Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 4, April 2018, pp. 621-630
DOI: 10.23940/ijpe.18.04.p4.621630

Exploiting Best Practice of Deep CNNs Features for National Costume Image Retrieval

Juxiang Zhou[a,b], Xiaodong Liu[a,*], and Jianhou Gan[b]

[a]*Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, 116024, China*
[b]*Key Laboratory of Education Informatization for Nationalities, Yunnan Normal University, Kunming, 650500, China*

Abstract

Convolutional neural networks (CNNs) have recently achieved remarkable success with superior performances in computer vision applications. In most CNN-based image retrieval methods, deep CNNs features are verified as discriminative descriptors for effective image representation. This paper exploits the best practice for CNNs application to national costume image retrieval. Several important aspects that affect the discriminative ability of deep CNNs features are investigated thoroughly, including layers selection, aggregation and weighting methods. Firstly, an effective weighting method for sum-pooling features aggregation is given, which is more suitable for national costume image than some typical aggregation methods such as SPoC and SCDA. Secondly, in view of the complementary strengths, compact multi-layer CNN features combined with low dimensions are proposed and proven to be effective for national costume expression. Finally, a re-ranking strategy of diffusion process is applied to further enhance the performance for national costume images retrieval. The experimental results show that the proposed method outperforms the existing methods remarkably, which will provide some new research ideas and technical references for researchers in the field of national costume image retrieval.

Keywords: CNN; image retrieval; national costume; aggregation; diffusion process

(Submitted on December 17, 2017; Revised on January 29, 2018; Accepted on March 8, 2018)

1. Introduction

With the rapid development of computer networks and the application of digital information technology and multimedia technology, a great deal of information from all walks of life are digitized in the form of multimedia information. Rapid growth of digital media resources makes it harder for people to obtain effective and valuable information from large data. For digital image resources, content-based image retrieval (CBIR) technology has become a hot research topic in the field of digital image application in recent years [19,23,24]. The CBIR technology has successfully been applied in many areas, from natural image retrieval to special field of image retrieval, including exhibition and museum management, construction and engineering design, remote sensing and geographical resources management, GIS, weather forecast, cultural heritage image retrieval, Chinese Ink Paintings image retrieval and national costume image retrieval, etc. [2,14,25,26,33].

The national costume is the unique ethnic minority culture, which was created by the large minority compatriots in the process of historical development [31]. Compared to visual characteristics of common natural images, the national costume is embedded with prominent characteristics mixed with flowery color, complex texture and rich totem, which makes it more challenging for image processing. Analyzing the visual characteristics of the national costume from the perspective of computer vision has significant advantages in two respects. On one hand, it has prominent superiority with accurate description, quick analysis and reliable judgment, which provides beneficial reference and basis for the identification and cognition of national costumes. On the other hand, it can dig out deeper potential information that cannot be identified by human eyes, which can assist scholars in deeply researching national costumes more objectively and more comprehensively.

* Corresponding author.
E-mail address: zjuxiang@126.com

However, most scholars research and analyze the characteristics of national costumes, such as color, texture, and shape only from the perspective of aesthetics, literature and art. Few do from the angle of computer vision. At the same time, in terms of national costumes image resources retrieval application, the demand of national costumes also continues to expand in the field of academic research, art design and more fields with rapid growth of multimedia technology. But the fact is that we cannot get satisfactory results when seeking adequate national costumes image resources in large databases. The main reason is that image resources of national costumes is rarely online. Because the visual information contained in national costumes is richer and more abstract, traditional retrieval methods based on analysis and extraction for low-level designed features have great limitations in the retrieval application for national costume images.

In recent years, with further research and successful application of deep learning in the field of image processing, the feature extraction method based on deep learning has received great attention. Compared with the traditional designed characteristics, the learning features based on deep learning can automatically learn the effective features from the image. Many studies have shown that using depth, CNNs can autonomously learn features ranging from low-level to high-level from huge amounts of image, making image classification close to human level [1,10]. Considering the advantages of features based on CNNs and the complexity of visual features of national costumes, this paper exploits and investigates the best practice of deep convolutional features for national costume image retrieval. The main contributions include:

- CNNs are introduced to national costume images retrieval for the first time, exploring the potential ability of CNNs in special application fields.
- An effective weighting before sum-pooling is proposed to create a final aggregation by evaluating the usefulness of the obtained deep descriptors, which can enhance the power of useful deep descriptors and weaken the negative effects brought by background or noise.
- With suitable aggregating and effective weighting, compact multi-layer CNN features combined with low dimension are proposed, which achieves remarkable improvement compared to some existing methods for national costume images retrieval.
- Diffusion process as a re-ranking strategy is introduced to enhance the performance of the proposed national costume images retrieval framework.

The rest of the paper is organized as follows: In section 2, the related works is summarized from two aspects of review on clothing retrieval techniques and research status on national costume image retrieval. In section 3, the proposed approach for national costume image retrieval is presented, including the proposed weighting method for feature aggregation, multi-layers deep CNNs features extraction and ranking optimization for feature matching. Section 4 shows experimental results and discussion. The conclusions and future works are given in section 5.

2. Related works

2.1. Review on clothing retrieval techniques

Huge amounts of clothing images are easily available from the Internet, especially from e-commercial platforms. Clothing image retrieval is of importance for commercial and social applications, which has recently received tremendous attention such as multimedia processing and computer vision [18]. However, sifting through massive amounts of available products to find an item that suits one's tastes and preferences can be an arduous, time-consuming task [16]. Although CBIR methods can search for visual information with visual information input, it is still challenging to extract the image features that capture the design of the clothing.

Recently, clothing retrieval has drawn more attention due to the huge fashion market. Many scholars attempt to realize clothing retrieval based on traditional low-level visual features. A project by Hsu et al. [11] used images with uniform backgrounds as queries for retrieving a limited scope of clothing items that one might find on an online shop. Their approach involved comparing items based on the features — color, texture, SIFT features and outline — that the pixels in the clothing regions of the given images form. The system proposed by Chen et al. [5] describes clothing by semantic attributes, where a list of nameable clothing attributes is generated. They extract low-level features in a pose-adaptive way and associate complementary features for learning attribute classifiers. [8] proposed a color, texture and shape-based hybrid approach for clothing retrieval techniques. Though various researches have been done for clothing matching and retrieval, low-level based methods cannot get satisfactory results because there are deficiencies for describing high-level semantic information, especially when the clothes have varieties in garment appearance, layering, style, deformation, body shape and posture.

CNN-based visual representation has also shown improved performance over handcrafted features on digit recognition [4]. Considering the superiority of CNNs features, Lin K et al. [17] presented a deep CNN framework for rapid clothing retrieval in a recommendation system. Ruifan Li et al. [18] proposed multi-weight deep convolutional neural networks to cope with the noisy and imbalanced clothing image data in the real world, which contains two group layers, common layers and task dependent layers with a group of category relevant parameters. Lixuan Yang et al. [30] introduced a new method for extracting deformable clothing items from still images by extending the output of a Fully Convolutional Neural Network (FCN) to infer context from local units (super pixels). In addition, there are also many research achievements based on depth learning for clothing attribute description, landmarks detection, fashion style classification and other applications [13,20,22].

2.2. Research in National costume image retrieval

The national costume is a special kind of clothing and is one of the important embodiments of the national culture. The national costume has prominent visual characteristics mixed with a flowery color, complex texture and rich totem, which bring more challenges for feature description. There are some related works about national costume image retrieval. Xumei Shen et al. [26] proposed an image feature representation method to describe the rich information of national costume image based on color histogram and edge orientation histogram. The method was verified in their self-built national costume image dataset, which contained six national costume images including those from Bai nationality, Jingpo nationality, Hani nationality, Miao nationality, Bouyei nationality and Va nationality. Aimed at the shortcomings of the traditional image retrieval based on global features, Weili Zhao et al. [33] presented an effective method for national costume images retrieval based on integrated regional matching. This method used significant visual features of multi-feature fusing color, texture and shape feature for each region. But, in general, both methods mentioned above did not achieve ideal performances as both had retrieval accuracies of less than 70%. It is obvious that the traditional low-level based features are not sufficient enough to describe the gorgeous and complicated visual features implied in national costume images. So, more competent and strong features eagerly need to be explored to describe national costume images.

Deep CNNs features are also used to solve the complicated cross-domain problems because of their strong generalization ability. They contain high-level context information that has proven to be prominent in discriminating visual features for images. Nevertheless, there is still no research and investigation for content-based image retrieval of national costume images based on CNNs. This paper exploits the best practice of deep CNNs features for national costume image retrieval. Also, we look into the availability of CNNs for real application and present compact multi-layer CNN features for national costume with a proposed weighting method for feature aggregation. Finally, a diffusion process as a re-ranking strategy is applied to facilitate the retrieval performance.

3. The proposed approach for national costume image retrieval

In this section, we present the proposed approach for national costume image retrieval, which includes two stages. The framework of the proposed method is shown in Figure 1.

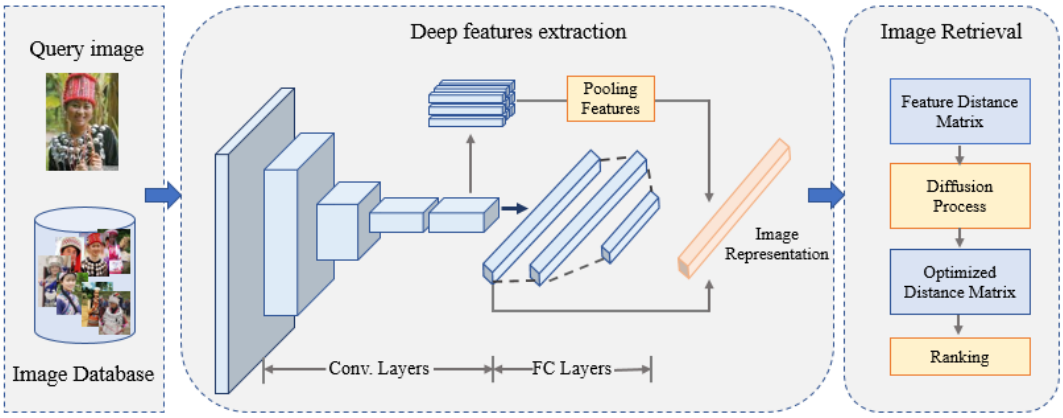

Figure 1. The framework of the proposed retrieval approach for national costume images

In the stage of deep feature extraction, an effective weighting method for sum-pooling aggregation is given, and a multi-layer features extraction method is proposed. Based on the feature similarity, we introduce a diffusion process for ranking optimization in order to enhance the retrieval performance of our proposed framework.

3.1. Deep features extraction

There are several approaches to retrieval based on CNNs, including obtaining features by existing CNNs from convolutional (conv.) layers or fully-connected (FC) layers with domain-relevant datasets as feature extractors. Because conv. and FC features are derived from different depths in CNN architecture, they are at various representation levels. Conv. features correspond to the local responses of every image region, while FC features contain global information of the holistic image. This diversity may lead to different impacts on different data [29]. In this paper, we adopt deep features derived from conv. and FC layers of a pre-trained VGG-F model [6] to obtain image descriptors for national costumes.

3.1.1. Conv. Features and FC features

The activations of intermediate conv. layers reflect relatively low-level features because of insufficient feature learning from the training data. Thus, conv. activations tend to catch finer grained information that may be beneficial to content-based retrieval [29]. When passing an input image *I* through a CNN, the outputs from conv. layers are feature maps, in which each element corresponds to a receptive field of the input image. Suppose the responses of a certain conv. layer form D feature maps and the size of each feature map is H×W, the activations of a conv. layer are formulated as order-3 tensor T with H×W×L elements that describe local features for *I*. From another point of view, T can be also considered as a H×W L-dimensional deep descriptor. The channel number of L depends on the inherent structure of CNN, and the feature map spatial resolution H×W rests with the architecture of CNN, the adopted layer and the size of image I.

The activations derived from deep FC layers convey category-level features that can be interpreted as global representations for the input image. When the last FC layer, namely the SoftMax layer, is removed, the rest of a CNN can be regarded as a generalized feature extractor. Compared with conv. layers, FC layers can only process images with a fixed size, e.g., 227×227 pixels or 224×224 pixels. For a resized or cropped input image I, FC layers can directly generate discriminating descriptors that originally are in the form of single vectors, which has a settled dimension that merely depends on the inherent structure of the CNN model, e.g., 4096-dimension or 1024-dimension.

3.1.2. Feature aggregation and the proposed weighting method

For conv. features, the activation maps cannot be directly used to describe images, and they need to be normally aggregated into compact global descriptors by aggregation methods. There are several promising aggregation strategies, including some encoding or pooling methods. BoW [15] and VLAD [12] are commonly used feature encoding approaches. BoW describes image information with statistics on the spatial distribution of local feature vectors through a codebook learned from feature set by k-means clustering. The output of BoW is a k-dimensional vector. VLAD considers the statistical distribution of local features as well as the vector difference between local features and centroids simultaneously. VLAD encodes activations at each level separately, and then it finally concatenates the features. The most familiar pooling methods are max-pooling and sum-pooling.

However, these approaches directly used the CNN activations/descriptors and encodes them into a single representation without considering the weighting of different spatial positions in a feature map. A. Babenko et al. [3] found that using sum-pooling to aggregate deep features on the last convolutional layer leads to a better performance, and they then proposed the sum-pooled convolutional (SPoC) features. Also, SPoC is acquired with center-prior Gaussian weighting on spatial of feature maps followed by sum pooling. XiuShen Wei et al. [28] proposed the selective convolutional descriptor aggregation (SCDA) method with evaluating the usefulness of the obtained deep descriptors, which can select the useful deep descriptors and remove background or noise by localizing the main unsupervised object.

Obviously, prominent superiorities of SPoC and SCDA can be better reflected by dealing with images that have important information in the center or ones that contain some objects. However, our national costume images may not have such peculiarity that can be seen from the image samples. A new weighting is proposed before sum-pooling to create a final aggregation. The difference is that we cannot directly remove certain features in a feature map, but we can enhance or weaken the influence by a weighting function ω that reflects the degree of usefulness by comparing it with the average value. Specifically, it is defined in Equation (1).

$$V_m(I) = \sum_{y=1}^{H} \sum_{x=1}^{W} \omega(x,y) f(x,y) \tag{1}$$

$$\omega(x,y)=1+\frac{s(x,y)-Avg}{Avg},\quad s(x,y)=\sum_{m=1}^{L} f(x,y),\quad Avg=(\sum_{y=1}^{H}\sum_{x=1}^{W} s(x,y))/L$$

where $H\times W\times L$ is the size of the feature map from convolutional layer, $V_m(I)$ is the value after *aggregation for m-th* map, m=1, 2, …, L.

3.1.3. Multi-layer features extraction

Different layers focus on extracting different information from images. High-layer feature is not powerful enough to describe detailed information, while low-layer feature suffers from background clutter and semantic ambiguity [32]. So, we take advantage of the complementarity of low-layer and high-layer to improve the image retrieval task. Then, the last convolutional layer of pool5 is selected as conv. features, and the full connected layer of relu6 is selected as FC features. For pool5 feature maps, we use the proposed weighting on the spatial of feature maps followed by sum pooling. Although FC features can be used directly, in order to achieve higher effectiveness, we preprocess both conv. and FC activations with PCA reduction followed by L2-normalization respectively, and then concatenate them together. The experimental results will prove the effectiveness of the PCA-compressed deep features.

3.2. Feature matching based on similarity metric *and ranking optimization* with diffusion process

To achieve the final retrieval, feature matching is an indispensable link after feature extraction by comparing the similarity of query images and images in a database. Similarity metric is defined by computing the distance between visual feature vectors, which is one of the core issues in content-based image retrieval. The resulting PCA-compressed deep features are normalized by L2-normalization. Euclidian distances is used to compute the distances of pair deep features to form a distance matrix D (lower is more similar) for image retrieval. Suppose the number of images in a database is N, and distance matrix D is symmetric with $N\times N$ elements, then all primary diagonal elements are 0.

Traditionally, in an image retrieval system, the retrieval task will be conducted by sorting the distance matrix D. However, it is just based on analyzing pairwise difference values, which measure the distance between two elements. That is to say, the performance is mainly defined by the pairwise measures. It is obvious that such a retrieval approach ignores the structure of the underlying data manifold without considering the context sensitive similarities. In view of this limitation, several diffusion processes [9] have been proposed to exploit the context between all elements of the database, as ranking optimization for distance metric of features has been used to improve retrieval performance. In general, the aim of these methods is to capture the intrinsic manifold structure defined by pairwise affinity matrices. Usually, the diffusion process includes three critical factors: initialization, definition of transition matrix, and diffusion process. Many different diffusion processes can be established with different tactics. In this paper, one type of diffusion process is introduced to optimize the distance matrices formed by deep features. The selected diffusion process is the same as in [9], which has been effectively verified to improve the performance in large-scale retrieval applications, while having a low computation time. The main steps are shown below:

Input: Distance matrix D, the number of the nearest neighbors N_k , iteration stop error ε .

Output: A diffused matrix A^*

Step 1: Normalize the distance matrix D to an affinity matrix A, in which the values of all elements are between 0 (the smaller the difference) and 1 (the larger the similarity).

Step 2: Define P_{kNN} matrix, which can be obtained by setting all positions to 0 except the positions of N_k nearest neighbors for every element in affinity matrix A, and then normalizing the sum over all rows to 1.

Step 3: Initialize the diffusion process W_0 by P_{kNN} matrix.

Step 4: Define transition matrix $T = P_{kNN}$ to constrain the transition matrix to the N_k -nearest neighbors.

Step 5: Update W by $W_{t+1} = TW_tT^T$.

Step 6: Compute the difference of the average number v of changing elements between two subsequent iterations after comparing the old rankings with the new rankings of the obtained diffused matrices.

Step 7: Stop update until v is below ε , finally yields the final matrix A^* .

In the above process, P_{kNN} is a Markov Chain transition matrix, the core idea of which is to obtain the transition probabilities, setting them to high values if and only if all paths between two nodes in k-NN neighbor are short [9]. It is worth noting that A^* may no longer be symmetric, and every element $a^*_{i,j}$ (i=1,2,3,…,N; j=1,2,3,…,N) represents an affinity value

between *i-th* query image and *j-th* images in the database. The larger the value $a^*_{i,j}$, the more similar the images *i* and *j* are. Then, the final retrieval task can be completed through sorting A^* in a row.

4. Experimental results and discussions

4.1. National costume image database

The dataset [21] used in our experiments is built by [26], which contains six classes of national costume images including the Bai nationality, Jingpo nationality, Hani nationality, Miao nationality, Bouyei nationality and Va nationality. Each nationality has 100 images, and the total number of images is 600. The sizes of the images are all 128×96 or 96×128. Some samples are shown in Figure 2.

Figure 2. Image samples in national costume image database

4.2. Performance evaluation and experiment setup

In all the experiments, each image in the four databases is used as a query image. The top relevant images are retrieved and ranked according to the similarity scores, and the mean average precision (mAP) is computed to measure the retrieval accuracy based on the ground truth of the dataset. The precision of ranked top R_k images is computed to measure the retrieval accuracy for any query as in Equation (2).

$$P(R_k) = \frac{The\ number\ of\ (relevant\ images \cap retrieved\ images)}{The\ number\ of\ retrieved\ images} \tag{2}$$

where the relevant images are only images relevant to the query image according to its ground truth. The average precision (AP) is the average of the precision value obtained for the set of top *k* images over all queries. Suppose the number of relevant images in the database for every query image *q* is N_q, and the number of query images is $|Q|$, then the mAP is defined as Equation (3).

$$mAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{N_q}\sum_{k=1}^{N_q}P(R_k) \tag{3}$$

In experiments, for the pre-trained deep model, the publicly available VGG-F model [7] is employed to extract deep convolutional descriptors using the open-source library MatConvNet [27]. All images in database are resized to 224×224 and sent to the network. Then, a group of deep features can be obtained from 20 different layers excluding the last 'SoftMax' layer. Among them, the dimensions of pool5 and relu6 are 6×6×256 and 1×1×4096 respectively.

4.3. Results and discussions

In order to explore the best deep CNNs features for national costume image retrieval, we perform experiments on national costume image database and investigate some variables concerned to the property of CNN representations.

4.3.1. Performance investigation of different CNN layers

Conv. and FC features are derived from different depths with various representation levels in CNN architecture. Conv. features represent local responses of every image region, while FC features contain global information of the holistic image. This diversity may lead to a different impact on different data. In order to investigate their feasibility for instance retrieval, we evaluate the performances using features from 9 conv. layers with sum-pooling and 4 FC layers. All these features are compacted by PCA with 128 dimensions and normalized by L2-normalization.

The retrieval performances of deep features extracted from different layers are shown in Figure 3. From the results, we can see that the retrieval performances increase as the layer increases from the lower layer conv2 to the higher layers rule6. Then, it begins to decrease from the layer fc7. Such results verify the fact that features from lower layers, such as conv2 and conv3, are too generic and lack the semantic meanings of the object in the image, while features from the highest layer (fc7) contain the semantic meaning of objects but lack the detailed and local information needed to match two similar images. For conv. layers and FC layers, the best results are obtained in rule5 and relu6, respectively, where the feature vectors combine both the low-level detailed information and high-level semantic meanings of the image. So, they are later considered for multi-layer fusion.

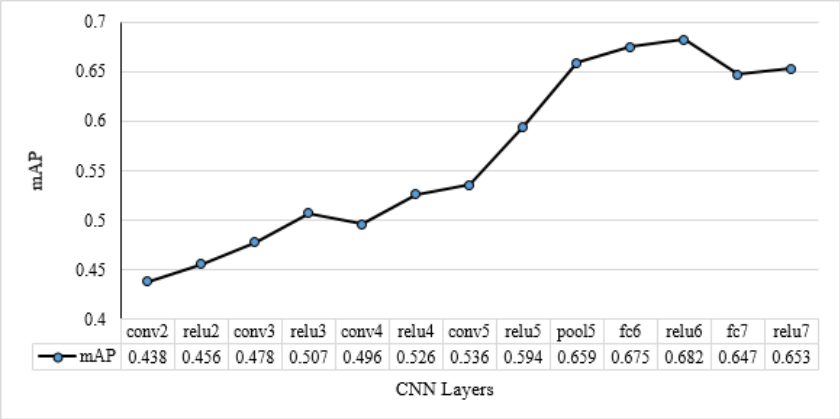

Figure 3. Performance investigation of different CNN layers

4.3.2. Performance comparison with different aggregation methods

The aggregation methods determine the final vector of conv. features. We verify several commonly used aggregation approaches for pool5 conv. features, including BoW [15], VLAD [12], Max-pooling, Sum-pooling, SPoC [3], SCDA [28] and our proposed aggregation method. We set the cluster number of k-means for BoW and VLAD at 128 and 100 respectively. The final aggregated features of all these methods are compacted by PCA with 128 dimensions, and they are normalized by L2-normalization except for BoW.

The results are shown in Table 1. It is obvious that our proposed aggregation method achieves the best performance. More specifically, the sum-pooling method is best suited for conv. layer when comparing with three traditional encoding or pooling methods for aggregating feature maps such as BoW, VLAD and Max-pooling. It is worth noting that the SPoC method cannot outperform Sum-pooling even though the center-prior Gaussian weighting is followed by Sum-pooling. This illustrates that SPoC is not suitable for national costume images with complex backgrounds and multi-objects that are not center-prior. In addition, the performance of the SCDA method is a bit better than Sum-pooling, but it is still lower than our method. For pool5 of conv. with a 6×6 feature map in each channel, after removing some useless deep descriptors by evaluation, only a few deep descriptors in the feature map are left. This could lead to the risk of information loss, especially for some images without obvious objects.

Table 1. Performance comparison with different aggregating methods for pool5 conv. features

Aggregation Methods	BoW	VLAD	Max-pooling	Sum-pooling	SPoC	SCDA	Our
mAP	0.406	0.512	0.610	0.659	0.609	0.664	**0.688**

4.3.3. Investigation of PCA compressed deep features

Dimensionality reduction is one of the key factors concerned to the property of feature representations. In above experiments, we only see its performance with 128-PCA compressed features. Here, we investigate the impact of dimensions of PCA compressed deep features on retrieval performance.

We plot the change curves of mAP with different numbers of principal components reserved in Figure 4. This picture represents two parts of content, including the general and local changes of performances with different PCA compressed dimensions from 1 to 256 and 1 to 32 respectively. From the general change trends, we can see that the best mAP for the pool5 of conv. features and relu6 of FC features is consistently achieved when feature dimensions are within 50. The performances tend to stabilize with a slight drop. Although reduced image representations retain less information, they can

do even better than uncompressed vectors in retrieval tasks. This is perhaps because the discarded secondary components carry redundant information and are useless for image retrieval. From the local changes, it can be seen that as the dimension increases, the mAP is dramatically improved until the highest values for pool5 and relu6 are achieved, which are 8 and 12 respectively. This further suggests that the energy of deep features for national costume images are intensively focused on a few principal components.

To fuse the detail information and semantic information, multi-layers deep features of pool5 and relu6 are concatenated. Figure 5 shows grid searching for the best combination of the principal component reserved when fusing pool5 and relu6 features. From the results, we can see that the best combination of principal components with high performance falls in a local compact region. Considering the low computation, the first 8 and 12 principal components are selected for pool5 and relu6 layers respectively in the multi-layer fusion.

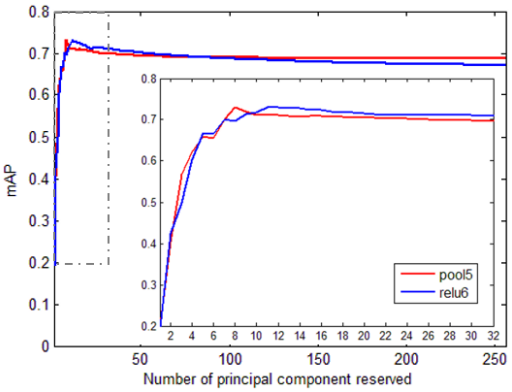

Figure 4. The number of principal component reserved vs. mAP

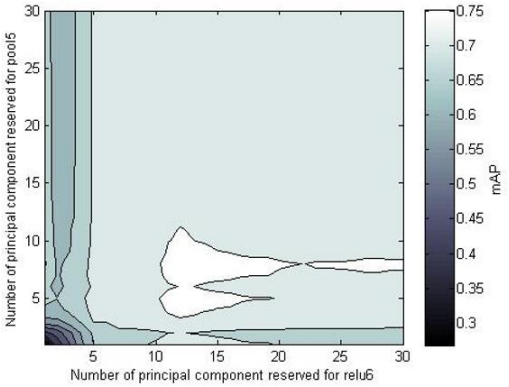

Figure 5. Grid searching for best combination of principal component reserved when fusing pool5 and relu6 layers

4.3.4. Exploiting the effectiveness for re-ranking strategy

Sometimes, even with discriminative visual features and an adaptive similarity metric, the ranking results of CBIR may not be very satisfying. Therefore, some re-ranking techniques become essential to efficiently and effectively retrieve required images. Diffusion processes have shown to be an indispensable tool for improving retrieval performance. Here, we exploit the effectiveness of the diffusion process, comparing it with the common used query expansion (QE) [7] method. From the results shown in Table 2, we can conclude that all re-ranking strategies can dramatically improve the final retrieval performance whether it is for single deep features or multi-layers deep features. In addition, it is obvious that the diffusion process achieves a higher capacity for enhancing the retrieval performance. That is due to its ability to capture intrinsic manifold structure implied in similarity relationship between images in databases.

Table 2. Exploiting the effectiveness for re-ranking strategy

mAP	pool5 (PCA-8)	relu6 (PCA-12)	pool5+relu6
None	0.731	0.730	0.759
with QE (Query Expansion)	0.771	0.773	0.800
with DP (Diffusion Process)	0.808	0.808	0.817

4.3.5. Comparison with existing methods

Based on the previous experimental results, we propose a new approach for national costume image based on multi-layers deep feature representations and a diffusion process. We compare our proposed scheme with some recent methods using low-level visual features in terms of AP while considering 12 retrieved images. Among them, single features such as LBP, HoG and color histogram (CH) are used, and multi-features such as color histogram fused with edge orientation histogram [26] and multi-features fusing color, texture and shape in multi-regions are also used [33].

From the results listed in Table 3, it is obvious that the deep CNN features based on off-the-shelf without fine-turning perform far better than the traditional low-level handcraft features. The proposed method achieves the remarkable improvement comparison with some existing works for national costume images retrieval.

Table 3. Comparison with some existing methods

	Low-level visual features					Deep Convolutional Features			
Method	LBP	HoG	CH	Xu. [26]	Zhao. [33]	pool5	relu6	pool5+relu6	pool5+relu6+DP
AP@12	38.90	41.86	59.78	66.06	68.15	83.58	85.39	88.96	90.82

5. Conclusions

This paper conducts exploratory research based on the practical application of deep learning in the field of image retrieval for national costumes. It then proposes an effective image retrieval approach. In our works, a multi-layer deep feature representation is presented after investigating the performance comparison of deep features extracted from different layers. Meanwhile, a new feature aggregation method is proposed for conv. feature maps. Also, using a diffusion process as a re-ranking strategy is introduced to further enhance the retrieval performance. Compared with some current related methods in the same scope of research, ours achieved more than 20% improvement in retrieval precision. This work will provide some new research ideas and technical reference for researchers in the field. In future works, we will continue to expand the national costume database and consider using fine-turning or transfer learning technology to refine the application of deep learning in the field of image retrieval for national costumes.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and constructive recommendations in improving this manuscript. This work is supported by the National Natural Science Foundation of China (Nos. 61673082, 61462097 and 61562093).

References

1. Ahmad Alzu'bi, Abbes Amira, and Naeem Ramzan, "Content-based Image Retrieval with Compact Deep Convolutional Features," Neurocomputing, vol.249, pp: 95-105, 2017
2. Flora Amato, et al., "Building and Retrieval of 3D Objects in Cultural Heritage Domain," Complex, Intelligent and Software Intensive Systems (CISIS), 2012 Sixth International Conference on. IEEE, pp:816-821, 2012
3. Artem Babenko, and Victor Lempitsky, "Aggregating Local Deep Features for Image Retrieval," Proceedings of the IEEE international conference on computer vision, pp:1269-1277, 2015
4. Dan Ciregan, Meier Ueli, and Jürgen Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," IEEE conference on Computer vision and pattern recognition (CVPR), pp: 3642-3649, 2012
5. Huizhong Chen, Andrew Gallagher, and Bernd Girod, "Describing Clothing by Semantic Attributes," European conference on computer vision. Springer, Berlin, Heidelberg, pp: 609-623, 2012
6. Ken Chatfield, et al., "Return of the Devil in the Details: Delving Deep into Convolutional Nets," arXiv preprint arXiv:1405.3531, 2014
7. Ondrej Chum, et al., "Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval," IEEE 11th International Conference on Computer Vision, ICCV, pp:1-8, 2007
8. Partima Chaudhary, and Sakshi Sharma, "A Color, Texture and Shape Based Hybrid Approach for Clothing Retrieval Techniques," IJMCA, vol.4, no.6, pp: 382-387, 2016
9. Michael Donoser, and Horst Bischof, "Diffusion Processes for Retrieval Revisited," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp:1320–1327, 2013
10. Albert Gordo, et al., "Deep Image Retrieval: Learning Global Representations for Image Search," European Conference on Computer Vision. Springer, Cham, pp: 241-257,2016
11. Esther Hsu, Christie Paz, and Shizhe Shen, "Clothing Image Retrieval for Smarter Shopping," EE368, Department of Electrical and Engineering, Stanford University, pp:1-6, 2011
12. Hervé Jégou, et al., "Aggregating Local Descriptors into a Compact Image Representation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp:3304-3311, 2011
13. M. H. Kiapour, et al., "Where to Buy It: Matching Street Clothing Photos in Online Shops," ICCV, pp: 3343-3351, 2015
14. Chawee LivioDeLuca, Chiara Busayarat Stefani, Philippe Veron, Michel Florenzano,‟ A Semantic-based Platform for the Digital Analysis of Architectural Heritage," Computers & Graphics, vol.35, pp.227-241, 2011
15. Fei-Fei Li, and Pietro Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR. vol. 2, 2005
16. Honglin Li, et al., "Retrieval of Clothing Images based on Relevance Feedback with Focus on Collar Designs," The Visual Computer, vol.32, no.10, pp: 1351-1363, 2016
17. Kevin Lin, et al., "Rapid Clothing Retrieval via Deep Learning of Binary Codes and Hierarchical Search," Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM, pp: 499-502, 2015
18. Ruifan Li, et al., "Retrieving Real World Clothing Images via Multi-weight Deep Convolutional Neural Networks," Cluster Computing pp: 1-12,2017
19. Ying Liu, et al., "A Survey of Content-based Image Retrieval with High-level Semantics," Pattern recognition, vol.40, no.1, pp: 262-282, 2007
20. Ziwei Liu, et al., "Deepfashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 1096-1104, 2016
21. "National Costume Images Database," Available at, http://einapp.ynnu.edu.cn/mzfs/, Last accessed on November 10, 2017

22. Edgar Simo-Serra, and Hiroshi Ishikawa, "Fashion Style in 128 Floats: Joint Ranking and Classification Using Weak Data for Feature Extraction," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp: 298-307, 2016
23. Nidhi Singhai, and K. Shandilya. Shishir. "A Survey on: Content based Image Retrieval Systems." International Journal of Computer Applications, vol.4, no.2, pp: 22-26,2010
24. S. R. Surya, and G. Sasikala, "Survey on Content based Image Retrieval," Indian Journal of Computer Science and Engineering (IJCSE), vol.2, no.5, pp: 691-696,2011
25. Tobias Seppenhauser, Alexander Stenzer, and Burkhard Freitag, "Retrieving Cultural Heritage Information with Google Earth," Web Conference (APWEB), 2010 12th International Asia-Pacific. IEEE, pp:92-98, 2010
26. Xumei Shen, Juxiang Zhou, and TianWei Xu, "Minority Costume Image Retrieval by Fusion of Color Histogram and Edge Orientation Histogram," Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE, pp:365-372, 2016
27. A. Vedaldi and K. Lenc, "MatConvNet – Convolutional Neural Networks for MATLAB," in Proceeding of ACM International Conference on Multimedia, Brisbane, Australia, pp:689–692, 2015
28. XiuShen Wei, et al., "Selective Convolutional Descriptor Aggregation for Fine-grained Image Retrieval," IEEE Transactions on Image Processing, vol.26, no.6, pp: 2868-2881, 2017
29. GuiSong Xia, et al., "Exploiting Deep Features for Remote Sensing Image Retrieval: A Systematic Investigation," arXiv preprint arXiv:1707.07321, 2017
30. Lixuan Yang, et al. "Fully Convolutional Network with Superpixel Parsing for Fashion Web Image Segmentation." International Conference on Multimedia Modeling. Springer, Cham, pp: 139-151, 2017
31. Pan Yang, "Interpretation and Protection of the National Costume Culture," Guizhou Ethnic Studies, vol.37, no.11, pp: 111-114, 2016
32. Wei Yu, et al., "Exploiting the Complementary Strengths of Multi-layer CNN Features for Image Retrieval," Neurocomputing, vol.237, pp: 235-241, 2017
33. Weili Zhao, Juxiang Zhou, Tianwei Xu, "National Costume Image Retrieval Based on Integrated Region Matching," 2nd International Conference on Signal and Image Processing, pp:172-177, 2017

Juxiang Zhou received her Master's degree in 2011 from Yunnan Normal University, Kunming, China. Currently, she is an Assistant Research Fellow at Yunnan Normal University and is pursuing her PhD degree at Dalian University of Technology, Dalian, China. Her research interests include image retrieval and pattern recognition.

Xiaodong Liu received his B.S. and M.S. degrees in Mathematics from Northeastern Normal University in 1986 and Jilin University, Jilin, in 1989, P. R. China respectively. He received his Ph.D. degree in Control Theory and Control Engineering from Northeastern University, Shenyang, P. R. China in 2003. He is currently a professor at the Research Center of Information and Control, Dalian University of Technology and Department of Applied Mathematics, and Dalian Maritime University. He is also a guest professor of the ARC Research Center of Excellence in PIMCE and Curtin University of Technology, Australia. He was a Senior Visiting Scientist in the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton Canada in 2003 and a Visiting Research Fellow in the Department of Computing at the Curtin University of Technology, Perth Australia in 2004.

Jianhou Gan received his B.S. degree in Computer Science Education and M.S. degree in Mathematics from Yunnan Normal University, China, in 1998 and 2004. He received his Ph.D. degree in Computational Metallurgy from Kunming University of Science and Technology, China, in 2016. He is the Vice Director of the Key Laboratory of Educational Informatization for Nationalities, Yunnan Normal University, Kunming, P.R. China. His current research interests include knowledge engineering and educational informatization for nationalities.