International Journal of Machine Learning and Cybernetics (2019) 10:1863–1873
https://doi.org/10.1007/s13042-018-0898-2

ORIGINAL ARTICLE

Multi-center convolutional descriptor aggregation for image retrieval

Jie Zhu[1] · Shufang Wu[2] · Hong Zhu[3] · Yan Li[4] · Li Zhao[3]

Received: 15 May 2018 / Accepted: 26 November 2018 / Published online: 5 December 2018

Abstract
Recent works have demonstrated that the convolutional descriptor aggregation can provide state-of-the-art performance for image retrieval. In this paper, we propose a multi-center convolutional descriptor aggregation (MCDA) method to produce global image representation for image retrieval. We first present a feature map center selection method to eliminate the background information in the feature maps. We then propose the channel weighting and spatial weighting schemes based on the centers to boost the effect of the features on the object. Finally, the weighted convolutional descriptors are aggregated to represent images. Experiments demonstrate that MCDA can produce state-of-the-art retrieval performance, and the generated activation map is also effective for object localization.

Keywords Multi-center · Descriptor aggregation · Feature map · Feature weighting

1 Introduction

Image retrieval has been evolving rapidly over the last decade. Many existing methods adopt some low level descriptors, and encode them using bag-of-words (BoW) or some others methods. After the seminal work of Krizhevsky [1], deep learning has demonstrated the advantages in many areas of artificial intelligence [2–5].

Many works have applied pre-trained convolutional neural networks (CNNs) models to extract generic features for image retrieval and obtained excellent performances [5–7]. In all these methods, the activations in the convolutional layers or pooling layers which can capture semantic features are used to represent images. Usually there are three steps, first, the descriptors are extracted and selected, and second, these descriptors are aggregated to represent images. Finally, the retrieval results are obtained by calculating the similarities between images. In addition, the activation map, which is generated by summing the feature maps in the same layer, is effective to describe the object region in the image.

Although CNN has been successful applied on image retrieval, a few questions still remain unaddressed. First, the positions of the top few highest responses in a CNN activation map usually correspond to different object regions in an image, and previous work [6] also demonstrated that the positions with the top few highest responses in some feature maps also correspond to the object regions. Therefore, it is questionable whether it is best to use the responses in the feature maps to localize the objects. Second, some methods select a few descriptors and then weight and aggregate them to represent images; however, the background elements in the descriptors are not eliminated. Whether descriptors can be better represented by all the responses or some of the responses across all the channels is also not clear.

To meet these challenges, we propose a simple way of producing image representation via feature map center selection. The proposed multi-center convolutional descriptor aggregation (MCDA) method can localize the object by selecting few high responses in each feature map and also weight the descriptors based on these responses for image representation. Figure 1 demonstrates that MCDA activation map can localize the object more accurately than CNN activation map.

As many previous works, the convolutional descriptors are extracted based on pre-trained CNN model. Extensive experiments were conducted on three challenging

✉ Shufang Wu
Shufang_44@126.com

[1] Department of Information Management, The National Police University for Criminal Justice, Baoding, China

[2] College of Management, Hebei University, Baoding, China

[3] College of Computer Science and Software Engineering, Shen Zhen University, Shenzhen, China

[4] School of Applied Mathematics, Beijing Normal University Zhuhai, Zhuhai, China

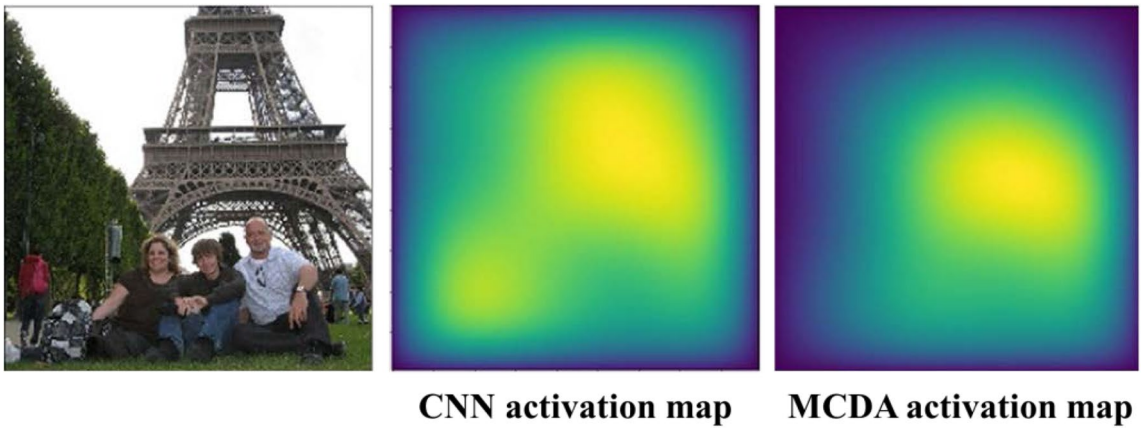

Fig. 1 Comparison between CNN activation map and MCDA activation map. CNN activation map highlights the regions of both the Eiffel and background, while MCDA activation map can describe the Eiffel more accurately

image retrieval dataset, i.e., Holiday [8], Oxford Paris [9], Oxford5K [10] and Oxford 100K [10], and the retrieval experiments verify the effectiveness of MCDA. The major contributions are summarized as follow.

1. We present an effective method to localize the object. Different from most existing methods, which localize the object by using few highest responses in the activation map, MCDA attempts to localize the object by using few highest responses in the feature maps.
2. We present an optimal feature map center selection method based on the descriptor dissimilarity among the high response positions and the spatial relationship among these positions.
3. We present channel weighting and spatial weighting schemes based on the selected feature map centers to boost the contribution of the object features in image representation.

This paper is organized as follows: in Sect. 2 we briefly introduce the related work, while in Sect. 3 we present a Multi-center Convolutional Descriptor Aggregation method to localize the object and represent images. In Sect. 4 we present experimental results for visual search and object localization, and we conclude this paper in Sect. 5.

2 Related work

Machine learning is developing very fast in recent years. Wang [11] studied the relation between generalization and uncertainty of classifiers, and some guidelines are also given for enhancing the generalization ability of classifiers. Wang [12] combined the frequency and segment strategies for splitting the nodes with continuous valued attributes in decision trees. In addition, active learning is a hot topic in machine learning, an ambiguity-based multiclass active learning strategy [13] is proposed for informative unlabeled samples selection. The diversity of the training bag is usually neglected in active learning, in the following work [14], clustering-based diversity and fuzzy rough set based diversity are proposed for bag selection. These works become the theoretical bases of some image classification and image retrieval methods.

Most of the traditional image retrieval methods are proposed based on local features, which are used to construct histogram. These researches mainly focus on the construction of descriptor and the integration of different kinds of information into BoW framework. Other feature aggregation methods such as VLAD [15] and FV [16] can generate compact image representations and have also obtained excellent retrieval results.

Deep learning based models have been widely applied to almost every computer vision related task. Recent studies showed that deep descriptors extracted from pre-trained deep network can be aggregated and have achieved state-of-the-art results in image retrieval. Babenko et al. [17] first investigated the use of neural codes in image retrieval and a similar work of Razavian et al. [18] introduced a basic pipeline of the aggregation of the responses from fully connected layer and convolutional layer for image representation. Ng et al. [7] encoded the convolutional features from different layers into a single vector by VLAD, and the experiments demonstrated that intermediate or higher layers can produce better retrieval results, compared to the last layer. However, the descriptors were aggregated without considering the relationship among local responses. Gong et al. [19] extracted CNN activations for local patches at multiple scale, the activations at each level were pooled by VLAD and concatenated to

represent images. These methods aggregated all the deep features without judging the importance of them, while some works attempted to highlight the deep features in the object region so as to enhance the discrimination of image representation.

The objects tend to be located close to the centers of images, Babenko et al. [20] presented the SPoC descriptor, and the distance between each position and the image center was used to compute the Guassian weight for the descriptor corresponding to the position. Due to the fact that objects can be in any area of an image in reality, activation map is used to localize the objects in some works. Wei et al. [6] discovered that the higher response a particular position is, the more possibility of its corresponding region being part of the object in the activation map, and the positions whose responses are higher than the mean value are considered as the location of the object. Kalantidis et al. [5] applied L2 normalization and power-scaling to activation map for computing the spatial weights, which can boost the features on the object. The large weights were assigned to the positions with high responses in the activation map. These researches utilized the relationship between responses in activation map and object location to assign reasonable weights for all the local descriptors.

It has been discovered that the high response positions in each feature map may indicate object locations, but can also indicate some background locations [6]. Therefore, it is difficult to directly localize the object using the feature maps. Nevertheless, high responses in feature maps are effective for image representation. Maximum response of each feature map can effectively encode an image. Tolias et al. [21] sampled square regions at different scales, and the maximum activations of convolutions (MAC) in these regions were then used for image presentation. Radenović et al. [22] showed that patches corresponding to the MAC vector components have the highest contribution to the pairwise image similarity. However, the question how to select the positions in feature maps to localize the object was not discussed in these papers.

Channel sparsity is an easy and effective way to evaluate the importance of each feature map in image representation, because channel sparsities are highly correlated for images of the same category and less correlated for images of different categories. Kalantidis et al. [5] evaluated the channel sparsity based on the proportion of zero responses of each feature map, however, the difference among non-zeros responses were neglected in this way. Channel sparity was obtained through retraining the network weights in recent works [23], however, this method is hard to be deployed on resource constrained devices. Spatial weighting can be used to evaluate the important of a position. Boscaini et al. [24] used anisotropic heat kernels as spatial weighting functions, and Kalantidis et al. [5] used normalized total response across all channels to compute the spatial weight. The weight computation is mainly dependent on the responses in the feature maps, so it is better to neglect small responses to compute the weights more accurately.

Overall, compared to the works mentioned above, we select few high response positions from each feature map as feature map centers to weight local descriptors and represent images.

3 Approach

We aim to provide a simple method of extracting, weighting and encoding convolutional features for image retrieval. In this section, we first introduce the background knowledge about the activations in CNN, and then introduce the feature map center selection method and how to use these centers to compute the channel weights and spatial weights; lastly, we aggregate the weighted descriptors to represent images.

3.1 Background

Given a pre-trained deep network, an image I of size $H_I \times W_I$ is input into this network, the activations of a convolutional/pooling layer, which is denoted as S, form a 3-D tensor of $H \times W \times K$ dimensions, where H and W represent the spatial dimension of the layer and K represents the number of channels. S contains K feature maps, and each $s_i, \quad i = 1, \ldots, K$ represents the feature map of the ith channel. S can also be considered as having $H \times W$ positions, and each position corresponds to a K-D descriptor. Here we denote the descriptor of position p as $d(p)$.

MAC [25] represents an image by concatenating the highest response in each channel

$$F = [f_1 \cdots f_i \cdots f_K], \quad with\, f_i = \max_{p \in H_i \times W_i} (s_i(p)), \tag{1}$$

f_i is the highest response over all the positions in the ith feature map, $s_i(p)$ is the response of position p, and $H_i \times W_i$ is the size of the feature map. The positions corresponding to the MAC vector components have the highest contribution to the pairwise image similarity, because each filter is interested in one kind of feature and the highest response can be considered as having a highest possibility to possess this feature.

Activation map is constructed by summing the feature maps in the same layer as Eq. 2. Activation map is widely used to describe the region of an object. The higher response a particular position is, the more possibility of its corresponding region being part of the object

$$S' = \sum_{i=1}^{K} s_i. \tag{2}$$

Because the zero responses in the feature maps are sparse, if a position has a high response in a feature map, it is likely that the response of this position in the activation map is also high, and the position is likely to correspond to an object region. Therefore, high response positions in feature maps are important for both image representation and object localization.

3.2 Feature map center selection

In our opinion, the position with the highest response can be considered as the center of this feature map, so MAC can be considered as a single center method. On the contrary, our MCDA method selects positions with the top few highest responses from each feature map to be the feature map centers. Compared with MAC, MCDA can be considered as a multi-center method.

As shown in Fig. 2, we color the positions whose responses are higher than zero on feature maps, red, green, blue and black regions represent the positions with the first, second, third and fourth highest responses respectively. MCDA selects some high response positions in each feature map as centers. The responses of centers are preserved, and at the same time, the rest of the responses are set to be zeros. Note that, the positions of centers correspond to the object region in the original image. The descriptor of a center is represented by concatenating all the responses that have the same position as the center across all the channels.

In order to discover these centers, we rank all the responses in each feature map in descending order, and the positions corresponding to the top $n_i, \quad i = 1, 2, \ldots, K$ highest responses are considered as the centers of the ith feature map, where K is the number of feature maps. Note that $0 \leqslant n_i \leqslant \varphi(i)$, where $\varphi(i)$ is the number of non-zero responses in the ith feature map.

To obtain n for each feature map, we formulate an objective function as

$$\begin{aligned} &\min_{n_i=0,\ldots,\varphi(i)} \sum_{j=1}^{T-1} \sum_{l>j}^{T} (\alpha D(p_j, p_l) - (1-\alpha) D(d(p_j), d(p_l))), \\ &s.t. \quad T = \sum\nolimits_i n_i, \end{aligned} \tag{3}$$

where $D(\cdot, \cdot)$ is the Euclidean distance between two vectors. Here $D(p_j, p_l)$ is the distance between a pair of centers, and $D(d(p_j), d(p_l))$ is the distance between the descriptors of two centers. Usually, the pixels on the object are close to one another. Therefore, the center positions should also be close. Minimizing the first term aims to make the range of centers relatively concentrated. On the other hand, MCDA hopes to explore all the features on the object, and the descriptors of the centers should cover all parts of the object. Minimizing the second term is to obtain the centers with diverse characteristics. α is used to leverage the contribution of the two terms. We experimented with different α for the objective function, and we achieved the best retrieval performance when we choose $\alpha = 0.7$.

Given the number of non-zero responses in each feature map, this can be cast as an optimization problem over discrete variables, and this optimization problem can be solved by coordinate descent approach [26] to update the number of centers for each feature map, as in Eq. 4, $n_q^{(t)}$ and $n_q^{(t+1)}$ denote the updated center numbers for the qth feature map in the tth and $(t+1)$th iteration. All the possible numbers of centers in a feature map should be enumerated in the iteration. During this process, the position of a possible center is denoted as p'

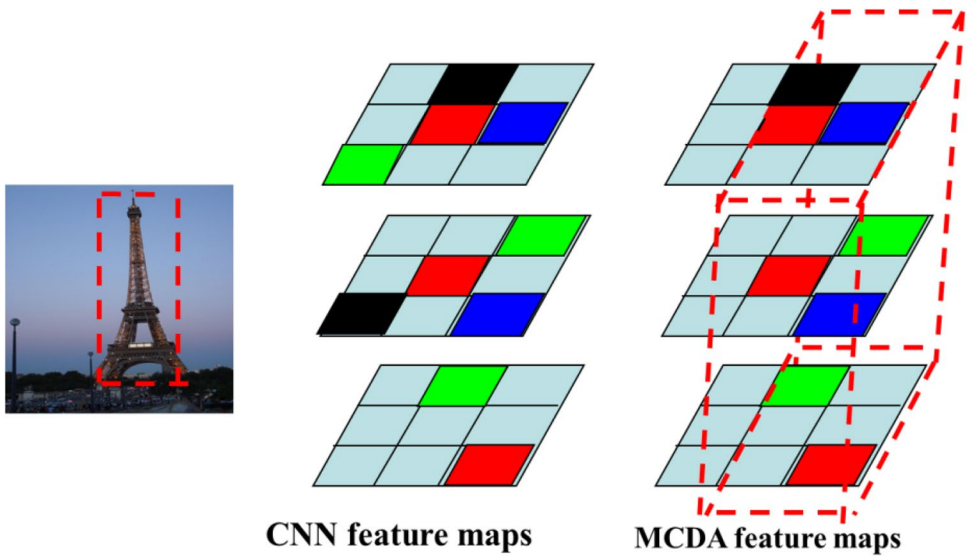

Fig. 2 The difference between CNN feature maps and MCDA feature maps

$$
\begin{aligned}
n_1^{(t+1)} = \arg\min_{r=0}^{\varphi(1)} & \sum_{j=n_1^{(t)}+1}^{T^{(t)}-1} \sum_{l>j}^{T^{(t)}} (\alpha D(p_j^{(t)}, p_l^{(t)}) - (1-\alpha) D(d(p_j^{(t+1)}), d(p_l^{(t+1)}))) \\
& + \sum_{i=0}^{r} \sum_{j=n_1^{(t)}+1}^{T^{(t)}} (\alpha D(p'_i, p_j^{(t)}) - (1-\alpha) D(d(p'_i), d(p_j^{(t+1)}))) \\
& + \sum_{i=0}^{r-1} \sum_{j>i}^{r} (\alpha D(p'_i, p'_j) - (1-\alpha) D(d(p'_i), d(p'_j))) \\
& \cdots \\
n_q^{(t+1)} = \arg\min_{r=0}^{\varphi(q)} & \sum_{j=1}^{n_1^{(t)}+\cdots+n_{q-1}^{(t)}-1} \sum_{l>j}^{n_1^{(t)}+\cdots+n_{q-1}^{(t)}} (\alpha D(p_j^{(t)}, p_l^{(t)}) - (1-\alpha) D(d(p_j^{(t+1)}), d(p_l^{(t+1)}))) \\
& + \sum_{j=n_1^{(t)}+\ldots+n_{q-1}^{(t)}+1}^{T^{(t)}-1} \sum_{l>j}^{T^{(t)}} (\alpha D(p_j^{(t)}, p_l^{(t)}) - (1-\alpha) D(d(p_j^{(t+1)}), d(p_l^{(t+1)}))) \\
& + \sum_{i=1}^{n_1^{(t)}+\ldots+n_{q-1}^{(t)}-1} \sum_{j=n_1^{(t)}+\ldots+n_{q-1}^{(t)}+1}^{T^{(t)}} (\alpha D(p_i^{(t)}, p_j^{(t)}) - (1-\alpha) D(d(p_i^{(t+1)}), d(p_j^{(t+1)}))) \\
& + \sum_{j=1}^{n_1^{(t)}+\ldots+n_{q-1}^{(t)}-1} \sum_{l=0}^{r} (\alpha D(p_j^{(t)}, p'_l) - (1-\alpha) D(d(p_j^{(t+1)}), d(p'_l))) \\
& + \sum_{j=n_1^{(t)}+\ldots+n_{q-1}^{(t)}+1}^{T^{(t)}} \sum_{l=0}^{r} (\alpha D(p_j^{(t)}, p'_l) - (1-\alpha) D(d(p_j^{(t+1)}), d(p'_l))) \\
& + \sum_{i=0}^{r-1} \sum_{j>i}^{r} (\alpha D(p'_i, p'_j) - (1-\alpha) D(d(p'_i), d(p'_j))) \\
& \cdots \\
n_K^{(t+1)} = \arg\min_{r=0}^{\varphi(K)} & \sum_{j=1}^{T^{(t)}-n_K^{(t)}-1} \sum_{l>j}^{T^{(t)}-n_K^{(t)}} (\alpha D(p_j^{(t)}, p_l^{(t)}) - (1-\alpha) D(d(p_j^{(t+1)}), d(p_l^{(t+1)}))) \\
& + \sum_{i=0}^{r} \sum_{j=1}^{T^{(t)}-n_k^{(t)}-1} (\alpha D(p'_i, p_j^{(t)}) - (1-\alpha) D(d(p'_j), d(p_j^{(t+1)}))) \\
& + \sum_{i=0}^{r-1} \sum_{j>i}^{r} (\alpha D(p'_i, p'_j) - (1-\alpha) D(d(p'_i), d(p'_j))).
\end{aligned}
\tag{4}
$$

3.3 Channel weighting

With the assumption that similar images have similar occurrence frequencies of a given feature, channel sparsity, which is the proportion of zero responses in a feature map, is then used to measure the channel weight [5]. A small response means that the confidence of having a certain characteristic is very low in the corresponding position, however, the response difference is not considered in the computation of sparsity.

Here we compare the regions of two images with the same object, and the corresponding regions with the top five highest responses in different channels are shown in Fig. 3. There are two things to be noted, first, the first 4 pairs of regions can be matched in the 12th channel while only the first pair of regions can be matched in the 45th channel. Second, the matching results decrease as the responses decrease. Figure 3 indicates that the channel weights calculated depending on the number of zero responses are not

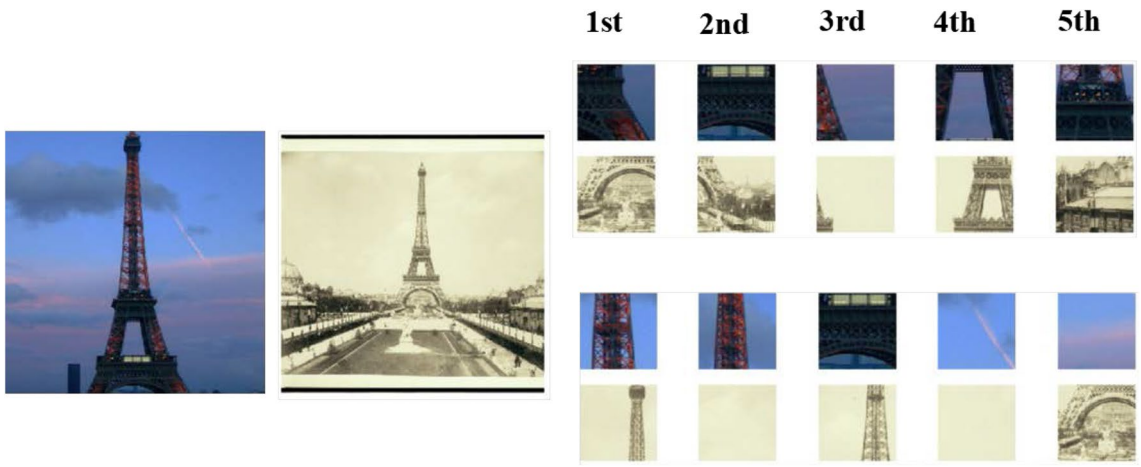

Fig. 3 Visualization of the corresponding regions of the top five highest responses. On the left we show two Eiffel images. On the right we show the comparison of the corresponding regions in the 12th (right top) and 45th (right bottom) channels

accurate enough, because the small non-zero responses can decrease the ability to describe the feature of objects.

Feature map center selection can eliminate small responses from each feature map. As a result, the channel sparsity can be computed more accurately using the number of centers, and finally used to evaluate the weight for each channel. For each feature map s_i, only the responses of n_i centers are preserved, and the rest of the responses are set to zeros, so we denote the sparsity as $sp_i = \frac{W_i \times H_i - n_i}{W_i \times H_i}$. Infrequently occurring features can also provide important information. Motivated by boosting the contribution of these features, we present a channel weighting scheme based on the channel sparsity, and the weight wc_i is $e^{-(1-sp_i)}$.

3.4 Spatial weighting

Inspired by [5], we propose a spatial weighting method based on the normalized activation map and the number of centers in the position across all channels. Let S' be the matrix of activation map as Eq. 2. For a given position, there are two factors can lead to an increase in spatial weight. First, the response of this position in activation map should be high. Second, the number of centers of this position across all the channels should be large. A large number means this position has large contribution to localize the object. We use L2 normalization and power-scaling to generate the spatial weight as

$$ws_{ij} = \left(\frac{S'_{ij} * k_{ij}}{\left(\sum_{m,n} (S'_{mn} * k_{mn})^a \right)^{1/a}} \right)^{1/b}, \tag{5}$$

where k_{ij} is the proportion of the centers across all the channels at position (i, j), and the parameters $a = 0.5$ and $b = 2$ [5].

3.5 Image representation

After feature map center selection, MCDA first weight the responses by the channel weight and spatial weight, and then all the descriptors are aggregated by sum pooling to construct image representation. The dimensionality of the image representation is the same as the number of channels.

4 Experiments

Our experiments aim to show the effectiveness of MCDA in image retrieval. We experiment on three challenging image retrieval datasets. INRIA Holiday [8] consists of 1491 images of 500 scenes or objects, which are collected from the personal holiday trip. Oxford5K [10] consists of 5062 Oxford landmarks images, which are collected from Flicker. There are 11 categories of landmarks, and five queries are selected from each category. Furthermore, additional 100,071 distractor images are combined with this dataset to be Oxford105K. Oxford Paris [9] consists of 6412 images of the landmarks of Paris.

For Oxford datasets, we follow the protocol that the cropped queries are used as the inputs of CNN. We employ the pre-trained deep model VGG16 [27] and the feature maps of the last pooling layer (pool5) are used to extract deep descriptors, and then the L2-normalized descriptors are weighted and aggregated for image representation,

the dimensionality of the image representation is 512. We adopt the Euclidean distance to measure the similarity between each pair of images, and mAP to measure the retrieval performance. In addition, we simply use the query expansion technology with MCDA features, we sum the MCDA features with the top $M = 10$ most similar retrieval results, and then L2-normalize this feature for re-query.

4.1 Image search

Max pooling and sum pooling are usually adopted for descriptor aggregation. We analyze and test our method by using these two schemes. When we use max pooling scheme to select the highest response from each channel, and the method is similar to MAC, we denote this method as MCDA_M. Here we denote our method based on sum pooling scheme as MCDA. The comparison of mAP is shown in Table 1. Note that, MCDA consistently outperforms MCDA_M in all the datasets greatly. This is because more responses are preserved in MCDA, so MCDA can contain the global information of the object. It is worth noting that the center number of a feature map can be zero, which means that the feature map is useless to represent the object. On the contrary, MCDA_M has to choose one response from each feature map. Information is lost in the feature maps where more than one response contains the object information, and also useless information is preserved in the feature maps where there is no response can contain the object information. Therefore, MCDA_M can be considered as a special case of MCDA.

We compare our results with some state-of-the-art methods, which are proposed based on pre-trained CNN model in Table 2. The convolutional descriptors in different layers are encoded using standard VLAD encoding in [7]. The object region and importance of descriptors are not considered in this method, so the performance is inferior to most of the methods. The method in [7] outperforms MCDA in Holiday, but MCDA outperforms [7] in all the other three datasets. Holiday dataset contains many scene images; all the contents are important to represent images. The method in [7] uses all the descriptors in different levels, so all the information is preserved. While MCDA only selects some descriptors to represent images and therefore some information is lost.

SPoC [20] assigns large weights to the descriptors close to the image center. However, the objects can be located in any part of an image. MCDA can find the object descriptors based on feature map centers. Moreover, SPoC weights and aggregates all the descriptors, including both the object and the background, whereas, MCDA aggregates the descriptors corresponding to the object region. Therefore, MCDA outperforms SPoc in all the datasets. R-MAC [21] samples squares in different scales, and the image representation is constructed by summing and normalizing the region descriptors. These region descriptors contain the context information, however, the descriptors of overlapping regions may contain duplicate information, and the importance of descriptors is not considered. MCDA can represent the deep features without duplicate information and assign reasonable weights for the descriptors. Similar to MCDA, CroW [5] uses spatial weight and channel weight to highlight the descriptors corresponding to the object region. All the responses in the feature maps are used for the weight computation. Compared with Crow, our spatial weight and channel weight are proposed based on feature map centers, so these two weights can be computed more accurately. As a result, MCDA achieves a greater than 1% improvement in mAP.

CroW [5] assigns weights based on the responses of the feature maps, and the time complexity is $O(WHK)$. The time complexity of R-MAC [21] is $O(WHR)$, where R is the number of sampled regions. In contrast, MCDA spends extra time in solving the center selection problem, and the time complexity of each iteration is $O(WHKT'^2)$, where T' is the initial number of centers in the iteration. T' is usually larger than K and R, so the time complexity of MCDA is usually larger than CroW [5] and R-MAC [21]. Because of the optimization, MCDA outperforms some state-of-the-art methods in Table 2 and some retrieval examples are shown in Fig. 4. Note that these images are under different viewpoints and lightings. The retrieval results demonstrate that MCDA is robust to viewpoint and lighting variance.

Table 1 Comparison of the mean average precision when using different aggregation method for MCDA

Method	Holiday	Oxford5k	Oxford105k	Paris
MCDA_M	0.701	0.585	0.573	0.789
MCDA	0.839	0.768	0.717	0.862

Table 2 Comparison with state-of-the-art on Oxford and Holiday

Method	Holiday	Oxford5k	Oxford105k	Paris
Neural Codes [17]	0.749	0.435	0.329	–
Ng et al. [7]	0.840	0.581	–	0.688
Razavian et al. [28]	0.716	0.533	0.489	0.670
SPoC [20]	0.802	0589	0.578	–
R-MAC [21]	–	0.669	0.616	0.830
CroW [5]	0.828	0.749	0.706	0.848
MCDA	0.839	0.768	0.717	0.862

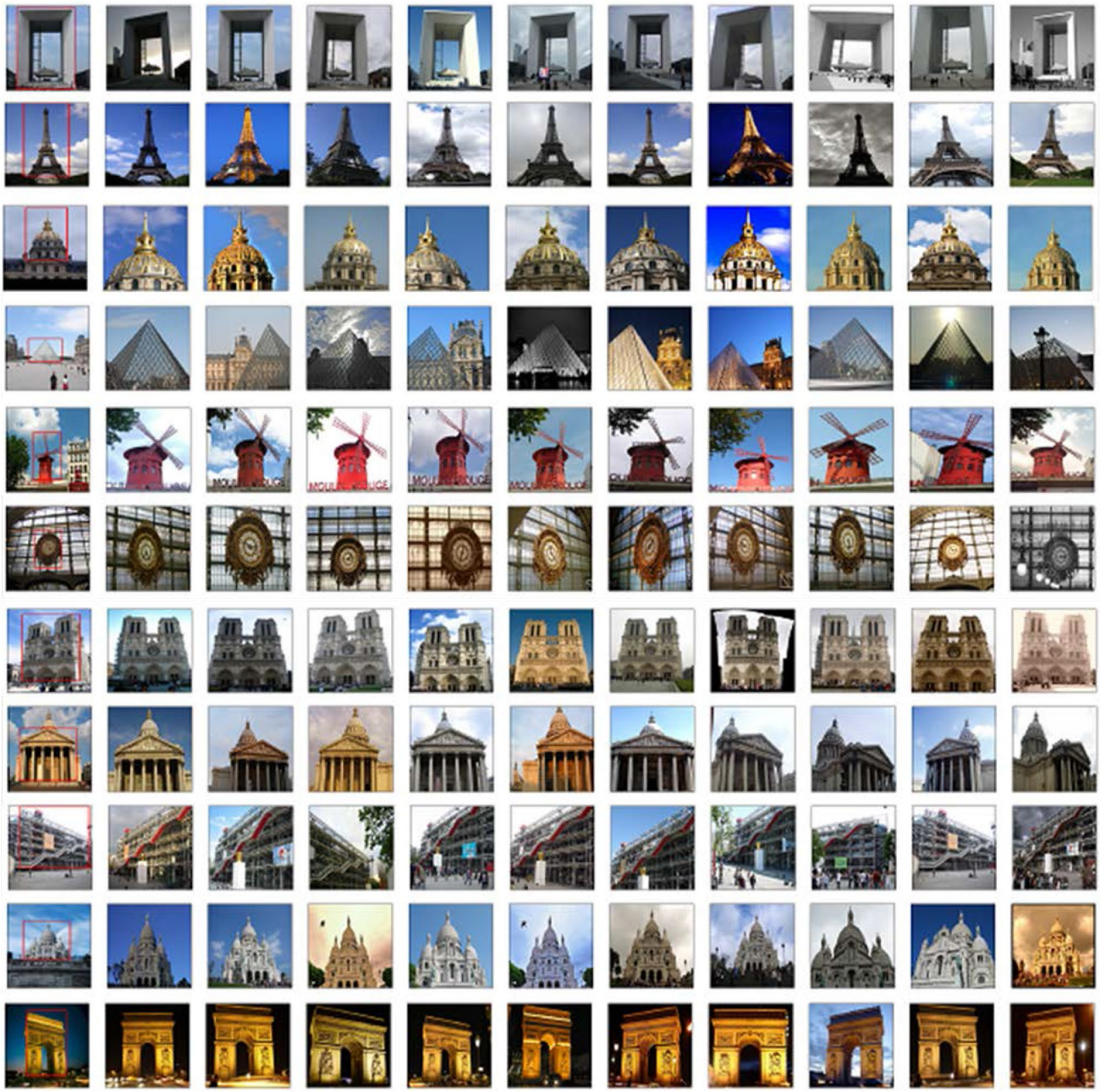

Fig. 4 Retrieval examples using MCDA on Oxford Paris dataset. The query images are shown on the leftmost, and the objects are marked with bounding boxes

4.2 Object localization

The activation map of MCDA can localize the object at any location, we select some images from Paris, Oxford 5K and Holiday, and the corresponding activation maps are shown in Fig. 5.

α is employed to adjust the difference in importance between the position similarity and descriptor similarity during the feature map center selection, and its value is tuned over the values $\{0, 0.1, 0.2,..., 0.9\}$. The reason that α cannot be 1 is that no center will be selected in each channel in this case. A larger α means that MCDA wants the selected centers to be more compact. In Fig. 6 we present mAP when varying the value of α. When we choose $\alpha = 0.7$, the best results can be obtained in all the datasets. In Fig. 7 we present the activation maps when varying the

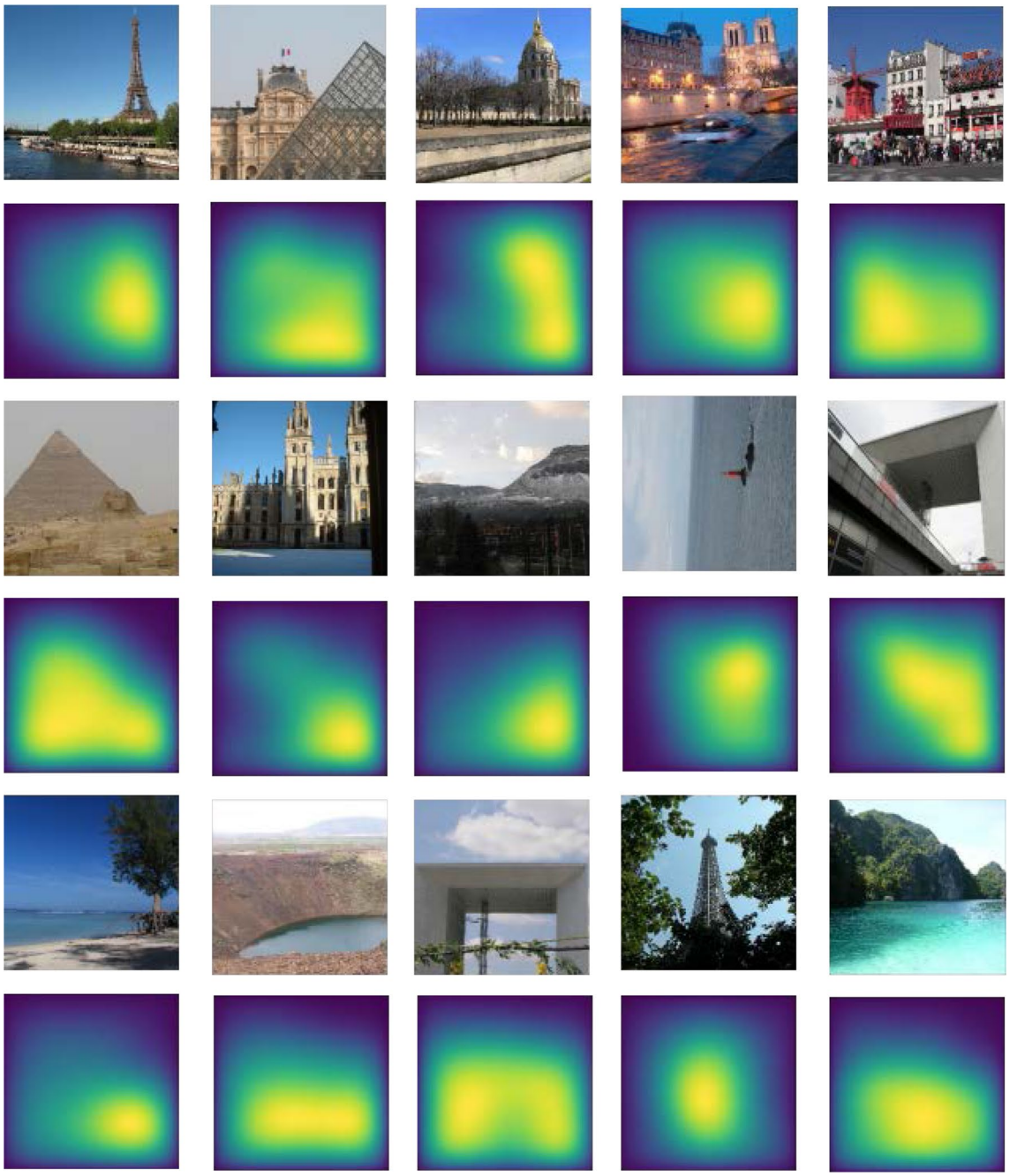

Fig. 5 Activation maps of MCDA

value of α. It is obviously that the object region can be better localized when we choose $\alpha = 0.7$. We can also observe that, there is a sharp decrease in localization accuracy when α is larger than 0.7.

Descriptors are usually extracted from the last pooling or convolutional layer in most of the methods. We test the retrieval performance on these two cases, and find that the mAP of pool5 layer consistently outperforms that of conv5-3. In addition, the size of the feature maps in conv5-3 is 14*14, while the size is only 7*7 in pool5, therefore, the feature map center selection process on pool5 is much faster than on conv5-3. It is effective and efficient to use pool5 instead of Conv5-3 in our method.

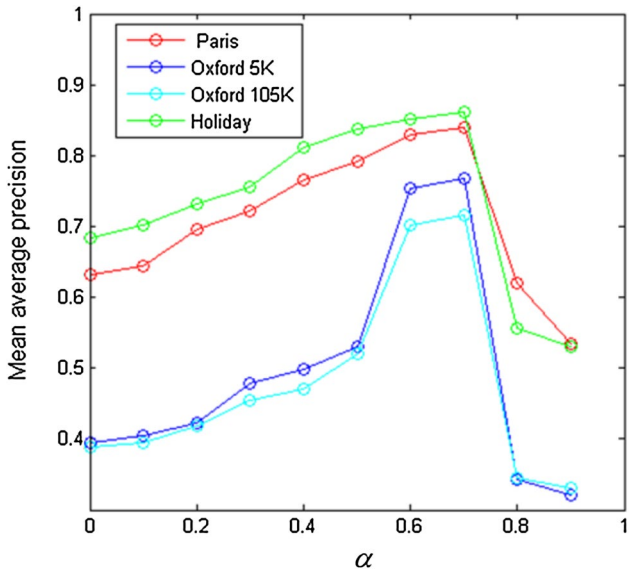

Fig. 6 Mean average precision on Holiday, Oxford and Paris when varying the value of α

5 Conclusion

In this paper, we propose a feature map center selection method for aggregating and weighting the deep features. The experiments demonstrate that our MCDA method can achieve state-of-the-art retrieval results. Moreover, the activation maps generated using these centers are also effective for object localization. However, our image representation is constructed based on the pre-trained network VGG16, and the parameters are learned from ImageNet. Tuning these parameters for a particular retrieval task is a promising future direction. We will perform research in tuning the parameters based on feature map center selection and ranking loss for image retrieval in the future. In addition, semi-supervised [29, 30] and unsupervised [31–33] learning based methods has shown their advantages in machine learning, we will also try to perform research based on these excellent methods.

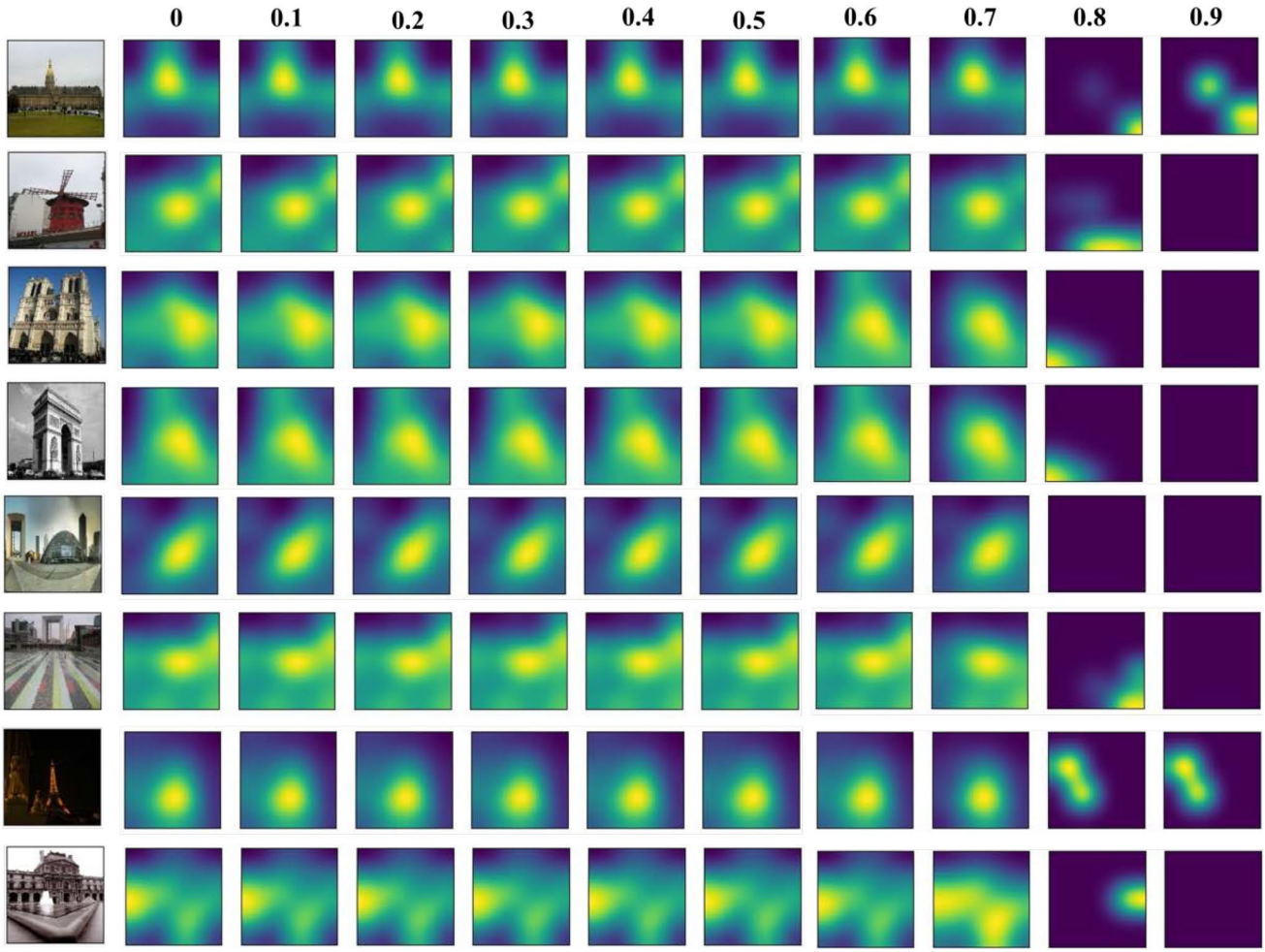

Fig. 7 The activation maps when varying the value of α

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant 61772344 and Grant 61732011), in part by the Natural Science Foundation of SZU (Grant 827-000140, Grant 827-000230, and Grant 2017060), the National Social Science Foundation of China (17BTQ068), the Youth Foundation of Education Bureau of Hebei Province (Grant QN2015099), China Postdoctoral Science Foundation funded project (Grant 2017M621078). The funding project of midwest colleges and universities promoting comprehensive strength of Hebei University.

References

1. Krizhevsky A, Sutskever I, Hinton GE(2012) Imagenet classification with deep convolutional neural networks. In: Annual conference on neural information processing systems, pp 1097–1105
2. Cui L, Yang S, Chen F et al (2018) A survey on application of machine learning for internet of things. Int J Mach Learn Cybernet 9(8):1399–1417
3. Banharnsakun A (2018) Towards improving the convolutional neural networks for deep learning using the distributed artificial bee colony method. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0811-z
4. Cui Y, Jiang J, Lai Z et al (2018) Supervised discrete discriminant hashing for image retrieval. Pattern Recogn 78:79–90
5. Kalantidis Y, Mellina C, Osindero S (2016) Cross-dimensional weighting for aggregated deep convolutional features. European conference on computer vision. Springer, Cham, pp 685–701
6. Wei XS, Luo JH, Wu J (2017) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans Image Process 26(6):2868–2881
7. Ng JYH, Yang F, Davis LS (2015) Exploiting local features from deep networks for image retrieval. arXiv preprint: 1504.05133
8. Jegou H, Douze M, Schmid C (2008) Hamming embedding and weak geometric consistency for large scale image search. European conference on computer vision, pp 304–317
9. Philbin J, Chum O, Isard M (2008) Lost in quantization: Improving particular object retrieval in large scale image databases. International conference on computer vision and pattern recognition, pp 1–8
10. Philbin J, Chum O, Isard M (2007) Object retrieval with large vocabularies and fast spatial matching. International conference on computer vision and pattern recognition, pp 1–8
11. Wang XZ, Wang R, Xu C (2018) Discovering the relationship between generalization and uncertainty by incorporating complexity of classification. IEEE Trans Cybern 48(2):703–715
12. Wang R, Kwong S, Wang XZ et al (2015) Segment based decision tree induction with continuous valued attributes. IEEE Trans Cybern 45(7):1262–1275
13. Wang R, Chow CY, Kwong S (2016) Ambiguity-based multiclass active learning. IEEE Trans Fuzzy Syst 24(1):242–248
14. Wang R, Wang XZ, Kwong S et al (2017) Incorporating diversity and informativeness in multiple-instance active learning. IEEE Trans Fuzzy Syst 25(6):1460–1475
15. Jégou H, Douze M, Schmid C (2010) Aggregating local descriptors into a compact image representation. International conference on computer vision and pattern recognition, pp 3304–3311
16. Sánchez J, Perronnin F, Mensink T et al (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vis 105(3):222–245
17. Babenko A, Slesarev A, Chigorin A (2014) Neural codes for image retrieval. European conference on computer vision, pp 584–599
18. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition. International conference on computer vision and pattern recognition workshops, pp 512–519
19. Gong Y, Wang L, Guo R (2014) Multi-scale orderless pooling of deep convolutional activation features. European conference on computer vision, pp 392–407
20. Babenko A, Lempitsky V (2015) Aggregating local deep features for image retrieval. IEEE international conference on computer vision, pp 1269–1277
21. Tolias G, Sicre R, Jégou H (2015) Particular object retrieval with integral max-pooling of CNN activations. arXiv preprint 1511.05879
22. Radenović F, Tolias G, Chum O (2016) CNN image retrieval learns from BoW: unsupervised fine-tuning with hard examples. European conference on computer vision, pp 3–20
23. Liu Z, Li J, Shen Z (2017) Learning efficient convolutional networks through network slimming. IEEE international conference on computer vision, pp 2755–2763
24. Boscaini D, Masci J, Rodolà E (2016) Learning shape correspondence with anisotropic convolutional neural networks. Adv Neural Inf Process Syst 3189–3197
25. Azizpour H, Sharif Razavian A, Sullivan J, Maki A, Carlsson S (2015) From generic to specific deep representations for visual recognition. International conference on computer vision and pattern recognition workshops, pp 36–45
26. Fu Z, Robles-Kelly A, Zhou J (2011) MILIS: Multiple instance learning with instance selection. IEEE Trans Pattern Anal Mach Intell 33(5):958–977
27. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. International conference on learning representations, pp 1–14
28. Razavian AS, Sullivan J, Carlsson S (2016) Visual instance retrieval with deep convolutional networks. ITE Trans Media Technol Appl 4(3):251–258
29. Huang R, Zhang G, Chen J (2018) Semi-supervised discriminant Isomap with application to visualization, image retrieval and classification. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0809-6
30. Zhu Q, Yuan N, Guan D et al (2018) An alternative to face image representation and classification. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0802-0
31. Liu J, Liu W, Ma S et al (2018) Image-set based face recognition using K-SVD dictionary learning. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-017-0782-5
32. Ding S, Zhang N, Zhang J et al (2017) Unsupervised extreme learning machine with representational features. Int J Mach Learn Cybernet 8(2):587–595
33. Fang J, Xu X, Liu H et al (2018) Local receptive field based extreme learning machine with three channels for histopathological image classification. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-018-0825-6