Image Retrieval Based on the Weighted and Regional Integration of CNN Features

Kaiyang Liao[1], Bing Fan[1*], Yuanlin Zheng[1], Guangfeng Lin[1], Congjun Cao[1,2]
[1] Xi'an University of Technology, School of Printing, Packaging and Digital Media, Xi'an, China
[e-mail: 1120974840@qq.com]
[2] Printing and Packaging Engineering Technology Research Centre of Shaanxi Province Xi'an, China
[e-mail: caocongjun@xaut.edu.cn]
[*]Corresponding author: Bing Fan

Received July 6, 2021; revised January 30, 2022; accepted March 18, 2022; published March 31, 2022

Abstract

The features extracted by convolutional neural networks are more descriptive of images than traditional features, and their convolutional layers are more suitable for retrieving images than are fully connected layers. The convolutional layer features will consume considerable time and memory if used directly to match an image. Therefore, this paper proposes a feature weighting and region integration method for convolutional layer features to form global feature vectors and subsequently use them for image matching. First, the 3D feature of the last convolutional layer is extracted, and the convolutional feature is subsequently weighted again to highlight the edge information and position information of the image. Next, we integrate several regional eigenvectors that are processed by sliding windows into a global eigenvector. Finally, the initial ranking of the retrieval is obtained by measuring the similarity of the query image and the test image using the cosine distance, and the final mean Average Precision (mAP) is obtained by using the extended query method for rearrangement. We conduct experiments using the Oxford5k and Paris6k datasets and their extended datasets, Paris106k and Oxford105k. These experimental results indicate that the global feature extracted by the new method can better describe an image.

Keywords: Image retrieval, Weighting feature, Global feature, Convolutional neural network, Regional integration.

This work is supported by the National Natural Science Foundation of China Project Nos. 61771386, 52075435 and Natural Science Foundation of Shaanxi Province No.2021JM-340.

http://doi.org/10.3837/tiis.2022.03.008 **ISSN : 1976-7277**

1. Introduction

In recent years, content-based image retrieval technology (CBIR) [1] has developed rapidly and obtained abundant research results. The rough CBIR process is to extract the underlying visual features, such as the image color, texture and shape; calculate the distance between the query image and the test image features; and return the image that is the most identical or similar to the query image [2,3]. However, there is a semantic gap between the low-level features and the high-level semantics, which cannot accurately reflect the user's retrieval intention. With the rapid development of social information technology and the arrival of the era of big data, not only has the number of pictures increased, but the content of pictures has also become more complex and diverse. The traditional text-based and content-based retrieval technology has been unable to meet the needs of users.

In the ILSVRC 2012 competition, Krizheysky et al. designed a deep convolutional network model called AlexNet [4], which reduced the error rate of image classification from 26.2% to 15.3%, notably better than other algorithms. This finding caused the Convolutional Neural Networks (CNNs) [5] to receive considerable attention in the visual image field, and it became the first choice as the basic image retrieval model. Recent research shows that the CNN feature has achieved unexpected results in the image retrieval field. Compared with traditional methods, the retrieval precision has been considerably improved [6]. Based on this background, this paper proposes a method of generating a one-dimensional eigenvector to describe a whole picture based on the CNN model. The method weights each feature element and finally integrates it into a one-dimensional global feature vector. In this paper, the target recognition CNN model was directly used. Without retraining the network parameters, the retrieved mAP on the Paris6k dataset reaches 0.87 and that on the Oxford5k dataset reaches 0.75.

In general, the CNN is composed of convolutional layers and fully connected layers. After the image data are fed into the CNN, three-dimensional convolutional features and one-dimensional fully connected feature vectors can be generated. When images with different sizes are input, the feature length of the fully connected layer is the same, and the convolutional layer size is different. Therefore, the convolution layer feature can better reflect the original features of the image than the fully connected layer feature. In this paper, the network model only intercepts the convolutional layer, which avoids adjusting and clipping the size of the input image and keeps the spatial characteristics of images with different scales unchanged [7]. Convolutional features are high-dimensional features. Usually, the size of the convolutional features extracted from an image is hundreds of times that of traditional features. It will take considerable time to directly match images from a large dataset, which is clearly not applicable. Therefore, how to simplify the features and maintain their original properties is the core of this paper. The basic idea is to compress the three-dimensional convolutional features into a one-dimensional eigenvector and allow the vector to still contain edge information and position information. Therefore, this paper uses the feature weighting and region integration method to address the convolutional special features. Finally, appropriate image matching and rearrangement methods are selected to calculate the similarity between images. Based on the above ideas, three aspects of work should be done: 1. Convolution feature extraction and weighting; MatConvNet [8] (Convolutional Neural Networks for MATLAB) in this paper are used to extract the convolutional features of the image and to weight the features so that the elements including edge information and location information are given a greater weight. 2. Regional integration; we transform high-dimensional features into simple global feature vectors. 3. Retrieval and rearrangement; we match the similarity between the feature vectors of the query image and the images in the dataset to get the initial ranking, and then to rearrange

to get the final result.

This paper is tested on the Oxford5k [9] and Paris6k [10] datasets and on the Paris106k and Oxford105k datasets extended with 100k images. The results show that our method achieves better results than such methods as CroW [11] and R-MAC [12]. In other words, the features obtained by this method can better describe an image. The structure of this paper is as follows: we review the related work in Sec.2. We describe proposed method in detail in Sec.3. We describe experiment and discussion in Sec.4. Conclusion is provided in Sec. 5.

2. Related work

Most early image retrieval methods were based on low-level features, such as SIFT features [13], GIST characteristics [14,15], and color features and texture features [16], and were combined with the BOF [17] model to retrieval. However, SIFT, GIST and other features cannot express the deep semantics of images, and the high complexity of the BOF model limits the room for improvement of such methods. The Vector of Aggregate Locally Descriptor (VLAD) [18] and Fisher Vectors [19] integrated local features into global features, which reduced the complexity of features and accelerated the retrieval speed, but the underlying features with small amounts of information lose information again. However, literature [20] using SIFT features to conduct image retrieval has achieved good results, but the method is far from sufficient.

Next, in the ILSVRC-2012 competition, Krizheysky et al. designed the AlexNet network, and the deep convolutional neural networks began to enter people's focus. A large number of studies have shown that CNN features have considerable advantages over traditional features. Today's image retrieval technology has shifted from traditional features to the study of convolutional neural network features. Neural codes [21] were among the earliest ways to use CNN features for image retrieval, and then this field developed rapidly. Because the features from deep learning have high dimensionality, which leads to the large time and memory consumption, it is urgently important to simplify deep learning methods. At present, deep learning methods can be roughly divided into two processing methods: one method uses a hashing algorithm to handle features, such as literature [22]; and the other method compresses high-dimensional features to reduce the dimension to form global features to describe the whole image, such as CroW [11]. The second method is used in this paper.

The SPoC's [23] method solved the problem of how to turn the feature image of convolutional layers into a single feature vector, and its use and sum-pooling [24] technique also achieved good results. However, one problem was that even the best features are noisy. SPoC only used the radial basis function to calculate the weight of each pixel, and it ignored the weight of each channel. Later, Yannis et al. proposed the CroW feature extraction method, which gives weights to the last convolutional layer feature of the CNN. The method not only changed the weights of the feature image elements of each layer, but it also weighted each channel, to achieve better results. Feature weighting is similar to the attention strategy. However, the weight sources are different. The feature weights are equal, each dimension is multiplied by the same weight parameter. Attention strategy is equal that each feature is multiplied by a different weight parameter, weight parameter is adaptive learning. Recent studies have shown that using the sliding window principle is beneficial to target localization. R-MAC [12] uses sliding windows with different scales to conduct the max pooling of feature maps and then integrates the three-dimensional features of multiple channels into single-dimensional vectors to describe global information. We find that the two methods can be used together; therefore, we propose a new method to address features and improve the retrieval

performance.

Currently, rearrangement has become an essential step in the image retrieval process. Previous research [25] has greatly improved the recognition accuracy through complex rearrangement algorithms. However, the focus of this paper is to test the performance of the processed global eigenvectors. Therefore, the rearrangement method selects the simple Query Expansion (QE) algorithm.

3. Proposed method

3.1 Overall Framework of The Method

This section describes the proposed in this paper in detail. The method inputs a pair of images and outputs a retrieval ranking as shown in **Fig. 1**. The method in this paper is divided into three parts: the first is feature extraction and weighting, the second is feature integration, and the third is retrieval and rearrangement.

Formerly, when extracting local features (such as SIFT features), the constructed BoW, VLAD and Fisher Vectors, the MSER [26], saliency [27] and other methods could be used to limit the SIFT features to the regions with objects. Similarly, in CNN-based image retrieval, we want to highlight regions with object feature maps. There were usually two ways to refine the features for image retrieval: one is to perform object detection first and then extract the CNN features in the detected object region; and the other is to increase the weight of the object region and reduce the weight of the nonobject region using some adaptive weighting method. This paper adopts the latter method. The proposed Multiscale Order less Pooling of deep convolutional activation features (MOP) [28] algorithm by Y Gong et al. was used a multi-scale sliding window to process an original image. In this paper, a similar method is used to process the sliding window on the feature map to get the region feature vector. The region vector is added directly to obtain the global feature. Finally, the similarity score between the feature vectors is calculated to retrieve the ranking.

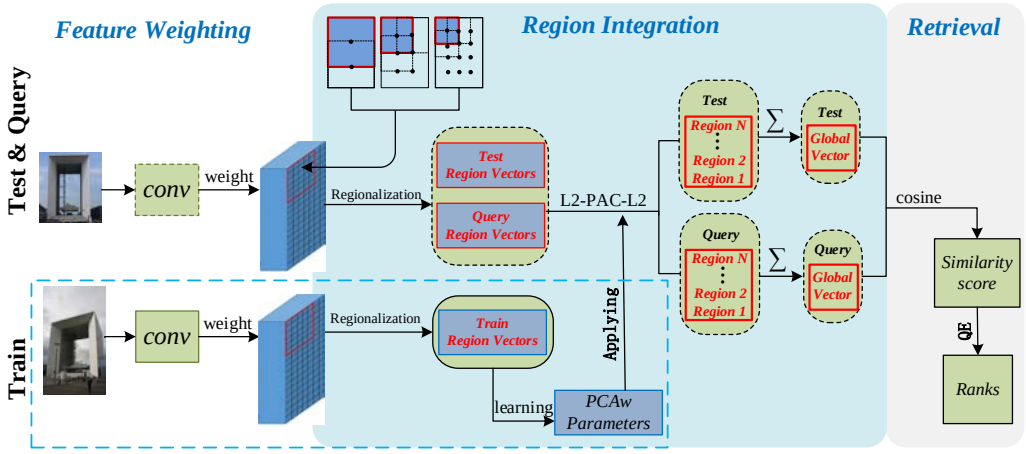

Fig. 1. The structure of this algorithm

As shown in **Fig. 1**, the entire method can be divided into three parts:

(i)Feature extraction with weighting: In the front part of **Fig. 1**, will extract the last convolutional layer feature for spatial weighting and channel weighting. When space is weighted, we sum up multiple channels to highlight nonzero and largely responsive regions, which are also generally the regions where the object is located. When the channel is weighted,

we use the idea of Inverse Document Frequency (IDF) for assignment.

(ii)Regional integration: In the middle part of **Fig. 1**, multiple sliding windows with different sizes are designed. The elements in each sliding window are added together to obtain multiple local feature vectors. After L2 normalization, PCA-whitening dimensionality reduction is conducted, L2 normalization is conducted again and finally the results are directly added together to get global features.

(iii)Retrieval and rearrangement: In the back part of **Fig. 1**, first, we use cosine similarity matching, and then we subsequently rearrange with the QE method to obtain the final result.

3.2 Feature Weighting

In this paper, the 3D features are extracted by a convolutional neural network are represented by $X \in R^{(N\times W\times H)}$,where W and H are the size of the feature map of each layer, and N is the number of channels. We note that the size of W and H varies with the size of the input network image. x_{kij} represents the element at position (i,j) of the feature plane on the kth channel. $C^{(k)}$ represents the whole feature map of the kth channel. First, we weight each element on $C^{(k)}$.Then, there are a total of $i \times j$ weights, and each weight is represented by a_{ij}. Similarly, $\lambda_{ij}^{(k)}$ represents element at (i,j) in the feature map of the kth channel, $\lambda_{ij}^{(k)} = x_{kij}$. We assign a weight to each channel, and each weight is represented by b_k. After weighting, the value of each element becomesx'_{kij}, which is represented by the following expression:

$$x'_{kij} = a_{ij} b_k x_{kij} \tag{1}$$

The purpose of map weighting and channel weighting is to increase the weight of the region of interest and to reduce the weight of the nonobject regions. When a map is weighted, the feature maps of each channel are summed directly. Usually, the places with strong responses are generally the edges of the object after convolutional filtering. After summing the multiple channels, the regions with nonzero and large responses are generally the regions where the object is located; therefore, they can be used as the weights of the feature maps. s'_{ij} represents the sum of the elements located at (i,j) on each feature map, and it is defined as:

$$s'_{ij} = \sum_k c_{ij}^{(k)} \tag{1}$$

The k is in the range of [1, N].

Next, we can obtain the map weight a_{ij}, which is located at (i,j) on the feature map. The expression is as follows:

$$a_{ij} = \left(\frac{s'_{ij}}{\left(\sum_{m,n} {s'_{mn}}^{\alpha}\right)^{\frac{1}{\alpha}}} \right)^{\frac{1}{\beta}} \tag{3}$$

In formula 3, the range of values of m is [1, W], and the range of values of n is [1, H], $\left(\sum_{m,n} {s'_{mn}}^{\alpha}\right)^{\frac{1}{\alpha}}$ can be regarded as the norm of the matrix. According to experimental data, the results are the best when α=2 and β=2.

Channel weighting follows the idea of IDF weighting. If each element value on the feature map of a channel is non-zero and large, then the strong response region occupies the whole feature map visually; therefore, the feature map of this channel is not conducive to locating

the region where the object is located, and its channel weight needs to be reduced. However, for the channels whose strong response area is very small, it can be considered that the channels contain the image object position information, therefore, we need to increase the weights of these channels. $N_k$represents the number of nonzero elements on the *k*-th channel as follows:

$$N_k = \sum_{ij} 1\left[\lambda^{(k)}_{(ij)} > 0\right] \tag{2}$$

It has been indicated that the nonzero rare element channel can describe the location information of the image more; therefore, it can be weighted according to the number of nonzero elements. The expression b_k is as follows:

$$b_k = log\left(\frac{\sum_h N_h}{\varepsilon + N_k}\right) \tag{3}$$

To ensure that the denominator is not zero, the minimum value ε is added. Because the molecules are relatively large, $\varepsilon = 1$ is used in this experiment, and it has little effect on the results.

3.3 Regional Integration

This section introduces the method that integrates 3D features into global feature vectors. CroW added the elements of each channel directly and gave the following expression for the characteristic vector *F:* $F = [f_1, f_2, \ldots, f_k]$. The representation of f_k is as follows:

$$f_k = \sum_{i=1}^{W} \sum_{j=1}^{H} x'_{kij} \tag{4}$$

A CroW feature only considers the global feature map without considering the locality. Therefore, we divided the feature map into several different regions for separate calculations and then integrated them. Part of **Fig. 1** indicates the sizes of three windows. As shown in **Fig. 1**, '·'represents the center of the window. The feature map area corresponding to each window was processed via summation. In this paper, the region of the features map was divided by using *L* different sliding windows. For example, when the *L* =3, 20 regional features can usually be obtained. In addition, the whole feature map was processed using the additive summation method to obtain a feature vector, therefore, an image can obtain 21 regional features. The 21 regional features were directly added together to obtain the final global feature. There were some overlapping areas between windows, and this paper adds them together to generate global features. Therefore, it can be considered that those overlapping areas were assigned greater weights. Each sliding window is a square, and we adopted uniform sampling and automatically adjusted the center position to ensure a 40% overlap area. The size of the sliding window was determined by the short side $min\ (W, H)$ of the feature map, and the expression of the length of the sliding window is as follows:

$$l = \frac{2 \times min(W,H)}{L+1} \tag{5}$$

As shown in **Fig. 1**, when *L* =3, there are three sliding windows with different scales on the feature map, and adding the elements in the sliding window are directly added. A sliding window generates a feature vector $\boldsymbol{F'}$, similar to CroW, where $\boldsymbol{F'} = [f_1', f_2', \ldots, f_N']$. Only

$f_k'$is limited to the addition of elements within the sliding window and not the addition of elements of the entire feature image. After applying *n* sliding windows, the sequence of L2 normalization, PCA-whitening dimensional reduction and L2 normalization was used successively to optimize, and the three-dimensional features became n regional feature vectors. Finally, the global feature vector G was generated by adding all the regional feature vectors together. The expression of G is as follows:

$$G = F_1' + F_2' + \cdots + F_n' \tag{6}$$

In this paper, the method is progressive in turn. First, the extracted convolutional feature maps are weighted and then the channels are weighted. Finally, the multiscale sliding window is used to integrate the features to obtain the image descriptor. Each step can gradually improve the final retrieval effect.

3.4 Retrieval and Rearrangement

The cosine distance, which is also known as the cosine similarity, is a measure of the difference between two individuals by using the cosine of the angle between two vectors in vector space. The Euclidean distance can reflect the absolute difference of individual numerical characteristics; therefore, more is needed to analyze the differences in the sizes of the dimensions. The cosine distance uses the direction more to distinguish differences, but it is not sensitive to the absolute value. The cosine distance is used more for users' score content to distinguish between similarities and differences of interest. Moreover, revised user disunity problems that may exist between metrics (because the cosine distance is not sensitive to the absolute value). The global feature vector was regarded as a directional line in multidimensional space. If the direction of two vectors is the same, which means that the included angle is close to zero, then the two vectors can be considered as close to each other. To ensure that two vectors are in the same direction, we used the law of cosines to calculate the angle between the vectors. Therefore, the cosine distance was used in this paper to identify the similarity of global feature vectors, and the similarity of pictures X and Y is expressed by the following expression:

$$sim(X,Y) = cos\,\theta = \frac{\overrightarrow{G_X}\cdot\overrightarrow{G_Y}}{\|G_X\|\cdot\|G_X\|} \tag{7}$$

Because the cosine distance represents in the difference in direction between vectors, the use of L2 normalization or PAC-whitening on the job would not affect the final similarity assessment.

QE (query expansion) is a simple rearrangement method. The method selects the top image (including the query image), calculates the mean vector of its feature vector, and finally uses the mean vector to conduct the final rearrangement of the result. Although the method is simple, it can considerably improve the retrieval recall rate. As shown in **Fig. 2**, the experiment on the Oxford5k dataset uses the VGG16 model [29]. On the left is the query image, and on the right are the 6th,7th,8th,9th,10th charts corresponding to the initial ranking and reranking. It can be seen that some of the correlation image behind the initial ranking after using QE are at the top of the ranking, which improves the retrieval recall rate. Because *qe*=5 is selected, the initial and reranked top five change only slightly.

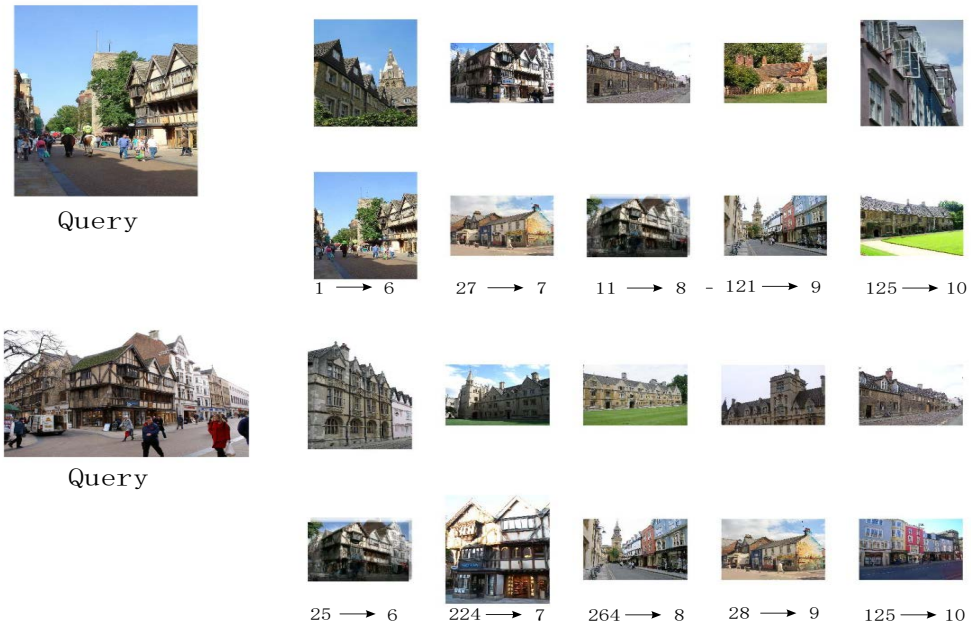

Fig. 2. The effect after using QE

4. Experiment and discussion

4.1 Datasets and CNN model

This paper used the Oxford Building dataset and Paris dataset for experiments. The two datasets contained 5063 images and 6412 images, respectively. In addition, we added 100k Flickri [30] images to the Oxford5k and Paris6k datasets to extend the datasets. We used Oxford105k and Paris106k to represent the extended datasets. Principal component analysis (PCA) is a common method for data preprocessing method in statistical machine learning and data mining. There are two functions of PCA: one is data dimensionality reduction, and the other is data visualization. A preprocessing step associated with PCA is whitening. Because the adjacent pixels of an image are related, the inputs used for training are redundant. Whitening reduces the redundancy of the inputs to make each feature have the same variance. Therefore, PCA-whitening was used to optimize the feature vectors in this paper. When we used Oxford5k dataset as the test dataset, the Paris6k dataset served as the training dataset to train PCA-whitening and vice versa. For the query image dataset, we defined a query image according to the standard protocol area box, and the query area was cropped and subsequently fed into the CNN model. To make the final result persuasive, 5 images of each type image were removed as the query images and formed a query dataset of 55 images. In this paper, the mAP was used to evaluate the performance of the methods.

In the whole experiment, the CNN network parameters were not retrained. Instead, convolutional features were extracted and optimized for image retrieval. In this paper, two CNN models were used for feature extraction: AlexNet and VGG16. Since the features were extracted from the last convolutional layer, the features that were generated by AlexNet have 256 channels, and those generated by VGG16 have 512 channels. Finally, the global features processed are 256 and 512 dimensions respectively. Our feature was quite similar to the CroW feature, and both of them form the global feature to describe a whole image in the end.

However, the features that were extracted by the method in this paper also took local areas into account, therefore, they had better performance. We used the MatConvNet tool to extract features.

4.2 Parameter Selection

We compared the performance of CroW features and the features extracted by the method of this paper. Section 2.3 detailed the differences between the features of extracted by the method of this paper and CroW. The performance of the two features for image retrieval is shown in **Table 1**. As shown in the region integration section of **Fig. 1**, there were certain overlapping regions that we can consider to be providing a large weight to those overlapping regions between the windows. Finally, summation was used to form the global feature. Therefore, we did not to divide the feature maps as fine as possible. In this paper, the overlap rate between sliding windows was 40%. In the experiment, we used L sliding windows to process feature maps with different scales, and we applied the VGG16 model to the Oxford5k dataset. The experimental results are shown in **Table 1**, which shows that the retrieval effect is the best when L =3.

Table 1. The performance with different scales of sliding windows

The Network	CroW	Ours			
		L = 1	L = 2	L = 3	L = 4
AlexNet	/	0.710	0.734	0.743	0.738
VGG16	0.797	0.798	0.809	0.805	0.797

We extracted the entire dataset and the global features of the query images. In this paper, the feature vector of the query images and the feature vector of the images in the dataset were assinged cosine values, and the cosine values were used as the standard to measure the similarity of the two images. The image ranking was obtained as the initial result. Finally, QE was used to rearrange the results. The method selected the top-ranked *qe* image that included the query image, calculated the final mean vector of their feature vector, and finally used the average value vector to conduct the final rearrangement of the results. **Fig. 3** shows the result of rearrangement when taking *qe*=1:50. It can be seen that after rearrangement, the result was been improved, but a higher qe value does not necessarily mean a better result. For the Paris6k dataset, *qe* <30, and as *qe* increases, the mAP has a slow growth. For Oxford5k, *qe*>5, as *qe* increases, the mAP rapidly declines. Therefore, this paper selected *qe*=5 for the final rearrangement when the methods are tested on the Oxford dataset, and *qe*=10 for final rearrangement when is the methods are tested on the Paris dataset.

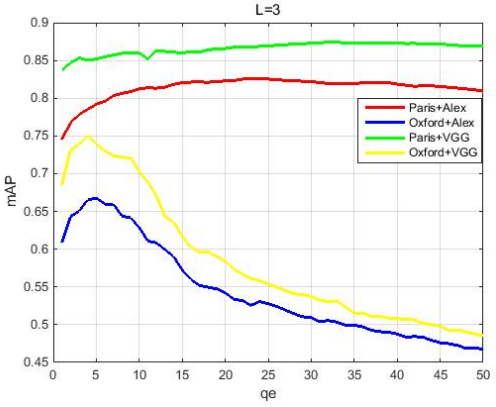

Fig. 3. Changes in mAP with *qe* values

4.3 Experimental results

We compared results of the proposed methods with those of Neural Codes, SPoC, R-MAC, and CroW in this section, these methods are briefly described in section 1. The *d* =256 and *d* =512 of the second column in **Table 2** represented retrieval using the features extracted from the AlexNet and VGG16 network models, respectively. The third and fourth columns represented the mAPs of the methods tested on the Paris6k and Oxford5k datasets. The fifth and sixth columns indicated that 100k of the wrong images in the Oxford100k dataset were mixed in the Paris6k (Oxford5k) dataset to extend into a new Paris106k (Oxford 105k) dataset. Then, the results of the tests were calculated. It can be seen in **Table 2** that the method that achieves the best results on all four datasets is that using the AlexNet model. The method achieved an R-MAC accuracy that was approximately 3% better on average. The method using the VGG16 model achieves the best test result on the Paris datasets. After using QE to rearrange the initial retrieval results, on the Paris106k dataset, the method in this paper achieved a mAP 2% better than that of the CroW method; and for the Oxford5k dataset, the method in this paper achieved a mAP 1-2% better.

The final results of the tests in the four datasets outperformed the CroW, and the mAP averaged increased by 1%. The last three lines * were denoted as the results obtained at *qe* =30. For the Paris dataset, when *qe* was 30, the method in this paper achieved a mAP approximately 3% better than that of the CroW method.

Table 2. mAP values for different methods

Method	d	Paris6k	Oxford5k	Paris106k	Oxford105k
Tr. Embedding[20]	1024	/	0.560	/	0.502
Neural Codes[21]	256	/	0.435	/	0.749
Razavian et al.[31]	256	0.670	0.533	/	0.489
SPoC[23]	256	/	0.531	/	0.501
R – MAC[12]	256	0.729	0.561	0.601	0.470
Ours	256	**0.743**	**0.593**	**0.614**	**0.532**
Neural Codes[21]	512	/	0.435	/	0.392
R – MAC	512	**0.830**	0.669	**0.757**	0.616
Ours	512	0.805	**0.683**	0.733	**0.626**
Ours + QE.	256	0.812/0.822 *	0.670	0.702/0.720 *	0.630
CroW + QE[11]	512	0.848	0.749	0.794	0.706
Ours + QE.	512	**0.860/0.873 ***	**0.751**	**0.813/0.822 ***	**0.717**

Fig. 4 shows the top-10 retrieval results for 11 query images (one for each type of image) for the method that used the VGG16 model applied on the Paris6k dataset. The first column was the query image, and it was followed by the retrieval results ranked 1-10 for the corresponding image. It can be seen that the image retrieved by the method proposed in this paper is mostly accurate.

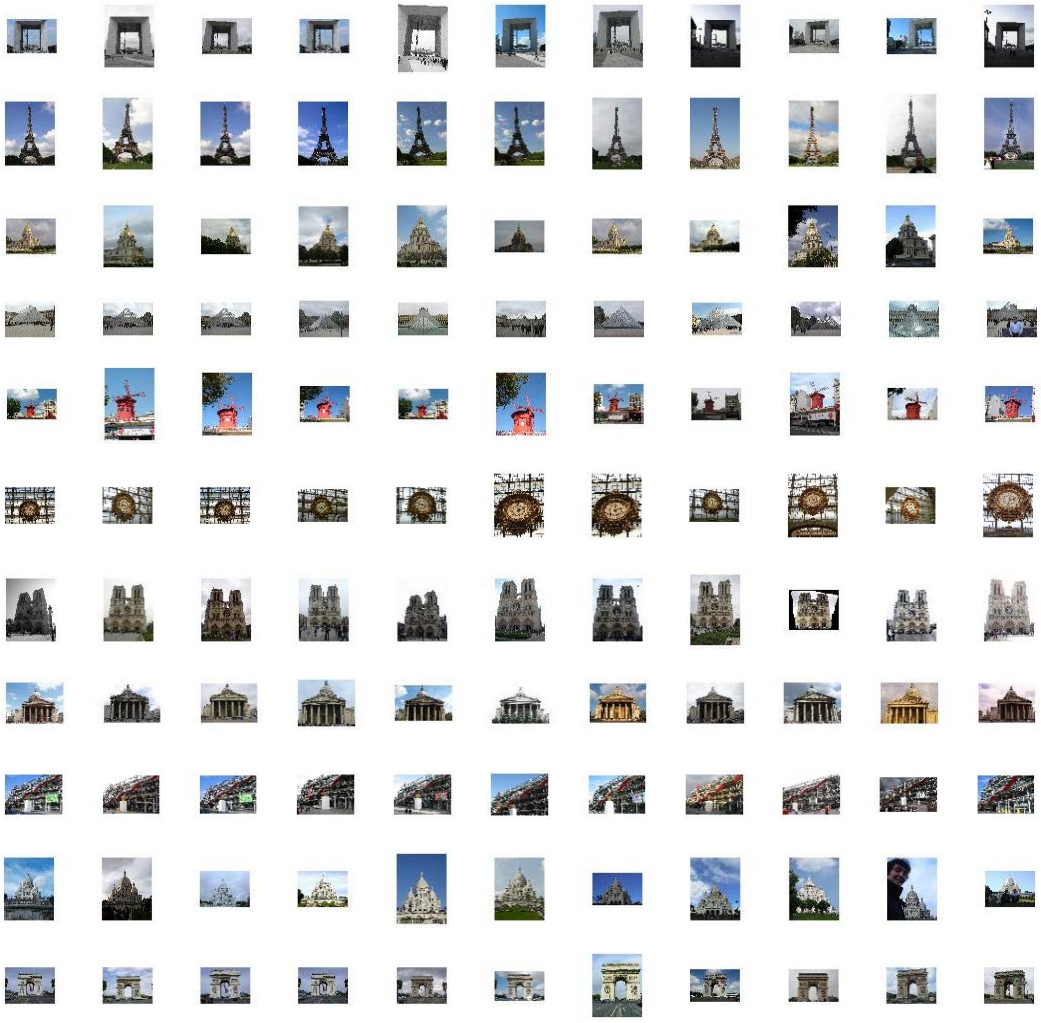

Fig. 4. Top-10 results returned for all 11 queries

5. Conclusion

The main research content of this paper concerns how to generate features that can accurately describe an image. This paper constructs a complete image retrieval model: first, feature extraction and processing are carried out; next, similarity matching is performed; and finally, the retrieval results are rearranged and optimized. Using the convolutional layers of the AlexNet and VGG16 models to partially extract features, a whole image is described by the global features generated by feature weighting and feature integration. Next, the cosine similarity of the query image and the image in the dataset is calculated to get the initial retrieval result. Finally, the QE algorithm is used to rearrange the initial ranking to get the final mAP.

Testing of the method proposed in this paper on the Paris6k and Oxford5k datasets and the extended Paris106k and Oxford105k datasets, respectively, achieved advanced performance levels without retraining the network. Therefore, it is concluded that the global feature vector extracted by this method can accurately describe an image's properties.

References

[1] K. Ahmad, M. Sahu, M. Shrivastava, M. A. Rizvi, and V. Jain, "An efficient image retrieval tool: query based image management system," *International Journal of Information Technology*, vol. 12, no. 1, pp. 103-111, May. 2020. Article (CrossRef Link)

[2] B. Khaldi, O. Aiadi, and K. M. Lamine, "Image representation using complete multi-texton histogram," *Multimedia Tools and Applications*, vol. 79, no. 11, pp. 8267-8285, Jan. 2020. Article (CrossRef Link)

[3] P. Jitesh, A. K. Pal, H. Banka, and P. Dansena, "Fusion of region based extracted features for instance-and class-based CBIR applications," *Applied Soft Computing*, vol. 102, pp. 1-24, Apr. 2021. Article (CrossRef Link)

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, Jun. 2017. Article (CrossRef Link)

[5] J. Kim, H. Park, and J. I. Park, "CNN-based image steganalysis using additional data embedding," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 1355-1372, Oct. 2020. Article (CrossRef Link)

[6] Q. Liu, X.Y. Xiang, J. H. Qin, Y. Tan, and Q. Zhang, "Reversible Sub-Feature Retrieval: Toward Robust Coverless Image Steganography for Geometric Attacks Resistance," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 3, pp. 1078-1099, Feb. 2021. Article (CrossRef Link)

[7] X. D. Wang, Z. D. Zheng, Y. He, F. an, Z. Z. Zeng, and Y. Yang, "Progressive local filter pruning for image retrieval acceleration," *arXiv preprint arXiv:2001.08878*, Jan. 2020.

[8] A. Vedaldi, and K. Lenc, "Matconvnet: Convolutional neural networks for matlab," in *Proc. of the 23rd ACM international conference on Multimedia*, pp. 689-692, Oct. 2015. Article (CrossRef Link)

[9] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of 2007 IEEE conference on computer vision and pattern recognition*, pp. 1-8, Jun. 2007. Article (CrossRef Link)

[10] J. Philbin, O. Chum, M. Isard, S J. ivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of 2008 IEEE conference on computer vision and pattern recognition*, pp. 1-8, Jun. 2008. Article (CrossRef Link)

[11] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," in *Proc. of European conference on computer vision*, pp. 685-701, Oct. 2016. Article (CrossRef Link)

[12] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *arXiv preprint arXiv:1511.05879*, Nov. 2015. Article (CrossRef Link)

[13] D. G. Lowe, "Local feature view clustering for 3D object recognition," in *Proc. of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 291-301, Dec. 2001. Article (CrossRef Link)

[14] Z. Yang, X. Gao, Z. Xie, and K. W. Wu, "Scene categorization of local Gist feature match kernel," *Chinese Journal of Graphics*, vol. 18, no. 3, pp. 264-270, Mar. 2013. Article (CrossRef Link)

[15] J. Chaki, N. Dey, *Texture feature extraction techniques for image recognition*, Springer, Singapore, 2020.

[16] J. Sivic, and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. of Ninth IEEE International Conference on Computer Vision*, vol. 2, pp. 1470-1477, Oct. 2003. Article (CrossRef Link)

[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. of 2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3304-3311, Jun. 2010. Article (CrossRef Link)
[18] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *Proc. of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3384-3391, Jun. 2010. Article (CrossRef Link)
[19] H. Jégou, and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 3310-3317, 2014. Article (CrossRef Link)
[20] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *Proc. of European conference on computer vision*, Springer, Cham, pp. 584-599, Sep. 2014. Article (CrossRef Link)
[21] F. Chen, Z. H. Lv, J. Li, X. D. Wang, and Y. Dou, "Multi-Label Image Retrieval by Hashing with Object Proposal," *Chinese Journal of Image Graphics*, vol. 22, no. 2, pp. 232-240, 2017. Article (CrossRef Link)
[22] A. Babenko, and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," in *Proc. of the IEEE conference on computer vision and pattern recognition, arXiv preprint arXiv:1510.07493*, 2015. Article (CrossRef Link)
[23] A. Raza, T. Nawaz, H. Dawood, and H. Dawood, "Square texton histogram features for image retrieval," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2719-2746, Feb. 2019. Article (CrossRef Link)
[24] F. Wei, W. N. Yi, P. Lin, and S. N. Hou, "A Method of License Plate Location and Character Recognition based on CNN," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 8, pp. 3488-3500, 2020. Article (CrossRef Link)
[25] C. Y. Kao, and S. A. Mohammadi, "Extremal rearrangement problems involving Poisson's equation with Robin boundary conditions," *Journal of Scientific Computing*, vol. 86, no. 3, pp. 1-28, 2021. Article (CrossRef Link)
[26] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 1155-1162, 2013.
[27] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale Orderless Pooling of Deep Convolutional Activation Features," in *Proc. of European Conference on Computer Vision, Springer International Publishing*, vol. 8695, pp. 392-407, 2014. Article (CrossRef Link)
[28] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, Apr. 2014.
[29] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International journal of computer vision*, vol. 87, no. 3, pp. 316-336, 2010. Article (CrossRef Link)
[30] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki, "Visual instance retrieval with deep convolutional networks," *ITE Transactions on Media Technology and Applications*, vol. 4, no. 3, pp. 251-258, 2016. Article (CrossRef Link)

Kaiyang Liao received the B.S. degree in computer science from the XIDIAN University, Xi'an, China, in 2004, the M.S. degree in computer science from the University of Science and Technology LiaoNing, Anshan, China, and the Ph.D. degree in Information and Communication Engineering from the Xi'an Jiaotong University, Xi'an, China, in 2013. He is currently a Full lecturer with the School of Printing and Packaging Engineering, Xi'an University of Technology, Xi'an, China. His research interests include data mining, pattern recognition, video analysis and retrieval.

Bing Fan received the B.E. degree in Printing Engineering from Hu'nan University of Technology, Hu'nan, China, in 2019. She is currently a postgraduate student in Xi'an University of Technology, Xi'an, China. Her research interests include image retrieval and digital image processing.

Yuanlin Zheng received the B.S. degree in Printing Engineering from Zhuzhou Institute of Technology, Zhuzhou, China, in 1999, the M.S degree in Pulp and Papermaking Engineering from Xi'an University of Technology, Xi'an, China, in 2002, and the Ph.D. degree in Pulp and Papermaking Engineering from Tianjin University of Science & Technology, Tianjin, China, in 2007. He is currently a fully associate professor with Printing Engineering in Xi'an University of Technology, Xi'an, China. His research interests include color management, evaluation of quality of color image and color science, pattern recognition.

Guangfeng Lin received the PhD in control theory and control engineering from Xi'an University of Technology. He is currently a lectorate of Department of Information Science at the Xi'an University of Technology. His research interests include digital image processing and pattern recognition. He is CCF professional member and ACM member.

Congjun Cao received the B.S. degree in printing machinery and technology from the Xi'an University of technology, Xi'an, China, in 1992, the M.S. degree in printing engineering from Xi'an University of technology, Xi'an, China in 1998, and the Ph.D. degree in computer software and theory from the Northwest University, Xi'an, China, in 2008. She is currently a Full professor in the School of Printing, Packaging Engineering and digital media technology, Xi'an University of Technology, Xi'an, China. Her research focuses on color management technology, quality control technology of printing image reproduction and functional printed materials, video analysis and retrieval.