Journal of Cultural Heritage 27 (2017) 60–71

Available online at
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com/en

Original article

Bodhisattva head images modeling style recognition of Dazu Rock Carvings based on deep convolutional network

Haiyan Wang[a,b,*], Zhongshi He[a], Yongwen Huang[c], Dingding Chen[a], Zexun Zhou[a]

[a] College of Computer Science, Chongqing University, 400044 Chongqing, China
[b] Department of Art History, Sichuan Fine Arts Institute, 401331 Chongqing, China
[c] School of Electrical and Information Engineering, Chongqing University of Science and Technology, 401331 Chongqing, China

ARTICLE INFO

Article history:
Received 23 January 2017
Accepted 21 March 2017
Available online 10 May 2017

Keywords:
Deep convolutional network
Modeling style recognition
Dazu Rock Carvings
VGGNet
Exemplar-based inpainting

ABSTRACT

As the only World Culture Heritage in Chongqing, the Dazu Rock Carvings have been suffering from natural environment erosion for almost one thousand years, so the inpainting of these carvings is desired. In order to reduce the inpainting risk and keep the carvings' original appearance, it is necessary to introduce digital image processing techniques to perform virtual inpainting initially. The key step in exemplar-based inpainting algorithm is to search for the most similar patch. Efficient modeling style recognition is therefore the basis of the carvings' inpainting. Focusing on Bodhisattva head images, this paper proposes a two-step recognition method for their modeling style: feature extraction utilizing VGGNet[1] and clustering with K-means algorithm. The proposed method obtains promising results compared with 5 classical feature extraction algorithms. According to experiment results, combining both art archaeology and image analysis, we conclude: (1) the modeling style is similar for the statues in the same cave or region; and the modeling style of the statues on the same subject is also similar, even though they are in different caves or regions; (2) the name of Beishan N° 180 should be "Cave of Eleven Incarnations of Avalokitesvara" instead of the previous "Cave of Thirteen Incarnations of Avalokitesvara". It is because the modeling style of two statues behind the major statue is quite different from the others. They were probably carved to substitute the two outmost damaged statues. Our work can be viewed as a reference to solve some art problems. Moreover, as it is efficient to search for similar images by our method, this paper can be as the basis for virtual exemplar-based inpainting of Dazu Rock Carvings in our future work.

1. Introduction and research aim

With the development of pattern recognition and image processing technologies in computer science, these techniques have been widely applied to heritage digital conservation and protection such as image preservation, classification, inpainting, digital museum and VR show etc. As the pinnacle of Chinese cave temple art dating from the 9th to the 13th century, Dazu Rock Carvings in Chongqing were recognized as the World Cultural Heritage in 1999 for their outstanding universal value. They are comprised of the cliffside carvings at Beishan, Baodingshan, Nanshan, Shizhuanshan and Shimenshan. Appraised by UNESCO [1], the Dazu Rock Carvings are remarkable for their aesthetic quality, rich diversity of subject matter, both secular and religious and the light that they shed on everyday life in China during this period. And they provide splendid evidence of the harmonious synthesis of Buddhism, Taoism and Confucianism.

Located outdoors and due to moist climate, shallow grotto and niche depth, the carvings have been inevitably suffering from surface weathering, water seepage, rock instabilities, biological erosion and other natural environment threats. Conservation measures are becoming extremely urgent. To preserve the integrity of the World Cultural Heritage, the conservation must be strictly adhered to the principle of "retaining the historic condition", such as the design, materials, technology and layout. So the protectors are cautious of the restoration to prevent the carvings from risks. Therefore, we take priority to seek a more efficient way of combining traditional methods and modern technologies. Scientific methods, such as digitalization, virtual imaging and multimedia, have been successfully applied to the inpainting of Thousand-armed Avalokitesvara at Baodingshan, lasting from the year 2008 to 2015, and the Museum of Dazu Rock Carvings, established in 2015. However, compared with the mature utilization of scientific

[*] Corresponding author. College of Computer Science, Chongqing University, Chongqing 400044, China. Tel.: +86 1 8523535225.
E-mail address: haiyanwang@cqu.edu.cn (H. Wang).

[1] VGGNet is a very deep convolutional network for large-scale image recognition proposed by Visual Geometry Group of Oxford University.

http://dx.doi.org/10.1016/j.culher.2017.03.006
1296-2074/© 2017 Elsevier Masson SAS. All rights reserved.

technologies in conservation and protection of Dunhuang Frescoes in China, the application of image processing techniques to Dazu Rock Carvings has just started.

Our work focuses on virtual exemplar-based inpainting of the carvings. The critical step of exemplar-based inpainting is to search the most similar patch. It requires an efficient recognition algorithm to match image similarity. Based on Bodhisattva statues, this paper addresses their modeling style recognition and constructs a digital image sample dataset, providing the basis for further researches on permanent digital conservation, virtual inpainting and display. The final aim is to preserve and spread Dazu Rock Carvings' culture across time and space.

The process of image recognition includes three steps:

- preprocessing;
- feature extraction;
- recognition.

Feature extraction is critical to the whole process, since the classifier could not recognize images based on poorly selected features. Criteria to feature selection given by Lippman is: "Features should contain information required to distinguish between classes, be insensitive to irrelevant variability in the input, and also be limited in number to permit efficient computation of discriminant functions and to limit the amount of training data required" [2]. Obviously, the goal of feature extraction is to obtain the most relevant information from the original data and represent it in a lower dimensionality space [3]. Effective characterization can be directly concluded from the original image with a few preprocessing with the Convolutional Neural Networks (CNNs). Pointed by Yann Lecun et al. [4], CNN is a dominant approach for almost all recognition and detection tasks [5–10]. Therefore, CNN becomes our first choice for Dazu Bodhisattva image recognition. VGGNet [10] was the champion for localization and the second place for classification on ILSVRC[2] in 2014 and it has been successfully applied to face recognition. We choose VGGNet to extract features because our research mainly involves the recognition and restoration of the Bodhisattva statue's head.

2. Related work

2.1. Convolutional neural networks

Inspired by biological natural visual perception mechanism [11] and neocognitron [12], deep CNN architecture with multiple layers is designed. Its idea is to stack up multiple layers and the output of the previous layer is the input of the next layer. Thus, it could obtain high-level features by combining the low-level features and learn the abstracted feature presentation of the input data. Due to its efficiency, CNN has been widely applied in image vision. LeCun et al. [13] first established a CNN architecture named LeNet-5 to make classification of handwritten numbers in 1990s. Following LeNet-5, various architectures have been designed to overcome the difficulties in training depth of CNN, as the network can better approximate the target function and conclude better feature representations by increasing the depth. In 2012, Krizhevsky et al. [14] proposed a classic CNN architecture named AlexNet, which was deeper than LeNet-5, and made a breakthrough in image recognition. Based on AlexNet, more CNN architectures have emerged, such as ZFNet [15], VGGNet [10], GoogleNet [5] and ResNet [16]. The recent advances in CNN were reviewed in [17].

The basic components are very similar for different variants of CNN architectures. It is comprised of 3 layer types:

- convolution (Conv.) layer;
- pooling (P) layer;
- full-connected (FC) layer.

Conv. layer is composed of several Conv. kernels. Its main task is to extract features and learn their representations. Each unit in a Conv. layer is connected to local patches in the feature maps of the previous layer by a set of weights called filter bank. The result of the local weighted sum is then passed through an activation function. Thus, the network can get quantities of different features and extracts the features in a loop to get highly complex feature in combination. According to [17], the feature value $z^l_{i,j,m}$ at location (i,j) in the m-th feature map of l-th layer is calculated by:

$$z^l_{i,j,m} = (w^l_m)^T x^l_{i,j} + b^l_m \tag{1}$$

where w^l_m and b^l_m are the weight vector and bias term of the m-th filter of the l-th layer respectively; $x^l_{i,j}$ is the input patch centered at location (i,j) of the l-th layer; T denotes the transpose of matrix. In the same feature map m, all the feature values share the same weight vector.

P layer, also called subsampling, is used to aggregate features, i.e., merging semantically similar features into one. It is the most important component of CNN, as it is an efficient method to reduce the feature dimension. On this stage, the local correlation of the image is used for sampling and meanwhile keeps its useful information. It can reduce the computation complexity.

FC layer is used to stack up the Conv. and P layers to form a full connection layer or multi-layers, generating global semantic information. Softmax operator is commonly added after the FC layers for classification tasks.

Thus, the information is transmitted to different layers sequentially in CNN. The input of the lowest level of hierarchy is a small part of the image. The most distinguishing features of the input are captured through a digital filter on each layer.

2.2. VGGNet

Inheriting from the architecture of Lenet-5 and Alexnet, numerous improvements have been made in better accuracy [8,16], networks dense [8,18] and the depth. VGGNet addressed how the depth of the CNN architecture had affected the performance. In order to make a fair evaluation, VGGNet utilized the same generic design, fixed other parameters of the architecture, but steadily increased the depth of the network from 11 (A architecture) to 19 (E architecture) by adding more Conv. layers. It was feasible due to the use of very small (3 × 3) Conv. filters in all layers.

Among 5 different architectures, D includes 16 weight layers (called VGGNet-16 in this paper) and E includes 19 weight layers. According to the comparison experiments in [10], E performs little better than D, even though it adds 3 Conv. layers at the expense of time complexity increase. Hence, we utilize D architecture in this research, which includes 13 Conv. layers and 5 P layers, 3 FC layers and 1 soft-max. Conv. layers and P layers construct 5 stages (differing from 3 stages in popular networks), combining 2 FC layers to extract features, 1 FC layer and soft-max are used for classification.

3. Proposed method

The recognition in this paper is first to aggregate Bodhisattva head images into different clusters and then identify which cluster or modeling style one specific image belongs to. Therefore, we can

[2] ILSVRC is short for the annual ImageNet Large Scale Visual Recognition Challenge. It is a famous software contest where software programs compete to correctly classify and detect objects and scenes.

search its most similar image in a specific cluster instead of different clusters in sequence.

For efficiently finding the most similar image, our proposed algorithm includes 4 stages:

- image preprocessing;
- feature extraction;
- clustering;
- image recognition.

Bohdisattva head images are firstly resized for normalization. Prominent features of images are then extracted with VGGNet-16. Style recognition, including similarity comparison and similar image search, is performed by K-means clustering.

3.1. *VGGNet-16-based feature extraction*

We first extract prominent feature representations of Bohdisattva head images based on VGGNet-16. The pseudo code of the VGGNet-16-based feature extraction algorithm is shown below.

Algorithm 1 VGGNet-16-based feature extraction

```
Input:
Image set: $\boldsymbol{I} = \{p_1, p_2, \ldots, p_n\}$
VGGNet-16 parameter and construction
Output:
Feature vector set of: $\boldsymbol{I} : \boldsymbol{F} = \left\{f_1, f_2, \ldots, f_n\right\}$
load $\boldsymbol{I} = \{p_1, p_2, \ldots, p_n\}$, VGGNet-16 parameter and construction
initialize $\boldsymbol{F} = \phi$
for each $p_i \in \boldsymbol{I}$ do
    preprocess $p_i$
    read$p_i$ into input data layer of VGGNet-16
    extract feature vector $f_i$ of $p_i$
    add $f_i$ *into* $\boldsymbol{F} : \boldsymbol{F} = \boldsymbol{F} \cup \left\{f_i\right\}$
end
output $\boldsymbol{F} = \left\{f_1, f_2, \ldots, f_n\right\}$
```

Our VGGNet-16-based feature extraction architecture is shown in Fig. 1. The input is a 224 × 224 RGB image. After it passes through 13 Conv. layers (2 Conv. layers and 1 P layer in C1-2 respectively, 3 Conv. layers and 1 P layers in C3-5 respectively), we get its feature maps from low-level to high-level. These feature maps are combined to a 1 × 4096 feature vector through 2 FC layers (FC6-7). At last, we obtain the output of a 1 × 4096 feature vector corresponding to the input image.

3.2. *K-means Clustering*

K-means clustering, one of the most popular and widely used learning algorithms for clustering analysis, is applied to cluster images. This method aims to find the best division of n entities into k clusters with the minimization of the distance between the cluster members and its corresponding centroid (representative of the group).

After loading data set $\boldsymbol{D} = \left\{f_1, f_2, \cdots, f_n\right\}$, set of entities to be clustered, we initialize clusters' centroids with the value of k and set iteration times $rep = 15$, best performance in executing. We calculate the similarity between two entities with *Cosine similarity* instead of commonly used *square value of Euclidean distance* to find the 5 most similar images to each input image. *Cosine similarity* is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The smaller the angle is, the more similar they are, and the value is more approximate 1. It defined as

$$\cos\theta = \frac{\sum_{k=1}^{n} f_{1k} f_{2k}}{\sqrt{\sum_{k=1}^{n} f_{1k}^2 f_{2k}^2}} \tag{2}$$

where $\theta = \left\langle f_1, f_2 \right\rangle, f_i = \left(f_{i_l}, \ldots f_{i_k}, \ldots f_{i_n}\right), i = 1, 2.$

According to the most common algorithm, our clustering can be described as follows:

Algorithm 2 K-means clustering

```
Input:
  Data set: $\boldsymbol{D} = \left\{f_1, f_2, \ldots, f_n\right\}$
Output:
  Cluster centroids: $\boldsymbol{C} = \{c_1, c_2, \ldots, c_k\}$
Cluster labels for $\boldsymbol{D}$:$\boldsymbol{L} = \left\{lf_i\right\} | i = 1, 2, \ldots, n$
1: load $\boldsymbol{D} = \left\{f_1, f_2, \ldots, f_n\right\}$
2: initialize $k$ and $c_i (1 \le i \le k)$
3: set iteration times $rep = 15$
4: for each $f_i \in \boldsymbol{D}$ do
5:    calculate $d_{ij} = \cos\theta$ where $f_1 = f_i, f_2 = c_j$
6:    $l(f_i) = j^*$ where $d_{ij*} = \min_j \left\{d_{ij} | j = 1, \ldots, n\right\}$
7: end
8: update cluster centroid $c_j (j = 1, 2, \ldots, k)$
9: repeat step 4–7 till $rep$ or $\boldsymbol{C}$ no longer change
10: output $\boldsymbol{C} = \{c_1, c_2, \ldots, c_n\}, \boldsymbol{L} = \left\{lf_i\right\} | i = 1, 2, \ldots, n$
```

4. Experiments and analysis

To validate the efficiency of the proposed method, we test it on abundant image data. The image data composes of Bohdisattva head RGB images of 6 different caves or niches at 3 different hills (Beishan, Shimenshan and Baodingshan). There is only 1 Taoism statue cave at Nanshan and Shizhuanshan is famous for its Confucianism carvings, so we do not take them into account in this paper. The digital images are partly provided by the Dazu Institute and mainly photographed by the authors. And the head (face and crown) part of each Bodhisattva statue image is captured in Photoshop, with length-width ratio of 5:8. Feature extraction based on VGGNet-16 is performed on Convolution Architecture for Feature Extraction. The clustering and style recognition are implemented on Matlab7.8.0 (R2009a).

We first collect 114 digital images, as shown in Table 1, in which 3 caves (BS136, BS180, SM6) are on the same subject, the others are on different subjects.

4.1. *Modeling style clustering of all the statues*

We initialize predicted categories k to be 6 (6 different caves/niches), 5 (BS136 and BS180 are on the same hill and on the same subject), 4 (BS136, BS180 and SM6 are on the same subject), 3 (3 different hills) respectively for the clustering experiments with features extracted by VGGNet-16. The clustering results in Fig. 2 illustrate:

- the modeling style of the statues at the same hill is much similar to each other and the modeling style of the statues on the same subject shares a high similarity. In Fig. 2(a), we can see 88% images of BD11 and 100% of BD18 are clustered to be one group, 100% of SM6 are in another group. Fifty percent of BS136 and 20% of BS180 are always within the group including SM6, shown in Fig. 2(a)–(d), because statues of the 3 caves are on the same subject of incarnations of Avalokitesvara, and Buddhist images are

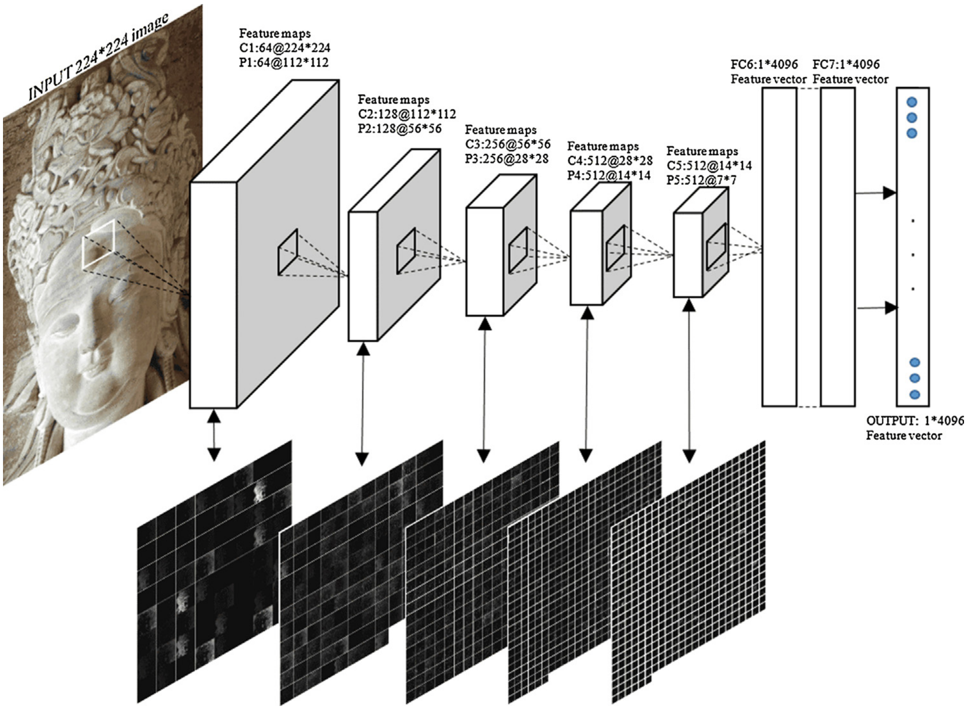


Fig. 1. Architecture of VGGNet-16-based feature extraction.

Table 1
Data 1: statues of Beishan Nº 136 & 180, Shimenshan Nº 6 and Baodingshan Nº 29,11&18.

Number of the cave\niche	Total images
Beishan Nº 136 (BS136), Cave of Prayer Wheel, A.D. 1142–1146.	8
Beishan Nº 180 (BS180), Cave of Thirteen Incarnations of Avalokitesvara, A.D. 1116–1122.	14
Shimenshan Nº 6 (SM6), Cave of Ten Avalokitesvaras, A.D. 1136–1141.	20
Baodingshan Nº 11 (BD11), Niche of Sakyamuni Entering Nirvana, A.D. 1174–1252.	31
Baodingshan Nº 18 (BD18), Sutra of Amitabha and His Pure Land, A.D. 1174–1252.	26
Baodingshan Nº 29 (BD29), Cave of Full Enlightenment, A.D. 1174–1252.	15

always with a fixed pattern for a specific Bodhisatva. As shown in Fig. 3(a)–(c), their face modeling style is almost the same;

- the statue images of BS136 and BS180 are difficult to be aggregated into different clusters. From Fig. 2(a)–(d), we can find that 50% images of BS136 and 80% of BS180 are always within the same cluster. Because they are not only both at the same hill but also on the same subject, shown in Fig. 3(a)–(b);
- when we set k to be 4 and 3, 86% of BD29 are clustered into the group including SM6, shown in Fig. 2(a)–(b). Maybe the statues in both caves are suffering from the same weathering disease. The former is the squamous flaking and curling of gold fods, whereas the latter is the peeling of color surface, referenced in Fig. 3(c)–(d).

4.2. Similarity comparison of statues at different hills

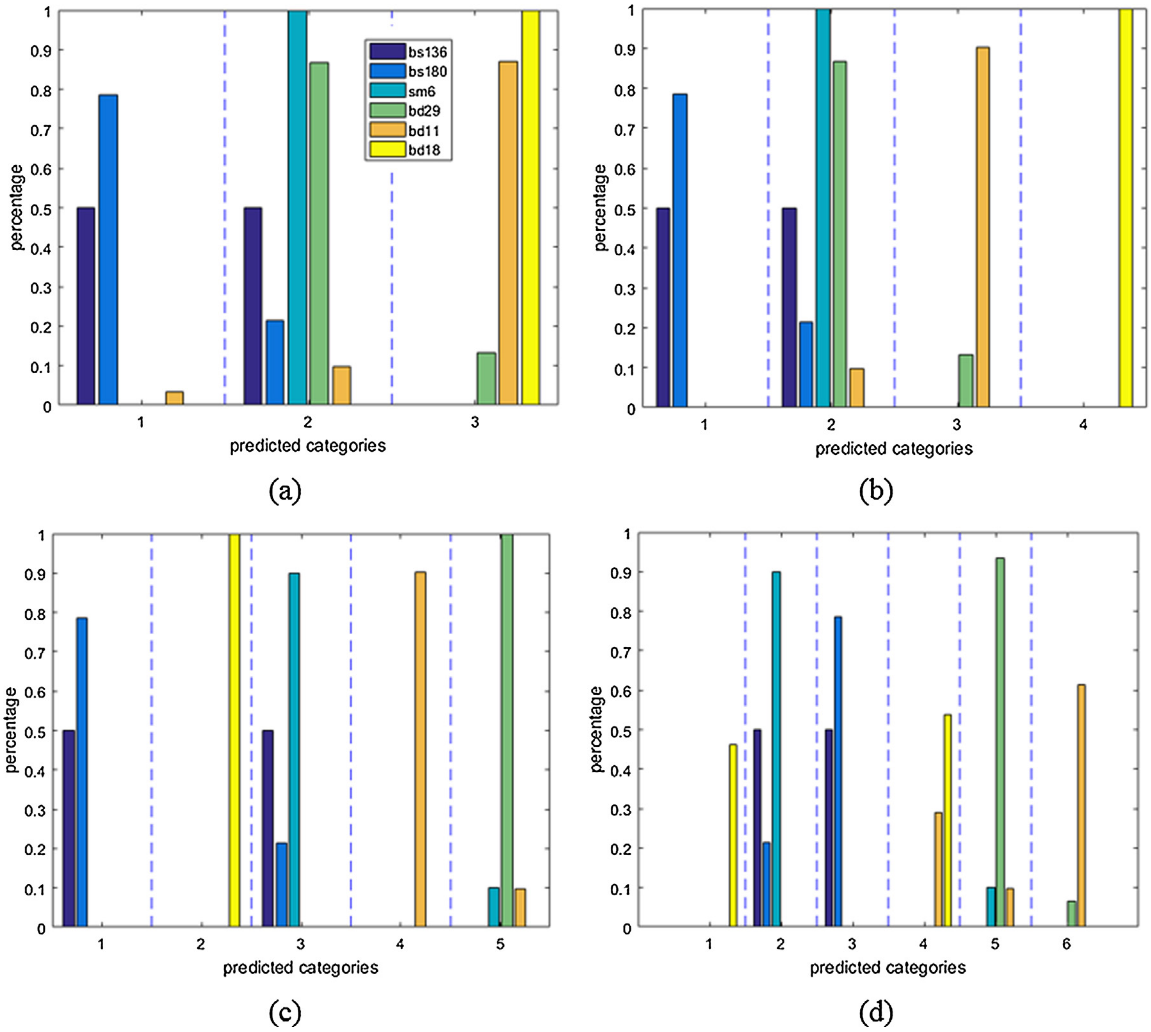
For the modeling style recognition experiments, the authors first choose the statues in two caves at different hills to compare the similarity in this section. Similarity between two images is calculated with *Cosine similarity*. The similarity value varies from 0 to 1 and the higher it is, the more similar the two images are. The input images, shown in Fig. 4, are 6 statue images (Nº 1-6) of BS136 and 6 statue images (Nº 7-12) in the east of BD29.

Suppose image Nº (Img.)1-6 and Img.7-12 are in two groups as they belong to different caves at different hills, paired similarity of 12 images is reviewed in Table 2. Except the value of 1, we highlight those similarity value greater than 0.9 in bold, in which the highest in each row is marked in italics. It illustrates that the most similar pictures of Img.1-6 nearly all fall into the first group, and the similarity values greater than 0.9 are evenly distributed in the first group as well. Similar results are also revealed for Img.7-12. Moreover, for the comparison of the images between two groups, most results are below 0.9 and a great more below 0.85. Therefore, the paired comparison results well verify the statues' difference between two caves in different regions.

There is an outlier in Table 2. Img.12 is the most similar image to Img.1, with a similarity value 0.90578, though they belong to different groups. It may be attributed to less valuable face information in Img.1. Even so, the 2nd and 3rd most similar images to Img.1 are Img.5 and Img.3, both belonging to the same group; the most similar image to Img.12 is not Img.1 but Img.10, with a similarity value 0.94721. Moreover, the similarity value between Img.12 and Img.8, Img.11 are 0.93415, 0.91716 respectively, exceeding the value 0.90578.

The comparison between computer results and human experts in estimating similarity is shown in Table 3. There are 3 differences among 12 comparisons. The relative error of Img.3 is 2.2% = (0.93967–0.91832)/0.93967 × 100% and the relative error of Img.1 and Img.6 are 4.4% and 3.2% respectively. Obviously, the accuracy of similarity assessment of Img.7-12 is much better than that of Img.1-6, because Img.7-12 are all front head images while Img.1-6 are multi-view head images.

These errors indicate that the head rotation angle affects its recognition and human eyes are more interested in the face part (local information), while computers focus on the full image (global



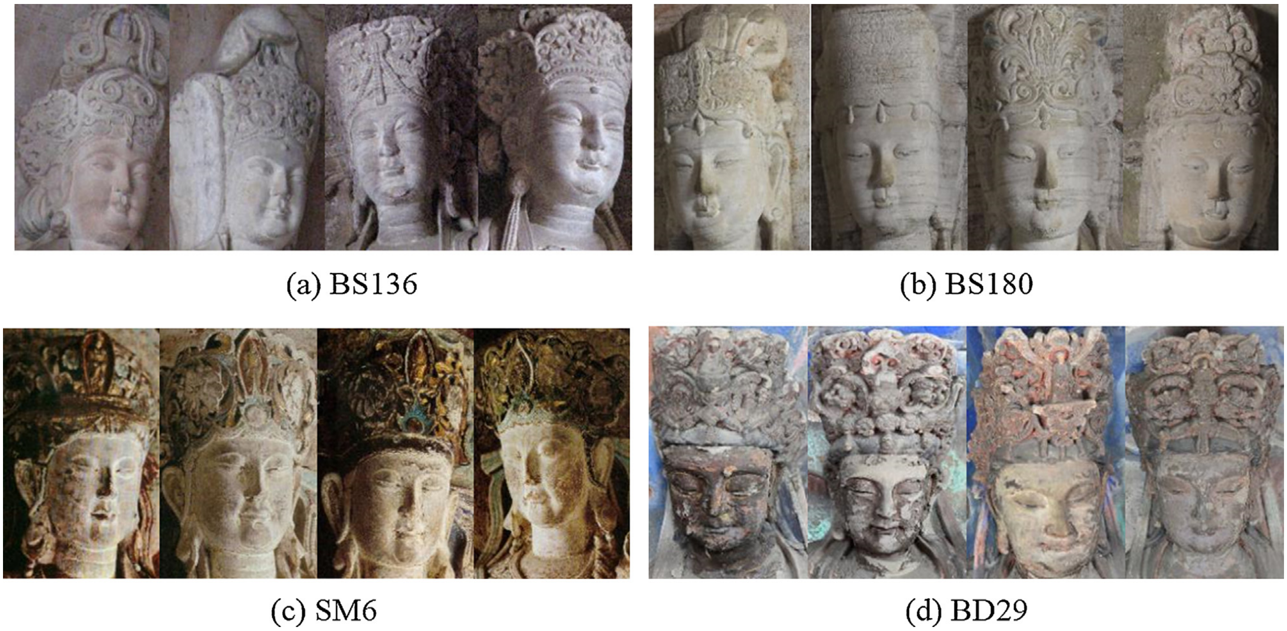
(a)

(b)

(c)

(d)

Fig. 2. Clustering results of BS136 & 180, SM6 and BD29, 11 & 18.



(a) BS136

(b) BS180

(c) SM6

(d) BD29

Fig. 3. Some images of BS136, BS180, SM6, and BD29.

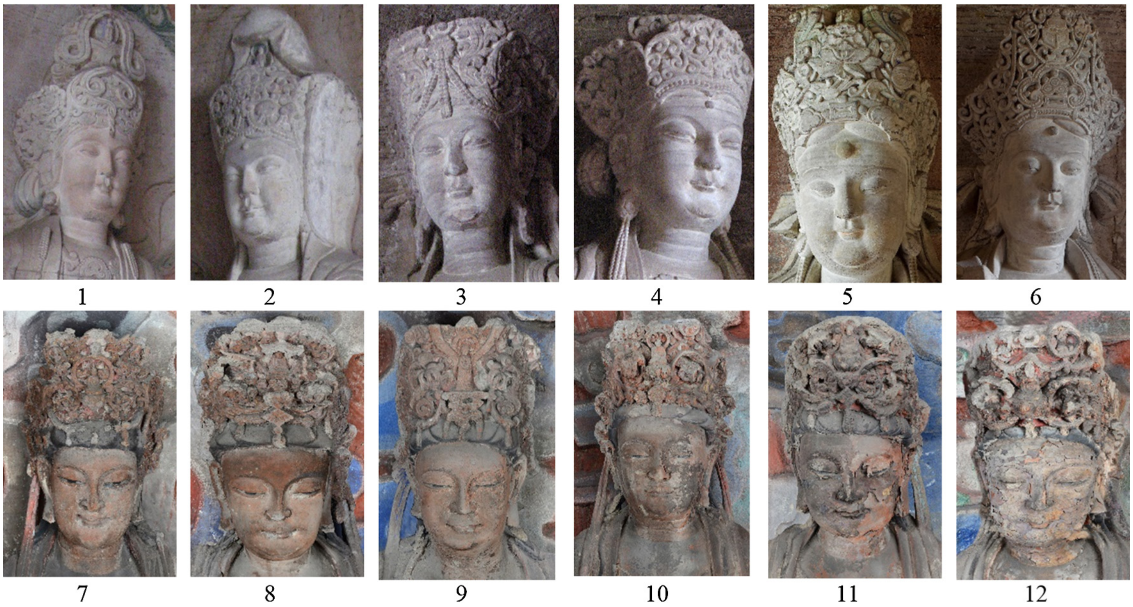


Fig. 4. Data 2: BS136 and BD29.

Table 2
Paired similarity value between BS136 and BD29.

Img.	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.86367	0.89246	0.86542	0.89606	0.86743	0.85148	0.88614	0.86316	0.88352	0.83915	***0.90578***
2	0.86367	1	**0.90226**	0.88538	***0.90657***	0.88842	0.82807	0.82380	0.85029	0.81005	0.81858	0.78807
3	0.89246	**0.90226**	1	**0.91832**	**0.93043**	***0.93967***	0.85559	0.86220	0.84096	0.81228	0.80010	0.81135
4	0.86542	0.88538	***0.91832***	1	0.8614	**0.91518**	0.82958	0.77637	0.80586	0.78061	0.76621	0.78496
5	0.89606	**0.90657**	***0.93043***	0.8614	1	**0.90882**	0.8293	0.87987	0.83616	0.84365	0.87027	0.85020
6	0.86743	0.88842	***0.93967***	**0.91518**	**0.90882**	1	0.81355	0.83454	0.80924	0.77267	0.80466	0.79198
7	0.85148	0.82807	0.85559	0.82958	0.8293	0.81355	1	0.87544	*0.89242*	0.85614	0.81358	0.83961
8	0.88614	0.82380	0.86220	0.77637	0.87987	0.83454	0.87544	1	**0.92618**	**0.92255**	**0.92095**	***0.93415***
9	0.86316	0.85029	0.84096	0.80586	0.83616	0.80924	0.89242	***0.92618***	1	**0.92340**	0.88214	0.89845
10	0.88352	0.81005	0.81228	0.78061	0.84365	0.77267	0.85614	**0.92255**	**0.92340**	1	**0.90326**	***0.94721***
11	0.83915	0.81858	0.80010	0.76621	0.87027	0.80466	0.81358	***0.92095***	0.88214	**0.90326**	1	**0.91716**
12	**0.90578**	0.78807	0.81135	0.78496	0.8502	0.79198	0.83961	**0.93415**	0.89845	***0.94721***	**0.91716**	1

Table 3
Comparison between computers and human experts in assessing similarity.

Img.	Similar image searched by computers		The most similar image searched by human experts	Consistent result (Y/N)
	1st similar Img.	2nd similar Img.		
1	12 (0.90578)	5 (0.89606)	4 (0.86542)	N
2	5 (0.90657)	3 (0.90226)	3	Y
3	6 (0.93967)	5 (0.93043)	4 (0.91832)	N
4	3 (0.91832)	6 (0.91518)	3	Y
5	3 (0.93043)	6 (0.90882)	6	Y
6	3 (0.93967)	4 (0.91518)	5 (0.90882)	N
7	9 (0.89242)	8 (0.87544)	9	Y
8	12 (0.93415)	9 (0.92618)	9	Y
9	8 (0.92618)	10 (0.92340)	8	Y
10	12 (0.94721)	9 (0.92340)	9	Y
11	8 (0.92095)	12 (0.91716)	8	Y
12	10 (0.94721)	8 (0.93415)	10	Y

information). Thus, our future work should first adjust statue angles and extract local features to improve the accuracy of similarity assessment.

4.3. Similarity comparison of statues in different caves at the same hill

The authors then choose the statues in two different caves at the same hill to compare the similarity and search for the 5 most similar images to each image. The input data 3 comprises of 31 statue images of BD11 (Img.1-31) and 26 images of BD18 (Img.32-57).

Assuming Img.1-31 in the 1st group and Img.32-57 in the 2nd group, shown in Fig. 5, we can find that the 5 most similar images of Img.1-31 are evenly distributed in the 1st group, whereas the same result is available for Img.32–57. From detailed analysis in Fig. 6, the 5 most similar images (Img.6,4,17,25,26) to Img.7 are from the 1st group and the 5 most similar images (Img.52,57,42,35,38) to Img.53 are from the 2nd group. These results illustrate that the statues in the same cave are much similar to each other, but much different

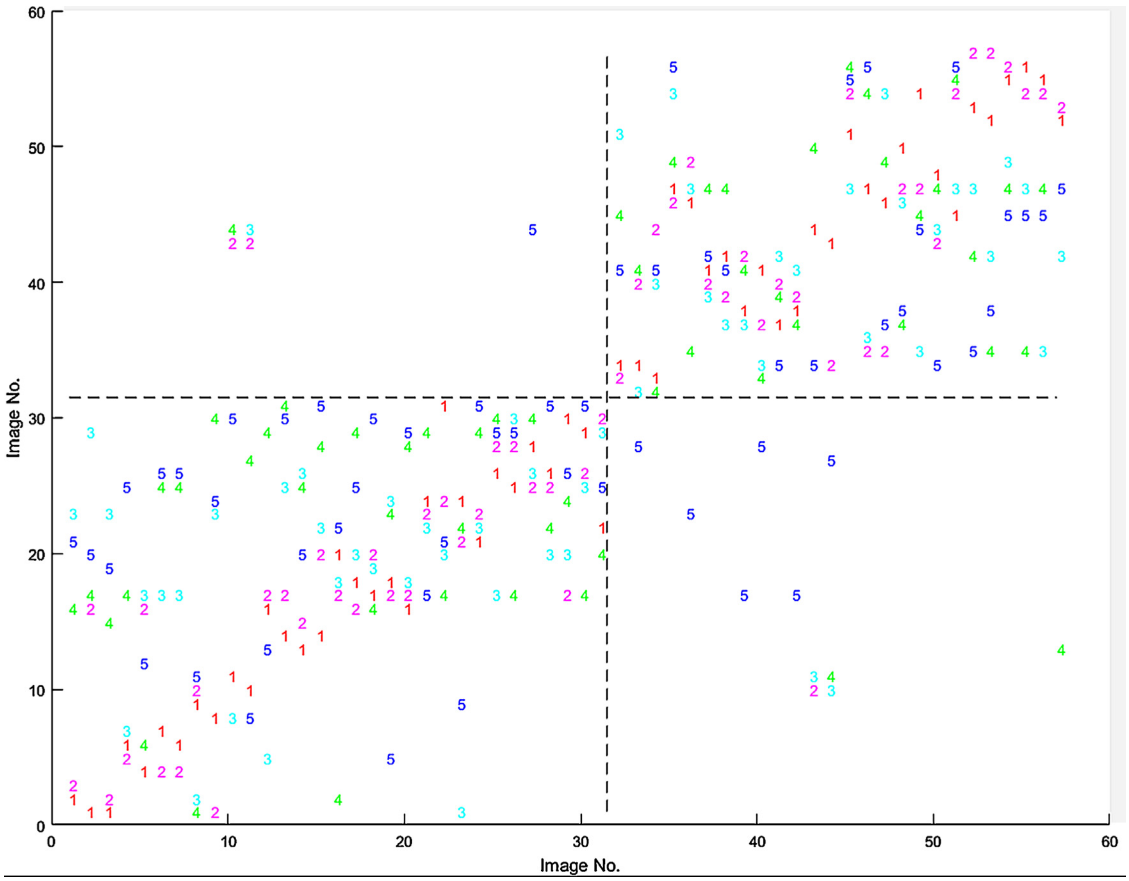


Fig. 5. Distribution of the 5 most similar images to every image of BD11&18, figures 1–5 stand for the 1st, the 2nd, the 3rd, the 4th and the 5th similar respectively.

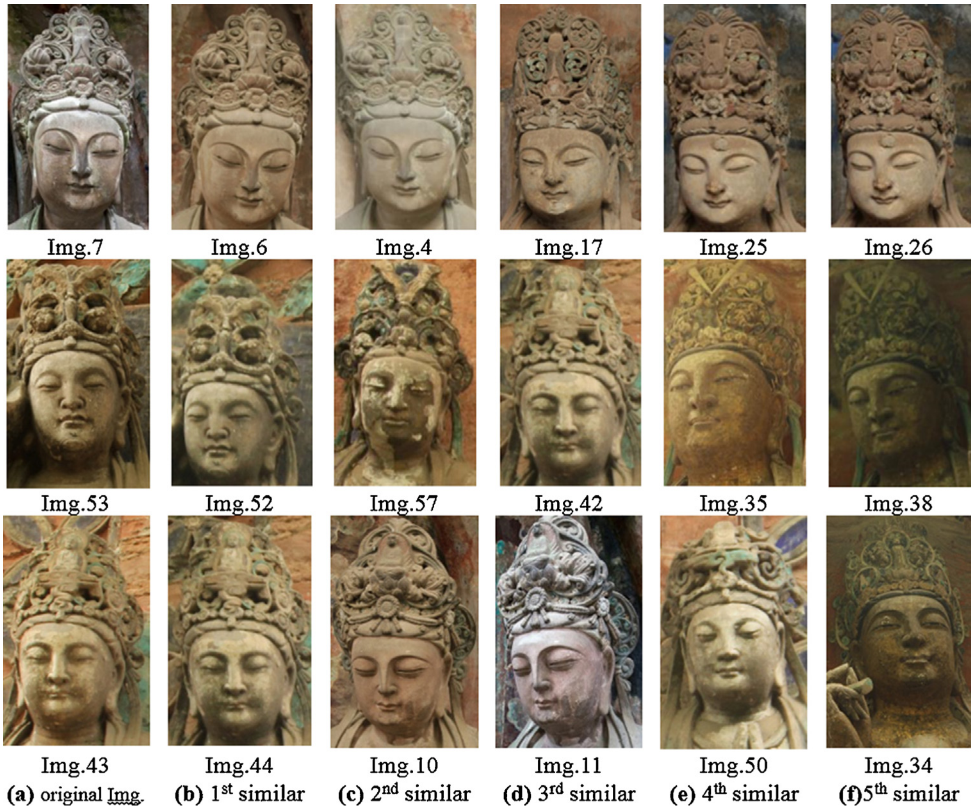


Fig. 6. The 5 most similar images to Img.7, 53, 43 respectively.

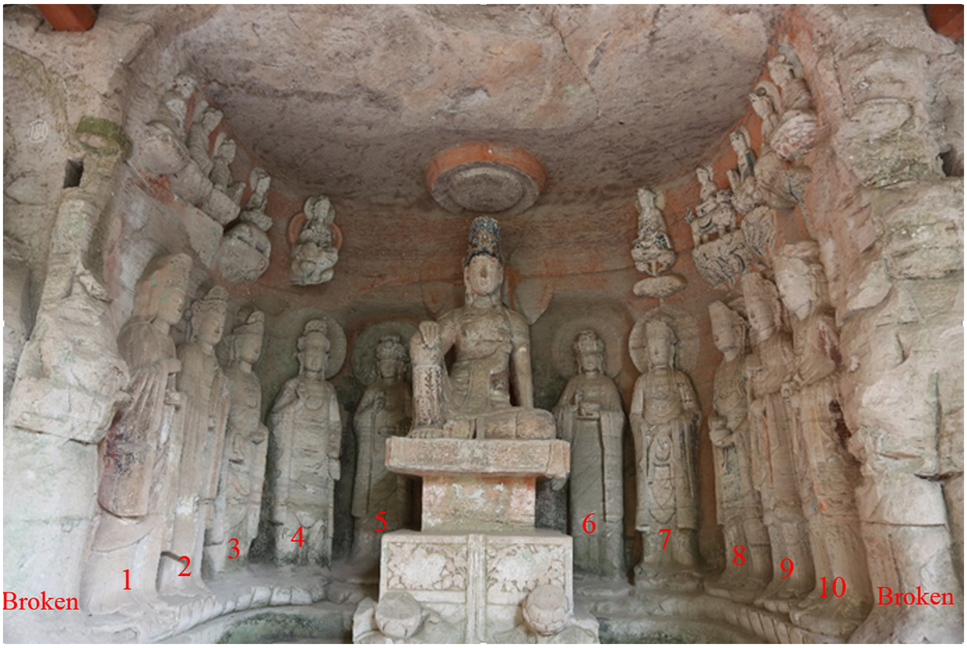


Fig. 7. Cave of BS180.

Table 4
Clustering comparison of 6 different algorithms.

	Clust							
	3		4		5		6	
Algor	Intra-dist	Inter-dist	Intra-dist	Inter-dist	Intra-dist	Inter-dist	Intra-dist	Inter-dist
VGGNet-16	**0.1964**	0.2507	**0.1354**	0.2678	**0.1015**	0.2669	0.0813	0.2393
LGBP_p	**0.1964**	0.2507	**0.1354**	0.2678	0.1020	0.2385	**0.0794**	0.2379
LBPH	1.6303	0.5283	1.1400	0.4944	0.8647	0.5387	0.6974	0.5621
LDA	6.9337	**1.3504**	4.1801	**1.3056**	2.8884	**1.3056**	2.0664	**1.3183**
LGBP_m	0.3751	0.2736	0.2564	0.2462	0.1943	0.2732	0.1544	0.2547
PCA	9.7420	1.2631	5.6528	1.2709	4.0052	1.2424	2.9797	1.2473

from the statues in the other caves, although they are located on the same hill.

There are a few outliers, e.g., Img.43,44 and Img.10,11 have two similar images in the other group. Picking up these images in bottom row of Fig. 6, we can see that they have many similar features, such as round face, long and thing eyebrows and eyes, straight nose, slightly raised mouth corner and gentle smile. It is because Baodingshan Rock Carvings were project designed by a famous monk Zhifeng Zhao in Southern Song Dynasty with an overall blueprint, and implemented carvings across 79 years.

4.4. Efficiency comparison

This paper applies 5 classical feature extraction algorithms, including local Gabor binary patterns-phase (LGBP-p) [19], local binary patterns histograms (LBPH) [20], linear discriminant analysis (LDA) [21], local Gabor binary patterns-magnitude (LGBP-m) [22], principle component analysis (PCA) [23], to compare the clustering efficiency with the above proposed method. Firstly, we extract features of Data 1 with these feature extraction methods; then perform the clustering based on K-means method. All the clustering experiments are executed with $k = 3, 4, 5, 6$ respectively and 15 iterations. We compute the average of the intra-cluster distance (intra-dist) and inter-cluster distance (inter-dist). Shown in Table 4, the value in bold stands for the best performance in column. We can find that the clustering based on VGGNet-16 and LGBP-p are not only with the minimum average intra-dist but also with the minimum average inter-dist. Whereas the clustering based on LDA is with the maximum average inter-dist but with high average intra-dist. Therefore, future work combining VGGNet and LDA to take into account both intra-aggregation and inter-separation is recommended, so as to enhance the recognition accuracy.

5. Renaming Beishan NO.180

5.1. Clustering and similarity comparison

In this section, our image data is 217 images of 10 Avalokitesvara statues with different angles of BS180, “Cave of Thirteen Incarnations of Avalokitesvara”. Saint Avalokitesvara is in the middle of the cave. There are six incarnations of Avalokitesvara on both sides, all barefoot on the holy lotus flowers. Two statues outmost the left and right wall are broken, leaving only two discernible bodies, as shown in Fig. 7. Therefore, our image data contains only 10 statues, illustrated in Table 5.

5.1.1. Clustering results

We set k to be 2 to 10 clusters respectively; the clustering results are illustrated in Fig. 8.

From Fig. 8, we can learn that no matter how many clusters (2–10) to set, the images of statue (Sta.)5 and Sta.6 can always be clearly and entirely aggregated into one cluster 6 times and

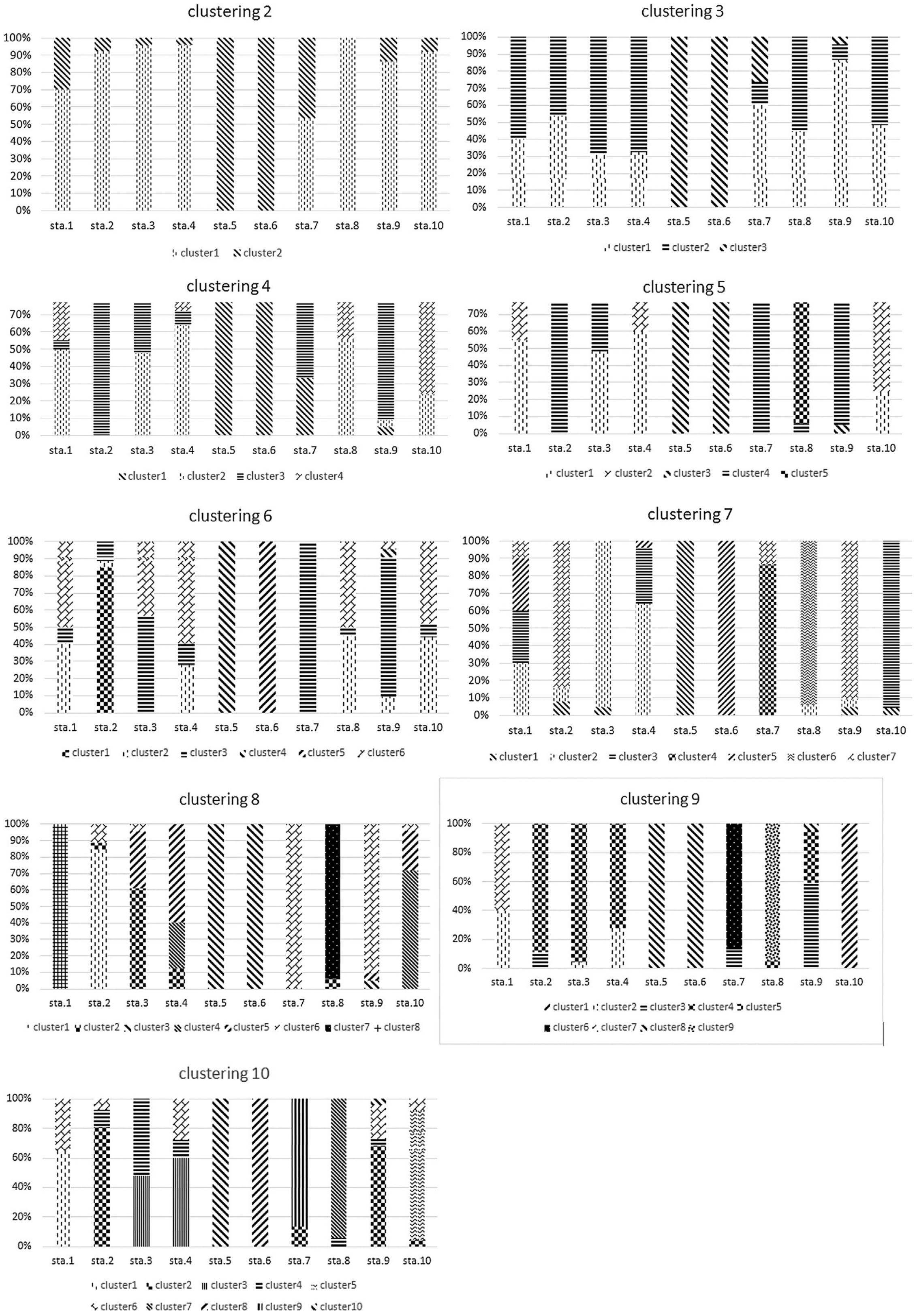
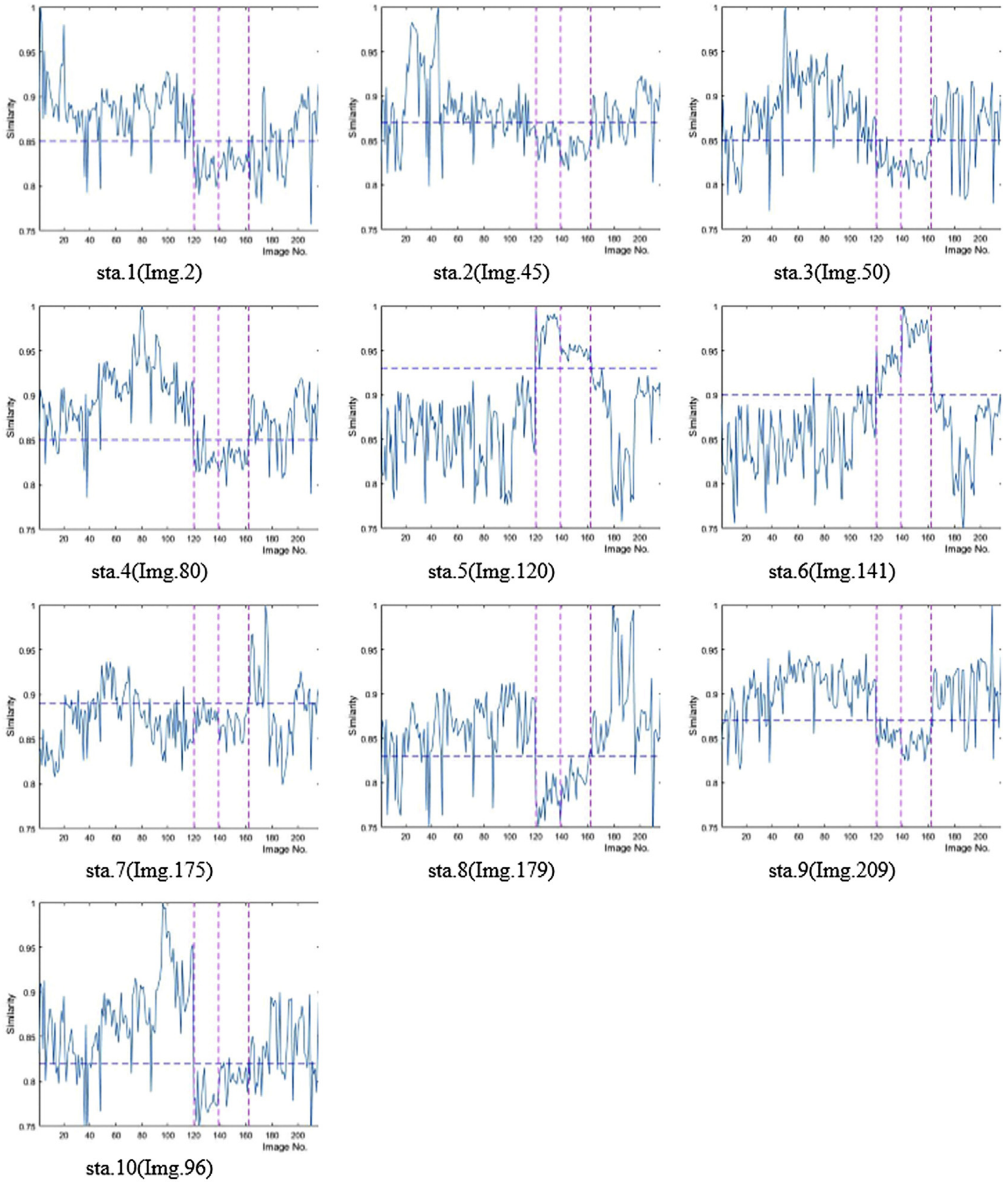


Fig. 8. Clustering results of BS180 with 2–10 clusters respectively.

different independent clusters 3 times. Whereas, the images of the other 8 statues are difficult to be completely clustered, but are always integrating and blending with each other. These results indicate that Sta.5&6 are quite different from the other 8 statues, while the other 8 statues are much similar to each other.

5.1.2. Similarity comparison

We randomly chose one image of each statue to compare them with the other images respectively: Img.2&45 stand for Sta.1&2, Img.50&80 stand for Sta.3&4, Img.120&141 stand for Sta.5&6, Img.175&179 stand for Sta.7&8, Img.209&96 stand for Sta.9&10.



sta.1(Img.2)

sta.2(Img.45)

sta.3(Img.50)

sta.4(Img.80)

sta.5(Img.120)

sta.6(Img.141)

sta.7(Img.175)

sta.8(Img.179)

sta.9(Img.209)

sta.10(Img.96)

Fig. 9. Similarity distribution curve of each statue in BS180.

Table 5
Data 4: ten statues of BS180.

Statue number & Image label	Total images	Place
Sta. 1(1–20)	20	Right wall from outside to inside, 5 statues
Sta. 2 (21–46)	26	
Sta. 3 (47–69)	23	
Sta. 4 (70–94)	25	
Sta. 5 (120–139)	20	
Sta. 6 (140–162)	23	Left wall from inside to outside, 5 statues
Sta. 7 (63–177)	15	
Sta. 8 (178–195)	18	
Sta. 9 (196–217)	22	
Sta. 10 (95–119)	25	

The similarity distribution curve of every image is shown in Fig. 9.

Regarding of the similarity level between Img.120&141 and the other images, the similarity value to the rest images of Sta.5&6 are all above 0.9; while the measurements to the other 8 statues' images are below 0.89 utmost, e.g., Sta.1 < 0.85, Sta.2 < 0.87, Sta.3&4 < 0.85, Sta.7 < 0.89, Sta.8 < 0.83, Sta.9 < 0.87, and Sta.10 < 0.82. Meanwhile, the similarity distribution curves of the other 8 images are almost uniform except among the image range of Sta.5&6. These results also indicate that Sta.5&6 are very similar, but different from the other 8 statues.

5.2. Art archeology and image analysis

The above experimental results can be verified by art archeology and image analysis. Firstly, from the images shown in Fig. 10, we

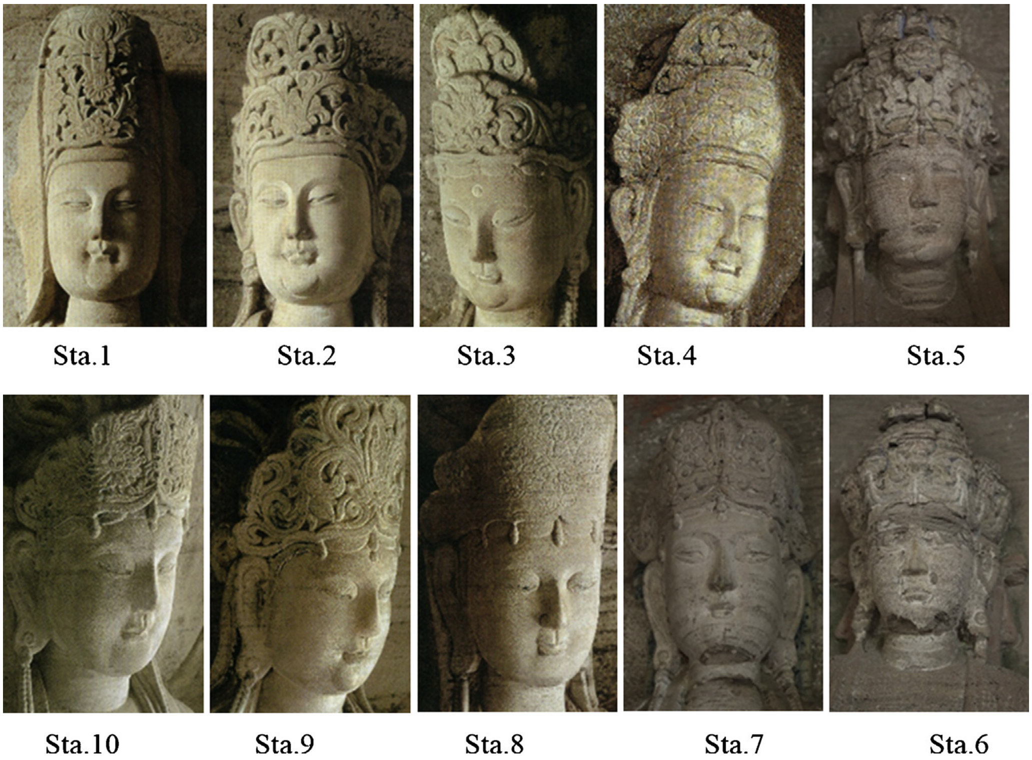


Fig. 10. Ten Avalokitesvara statues of BS180.

can see the difference between Sta.5&6 and the other 8 statues. For the other 8 statues, the lines of the face and crown are very smooth and soft, making the statues mellow and full, while the lines of Sta.5&6 look rigid and hard, making them stiff. Secondly, Sta.5&6 look like suffering from serious disease of weathering, but the other 8 statues are not. It is illogical and outrageous as Sta.5&6 are in the deepest of the cave, behind the main statue, shown in Fig. 7. Moreover, the overall design does not look harmonious. From Fig. 7, we can see that the cave looks like a half arc. Unlike other 8 statues, Sta.5&6 are not on the right or left side of the Saint Avalokitesvara respectively, but almost behind it, as somewhat needless statues. Simultaneously, besides of the two damaged statues located outmost the cave, there is a smaller Avalokitesvara statue on a holy lotus or an auspicious cloud overhead the eight statues respectively, whereas there is nothing overhead Sta.5&6. Furthermore, the feet of Sta.5&6 are not on the same horizontal line as the eight statues, but a little lower. Last but not least, their position on the wall is not in the same plane with the eight statues, but sunken in the wall. Each evidence indicates that Sta.5&6 were not designed to carve with the other statues together initially. Maybe, they were added later to keep the integrity of the cave since the outmost two statues were broken.

Avalokitesvara statues are the most common Buddhist statues, with the most types and the biggest change, as Buddhist scriptures say there are 33 incarnations for Avalokitesvara. They present different morphology of Avalokitesvara, with different gestures, scenes and ritual apparatus. The common form for the incarnations of Avalokitesvara are six Avalokitesvaras, fifteen Avalokitesvaras, thirty-two or thirty-three Avalokitesvaras etc. Thus, thirteen Avalokitesvaras are not a fixed form. And with the reference to the form of Shimenshan N^{o} 6, in which 5 Avalokitesvara statues are on both left and right sides, we can make a bold conjecture that the number of Avalokitesvara statues of Beishan N^{o} 180 might also be 10.

According to the computer experiments, art archaeology and image analysis, we therefore can confirm that there were only 5 statues on both sides in Beishan No. 180 initially. Thus its name should not be “Cave of Thirteen Incarnations of Avalokitesvara”, but be “Cave of Eleven Incarnations of Avalokitesvara”.

6. Discussion

In order to establish a recognizable and permanent digital image data, preparing for our future research of virtual inpainting, it is worthy to combine VGGNet and LDA, or adopt other excellent deep convolution networks, e.g., GoogleNet and ResNet to enhance the recognition accuracy. Furthermore, Bodhisattva head image comprises face and crown, the wide-length ratio of the image is 5:8. The input image in VGGNet is resized to be 224 × 224, which might distort our images. A subtle difference is critical for art image recognition, so the size of the input image is also worth considering in our future work. Moreover, Taoism and Confucianism statue images have to be added to our image data for the integrity of our Dazu carvings image dataset.

7. Conclusion

In this paper, we propose a VGGNet-16-based algorithm to recognize the modeling style of Dazu Bodhisattva head images. Abundant experiments combined with art archaeology and image analysis show its efficiency and superiority over five classical feature extraction algorithms. We therefore obtain several valuable conclusions: the modeling style is similar for the statues in the same cave or region, and it is also similar for the statues on the same subject, even though they are in different caves or regions; we rename BS180 as “Cave of Eleven Incarnations of Avalokitesvara” instead of the previous “Cave of Thirteen Incarnations of Avalokitesvara”, for two statues behind the major statue are quite different from the others. They might be carved to substitute the damaged statues outmost the cave for the integrity of the cave. The conclusions also prove our algorithm performs excellent in solving some art problems.

Furthermore, as the excellent performance in Bodhisattva head image modeling style recognition with our method, it is easy to find the most similar image to a specific image, which ensures future exemplar-based inpainting for rock carvings and cultural heritage preservation.

Acknowledgements

This work was supported by Chongqing Municipal Education Commission Humanities and Social Science Fund [grant number 14SKJ01].

References

[1] http://www.whc.unesco.org/en/list/912.

[2] R.P. Lippmann, Pattern classification using neural networks, IEEE Commun. Mag. 27 (11) (1989) 47–64.

[3] G. Kumar, P.K. Bhatia, A detailed review of feature extraction in image processing systems, in: International Conference on Advanced Computing & Communication Technologies, 2014, pp. 5–12.

[4] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, CoRR, 2014, abs/1409.4842.

[6] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, IEEE Comput. Vision Pattern Recognit. (2015) 648–656.

[7] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, in: Proc. Conference on ComputerVision and Pattern Recognition, 2014, pp. 1701–1708.

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition. Localization and detection using convolutional networks, EprintArxiv, 2013.

[9] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014) 580–587.

[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, CoRR, abs/1490.1556,2014.

[11] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154.

[12] K. Fukushima, S. Miyake, Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybern. 36 (1980) 193–202.

[13] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, 1998, pp. 1–46.

[14] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2) (2012) 1090–1098.

[15] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, Springer International Publishing, 2014, pp. 818–833.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern recognition. IEEE Computer Society, 2016, pp. 770–778.

[17] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, Recent advances in convolutional neural networks, arXiv preprint arXiv:1512.07108 (2015).

[18] A.G. Howard, Some improvements on deep convolutional neural network based image classification, *In Proc. ICLR* (2014).

[19] W. Zhang, S. Shan, L. Qing, Are Gabor phases really useless for face recognition? Pattern Anal. Appl. 12 (3) (2009) 301–307.

[20] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns computer vision–ECCV (2004), Springer, Berlin Heidelberg, 2004, pp. 469–481.

[21] L. Chen, H. Liao, M. Ko, J. Lin, G. Yu, A new LDA-based face recognition system which can solve the small sample size problem, Pattern Recognit. 33 (10) (2000) 1713–1726.

[22] W. Zhang, S. Shan, W. Gao, X. Chen, Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition[C]//, in: Tenth IEEE International Conference on Computer Vision, 2005, pp. 786–791.

[23] M.A. Imran, M.S.U. Miah, H. Rahman, A. Bhowmik, D. Karmaker, Face recognition using eigenfaces, Proc. IEEE Conf. Comput. Vision. Pattern Recognit. 84 (9) (2011) 586–591.