



Matching a composite sketch to a photographed face using fused HOG and deep feature models

Jiayi Xu¹ · Xinying Xue¹ · Yitiao Wu¹ · Xiaoyang Mao² 

Accepted: 5 September 2020 / Published online: 30 September 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

In this paper, we focus on the research of matching a computer-generated composite face sketch to a photograph. This is of great importance in the field of criminal investigation. To blend the different facial representation modalities, we propose a robust feature model by combining pixel-level features extracted from multi-scale key face patches and high-level features learned from a pre-trained deep learning-based model. At first, texture features are captured by a two-level histogram of oriented gradient descriptor, considering both the overall structure and local details. The semantic-level facial characteristics are analyzed through the high-level features of the Visual Geometry Group-Face (VGG-Face) network. Next, feature similarities between each sketch/photograph pair are measured by feature distance. Then, adaptive weights are assigned to each feature similarity, and score level fused according to their visual saliency contribution. Finally, the fused feature similarity is evaluated for matching purposes. After experimenting on the Pattern Recognition and Image Processing-Viewed Software-Generated Composite (PRIP-VSGC) database and the expanded University of Malta Composite Face Sketch (UoM-SGFS) database, it is found that this framework could achieve more satisfying results compared to the existing methods.

Keywords Composite sketch · HOG feature · VGG-face feature · Adaptive feature weight

1 Introduction

With the development of facial recognition technology, law enforcement departments are able to identify the information regarding suspects with greater effectiveness. Sometimes, a clear picture of a suspect's face cannot be obtained. In such cases, a facial sketch based on an eyewitness's description of the suspect's characteristics becomes the primary way to determine the suspect's facial information. It would be different to match a photograph of a person's face to a sketch of the face than to match a photograph to another photograph. This is due to the fact that a sketch has a higher abstraction

degree and fewer details than a photograph. Besides, since a sketch is influenced by the eyewitness's memory and synthesis, only some of the facial features can be restored in the resulting sketch. In such cases, the direct application of photograph-to-photograph facial recognition algorithms cannot achieve high recognition rates. In the sketch-photograph recognition area, feature representation that could precisely capture invariant features under different modalities is the key to success. In order to directly match facial sketches to facial photograph, this paper proposes a robust cross-modal feature model containing multiple levels of conventional textures and features based on deep learning. By carefully fusing the features, our method resulted in a more accurate retrieval of faces than in the majority of cases using current methods, and there may even be a great difference between the sketch and the photograph.

The remainder of the paper is organized as follows: After reviewing the related works in Sects. 2, 3 describes the overall structure of the proposed recognition method. Section 4 explains the hybrid-feature model and recognition process in detail. Section 5 provides several experiments and compares our method with those of others. Section 6 concludes the paper and discusses possible future work in this field.

✉ Xiaoyang Mao
mao@yamanashi.ac.jp

Jiayi Xu
xujiayi@hdu.edu.cn

Xinying Xue
xxy1202@hdu.edu.cn

Yitiao Wu
uuii210@163.com

¹ Hangzhou Dianzi University, Hangzhou, China

² University of Yamanashi, Kofu, Japan

2 Related works

The research in matching facial photograph from sketches falls under the research field of cross-modal facial recognition. Face sketches are often drawn by portrait experts and are widely used by law enforcement departments to determine the identity of suspects. Assuming an eyewitness could express the facial characteristics precisely, the sketch drawn by an expert would be highly similar to the original face. Thus, matching a photograph with a hand-drawn sketch usually achieves a high recognition rate. However, it may take the expert a lot of time to draw a sketch that accurately represents the face of a suspect, not to mention that it takes more time to train a professional portrait artist. Therefore, some researchers proposed generating face sketches using computer software based on the description provided by eyewitnesses or from their own observations. This way of composing a sketch is quite efficient and easy to use, and it is gradually favored by law enforcement departments. However, as the quality of a sketch synthesized by software is limited by the number of samples and the fact that the precise relative position or arrangement of facial components plays an important role, the computer-generated sketches may be less authentic than hand-drawn sketches generated by portrait experts. The identification rate still needs to be improved. According to the generation method, we divide the face sketches into two categories: hand-drawn sketches and software composite sketches.

Earlier, researchers proposed using facial composition based on a recognition method to create facial images collected based upon approaches to the same process by transforming the sketch to a photograph, and vice versa, and then carrying out facial matching procedures. The disadvantage of this method is the low efficiency due to the extra step of transformation. Many well-known research teams have tried this approach. Tang et al. first proposed the eigenface algorithm [1], the locally linear embedding (LLE) algorithm [2], and the Markov random field (MRF)-based algorithm [3]. Gao et al. proposed a face synthesis algorithm based on sparse representation [4] and based on random sampling [5]. Zhang et al. [6] proposed face synthesis algorithm based on the end-to-end convolutional neural network (CNN) model. Kazami et al. [7] proposed an unsupervised face geometry learning model based on cycle-consistent adversarial networks (CycleGAN). Isola et al. [8] proposed a learning model based on conditional adversarial network (CGAN). Pallavi et al. [9] proposed a synthesized sketch algorithm based on an image enhancement algorithm.

Later, researchers realized that facial matching could be performed without modal changes. Different representation modes could be mapped to the same feature space and modal invariance features would be extracted. In early studies, researchers employed low-level feature extraction models.

Klare et al. [10] measured the similarity between a sketch and photograph directly by the scalar invariant feature transform (SIFT) feature distance. Non-overlapping multi-scale SIFT representation of size 32 and 16 was used, and the features were fused in the score level. They further proposed a framework [11] in which SIFT and a multi-scale local binary pattern (LBP) descriptor-based feature model were incorporated, and linear discriminant analysis (LDA) projection was used for minimum distance matching. Although these methods used a mixture of local feature descriptors, the importance of different facial components was not evaluated. Later, they proposed the P-RS method [12], short for the heterogeneous face recognition using kernel prototype similarities, which projected the feature from different modalities into a linear discriminant subspace to measure the similarities using a kernel-based method. Mittal et al. proposed a self-similar descriptor (SSD) [13] algorithm based on dictionary learning. Han et al. [14] proposed a recognition algorithm based on local features; multi-scale LBP features were extracted from six facial components separately. Based on the experiments, the four most dominant features, from the eyebrow, nose, hair, and mouth, were fused at the score level for facial matching. However, most of the methods mentioned above only consider the texture features, and they ignore the structural features. To deal with this problem, Liu et al. [15] proposed a component-based representation approach (CBR) that incorporated the SIFT feature and the HOG feature to represent local features from six facial components, and different feature weights were assigned empirically. However, the adjustment to the weight in an adaptive way was not considered. Mittal et al. [16] addressed a hybrid-feature description model fusing DAISY and HOG features. Different from the former methods, global and local texture features were extracted based on a saliency map, and several semantic attributes were considered. Manually labeled attributes included ethnicity, skin color, gender, and age. The attributes and texture features were fused at the score level to get the ranked list. Xu et al. [38] used a multi-scale HOG descriptor to extract the features and fused them at the score level. Three manually labeled semantic attributes, including gender, glasses, and mustache, were used to filter the matching result.

With the success of the deep learning method, more and more researchers considered introducing deep learning-based models. To deal with the problem of limited training data, several researches tried transfer learning-based methods. After converting color images into grayscale images, Mittal et al. [17] proposed a transfer learning-based algorithm with autoencoder and deep belief network representation. Galea et al. [18] used the fine-tuned VGG-Face [19] model for composite facial sketch matching. Chugh et al. [20] addressed a new feature descriptor, the histogram of image moments, and combined it with the HOG descriptor to obtain the

features in local regions. They used a genetic approach to match hand-drawn sketches to digital photographs. After using inductive transfer learning to re-weight the top 20 weight vectors, the method could be extended to perform a composite sketch to digital photograph matching. Hadi Kazemi et al. [21] proposed a deep coupled convolutional neural network (DCCNN) using 180 manually labeled biological attributes. The identification result was quite good, but obviously, labeling such a large number of attributes was rather difficult. The coupled network composed of a photograph-DCNN and a Sketch-attribute-DCNN learned to map the two face modalities onto a shared subspace. Wan et al. [22] proposed a three-channel CNN architecture based on the VGG-Face network, whose inputs are an anchor (sketch), a positive (photograph with same identity), and a negative (photograph with different identity). Triplet loss function minimizing inter-subject distance and maximizing intra-subject distance was adopted to fine-tune the shared weights. Peng et al. [23] addressed a sparse graphical representation-based discriminant analysis approach. A Markov network model was constructed and refined to generate adaptive sparse vectors for face matching. Peng et al. [24] also proposed a deep local descriptor learning framework called DLFace, in which an effective triplet loss was designed and aimed at decreasing the distance between the same subject and increasing the distance between different subjects.

Above all, there are still a lot of problems to be solved in the research of facial recognition using face sketches. For example:

1. Traditional methods often try to extract cross-modal features from texture details using descriptors, such as HOG, SIFT, and LBP, and combine them. However, low-level features are not adequate due to sketch details not being sufficient in highly abstract parts of a sketch, and finding a fixed feature weight suitable for all facial data is not feasible.
2. Deep learning-based methods often gain higher recognition accuracy than conventional methods, but the combination with other feature models is usually not considered. Besides, the preparation work of collecting thousands of sketch/photograph pairs and training the network is rather time-consuming.

In this paper, we propose a composite sketch–photograph matching algorithm with both traditional low-level texture features and deep learning-based high-level features. To catch the features from abstracted sketches and detailed photographs, a hybrid-feature model with global- and local-scale HOG texture features and CNN-based features was constructed. Instead of training a deep network from scratch, we use the pre-trained VGG-Face network to extract deep features. Feature similarity is measured by feature distance.

Between a sketch and a photograph, the similarities between each feature can be computed. After analyzing the importance of single features and carefully fitting a weight function, the reliability of the each feature similarity was determined. Finally, adaptively fusing the effect of all features, matching between face sketch and photograph is realized. Compared with the methods based solely on either low-level features or very deep neural networks, the proposed method ensures high identification accuracy while keeping low computation time.

The main contributions of our method are summarized as follows:

1. A multi-scale HOG feature model is proposed, where both local details and global-structure features are considered. The incorporation of the pre-trained VGG-Face feature, an effective and efficient hybrid-feature model is built.
2. The importance of different features is discussed, and the feature weights are adaptively learned from the sample data.
3. Instead of score level fusing all the features and then matching, our method obtains individual matching results using each feature and then fuses them based on dynamically adjusted weights.

3 Proposed system

This paper proposes a hybrid-feature model composed of pixel-level texture features and high-level facial features extracted from a deep network, which could be used to efficiently match a sketch with a photograph or vice versa. Based on the HOG descriptor and VGG-Face descriptor, which is based on the VGG network [25], the influence of each feature is dynamically adjusted according its visual contribution. The sketch/photograph similarity is measured by score-level fusing all the feature similarities. Based on the feature similarities, a list of matching results is obtained. More specifically, as shown in Fig. 1, the overall framework of the proposed method includes the following steps:

– Step 1. Facial image preprocessing

Perform processes such as background removal, face alignment, normalization, and image cropping. In the cropping step, hair is moved above the forehead. The profile is retained for subsequent hairstyle extraction.

– Step 2. Hybrid feature model

1. *Segmentation of facial components* Locate the facial landmarks from which to segment image patches of facial components including hair, eyebrows, eyes, nose, mouth, and chin contour.

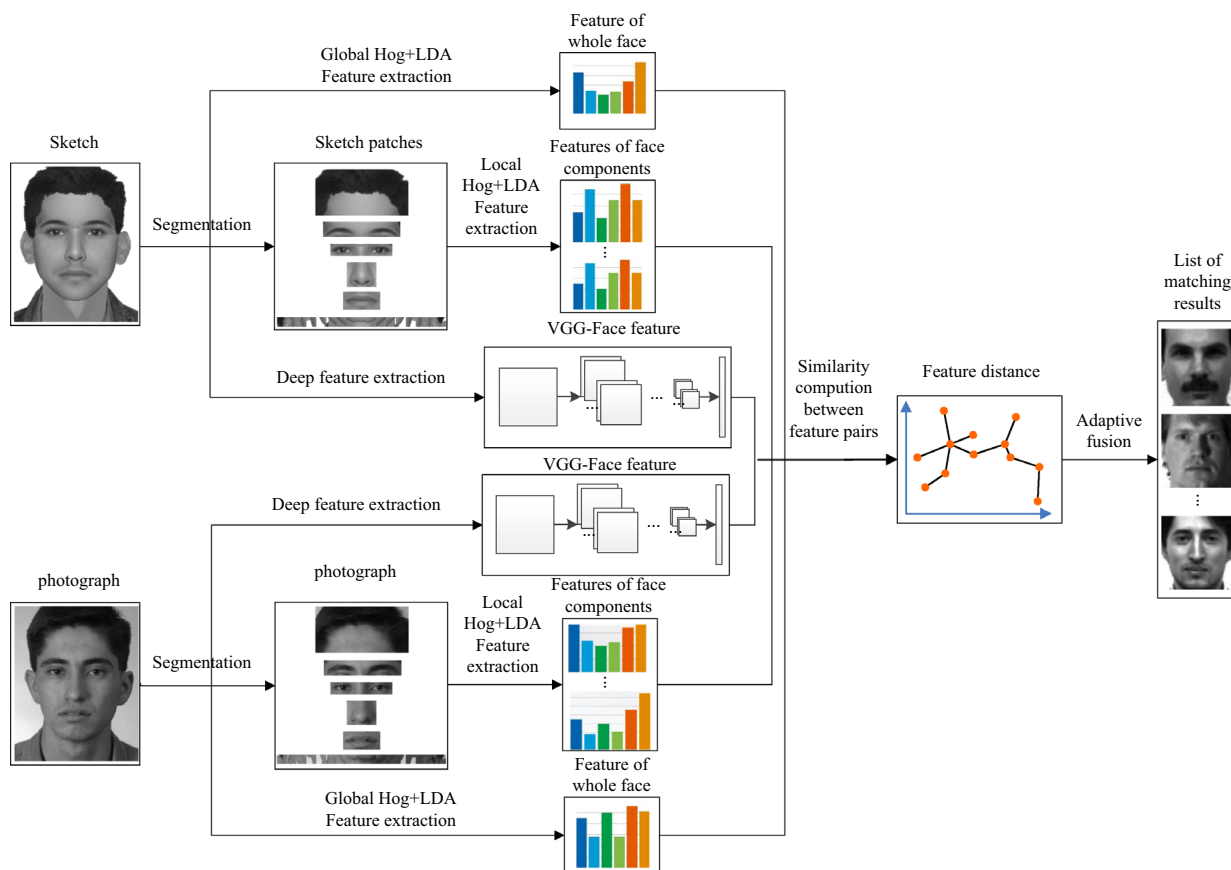


Fig. 1 Sketch–photograph recognition framework

2. *Multi-scale HOG feature extraction* Perform global HOG feature extraction of the whole face, and local HOG feature extraction of the key components from sketch and photograph images. For each facial image, compute a feature vector including one global feature mainly describing the facial contour and hairstyle, and six local features extracted from segmented image patches.
 3. *VGG-Face feature extraction* Feed the entire face into the VGG-Face network, and extract the high-level features of fc8 layer.
- *Step 3. Adaptive score-level fusion of feature similarities and sketch–photograph matching*
1. *Compute similarity matrices for HOG and VGG-Face features* Between each sketch and photograph, calculate its feature distances for global HOG features, local HOG features, and VGG-Face features, respectively. In this way, a total of eight similar matrices between every sketch and photograph image in the database are constructed.
 2. *Normalize the similarity matrices* Normalize similarity matrices by referencing the average face that

has been computed by averaging all the faces in the database.

3. *Adaptively assign weights to feature similarities* For a given sketch, according to its deviation from the average face, adaptively set the weights of each similarity matrix, and vice versa. Score level fuses similar matrices into one overall matrix that stores the weighted feature distance.
4. *Find the match list* For the given sketch, reference the fused feature similarity matrix to find the closest photographs, and arrange them according to feature distance to form a matching list.

4 Multi-level feature extraction

4.1 Facial image preprocessing

Due to different acquisition resources, collected facial images may have different sizes and backgrounds that will affect the subsequent feature extraction and recognition. Therefore, preprocessing procedures, including face alignment, background removal, and normalization, should be per-



Fig. 2 Facial photographs before and after preprocessing



Fig. 3 Facial sketches before and after preprocessing

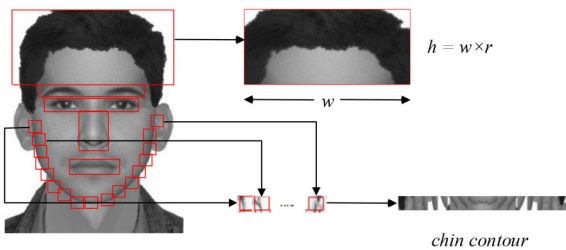


Fig. 4 Segmentation of key facial regions

formed. Figures 2 and 3 show an example of grayscale photographs and sketches before and after preprocessing. More specifically, the preprocessing is divided into three steps:

1. *Face alignment* Affine transformations of sketches and photographs are carried out so that the eyes are on the same horizontal line. For preparation of HOG feature extraction, color photographs are transformed into grayscale images.
2. *Background removal* The aligned image is clipped and a fixed ratio of facial length to width reserves only the facial area while eliminating the background. In our clipped result, the hair rests above the forehead so that the sides of the face are retained.
3. *Scale normalization* Scales the clipped facial sketch and photograph into predefined sizes. Different image sizes are used for HOG and VGG-Face feature detection.

4.2 Hybrid feature model

4.2.1 Multi-scale HOG features

As facial images contain a lot of detailed features, we use the HOG descriptor [26] to catch texture features for facial

Type	Name	Kernel	Stride	Ch I/O	Input
conv	conv1_1	3	1	3/64	image
conv	conv1_2	3	1	64/64	conv1-1
mconv	pool1	2	2	64/64	conv1_2
conv	conv2_1	3	1	64/128	pool1
conv	conv2_2	3	1	128/128	conv2_1
mconv	pool2	2	2	128/128	conv2_2
conv	conv3_1	3	1	128/256	pool2
conv	conv3_2	3	1	256/256	conv3_1
conv	conv3_3	3	1	256/256	conv3_2
mconv	pool3	2	2	256/256	conv3_3
conv	conv4_1	3	1	256/512	pool3
conv	conv4_2	3	1	512/512	conv4_1
conv	conv4_3	3	1	512/512	conv4_2
mconv	pool4	2	2	512/512	conv4_3
conv	conv5_1	3	1	512/512	pool4
conv	conv5_2	3	1	512/512	conv5_1
conv	conv5_3	3	1	512/512	conv5_2
mconv	pool5	2	2	512/512	conv5_3
conv	fc6	7	1	512/4096	pool5
conv	fc7	1	1	4096/4096	fc6
conv	fc8	1	1	4096/2622	fc7
softmax	prob	—	1	—	fc8

Fig. 5 The architecture A of the VGG-Face network

sketches and photographs. Local HOG feature extraction is performed on the image patches for each key facial component, while global HOG feature extraction is performed on the entire face to capture the characteristics, such as facial structure and hairstyle. Image patches corresponding to key facial components are segmented for preparation by the local HOG feature extraction. First, 68 landmarks are detected on the facial image. Then, for components including hair, eyebrows, eyes, nose, and mouth, one rectangular image patch is created for each landmark from the preprocessed sketch and photograph. As illustrated in Fig. 4, the width of each rectangular image patch is calculated by the horizontal distance difference between the rightmost landmark and leftmost landmark, while the height is calculated by the product of the width and the ratio derived from the average face. In order to precisely describe the shape of chin contour, multiple patches are cut and merged into a long rectangle. In the current implementation, 17 patches are located according the position of 17 landmarks along the chin and are linked together to represent the chin contour.

The extraction process of the HOG feature can be divided into the following five steps:

1. Separate the rectangular image patch of the facial component into $s \times s$ sized blocks;
2. Further cut each patch into four equal cells;
3. Compute the gradient magnitude and direction of each pixel in the cell;
4. The range of gradient directions is divided into eight directions, and then, the gradient magnitude of the pixel

in each cell in the same direction is accumulated to obtain an eight-dimensional cell feature vector;

- The cell feature vector in each image block is calculated and linked to form the block feature vector. Next, the block feature vector in each block is calculated and concatenated to form the feature vector F_m for each facial component:

$$F_m = (F_1, F_2, \dots, F_n) \quad (1)$$

where n is the number of blocks for each image.

After extracting the global and local HOG feature vectors of the images, the LDA algorithm [27] is used to reduce the feature dimensions of each feature vector. Finally, for each sketch and photograph image, respectively, seven feature vectors are obtained (six for facial components and one for the whole face).

4.2.2 VGG-face based features

High-level semantic features that more closely correspond to visual recognition processes may well supplement the aforementioned texture-based facial features. In this work, we employ the VGG-Face model [19] and use the high-layer features extracted from the network to represent high-level facial features. As shown in Fig. 5, the architecture A of the VGG-Face network consists of 11 blocks. The first eight blocks are convolutional layers, while the last three blocks are fully connected layers. The output features of the fc6, fc7, and fc8 layers are tested to represent semantic facial features, and features of the fc8 layer are chosen to be incorporated in our framework.

4.3 Dynamic feature fusion and sketch–photograph matching

4.3.1 Feature similarity measurement

To measure the similarity between a sketch image and a photograph image, Euclidean distance between corresponding features is used. The smaller the Euclidean distance, the more similar the two images are. Before comparison, each feature vector F_m is normalized according to Eq. 2. Then, feature matrices for sketch and photograph images are constructed, which correspond to the hair, eyebrows, eyes, nose, mouth, chin contour, the whole face, and the VGG-Face network. For example, Eq. 3 and Eq. 4 show the computation of the sketch and photograph feature matrix, respectively.

$$F_{norm} = \frac{F_m - F_{min}}{F_{max} - F_{min}} \quad (2)$$

$$F_j^S = (F_{norm}^{S_1}, F_{norm}^{S_2}, \dots, F_{norm}^{S_{m1}})^T \quad (3)$$

$$F_j^P = (F_{norm}^{P_1}, F_{norm}^{P_2}, \dots, F_{norm}^{P_{m2}})^T \quad (4)$$

where F_{max} and F_{min} are the features with the largest and smallest length in the database, j refers to the feature, which may correspond to hair, eyebrows, eyes, nose, mouth, chin contour, the whole face, and the VGG-Face network, $m1$ represents the number of sketch images, and $m2$ represents the number of photo images in the database.

A total of eight feature similarity matrices between the sketch and photograph images are computed using Eq. 5. By indexing the matrix D_j , we could calculate the similarity between the corresponding features in any sketch/photograph pairs in the database.

$$D_j = \|F_j^S - F_j^P\| \quad (5)$$

4.3.2 Adaptive feature fusion and facial recognition

In the observation that facial components with large deviation to the average face may attract special attention, we evaluated the significance of different features based on their visual saliency. According to Eq. 6, different weights $w(i)$ are assigned adaptively to every feature, while a relatively large weight $w(0)$ is assigned to the VGG-Face feature empirically.

$$w(i) = \begin{cases} \frac{K}{\sqrt{2*i-1} \cdot Sum}, & i \subseteq [1, 7], \text{ for HOG features} \\ \frac{2 * K}{Sum}, & i = 0, \text{ for VGG - Face feature} \end{cases} \quad (6)$$

$$F_{dev} = \frac{F_{norm} - F_{average_norm}}{F_{average_norm} - F_{min_norm}}. \quad (7)$$

Here, i is the order of feature significance ranging from one to seven. The order of feature significance is evaluated with its deviation F_{dev} as illustrated in Eq. 7. In this equation, $F_{average_norm}$ refers to the average of all normalized features and F_{min_norm} refers to the smallest normalized feature. For example, the order of the most significant HOG feature with the largest F_{dev} is set to 1. Constant $K = 14$ is used in the current implementation. As a result, a score-level fused feature similarity matrix D is obtained from the above mentioned eight component-based similarity matrices:

$$D = (D_{all}, D_h, D_{eb}, D_e, D_n, D_m, D_c, D_{vgg}) \times W^T, \quad (8)$$

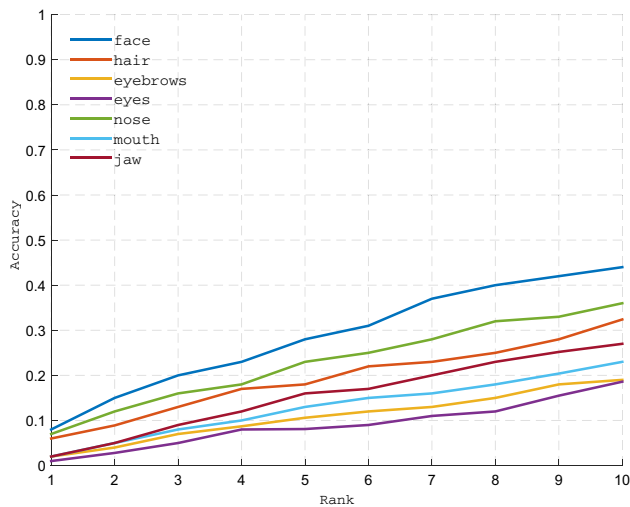
where W^T represents the vector composed of the above-mentioned feature weights and $D_{all}, D_h, D_{eb}, D_e, D_n, D_m, D_c, D_{vgg}$ are the similarity matrices for the whole face, hair, eyebrows, eyes, nose, mouth, chin, and VGG-Face based features, respectively.

Table 1 Parameter setting in HOG feature extraction

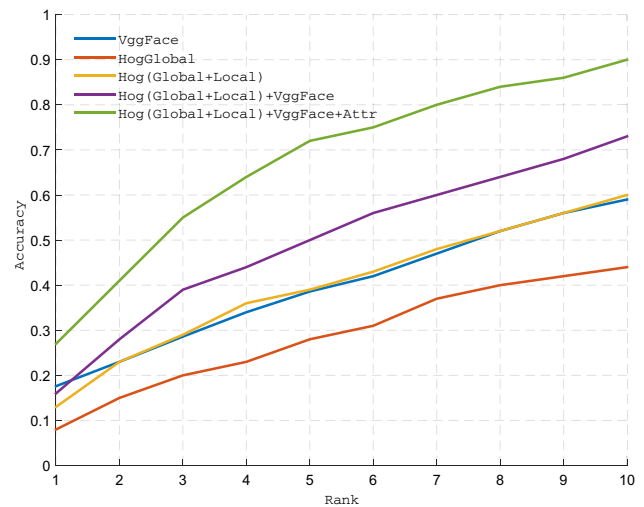
Component	Patch size of this component	Overlapping	Block size for HOG descriptor (using padding)
Whole face	200×250	yes	32×32
Hair	150×75	no	40×40
Eyebrows	125×25	no	20×20
Eyes	100×20	no	20×20
Nose	45×45	no	20×20
Mouth	60×30	no	20×20
Chin contour	272×16	no	16×16

Table 2 Comparison of rank 10 accuracy on different feature layers of VGG-Face network

Layer	Recognition rate (%) on the PRIP-VSGC database	Recognition rate (%) on the Uom-SGFS database(Set A)	Recognition rate (%) on the expanded Uom-SGFS database(Set A)
fc6	57.2	56.7	46.2
fc7	57.0	54.3	45.3
fc8	59.0	62.8	47.0



(a) Recognition rate of a single feature



(b) Recognition rate of different feature models

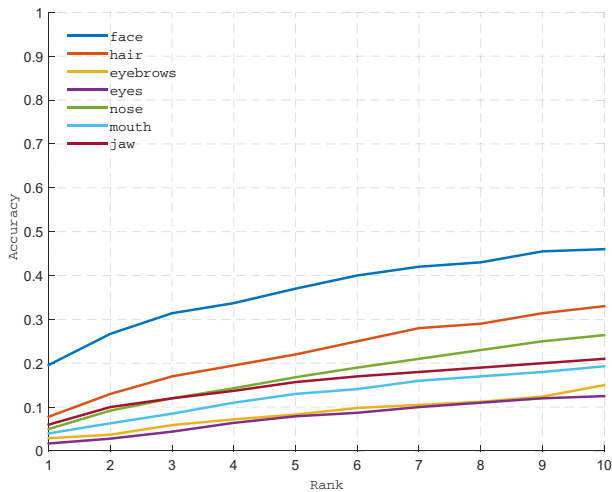
Fig. 6 Recognition performance on the PRIP-VSGC database

5 Experimental results

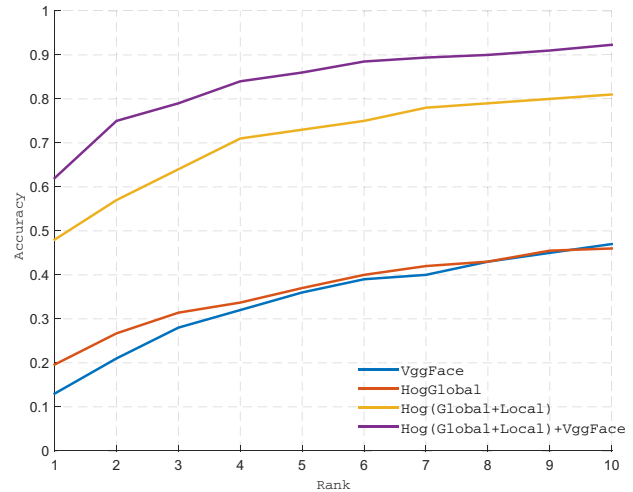
In this section, we evaluate the proposed algorithm on two commonly used public software-generated facial sketch databases: the PRIP-VSGC database [28], and the Uom-SGFS database [29]. We expanded the Uom-SGFS database from 600 samples to 2,180 (500 from the original Uom-SGFS database, 188 from the CUHK Face Sketch database (CUFS) [3], 121 from the AR database [30], 406 from the Facial Recognition Technology (FERET) database [31], 100 from the Memory Gap Database (MGDB) [32], 275 from the CelebFaces Attributes Dataset (CELEBA) [33], 140 from the

FEI Face Database [34], and 350 from the Multiple Encounter Dataset II (MEDS-II) [35]).

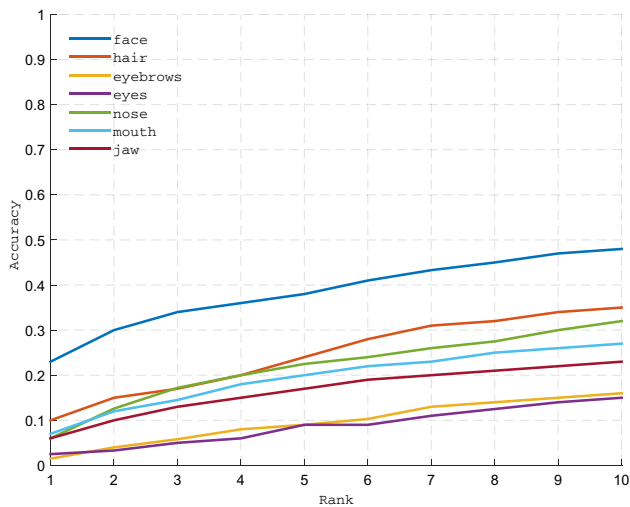
For our experiments, the dlib library [36] was employed to detect the 68 landmarks of the face. Based on the horizontal center of the eyes, the images were aligned and scaled. All the images were cropped to the size of 200 × 250 pixels. For each face, six component patches were generated, and HOG feature extraction was performed using the setting shown in Table 1. More specifically, based on the patch size of each facial component, different block sizes for HOG descriptors were used. If the block size could not be divided by patch size, padding was employed. For extracting the global HOG fea-



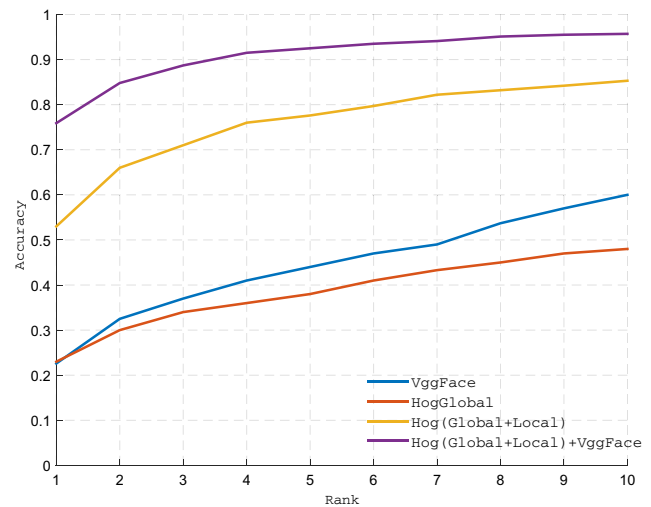
(a) Recognition rate of a single feature



(b) Recognition rate of different feature models

Fig. 7 Recognition performance on the expanded Uom-SGFS database (Set A)

(a) Recognition rate of a single feature



(b) Recognition rate of different feature models

Fig. 8 Recognition performance on the expanded Uom-SGFS database (Set B)

ture from the whole face, overlapped scheme was used. For extracting the local HOG features, non-overlapped scheme was used. Then, to gain deep features, the VGG-Face network pre-trained using the labeled faces in the wild (LFW) database [37], as shown in Fig. 5, was used. Resized facial images of 224×224 pixels were used for the network input value. The network optimization was realized by stochastic gradient descent using mini-batches of 64 samples and a momentum coefficient of 0.9. The learning rate was initially set to 10^{-2} and then decreased by a factor of 10 when the validation set accuracy stopped increasing. The coefficient of weight decay was set to 5×10^{-4} . The dropout was applied after the two fully connected layers at a rate of 0.5. To evaluate the effectiveness of different feature layers, fc6,

fc7, and fc8 were tested. As can be seen from Table 2, using the features from fc8 layer was most effective, so we decided to employ the 2, 622-dimensional output of fc8 as the VGG-Face feature in our framework. In the final feature fusion stage, adaptive weights that corresponded to the HOG and VGG-Face features were computed according to the importance of the components using Eq. 6.

5.1 Experiments on the PRIP-VSGC database

In this section, the PRIP-VSGC database was used to verify the proposed algorithm. The PRIP-VSGC database contains 123 photographs from the AR database [30], and 123 corresponding sketches were generated using the Identi-Kis

Table 3 Comparison of rank 10 accuracy on the PRIP-VSGC database

Algorithm	Recognition rate (%)
Eigenfaces [40]	21.9±4.4
SSD [13]	45.3
P-RS [12]	53.7±6.1
G-HFR [41]	54.9±3.1
CBR [15]	61.6±2.7
DCCNN without attributes [21]	67.4±1.9
DCCNN with attributes [21]	72.6±0.9
SGR-DA [23]	70.0±1.98
DLFace [24]	76.4
Our previous model [38]	88.6±1.9
Ours	72.9±2.49
Ours with semantic attributes	90.4±1.11

software [39]. Because the LDA algorithm that we used for feature reduction relies on prior knowledge of classification, we separated the images into a testing set for identifying the classification information and a training set for feature reduction. Similar to the parameters used in the work of Liu et al. [15], for each experiment, we randomly selected 48 pairs of photographs and sketches as the training set, leaving 75 pairs as the testing set.

Five groups of experiments were carried out, the average recognition accuracy was calculated, and the data were plotted as a cumulative match characteristic (CMC) curve as depicted in Fig. 6. First, based on the HOG descriptor, the global feature (whole face) and each local component feature were used for recognition verification, and the identification accuracy is shown in Fig. 6a. Among the features, rank 10 identification accuracy achieved the best value of 44.4% using global feature, which revealed the whole face was the most important single feature.

The matching results using single feature were significantly low; therefore, a proper mixture of single features was crucial. Experiments were then designed using fused local features, fused global and local features, and adding attribute constrains based on fused global and local features as shown in Fig. 6b. As can be seen from the figure, rank 10 identification accuracy using fused local features achieved an accuracy rate of 52%, rank 10 identification accuracy using fused global and local features achieved 59.7%, and rank 10 identification accuracy using VGG-Face feature achieved an accuracy rate of 59.0%. By adaptively combing HOG and VGG-Face features, the rank 10 identification accuracy achieved an accuracy rate of 72.9%. By further introducing three manually semantic attributes (including gender, glasses, mustache) as illustrated in our previous method [38], the rank 10 identification accuracy reached the highest value at 90.4%. However, these attributes were not automatically

Table 4 Comparison of rank 10 accuracy on the expanded Uom-SGFS database

Algorithm	Sample number	Recognition rate (%)
Fisherface [40]	1,522	5.3±1.2
VGG-Face [19]	1,522	33.3±2.9
LGMS [42]	1,522	60.4±3.1
CBR [15]	10,000	68.3±3.8
DLFace(Set A) [24]	1,521	92.1
DLFace(Set B) [24]	1,521	94.8
Ours(Set A)	2,180	92.3±1.15
Ours(Set B)	2,180	95.8±1.28

labeled, and this method was unable to be extended to a large database.

We compared our algorithm with several state-of-the-art algorithms, such as the classical eigenfaces model and the recently proposed deep learning models. As shown in Table 3, the proposed algorithm with three semantic attributes achieved a rank 10 identification accuracy of 90.4%, which performs much better than many CNN based models, including SGR-DA [23], DLFace [24], and DCCNN [21], in which 180 attributes were adjusted manually.

5.2 Experiments on the expanded Uom-SGFS database

In this section, the enlarged Uom-SGFS database was used to verify the proposed algorithm. Each face photograph corresponds to two face sketches (Set A and Set B). Set B was modified based on the software generated in Set A, which was more similar to the facial photograph. For feature reduction procedures, the LDA algorithm was trained with 450 sketch/photograph pairs from the original Uom-SGFS database and was tested on the remaining 1,730 sketch/photograph pairs.

Five groups of experiments were carried out, the average recognition accuracy was calculated, and the data were plotted as a CMC curve as depicted in Fig. 7 for Set A and Fig. 8 for Set B. First, based on the HOG descriptor, global feature, and each local component feature used for recognition verification, the identification accuracy is shown in Fig. 7a for Set A and Fig. 8a for Set B. Among the features, rank 10 identification accuracy achieved the best value of 46.0% for Set A and 48.0% for Set B using global features. Similar to the experimental results on the PRIP-VSGC database, the matching results using single local features were significantly low.

It was insufficient to use single feature for identification, as a hybrid-feature model was more suitable. Later, experiments were designed by using fused local component features, fused global and local features, VGG-Face features, and

Table 5 Comparison of rank 10 accuracy on different HOG feature weights

Whole:hair:eyebrows: eyes:nose:mouth:chin	9:5:2:1:5:5:5 our previous model [38]	2:1:1:1:1:1	Adaptive weights
Database			
Prip-VSGC database	60.27±2.19%	59.16±2.97%	59.72±0.99%
expanded Uom-SGFS database(Set A)	79.73±4.21%	80.53±4.11%	80.53±0.96%
expanded Uom-SGFS database(Set B)	84.53±1.92%	84.26±3.52%	85.28±0.81%

Table 6 Comparison of rank 10 accuracy on different HOG and VGG-Face feature weights

Whole:hair:eyebrows: eyes:nose:mouth:chin: VGG-Face	9:5:2:1:5:5:5:18 Our previous model [38]	2:1:1:1:1:1:4	Adaptive weights	Adaptive weights with three attributes
Database				
Prip-VSGC database	71.77±1.76%	68.77±1.98%	72.86±2.49%	90.40±1.11%
expanded Uom-SGFS database(Set A)	90.0±2.93%	89.73±3.14%	92.27±1.15%	×
expanded Uom-SGFS database(Set B)	94.67±1.34%	94.93±1.39%	95.74±1.28%	×

combing all the above mentioned features as shown in Fig. 7b for Set A and Fig. 8b for Set B. As can be seen from these figures, for expanded Set A, rank 10 identification accuracy using fused local features achieved an accuracy rate of 80.5%, rank 10 identification accuracy using VGG-Face features only achieved an accuracy rate of 47.0%, rank 10 identification accuracy using fused HOG and VGG-Face features achieved an accuracy rate of 92.3%. For the expanded Set B, rank 10 identification accuracy using fused local features achieved an accuracy rate of 85.3%, rank 10 identification accuracy using VGG-Face features only achieved an accuracy rate of 60.0% and rank 10 identification accuracy using fused HOG and VGG-Face features achieved an accuracy rate of 95.8%.

We compared our model with the classical Fisherface model, the original VGG-Face model, and several latest deep learning models. As shown in Table 4, our algorithm could achieve rank 10 recognition of 92.3% (Set A) and 96.8% (Set B) without the burden of training the network, which outperforms the deep learning methods of Liu et al. [15] and Peng et al. [24].

5.3 Experiments on different weight schemes

To evaluate whether our adaptive weight scheme outperformed the fixed weights set based on empirical values, two experiments were conducted to compare the effect of different weights (rank 10 identification accuracy in current implementation) and showed the applicability of our proposed weight function (Eq. 6). In the first experiment, we

compared the identification accuracy using mixed HOG features on two databases. As can be seen from the data (Table 5), the performance of adaptive HOG weights was similar to that of the fixed weights with a small increase in stability. In the second experiment, a hybrid feature model with both HOG and VGG-Face features was used. As shown in Table 6, a larger weight was assigned to the VGG-Face-based feature in our model and achieved better accuracy and stability than our previous method.

6 Conclusion

In this paper, we proposed a multi-scale HOG descriptor and VGG-Face combined framework for matching software-generated facial sketches to photographs. Based on the traditional feature extractor theory, both local features of key facial components and global features of face structure, including hairstyle, were extracted. Furthermore, we employed the pre-trained neural network which runs in high efficiency to collect facial features from a high semantic level. After the score level fusing all these features, recognition was carried out based on the weighted feature similarity distance to form the matching list. Experiments based on the PRIP-VSGC database and the expanded Uom-SGFS database reflected that the proposed algorithm achieved a higher recognition accuracy than most of the existing algorithms.

In the future, we plan to construct a new deep network to add more effective features in further studies. We will

also consider introducing more visually salient semantic attributes to classify them automatically.

Acknowledgements This paper was sponsored by the Public Welfare Research Project of Zhejiang Province, China (Grant No. LGF18F020015), JSPS Grants-in-Aid for Scientific Research, Japan (Grant No. 17H00737), and Opening Foundation of Key Laboratory of Fundamental Science for National Defense on Vision Synthesis, Sichuan University, China (Grant No. 2020SCUVS007).

References

- Tang, X., Wang, X.: Face photo recognition using sketch. *Proceedings of International Conference on Image Processing*, pp. 257–260 (2002)
- Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. *Int. Conf. Comput. Vis. Pattern Recogn.* **1**, 1005–1010 (2005)
- Tang, X., Wang, X.: Face photo-sketch synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(11), 1955–1967 (2009)
- Gao, X., Wang, N., Tao, D., Li, X.: Face sketch-photo synthesis and retrieval using sparse representation. *IEEE Trans. Circuits Syst. Video Technol.* **22**(8), 1213–1226 (2012)
- Wang, N., Gao, X., Li, J.: Random sampling for fast face sketch synthesis. *Pattern Recogn.* **76**, 215–227 (2018)
- Zhang, L., Lin, L., Wu, X., Ding, S., Zhang, L.: End-to-end photo-sketch generation via fully convolutional representation learning. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 627–634 (2015)
- Kazemi, H., Taherkhani, F., Nasrabadi, N.M.: Unsupervised facial geometry learning for sketch to photo synthesis. *International Conference of the Biometrics Special Interest Group*, pp. 1–5 (2018)
- Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134 (2017)
- Pallavi, S., Sannidhan, M.S., Sudeepa, K.B., Bhandary, A.: A novel approach for generating composite sketches from mugshot photographs. *IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 460–465 (2018)
- Klare, B., Jain, A.K.: Sketch-to-photo matching: a feature-based approach. In *Proceedings of SPIE - The International Society for Optical Engineering*, pp. 7667–7702 (2010)
- Klare, B., Li, Z., Jain, A.K.: Matching forensic sketches to mugshot photos. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(3), 639–646 (2011)
- Klare, B., Jain, A.K.: Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(6), 1410–1422 (2013)
- Mittal, P., Jain, A., Goswami, G., Singh, R., Vatsa, M.: Recognizing composite sketches with digital face images via SSD dictionary. In *International Joint Conference on Biometrics*, pp. 1–6 (2014)
- Han, H., Klare, B.F., Bonnen, K., Jain, A.K.: Matching composite sketches to face photos: a component-based approach. *IEEE Trans. Inf. Forensics Secur.* **8**(1), 191–204 (2013)
- Liu, D., Li, J., Wang, N., Peng, C., Gao, X.: Composite components-based face sketch recognition. *Neurocomputing* **302**, 46–54 (2018)
- Mittal, P., Jain, A., Goswami, G., Vatsa, M., Singh, R.: Composite sketch recognition using saliency and attribute feedback. *Inf. Fusion* **33**, 86–99 (2017)
- Mittal, P., Vatsa, M., Singh, R.: Composite sketch recognition via deep network-a transfer learning approach. *International Conference on Biometrics*, pp. 251–256 (2015)
- Galea, C., Farrugia, R.A.: Matching software-generated sketches to face photographs with a very deep CNN, morphed faces, and transfer learning. *IEEE Trans. Inf. Forensics Secur.* **13**(6), 1421–1431 (2017)
- O.M. Parkhi, A. Vedaldi, and A. Zisserman, *Deep face recognition*. *British Machine Vision Conference (BMVC)*, 2015
- Chugh, T., Singh, M., Nagpal, S., Singh, R., Vatsa, M.: Transfer learning based evolutionary algorithm for composite face sketch recognition. *International Conference on Computer Vision and Pattern Recognition Workshops*, pp. 117–125 (2017)
- Kazemi, H., Soleymani, S., Dabouei, A., Iranmanesh, M., Nasrabadi, N.M.: Attribute-centered loss for soft-biometrics guided face sketch-photo recognition. *International Conference on Computer Vision and Pattern Recognition Workshops*, pp. 499–507 (2018)
- Wan, W., Gao, Y., Lee, H.J.: Transfer deep feature learning for face sketch recognition. *Neural Comput. Appl.* pp. 1–10 (2019)
- Peng, C., Gao, X., Wang, N., Li, J.: Sparse graphical representation based discriminant analysis for heterogeneous face recognition. *Sig. Process.* **156**, 46–61 (2019)
- Peng, C., Wang, N., Li, J., Gao, X.: DLFace: Deep local descriptor for cross-modality face recognition. *Pattern Recogn.* **90**, 131–171 (2019)
- Simonyan, K., Zisserman, A.: *Very Deep Convolutional Networks for Large-Scale Image Recognition*. (2015)
- Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. *Int. Conf. Comput. Vis. Pattern Recogn.* **1**, 886–893 (2005)
- Chen, L.F., Liao, H.Y.M., Ko, M.T., Lin, J.C., Yu, G.J.: A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recogn.* **33**(10), 1713–1726 (2000)
- Klum, S., Han, H., Klare, B., Jain, A.K.: The FaceSketchID system: Matching facial composites to mugshots. *IEEE Trans. Inf. Forens. Secur.* **9**(12), 2248–2263 (2014)
- Galea, C., Farrugia, R.A.: A large-scale software-generated face composite sketch database. *International Conference of the Biometrics Special Interest Group*, pp. 1–5 (2016)
- Martinez, A., Benavente, R.: The AR face database. *CVC, Barcelona, Spain, Tech. Rep. 24, Jun.* (1998)
- Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.* **16**(5), 295–306 (1998)
- Ouyang, S., Hospedales, T.M., Song, Y.Z., Li, X.: Forgetmenot: Memory-aware forensic facial sketch matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5571–5579, (2018)
- Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, (2015)
- FEI. <http://fei.edu.br/~cet/facedatabase.html>
- MEDS-II. <http://www.nist.gov/itl/iad/ig/sd32.cfm>
- Dlib. <http://dlib.net/>
- Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Teh, Y.W., Forsyth, D.A.: Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, p. 2 (2004)
- Xue, X., Xu, J., Mao, X.: Composite sketch recognition using multi-scale HOG features and semantic attributes. *Cyberworlds*, (2018)
- Identi-kit, identi-kit solutions. <http://www.identikit.net/>
- Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **7**, 711–720 (1997)

41. Peng, C., Gao, X., Wang, N., Li, J.: Graphical representation for heterogeneous face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(2), 301–312 (2017)
42. C. Galea, and R.A. Farrugia, Face photo-sketch recognition using local and global texture descriptors. In 2016 24th European Signal Processing Conference (EUSIPCO), pp. 2240–2244, (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Jiayi Xu received her B.S. in Computer Science from Zhejiang University, M.S. and Ph.D. in CAD&CG State Key Lab of Zhejiang University. She is currently an assistant professor at the college of Computer Science and technology, Hangzhou Dianzi University, China. Her current research interests include image processing, and pattern recognition.



Xinying Xue received the B.S. degree in software engineering from Xinyang Normal University and M.S. degree in Computer Science from Hangzhou Dianzi University in 2017 and 2020 respectively in China. Her research interests include computer vision and pattern recognition.



Yitiao Wu is currently an undergraduate student in the college of Computer Science, Hangzhou Dianzi University. His research interests include image processing.



Xiaoyang Mao received her B.S. in Computer Science from Fudan University, M.S. and Ph.D. in Computer Science from University of Tokyo. She is currently a professor at Department of Computer Science and Engineering, University of Yamanashi, Japan and also holds an adjunct professor position at the College of Computer Science, Hangzhou Dianzi University, China. She received Computer Graphics International Career Achievement Award in 2018. Her current research interests include image processing, visual perception, non-photo-realistic rendering and their applications to e-health.