Neurocomputing 363 (2019) 17–26

Contents lists available at ScienceDirect

Neurocomputing

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

Brief papers

Improving cross-dimensional weighting pooling with multi-scale feature fusion for image retrieval

Qi Wang, Jinxiang Lai, Zhenguo Yang*, Kai Xu, Peipei Kan, Wenyin Liu*, Liang Lei

School of Computers, School of Physics & Optoelectronic Engineering, Guangdong University of Technology, Guangzhou, China

ARTICLE INFO

Article history:
Received 11 December 2018
Revised 31 July 2019
Accepted 8 August 2019
Available online 12 August 2019

Communicated by Zhiyong Wang

Keywords:
Convolutional neural network
Image representations
Multi-scale feature fusion
Feature aggregated
Weight constraints

ABSTRACT

In this paper, we aim to achieve effective image representation for image retrieval in an unsupervised manner. To this end, we propose a fully cross-dimensional weighting pooling (FCroW) method to improve the weight strategy of the cross-dimensional weighting pooling (CroW). More specifically, FCroW weights both the non-zero parts and zero-parts of convolutional layers, aiming to obtain robust image representations. In particular, we aggregate multi-scale features extracted by convolutional neural networks using the proposed FCroW, taking into account multiple aspects of visual features captured by the networks. Different weights can be assigned to the features extracted by different layers of the networks. To reduce the effort for parameter tuning, we propose an initial strategy to prune the searching space of the weights, which is achieved by designing constraint rules based on the prior knowledge on relations between the layers of the networks. Based on this, we propose weighted multi-layer feature fusion for similar image representations. Extensive experiments conducted on four public real-world datasets demonstrate the effectiveness of the proposed FCroW method and the pruning strategy for image retrieval.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, deep learning has made significant breakthroughs in various computer vision tasks. One of the important factors is its powerful ability of learning nonlinear representation, which is beneficial to the understanding of the content of images. Deep convolutional neural networks (CNN) are popularized by the revolutionary work of Krizhevsky et al. [1]. It has shown that CNN can "effortlessly" improve the state-of-the-art in most areas of computer vision [2,3] etc., beating many highly optimized approaches of the traditional machine learning. And in the field of image retrieval, the development is more particularly rapid. Content based image retrieval has received sustained attention over the last decade, leading to rapid expansion in visual instance retrieval [4–8].

Some traditional methods based on local features are mainly the variants of the bag-of-words (BOW) model, and the representative is SIFT [9]. Some others such as soft assignment [10], spatial matching [11,12], query expansion [3,6,13,14], better descriptor normalization [6], feature selection [15,16] and very large vocabularies [17] are difficult to apply on large scale data because of their high cost computation and memory space. But image retrieval based on deep learning does not have this problem, which mainly relies on compact image feature descriptors [18–21]. Most of the works focus on fine-tuning the networks, i.e. initialization by a pretrained neural network, and apply pooling methods to extract features. At present, most of these methods are based on global features, and understand the image and generate global features for image retrieval through the high-level features extracted by CNN.

This work is an extension of our previous work [22] that focused on the Grand Challenge of "AI Meets Beauty" in conjunction with the ACM Multimedia 2018 [23]. In the challenge, the competitors were required to find the beauty product in images captured by users from half a million beauty product images provided by different online shopping sites. To deal with the challenge, we extracted the features of multiple deep models and multiple layers of the individual models, and integrated them to obtain image representations, which achieved the third place with a bronze certificate. In this work, we improve the previous work [22] mainly from three aspects. Firstly, we refine the CroW [18] pooling method by considering the weights of both zero parts and non-zero parts. Secondly, we introduce an initial strategy of weight constraint to assign optimal combinations quickly and propose weighted

* Corresponding authors.
E-mail addresses: wangqi_6414@sina.com (Q. Wang), 1048703768@qq.com (J. Lai), zhengyang5-c@my.cityu.edu.hk (Z. Yang), kaixu.gdut@foxmail.com (K. Xu), ppkanggdut@126.com (P. Kan), liuwy@gdut.edu.cn (W. Liu), leiliang@gdut.edu.cn (L. Lei).

https://doi.org/10.1016/j.neucom.2019.08.025
0925-2312/© 2019 Elsevier B.V. All rights reserved.

multi-layer feature fusion for similar image representations. Thirdly, we evaluate the performance of the proposed method on four public data sets for image retrieval.

More specifically, we propose a fully cross-dimensional weighting pooling (FCroW) method to weight both non-zero parts and zero parts, where the zero parts can increase the region of interest and the non-zero parts can restrict the oversize weights of the non-zero. Therefore, FCroW aggregates convolutional features layer through a combination of two different channel weights, and gets stronger representation of image features. Furthermore, we take the advantage of multi-convolution feature layer fusion to integrate the multi-aspects of image features. A strategy of constraint is introduced to assign the weights of the features extracted from different layers of the deep networks effectively. Extensive experiments conducted on four public real-world data sets demonstrate the effectiveness of the proposed FCroW method and the strategy of assigning weights to the multi-aspect feature. In summary, the main contributions of this work are as follows.

1) We extend CroW pooling by taking into account both the zero parts and non-zero parts in channel weighing, which can improve the spatial response to get comprehensive features.
2) We propose a weighted multi-feature fusion method and a weighting strategy to combine different feature layers for image representations, which are more robust than using single-layer feature merely.
3) We conduct extensive experiments on four public data sets, and the experimental results demonstrate the effectiveness of the proposed FCroW pooling for image retrieval.

The paper is structured as follows. In Section 2, we discuss the related work on image retrieval. Section 3 presents a general framework and the proposed method in details. Section 4 shows the experimental results. Section 5 concludes the paper.

2. Related work

The approaches of learning image representations can be divided into two categories [21]: BoF-based approaches [24] and CNN-based approaches. Conventional image retrieval methods were conducted under the framework of BoF [24] combined with the SIFT feature or other image features like SURF, DoG [9], MAER [25], Hessian affine [26], etc. Recently, with the success of CNN in a wide range of areas, there are increasing attentions to utilizing CNN features for image retrieval. This section reviews the related works of both categories.

2.1. The BoF-based methods

Typical BoF-based methods extracted SIFT features from images, and constructed codebook for encoding, which had been predominantly studied before 2012 [1]. For instance, Jégou et al. [24] introduced a graph-structured quantizer, which significantly speeded up the assignment of the descriptors to visual words on large scale image search, based on the bag-of-features approach in the framework of approximate nearest neighbor search. And he also proposed a simple yet efficient way of aggregating local image descriptors into a vector of limited dimension, and showed how to jointly optimize the dimension reduction and the indexing algorithm on a very large-scale dataset [27]. Perronnin et al. [28] proposed to use an alternative fisher kernel framework, and compressed fisher vectors to reduce their memory footprint, which speeded the retrieval for the problem of large-scale image search. And he also used small codebooks for fewer than several thousand visual words, and compact vectors were generated before dimension reduction and coding [5]. Philbin et al. [29] exploited the sparse BoW histograms, the inverted index and memory-friendly signatures. The BoF-based methods [10,30,31,32] have been the dominated approaches on image retrieval tasks in the past a few decades. However, most of them usually perform not well when dealing with large-scale data.

2.2. The CNN-based methods

For CNN-based methods, pre-trained, fine-tuned CNN models and hybrid methods are often used in the process of image search. And fixed-length compact feature vectors are usually produced by pooling or some other measures of convolution layer. CNN-based [3,33] image representations take the advantage of deep learning and can be seemed as global features, which achieve remarkable performance on image retrieval [34].

In terms of the methodologies, the deep-learning based image retrieval methods can be divided into supervised ones and unsupervised ones. More specifically, end-to-end [13,36–38] models are typical supervised approaches. For instance, GEM pooling [35] was proposed to fine-tune CNNs for image retrieval on a large collection of unordered images. However, the supervised methods require a large number of annotated data for training. In terms of the unsupervised approaches, they usually extract the outputs of fully connected layers or the convolution layers and use pooling strategies to obtain image representations. As our work can be seen as the second category, we investigate more approaches in unsupervised manners.

More specifically, a number of pooling strategies have been proposed, such as global max pooling, global average pooling, CroW pooling [18], r-mac pooling [19], etc. In the context of content-based image retrieval, a number of approaches directly use pooling strategies to obtain image features and perform image search successfully. For instance, Razavian et al. [39] divided the original images into many patches, and then extracted the features from the segmented image patches. Finally, the feature vectors extracted from all patches of the image were put together for post-processing. Babenko et al. [40] aggregated local deep features to produce compact global descriptors for image retrieval. Kalantidis et al. [18] mainly proposed a direct and effective image representation method, which was based on the cross-dimensional weighting and aggregation of deep convolutional neural network layer outputs. Tolias et al. [19] built compact feature vectors that encoded several image regions without the need to feed multiple inputs to the neural network, and proposed an R-MAC pooling method for image retrieval.

In terms of the dataset, the Oxford buildings [31], Paris [10] and Holidays [12], etc., are representative data sets for image retrieval. For the Perfect-500 K dataset [23] released by the ACM Multimedia 2018, the discrepancy among different classes is not as distinct as the previous data sets. The beauty products in the Perfect-500 K dataset are harder to distinguish. Therefore, not only the global features are expected to be used to distinguish the inter-class instances, but also the local features are needed to distinguish the intra-class instances [41–43].

3. Methodology

This section is divided into three parts. Section 3.1 is the framework of our approach. Section 3.2 introduces our fully cross-dimensional weighting pooling (FCroW). Section 3.3 presents the strategy of weighted multi-layer feature in details.

3.1. The framework of our approach

The framework of our approach is shown in Fig. 1. In the offline stage, we use a general image retrieval framework and pre-trained VGG16 [20] CNN model on the ImageNet. The CNN model can also

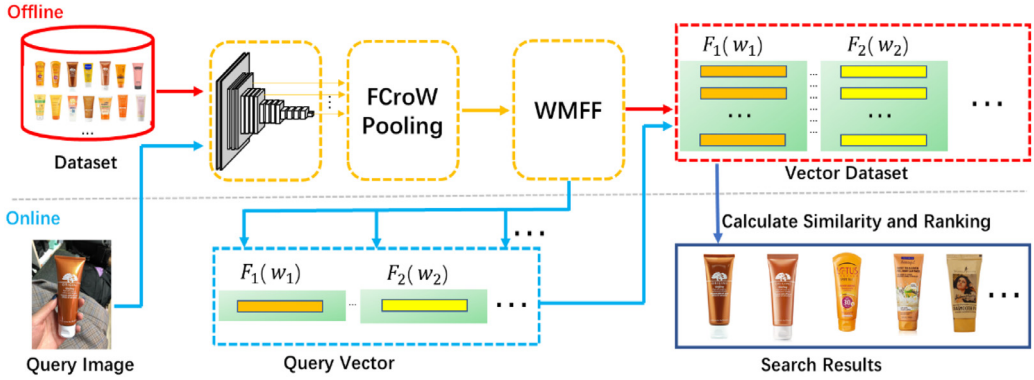


Fig. 1. The framework of our proposed approach. Offline stage: parameters of convolutional layers trained on the ImageNet dataset are transferred as the initial settings for our model. We propose the FCroW pooling method to extract features from the selected feature layers. Furthermore, we use weighted multi-layer feature fusion (WMFF) for image representation. Online stage: search result is obtained by measuring the cosine distances between the query image and images in the vectorized dataset.

be pre-trained on other deep learning models for image feature extraction without fine-tuning. Then, we extract different convolution layers by using our FCroW pooling method. The different features are merged to obtain the final feature vector for an image, and are assigned corresponding weights according to different CNN layers. Thus, we obtain the vectorized dataset. In the online stage, given a query image with a vector representation, we use the cosine similarity to measure the distances compared to the images in the dataset. Finally, we rank the images in the dataset according to the similarity scores to obtain similar ones as the result of output.

3.2. Fully cross-dimensional weighting pooling (FCroW)

In this section, we propose the fully cross-dimensional weight (FCroW) pooling to improve the CroW [18] method by taking into account the non-zero parts. The CroW pooling method increases the weight of the object area and reduces the weight of the non-object area by weight self-adaptive approach. This method constructs the spatial weight and the channel weight to aggregate feature map, and can increase the region of interest to a certain extent and restrain the weight of the non-object region. Based on CroW, we propose to aggregate spatial weight and new channel weight to get an improved pooling method, which consists of three main steps, i.e., spatial response, channel response, and calculating image vectors. Intuitively, the overview of FCroW pooling is shown in Fig. 2. In particular, we aggregate the (n, n, k) convolutional layer into a $(1, k)$ feature vector.

1) *Spatial response:* we mainly calculate spatial and channel response for spatial weight and channel weight. Firstly, the spatial weight is calculated directly by accumulating the feature maps of each channel. The superimposed spatial weight can be considered as a saliency map. Furthermore, we can calculate the spatial response for spatial weight as follows.

$$S_{ij} = \left(\frac{S'_{ij}}{\left(\sum_{m,n} {S'_{mn}}^{p} \right)^{1/p}} \right)^{1/2} \tag{1}$$

where S_{ij} is the spatial weight, and S' is the sum of all feature maps. (i, j) and (m, n) are spatial location label. $p = 2$ represents the $L2$ norm.

We use a computational approach similar to the CroW method to obtain spatial response, and its response reflects the approximate location of the target in the feature layer. The spatial response mainly superimposes all the feature maps to obtain S. Then calculate the spatial response map according to the spatial weight as in Eq. (1).

2) *Channel response:* In the calculation process of channel weight, CroW calculates the weight of channel weight by using TF-IDF. If the feature map data is not zero, the image can be considered as containing key image information, and can be used for representation for retrieval. However, the part of null range is obtained by pooling of the previous convolution layer. In the process of upsampling, it also contains part of the information of the image, which may provide critical details for image retrieval between two extremely similar classes. While in the process of CroW pooling and convolution, the image data information of this part is lost step by step. Therefore, we propose to calculate the channel response for channel weight as follows.

$$C_j = \log\left(\frac{\sum_{i=1}^{k} x_i}{x_j} + \frac{y_j}{\sum_{i=1}^{k} y_i} \right) \tag{2}$$

where C_j is the channel weight, and i is the label of channel. k is the number of channel in Fig. 2. x_j is the sum of non-zero responses of feature map, and y_j is the sum of zero responses of feature map.

We have added weights of the part of null range based on the original foundation, and these weights can restrict the oversize weights of the non-zero. This combination could aggregate convolutional layers of image feature better for more robust image representations. The channel response mainly expands each channel feature layer and calculates the channel response vector according to the channel weight in Eq. (2).

3) *Calculating image vectors:* The aforementioned weight and channel responses are similar to the attention mechanism in the model, aiming to find the dominant response and the key features. We calculate the above two responses and the original feature map to obtain the feature vector of the image. The original feature map is multiplied by the channel response map and the spatial response map to obtain the final feature vector. Therefore, we obtain a fixed dimension vector from the convolution feature map. Furthermore, we normalize the eigenvectors by L2-normalize for similarity calculation.

In this section, we improve the CroW pooling method. Spatial weighting tends to boost locations response of images compared to other spatial locations response of the similar images. These responses are more discriminative benefiting from the diversity of the features of different layers in the neural network. Channel weight tends to boost the whole object response, which gives a numerical indicator of the convolutional feature layer of the

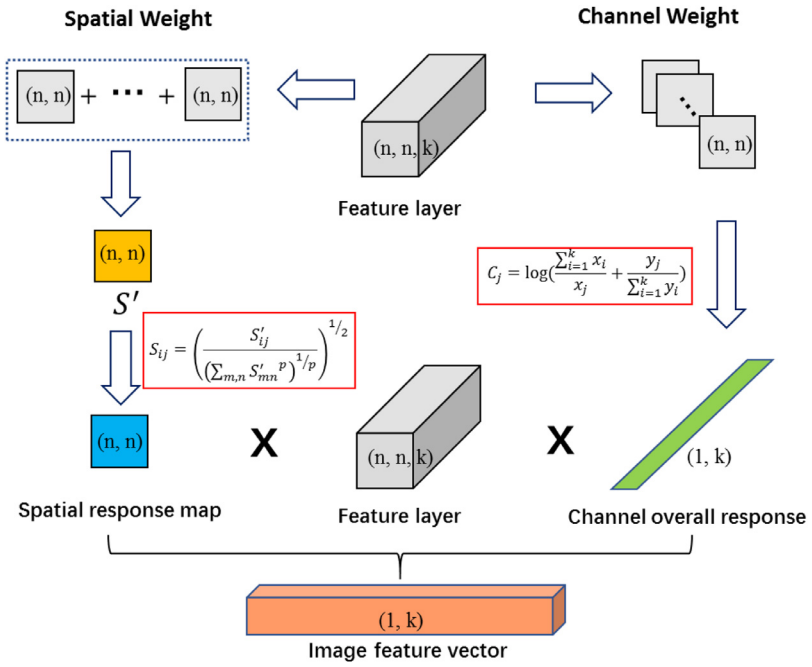


Fig. 2. The steps of the FCroW method: pool a feature layer of (n, n, k) to obtain a feature vector of $(1, k)$. We calculate the spatial response and the channel response separately for the feature layer, and then perform data point multiplication with the original feature layer to obtain the image feature vector.

image object regions. Our FCroW pooling method combines both responses to obtain the final eigenvector.

3.3. Weighted multi-layer feature fusion (WMFF)

1) *Multi-layer feature fusion.* Based on pooling methods, the convolution layers generate their corresponding feature maps, which can be used for feature representation. The FCroW pooling method assigns the weight ratio of the region where the target details are located, and can express the features of images well. For the unsupervised methods, most of them use only one convolutional layer (usually the convolutional layer close to the end rather than the fully connected layer) of the neural network for feature extraction. In fact, the former convolutional layers represent the specific information of the images, while the subsequent fully connected layers express the abstractive information of the images [18,19,44,45]. Therefore, it is necessary to combine the different layers for image representation, making the individual feature layers complement with each other.
2) *Weight constriction of feature layer.* The features extracted from different layers express different aspects of images, thus they are various in terms of the performance being used for image retrieval. Therefore, we propose a weighted multi-layer feature fusion (WMFF) to weight the importance of the features extracted from different layers. More specifically, we firstly select some specific convolutional layers to extract the feature maps. Furthermore, we assign different weights to the features extracted from different layers by introducing some strategies to reduce the weight space. Finally, we fuse these weighted features to get the final feature representation. In particular, as ZFNet [46] shows, the higher the network layer is, the more abstract the content is, and the more specific things can be expressed. We propose an intuitive constraint to quickly assign the weights as follows.

$$w_n * s_n \geq w_{n-1} * s_{n-1} \geq \ldots \geq w_2 * s_2 \geq w_1 * s_1 \tag{3}$$

$$w_1 + w_2 + \ldots + w_n = 1 \tag{4}$$

where w_n is a weight for the nth layer, s_n is the sum of the distances between the query image and the images in the dataset on the n-th feature map.

In the context of image retrieval [21,44], it is shown that the latter feature layers tend to perform better than the former feature layers. The constraint can be other forms but it is supposed to follow this principle. The constraint indicates that the feature information of the later layers is more important than the previous layers for image retrieval.

3) *Discussions and an example.* We give a detailed explanation about the weight constriction in Eqs. (3) and (4). Without loss of generality, let $F = \{f_1, f_2 \ldots f_n\}$ denote a feature vector, where n is the number of feature layers. $\{d_1, d_2, \ldots d_n\}$ are the dimensions corresponding to feature layers, where the dimension d_n is usually an integer power of 2. Given the extracted feature vectors, assume that each image in the dataset can be expressed as $\{P_1, P_2, \ldots P_n\}_{num}$, where P_n is the nth feature vector, *num* is the number of images in the dataset.

To retrieve relevant images for a query Q, we can calculate the cosine distances between Q and the images in the database by the following equation.

$$S = P \cdot Q/(|P| \cdot |Q|). \tag{5}$$

In terms of the distances on the features extracted from each layers, the distance between each feature layer of Q and P_n can be calculated as follows.

$$s_n = P_n \cdot Q/|P_n||Q| \tag{6}$$

In practical terms, Eq. (6) measures the degree of similarity between the various levels and the query through a specific quantitative relationship. Specifically, the formula can calculate the similarity distance between the specific number of layers of the CNN model and a given query.

Let $\{s_1, s_2 \ldots, s_n\}$ denote the distance on each feature layer, and $\{w_1, w_2, \ldots w_n\}$ denote the weights being assigned to each distance on the layers, where $w_1 + w_2 + \ldots + w_n = 1$. Therfore, we can calculate resulting distance between the query image and ith image of dataset by

$$Sum_i = w_1 * s_1 + w_2 * s_2 \ldots + w_n * s_n \quad (7)$$

Finally, the matching result of the query is determined by ranking the images according to all the distances $\{Sum_1, Sum_2, \ldots, Sum_i\}$.

In this section, we first propose the weighted multi-layer feature fusion for more complementary image representation. The weight constraint of the feature layer is given to obtain a better weight value for the similarity calculation. Finally, an example shows that the weight ratio can play a vital role in image retrieval.

4. Experiments

4.1. Task and datasets

(1) Task

For the actual image retrieval in general sense, the search datasets and the query datasets are given. For each single query, we need to return all relevant images in the dataset by calculating their similarities. By check the label consistency between query and retrieved images, we can get the precision performance. MAP is the mean average precision on all the queries.

(2) Datasets

We evaluate our method on four public data sets for image retrieval, including Oxford [31], Paris [10] and Holidays [12], and Perfect-500 K [23]. For Oxford and Paris, we use crop query by following existing image retrieval methods [18,19,35,40], while we search directly on Holidays.

1) *Oxford Buildings* [31]**.** It consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated for 11 different landmarks, each of which provides 5 queries, resulting in a set of 55 queries in total.
2) *Paris* [10]. It consists of 6412 images collected from Flickr by searching for particular Paris landmarks. There is a set of 12 query images. The structure of this data sets is almost the same as Oxford Buildings, and the main difference is that the content is about buildings in Paris.
3) *Holidays* [12]. The Holidays dataset is a set of images which mainly contains some of our personal holidays photos. The images were taken with multiple variations, such as rotations, viewpoint and illumination changes, blurring, etc., which can be used to test the robustness. The dataset includes a very large variety of scene types (natural, manmade, water and fire effects, etc.) and images are in high resolution. The dataset contains 500 image groups, each of which represents a distinct scene or object. The first image of each group is the query image and the expected retrieval results are the other images in the group.
4) *Perfect*-500 K [23]. It is a collection of beauty and personal care items from 14 popular e-commerce sites, including Amazon (USA, India), Cult Beauty, Flipkart, Galleria, Gmarket, JD.COM, Nordstrom, Sephora, Strawberrynet, Target, Walgreens, Walmart and Yahoo Shopping Mall. There are 520,000 beauty images. The query images are taken by users in real-life, and there are about 100 query images. The dataset is released by the Grand Challenge of ACM Multimedia 2018.

Table 1
MAP performance of adopting different distance metrics in our FCroW.

	Perfect-500K	Oxford5K	Paris6K	Holidays
KL-divergence	0.209	0.729	0.776	0.686
Euclidean distance	0.226	0.764	0.801	0.710
cosine distance	0.279	0.868	0.887	0.791
normalized correlation	**0.286**	**0.878**	**0.895**	**0.801**

4.2. Performance metrics

We evaluate the performance of the approaches by mean average precision (MAP) over all queries on Oxford, Paris and Holidays data sets. As is standard practice, in Oxford and Paris one uses only the annotated region of interest of the query, while for Holidays one uses the whole query image. And we use MAP-7 for the Perfect-500 K dataset by following the metric used by the Grand Challenge of ACM Multimedia 2018 [23]. In this dataset, there are many categories, and one query image may correspond to only a few expected images. MAP on top-7 results is adopted as the evaluation standard officially.

4.3. Implantation details

In the context of image retrieval, there are some strategies being widely adopted, such as query expansion [14,50], Whitening, etc. We also introduced them by following the existing works.

1) *Query Expansion.* We select the top-K retrieved images, and aggregate their features as well as the query feature by average. We use the aggregated feature as a new query and perform the retrieval again to obtain the final ranking list. The extended query data can be the mean value of the top-n results of the first query [14,51]. Query expansion alleviates the uncertainty of a single query, resulting in more robust queries.
2) *Whitening.* Whitening is usually used to reduce the feature dimension for computation. In the context of image retrieval, not all the pooling methods can benefit from whitening. In our experiments, we find SPOC [40] and RMAC [819] are better to be combined with whitening, while CroW [18] and the proposed FCroW are not.

4.4. Selection of network and feature layers

In terms of the network architectures for feature extraction, FCroW features are derived from the VGG16 model with weights pretrained on ImageNet. We have tried some other advanced networks, such as GoogleNet [47], ResNet [48], InceptionResNetV2 [49], and VGG network [20] is more effective for the current tasks. Compared with the VGG network, these networks are more cumbersome in selecting feature layers, which impacts the performance of image retrieval.

In terms of the selection of feature layers, we select all the convolution layers before the network pooling layers to extract the image features. Different feature layers can be helpful to express comprehensive aspects of images. Like existing methods [18,19,40], we use conv5 of VGG net as a baseline. Furthermore, we extract other convolutional layers including conv1 to conv4 and merge them together for multi-feature fusion.

4.5. Similarity measurement

Generally speaking, similarity measurements have an influence on the final performance. As shown in Table 1, we report the performance of different measurements on the current tasks, such

Table 2
Time costs of two similarity measurements.

	Perfect-500K	Oxford5K	Paris6K	Holidays
cosine distance	23,000 ms	190 ms	210 ms	49 ms
Normalized correlation	67,000 ms	660 ms	710 ms	163 ms

Table 3
Performance of FCroW by assigning different weights to the features on Perfect-500 K.

No.	conv1	conv2	conv3	conv4	conv5	Weight ratio	Dimension	MAP
1	0.20	0.20	0.20	0.20	0.20	1: 1: 1: 1: 1	1×1472	0.302
2	0	0	0	0	1	0: 0: 0: 0: 1	1×512	0.267
3	0.125	0.125	0.25	0.25	0.25	1: 1: 2: 2: 2	1×1472	0.355
4	0.10	0.10	0.20	0.20	0.40	1: 1: 2: 2: 4	1×1472	0.364
5	0.08	0	0.42	0.16	0.34	1: 0: 5: 2: 4	1×1344	0.393
6	0.077	0.077	0.384	0.154	0.308	1: 1: 5: 2: 4	1×1472	**0.419**

as KL-divergence, Euclidean Distance, Cosine Distance, and Normalized Correlation. In this experiment, the network we used is VGG16, and the conv3-3, conv4-3, and conv5-3 layers are used to extract features. Its weight ratio is 1:1:1. We fuse the features extracted by the three layers with equal weights for simplicity, and the resulting dimension of the feature is 1280.

From Table 1, we can make two observations: 1) In terms of the distance metrics, normalized correlation and cosine distance perform better than Euclidean distance and KL-divergence for image retrieval. The performance of using cosine distance and normalized correlation are quite close to each other. 2) Normalized Correlation performs the best on the current task.

Furthermore, we report the time costs of FCroW using normalized correlation and cosine distance on the four datasets in Table 2. From the table we can see that using cosine distance is more efficient. Overall, normalized correlation is the most effective measurement, while cosine distance is quite efficient.

4.6. Effectiveness of weighted multi-layer feature fusion (WMFF)

The feature maps extracted from different CNN layers conveys image information in different levels. The former ones express specific image characteristics, while the latter ones express abstractive image characteristics. For the images that are quite similar to a query image, especially the ones showing different items with the query, image features extracted from single layers cannot make sure the vital information has been included, thus it may not be enough to differentiate them. In the context of unsupervised learning, there is not any prior knowledge that is available for choosing effective features. Therefore, using the combinations of the features tends to be more robust, and it comes the issue of assigning weights to the multi-scale features according to different layers.

1) *On the Perfect-*500k *dataset*. In Table 3, we report some representative results of assigning weights to the layers according to the constraint in formulas (3) and (4). Following formulas (3) and (4), we have selected a part of the representative configuration by control variable method in a narrow limited domain. We try to adjust parameter of weights for the readability of the paper, and explain the validity of multi-layer weight fusion. Through the above principles, we obtain the results of Table 3. We follow the selected principle of Section 4.4, extract the features of each convolution layer, and obtain the following results. The weight ratios are as shown in Table 3. No. 1 is our preliminary experiments, and others is fine-tuned on the baseline. From the table, we can conclude some observations. 1) By comparing No.1 and No. 2, it shows that using multiple feature layers are more effective than using single feature layer. 2) By comparing No. 3 and No. 4, we can see that assigning large weight to latter layers tend to be more effective, which indicates the abstractive features are more discriminative. 3). By comparing No. 5 and No. 6, we can conclude that the former layers cannot be neglected totally, as it can be quite useful with appropriate weight being assigned.
2) *On the Holidays dataset*. For illustration, we take the Holidays dataset as an example to show the effectiveness of WMFF. Fig. 3(a) shows the distribution of performance under the weight constraint of $x + y + z = 1$. Fig. 3(b) is the distribution of performance in the finite field of WMFF as introduced formulas (3) and (4) in Section 3.3. Following the formula (4), we can find s_n of the corresponding layer at the same time. Fixing s_n in the formula (3), we can get the range of xyz in Fig. 3(b). In particular, we choose three convolution layers, i.e., {x: conv5-3, y: conv4-3, z: conv3-3}. The coordinate axes X, Y, and Z represent the weights of their respective parts. We choose the last three layers because they are more capable to represent the image features and give stereoscopic observability and contrast.

Fig. 3(a) shows that different weights being assigned to the layers result in different performance on the current task. In particular, when the proportion of Z axis (i.e. the weights for the former feature layers) increases, the MAP will decrease sharply, which coincides with the intuition that the latter convolutional layers are more effective than the former convolutional layers. Fig. 3(a) includes many of the statistical results of the different convolutional layers. Comparing the distribution between the depth of colors, we can directly find that different feature layers have complementarity, which can also indirectly expresses the significance of different weights.

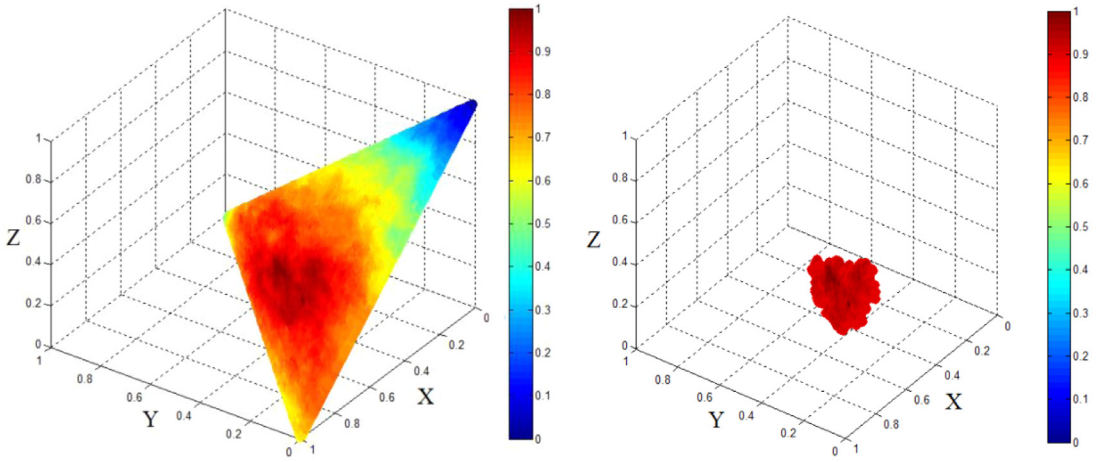
As a result, we propose the strategy of weighted multi-layer feature fusion (WMFF) to find the best combinations of the layers with appropriate weights as shown in Fig. 3(b). By comparing (a) and (b), we can see that the MAP region of (b) almost approximates the maximum MAP region of (a), indicating that the strategy of WMFF is quite effective to assign most effective weights to the different layers for the current task.

Intuitively, conducting pooling operations may lose some data information inevitably, while some lost information may be useful to distinguish the differences between images. In other words, there are risks of losing critical information for image retrieval. Therefore, we propose the strategy of WMFF to give high weights to the important layers and low weight to the supplementary layers, in order to obtain robust and strong image representations.

4.7. Comparison with the state-of-the-art unsupervised pooling methods

As the proposed FCroW is in an unsupervised manner, we compare our method with the state-of-the-art unsupervised pooling strategies for the sake of fairness. The baselines include SPOC [40], CroW [18] and RMAC [19], and all of them adopt the VGG16 networks. We take conv3-3, conv4-3 and conv5-3 layers from the VGG network to extract image features, and evaluate the performance of the pooling strategies, respectively. The experimental results are shown in Fig. 4, where SUM indicates that the three layers are fused by following the weight ratio 1:1:1.

From the figures, we have some observations. Firstly, the proposed FCroW performs well in term of using single image features extracted from different layers in most cases on the four data sets. The experimental results show the effectiveness of the FCroW method. Secondly, all the pooling methods combined with the strategy of multi-scale feature fusion achieve significant



(a) Following the constraint in Equation (4) (b) Following the constraint in Equations (3) and (4)

Fig. 3. The MAP distribution of the Holidays dataset by using FCroW.

	FCroW	CroW	SPOC	RMAC
conv3-3	**0.094**	0.052	0.05	0.041
conv4-3	**0.169**	0.146	0.131	0.132
conv5-3	**0.208**	0.194	0.184	0.176
SUM	**0.279**	0.218	0.201	0.196

(1) Perfect-500K

	FCroW	CroW	SPOC	RMAC
conv3-3	**0.298**	0.211	0.223	0.169
conv4-3	**0.616**	0.601	0.398	0.324
conv5-3	**0.76**	0.749	0.531	0.669
SUM	**0.791**	0.768	0.726	0.721

(2) Oxford5k

	FCroW	CroW	SPOC	RMAC
conv3-3	**0.357**	0.254	0.297	0.249
conv4-3	0.692	**0.714**	0.601	0.584
conv5-3	0.831	**0.848**	0.703	0.83
SUM	**0.878**	0.857	0.809	0.866

(3) Holidays

	FCroW	CroW	SPOC	RMAC
conv3-3	**0.418**	0.298	0.334	0.124
conv4-3	**0.703**	0.667	0.668	0.392
conv5-3	**0.867**	0.855	0.802	0.852
SUM	**0.897**	0.866	0.877	0.886

(4) Paris6k

Fig. 4. The performance of different pooling methods.

improvement on the performance, outperforming the methods using single image features. The experimental results demonstrate the effectives of feature fusion.

In addition, comparing the data of each group in Fig. 4, the performance is better when higher convolutional layer is used. This is because the high-level convolutional layer has rich semantic information and can perform image representation from a global perspective. In contrast, the performance of underlying conv3-3 is not outstanding. Anyway, this layer can still obtain certain image feature information, which provides a basis for multi-scale fusion.

4.8. Performance on the grand challenge of ACM multimedia

We show the top-5 performance of the teams participating in the Grand Challenge of ACM Multimedia 2018 in Table 4. The team achieving the first place does not release their method. The GAN [51] team proposes a regional maximum activations of convolutions with attention (RA-MAC) descriptor to extract image features for retrieval. Our proposed FCroW achieves the third place on the challenge. In particular, on the verification set, we are able to get a MAP of 0.4394. Due to the labels of the test set having not been released, we cannot report the newest performance accordingly.

Table 4
The performance ranking of the grand challenge of ACM multimedia [52].

Rank	Team name	Score (MAP)
1	VCA	0.578
2	GAN [51]	0.291
3	WISLAB (FCroW) [22]	0.271
4	CISIP [53]	0.252
5	Ivlab2017	0.229

4.9. Quantitative evaluations

In Table 5, we report the performance of the pooling approaches combined with the proposed weighted multi-layer feature fusion (WMFF) strategy, whitening, and query expansion (QE) on the three data sets. Since the feature layer number and weight ratio of NO.6 in Table 3 are the best, the following methods use this configuration, and both can achieve better results than

Table 5
The best performance of the pooling methods on different datasets.

Method	Dim	Perfect500K	Oxford5K	Paris6K	Holidays
FCroW+ WMFF	1472	0.388	0.789	0.775	0.827
FCroW+ WMFF + QE	1472	**0.439**	**0.829**	**0.805**	**0.907**
CroW+ QE [18]	512	0.218	0.749	**0.848**	0.871
CroW+ QE + WMFF	1472	**0.322**	**0.849**	0.796	**0.899**
RMAC+ QE [19]	512	0.201	0773	**0.865**	0.889
RMAC+ QE + WMFF+ Whitening	1024	**0.292**	**0.795**	0.771	**0.918**
SPOC+ QE [40]	512	0.196	0.681	0.782	0.839
SPOC+ QE + WMFF+ Whitening	1024	**0.261**	**0.731**	**0.801**	**0.855**

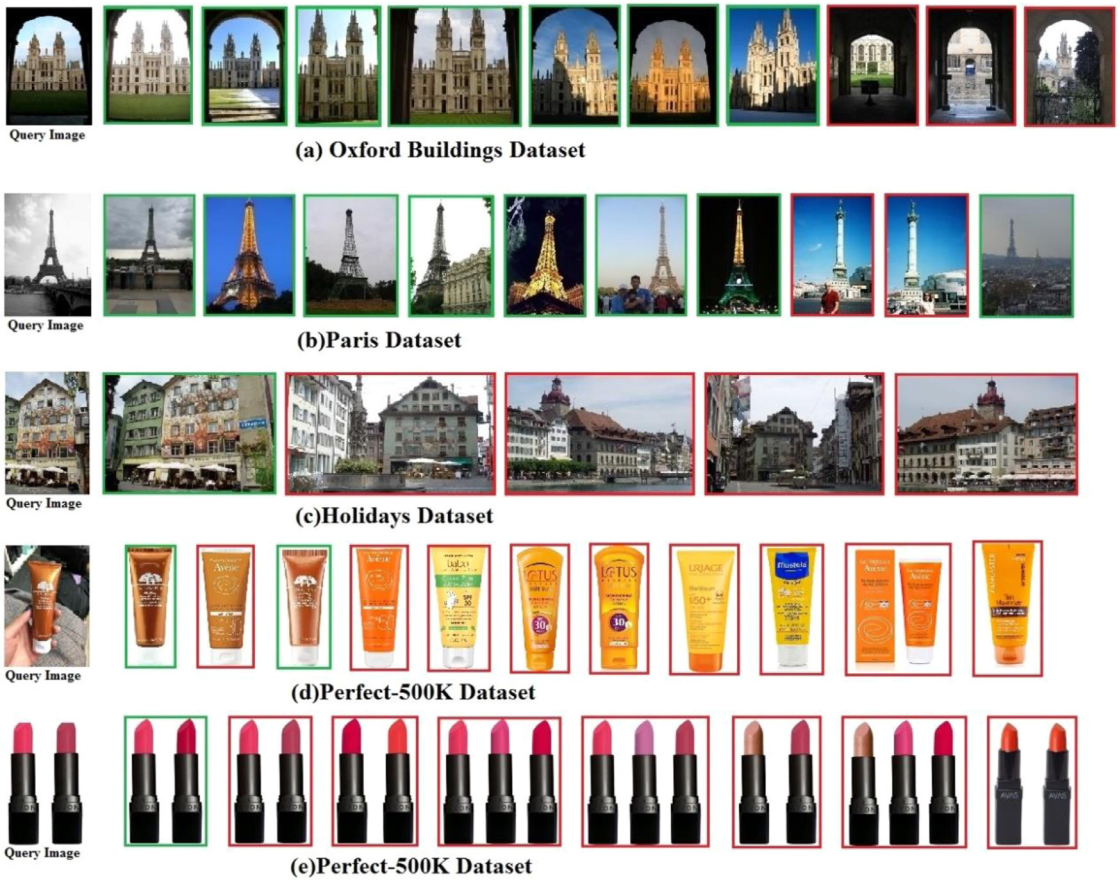


Fig. 5. Image retrieval examples. The images in green box are correct results, and the ones in red box are wrong.

before. From the experimental results, we can conclude some observations. 1). In terms of the FCroW, it can obtain better performance by introducing both WMFF and QE on all the data sets. 2). For most of the pooling methods, using more strategies like WMFF, QE, and Whitening tends to achieve better performance.

4.10. Qualitative evaluations

Intuitively, we give some result examples of the proposed FCroW on the four data sets as shown in Fig. 4. From the figure we can observe that the correct results usually are in the front of the rankings. In particular, the incorrect results are quite similar to the query images, indicating the difficulty of the image retrieval tasks. In particular, for the incorrect results returned by our FCroW on the Perfect-500 K dataset, it is even hard to judge by human that whether they indeed are mismatched products. It is hard to distinguish the differences between the incorrect images and the query image (Fig. 5).

5. Conclusion

In this work, we propose a cross-weighting pooling method to extract image features named as FCroW, which inherits the advantages of CroW. In addition, we propose a weighted multi-layer feature fusion (WMFF) strategy to combine image features extracted from different layers in deep networks for image retrieval. In particular, we introduce a constraint to quickly assign the weights of the features from different layers, which benefits to some classical pooling methods as well. The experimental results on four public data sets demonstrate the effectiveness of the proposed FCroW pooling and WMFF strategies.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that

could be construed as influencing the position presented in, or the review of, the manuscript entitled fImproving Cross-dimensional Weighting Pooling with Multi-scale Feature Fusion for Image Retrieval". Signed by all authors as follows: Qi Wang, Jinxiang Lai, Zhenguo Yang, Kai Xu, Peipei Kang, Wenyin Liu, Liang Lei.

Acknowledgments

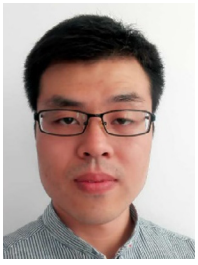
This work is supported by the National Natural Science Foundation of China (No.61703109, No.91748107, No.500160134), China Postdoctoral Science Foundation (No.2018M643024), the Guangdong Innovative Research Team Program (No. 2014ZT05G157), and Entrusted Projects by Enterprises and Institutions (No.18HK0336).

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, NIPS, 2012, doi:10.1145/3065386.
[2] R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587, 2014.
[3] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2014) PP(99):1-1, doi:10.1109/TPAMI.2016.2572683.
[4] H. Azizpour, A.S. Razavian, J. Sullivan, A. Maki, S. Carlsson, From generic to specific deep representations for visual recognition, in: CVPR, 2015, pp. 36–45, doi:10.1109/CVPRW.2015.7301270.
[5] F. Perronnin, T. Mensink, Improving the fisher kernel for large-scale image classification, in: ECCV, 115, 2010, pp. 143–156, doi:10.1007/978-3-642-15561-1_11.
[6] R. Arandjelovic, A. Zisserman, Three things everyone should know to improve object retrieval, in: CVPR, 6, 2012, pp. 2911–2918, doi:10.1109/CVPR.2012.6248018. 1.
[7] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, K. Keutzer, Densenet: Implementing efficient convnet descriptor pyramids, 2014 arXiv preprint arXiv:1404.1869.
[8] X. Yang, X. Gao, B. Song, et al., Aurora image search with contextual CNN feature, Neurocomputing 281 (2018) 67–77 3. 15, doi:10.1016/j.neucom.2017.11.059.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110 doi:10.1023/B:VISI.0000029664.99615.94.
[10] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Lost in quantization: improving particular object retrieval in large scale image databases, CVPR, 8, 2008, doi:10.1109/CVPR.2008.4587635.
[11] Y. Avrithis, G. Tolias, Hough pyramid matching: speeded-up geometry re-ranking for large scale image retrieval, Int. J. Comput. Vis. 107 (1) (2014) 1–9 3.1, doi:10.1007/s11263-013-0659-3.
[12] H. Jegou, M. Douze, C. Schmid, Hamming embedding and weak geometric consistency for large scale image search, in: ECCV, 10, 2008, pp. 304–317, doi:10.1007/978-3-540-88682-2_24. 12.
[13] O. Chum, A. Mikulik, M. Perdoch, J. Matas, Total recall II: query expansion revisited, in: CVPR, 20, 2011, pp. 889–896, doi:10.1109/CVPR.2011.5995601.
[14] G. Tolias, H. Jégou, Visual query expansion with or without geometry: refining local descriptors by feature aggregation, Pattern Recognit. 47 (10) (2014) 3466–3476, doi:10.1016/j.patcog.2014.04.007.
[15] G. Tolias, Y. Kalantidis, Y. Avrithis, S. Kollias, Towards large-scale geometry indexing by feature selection, Comput. Vis. Image Underst. 120 (2014) 31–45, doi:10.1016/j.cviu.2013.12.002.
[16] P. Turcot, D.G. Lowe, Better matching with fewer features: the selection of useful features in large database recognition problems, in: ICCV, 27, 2009, pp. 2109–2116, doi:10.1109/ICCVW.2009.5457541.
[17] A. Mikulík, M. Perdoch, O. Chum, J. Matas, Learning a fine vocabulary, in: ECCV, 5, 2010, pp. 1–14, doi:10.1007/978-3-642-15558-1_1. 9.
[18] Y. Kalantidis, C. Mellina, S. Osindero, Cross-dimensional weighting for aggregated deep convolutional features, in: ECCV, 2016, pp. 685–701, doi:10.1007/978-3-319-46604-0_48.
[19] G. Tolias, R. Sicre, H. Jégou, Particular object retrieval with integral max-pooling of CNN activations, 2015 arXiv preprint arXiv:1511.05879.
[20] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition 2014, arXiv preprint arXiv:1409.1556.
[21] L. Zheng, Y. Yang, Q. Tian, SIFT meets CNN: a decade survey of instance retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1224–1244 5. 1, doi:10.1109/TPAMI.2017.2709749.
[22] Q. Wang, J. Lai, K. Xu, W. Liu, L. Lei, Beauty product image retrieval based on multi-feature fusion and feature aggregation, in: ACM MM Conference, 10, 2018, pp. 2063–2067, doi:10.1145/3240508.3266431.
[23] W.H. Cheng, J. Jia, S. Liu, etc., Perfect Corp. Challenge 2018: half million beauty product image recognition, in: https://challenge2018.perfectcorp.com/index.html.
[24] H. Jégou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search, Int. J. Comput. Vis. 87 (3) (2010) 316–336, doi:10.1007/s11263-009-0285-2.
[25] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust wide baseline stereo from maximally stable extremal regions, Image Vis. Comput. 22 (10) (2004) 761–767, doi:10.1016/j.imavis.2004.02.006.
[26] K. Mikolajczyk, C. Schmid, Scale & affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86 doi:10.1023/B:VISI.0000027790.02288.f2.
[27] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, Comput. Vis. Pattern Recognit., 238, 3304–3311, doi:10.1109/CVPR.2010.5540039.
[28] F. Perronnin, Y. Liu, J. Sanchez, H. Poirier, Large-scale image retrieval with compressed fisher vectors, Comput. Vis. Pattern Recognit., Vol. 26, 3384–3391. doi:10.1109/CVPR.2010.5540009.
[29] N. Er, H. Stewenius, Scalable recognition with a vocabulary tree, in: CVPR, 2, 2006, pp. 2161–2168. doi:10.1.1.61.9520.
[30] H. Jégou, M. Douze, C. Schmid, On the burstiness of visual elements, in: CVPR, 20, 2009, pp. 1169–1176, doi:10.1109/CVPR.2009.5206609. 6.
[31] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: CVPR, 17, 2007, pp. 1–8, doi:10.1109/CVPR.2007.383172.
[32] J. Yu, Z. Qin, T. Wan, X. Zhang, Feature integration analysis of bag-of-features model for image retrieval, Neurocomputing 120 (2013) 355–364, doi:10.1016/j.neucom.2012.08.061.
[33] A.S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: CVPR, 2014, pp. 806–813, doi:10.1109/CVPRW.2014.131.
[34] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, ECCV, 2014, doi:10.1007/978-3-319-10584-0_26.
[35] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 6 (2018) 12, doi:10.1109/TPAMI.2018.2846566.
[36] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, NetVLAD: cnn architecture for weakly supervised place recognition, in: CVPR, 2016, pp. 5297–5307, doi:10.1109/CVPR.2016.572.
[37] H. Liu, Y. Tian, Y. Yang, L. Pang, T. Huang, Deep relative distance learning: tell the difference between similar vehicles, in: CVPR, 2016, pp. 2167–2175, doi:10.1109/CVPR.2016.238.
[38] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, DeepFashion: powering robust clothes recognition and retrieval with rich annotations, in: CVPR, 2016, pp. 1096–1104, doi:10.1109/CVPR.2016.124.
[39] A.S. Razavian, J. Sullivan, S. Carlsson, A. Maki, Visual instance retrieval with deep convolutional networks, ITE Trans. MTA (2016), doi:10.3169/mta.4.251.
[40] A. Babenko, V. Lempitsky, Aggregating deep convolutional features for image retrieval, ICCV, 2015, doi:10.1109/ICCV.2015.150.
[41] O. Chum, J. Matas, Large-scale discovery of spatially related images, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2) (2010) 371–377, doi:10.1109/TPAMI.2009.166.
[42] T. Weyand, B. Leibe, Discovering details and scene structure with hierarchical iconoid shift, in: ICCV, 2013, pp. 3479–3486, doi:10.1109/ICCV.2013.432.
[43] J. Philbin, J. Sivic, A. Zisserman, Geometric latent dirichlet allocation on a matching graph for large-scale image datasets, Int. J. Comput. Vis. 95 (2) (2011) 138–153, doi:10.1007/s11263-010-0363-5.
[44] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, in: ICML, 2014, pp. 647–655. arXiv:1310.1531.
[45] J.Y. Wang, Z. Zhu, Image retrieval system based on multi-feature fusion and relevance feedback, in: ICMLC, 4, 2010, pp. 2053–2058, doi:10.1109/ICMLC.2010.5580505.
[46] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: ECCV, 6, 2014, pp. 818–833, doi:10.1007/978-3-319-10590-1_53.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9, doi:10.1109/CVPR.2015.7298594.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.
[49] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI, 4, 2017, p. 12.
[50] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: Automatic query expansion with a generative feature model for object retrieval, in: Proceedings of IEEE International Conference on Computer Vision, 2007, pp. 1–8.
[51] Z. Lin, Z. Yang, F. Huang, J. Chen, Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval, in: ACM MM Conference, 10, 2018, pp. 2073–2077, doi:10.1145/3240508.3266436.
[52] https://mm18grandchallenge.github.io/Perfect/index.html#item11.
[53] J.H. Lim, N. Japer, C.C. Ng, C.S. Huang, J. Chen, Unprecedented usage of pre-trained CNNs on beauty product, in: ACM MM Conference, 10, 2018, pp. 2073–2077.

Qi Wang is currently pursuing the Ph.D. degree with the School of Computer, Guangdong University of Technology, Guangzhou, China. His-current research interests include computer vision and pattern recognition and machine learning.

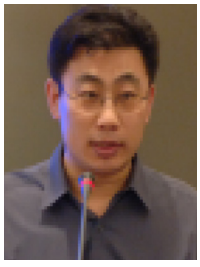
Jinxiang Lai received the B.S. degree in Optoelectronic information science and Engineering from Guangdong University of Technology, Guangzhou, China, in 2017. He is currently pursuing the M.S. degree in the School of Physics and Optoelectronic Engineering, Guangdong University of Technology, China. His-research interests include neural network, deep learning, pattern recognition and computer vision.



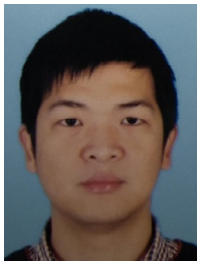
Zhenguo Yang is a Associate Professor at Guangdong University of Technology. He was a Postdoctoral Fellow at Guangdong University of Technology from 2017 to 2019. He received his Ph.D. degree in Computer Science from City University of Hong Kong, in 2017, and the M.E. degree in Computer Science from Zhejiang Normal University, China, in 2013, and the B.E. degree in Computer Science from Shandong Normal University, China, in 2010. His-research interests include online event detection and transfer learning, etc. He has published papers on prestigious venues like IEEE TPAMI, ACM TOIT, and ACM Multimeda, etc.

Kai Xu is currently pursuing the Ph.D. degree with the School of Computer, Guangdong University of Technology, Guangzhou, China. His-current research interests include deep learning and natural language processing.

Peipei Kang received the M.S. degree in School of Computer Science and Technology from Guangdong University of Technology, Guangzhou, China. She is currently pursuing the Ph.D. degree in School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. Her current research interests include pattern recognition and machine learning.



Wenyin Liu is currently a Professor in School of Computer Science and Technology, Guangdong University of Technology. He was Deputy Director of Multimedia software Engineering Research Centre at the City University of Hong Kong from 2013 to 2016, an assistant professor in the Department of Computer Science at the City University of Hong Kong between from 2002 to 2012, and a full time researcher at Microsoft Research China/Asia from 1999 to 2001. His-current research interests include blockchain, anti-phishing, Web identity authentication and management. He has a BEng and MEng in computer science from Tsinghua University, Beijing and a DSc from the Technion, Israel Institute of Technology, Haifa. In 2003, he was awarded the International Conference on Document Analysis and Recognition Outstanding Young Researcher Award by the International Association for Pattern Recognition (IAPR). He had been TC10 chair of IAPR for 2006–2010. He had been on the editorial boards of the International Journal of Document Analysis and Recognition (IJDAR) from 2006 to 2011 and the IET Computer Vision journal from 2011–2012. He is a Fellow of IAPR and a senior member of IEEE.



Liang Lei is currently a Professor in Guangdong University of Technology. He received his Ph.D. degree in Optics in 2007 from SUN YAT-SEN UNIVERSITY. His-research interests include computer vision, machine learning, and Photoelectric vision inspection, etc.