling, nonlinear activation function, which is a hierarchical structure. From the bottom layer to the top layer, an input image will undergo convolution operations with filters of different sizes. Thus, CNN-based models can yield more compact global image features ("compact" denote a low-dimension descriptor vector with strong

[*] Corresponding author.

https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-649-2022 | © Author(s) 2022. CC BY 4.0 License.

discriminability), the output of the corresponding deep layer is in principle estimated by a larger receptive field which constitute the "global" feature in this paper and can contribute to a more informatic description on the image level, in addition, the whole image is represented by just one multi-dimensional vector as the related descriptor. In consequence, the global feature can improve the time efficiency of the retrieval procedure, whereas, the accuracy efficiency might be dropped off as some sensitive local information is discarded when running the recursive convolution and pooling operations.

The most reliable and accurate strategy is the widely-applied "local" feature matching method (local features often appear on the salient points on the image and described by the content located around small patches), for example, the well-known SIFT features are firstly extracted on every images, and pairwise image matching is performed on all features from every images, the mathematical logic of the SIFT features is rigorous and explainable, thus, typically, it can work well on most image datasets. However, the underlying limitation is the time efficiency, the main reason is that one image typically can generated hundreds or thousands of local features and each feature is described by a high dimensional vector, which can be very problematic when dealing with ten or hundred thousands of images.

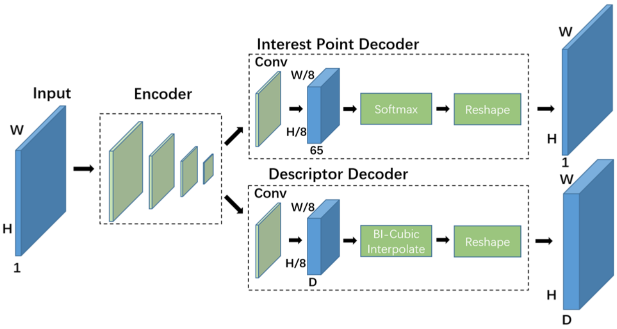


Figure 2. SuperPoint Network Architecture. The feature detector and the descriptor sub-network share a single forward encode, but decoder contains different structures and different model parameters are learned according to different tasks.

In this paper, we comprehensively study the advantages of learning-based global feature and local feature and propose hierarchical image retrieval solutions, the main idea is: the leaning-based compact global features are firstly extracted and the candidate images can be figured out by calculating the corresponding Euclidean distance between global features in a very fast way, to further refine the retrieved candidate images, the learning-based local features are employed to rearrange the previous obtained result (i.e., global feature matching). Based on the above idea, three sets of hierarchical image retrieval solutions are designed, namely, ResNet101 and SuperPoint, VGG16, NetVLAD and SuperPoint, and YOLO, VGG16, NetVLAD and SuperPoint. The above three solutios will be introduced in detail later. The contributions of this work are as follows:

1. This paper introduces three hierarchical image retrieval solutions that integrate multiple learning-based global features and one learning-based local feature.
2. We extensively tested the three solutions on various benchmark datasets (one contains more than 10000 images), the results show that both solutions B and C performs in real-time on a modern GPU and can be readily integrated into different SfM or SLAM systems.

2. RELATED WORK

In general, the current mainstream image retrieval technologies are based on global features represented by relevant CNN architectures and local features represented by SIFT, SURF (Bay et al., 2006) etc. A more comprehensive review of image retrieval topics can be referred to Zheng et al. (2017) and Hartmann et al. (2016). In this section, we will give detailed explanations of these related works that are the basics of our hierarchical image retrieval solutions.

2.1 SuperPoint and relevant networks

SuperPoint was proposed by the group of MagicLeap (Detone et al., 2018) and has been widely used in autonomous driving and other applications as a replacement for SIFT. They designed a self-supervised network framework that can extract the location and descriptor of feature points at the same time, in which the positions of extracted features are with pixel-level precision in the original image. In addition, a Homographic Adaptation strategy was proposed to enhance the feature point recheck rate and cross-domain practicability, where cross-domain refers to the generalization ability from synthetic to real. The SuperPoint network structure is shown in Fig. 2, it consists of four parts: "Shared Encoder", "Interest Point Decoder", "Descriptor Decoder", and "Error Construction", from this figure it can be seen that the feature point detector and the descriptor sub-network share a single forward encoder, but different structures are used in the decoder, and different network parameters are learned according to different tasks. This end-to-end architecture is different from other networks - LIFT (Yi et al., 2016), UCN (Choy et al., 2016), which train feature detection and feature descriptor in two separated networks continuously.

2.2 ResNet101, VGG16 and NetVLAD

ResNet was proposed by He et al. (2016). Compared with AlexNet (Krizhevsky et al., 2012) including five convolution layers and other classical convolutional neural networks (GoogLeNet (Szegedy et al., 2015), LeNet (Lecun et al., 1998)), ResNet introduced the structure of Residual blocks (as shown in Fig. 3), which makes it possible to train hundreds or even thousands layers, while avoid the problem of gradient vanishing problem during the training epoch increases. Thanks to its superior performance on various tasks including image classification, target detection and face recognition and etc., many computer vision applications' record had been improved, it become one of the most popular network architecture. The structure of this network is shown in Tab. 1.

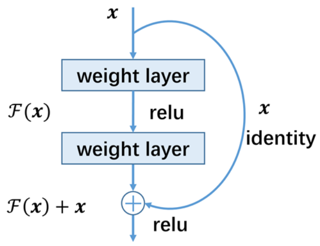


Figure 3. ResNet residual module structure.

VGGNet (Simonyan and Zisserman, 2014) is a deep convolutional neural network, who main idea is exploring the relationship between the depth of convolutional neural network and the resulted performance. Via recursively stacking convolutional kernels of size 3×3 and maximum pooling layers of size 2×2, VGGNet successfully builds 16 to 19 layers of deep convolutional neural network. Compared with the previous stateof-the-art network architectures (Krizhevsky et al., 2012), VGGNet was demonstrated to be of higher generalization ability and obtain advanced performance on various tasks and datasets. Up to date, VGGNet often appears in some missions that need to extract feature images.

Layer name	101-layer
Conv1	7×7，64，stride 2
Conv2_x	3×3 max pool，stride 2 1 x 1, 64 [3 x 3, 64] × 3 1 x 1, 256
Conv3_x	1 x 1, 128 [3 x 3, 128] × 4 1 x 1, 512
Conv4_x	1 x 1, 256 [3 x 3, 256] × 23 1 x 1, 1024
Conv5_x	1 x 1, 512 [3 x 3, 512] × 3 1 x 1, 2048
	Average pool，1000-d fc，softmax

Table. 1 ResNet101 network structure

NetVLAD (Arandjelovic et al., 2017) was proposed to improve the method of Vector of Locally Aggregated Descriptors (VLAD, Jégou et al., 2010). First of all, VLAD is a feature description method similar to Bag of Features (BoF, Lazebnik et al., 2016), and its main purpose was to aggregate local features into one global feature. However, BoF usually take SIFT feature as input to describe images and is widely used in image retrieval. Compared with BoF, VLAD is able to cluster local features into global feature in a more general manner, which means the corresponding image can be accurately represented by the obtained global feature with higher discriminability, and dimension reduction is facilitated as well. However, VLAD method does not have differentiability and the corresponding training procedure - back propagation is not doable. To cope with the inherent limitation of VLAD, NetVLAD includes exactly the same structure of VLAD embedded in convolutional neural network, in which the convolutional neural network is connected as the basic feature extraction structure to realize end-to-end training. The network structure is shown in Fig. 4, it consists of two parts: the first part is selected from the last convolution layer of a convolutional neural network, and its output are the feature maps with size of $H * W * D$ (H, W are the height and width of the image, and D is the feature dimension); The last part is actually functioned as the pooling layer, which is based on VLAD.

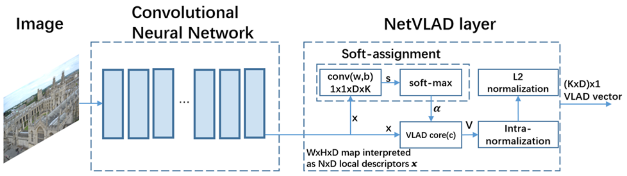


Figure 4. NetVLAD network structure.

3. METHODLOGY

The goal of this paper is to provide a time and accuracy efficient solution for image retrieval on more than ten thousand images. To achieve this goal, this paper investigates the advantages and disadvantages of global and local image features, several representative methods are discussed, and a hierarchical image retrieval solution is proposed, which combines the local features represented by SuperPoint (DeTone et al., 2018) with the global features that estimated from ResNet101 and VGG16+NetVLAD network.

The basic idea of " hierarchical" is to guarantee both the time and accuracy efficiency. Specifically, the efficient and compact global feature vectors based on CNN model are first considered, the distance between the global feature vectors denote the similarity degree of two corresponding images, thanks to the fact that global feature is just one feature per image with fixed size vector, this distance can be extremely fast computed, some candidate similar images which contain only a small subset of the original dataset can be roughly found with very high time efficiency. To take care of the accuracy, local features are extracted on the small candidate subset, the local feature matching is performed to rearrange the most similar images and refine the result, the superiority of the suggested hierarchical solution can avoid the consuming time on running exhaustive pairwise local feature matching, just the local features in some limited candidate images which are detected by using the global features are matched to further refine the retrieval results.

Based on the abovementioned statement, we introduce three sets of hierarchical image retrieval solutions which integrate with various global and one local features:

--Solution A: In this solution, the fully connected output layer of ResNet101 model after pre-trained on ImageNet (Deng et al., 2009) is selected as the global feature descriptor, and the global feature descriptor dimension is 1000. SuperPoint is selected as the local feature descriptor in the second step of the hierarchical solution. The overall working flow is shown in Fig. 5. First, the Top-N images are quickly determined based on the Euclidean distance between the global feature descriptor vectors extracted by the ResNet101, and then the candidate coarse retrieved images are rearranged based on the SuperPoint feature matching results;

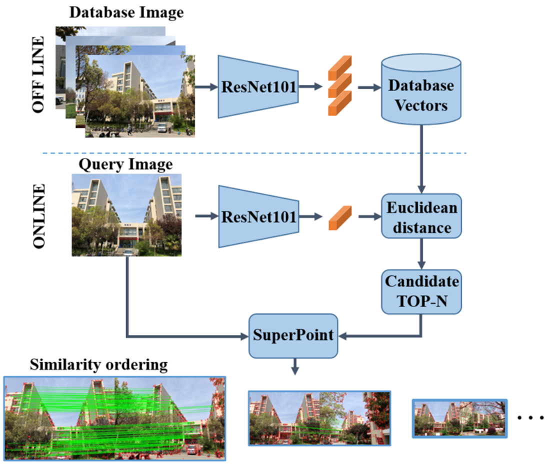


Figure 5. Solution A. Extracting global feature based on ResNet101 for rapid preliminary searching, and then rearrange candidate images based on the results of SuperPoint feature matching.

--Solution B: The pre-trained model of VGG16 and NetVLAD (based on the training data ImageNet) is selected to extract global feature descriptors. The initial descriptors dimension after aggregation by NetVLAD is 4096. SuperPoint was selected as a local feature descriptor, and the overall solution flow was shown in Figure 6. First, the global features are generated from the shallower VGG16 model and reduced by the NetVLAD network clustering for a more compact feature, the reduced global feature vector Euclidean distance is again computed to obtain the Top-N nearest images, and then rearranges the Top-N similar images based on the SuperPoint feature matching result, as shown in Fig. 6;

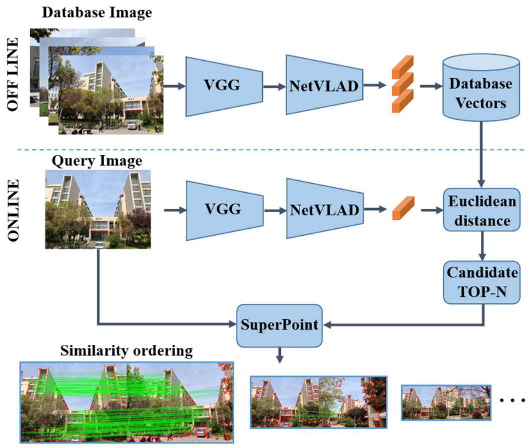


Figure 6. Solution B. Extracting global feature based on VGG16 and NetVLAD network for rapid preliminary searching, and then rearrange candidate images based on the results of SuperPoint feature matching.

--Solution C: This solution is basically identical with the previous one, the difference is that the YOLOv3 network is employed to remove the areas that are irrelevant to image retrieval task and fill the corresponding areas with blanks, and the following processing is just as the same as what solution B execute. The overall solution flow can be seen in Fig. 7.

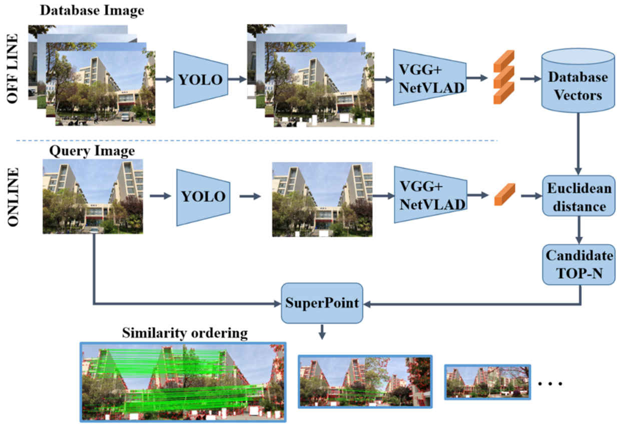


Figure 7. Solution C. After removing the interference area in the image, the subsequent retrieval process is carried out according to Solution B.

As Fig. 5,6,7 show, this work consists of offline and online processing module for these three hierarchical retrieval solutions. The offline mode extracts the global features for all database images, and builds the corresponding global feature database. In the online retrieval module, global features and local features are successively extracted from the query image. Then, the candidate TOP-N images are fast figured out based on the Euclidean distance among the global feature vectors, and then the obtained candidate images from previous step are rearranged and refined based on the matching results of the corresponding local features.

In solution C, before extracting features, it employs YOLOv3 model pre-trained on the COCO dataset for common target detection (people, tables and chairs, green vegetation decoration etc., this is motivated by the fact that some datasets, e.g., crowdsourced images, are often covered by pedestrians, tourists or moving objects etc., which are not useful for image retrieval) and then replace these target areas with blanks and insert the blanked images into the corresponding networks.

4. EXPERIMENTS

To evaluate the proposed hierarchical image retrieval solutions, a total of nine sets of comparison experiments are conducted on the benchmark Oxford Building (accessed on 2021.12.22) and a mixed dataset (accessed on 2021.12.22) of over 10000 images. In addition, the most popular dimensionality reduction algorithm PCA (Principal Component Analysis) is utilized to simplify the extracted global features from VGGNet+NetVLAD.

To demonstrate our method's effectiveness, the all reported experimental results include the searching time efficiency and searching precision. The next subsection 4.1 introduces the calculation of evaluation criterion, followed by the experiments of Oxford building in subsection 4.2, and the experiment of the mixed dataset (in subsection 4.3) closes this section.

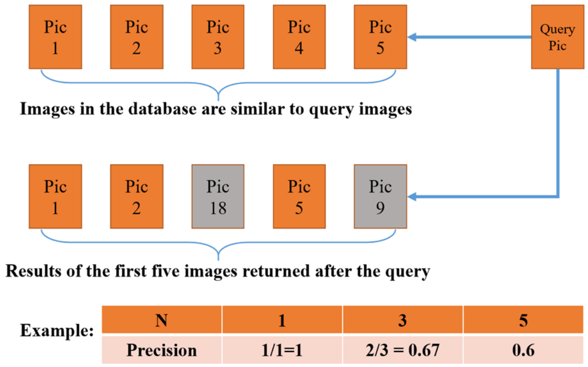
4.1 Evaluation criterion

In terms of the retrieval quality, the precision is considered, which indicates the proportion of correctly found images in the retrieval result. Basically, for a tradition classification problem, precision is computed by formula (1) given the confusion matrix shown in Tab. 2, where TP is the number of samples that are correctly predicted as positive, TN is the number of samples that are correctly predicted as negative, FN is the number of samples that are wrongly predicted as negative, and FP is the number of samples that are incorrectly predicted as positive.

Forecast Result	Label results	
	Positive example	Negative number
Positive example	TP	FP
Negative number	FN	TN

Table 2. Confusion matrix

$$Precision = \frac{TP}{TP + FP} \tag{1}$$


Pic 1
Pic 2
Pic 3
Pic 4
Pic 5
Query Pic
Images in the database are similar to query images
Pic 1
Pic 2
Pic 18
Pic 5
Pic 9
Results of the first five images returned after the query
Example:

N	1	3	5
Precision	1/1=1	2/3 = 0.67	0.6

Figure 8. Precision calculation.

As for our image retrieval issue, taking retrieved 5 images from one query image as an example, the process of calculating the precision of image retrieval is intuitively illustrated in Fig. 8. The symbol "Top1" means to the top 1 nearest image, the symbol "Top2" means to the top 2 nearest images and so on. Assuming

	4096			1024/1000			512		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
ResNet101-FC1000 + SuperPoint	/	/	/	0.7345	0.7173	0.6536	/	/	/
Vgg16+NetVLAD	0.9091	0.8667	0.8236	0.8727	0.8364	0.7856	0.8636	0.8209	0.7786
YOLOV3+VGG16+NetVLAD	0.9122	0.8705	0.8206	0.8716	0.8391	0.7819	0.8654	0.8273	0.7818
Vgg16 + NetVLAD + rerank-top5 by SuperPoint	0.9216	0.8946	**0.8512**	0.8913	0.8557	0.8013	0.8813	**0.8436**	**0.7896**
Vgg16 + NetVLAD + rerank-top50 by SuperPoint	0.8526	0.7985	0.7456	0.8299	0.7856	0.7253	0.8287	0.7826	0.7256
Vgg16 + NetVLAD + rerank-top100 by SuperPoint	0.8320	0.7756	0.7257	0.8156	0.7607	0.7062	0.8136	0.7616	0.7056
YOLOv3 + Vgg16 + NetVLAD + rerank-top5 by SuperPoint	**0.9227**	**0.8952**	0.8493	**0.8924**	**0.8569**	**0.8036**	**0.8822**	0.8407	0.7841
YOLOv3 + Vgg16 + NetVLAD + rerank-top50 by SuperPoint	0.8536	0.7997	0.7485	0.8306	0.7866	0.7241	0.8297	0.7846	0.7266
YOLOv3 + Vgg16 + NetVLAD + rerank-top100 by SuperPoint	0.8306	0.7766	0.7236	0.8188	0.7635	0.7093	0.8127	0.7630	0.7072

Table 3. Precision results of various hierarchical retrieval solutions on Oxford Buildings dataset. The precision of nine tests using global feature descriptors in dimension of 4096, 104/1000 and 512 were provided, and best results are highlighted in bold font.

that the number of retrieved "Top1" image that is identical with the provided ground truth is n_1, and the number of retrieved "Top3" image that is identical with the provided ground truth is n_3, the corresponding precision of "Top1" is $n_1/1$, and that of "Top3" and “Top5” are $n_3/3$, $n_5/5$, respectively. All the tested methods are assessed by the average precision values of "Top1", "Top3", and "Top5" estimated from all query images.

In terms of retrieval time efficiency, the consuming time (in millisecond) from feature acquisition to the final result for a single image is recorded, which mainly includes: input images preprocessing、global feature extraction、Retrieval based on global features、Candidate image local feature extraction、Rearrange based on local features.

4.2 Experiments on Oxford Building Dataset

4.2.1 Oxford Building Dataset

The Oxford Buildings dataset contains 5063 images collected from a public photo sharing website Flickr for specific Oxford landmarks. This dataset has been manually labeled for a total of 11 POIS (Point of Interests), which is abundant of multiple outdoor buildings., and 5 images are selected as query images to be retrieved under each POI, i.e., 55 query images in total, and the ground truth similar images are provided as well for evaluation.

4.2.2 Experimental Results and Analysis of Hierarchical Retrieval

To evaluate these three image retrieval solutions suggested in this paper, as Tab. 3 lists, a total of 9 sets of comparison experiments were conducted on the Oxford Buildings dataset, which are as follows: (1)ResNet101-FC1000 and rerank-top5 by SuperPoint; (2)VGG16 + NetVLAD; (3)YOLOv3 + VGG16 + NetVLAD; (4)VGG16 + NetVLAD and rerank-top5 by SuperPoint; (5)VGG16 + NetVLAD and rerank-top50 by SuperPoint; (6)VGG16 + NetVLAD and rerank-top100 by SuperPoint; (7)YOLOv3 + VGG16 + NetVLAD and rerank-top5 by SuperPoint; (8)YOLOv3 + VGG16 + NetVLAD and rerank-top50 by SuperPoint; (9)YOLOv3 + VGG16 + NetVLAD and rerank-top100 by SuperPoint.

In abovementioned experiments, test (1) is our solution A, test (2) is the global feature retrieval that only uses VGG16 + NetVLAD output, test (3) improves test (2) by adding YOLOv3 to pre-process the interference area; Test (4) to (6) are our solution B, in which only the number of TOP-N candidate from global feature searching is different, in particular, TOP- 5, 50, and 100 candidate images are selected. Test (7) to (9) correspond to the ideas of solution C. On the basis of the three test of experiments in solution B, YOLOv3 preprocessing is performed on the images to remove retrieval interferences.

In our experiment, the TOP-N candidate retrieved images was set with 5, 50, and 100 for subsequential rearrangement by using SuperPoint, this is to investigate the influence of the number of TOP-N on the accuracy of subsequent rearrangement based on local features. At the same time, in solution B and C, we also conducted experiments by exploring effect of the dimension of the global feature vector. The dimension of the global feature descriptor by the original VGG16 + NetVLAD is 4096. In order to further speed up the calculation of Euclidean distance, the popular PCA algorithm is used to reduce descriptor dimension to 1024 and 512, respectively, and the corresponding retrieval precision and time consuming are compared with. Due to the fact that the dimension of global feature descriptor dimension in test (1) is 1000 which is so closed to 1024, we report the corresponding results together in Tab. 3.

From Tab. 3, the hierarchical retrieval solution based on learning-based global features and local features suggested in this paper (in the case of top-5 candidate images) is better than the method that only relies on global features in terms of the accuracy, in particular, in the case of 4096-dimension, 1024-dimension, and 512-dimension, the average improved accuracy is 2.27%, 1.79% and 1.71%, respectively. However, we find that it is necessary to strictly control the number of images initially selected based on global features in the first step, because the more resulted similar images, the more likely it is to select some false positive similar images and negatively affect the precision of subsequent rearrangement based on local features. In solution B, when the number of top-N images is 50 and 100, it is about 7% and 9% lower than when the number of top-N images is 5. Comparing the results of test 4-6 and 7-9, it can be seen that using YOLOv3 to remove the interference area has a tiny improvement in the overall precision. Comparing test (1) with test (4), it is relatively explicit that the result of ResNet101 is inferior to Vgg16 + NetVLAD by a large margin in precision. At the same time, from the reported results of different feature descriptor dimensions, the reduction of the global feature dimension will result in a certain degree of precision reduction, i.e., the average precision of solution B is reduced by 2.51%, 2.94% when reducing the global feature descriptor dimension from 4096 to 1024 and 512, for solution C, the corresponding reduced precisions are 2.38%, 2.99%; In conclusion, solution A of the three hierarchical retrieval solutions proposed in this paper is worst, while solution B and C have basically similar performance.

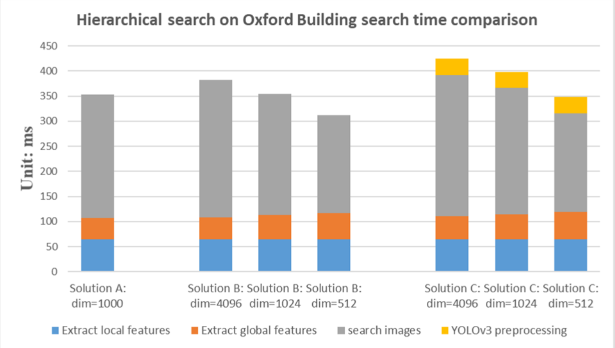


Figure 9. Time-consuming comparison of three hierarchical retrieval solutions on the Oxford Buildings dataset.

Fig. 9 reported the consuming time of the proposed three hierarchical image retrieval solution. All the experiments are conducted with hardware of one NVIDIA 1080Ti graphics card and Intel I7-8700K. As Fig. 9 shows, we can find that for the Oxford Buildings dataset, in terms of overall consuming time, among these three proposed solutions, solution B is the best, solution A is the second, and solution C is the worst, and their specific running time is 349ms,353ms,392ms, which are far lower than the Brute-Force matching, Bag-of-Words model, and Multi-Vocabulary Trees based on SIFT feature, the time consumption of these three methods are: 3690ms, 4434ms, 1519ms (these were test with the same machine). In solution C, due to the processing of YOLOv3, and the increased running is about 32ms, but it in turn benefits the precision to a small certain extent (see Fig. 9 for more details). Excluding the time for YOLOv3, the remaining part of solution C and solution B are nearly the same, so in practical applications, solution B or solution C can be flexibly selected up to the dataset.

In addition, we investigate the time consumption under different global feature dimensions in solution B and C, it can be seen that dimensionality reduction can indeed reduce the retrieval time. In fact, the time consumption for global feature extraction increases slightly, and the increased part is the due to the processing of PCA.

4.3 Experiments on Mixed Dataset

4.3.1 Oxford Building Dataset

To further test the performance of our proposed three hierarchical solutions, we simulate a dataset with more than 10,000 images by mixing the Oxford Building Dataset and Paris Dataset into a dataset, which we call the Mixed dataset. The Mixed dataset contains 11,475 images. The data set has been manually labeled for a total of 22 POIS. The dataset provides 5 images as images to be retrieved under each POI, a total of 110 images to be queried. The reference of the nearest similar images is generated by using SIFT feature matching, and the pairs with more found correspondences are supposed to be more similar.

4.3.2 Experimental Results and Analysis of Hierarchical Retrieval

Similar to the reported experiment in Tab. 3, a total of 9 sets of comparison experiments were also conducted on the Mixed dataset. The precision of the above nine test of Top1, Top3 and Top5 with three different global feature vector dimensions are evaluated, as shown in Table 4. The highest precision is marked in bold font.

The hierarchical retrieval solution based on learning-based global features and local features suggested in this paper (under the condition of five candidate images) is better than the method that only relies on global features in terms of the precision, in particular, in the case of 4096-dimensional, 1024-dimensional, and 512-dimensional features, the average improved precision is 1.04%, 0.93% and 1.01%, respectively. Nevertheless, it is necessary to point out that when choosing the number of images that are initially selected from global features in the first step. Because the more the initial candidate images, the more likely it

	4096			1024/1000			512		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
ResNet101-FC1000 + SuperPoint	/	/	/	0.7636	0.7318	0.6864	/	/	/
Vgg16+NetVLAD	0.9273	0.9152	0.8855	0.9091	0.8758	0.8327	0.9009	0.8606	0.8227
YOLOV3+VGG16+NetVLAD	0.9273	0.9170	0.8891	0.9118	0.8727	0.8364	0.9091	0.8667	0.8236
Vgg16 + NetVLAD + rerank-top5 by SuperPoint	0.9332	0.9224	**0.9034**	0.9182	**0.8818**	**0.8455**	0.9091	**0.8727**	0.8327
Vgg16 + NetVLAD + rerank-top50 by SuperPoint	0.8636	0.8030	0.7545	0.8545	0.7848	0.7473	0.8364	0.7788	0.7273
Vgg16 + NetVLAD + rerank-top100 by SuperPoint	0.8364	0.7818	0.7291	0.8273	0.7667	0.7091	0.8273	0.7576	0.7055
YOLOv3 + Vgg16 + NetVLAD + rerank-top5 by SuperPoint	**0.9364**	**0.9273**	0.8991	**0.9207**	0.8727	0.8427	**0.9182**	0.8713	**0.8409**
YOLOv3 + Vgg16 + NetVLAD + rerank-top50 by SuperPoint	0.8636	0.8121	0.7582	0.8545	0.7939	0.7436	0.8455	0.7818	0.7418
YOLOv3 + Vgg16 + NetVLAD + rerank-top100 by SuperPoint	0.8545	0.7758	0.7309	0.8364	0.7697	0.7218	0.8273	0.7606	0.7109

Table 4. Precision statistics of hierarchical retrieval solutions on Mixed Buildings dataset, and The precision of nine tests using global feature descriptors in dimension of 4096, 104/1000 and 512 were provided, and best results are highlighted in bold font.

is to reduce the accuracy of subsequent rearrangement using local features. In solution B, when the number of Top-N images is 50 and 100, the precision is about 10% and 12% lower than the precision of Top-5. Comparing the results of test 4-6 and 7-9, it can be figured out that using YOLOv3 to remove the interference area has a slight improvement in precision as a whole. Analogous to Tab. 4, it is obvious that the result of ResNet101 is inferior to Vgg16 + NetVLAD by a large margin in precision. At the same time, from the test results of different feature dimensions, the overall reduction of the global feature dimension will bring about a certain degree of accuracy reduction.

In general, comparing with the previous experiment, similar conclusion can be drawn that solution B and C perform similarly, and both of them are superior to solution A.

In terms of retrieval time, the performance of the above three hierarchical image retrieval solutions on the mixed dataset is shown in Fig. 10. Solution B using global feature descriptor of 512 dimension is the fastest, followed by solution B with 1000-dimension global feature and Solution A, and solution C is the slowest, running time are 582ms,605ms,648 ms (see Fig. 10). In solution C, the time consumption of YOLOv3 preprocessing is required, which is about 34ms, and this bring a slight improvement of precision. In practical application, solution B and C can be flexibly selected according to the dataset. In addition, we calculated the time consumption under different global feature dimensions in Solutions B and C, it can be seen that dimensionality reduction will reduce the retrieval query time, but, the time consumption for global feature extraction increases slightly, and the increased part is due to PCA dimensionality reduction. From the test on more than 10,000 image datasets, we found that solution B and C performed very well, which are supposed to be able to satisfy real-time and reliable image retrieval.

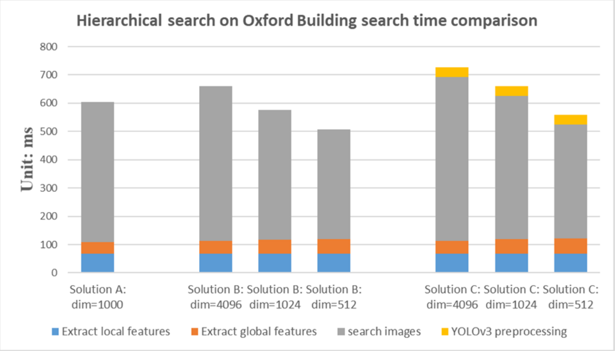


Figure 10. Time-consuming comparison of three hierarchical retrieval Solutions on the Mixed dataset.

5. CONCLUSION

In this work, we propose hierarchical image retrieval solutions based on learning-based global and local features. In particular, global feature extracted from ResNet101 and VGGNet16 +NetVLAD are explored, and the deep-learning based local feature Superpoint is also studied. To improve the search speed and also guarantee the precision, the global features are applied for fast retrieving the initial candidate images and the local feature are used to refine the initial retrieved results. Thus, we present three hierarchical solutions that use different combinations of global and local features. Our experimental results shows that the global features based on VGG16+NetVLAD significantly outperformed those based on ResNet101, the retrieval precision is always around 90% and the consuming time for querying one image from about 10000 images only takes 0.5 second.

In the future, we would like to first test more dataset (such as, photogrammetric benchmarks, UAV or close-range images etc.) and then integrate our image retrieval method into practical SfM or SLAM system to verify its real performance for image orientation tasks.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61871295).

REFERENCES

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297-5307.

Bay, H., Tuytelaars, T., & Gool, L. V., 2006, May. Surf: Speeded up robust features. In Proceedings of the European conference on computer vision, Springer, Berlin, Heidelberg, pp. 404-417.

Choy, C. B., Gwak, J., Savarese, S., & Chandraker, M., 2016. Universal correspondence network. Advances in neural information processing systems, 29.

Dalal, N., & Triggs, B., 2005. Histograms of oriented gradients for human detection. In Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1, pp. 886-893.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In Proceedings of IEEE conference on computer vision and pattern recognition, pp. 248-255.

DeTone, D., Malisiewicz, T., & Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 224-236.

Harris, C., & Stephens, M., 1988. A combined corner and edge detector. In Alvey vision conference, Vol. 15, No. 50, pp. 10-5244.

Hartmann, W., Havlena, M., & Schindler, K., 2016. Recent developments in large-scale tie-point matching. ISPRS Journal of Photogrammetry and Remote Sensing, 115, 47-62.

He, K., Zhang, X., Ren, S., & Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Jégou, H., Douze, M., Schmid, C., & Pérez, P., 2010. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp. 3304-3311.

Krizhevsky, A., Sutskever, I., & Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In Proceedings of Advances in neural information processing systems, 25.

Lazebnik, S., Schmid, C., & Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Vol. 2, pp. 2169-2178.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P., 1998. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, 86(11), 2278-2324.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), pp. 91-110.

Rublee, E., Rabaud, V., Konolige, K., & Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International conference on computer vision, pp. 2564-2571. analysis and machine intelligence, 40(5), pp. 1224-1244.

Simonyan K, Zisserman A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9.

Torralba, A., Murphy, K. P., Freeman, W. T., & Rubin, M. A., 2003. Context-based vision system for place and object recognition. In Proceedings of the International conference on computer vision, Vol. 2, pp. 273-273.

Wang, X., Zhan, Z. Q., Heipke, C., 2017. An Efficient Method to Detect Mutual Overlap of a Large Set of Unordered Images for Structure-From- Motion. In Proceedings of ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-1/W1, pp. 191-198.

Yi, K. M., Trulls, E., Lepetit, V., & Fua, P., 2016. Lift: Learned invariant feature transform. In Proceedings of European conference on computer vision, Springer, Cham, pp. 467-483.

Zheng, L., Yang, Y., & Tian, Q., 2017. SIFT meets CNN: A decade survey of instance retrieval. IEEE transactions on pattern